DAUPHINE
UNIVERSITÉ PARIS

Executive Master

"Statistique et Big data"

Cours "Modèle linéaire gaussien".

Notes de cours.

Version du 15 juin 2021

Avant-propos

Ce cours a été écrit successivement par Laetitia Comminges et Sophie Donnet. Il est très inspiré des chapitres 1, 2, 3 et 5 du livre de Pierre-André Cornillon et Eric Matzner-Lober, **Régression avec R**, paru chez Springer.

Les excellentes références suivantes ont également été utilisées :

- **Le modèle linéaire par l'exemple**, Jean marc Azaïs, Jean-Marc Bardet, Dunod, 2005.
- `https://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf`
- **An Introduction to Statistical Learning with Applications in R**
- **Le modèle linéaire par l'exemple** de J.-M. Azais et J.-M. Bardet (Dunod)
- **Le modèle linéaire et ses extensions** de J.-J. Daudin (Ellipse)

Vous trouverez aussi de très bonnes références sur le web. En voici une sélection.

- Jeux de données pour modèle linéaire : `http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html`
- `http://www.math.univ-toulouse.fr/~azais/styles/other/student/modlin.pdf`
- `https://perso.univ-rennes2.fr/system/files/users/fromont_m/Poly_Reg.pdf`
- `https://www.agroparistech.fr/IMG/pdf/ExemplesModeleLineaire-AgroParisTech.pdf`
- Cours de C Chouqet (Toulouse) : `https://www.math.univ-toulouse.fr/~barthe/M1modlin/poly.pdf`

Le cours sera centré sur l'utilisation et la compréhension du modèle linéaire gaussien. Nous éluderons tant que possible les démonstrations mathématiques. Cependant, si vous le souhaitez, vous pouvez trouver une version plus mathématique du cours sur l'espace `mycourse` (cours des M1 dans les documets supplémentaires).

Quelques notations

Soit un vecteur $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ de $\mathbb{R}^n$. On note

- x^T le vecteur ligne correspondant.
- $\|x\|^2 = \sum_{i=1}^n x_i^2$ (la norme euclidienne).

Soit $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ un autre vecteur. On note

- $\langle x, y \rangle = \sum_{i=1}^n y_i x_i$
- I_n est la matrice identité de taille $n \times n$.

Pour un vecteur aléatoire Z, on note Σ_Z sa matrice de variance-covariance.
Dans la mesure du possible on utilisera des majuscules pour la variables aléatoires et des minuscules pour leur réalisation.

Table des matières

Chapitre 1

Préambule : motivation et exemple introductif

Ce chapitre va nous permettre de motiver le cours. Nous nous intéressons au jeu de données suivant `https://www.economicswebinstitute.org/data/wagesmicrodata.xls` tiré du Economics Web Institute.
Dans ce jeu de données, on a relevé le salaire horaire de 534 personnes (en dollars) ainsi que d'autres caractéristiques économiques telles que l'occupation (divisée en 6 catégories, 1=Management, 2=Sales, 3=Clerical , 4=Service, 5=Professional, 6=Other), le nombre d'années d'éducation, le nombre d'années d'expérience, l'âge, le sexe, le statut marital... Un extrait des données est donné dans le Tableau 1.1

ID	WAGE	OC.	SECT.	EDUC.	EXPER.	AGE	SEX	MARR	RACE	SOUTH
1	510.00	6	1	8	21	35	1	1	2	0
2	495.00	6	1	9	42	57	1	1	3	0
3	667.00	6	1	12	1	19	0	0	3	0
4	400.00	6	0	12	4	22	0	0	3	0
5	750.00	6	0	12	17	35	0	1	3	0
6	1307.00	6	0	13	9	28	0	0	3	0

TABLE 1.1 – Extrait du jeu de données "Wages"

L'objectif de cette étude (et du cours) est d'évaluer l'effet éventuel des caractéristiques sociodémographiques sur le salaire des employés.

1.1 Description des données

Avant toute étude statistique plus élaborée, il faut d'abord procéder à une étude descriptive des données. Les indices à calculer et les éventuelles représentations graphiques dépendent du type de variables considérées (qualitatives ou quantitatives).
Pour les variables quantitatives, on calculera par exemple les moyennes et écarts-types empiriques, médiane, valeurs extrémales ... Pour le jeu de données qui nous intéresse, ces statistiques sont regroupées dans la Table 1.2.

Remarque 1.1. *Pour rappel :*

Définition 1.1. *On appelle moyenne empirique et on note $\bar{x}$ la quantité*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

On appelle variance empirique de x et on note σ_x^2 la quantité

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

	MOYENNE	ECART TYPE	MINIMUN	MEDIAN	MAXIMUM
WAGE	9.20	4.96	1	8	26.29
EDUCATION	13.02	2.62	4.00	12.00	18.00
AGE	36.93	11.68	18.00	34.00	64.00
EXPERIENCE	17.82	12.38	0.00	15.00	54.00

TABLE 1.2 – Jeu de données "Wages" : statistiques pour variables quantitatives

Les variables quantitatives sont représentées graphiquement sous forme de boxplot ou histogramme sur la Figure 1.1 (à gauche).
Les variables qualitatives sont représentées sous forme de tableaux de fréquences (Tableau 1.3) ou graphiquement comme sur la Figure 1.1 (à droite).

	Modalités	Effectifs	Fréquences (%)
OCCUPATION	1	55	10.30
	2	38	7.12
	3	97	18.16
	4	83	15.54
	5	105	19.66
	6	156	29.21
SECTOR	0	411	76.97
	1	99	18.54
	2	24	4.49
SEX	0	289	54.12
	1	245	45.88

TABLE 1.3 – Jeu de données "Wages" : effectifs et fréquences pour les variables qualitatives

1.2 Régression linéaire

Maintenant qu'on a pris le jeu de données en main, on cherche à comprendre l'influence des variables quantitatives (Education, Age, Experience) sur le salaire (Wage). On trace sur la Figure 1.2 les trois nuages de points correspondants.

Covariance empirique On appelle covariance empirique et on note $\sigma_{x,y}$ la quantité

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

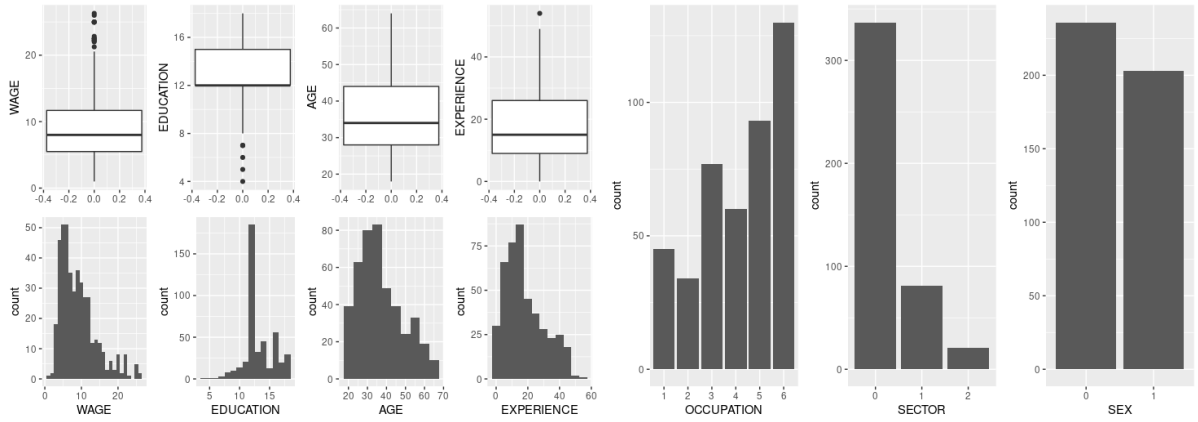


FIGURE 1.1 – Jeux de données "Wages" : Boxplots et histogrammes des variables quantitatives (gauche), diagrammes en bâtons pour variables qualitatives (droite)

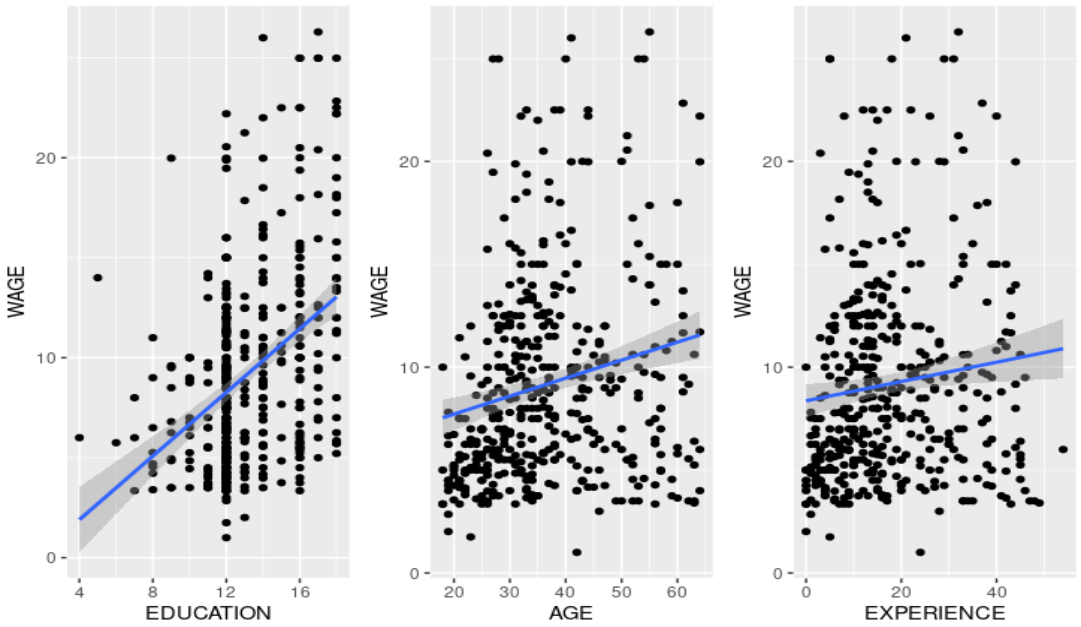


FIGURE 1.2 – Jeux de données "Wages" : Wages en fonction de l'éducation, l'âge et l'expérience.

Coefficient de corrélation linéaire Afin de quantifier la relation linéaire entre deux variables quantitatives X et Y, on peut calculer le coefficient de corrélation linéaire ρ_{XY} :

$$\rho_{XY} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i,j}(y_i - \bar{y})(x_j - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}} \tag{1.1}$$

où $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Par construction, $|\rho_{XY}| \leq 1$. Si les points sont parfaitement alignés alors $|\rho_{XY}| = 1$. Sur ce jeu de données, on trouve

ρ_{XY}	Education	Age	Experience
Wage	0.40	0.21	0.12

Question : Est-ce grand ? petit ? Significatif ?
Considérons à nouveau le nuage de points à gauche de la Figure 1.2. Si on cherche à résumer le nuage de points par une droite, appelée *droite de régression linéaire simple*, on écrira :

$$y_i \text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + e_i$$

où e_i est un terme d'erreur entre la droite de l'observation y_i.

Dans ce modèle très simple, on explique la variable d'intérêt (quantitative) par une variable quantitative dite *explicative* (ou *covariable*). La pente (β_1) et l'ordonnée à l'origine (β_0) de la droite sont *estimées* à partir des observations pour "placer" convenablement la droite. Dans ce cours, on verra comment estimer ces paramètres, quelles sont les propriétés d'un tel estimateur. Par ailleurs, on cherche à savoir si la pente est significativement différente de 0, i.e. on va chercher à écrire des tests sur les paramètres du modèles. Ce modèle sera traité au Chapitre 2.
On pourra aussi chercher à expliquer le salaire comme une combinaison linéaires des autres variables quantitatives :

$$\begin{aligned} \text{Wage}_i &= \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Age}_i + \beta_3 \text{ Experience}_i + e_i \\ &= \beta_0 + \sum_{k=1}^K \beta_k x_i^k + e_i \end{aligned}$$

où x_i^k est la valeur de la k-ième variable explicative de l'individu i. On parlera alors de *régression linéaire multiple* (voir Chapitre 3). Les mêmes questions que précédemment se posent : significativité des β_k, sélection des variables explicatives les plus pertinentes, etc.

1.3 Analyse de la variance

Analyse de la variance à un facteur Il peut être aussi intéressant d'étudier la relation entre le salaire (variable d'intérêt quantitative) et les variables qualitatives, par exemple le sexe, ou l'occupation (type de métier). Graphiquement, on pourra tracer des boxplots par modalité de la variable qualitative comme sur la Figure 1.3 De façon naturelle, pour comparer les salaires au sein des différentes populations, on cherchera à comparer les moyennes au sein des groupes :

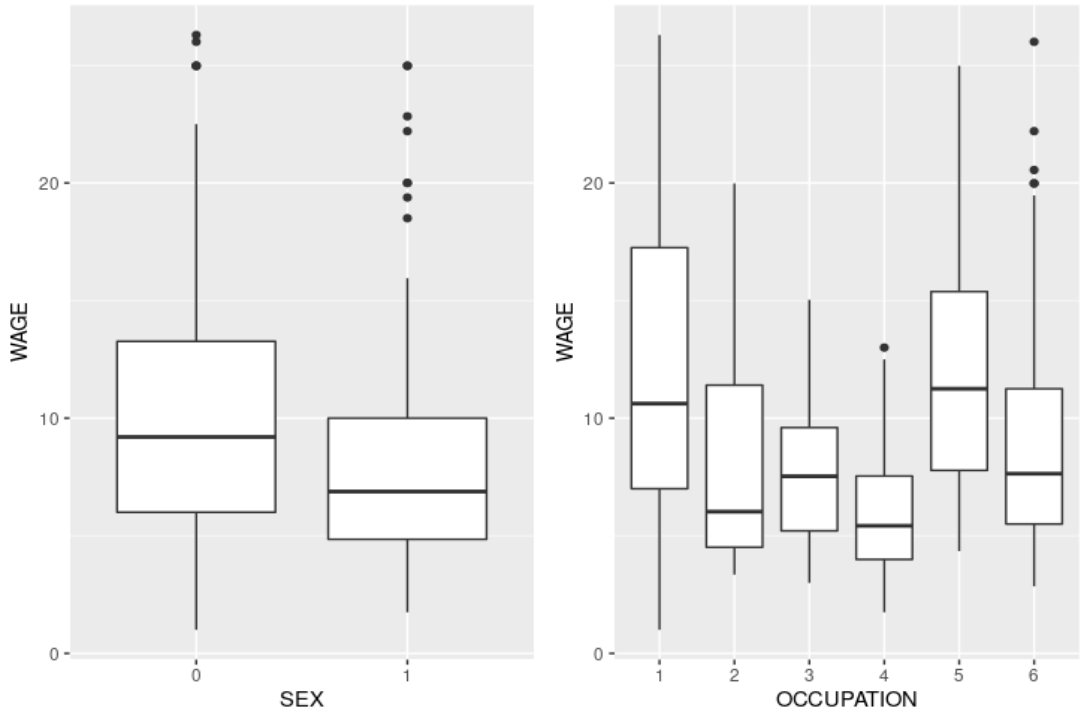


FIGURE 1.3 – Jeux de données "Wages" : Wages en fonction de du sexe (0=Homme) ou de l'occupation.

Occupation	1	2	3	4	5	6
Moyenne par Occupation	12.54	7.91	7.59	5.92	12.06	8.80

Des tests de comparaison de moyenne pourront alors être appliqués. Cependant, il est en fait possible d'écrire un modèle linéaire permettant d'étudier le salaire en fonction du sexe ou de l'occupation :

$$\text{Wage}_i = \underbrace{\beta_0}_{\text{Salaire des Hommes}} \mathbb{1}_{\text{Sexe}_i=0} + \underbrace{\beta_1}_{\text{Salaire des Femmes}} \mathbb{1}_{\text{Sexe}_i=1} + e_i$$

$$\text{Wage}_i = \sum_{l=1}^{l} \underbrace{\beta_l}_{\text{Salaire des occupations } l} \mathbb{1}_{\text{Occupation}=l} + e_i$$

Ce modèle est le *modèle d'analyse de variance à un facteur* (Anova) puisqu'on cherche à comprendre à la variation du salaire en fonction d'un facteur (sexe, ou bien l'occupation).

Analyse de la variance à deux facteurs On peut se demander si il n'y a pas un effet conjoint des deux facteurs. On cherchera alors à croiser les facteurs, en calculant par exemple les salaires moyens comme suit :

On cherchera à étudier l'influence de chaque facteur sur les salaires mais aussi leurs influences conjointes (éventuelles interactions). Nous écrirons un modèle linéaire.

Sexe	0	1
Moyenne par sexe	10.33	7.88

	Occupation	1	2	3	4	5	6	Moyenne par Sexe
Sex	0	14.42	9.72	7.95	6.12	12.78	9.42	10.32
	1	9.45	5.32	7.51	5.81	11.32	6.05	7.88
Moyenne par	Occupation	12.54	7.91	7.59	5.92	12.06	8.80	9.20

1.4 Analyse de la covariance

On peut penser que le lien entre salaire et éducation n'est pas le même selon que l'on est un homme ou une femme. On voudra alors écrire un modèle de régression pour chaque sexe (voir Figure 1.4. De la même façon, nous verrons qu'il est possible d'écrire un modèle linéaire répondant à ce besoin (voir Chapitre 4).

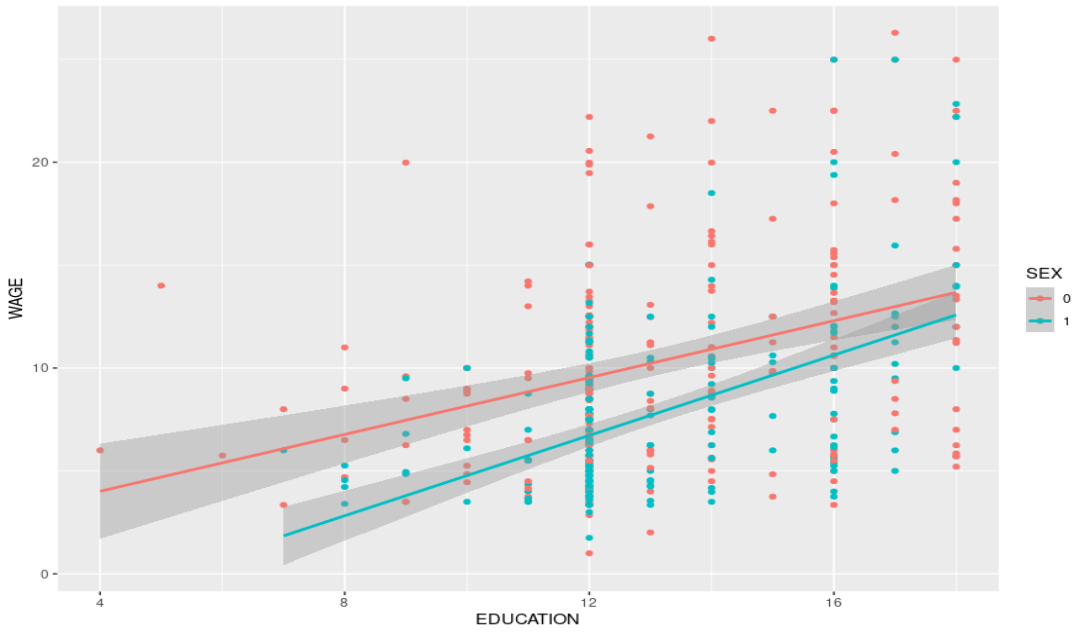


FIGURE 1.4 – Jeux de données “Wages” : Wages en fonction de l’éducation et du sexe

Chapitre 2

La régression linéaire simple

Nous nous intéressons d'abord au cas où nous voulons comprendre la relation entre une variable d'intérêt y et une variable explicative x. On suppose que les deux variables sont quantitatives.

2.1 Un exemple simple

Nous allons commencer par un exemple concret et très simple. Imaginons que nous nous intéressions au lien entre le poids d'une femme et sa taille. Le poids d'une personne semble naturellement être une fonction croissante de sa taille. On trouve dans le jeu de données women (disponible sous R) les valeurs du poids (en lbs) et de la taille (en in) de 15 femmes :

```
str(women)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

```
summary(women)
```

```
##      height         weight
##  Min.   :58.0   Min.   :115.0
##  1st Qu.:61.5   1st Qu.:124.5
##  Median :65.0   Median :135.0
##  Mean   :65.0   Mean   :136.7
##  3rd Qu.:68.5   3rd Qu.:148.0
##  Max.   :72.0   Max.   :164.0
```

Notre objectif est :

- d'expliquer le lien entre la variable poids et la variable taille, autrement dit, expliciter la fonction f qui lie les deux variables par `poids`$=f($`taille`$)$.
- prédire le poids d'un nouvel individu à partir de sa taille.

Avant toute analyse, il est intéressant de représenter les données, ce qui est fait dans la Figure 2.1. Cette figure a été obtenue avec le code suivant :

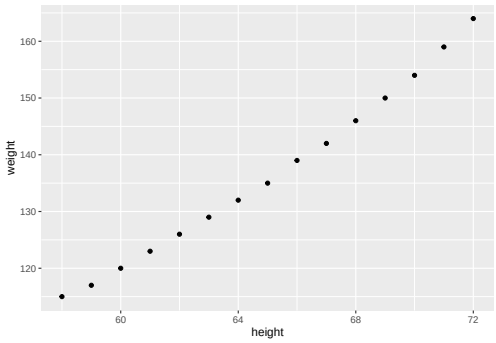


FIGURE 2.1 – poids et taille de 15 femmes

Code R

```
library(ggplot2)
ggplot(women) + geom_point(aes(x = height, y = weight))
```

D'après ce graphique, il semble clairement y avoir un lien linéaire entre le poids et la taille : la représentation graphique est quasiment une droite. Plus précisément si on note y_i et x_i le poids et la taille de la femme numéro i, il existe un couple (a, b) de réels tels que pour $1 \leq i \leq 15$:

$$y_i \approx ax_i + b$$

($y = ax + b$ est l'équation d'une droite approchant le graphique).

Démarche générale De manière générale, en statistique, on cherche souvent à expliquer, ou prédire, une variable d'intérêt qu'on note y, ici le poids, en fonction d'une autre variable notée x, ici la taille. On dit que y est la **variable à expliquer** et x la **variable explicative**.

On dispose pour cela d'un ensemble de n valeurs de ces variables $(x_1, \ldots x_n)$ et $(y_1, \ldots, y_n)$ mesurées sur n individus (ici n=15) pour *apprendre* la relation en x et y.

On ne sait en général pas à l'avance si x explique correctement y (x est-elle en particulier suffisante pour prévoir y ?). Et même si x explique bien y, c'est-à-dire si on a $y \approx f(x)$ pour une certaine fonction f, on ne connaît pas à l'avance la forme de la fonction f qui est censée lier y à x. On peut essayer de la "deviner" ou d'avoir une première idée en regardant le graphique quand on n'a qu'une seule variable explicative comme ici (infaisable avec beaucoup de variables explicatives). Parfois on a des connaissances a priori sur le domaine qui nous guident pour le choix de la "forme" de f (un polynôme ? une fonction sinusoïdale ? etc). Souvent c'est l'expérience d'un type de données qui nous aide. Quoiqu'il en soit, **si on n'a qu'une seule variable explicative comme ici, on commence toujours par une représentation graphique pour avoir une idée de la "forme" de cette fonction f.**

Estimation Dans ce chapitre, qui concerne ce qu'on appelle la *régression linéaire simple*, on suppose une forme très simple pour f. On suppose que f est une fonction affine

$$f(x) = ax + b.$$

Il faudra donc estimer, dans le contexte de la régression linéaire simple, deux réels a et b. Ce modèle peut sembler simplet au premier abord, surtout en comparaison de modèles plus complexes et sophistiqués que vous verrez certainement plus tard. Cependant, les modèles les plus simples sont dans certains cas ceux qui fonctionnent le mieux, et bien qu'étant une "vieille méthode", la régression linéaire est encore largement utilisée. De plus, beaucoup de méthodes plus sophistiquées sont des extensions ou des généralisations de cette méthode.

Afin de quantifier précisément le symbole $\approx$ dans la formulation $y \approx f(x)$, nous nous donnons une fonction de perte (ou fonction de coût) ℓ et nous cherchons la fonction affine f qui minimise

$$\sum_{i=1}^{n} \ell(y_i - f(x_i))$$

Nous cherchons donc une droite, qui soit "la plus proche possible du nuage de points", cette proximité étant mesurée par la fonction ℓ.
De nombreuses fonctions de coût existent, mais les deux principalement utilisées sont :

- le coût quadratique : $\ell(u) = u^2$
- le coût absolu : $\ell(u) = |u|$.

Ces deux fonctions de coût sont positives, nulles en zéro et symétriques.

A propos des fonctions de perte ℓ Il est évident que, par rapport au coût absolu, le coût quadratique accorde plus de poids aux points qui restent éloignés de la droite ajustée, la distance étant élevée au carré. En conséquence, le coût quadratique accorde plus d'importance aux points aberrants. Des observations aberrantes sont des observations peu fréquentes et très différentes du reste des observations de l'échantillon. Elles peuvent être dues à des erreurs de mesure ou à une autre anomalie. Par exemple, si on modifie une donnée dans le jeu de données précédent, en augmentant fortement la valeur de la variable poids et en gardant la même valeur pour la variable taille, on aura une observation aberrante, et le point correspondant ne sera pas aligné avec les autres points. La droite ajustée en minimisant le coût quadratique sera très différente de la droite initiale, alors que ce sera beaucoup moins le cas pour la droite ajustée par le coût absolu. On dit que le coût absolu est robuste.
Malgré cette robustesse du coût absolu, on utilise beaucoup plus le coût quadratique et ce pour plusieurs raisons : historique, calculabilité, propriétés mathématiques. Et c'est cette fonction de coût que l'on utilise dans ce cours. On parle alors d'**estimation par les moindres carrés** (least squares, LS).

Propriétés des estimations Une fois les paramètres a et b estimés sur nos données dites d'apprentissage, on peut se demander ce qu'il se serait passé si on avait pris un autre échantillon : aurait-on observé de grandes variations dans nos estimations ? Si oui, quelle validité donner à nos résultats ? Pour répondre à ces questions, la démarche statistique est de supposer que les observations sont la réalisation d'une variable aléatoire dont on spécifie (contrôle) la loi de probabilité, puis d'étudier les variations de notre estimation sous cette hypothèse. On va donc étudier les propriétés probabilistes de notre estimateur. On pourra ainsi avoir une idée de l'incertitude que l'on a sur nos valeurs estimées de a et b.

2.2 Modélisation statistique

Modéliser On dit que l'on *modélise les observations* $\boldsymbol{y} = (y_1, \ldots, y_n)$ quand on fait l'hypothèse (on décide) que $\boldsymbol{y}$ est la réalisation d'une variable aléatoire $\boldsymbol{Y}$ dont on décrit la loi de probabilité. Dans ce cours, on s'intéresse au *modèle linéaire*, spécifiant une relation affine entre la variable observée y et la variable explicative x à laquelle s'ajoute des termes d'erreur. *Définir le modèle consiste à spécifier la loi de probabilité suivie par les erreurs.*

En d'autres termes, on cherche à comprendre le lien entre la variable d'intérêt y la variable explicative x. On suppose que ce lien est affine :

$$y \approx ax + b$$

pour un certain couple (a, b) de réels.
Concrètement, nous avons un échantillon $((x_1, y_1), \ldots, (x_n, y_n))$ d'observations des deux variables sur un nombre n d'individus. La représentation graphique de ce nuage de points donnera rarement exactement une droite : le nuage pourrait ressembler à la Figure 2.2, qui est un exemple plus réaliste. En effet, d'une part la variable y n'est pas forcément expliquée uniquement par la variable x. D'autre part, les mesures effectuées dépendent de la précision de l'appareil de mesure, de l'opérateur et il arrive souvent que, pour des valeurs identiques de la variable x, nous observions des valeurs différentes pour y.

Nous modélisons les observations de la façon suivante. $\forall i = 1, \ldots, n$, **nous supposons que** y_i **est la réalisation d'une variable aléatoire** Y_i **telle que** :

$$Y_i = ax_i + b + \varepsilon_i.$$

où ε_i (bruit ou **erreur**) est une variable aléatoire dont on va décrire la loi de probabilité. Cette équation est appelée **modèle de régression linéaire simple**.

A propos des termes d'erreurs $\varepsilon_1, \ldots, \varepsilon_n$ On suppose que les termes d'erreur $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ sont d'espérance nulle, indépendants et de variance constante σ^2

$$\mathbb{E}[\varepsilon_i] = 0 \quad \text{et} \quad \mathbb{V}[e_i] = \sigma^2$$

Dans le cas particulier du **modèle linéaire gaussien**, les termes d'erreur $\boldsymbol{\varepsilon}$ sont distribués selon la loi d'un vecteur gaussien de taille n, d'espérance nulle et de matrice de variance-covariance diagonale :

$$\boldsymbol{\varepsilon}_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

Postulats La loi sur les $\boldsymbol{\varepsilon}$ implique plusieurs postulats concernant les observations.

- [P1] Les erreurs sont centrées : $\mathbb{E}[\boldsymbol{\varepsilon}] = 0_{\mathbb{R}}$. En pratique, cela veut dire que le modèle est correct et que l'on n'a pas oublié un terme pertinent. Un exemple d'erreurs centrées et non centrées est donné sur la Figure 2.2.
- [P2] Les erreurs $\boldsymbol{\varepsilon}$ sont de variance constante : $\mathbb{V}[\varepsilon_i] = \sigma^2, \forall i = 1 \ldots n$. On parle de modèle *homoscédastique*, par opposition à un modèle *hétéroscédastique* où le terme d'erreur n'aurait pas la même variance pour toutes les observations. Un exemple d'erreurs homoscédastiques et hétéroscédastiques est donné sur la Figure 2.3. Sur la première ligne,

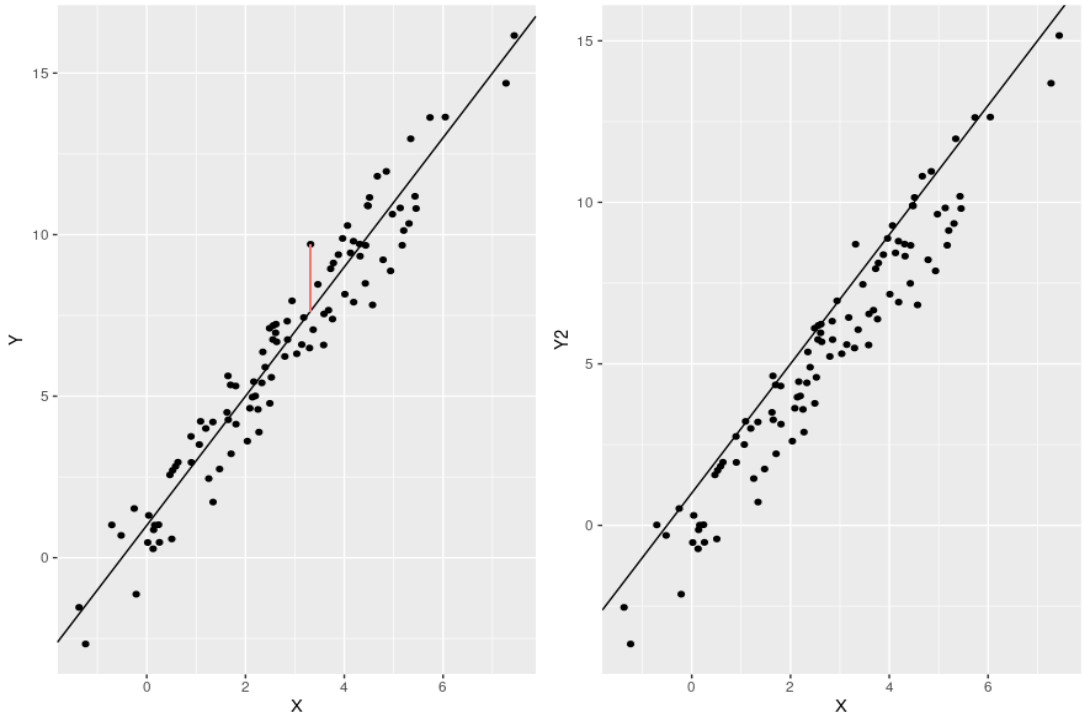


FIGURE 2.2 – Erreurs centrées (à gauche) et non centrées (à droite)

la variance des erreurs en constante (petite à gauche, grande à droite). Sur la ligne du bas à gauche, il y a deux tailles de variance d'erreur : certains points ont une erreur de variance petite, d'autres de variance grande. Ligne du bas à droite, la variance des erreurs augmente avec x.

- [P3] Les termes d'erreur sont supposés indépendants. Ainsi, les observations sont supposées indépendantes, i.e. correspondent à un échantillonnage indépendant ou aux résultats d'une expérience physique menée dans des conditions indépendantes. Des problèmes peuvent surgir quand le temps a une importance dans le phénomène. On constate sur la Figure 2.4 que les erreurs sont globalement centrées mais elles n'oscillent pas en permanence de part d'autre de 0 : il semble qu'elles s'attardent d'un côté de 0 puis de l'autre.
- [P4] Finalement, les erreurs sont supposées gaussiennes. Ce postulat est le moins important dans la mesure où on pourra s'en passer si le nombre d'observations est grand (au delà de 20 ou 30 observations). Il est difficile de détecter la non-gaussianité des erreurs. Nous verrons que **R** propose des outils graphiques pour tenter de valider ou non cette hypothèse.

Remarque 2.1. *Notez que l'on parle de* postulats *en ce sens que nous ne pouvons pas formellement montrer qu'ils sont vérifiés par des tests statistiques. Nous allons utiliser des outils graphiques pour les tester.*

Paramètres du modèle a et b sont appelés les paramètres du modèle ou **coefficients** du modèle (coefficient de régression et constante de régression ou intercept), ils sont fixes et inconnus. Ce sont ces paramètres que nous voulons estimer.

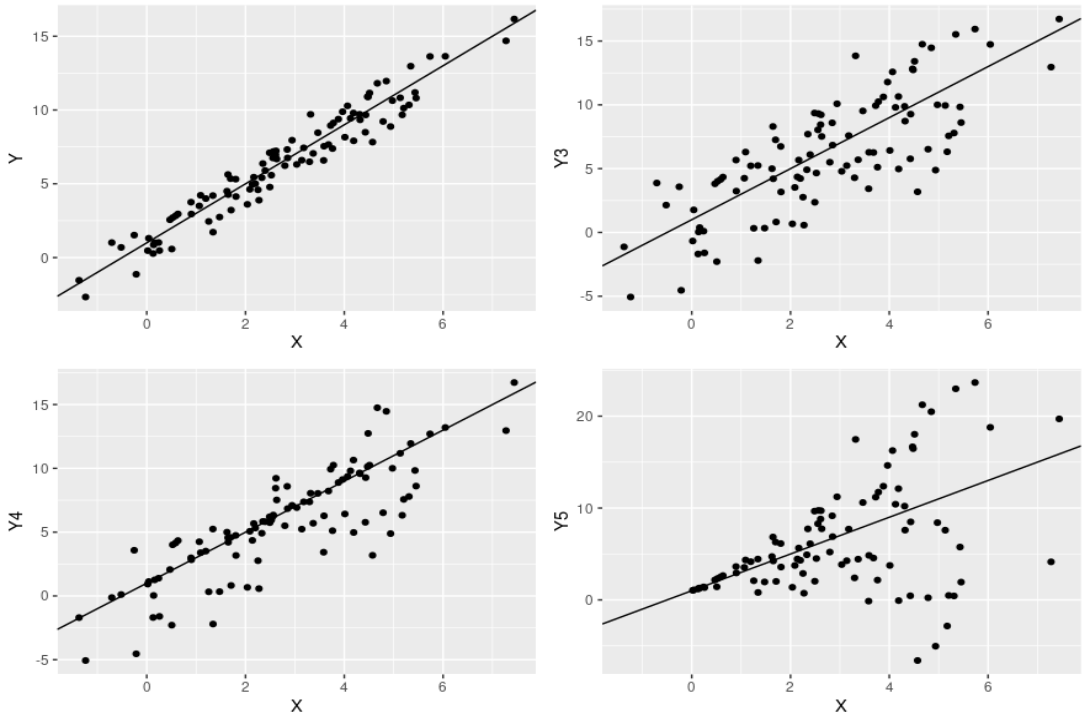


FIGURE 2.3 – ε de variance constante en haut, de variance variable en bas

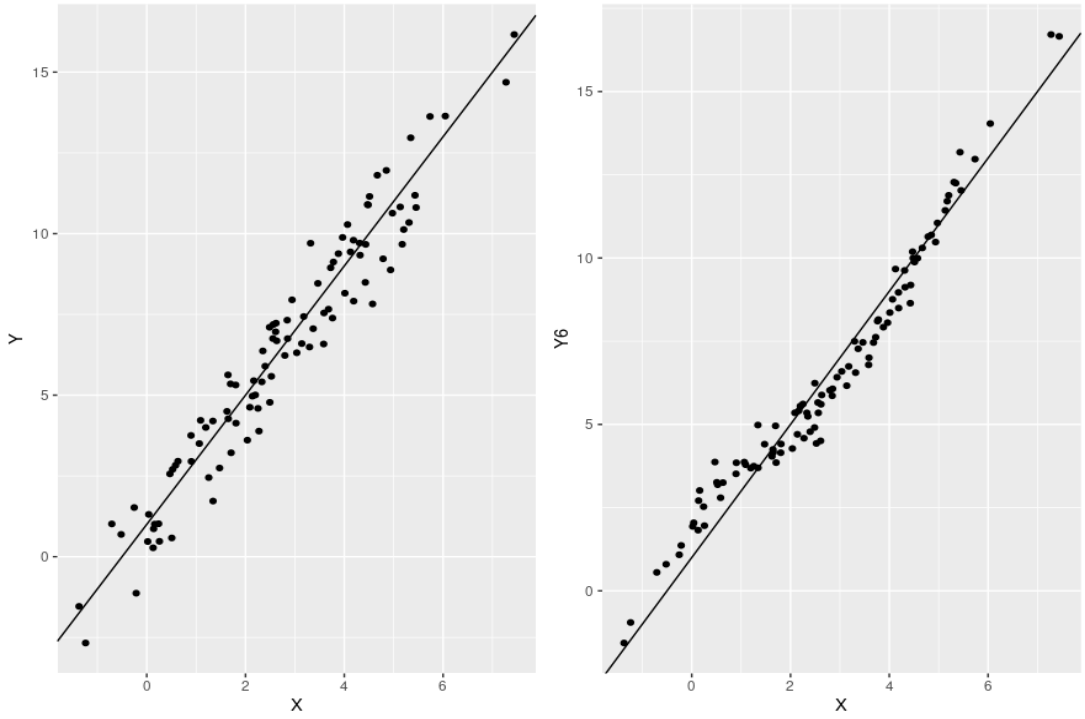


FIGURE 2.4 – Erreurs indépendantes à gauche, dépendantes à droite

Résumé du modèle :
On suppose que pour tout i, y_i est la réalisation de la variable aléatoire Y_i telle que

$$Y_i = ax_i + b + \varepsilon_i, i = 1, \dots, n$$

avec :
- les x_i sont connus et non aléatoires.
- a et b sont inconnus et non aléatoires.
- les ε_i sont aléatoires de loi $\mathcal{N}(0, \sigma^2)$.
- les Y_i sont en conséquence aussi gaussiens : $Y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$.

On va utiliser dans la suite la notation suivante

$$\beta = \begin{pmatrix} b \\ a \end{pmatrix}$$

On note ici EMC pour estimateur des moindres carrés. Il existe des variations sur cet estimateur, on trouve donc aussi dans la littérature la dénomination de "moindres carrés ordinaire" (MCO). En anglais on rencontre "ordinary least-squares", en abrégé "LS" ou "OLS".
Nous allons voir dans la prochaine section comment calculer des estimations de (a, b).

2.3 Estimation des moindres carrés (EMC)

2.3.1 Définition

On appelle *estimation des moindres carrés de a et b* les valeurs $\hat{a}$ et $\hat{b}$ obtenues par minimisation de la quantité :

$$S(a, b) = \sum_{i=1}^{n} (y_i - b - ax_i)^2$$

Ce que l'on note habituellement par

$$(\hat{a}, \hat{b}) = \underset{(a,b)\in\mathbb{R}\times\mathbb{R}}{\arg\min} \sum_{i=1}^{n} (y_i - b - ax_i)^2 \tag{2.1}$$

Remarque 2.2. *On peut utiliser l'écriture matricielle suivante.*
Notons

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_n \end{pmatrix}$$

Avec la notation $\boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}$, *on peut écrire*

$$\begin{pmatrix} ax_1 + b \\ \vdots \\ ax_n + b \end{pmatrix} = b\mathbf{1} + a\boldsymbol{x} = X\boldsymbol{\beta}$$

et

$$(\hat{a}, \hat{b}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$$

2.3.2 Calcul de l'estimation

Théorème 2.1. *Supposons qu'il existe au moins deux points d'abscisses différentes, i.e. il existe un couple (i,j) tel que $x_i \neq x_j$. Alors il existe une solution unique au problème de minimisation (2.1) dont l'expression est donnée par :*

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{2.2}$$

et

$$\hat{b} = \bar{y} - \hat{a}\bar{x} \tag{2.3}$$

Démonstration. Soit $f : (a,b) \to \sum_{i=1}^n (y_i - b - ax_i)^2$.
Remarquons tout de suite que la condition de l'énoncé "les x_i ne sont pas tous égaux" implique aussi qu'il existe au moins un x_i tel que $\bar{x} \neq x_i$. On note $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$
On cherche un point critique de f :

$$\left\{\begin{array}{l} \frac{\partial f}{\partial b}=0 \\ \frac{\partial f}{\partial a}=0 \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} \sum_{i=1}^n (y_i - ax_i - b) \ =0 \\ \sum_{i=1}^n x_i(y_i - ax_i - b)=0 \end{array}\right. \Leftrightarrow \left\{\begin{array}{rl} \bar{y} - a\bar{x} & =b \\ \sum_{i=1}^n x_i(y_i - \bar{y} - a(x_i - \bar{x}))=0 & \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} b= \ \bar{y} - a\bar{x} \\ a=\frac{\sum_{i=1}^n x_i(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} \end{array}\right.$$

On utilise ensuite le fait que $\sum_{i=1}^n (y_i - \bar{y}) = n\bar{y} - n\bar{y} = 0$ donc

$$\sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n x_i(y_i - \bar{y}) - \bar{x}\sum(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

On admet que la fonction est strictement convexe. Donc il s'agit bien de l'unique minimum. □

On peut noter ces formules de manière plus compacte en introduisant les moyennes, variances et covariances empiriques.
Avec les notations des quantités empiriques, l'EMC $(\hat{a}, \hat{b})$ peut s'écrire

$$\hat{a} = \frac{\sigma_{x,y}}{\sigma_x^2} \quad \text{et} \quad \hat{b} = \bar{y} - a\bar{x} \tag{2.4}$$

avec

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$$

Calcul de l'estimation sous R sur un exemple : hauteur des eucalyptus Lorsqu'un forestier essaie de quantifier le volume de bois fourni par un arbre, il est nécessaire de connaître sa hauteur. Or il est parfois impossible (ou trop long) d'effectuer une telle mesure. Une mesure plus simple est la mesure de la circonférence de l'arbre à une hauteur fixée du sol. Le forestier souhaite trouver une formule, si celle-ci existe, permettant de déduire la hauteur de l'arbre à partir de sa circonférence. Pour cela il dispose d'un ensemble de $n = 1429$ couples de mesures circonférence-hauteur effectuées sur n arbres.
Ces données sont contenues dans le fichier texte "eucalyptus.txt". Les noms des variables sont `ht` et `circ`. On commence par importer les données dans un dataframe :

```
euca<-read.table("eucalyptus.txt",header=T,sep=" ")
str(euca)
```

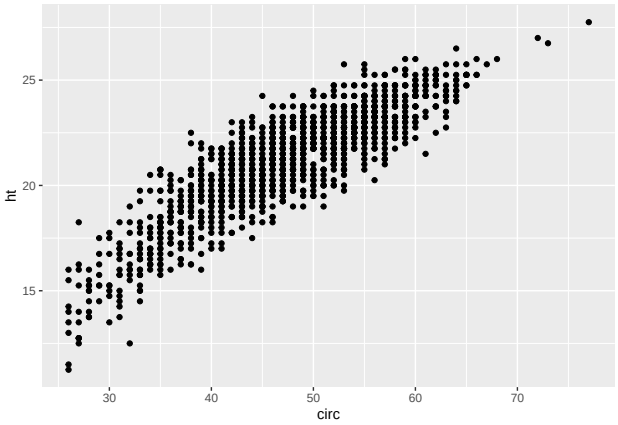


FIGURE 2.5 – Jeu de données *Eucalyptus* : représentation hauteur versus circonférence pour les 1429 eucalyptus mesurés

```
## 'data.frame':    1429 obs. of  3 variables:
##  $ ht  : num  18.2 19.8 16.5 18.2 19.5 ...
##  $ circ: int  36 42 33 39 43 34 37 41 27 30 ...
##  $ bloc: Factor w/ 3 levels "A1","A2","A3": 1 1 1 1 1 1 1 1 1 1 ...
```

```
names(euca)
```

```
## [1] "ht"   "circ" "bloc"
```

• Pour une régression simple comme ici (i.e. une seule variable explicative), on commence toujours par représenter le nuage de points (Figure 2.5). Cela nous permet de savoir qu'une régression simple semble indiquée, les points étant disposés grossièrement le long d'une droite.

```
library(ggplot2)
gg_euca <- ggplot(euca, aes(x = circ, y = ht)) + geom_point()
gg_euca
```

• Nous estimons ensuite les coefficients. Ceux-ci sont calculés, entre autres, par la fonction `lm` (pour linear model) :

Code : fonction lm

```
reg <- lm(ht~circ,data=euca)
```

Afin de consulter un résumé des résultats liés à cette régression linéaire nous effectuons :

```
summary(reg)
```

```
##
## Call:
## lm(formula = ht ~ circ, data = euca)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7659 -0.7802  0.0557  0.8271  3.6913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 1427 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.7682
## F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Cette sortie contient un certain nombre d'informations que l'on va expliquer au fur et à mesure dans ce chapitre. Nous pouvons consulter la liste des différents résultats de l'objet `reg` avec :

```
names(reg)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

Pour le moment, nous voulons l'estimation de $(\hat{a}, \hat{b})$: il se trouve dans la colonne `Estimate` du tableau `Coefficients` donné par la fonction `summary`. La première ligne, `Intercept`, correspond au coefficient $\hat{b}$ (la constante de régression) et la seconde au coefficient $\hat{a}$ de circ. On peut donc récupérer les coefficients avec la commande `reg$coefficients` ou `coef(reg)`.

Code : récupération des coefficients

```
coef(reg)
```

```
## (Intercept)        circ
##   9.0374757   0.2571379
```

• Nous pouvons alors tracer la droite de régression sur le graphique du nuage de points (Figure 2.6) :

Code : tracé de la droite de régression

```
gg_euca + geom_abline(intercept=coef(reg)[1], slope = coef(reg)[2])
```

Remarque 2.3. *On aurait pu avoir la même figure avec :*

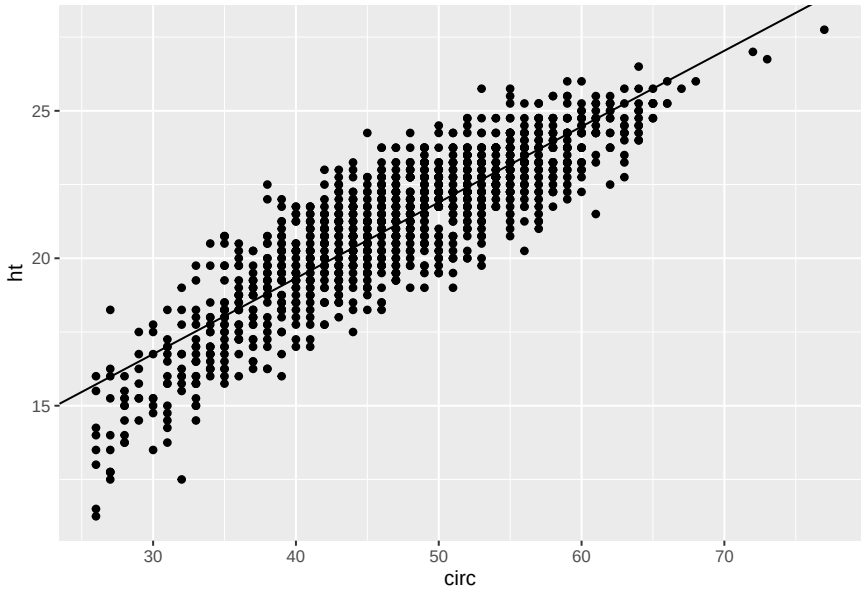


FIGURE 2.6 – Représentation hauteur versus circonférence pour les 1429 eucalyptus mesurés

```
gg_euca + geom_smooth(method='lm')
```

2.3.3 Propriétés de l'estimateur de (a, b)

De l'estimation à l'estimateur On a supposé que y_i est la réalisation d'une variable aléatoire donc on pense que si l'on refait la même expérience en faisant des mesures des variables x et y sur n autres individus, en prenant exactement les mêmes x_i, on aura des valeurs différentes pour les réalisations y_i et donc des estimations. On va donc s'intéresser aux variations de ces estimations. A partir de $(\hat{a}, \hat{b})$ est l'estimation des moindres carrés, on peut en déduire l'estimateur (où l'on remplace la réalisation y_i par la variable aléatoire Y_i :

$$\begin{aligned} \hat{A} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{B} &= \bar{Y} - \hat{A}\bar{x} \\ \text{avec } \bar{Y} &= \frac{\sum_{i=1}^n Y_i}{n} \end{aligned}$$

Etudier les propriétés de l'estimateur revient à étudier les variations aléatoires de l'estimation, i.e. si j'avais pris d'autres observations de la même expérience, aurais-je obtenu une estimation de a et b très différente. Pour cela on calculera la loi, l'espérance et la variance de notre estimateur $(\hat{A}, \hat{B})$.

Distribution de probabilité de l'estimateur $(\hat{A}, \hat{B})$

Théorème 2.2. *On pose* $\hat{\beta} = \begin{pmatrix} \hat{B} \\ \hat{A} \end{pmatrix}$ *l'EMC (estimateur des moindres carrés) de* $\beta = \begin{pmatrix} b \\ a \end{pmatrix}$. *On*

a

$$\hat{\beta} \sim \mathcal{N}_2(\beta, \sigma^2 V)$$

où

$$V = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Démonstration. • On peut d'abord remarquer que le vecteur $\hat{\beta} = (\hat{B}, \hat{A})^T$ est un estimateur linéaire, c'est-à-dire qu'il s'écrit comme une fonction linéaire du vecteur y :

$$\hat{A} = \sum_{i=1}^n \lambda_{i,1} Y_i \quad \text{et} \quad \hat{B} = \sum_{i=1}^n \lambda_{i,2} Y_i,$$

pour certaines constantes $\lambda_{i,1}, \lambda_{i,2}$ (constantes dépendant du vecteur x uniquement). Par exemple $\lambda_{i,1} = \frac{(x_i - \bar{x})(1 - \frac{1}{n})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

• Or on peut écrire $\boldsymbol{Y}$

$$\boldsymbol{Y} = ax + b\mathbf{1} + \varepsilon$$

où $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$. Le vecteur $ax + b\mathbf{1}$ est déterministe. Le vecteur ε ayant ses composantes iid de loi normale standard (i.e. de loi $\mathcal{N}(0, \sigma^2)$), c'est un vecteur gaussien standard i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ et donc

$$\boldsymbol{Y} \sim \mathcal{N}(ax + b\mathbf{1}, \sigma^2 I_n)$$

Le vecteur $\boldsymbol{Y}$ est donc bien gaussien.

• En conséquence, le vecteur $\hat{\beta}$ est aussi gaussien. Pour connaître complètement sa loi, il suffit de calculer son espérance et sa variance.
Calculons son espérance :

$$\mathbb{E}_\beta(\hat{A}) = \mathbb{E}_\beta\Big(\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}_\beta(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Or $Y_i = ax_i + b + \varepsilon_i$ donc $\mathbb{E}_\beta(Y_i) = ax_i + b$ et $\mathbb{E}_\beta(\bar{Y}) = \frac{1}{n}\sum_{i=1}^n (ax_i + b) = a\bar{x} + b$. Donc on a

$$\mathbb{E}_\beta(Y_i - \bar{Y}) = a(x_i - \bar{x})$$

ainsi

$$\mathbb{E}_\beta(\hat{a}) = \frac{\sum_{i=1}^n (x_i - \bar{x}) a (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = a$$

Et

$$\mathbb{E}_\beta(\hat{b}) = \mathbb{E}_\beta(\bar{Y} - \hat{a}\bar{x}) = \mathbb{E}_\beta(\bar{Y}) - \bar{x}\mathbb{E}_\beta(\hat{a}) = a\bar{x} + b - a\bar{x} = b$$

Le calcul de la matrice de variance-covariance V de cet estimateur est un peu fastidieux et admis. □

Biais et variance de $(\hat{A}, \hat{B})$ Rappelons que, pour mesurer la performance d'un estimateur, on peut utiliser la perte quadratique. C'est d'ailleurs la perte que nous avons utilisée pour définir l'EMC. Nous allons utiliser aussi cette perte pour mesurer la performance de l'estimateur. Rappelons la décomposition biais-variance vue dans le cours de statistique :

$$\mathbb{E}_\beta[(\hat{A} - a)^2] = \text{biais au carré de } a + \text{ variance de } a = [\mathbb{E}_\beta(\hat{A}) - a]^2 + \mathbf{Var}_\beta(A)$$

$$\mathbb{E}_\beta[\hat{B} - b)^2] = [\mathbb{E}_\beta(\hat{B}) - b]^2 + \mathbf{Var}_\beta(B)$$

Corollaire 2.1. *l'EMC* $\hat{\beta}$ *est* ***sans biais***

$$\mathbb{E}_\beta(\hat{\beta}) = \beta$$

Autrement dit $\mathbb{E}_\beta(\hat{A}) = a$ *et* $\mathbb{E}_\beta(\hat{B}) = b$

Corollaire 2.2. *Ses composantes, c'est-à-dire l'estimateur du coefficient* $\hat{A}$ *et l'estimateur de l'intercept* $\hat{B}$*, sont gaussiennes.*
Les variances et covariance de $\hat{A}$ *et* $\hat{B}$ *sont données par*

$$\mathbf{Var}_\beta(\hat{B}) = \sigma^2 \frac{\sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\mathbf{Var}_\beta(\hat{A}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\mathbf{Cov}_\beta(\hat{A}, \hat{B}) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Remarque 2.4. *En examinant les variances, on a une idée de la précision de l'EMC. On rappelle que plus la variance est faible, plus l'estimateur est précis. Pour avoir des variances petites, il faut avoir un numérateur petit et/ou un dénominateur grand. Les estimateurs seront donc de faible variance lorsque :*

- *la variance* σ^2 *est faible, cela signifie que le bruit est faible.*
- *la quantité* $\sum_{i=1}^n (x_i - \bar{x})^2$ *est grande : cela correspond à la variance empirique (multipliée par n). Les mesures* x_i *doivent être assez dispersées autour de la moyenne.*
- *la quantité* $\sum_{i=1}^n x_i^2$ *ne doit pas être trop grande.*

Remarque 2.5. *L'équation* $\hat{b} = \bar{y} - \hat{a}\bar{x}$ *se réécrit* $\bar{y} = \hat{a}\bar{x} + \hat{b}$ *et donc le point* $(\bar{x}, \bar{y})$ *est sur la droite de régression.*

2.4 Résidus et estimation de σ^2

Nous avons estimé jusqu'à présent les paramètres a et b : il reste à estimer le paramètre σ^2 inconnu lui aussi. Pour cela nous allons utiliser les résidus : ce sont des estimateurs des erreurs inconnues ε_i.

Définition 2.1. *On appelle valeur ajustée de* y_i *par le modèle et on note* $\hat{y}_i$ *la quantité*

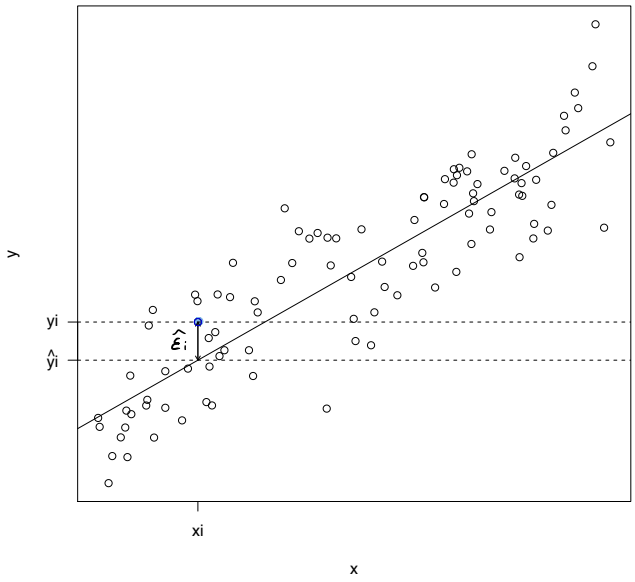


FIGURE 2.7 – Illustration des résidus

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

C'est donc l'ordonnée du point de la droite des moindres carrés correspondant à l'abscisse x_i (voir Figure 2.7). Pour l'exemple des eucalyptus, c'est la valeur de la hauteur que prédirait notre modèle quand la circonférence vaut x_i.

Code R : valeurs ajustées

On trouve les valeurs ajustées dans le vecteur `reg$fitted.values` ou plus simplement avec la commande `fitted(reg)`.

Définition 2.2. *(Résidus) Les résidus sont définis par*

$$\hat{e}_i = y_i - \hat{y}_i.$$

Code R : résidus

Dans la sortie de `summary(reg)`, on a un premier tableau appelé `Residuals` qui donnent des statistiques élémentaires sur le vecteur des résidus $\hat{\varepsilon}$ (minimum, médiane, etc). On peut trouver le vecteur des résidus avec la commande `reg$residuals` ou plus simplement avec `resid(reg)`.

Propriétés probabilistes des résidus De la même façon que précédemment, on définit la version "variable aléatoire" des valeurs ajustées et résidus.

$$\hat{Y}_i = \hat{A}x_i + \hat{B}, \quad \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

et on s'intéresse à la loi de ces quantités.
Remarquez que le vecteur $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$ est gaussien. En effet :

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{A}x_i + \hat{B}).$$

Or $\hat{A}$ et $\hat{B}$ sont des transformées linéaires du vecteur gaussien Y donc $\hat{Y}_i$ est transformée linéaire du vecteur gaussien Y, il est donc gaussien. On a admet la propriété suivante

les vecteurs $\hat{\varepsilon}$ et $\hat{\beta}$ sont indépendants

et donc, comme $\hat{Y} = \hat{A}x + \hat{B}$ est fonction de $\hat{\beta}$, on a aussi

les vecteurs $\hat{\varepsilon}$ et $\hat{Y}$ sont indépendants

Estimation de σ^2 Le paramètre σ^2 étant la variance des erreurs supposées centrées e_i, la quantité $\frac{1}{n}\sum_{i=1}^n e_i^2$ sera proche de σ^2 par la loi des grands nombres. Ces variables d'erreurs étant inconnues, on les remplace par les $\hat{e}_i$ et cela donne une première estimation $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$. On modifie légèrement cette estimation de façon à avoir un estimateur $\hat{S}^2$ sans biais (admis)

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{e}_i^2$$

Proposition 2.1. *(Estimateur de la variance du bruit)*

Posons

$$\hat{S}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2.$$

$\hat{S}^2$ *vérifie*

$$(n-2)\frac{\hat{S}^2}{\sigma^2} = \sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{\sigma^2} \sim \chi^2(n-2)$$

où $\chi^2(n-2)$ est une loi du Khi-deux à $n-2$ degrés de libertés. En particulier c'est un estimateur sans biais de σ^2.

Démonstration. On admet cette proposition mais voici l'idée :
Un rappel sur la loi du Khi-deux :

$$\text{Si } X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1) \text{ alors } \sum_{i=1}^n X_i^2 \sim \chi^2(n).$$

Et donc, comme $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ on a

$$\sum_{i=1}^n \frac{\varepsilon_i^2}{\sigma^2} \sim \chi^2(n).$$

Le remplacement de ε_i par son approximation $\hat{\varepsilon}_i$ fait baisser le nombre de degrés de libertés de n à $n-2$, pourquoi ?
On peut donner l'intuition de ce résultat en comparant avec un résultat vu dans le cours de statistique :

— si $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ alors $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$ car $\frac{X_i-\mu}{\sigma} \overset{iid}{\sim} \mathcal{N}(0,1)$.
— Quand on remplace le paramètre μ dans $\sum_{i=1}^n (X_i-\mu)^2$ par l'estimateur $\hat{\mu} = \bar{X}$ on perd un degré de liberté $\sum_{i=1}^n \left(\frac{X_i-\hat{\mu}}{\sigma}\right)^2 \sim \chi^2(n-1)$

Ici, ce qui joue le rôle de la variable X_i est la variable Y_i on a : $Y_i \sim \mathcal{N}(ax_i+b, \sigma^2)$ et les Y_i sont indépendants
donc

$$\frac{Y_i - (ax_i+b)}{\sigma} \overset{iid}{\sim} \mathcal{N}(0,1)$$

On a donc

$$\sum_{i=1}^n \left[\frac{Y_i - (ax_i+b)}{\sigma}\right]^2 \sim \chi^2(n)$$

Ici deux paramètres ont été estimés : $\hat{a}$ et $\hat{b}$. D'où le nombre $n-2$ de degrés de liberté :

$$\sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{\sigma^2} = \sum_{i=1}^n \left[\frac{Y_i - (\hat{a}x_i+\hat{b})}{\sigma}\right]^2 \sim \chi^2(n-2)$$

□

Code R : estimation de σ^2

Dans la sortie de `summary(reg)`, l'estimateur σ^2 correspond au carré de la `Residual standard error` et vaut donc $(1.199)^2$ (autrement dit "residual standard error" est l'estimateur de l'écart-type σ). On peut aussi y accéder par

```
> summary(reg)$sigma^2
[1] 1.438041
```

Remarque 2.6. *$\hat{S}^2$ étant une fonction de $\hat{\varepsilon}$, on a aussi :*

$\hat{S}^2$ est indépendant de $\hat{Y}$

Exercice 2.1.
1. *Charger le fichier "jouet1.txt" dans un dataframe nommé `jouet1`. Au vu du graphique de y contre x, une régression linéaire vous semble-t-elle indiquée ?*
2. *Faire la régression de y sur x et mettre le résultat dans un objet nommé `reg`. Afficher le résumé des résultats. Le résultat confirme-t-il la réponse à la question 1 ?*
3. *Afficher le graphique des résidus $\hat{\varepsilon}_i$ contre les valeurs ajustées $\hat{y}_i$. Que penser de ce graphique au vu de ce que l'on sait sur ces deux quantités ?*
4. *Afficher le graphique des résidus contre x. Identifier le problème.*

2.5 Intervalle de confiance et test sur les paramètres

2.5.1 Intervalles de confiance sur les paramètres

La valeur ponctuelle d'un estimateur est en général insuffisante et il est nécessaire de lui adjoindre un intervalle de confiance (IC) qui va prendre en compte la variabilité de l'estimateur. Nous allons donner des IC pour chaque paramètre a et b.

Définition 2.3. *Soient $I_1(Y)$ et $I_2(Y)$ deux variables aléatoires construites à partir d'un vecteur aléatoire Y dont la loi dépend de θ. $[I_1(Y), I_2(Y)]$ est un* intervalle de confiance *pour θ de niveau $1-\alpha$ si*

$$\mathbf{P}_Y\left([I_1(Y), I_2(Y)] \ni \theta\right) = 1-\alpha$$

Pour une réalisation y de Y, $[I_1(y), I_2(y)]$ est la fourchette de confiance *de θ. En général, on confond fourchette et intervalle.*

Construction d'un intervalle de confiance pour a

- On construit un intervalle de confiance à partir d'une statistique pivotale. Rappelons que nous avons montré que

$$\hat{A} \sim \mathcal{N}(a, \sigma^2_{\hat{A}}) \tag{2.5}$$

où

$$\sigma^2_{\hat{A}} = \frac{\sigma^2 \sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2 \rho_x$$

ce qui est équivalent à dire

$$\frac{\hat{A}-a}{\sigma_{\hat{A}}} \sim \mathcal{N}(0,1)$$

On a donc, si $q^{1-\alpha/2}_{\mathcal{N}(0,1)}$ est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{N}(0,1)$,

$$\mathbf{P}_\beta\Big(\frac{|\hat{A}-a|}{\sigma_{\hat{A}}} \leq q^{1-\alpha/2}_{\mathcal{N}(0,1)}\Big) = 1-\alpha$$

i.e.

$$\mathbf{P}_\beta\Big(\hat{A} - q^{1-\alpha/2}_{\mathcal{N}(0,1)}\sigma_{\hat{A}} \leq a \leq \hat{A} + q^{1-\alpha/2}_{\mathcal{N}(0,1)}\sigma_{\hat{A}} \leq \Big) = 1-\alpha$$

Donc $\left[\hat{A} - q^{1-\alpha/2}_{\mathcal{N}(0,1)}\sigma_{\hat{A}}, \hat{A} + q^{1-\alpha/2}_{\mathcal{N}(0,1)}\sigma_{\hat{A}}\right]$ est un intervalle de confiance de niveau $1-\alpha$ pour a.

- MAIS, cet intervalle de confiance dépend de σ^2 (dans $\sigma_{\hat{A}}$) qui est inconnu, donc pour être capable de calculer la fourchette, il faut qu'on le remplace par son estimateur $\hat{\sigma}^2$ (donc $\hat{\sigma}^2_{\hat{A}} = \hat{\sigma}^2\rho_x$) . Quand on remplace σ^2 (déterministe) par son estimateur, on introduit plus de variabilité dans la loi de la statistique pivotale. Par conséquent, sa distribution de probabilité change.
- Nous allons donc donner la loi de $\frac{|\hat{A}-a|}{\hat{\sigma}_{\hat{A}}}$

$$\begin{aligned}\frac{\hat{A}-a}{\hat{\sigma}_{\hat{A}}} &= \frac{\hat{A}-a}{\sigma_{\hat{A}}}\frac{\sigma_{\hat{A}}}{\hat{\sigma}_{\hat{A}}} = \frac{\hat{A}-a}{\sigma_{\hat{A}}}\frac{\sigma\cancel{\rho_x}}{\hat{S}\cancel{\rho_x}} \\ &= \frac{\frac{\hat{A}-a}{\sigma_{\hat{A}}}}{\sqrt{\frac{\hat{S}^2}{\sigma^2}}}\end{aligned}$$

  - Le numérateur suit une loi $\mathcal{N}(0,1)$;
  - D'après la Proposition 2.1,

$$\frac{\hat{S}^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{n-2};$$

  - D'après la remarque 2.6, le numérateur et le dénominateur sont indépendants.

— Donc on a (définition)

$$\frac{\hat{A}-a}{\hat{\sigma}_{\hat{A}}} \sim \mathcal{T}(n-2)$$

où $\mathcal{T}(n-2)$ est la loi de Student à $n-2$ degrés de liberté.

On a donc, si $q^{1-\alpha/2}_{\mathcal{T}(n-2)}$ est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{T}(n-2)$,

$$\mathbf{P}_\beta\Big(\frac{|\hat{A}-a|}{\hat{\sigma}_{\hat{A}}} \leq q^{1-\alpha/2}_{\mathcal{T}(n-2)}\Big) = 1-\alpha$$

i.e.

$$\mathbf{P}_\beta\Big(\hat{A} - q^{1-\alpha/2}_{\mathcal{T}(n-2)}\hat{\sigma}_{\hat{A}} \leq a \leq \hat{A} + q^{1-\alpha/2}_{\mathcal{T}(n-2)}\hat{\sigma}_{\hat{A}} \leq \Big) = 1-\alpha$$

Intervalles de confiance pour a et b

- Un IC du paramètre A est donné par

$$[\hat{A} - \hat{\sigma}_{\hat{A}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}, \hat{A} + \hat{\sigma}_{\hat{A}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}]$$

- Un IC du paramètre b est donné par

$$[\hat{B} - \hat{\sigma}_{\hat{B}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}, \hat{B} + \hat{\sigma}_{\hat{B}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}]$$

Code : fourchettes pour les paramètres

Le calcul des fourchettes de confiance des coefficients a et b est fait dans **R** par la fonction `confint`.

L'intervalle (bilatéral symétrique) de confiance à 95% est donné par défaut :

```
confint(reg)
```

```
##                 2.5 %    97.5 %
## (Intercept) 8.6847719 9.3901795
## circ        0.2498055 0.2644702
```

On a donc 95% de chances pour que la fourchette 0.2498055 et 0.2644702 contienne le vrai paramètre a.

On peut changer le niveau de confiance avec l'argument `level`, par exemple pour un niveau de confiance de 97%

```
confint(reg,level=0.97)
```

```
##                 1.5 %    98.5 %
## (Intercept) 8.6468993 9.4280520
## circ        0.2490182 0.2652575
```

On constate évidemment que plus on demande une confiance élevée et plus l'intervalle de confiance est large. Si on avait demandé une confiance de 100%, on aurait eu pour intervalle de confiance $]-\infty, +\infty[$.

2.5.2 Test de nullité de a

Une des questions importante est de savoir si la variable x a une influence sur la variable d'intérêt y. En d'autres termes, dans le cadre de notre modèle linéaire, on cherche à tester l'hypothèse

$$\mathcal{H}_0 : a = 0 \quad \text{versus} \quad \mathcal{H}_1 : a \neq 0.$$

On va chercher à construire un test statistique pour cette hypothèse.
Afin de tester si $a = 0$, on peut seulement regarder la valeur de notre estimation $\hat{a}$ et la comparer à 0. On va rejeter l'hypothèse $\mathcal{H}_0$ si $|\hat{a}| > s$ où le seuil s prend en compte la variabilité de notre estimation et l'erreur qu'on accepte de faire, i.e. la probabilité (sous Y) de rejetter $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie.
Pour cela, on va avoir encore une fois besoin d'une loi pivotale (i.e. une loi ne dépendant pas des paramètres inconnus). On sait que $\hat{a}$ est la réalisation de $\hat{A}$ et $\hat{A} \sim \mathcal{N}(a, \sigma^2 \rho_x)$. Or σ^2 est inconnu donc on a besoin de le remplacer par son estimateur. On a montré que

$$\frac{\hat{A} - a}{\hat{\sigma}_{\hat{A}}} \sim \mathcal{T}(n-2)$$

En particulier, sous $\mathcal{H}_0$, $a = 0$ donc

$$T = \frac{\hat{A}}{\hat{\sigma}_{\hat{A}}} \sim \mathcal{T}(n-2).$$

On va rejeter $\mathcal{H}_0$ si $|T| > q$ et on cherche q tel que $\mathbf{P}_{\mathcal{H}_0}(\text{rejeter } \mathcal{H}_0) = \alpha$, i.e. $\mathbf{P}_{\mathcal{H}_0}(|T| > q) = \alpha$. q est donc le quantile de niveau $1 - \frac{\alpha}{2}$ d'une loi de Student à $n-2$ degrés de libertés.
Test : La région de rejet $|T| > q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}$ fournit un test de niveau $1-\alpha$ de l'hypothèse $\mathcal{H}_0 : a = 0$ versus $\mathcal{H}_1 : a \neq 0$.

p-valeur En général, les logiciels fournissent la p-valeur, i.e. le niveau α le plus petit tel que pour tout niveau au dessus on rejette $\mathcal{H}_0$. Si la p-valeur est plus petite que 5% alors on rejette au niveau 5%. Sinon, on n'a pas assez d'information suffisante pour rejetter $\mathcal{H}_0$, soit parce que $\mathcal{H}$ est fausse, soit parce que notre estimateur a une variance trop grande (trop d'incertitude).

Code R : test de l'hypothèse $a = 0$

Sous R les résultats des tests sont fourni dans le summary :

```
summary(reg)
```

```
##
## Call:
## lm(formula = ht ~ circ, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7659 -0.7802  0.0557  0.8271  3.6913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 1427 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.7682
## F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

On constate que la p-valeur du test $a = 0$ est $< 2e-16$. Donc même pour une erreur de $\alpha = 2e-16$ on rejette $\mathcal{H}_0$. La circonférence du tronc a bien un effet sur la taille de l'arbre.

2.6 Prévision, prédiction

2.6.1 Définition et propriétés

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer y. C'est particulièrement le cas de notre exemple lié à la hauteur des arbres. L'idée est d'utiliser un échantillon de n données où on a mesuré la hauteur et la circonférence de chaque arbre, pour ensuite "deviner" le lien entre hauteur et circonférence, pour ensuite "deviner" la hauteur de n'importe quel eucalyptus à partir de sa circonférence.
Soit x_{n+1} une nouvelle valeur de la variable explicative, c'est-à-dire, pour notre exemple, la circonférence d'un nouvel arbre a été mesurée. Nous voulons prédire y_{n+1}, c'est-à-dire prédire sa hauteur à l'aide de notre modèle. Nous allons noter $\hat{y}^p_{n+1}$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue. On observe seulement x_{n+1}.
Le modèle indique que $y_{n+1} = b + ax_{n+1} + e_{n+1}$.
Nous pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{y}^p_{n+1} = \hat{a}x_{n+1} + \hat{b}$$

En utilisant la notation $\hat{y}^p_{n+1}$, nous souhaitons insister sur la notion de prévision : la valeur pour laquelle nous effectuons la prévision, ici la $(n+1)$ème n'a pas servi pour le calcul de l'estimateur $\hat{\beta}$.

Code R : calcul des prédictions

Pour prédire de nouvelles valeurs $(y_{n+1}, \ldots, y_{n+p})$ à partir de $(x_{n+1}, \ldots, x_{n+p})$, on utilise la fonction `predict`, qui prend en entrée le résultat de la fonction `lm`, c'est-à-dire pour notre exemple `reg` et les nouvelles valeurs des variables explicatives $(x_{n+1}, \ldots, x_{n+p})$ qui doivent être entrées dans la colonne d'un data.frame. Cette colonne doit avoir pour nom le nom de la variable, ici `circ`. Imaginons qu'on ait mesuré les hauteurs de trois nouveaux arbres : 46, 35 et 67. Cela donne le code suivant :

```
xnew=c(46,35,67)
xnew=data.frame(circ=xnew)
predict(reg,new=xnew)
```

```
##        1        2        3
## 20.86582 18.03730 26.26571
```

De la même façon, on peut s'intéresser aux variations de cette quantité.

Proposition 2.2. *Posons :* $\hat{Y}^p_{n+1} = \hat{A}x_{n+1} + \hat{B}$. $\hat{Y}^p_{n+1}$ *est une variable aléatoire gaussienne de variance :*

$$\mathbf{Var}(\hat{Y}^p_{n+1}) = \sigma^2\Big(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big)$$

Admis.

Pour avoir une idée de cette variance, il faudrait connaitre le paramètre σ^2. Ce paramètre étant inconnu en général, on va le remplacer par son estimateur $\hat{\sigma}^2$. On peut alors calculer une estimation de cette variance $\widehat{\mathbf{Var}}(\hat{Y}^p_{n+1}) = \hat{\sigma}^2\Big(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big)$. La racine carrée de cette variance (l'écart-type="standard error") est calculée par R. Il suffit de rajouter l'argument `se.fit=T` dans l'appel de la fonction `predict`.

Code R : Prédiction avec écart type

```
predict(reg,new=xnew,se.fit=T)
```

```
## $fit
##        1        2        3
## 20.86582 18.03730 26.26571
##
## $se.fit
##          1          2          3
## 0.03212020 0.05600524 0.08001489
##
## $df
## [1] 1427
##
## $residual.scale
## [1] 1.199183
```

En prévision, on s'intéresse généralement à l'erreur $\hat{\varepsilon}^p_{n+1}$ que l'on commet entre la vraie valeur (inconnue) à prévoir y_{n+1} et celle que l'on prévoit $\hat{y}^p_{n+1}$

$$\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$$

Cette erreur est inconnue puisque y_{n+1} est inconnue. Nous l'appellerons l'erreur de prévision. Cette erreur de prévision quantifie la capacité du modèle à prévoir. Nous avons sur ce thème la proposition suivante

Proposition 2.3. *(Erreur de prévision) L'erreur de prévision satisfait les propriétés suivantes :*

$$\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0$$

$$\mathbf{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\Big(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big)$$

Remarque 2.7. *On a donc :*

$$\mathbb{E}\Big[(\hat{\varepsilon}^p_{n+1})^2\Big] = \mathbf{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\Big(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big)$$

Ceci nous donne une idée de la performance de la prévision. La variance augmente quand x_{n+1} s'éloigne du centre de gravité $\bar{x}$. Autrement dit, faire de la prévision lorsque x_{n+1} est "loin" de $\bar{x}$ est périlleux. Ceci s'explique intuitivement par le fait que plus une observation x_{n+1} est éloignée de la moyenne $\bar{x}$ et moins on a d'information sur elle.

2.6.2 Intervalle de confiance pour la prévision

Exactement comme précédemment, on peut construire un intervalle de confiance pour

$$y_{n+1} = ax_{n+1} + b + e_{n+1}$$

où l'erreur est centrée, réalisation de ε_{n+1} gaussienne de moyenne nulle et de variance σ^2. Nous prédisons y_{n+1} avec

$$\hat{y}^p_{n+1} = \hat{a}x_{n+1} + \hat{b}.$$

On peut étudier les variations de $\hat{Y}^p_{n+1} - y_{n+1}$ pour construire un intervalle de confiance sur y_{n+1}.

Proposition 2.4. *(IC pour y_{n+1})*
Un IC de y_{n+1} est donné par :

$$\Big[\hat{Y}^p_{n+1} \pm q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\Big]$$

Autrement dit, avec une probabilité $1-\alpha$, cet intervalle contient la vraie valeur y_{n+1}.

Exercice 2.2. *(si temps)*

1. *Montrer que $\hat{\varepsilon}^p_{n+1} \sim \mathcal{N}(0, \sigma^2 C)$ où on a noté $C = 1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}$*
2. *Montrer que $\hat{\varepsilon}^p_{n+1}$ est indépendant de $\hat{\sigma}^2$.*
3. *Prouver la proposition précédente.*

Code R : intervalle de confiance pour les prédictions

Si on veut les trois intervalles de confiance associées aux trois valeurs de `xnew` de l'exemple précédent :

```
> predict(reg,new=xnew,interval="pred",level=0.95)
      fit      lwr      upr
1 20.86582 18.51262 23.21901
2 18.03730 15.68239 20.39222
3 26.26571 23.90813 28.62329
```

Par exemple, pour une circonférence de 46, on est sûr à 95% que la hauteur va se trouver

entre 20.86582 et 23.21901 mètres.

On peut illustrer graphiquement la variance de cette prédiction. On repart de la droite d regression tracée sur la Figure 2.6. On souhaite tracer la bande de confiance correspondant aux bornes inférieures et supérieures de la fourchette de confiance.
Pour cela, on crée (par exemple) une grille de 1000 points régulièrement espacés appelée `xnew` (ce sont des $x_{n+1}, \ldots, x_{n+1000}$). Ensuite la fonction `predict` calcule leurs prédictions $(\hat{y}^p_{n+1}, \ldots, \hat{y}^p_{n+1000})$ ainsi que les intervalles de confiance des vraies valeurs $y_{n+1}, \ldots, y_{n+100}$, et on peut tracer le graphique attendu, cf figure 2.8).

Code R : tracé de la droite de prédiction avec intervalle de confiance

```
xnew <- seq(min(euca$circ), max(euca$circ),len=1000)
xnew <- data.frame(xnew); names(xnew) = 'circ'
ICpred <- as.data.frame(predict(reg,xnew,interval="pred",level=0.95))
res_pred <- cbind(ICpred,xnew)
ggplot(euca, aes(x = circ, y = ht))+ geom_point() + geom_smooth(method = lm)
+ geom_line(data=res_pred,aes(x=circ,y=lwr), color = "red",
linetype = "dashed")
+ geom_line(data=res_pred, aes(x = circ,y=upr), color = "red",
linetype = "dashed")
```

Nous constatons que les observations sont globalement bien ajustées par le modèle, sauf peut-être pour les faibles valeurs de circonférences, qui semblent en majorité situées en dessous de la droite. Ceci indique qu'un remplacement de cette droite par une courbe serait une amélioration possible. Peut être qu'un modèle de régression simple du type

$$\texttt{ht} = b + a\sqrt{\texttt{circ}} + \varepsilon$$

serait plus adapté. Remarquons aussi que les 3 circonférences les plus fortes (supérieures à 70 cm) sont bien ajustées par le modèle. Ces 3 individus sont donc différents en terme de circonférence mais bien ajustés par le modèle.
Il en découle que la qualité de l'estimation semble être très bonne, ce qui est normal car le nombre d'individus (i.e. le nombre d'arbres) est très élevé et les données sont bien réparties le long d'une droite.

Remarque 2.8. *De manière générale, le modèle linéaire ne permet pas uniquement de modéliser une relation linéaire entre la variable à expliquer et la variable explicative. En effet, on peut transformer la variable initiale par une transformation quelconque : par exemple on a suggéré de transformer, après visualisation du graphique, la variable* `circ` *(cf aussi question 6 de l'exercice ci-dessous) par la variable* $\sqrt{\texttt{circ}}$*. On utilise ensuite le modèle linéaire avec cette nouvelle variable. En utilisant un modèle à plusieurs variables explicatives (régression multiple, chapitre 3), on peut même modéliser un polynôme.*

2.7 Validation du modèle

Tous les tests et intervalles de confiance sont valables sous l'hypothèse que nos observations sont la réalisation d'un modèle linéaire gaussien. Par conséquent, la première chose à faire est

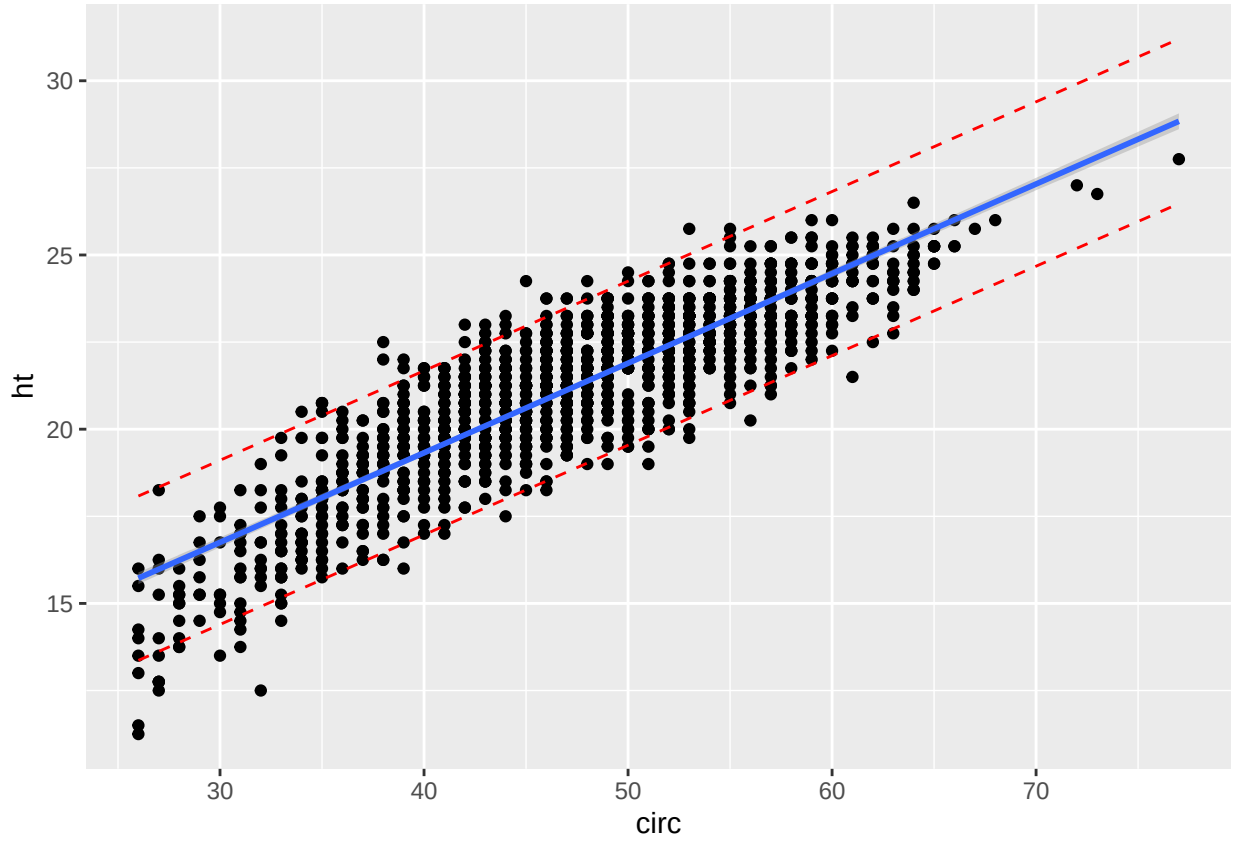


FIGURE 2.8 – Jeu de données *Eucalyptus* : droite de régression avec intervalle de confiance de prédiction

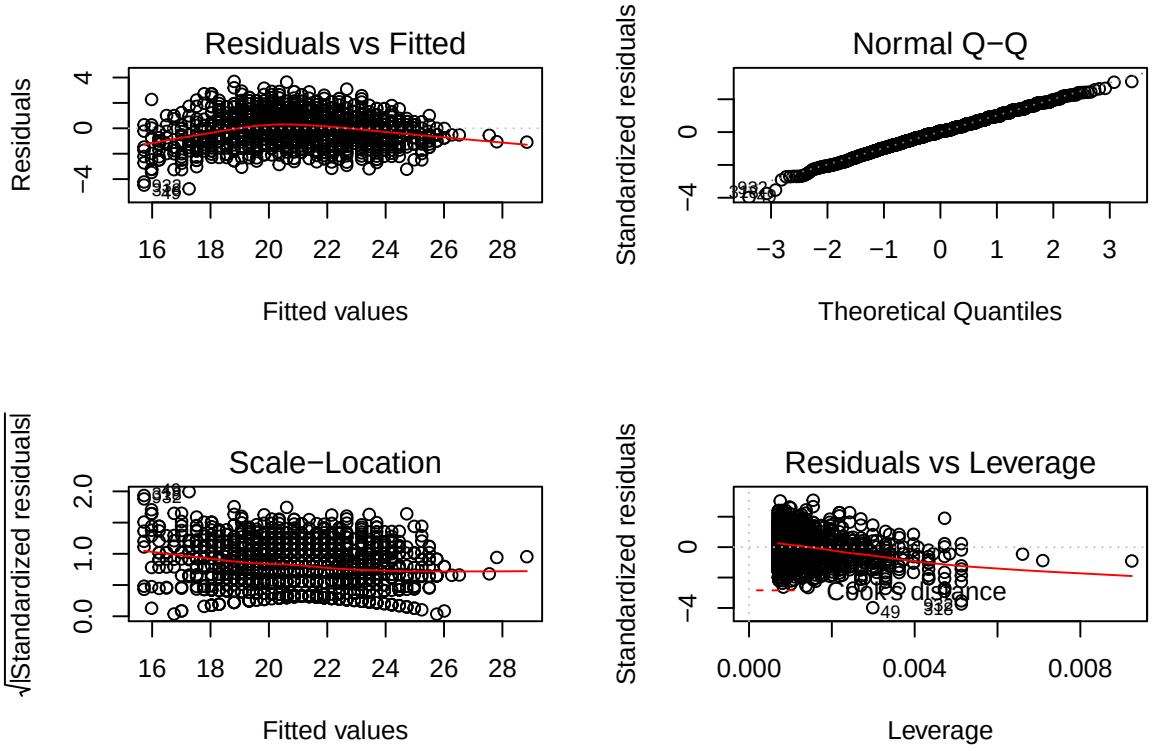


FIGURE 2.9 – Jeu de données *Eucalyptus* : graphes des résidus pour validation de modèle

de vérifier les postulats que l'on a défini sur les erreurs ε_i. Pour rappel, les 4 postulats sont les suivants.

- [P1] Les erreurs sont centrées : $\mathbb{E}[\boldsymbol{\varepsilon}] = 0_{\mathbb{R}}$.
- [P2] Les erreurs $\boldsymbol{\varepsilon}$ sont de variance constante : $\mathbb{V}[\varepsilon_i] = \sigma^2, \forall i = 1 \dots n$.
- [P3] Les termes d'erreur sont supposés indépendants.
- [P4] Finalement, les erreurs sont supposées gaussiennes.

Il est important de s'assurer qu'ils sont vérifiés *sur nos données*. Cependant, nous n'avons pas accès à ces résidus donc nous ne pouvons pas faire de tests statistiques. Nous allons les valider grâce à des outils graphiques utilisant les résidus estimés $\hat{e}_i = y_i - \hat{y}_i$. Les 4 graphiques utiles sont donnés par la commande **R** suivante :

Code R : graphique pour validation de modèle

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

On obtient les 4 graphes suivants, qu'il s'agit d'interpréter.

- [P3] L'indépendance des données ne peut être assurée que par le protocole expérimental.
- [P1] Valider [P1] revient à valider que la relation en y et x est bien affine. On remarquera que par construction les $\hat{e}_i$ sont de moyenne nulle donc on ne peut pas mettre en défaut le postulat [P1] en moyennant les résidus estimés. Cependant, on pourra s'intéresser au

graphe des résidus estimés $\hat{\boldsymbol{e}}$ en fonction des valeurs prédites $\hat{\boldsymbol{y}}$ (graphique en haut à gauche). Si les 4 postulats sont bien respectés alors ces deux vecteurs sont indépendants. Si l'on observe un nuage de points centré et aligné sans structure particulière, alors on se satisfait. Par contre, si on observe une structure particulière (croissance des résidus en fonction des données prédites par exemple ou autre) alors on pourra penser que le modèle n'est pas adapté aux données, qu'il manque une tendance dans le modèle. Plusieurs solutions sont possibles : travailler avec le log des observations, travailler avec le log des variables explicatives...

Dans le cas de nos données Eucalyptus, le premier graphe en haut à gauche semble montrer une tendance. Nous pourrons essayer travailler sur le log de la hauteur.

- [P2] Afin de vérifier que les résidus sont homoscédastiques (de même variance), on trace le graphe des résidus standardisés $\hat{r}_i = \frac{\hat{e}_i}{\hat{s}e_i}$ (où se_i est l'écart-type estimé des $\hat{e}_i$) contre les valeurs prédites (graphique en bas à gauche). De la même façon, si l'on observe un nuage de points centré et aligné sans structure particulière, alors on se satisfait. Par contre, si on observe une structure particulière (croissance des résidus standardisés en fonction des données prédites par exemple ou autre) alors on pourra penser que le modèle n'est pas homoscédastique.
- [P4] Le QQ-plot des résidus estimés (graphique en haut à droite) est une façon de tester le caractère gaussien des résidus. Cependant, ce postulat [P4] n'est pas obligatoire car si le nombre d'observations n est grand on peut obtenir des propriétés asymptotiques des estimateurs et tests.
- **R** fournit un quatrième graphe de diagnostique (en bas à droite) permettant de détecter les observations y_i ayant eu une grande influence sur l'estimation. Si une observation est très influente sur l'estimation alors c'est problématique puisque l'estimation ne sera pas stable si on considère une autre expérience. Ce graphe s'appelle le "residuals versus leverage".

  Revenons aux valeurs prédites $\hat{y}_i$. Elles s'écrivent comment combinaison linéaire de toutes les observations

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

  Si $|h_{ij}|$ est grand alors l'observation j pèse beaucoup dans la prédiction $\hat{y}_i$.

  Le graphe "residuals versus leverage" trace $\hat{r}_i$ en fonction de h_{ii}.

  — Si $\hat{r}_i$ grand, le point est mal ajusté, il est atypique par rapport aux autres données.
    — *Si $\hat{r}_i$ grand et h_{ii} petit*, il est atypique mais peu influent sur l'estimation des paramètres. Donc ce n'est pas très grave
    — *Si $\hat{r}_i$ grand et h_{ii} grand*, alors il est atypique ET influe beaucoup sur l'estimation. Cela devient problématique.

  Il est aussi possible de regarder la distance de Cook.

$$d_i = \frac{\sum_{j=1}^{n} (\hat{y}_j^{(-i)} - \hat{y}_j)^2}{(p+1)\hat{\sigma}^2} = \frac{\hat{r}_i^2 h_{ii}}{(p+1)(1-h_{ii})^2}$$

  où $\hat{y}_i^{(-j)}$ est la prédiction de la i-ème observation lorsque l'analyse a été faite en retirant l'observation j.

  Plus cette distance est grande et plus le point i est influent. En général on regarde de près les points dont la distance de Cook dépasse 1. On pourra décider d'enlever les points

aberrants.

La validation du modèle est faite de façon beaucoup plus étayée dans le Chapitre d'appendice C.

Exercice 2.3. *Au vu des graphiques de validation de modèle (Figure 2.9), on propose de changer de modèle et d'expliquer la hauteur en fonction du logarithme de la circonférence du tronc.*

1. *Ecrire les instructions* **R**.
2. *Afficher les* 4 *graphes de validation. A-t-on amélioré les choses ?*

Exercice 2.4. *On va utiliser les données contenues dans le fichier ozone.txt. On va étudier l'influence de la température sur la concentration en ozone :* `maxO3` *est la concentration en ozone et* `T12` *la température à 12H. On choisit donc* `maxO3` *comme variable à expliquer et* `T12` *comme variable explicative.*

1. *Importer le fichier texte ozone.txt dans un data.frame qu'on nommera* `ozone`. *Afficher ses variables. Que vaut n ici ?*
2. *Vous semble-t-il opportun d'utiliser le modèle linéaire ?*
3. *Afficher le résumé des informations de la régression. Donner l'EMC* $\hat{\beta}$. *Donner les intervalles de confiance à* 95%.
4. *Est-ce que la variable* `T12` *vous semble bien expliquer linéairement la variable* `maxO3` *? Donner le résultat du test.*
5. *Donner la prédiction d'ozone pour une température égale à 27 degrés ainsi que l'intervalle de confiance à 95% correspondant.*

Chapitre 3

Régression linéaire multiple

3.1 Introduction

La modélisation de la concentration d'ozone dans l'atmosphère évoquée dans le dernier exemple du Chapitre 1 est relativement simpliste. En effet, d'autres variables météorologiques sont susceptibles d'avoir une influence sur la variable `maxO3`, comme la quantité de vent, la température ou la nébulosité.
Pour analyser la relation entre la température (`T12`), le vent (`Vx12`), la nébulosité (`Ne12`) et l'ozone (`maxO3`), nous cherchons, comme dans le chapitre 1, une fonction $f : \mathbb{R}^3 \to \mathbb{R}$ telle que

$$\texttt{maxO3}_i \approx f(\texttt{T12}_i, \texttt{Vx12}_i, \texttt{Ne12}_i).$$

Si on suppose que le lien est linéaire :

$$f(x^1, x^2, x^3) = \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3$$

Pour estimer la fonction f, autrement dit ici, pour estimer les coefficients $\beta_1, \beta_2, \beta_3$, on dispose de n mesures. Pour notre exemple $n = 120$. De manière générale, dans ce chapitre, nous aurons toujours une variable y à expliquer mais on aura p variables explicatives.

Notations On introduit les notations suivantes.

- y_i est la i-ème observation.
- $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.
- x_i^j la mesure de la j-ème variable sur le i-ème observation.
- X est une matrice à n lignes et p colonnes telle que $X_{ij} = x_i^j$. Sur notre exemple, les observations sont les jours, et la quantité X_{21} par exemple représente la valeur de la température le deuxième jour.
- On note $x^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix}$ la j-ème colonne de X. x^j représente les mesures de la $j$ème variable explicative faites sur les n observations. On désignera alternativement x^j comme un vecteur ou comme "la j-ème variable explicative". Dans la suite, on va identifier la variable explicative au vecteur de ses mesures x^j.

- x_i est un vecteur ligne représentant les p variables de l'observation i.

Nous supposons donc dans ce chapitre qu'il existe p coefficients $(\beta_1, \ldots, \beta_p)$ tels que

$$y \approx \beta_1 x^1 + \ldots, \beta_p x^p$$

Cela signifie que pour tout $i \in \{1, \ldots, n\}$,

$$y_i \approx \beta_1 x_i^1 + \ldots + \beta_p x_i^p = x_i \beta$$

De la même manière que pour la régression simple, nous devons préciser le sens de cette approximation $\approx$. Nous nous donnons donc une fonction coût ℓ et nous cherchons à minimiser

$$\sum_{i=1}^{n} \ell(y_i - \beta_1 x_i^1 - \ldots - \beta_p x_i^p)$$

Comme pour la régression simple, et pour les mêmes raisons, nous utiliserons la fonction de coût quadratique $\ell(x) = x^2$.

3.2 Modélisation

Le modèle de régression multiple est une généralisation du modèle de régression simple. Nous avons à nouveau une erreur possible, représentée par un vecteur d'erreurs $(\varepsilon_1, \ldots, \varepsilon_n)^T$ sur chaque mesure, que nous supposerons aussi iid gaussiennes standard. Nous supposons donc que les données y_i sont la réalisation de Y_i telles que :

$$Y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i, \quad i = 1, \ldots, n \tag{3.1}$$

où

— Les x_i^j sont des nombres connus, déterministes (i.e. non aléatoires).
— Les paramètres β_j sont déterministes et inconnus.
— Les variables ε_i sont des variables aléatoires iid de loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 inconnu.

Pour simplifier l'introduction des résultats, nous avons supposé que la fonction f était linéaire. En réalité, l'immense majorité du temps, on la suppose affine, c'est-à-dire que

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

On a donc, comme pour la régression linéaire, une variable constante égale à 1, c'est-à-dire qu'on a un vecteur x^0 de coordonnées toutes égales à 1. Plus simplement on va écrire

$$Y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i, \quad i = 1, \ldots, n$$

et considérer que la première variable est l'intercepte, c'est-à-dire la variable constante égale à 1. Autrement dit, $x_i^1 = 1$ pour tout i, i.e. $x^1 = \mathbf{1}$ le vecteur de $\mathbb{R}^n$ de coordonnées toutes égales à 1.

Vocabulaire

On dit qu'on fait "la régression de y sur $x^1, \ldots, x^p$". Les variables explicatives sont appelées aussi des régresseurs. On trouve encore d'autres dénominations : variables d'entrée ou exogènes. La variable y est aussi appelée variable de sortie ou endogène.

Écriture matricielle Si on pose

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

on voit que les n équations de (3.1)

$$Y_i = \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i, \quad i = 1, \dots, n$$

se réécrivent

$$\boldsymbol{Y} = X\beta + \boldsymbol{\varepsilon}$$

où on note X la matrice de dimension $n \times p$ de terme général $X_{ij} = x_i^j$. On appelle cette matrice, la matrice de design. La matrice X a donc pour colonnes les variables explicatives, et pour lignes les individus.
Comme au chapitre 1, le vecteur d'erreurs $\boldsymbol{\varepsilon}$ vérifie

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Comme X et β sont déterministes, Y est aussi un vecteur gaussien et

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n). \tag{3.2}$$

Hypothèse sur X On va faire une hypothèse fondamentale concernant cette matrice dans le reste du cours : on suppose que

X est de plein rang en colonnes

Autrement dit, on suppose que les variables explicatives, qui constituent les colonnes de X, sont linéairement indépendantes. On a donc forcément

$$p \leq n$$

(voir rappels d'algèbre linéaire) Cette hypothèse est relativement contraignante et n'est pas toujours vérifiée. Elle sera difficilement vérifiée quand on a beaucoup de variables explicatives (p grand). On voit déjà que, pour qu'elle soit vérifiée, on doit avoir plus de données que de variables explicatives. Nous reparlerons de cette hypothèse un peu plus loin.
L'hypothèse de normalité est moins contraignante. En effet, on peut montrer que, même sans cette hypothèse, beaucoup de résultats que l'on va dériver de cette hypothèse sont valables pour n suffisamment grand, sous des conditions assez larges.
L'hypothèse que les erreurs sont toutes de même variance est appelée hypothèse d'homoscédasticité. Cette hypothèse, plus l'hypothèse d'indépendance de l'erreur, peuvent être non vérifiées dans les faits, et sont plus contraignantes que l'hypothèse de normalité.

Remarque 3.1. *On note x_i i-ième le vecteur ligne de X. L'équation (3.1) s'écrit aussi*

$$Y_i = x_i\beta + \varepsilon_i$$

3.3 Estimateur des moindres carrés ordinaire (EMC)

Comme pour la régression linéaire simple, nous choisissons la fonction de coût quadratique, d'où la dénomination d'estimateur des moindres carrés.

Définition 3.1. *L'estimation des moindres carrés $\hat{\beta}$ est défini comme suit*

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_i^j \beta_j \right)^2$$

Comme $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_i^j \beta_j)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}(X\beta)_i)^2 = \|y - X\beta\|^2$, on peut aussi écrire

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|^2$$

Dans la suite de cette section , nous allons donner l'expression de $\hat{\beta}$ ainsi que certaines de ses propriétés.

3.3.1 Calcul de l'EMC $\hat{\beta}$

Théorème 3.1. *On suppose que X est de plein rang en colonnes. Alors la solution $\hat{\beta}$ est donnée par*

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3.3}$$

Démonstration. (non exigée)

Rappels d'algèbre utiles pour la démonstration

Projecteurs orthogonaux

Soient E_1 et E_2 deux espaces supplémentaires dans E de sorte que $\forall \boldsymbol{y} \in E$ il existe un unique couple $(\boldsymbol{y}_1, \boldsymbol{y}_2) \in E_1 \times E_2$ tels que $\boldsymbol{y} = \boldsymbol{y}_1 + \boldsymbol{y}_2$. On appelle *projection de E sur E_1 parallèlement à E_2* l'application $p : E \mapsto E$ telle que $p(\boldsymbol{y}) = \boldsymbol{y}_1$.

Maintenant, considérons un sous espace vectoriel F et son supplémentaire orthogonal $F^\perp$ dans E de sorte que $\forall \boldsymbol{y} \in E$, $\exists! (\boldsymbol{y}_1, \boldsymbol{y}_2) \in F \times F^\perp$ tels que $\boldsymbol{y} = \boldsymbol{y}_1 + \boldsymbol{y}_2$.

On appelle *projecteur orthogonal* de E sur F le projecteur E sur F parallèlement à $F^\perp$. Soit $\boldsymbol{y} \in E$, notons $\pi_F(\boldsymbol{y})$ le projecteur orthogonal de E sur F. On a alors

- $\pi_F(\boldsymbol{y}) \in F$
- $\boldsymbol{y} - \pi_F(\boldsymbol{y}) \in F^\perp$

On peut montrer que

$$\pi_F(\boldsymbol{y}) = \arg\min_{v \in F} \|\boldsymbol{y} - \boldsymbol{v}\|^2$$

soit encore $\|\boldsymbol{y} - \pi_F(\boldsymbol{y})\|^2 = \min_{v \in F} \|\boldsymbol{y} - \boldsymbol{v}\|^2$.

En effet, pour tout $\boldsymbol{y} \in E$ et tout $\boldsymbol{v} \in F$, on a

$$\|\boldsymbol{y} - \boldsymbol{v}\|^2 = \| \underbrace{\boldsymbol{y} - \pi_F(\boldsymbol{y})}_{\in F^\perp} + \underbrace{\pi_F(\boldsymbol{y}) - \boldsymbol{v}}_{\in F} \|^2 = \|\boldsymbol{y} - \pi_F(\boldsymbol{y})\|^2 + \|\pi_F(\boldsymbol{y}) - \boldsymbol{v}\|^2$$

par le théorème de Pythagore. Par conséquent, $\|\boldsymbol{y} - \boldsymbol{v}\|^2$ atteint son minimum pour $\|\pi_F(\boldsymbol{y}) - \boldsymbol{v}\|^2 = 0$, i.e. $\pi_F(\boldsymbol{y}) - \boldsymbol{v} = 0_E$, c'est à dire $\boldsymbol{v} = \pi_F(\boldsymbol{y})$.

Theorème du rang

Soit X une matrice de taille (n, r). Par abus de notation on note aussi X l'endomorphisme associé à la matrice X. Alors on a

$$r = \dim \text{ Ker}(X) + rg(X).$$

Par conséquent, X est de rang plein si et seulement si X est injective.

1. Commençons par montrer que X de rang $r \Leftrightarrow X'X$ inversible.

   $X'X$ est une matrice carrée symétrique de taille r.

   — Soit $\boldsymbol{u} \in \mathbb{R}^r$ tel que $X'X\boldsymbol{u} = 0_{\mathbb{R}^r}$. Alors on a $\boldsymbol{u}'X'X\boldsymbol{u} = 0 \Leftrightarrow \|X\boldsymbol{u}\| = 0 \Leftrightarrow X\boldsymbol{u} = 0_{\mathbb{R}^n}$. Or puisque X est de rang plein, l'endomorphisme associé à X est injectif. D'où $\boldsymbol{u} = 0_{\mathbb{R}^r}$.

   — Réciproquement, supposons que $X'X$ est inversible. Soit $\boldsymbol{u} \in \mathbb{R}^r$ tel que $X\boldsymbol{u} = 0_{\mathbb{R}^n}$. Alors $X'X\boldsymbol{u} = 0_{\mathbb{R}^r}$. Or $X'X$ inversible, donc $\boldsymbol{u} = 0_{\mathbb{R}^r}$. D'où X est injective. D'où $rg(X) = r$. X est de rang plein.

2. Montrons maintenant que $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) = (X'X)^{-1}X'\boldsymbol{y}$.

   La démonstration de ce résultat repose sur la notion de projection orthogonale. Soit $\boldsymbol{y}$ un vecteur de $E = \mathbb{R}^n$ et F un sous-espace vectoriel de $\mathbb{R}^n$. Le projeté orthogonale de $\boldsymbol{y}$ sur F est le vecteur de F noté $\pi_F(\boldsymbol{y})$ est tel que $\pi_F(\boldsymbol{y}) = \arg\min_{v \in F} \|\boldsymbol{y} - \boldsymbol{v}\|^2$, soit encore $\|\boldsymbol{y} - \pi_F(\boldsymbol{y})\|^2 = \min_{v \in F} \|\boldsymbol{y} - \boldsymbol{v}\|^2$.

   Notons $F = [X]$ l'espace vectoriel engendré par les colonnes de X, autrement dit

   $$[X] = \{X\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^r\}.$$

   Ainsi minimiser $\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$ pour $\beta \in \mathbb{R}^r$ revient à minimiser $\|\boldsymbol{y} - \boldsymbol{v}\|^2$ pour $\boldsymbol{v} \in [X]$. Cette quantité est minimale en $\hat{\boldsymbol{v}} = \pi_{[X]}(\boldsymbol{y})$ (où $\pi_{[X]}(\boldsymbol{y})$ est le projeté orthogonal de $\boldsymbol{y}$ sur $[X]$).

Il existe un unique $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^r$ tel que

$$\pi_{[X]}(\boldsymbol{y}) = X\widehat{\boldsymbol{\beta}}$$

Par définition du projeté orthogonal, pour tout $k = 1, \ldots, r$, $\pi_{[X]}(\boldsymbol{y}) - \boldsymbol{y}$ est orthogonal à $[X]$, espace engendré par les vecteurs colonnes de X. Notons $X^1, \ldots, X^r$, ces vecteurs colonnes. $\forall k = 1 \ldots r$:

$$< X^k, \pi_{[X]}(\boldsymbol{y}) - \boldsymbol{y} >= 0 \quad \Leftrightarrow \quad < X^k, X\widehat{\boldsymbol{\beta}} - \boldsymbol{y} >= 0 \quad \Leftrightarrow \quad (X^k)'(X\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) = 0$$

Par conséquent $X'(X\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) = 0_{\mathbb{R}^r}$ d'où $X'X\widehat{\boldsymbol{\beta}} = X'\boldsymbol{y}$. Or $X'X$ est inversible car X de rang plein, donc

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) = (X'X)^{-1}X'\boldsymbol{y}.$$

□

Définition 3.2. *(Estimateur linéaire)*
Un estimateur $\hat{\beta}$ de β est dit linéaire s'il s'écrit $\hat{\beta} = Ay$ avec une matrice A dépendant du design X.

Remarque 3.2. *l'EMC $\hat{\beta}$ de β est un estimateur linéaire (prendre $A = (X^TX)^{-1}X^T$).*

Remarque 3.3. *Que se passe-t-il quand X n'est pas de plein rang en colonnes ? Le vecteur $\hat{y} = Py$ est toujours l'unique solution du problème (??) (il ne s'exprime plus comme $X(X^TX)^{-1}X^Ty$ puisque la matrice X^TX n'est plus inversible mais cette projection existe toujours et est unique). Ce vecteur peut toujours s'écrire sous la forme $\hat{y} = X\hat{\beta}$ mais cette écriture n'est pas unique. En effet, l'équation matricielle $X\hat{\beta} = \hat{y}$ a toujours au moins une solution (puisque $\hat{y} = P_{[X]}y \in \mathbf{Im}(X)$) mais dans ce cas on a même une infinité de solutions données par $\hat{\beta}_0 + \mathbf{Ker}(X)$, avec $\hat{\beta}_0$ une solution particulière. On met en général une contrainte sur β pour avoir une solution unique (non traité ici) et aussi réduire la variance.*

Code R : Estimation pour régression linéaire multiple

Revenons sur notre exemple `Ozone` : nous cherchons à expliquer la concentration en ozone `maxO3` par la nébulosité `Ne12`, le vent `Vx12` et la température `T12`. La syntaxe est presque la même. On n'a pas besoin de préciser que la première variable explicative est la constante **1**, R la met automatiquement. On peut cependant demander à R de l'enlever, en écrivant

```
maxO3~Ne12 + Vx12 + T12-1
```

On peut utiliser la fonction `attach` pour éviter de préciser dans la fonction lm qu'on utilise `data=ozone`.

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

```
##
## Call:
## lm(formula = maxO3 ~ Ne12 + Vx12 + T12)
##
```

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

Si X n'est pas de rang plein Donnons un exemple concret où la matrice de design n'est pas de plein rang en colonnes.
Supposons qu'on ait deux variables explicatives x^2 et x^3 corrélées (on note encore x^1 l'intercept) : $x^3 = 3x^1 + 2x^2$ par exemple. Introduisons un modèle linéaire avec ces deux variables et l'intercept $x^1 = \mathbf{1}$

$$y = 4x^1 + 2x^2 + x^3 + \varepsilon$$

Comme $x^3 = 3x^1 + 2x^2$, le modèle peut aussi s'écrire

$$y = 7x^1 + 4x^2 + \varepsilon$$

ou bien

$$y = x^1 + 2x^3 + \varepsilon$$

ou bien ....(il y a une infinité d'écritures). C'est ce manque d'identifiabilité qui fait que la formule donnée dans le théorème précédent ne fonctionne pas.

```
x2=runif(100)
x3=2*x2+3
y=4+2*x2+3*x3+rnorm(100)
donnees=data.frame(y=y,x2=x2,x3=x3)
summary(lm(y~x2+x3,data=donnees))

Call:
lm(formula = y ~ x2 + x3, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7121 -0.6585 -0.0272  0.6502  3.3645

Coefficients: (1 not defined because of singularities)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8054     0.1974   64.86   <2e-16 ***
x2            7.9613     0.3396   23.44   <2e-16 ***
x3                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.095 on 98 degrees of freedom
Multiple R-squared:  0.8487,Adjusted R-squared:  0.8471
F-statistic: 549.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

3.3.2 Propriétés de l'EMC $\hat{\beta}$

On s'intéresse maintenant aux propriétés probabilistes de l'estimateur $\hat{\boldsymbol{\beta}} = (X^T X)X^T \boldsymbol{Y}$. Comme précédemment, on a besoin de caractériser sa loi.

Théorème 3.2.

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

Démonstration. On pose $A = (X^T X)^{-1} X^T$. On a alors $\hat{\beta} = A\boldsymbol{Y}$, i.e. $\hat{\beta}$ est une transformée linéaire du vecteur $\boldsymbol{Y}$. Donc c'est un vecteur gaussien.
Pour le calcul de l'espérance et de la variance, on utilise les propriétés des vecteurs aléatoires (voir rappels au chapitre Annexe B). L'espérance de $\hat{\beta}$ est donnée par

$$\mathbb{E}(\hat{\beta}) = A\mathbb{E}(\boldsymbol{Y}) = (X^T X)^{-1} X^T (X\beta) = \beta$$

Le reste est admis.

□

Deux remarques peuvent être faites :

- L'EMC de $\hat{\beta}$ est sans biais.
- L'opérateur de covariance fait intervenir **l'inverse** de $X^T X$.

3.4 Résidus et variance résiduelle

3.4.1 Prédiction et Résidus

On appelle *valeur prédites* : $\forall i = 1, \ldots, n$

$$\hat{y}_i = \sum_{j=1}^{p} \hat{\beta}_j x_i^j = x_i \hat{\beta}$$

Donc, en adoptant l'écriture matricielle on a

$$\hat{\boldsymbol{y}} = X\hat{\beta} = X(X^T X)^{-1} X'Y = P_{[X]}\boldsymbol{y}$$

.

Comme dans le Chapitre 2, les *résidus* $\hat{e}_i$ sont définis, pour $i = 1, \ldots, n$, par

$$\hat{e}_i = \hat{y}_i - y_i, \quad \text{et} \quad \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

Si on note $\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$, on obtient

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I_n - P_{[X]})\boldsymbol{Y}$$

Donc le vecteur des résidus, étant une transformée linéaire du vecteur gaussien Y, est aussi gaussien. C'est un vecteur centré

$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbb{E}(\boldsymbol{Y} - X\hat{\beta}) = \mathbb{E}(\boldsymbol{Y}) - X\mathbb{E}(\hat{\beta}) = X\beta - X\beta = 0$$

On peut montrer que l'opérateur de variance-covariance $\hat{\boldsymbol{\varepsilon}}$ est égal à

$$\Sigma_{\hat{\boldsymbol{\varepsilon}}} = \sigma^2(I_n - P_{[X]})$$

Autrement dit cette matrice fait intervenir la variance σ^2 du bruit mais n'est en général pas diagonale : $P_{[X]}$ n'a aucune raison d'être diagonale. De plus ses termes diagonaux dépendent de σ^2 mais ne sont pas tous égaux. Or un vecteur gaussien a ses composantes indépendantes si et seulement si sa matrice de variance-covariance est diagonale.
En conclusion : **les résidus $\hat{\varepsilon}_i$ sont gaussiens centrés, comme l'erreur ε_i. Cependant ils n'ont pas tous la même variance, et ne sont pas indépendants en général**. On revient sur l'analyse des résidus en fin de chapitre. Une propriété importante est la suivante :

$\hat{\boldsymbol{\varepsilon}}$ et $\hat{\boldsymbol{Y}}$ sont indépendants

Code R : résidus

On trouve à nouveau ces résidus par la commande `resid(reg)`.

Exercice 3.1. *1. Charger le fichier "jouet2.txt". Faire la régression de y sur x.*

2. Afficher le graphique de $\hat{y}$ en ordonnée contre $\hat{\varepsilon}$ en abscisse .

3. Quelle conclusion tirer de ce graphique ?

3.4.2 Estimation de σ^2

On va à nouveau proposer un estimateur de σ^2 basé sur la variance empirique $\tilde{\sigma}^2$ du vecteur de résidu $\hat{e}$, c'est-à-dire $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$. De la même manière que pour la régression simple, on va modifier légèrement cet estimateur, en divisant par $n - p$ plutôt que par n.

Proposition 3.1. *(Estimateur de la variance du bruit)*
L'estimateur

$$\hat{S}^2 = \frac{1}{n-p}\sum_{i=1}^n \hat{\varepsilon}_i^2$$

est un estimateur sans biais de σ^2. On a plus précisément

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$$

où $\chi^2(n-p)$ est une loi du Khi-deux à $n-p$ degrés de libertés.

Remarque 3.4. *On peut aussi montrer que l'EMV de σ^2 est égal à la variance empirique $\tilde{\sigma}^2$ du vecteur $\hat{\varepsilon}$.*

Conséquence : estimation de la variance des estimateurs On a calculé que la variance théorique de chaque coefficient $\hat{\beta}_j$ pour $j \in \{1, \dots, p\}$ (Theorème 3.2). Par exemple

$$\sigma^2_{\hat{\beta}_j} = \sigma^2 \Big[(X^T X)^{-1}\Big]_{jj}$$

On obtient aussi la covariance entre les composantes de $\hat{\beta}$. Par exemple, la covariance entre $\hat{\beta}_j$ et $\hat{\beta}_k$ est donnée par

$$\sigma_{\hat{\beta}_j, \hat{\beta}_k} = \sigma^2 \Big[(X^T X)^{-1}\Big]_{jk}$$

Pour calculer ces quantités, il faudrait connaître σ^2. Or σ^2 est inconnue en général, on la remplace donc par un estimateur $\hat{\sigma}^2$. On obtient alors une estimation de la matrice de covariance. Par exemple

$$\hat{\sigma}^2_{\hat{\beta}_j} = \hat{\sigma}^2 \Big[(X^T X)^{-1}\Big]_{jj}$$

Code R : variance estimée des estimateurs et $\hat{\sigma}^2$

Dans le résumé `summary(reg)`, on trouve, de la même manière que pour la régression simple, l'estimation de la variance σ^2 en mettant au carré la quantité `Residual standard error`, que l'on obtient aussi par

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

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

Les écarts-types de chaque coefficient $\hat{\beta}_j$ sont donnés dans la colonne `Std.Error` du tableau `Coefficients` de la sortie de `summary`.

3.5 Prévision, prédiction

De la même manière que pour la régression simple, on voudrait prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.
De manière générale, nous avons au départ fait des mesures sur n individus. Ces individus correspondent aux lignes de la matrice de design X. On les a noté x_i, ce sont donc des vecteurs à p composantes.

On a une nouvelle donnée x_{n+1} qui correspond à un $(n+1)$-ème individu, et on ne connait pas le y_{n+1} correspondant. Dans l'exemple de l'ozone, il s'agit du vecteur de $\mathbb{R}^3$ composés des mesures de la température, la nébulosité et le vent sur cette journée.
Nous voulons prédire y_{n+1}. Pour notre exemple, on veut prédire la concentration en ozone sur une nouvelle journée à l'aide de notre modèle. Nous allons noter $\hat{y}^p_{n+1}$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue (non mesurée). y_{n+1} est la réalisation de Y_{n+1} :

$$Y_{n+1} = x_{n+1}\beta + \varepsilon_{n+1} \tag{3.4}$$

avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ et ε_{n+1} indépendant des $(\varepsilon_i)_{1\leq i\leq n}$.
Nous pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{Y}^p_{n+1} = x_{n+1}\hat{\beta}$$

et on s'intéresse à l'erreur de prévision $\hat{\varepsilon}^p_{n+1}$ que l'on commet entre la vraie valeur (inconnue) à prévoir y_{n+1} et celle que l'on prévoit $\hat{y}^p_{n+1}$

$$\hat{\varepsilon}^p_{n+1} = Y_{n+1} - \hat{Y}^p_{n+1}$$

Proposition 3.2. *(Erreur de prévision)*

L'erreur de prévision satisfait les propriétés suivantes :

$$\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0$$

$$\mathbf{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\Big(1 + x_{n+1}(X^TX)^{-1}x^T_{n+1}\Big)$$

Remarque 3.5. *On peut faire la même remarque que dans le chapitre sur la régression simple. Le terme $x_{n+1}(X^TX)^{-1}x^T_{n+1}$ est lié à la distance entre x_{n+1} et $\bar{x}$. Plus x_{n+1} est loin du centre de gravité du nuage de points, plus ce terme augmente et plus la précision de la prévision diminue. Cette distance n'est pas la distance euclidienne, c'est la distance de Mahaloanobis (représentée par l'inverse de la matrice de covariance des variables explicatives.)*

Exercice 3.2. *On a mesuré la valeur des trois variables* `T12`, `Ne12` *et* `Vx12` *pour une nouvelle journée : (20,6,-3). Donner le taux d'ozone prévu par le modèle linéaire. (Le code est le même que dans le chapitre 2)*

3.6 Intervalles de confiance

Comme dans le cas de la régression simple, on cherche à accompagner les estimations de fourchettes. Pour cela on construit des intervalles de confiance.

3.6.1 Intervalles de confiance sur les paramètres

Théorème 3.3. *(i) Un IC de niveau $1-\alpha$ pour le coefficient β_j est donné par*

$$[\hat{\beta}_j \pm q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-p)}\hat{\sigma}\sqrt{[(X^TX)^{-1}]_{jj}}]$$

(ii) Un IC de niveau $1-\alpha$ *pour* σ^2 *est donné par*

$$\left[\frac{(n-p)\hat{\sigma}^2}{q^{1-\frac{\alpha}{2}}_{\chi^2(n-p)}}, \frac{(n-p)\hat{\sigma}^2}{q^{\frac{\alpha}{2}}_{\chi^2(n-p)}}\right]$$

Code R : intervalle de confiance sur les paramètres

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

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

Avec une probabilité de 95%, $[-3.6368523, 0.399082]$ contient le coefficient de `Ne12`.

3.6.2 Intervalle de confiance pour la prédiction sur y_{n+1}

Soit $x_{n+1} \in \mathbb{R}^p$ un nouvel individu. On veut un intervalle de confiance sur y_{n+1}.

Théorème 3.4. *(IC de prévision)(non exigé)*

Un IC, de niveau $1-\alpha$ *pour* y_{n+1} *est donné par*

$$[x^T_{n+1}\hat{\beta} \pm q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-p)}\hat{\sigma}\sqrt{x_{n+1}(X^TX)^{-1}x^T_{n+1}+1}]$$

Code R : intervalle de confiance sur la prédiction

Dans l'exemple de l'ozone, imaginons qu'on ait les mesures de deux nouvelles journées x_{n+1} et x_{n+2} et qu'on veuille un IC de confiance sur la prévision de la concentration d'ozone correspondante pour chacune de ces deux journées. Sur la première journée, les valeurs respectives de la nébulosité, la quantité de vent et la température sont de 2, -1 et 20. Sur la deuxième journée, elles sont de 3, 0 et 23. Il faut là encore créer un dataframe et préciser le nom des colonnes de ce dataframe, qui doivent être les mêmes que les noms des variables explicatives dans l'appel de `lm`.

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

```
##        fit      lwr      upr
```

```
## 1 206.6379 166.8820 246.3938
## 2 157.0024 121.8401 192.1647
```

3.7 Tests statistiques

3.7.1 Tests sur la pertinence d'un coefficient

On se pose ici la question de l'utilité d'une variable explicative. Par exemple, pour le cas étudié dans ce chapitre, on peut se poser la question suivante : la quantité de vent influence-t-elle vraiment la concentration d'ozone ?
Dire que la variable est inutile revient à dire que son coefficient est nul. Le problème de test est donc le suivant

$$\mathcal{H}_0 : \beta_j = 0 \text{ contre } \mathcal{H}_1 : \beta_j \neq 0$$

Pour réexpliquer l'idée des tests, faisons d'abord l'hypothèse simplificatrice que σ^2 est connue. Nous devons faire un test sur le coefficient β_j. Comme d'habitude, on va partir d'un estimateur de β_j, ici on prend évidemment l'EMC de β_j, c'est-à-dire $\hat{\beta}_j$. On sait que

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

Ce qui donne, pour la composante $\hat{\beta}_j$ de $\hat{\beta}$,

$$\hat{\beta}_j \sim \mathcal{N}\Big(\beta_j, \sigma^2 [(X^T X)^{-1}]_{jj}\Big)$$

On rappelle qu'un test consiste à fabriquer une statistique, qu'on va appeler Z, qui suit une loi fixée et connue (ou qu'on peut tabuler) sous $\mathcal{H}_0$. Evidemment on souhaite que cette statistique ne se comporte pas de la même façon sous $\mathcal{H}_1$, ce qui permettra de plus facilement distinguer $\mathcal{H}_0$ et $\mathcal{H}_1$. **On commence donc toujours par regarder ce qui se passe sous $\mathcal{H}_0$** et on regarde quel est le comportement "typique" de la statistique choisie sous $\mathcal{H}_0$, ou autrement dit on se pose la question suivante : quelles sont les valeurs "aberrantes" de cette statistique quand on est sous $\mathcal{H}_0$? Ici, sous $\mathcal{H}_0$, nous avons $\beta_j = 0$. Donc

$$\text{sous } \mathcal{H}_0 \quad \hat{\beta}_j \sim \mathcal{N}\Big(0, \sigma^2 [(X^T X)^{-1}]_{jj}\Big)$$

On se ramène à une loi connue, la loi $\mathcal{N}(0,1)$, en standardisant

$$\text{sous } \mathcal{H}_0 \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2 [(X^T X)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

Or, si $Z \sim \mathcal{N}(0,1)$, on a

$$\mathbf{P}(|Z| \leq q_{\mathcal{N}(0,1)}^{1-\frac{\alpha}{2}}) = 1 - \alpha$$

où $q_{\mathcal{N}(0,1)}^{1-\frac{\alpha}{2}}$ est le quantile d'une loi normale centrée réduite de niveau $1 - \alpha/2$. Prenons $\alpha = 0.05$ pour fixer les idées. Cela signifie que, avec une grande probabilité, on a

$$|Z| \leq q_{\mathcal{N}(0,1)}^{0.975}$$

Ici, la variable Z est la variable $\frac{\hat{\beta}_j}{\sqrt{\sigma^2 [(X^T X)^{-1}]_{jj}}}$. Le quantile d'ordre 0.975 de la loi normale standard $\mathcal{N}(0,1)$ est d'environ 1.96.

On a donc 95% de chances pour que cette variable se trouve dans l'intervalle $[-1.96, 1.96]$, ce qui est un petit intervalle autour de 0, *si on est bien sous* $\mathcal{H}_0$.
Remarquez que si $\mathcal{H}_0$ est fausse, on espère que notre statistique sortira de cet intervalle ! Pour fixer les idées, imaginons que la vraie valeur de β_j soit 2. Alors notre statistique de test ne suit plus une loi $\mathcal{N}(0,1)$ (et heureusement, cela nous permettra de plus facilement distinguer $\mathcal{H}_1$ de $\mathcal{H}_0$). En effet, dans ce cas on a

$$\hat{\beta}_j \sim \mathcal{N}\Big(2, \sigma^2[(X^TX)^{-1}]_{jj}\Big)$$

i.e.

$$\frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}(\frac{2}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}}, 1)$$

c'est-à-dire que notre variable Z n'est plus concentrée autour de 0 : elle est, avec une grande probabilité, concentrée autour de la valeur $\frac{2}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}}$. Évidemment, plus on est "loin de $\mathcal{H}_0$", c'est-à-dire plus la valeur de β_j est grande, plus Z va se concentrer autour d'une valeur éloignée de 0. Et donc avec une grande probabilité, la procédure de test le verra et choisira $\mathcal{H}_1$. De la même manière, si β_j est non nulle mais très éloignée de 0 on aboutira souvent au non-rejet de $\mathcal{H}_0$, même si on est bien sous $\mathcal{H}_1$.
On voit aussi que cette valeur autour de laquelle la statistique de test $\frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}}$ se concentre, dépend non seulement de la vraie valeur de β_j (2 dans notre exemple) mais aussi de la variance du bruit σ^2 ainsi que du design à travers le terme $[(X^TX)^{-1}]_{jj}$.
Dans les faits, on ne connait pas σ^2 donc on la remplace par son estimateur $\hat{\sigma}^2$, ce qui modifie la loi : on passe d'une loi $\mathcal{N}(0,1)$ à une loi $\mathcal{T}(n-p)$, c'est-à-dire une Student à $n-p$ degrés de libertés.

Théorème 3.5. *On s'intéresse au problème de test suivant*

$$\mathcal{H}_0 : \beta_j = 0 \text{ contre } \mathcal{H}_1 : \beta_j \neq 0$$

On rejette l'hypothèse $\mathcal{H}_0$ *si*

$$\frac{|\hat{\beta}_j|}{\hat{\sigma}\sqrt{[(X^TX)^{-1}]_{jj}}} > q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-p)}.$$

On a alors construit un test d'erreur de première espèce α.

Démonstration. On a déjà vu (dans l'exercice sur l'IC de β_j) que

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X^TX)^{-1}]_{jj}}} \sim \mathcal{T}(n-p)$$

□

Dans les faits on regarde souvent la p-valeur p, et si on s'est fixé un seuil α, on conclut que :
- si $\alpha < p$ alors on conserve $\mathcal{H}_0$ au niveau α.
- si $\alpha > p$ alors on rejette $\mathcal{H}_0$ au niveau α.

De manière générale, plus la p-valeur est petite, plus on a envie de rejeter $\mathcal{H}_0$.
On considèrera donc une variable explicative comme pertinente au niveau α si la p-valeur correspondante est plus petite que α. Plus la p-valeur est petite, plus on aura confiance en notre rejet, en quelque sorte.

Remarque 3.6. *Attention cependant à la formulation : ne pas dire "la p-valeur est petite donc la probabilité que la variable est pertinente est grande". En effet, la p-valeur* ***n'est pas*** *la probabilité que $\mathcal{H}_0$ soit vraie sachant les données. Cette formulation "la probabilité de $\mathcal{H}_0$ sachant les données" est une formulation bayésienne, qui nécessité de se donner une loi a priori. La p-valeur est associée à ce qui se passe "sachant $\mathcal{H}_0$" et non "sachant les données".*

La p-valeur mesure essentiellement le côté atypique des données par rapport à ce qui est censé se passer sous $\mathcal{H}_0$.
On remarquera que c'est bien le test de nullité d'une des variables en présence de toutes les autres. On n'est pas en train de tester

$$\mathcal{H}_0 : Y_i = \mu + \varepsilon_i, \quad \text{versus} \quad \mathcal{H}_1 : Y_i = \mu + \beta_j x_i^j + \varepsilon_i.$$

Par conséquent, on est en train de chercher les variables dites "significatives" (i.e. qui ont une réelle influence, même quand les autres sont là). A contrario, les variables pour lesquelles on ne rejette pas $\mathcal{H}_0$ n'apporte pas grand chose par rapport aux autres, donc sont peut-être inutiles. Imaginons en effet que l'on ait inclus dans un modèle une variable inutile. Alors cette variable ne se contente pas être inutile, elle gêne aussi notre estimation. En effet elle induit une variance supplémentaire dans l'estimation des autres coefficients.

Remarque 3.7. *En particulier, si des covariables sont fortement corrélées, ce test est difficile à interpréter. Pour plus de détails, voir Chapitre D.*

Code R : test des variables

Ce test est fait automatiquement par R et le résultat se trouve dans le tableau `Coefficient` de la sortie de `summary`.
La valeur de la statistique de test se trouve dans la colonne `t.value` ("t" pour student). La p-valeur est donnée dans la dernière colonne `Pr(>|t|)` . On rappelle que plus la p-valeur est petite, plus on a envie de rejeter $\mathcal{H}_0$. Les variables `Vx12` et `T12` sont significatives au niveau 5%, c'est-à-dire qu'on considère que, au niveau 5%, la quantité de vent et la température influencent réellement (et linéairement) la concentration en ozone. On va considérer la variable `Ne12` comme non significative au niveau 5%.

```
summary(reg)
```

```
##
## Call:
## lm(formula = maxO3 ~ Ne12 + Vx12 + T12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.462 -11.448  -0.722   8.908  46.331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8958    14.8243   0.263   0.7932
## Ne12         -1.6189     1.0181  -1.590   0.1147
```

```
## Vx12          1.6290     0.6571   2.479   0.0147 *
## T12           4.5132     0.5203   8.674 4.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 108 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6518
## F-statistic: 70.25 on 3 and 108 DF,  p-value: < 2.2e-16
```

3.7.2 Test sur la pertinence d'un ensemble de variables explicatives

On peut aussi tester la pertinence d'un groupe de variables explicatives, autrement dit tester la nullité simultanée de plusieurs coefficients. Imaginons pour simplifier les notations que l'on veuille tester la nullité simultanée des q derniers coefficients.

$$\mathcal{H}_0 : \beta_{p-q+1} = \beta_{p-q+2} = \dots \beta_{p-1} = \beta_p = 0 \text{ contre } \mathcal{H}_1 : \exists j \in \{p-q+1, \dots, p\} : \ \beta_j \neq 0$$

Que signifie $\mathcal{H}_0$ en terme de modèle ? Si les q derniers coefficients sont vraiment nuls, alors le modèle est en réalité

$$Y_i = \sum_{j=1}^{p-q} \beta_j x^j + \varepsilon_i, \quad i = 1, \dots, n$$

Autrement dit

$$Y = X_0 \tilde{\beta} + \varepsilon \tag{3.5}$$

avec une nouvelle matrice X_0 de dimension $n \times (p-q)$, dont les colonnes sont les $p-q$ premières colonnes de la matrice X. On a aussi noté $\tilde{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{p-q} \end{pmatrix}$.

On va appeler $\mathcal{M}_1$ le modèle complet, c'est-à-dire avec toutes les variables explicatives de départ, et $\mathcal{M}_0$ le sous-modèle (défini par l'équation (3.5)), c'est-à-dire le modèle avec les q dernières variables explicatives enlevées. On soupçonne donc que le vrai modèle est le modèle $\mathcal{M}_0$.
On peut formuler ce test en terme de comparaison de modèles. Pour comparer les modèles $\mathcal{M}_0$ et $\mathcal{M}_1$ on va comparer leur ajustement aux données.
Définissons les sommes des carrées résiduels :

$$\begin{aligned} \text{SCR} &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \|Y - P_{[X]}Y\|^2 \\ \text{SCR}_0 &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(0)})^2 = \|Y - P_{[X_0]}Y\|^2 \end{aligned}$$

Or la matrice X_0 étant une sous partie des colonnes de X, on a $[X_0] \subset [X]$. Par conséquent,

$$\text{SCR} \ \leq \text{SCR}_0,$$

autrement dit, on s'ajustera mieux aux données avec le modèle plus riche. Cependant, on peut se demander si ce gain d'ajustement "vaut le coup" parce qu'un modèle plus "riche" implique plus de paramètres à estimer donc plus d'incertitude. Par conséquent, on va construire un test sur la différence $\text{SCR} - \text{SCR}_0$ et rejeter $\mathcal{H}_0$ si cette différence est significativement plus grande que 0. La statistique du test de Fisher repose sur cette différence et prend en compte la taille des modèles en compétition :

$$F = \frac{(\text{SCR}_0 - \text{SCR})/(n-p-(n-(p-q))}{\text{SCR}/(n-p)} = \frac{(\text{SCR}_0 - \text{SCR})/q}{\text{SCR}/(n-p)}$$

On admet que cette statistique suit, sous $\mathcal{H}_0$, une loi qu'on appelle loi de Fisher à q et $n-p$ degrés de libertés, ce que l'on note

$$F = \frac{(\text{SCR}_0 - \text{SCR})/q}{\text{SCR}/(n-p)} \sim \mathcal{F}_{q,n-p}$$

En résumé, on a le théorème suivant

Théorème 3.6. *Pour le problème de test*

$$\mathcal{H}_0 : \beta_{p-q+1} = \beta_{p-q+2} = \dots \beta_{p-1} = \beta_p = 0 \text{ contre } \mathcal{H}_1 : \exists j \in \{p-q+1, \dots, p\} : \ \beta_j \neq 0,$$

on rejette $\mathcal{H}_0$ *si*

$$F > q^{1-\alpha}_{\mathcal{F}_{q,n-p}}$$

où $q^{1-\alpha}_{\mathcal{F}_{q,n-p}}$ *est le quantile d'une loi de Fisher à* $(q, n-p)$ *degrés de liberté.*

Code R : test d'un sous modèle (1)

Comme expliqué plus haut, ce test est en fait un test de comparaison de modèles entre un modèle $\mathcal{M}_0$ et un modèle $\mathcal{M}_1$. Pour cela, dans R, il faut définir les deux modèles qui nous intéressent et utiliser la commande `anova` pour obtenir le test.
Nous l'appliquons d'abord sur le jeu de données Ozone.

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

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ T12
## Model 2: maxO3 ~ Ne12 + Vx12 + T12
##   Res.Df   RSS Df Sum of Sq    F   Pr(>F)
## 1    110 33948
## 2    108 29881  2    4067.1 7.35 0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Maintenant revenons à l'exemple concernant les eucalyptus. Nous avions remarqué, d'après le graphique, que le lien ne semblait pas vraiment linéaire (concavité du nuage de points), d'où

l'idée d'introduire la variable `sqrt(circ)` (i.e. $\sqrt{\texttt{circ}}$). Comme la courbe que forme le nuage de points ne semble être ni complètement une droite ni complètement le graphe de la fonction racine carrée, on a l'idée de faire un modèle avec les deux variables explicatives `circ` et `sqrt(circ)`. On appelle ce modèle, `modele1`. Le modèle de départ avec seulement la variable explicative sera appelé `modele0`. Le modèle0 est donc un cas particulier du modèle1. On parle de modèles emboités.
Pour comparer ces deux modèles, nous utilisons la fonction `anova`.

Code R : test d'un sous modèle (2)

```
modele0=lm(ht~circ,data=euca)
modele1=lm(ht~circ+I(sqrt(circ)),data=euca)
anova(modele0,modele1)
```

```
## Analysis of Variance Table
##
## Model 1: ht ~ circ
## Model 2: ht ~ circ + I(sqrt(circ))
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   1427 2052.1
## 2   1426 1840.7  1    211.43 163.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous obtenons une valeur très faible pour la p-valeur ($<$ 2.2e-16). Donc nous rejetons $\mathcal{H}_0$, autrement dit, nous choisissons le modèle incluant `sqrt(circ)`.
Remarquez que dans ce cas particulier, on aurait pu faire le test de student de pertinence de la variable explicative `sqrt(circ)` dans le modèle `modele1`. Ce sont des tests totalement équivalents.

```
summary(modele1)
```

```
##
## Call:
## lm(formula = ht ~ circ + I(sqrt(circ)), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1881 -0.6881  0.0427  0.7927  3.7481
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -24.35200    2.61444  -9.314   <2e-16 ***
## circ           -0.48295    0.05793  -8.336   <2e-16 ***
## I(sqrt(circ))   9.98689    0.78033  12.798   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.136 on 1426 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7919
## F-statistic:  2718 on 2 and 1426 DF,  p-value: < 2.2e-16
```

Remarque 3.8. *On peut aussi faire des tests sur plusieurs combinaisons linéaires des coefficients. Exemple avec $p = 4$ et des variables explicatives nommées* `x2`, `x3` *et* `x4`

$$\begin{cases} \beta_2 &= \beta_3 \\ 2\beta_3 + \beta_4 &= 1 \end{cases}$$

que l'on écrit

$$\begin{cases} \beta_2 - \beta_3 &= 0 \\ -1 + 2\beta_3 + \beta_4 &= 0 \end{cases}$$

Autrement dit aussi

$$\begin{cases} 0 + 1 \times \beta_2 - 1 \times \beta_3 + 0 \times \beta_4 &= 0 \\ -1 + 0 \times \beta_2 + 2\beta_3 + 1 \times \beta_4 &= 0 \end{cases}$$

Pour cela il suffit d'écrire la matrice associée à ce système d'équations, ici $\begin{pmatrix} 0 & 1 & -1 & 0 \\ -1 & 0 & 2 & 1 \end{pmatrix}$, *que l'on donne en argument à la fonction* `linearHypothesis`. *Le code est le suivant*

```
mat=matrix(c(0,1,-1,0,-1,0,2,1),nrow=2,byrow=T)
library(car)
linearHypothesis(lm(y~x2+x3+x4),mat)
```

3.7.3 Test du modèle ou test de Fisher global

Parmi tous les tests de sous-modèle, il y en a un en particulier que l'on fait à chaque fois, avant toute étude plus poussée, c'est le *test du modèle* global, qui est en fait le test de la pertinence de l'ensemble de variables explicatives. En d'autres termes, on veut comparer le modèle $\mathcal{M}_1$ faisant intervenir l'ensemble des variables explicatives, avec le modèle $\mathcal{M}_0$ ne faisant intervenir que l'intercept. Ce test est fait systématiquement par R et est donné dans le résumé (`summary`). Le test apparait à la dernière ligne et la statistique de test est appelée `F-statistic`. Il correspond donc au test suivant

$$\mathcal{H}_0 : \beta_2 = \ldots = \beta_p = 0 \text{ contre } \mathcal{H}_1 : \exists j \in \{2, \ldots, p\} : \ \beta_j \neq 0,$$

Cela correspond donc à faire un test sur la nullité de $q = p - 1$ variables explicatives (toutes les variables sauf l'intercept $x^1 = \mathbf{1}$) . La projection orthogonale $\hat{y}_0$ est dans ce cas la projection orthogonale sur le vecteur $\mathbf{1}$: elle est égale à $\bar{y}\mathbf{1}$. Ce test est donc associé à la statistique

$$F = \frac{(\text{SCR} - \text{SCR}_0)/(p-1)}{\|Y - \hat{y}\|^2/(n-p)} > q_{\mathcal{F}_{p-1,n-p}^{1-\alpha}}$$

qui suit une loi de Fisher à $p - 1$ et $n - p$ degrés de libertés.
Dans l'exemple précédent lié aux eucalyptus, avec les variables explicatives `circ` et `sqrt(circ)`, on obtient une valeur de la statistique F égale à à 2718. Ici $p = 3$ (les deux variables `circ` et

`sqrt(circ)` plus l'intercept). Les degrés de libertés sont 2 et $n-3$. Ici $n = 1429$ (qu'on peut vérifier avec `nrow(euca)`) Donc les degrés de libertés sont 2 et 1426 (`"on 2 and 1426 DF"`). L'information réellement utile est la p-valeur : ici elle est inférieure à `2.2e-16` et donc on rejette $\mathcal{H}_0$. Autrement dit, au moins l'une des variables explicatives est pertinente.

La démarche à adopter pour les tests.
- On commence par regarder le résultat du test du modèle. Si on rejette $\mathcal{H}_0$ alors au moins une des variables a un effet.
- On regarde ensuite variables par variables.
- On peut alors s'intéresser à des sous-modèles.
- Pour sélection une sous-ensemble des variables les plus pertinentes, il faut plutôt adopter une approche "Slection de modèles" qui sera vu dans un cours ultérieur

3.8 Ajustement du modèle aux données

3.8.1 Coefficient de détermination

Somme des carrés Sous le modèle défini par $X^{(0)} = \mathbf{1}_n$ (modèle sans covariable), $\hat{\beta}_0 = \bar{\boldsymbol{y}}$. Donc $\hat{y}_i = \bar{\boldsymbol{y}}$. Et

$$\mathrm{SCR}_0 = \sum_{i=1}^{n} (y_i - \bar{\boldsymbol{y}})^2 = \mathrm{SCT}$$

C'est la somme des carrés totale, autrement dit la variabilité des données. On peut montrer par le théorème de Pythagore que
a

$$SCT = \mathrm{SCM} + \mathrm{SCR} \quad \text{où} \quad \mathrm{SCM} = \sum_{i}^{n} (\hat{y}_i - \bar{\boldsymbol{y}})^2 \quad \text{et} \quad \mathrm{SCR} = \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

SCM est la somme des carrés du modèle, autrement dit, la variabilité expliquée par le modèle. SCR est la variabilité non-expliquée par le modèle. Plus SCM est grande par rapport à SCT , plus le modèle explique la variabilité des observations.

R^2 D'après la remarque précédente, il est intéressant de regarder le *coefficient de détermination*

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

qui donne le ratio entre la variabilité expliquée par le modèle et la variabilité totale. $R^2 \in [0, 1]$.

Limites du R^2 R^2 ne peut être utilisé pour sélectionner les covariables pertinentes car $q \mapsto R^2_q$ est une fonction croissante de q.
En effet, comparons SCR_q et SCR_{q+1}. Pour tout $q = 1 \dots p$, on construit X_q la matrice composée des q premières colonnes de X.

$$\mathrm{SCR}_{q+1} = \min_{\beta_{q+1} \in \mathbb{R}^{q+1}} \|\boldsymbol{y} - X_{q+1}\beta_{q+1}\|^2 \leq \min_{\beta_q \in \mathbb{R}^q} \left\| \boldsymbol{y} - X_{q+1} \begin{pmatrix} \beta_q \\ 0 \end{pmatrix} \right\|^2 = \min_{\beta_q \in \mathbb{R}^q} \|\boldsymbol{y} - X_q\beta_q\|^2 = \mathrm{SCR}_q$$

Par conséquent le coefficient de détermination augmente quelque soit la variable incluse. Dans le jeu de données ozone, si j'ajoute la covariable "nombre de naissances du jour" , j'ajusterai mieux mes observations alors que la covariable n'a aucun caractère explicatif du phénomène biologique. Par conséquent, le R^2 ne peut pas être utilisé tel quel pour la comparaison de modèles de tailles différentes ou la sélection des variables pertinentes. Cependant, le R^2 peut être intéressant pour comparer des modèles de même dimension.
En résumé :

$$0 \leq R^2 \leq 1$$

Si R^2 n'est pas assez proche de 1 alors cela signifie que le modèle n'approche pas bien y : soit il manque une variable explicative, qu'il faudrait donc introduire dans le modèle, soit l'une (ou plusieurs) des variables explicatives n'intervient pas de manière linéaire.

Code R : coefficient de détermination

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

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

Il est donc possible que certaines variables explicatives aient pu être oubliées ou que la relation ne soit pas linéaire avec l'une des variables explicatives.
A noter que sous **R**, le R^2 est donné dans le summary sous le nom

```
Multiple R-squared:  0.6968
```

R fournit une version corrigée du R^2 pour tenter de prendre en compte la dimension du modèle.

3.8.2 R^2 ajusté

Le coefficient du R^2 ajusté, que l'on va noter R^2_a est une modification du R^2 qui tient compte du nombre de variables. On définit le R^2 ajusté par

$$R^2_a = 1 - \frac{n-1}{n-p}\frac{SCR}{SCT}.$$

Autrement dit, R^2_a est la valeur amoindrie du R^2 : on a d'autant plus abaissé la valeur du R^2 que le nombre p de variables explicatives du modèle est grand.

Remarque 3.9. *Il existe bien d'autres critères de choix de modèles (non traités dans ce cours, cf cours ultérieur, exemples : BIC, AIC).*

Code
Le R^2 ajusté est également donné par `summary`. C'est le `"Adjusted R-squared"`.

Exercice 3.3. *Revenons à l'exemple lié aux eucalyptus.*

1. *Entre le modèle avec comme variable explicative* `circ`*, plus l'intercept, et le modèle avec comme variable explicative* `sqrt(circ)`*, plus l'intercept, lequel faut-il choisir ?*
2. *Entre le modèle que nous avons trouvé à la question précédente et le modèle avec les deux variables explicatives* `circ` *et* `sqrt(circ)`*, plus l'intercept, lequel faut-il choisir ?*

Exercice 3.4. *1. Importer le fichier donn.csv dans un dataframe appelé donn. Puis effectuer la régression de la première colonne y sur toutes les autres.*

2. Quelles semblent être les variables pertinentes ? Vérifier que le modèle obtenu en enlevant les variables déclarées non pertinentes par les tests de Student est meilleur que le modèle complet. Vérifier qu'il n'y a pas de problèmes de corrélation.

3. Comment améliorer encore cette régression ?

Remarque 3.10. *On peut en fait choisir comme variable explicative, toute transformée $g(x)$ d'une variable explicative existante. Par exemple, on peut prendre $g(x) = x^2$ ou de manière générale $g(x) = x^k$. De cette façon on peut, en ayant une seule variable explicative, approcher toute fonction continue f qui relirait y à x (la "fonction de régression") par un polynôme en la variable x en faisant une régression multiple : x^1 est l'intercept, x^2 la variable x, x^3 la variable $(x)^2$, x^4 la variable $(x)^3$ etc. On peut évidemment aussi choisir une autre base que la base des polynômes. Tout le problème étant ensuite de savoir à quel degré du polynôme s'arrêter et éventuellement aussi quelle base choisir (cf par exemple le cours de non-paramétrique). Pour faire la régression de y sur un polynôme en x de degré 5 par exemple, on utilise*

```
lm(y~poly(x,5)}.
```

Remarque 3.11. *De la même manière, on peut aussi introduire toute transformée de deux variables, ou même de plusieurs variables. Par exemple, si on dispose de deux variables explicatives x^2 et x^3 on peut former une quatrième variable $x^4 = x^2x^3$. On peut ainsi modéliser tous les polynômes à plusieurs variables.*
Si nous disposons d'un dataframe avec pour colonnes y, $x2$ et $x3$ et que nous souhaitons faire la régression de y sur $x2$, $x3$ et $x4 = x1 \times x2$, il suffit de faire

```
lm(y~x2+x3+x2:x3)
```

3.9 Feuille de route pour l'analyse de données par une régression multiple

1. Charger les données, vérifier que les variables sont bien de la nature attendue (qualitative / quantitative)
2. Ecrire le modèle linéaire
3. Valider le modèle par la lecture des graphes de résidus. Si problème, tenter de prendre le $\log(y)$ ou le log de certaines covariables
4. Faire le test du modèle global : si rejet on arrête là, le modèle linéaire n'est pas adapté.
5. Ensuite on peut tester les paramètres, faire de la prédiction.

3.10 Exercices récapitulatifs

Exercice 3.5. *[Abondance de papillons]*
On s'intéresse à l'abondance d'une espèce menacée de papillons. Pour cela, on a mesuré sur $n = 100$ *unités géographiques (supposées indépendantes) d'un part un indice d'abondance et d'autres part plusieurs indices écologiques (indice d'humidité, indice d'altitude et indice d'orientation). Les données sont représentées dans les graphiques de la Figure 3.1*

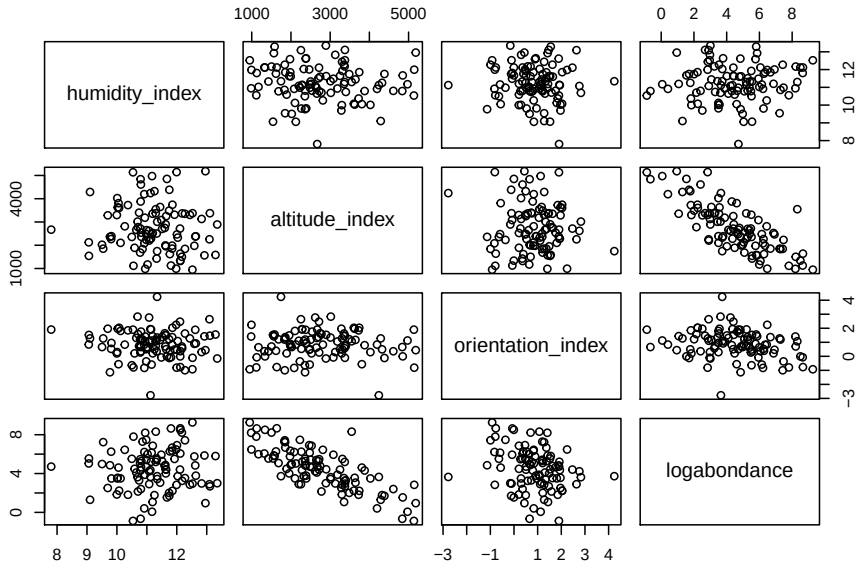


FIGURE 3.1 – Exercice Papillons : Représentation graphique des données

On cherche à modéliser le logarithme de l'abondance en fonction d'une combinaison linéaire des variables écologiques. Pour cela, on utilise la commande suivante :

```
lm_estim <- lm(logabondance~ humidity_index + altitude_index + orientation_index,
data=data_butterflies)
```

1. *Ecrire le modèle statistique correspondant aux commandes R précédentes. Rappeler ses hypothèses.*
2. *Commentez les graphes des résidus donnés dans la Figure 3.2. Les 4 postulats sont ils vérifiés ? Que pouvez-vous dire de l'observation* $i = 3$ *?*

   Les résultats de l'estimation sont donnés ci-dessous.

   ```
   summary(lm_estim)
   ```

   ```
   ##
   ## Call:
   ## lm(formula = logabondance ~ humidity_index + altitude_index +
   ##     orientation_index, data = data_butterflies)
   ##
   ## Residuals:
   ##     Min      1Q  Median      3Q     Max
   ```

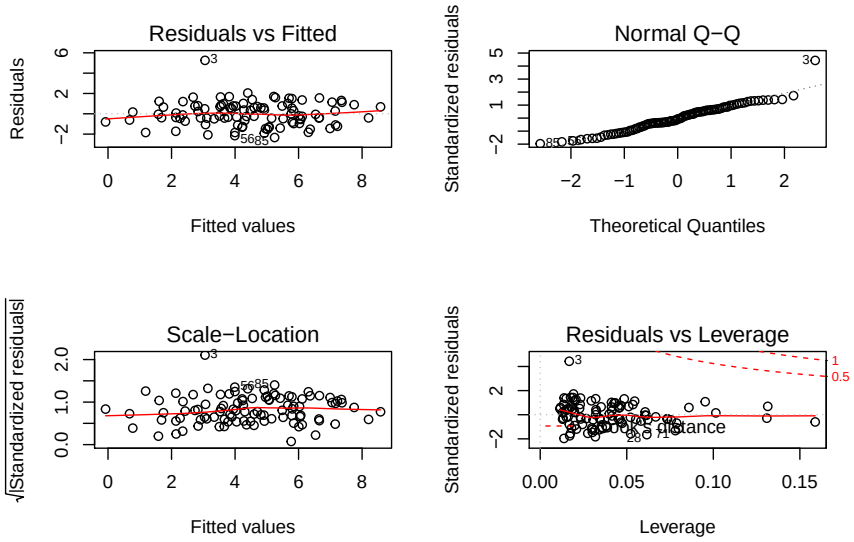


FIGURE 3.2 – Exercice Papillons : Graphes de résidus

```
## -2.3368 -0.7446 -0.0876  0.7586  5.2587
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.4345886  1.4081957   6.700 1.42e-09 ***
## humidity_index     0.0106327  0.1182602   0.090    0.929
## altitude_index    -0.0016437  0.0001171 -14.034  < 2e-16 ***
## orientation_index -0.6217292  0.1196374  -5.197 1.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 96 degrees of freedom
## Multiple R-squared:  0.6956, Adjusted R-squared:  0.6861
## F-statistic: 73.11 on 3 and 96 DF,  p-value: < 2.2e-16
```

3. *On cherche à tester le modèle. Rappelez la définition de ce test. Rappelez l'expression de la statistique de test ainsi que sa loi sous l'hypothèse $\mathcal{H}_0$. Retrouvez les degrés de libertés donnés dans les sorties R précédentes. Donner la conclusion du test du modèle.*
4. *Interpréter la quantité* `Pr(>|t|)` *pour la covariable* `humidity index`.
5. *Donnez l'estimation de σ^2 pour le modèle contenant toutes les covariables.*
6. *On souhaite construire un intervalle de confiance pour la quantité $x_3\beta$ où x_3 sont les covariables de la parcelle 3.*
   (a) *Rappeler la loi de $\hat{\boldsymbol{\beta}}$ dans le cas du modèle linéaire gaussien.*
   (b) *En déduire la loi de $x_3\hat{\beta}$.*
   (c) *Construire alors un intervalle de confiance pour $x_3\beta$.*
   (d) *On obtient la fourchette $[2.747, 3.362]$ alors qu'on avait observé $y_3 = 8.313263$. Commentez.*

Exercice 3.6. [Performances de processeurs] *Nous nous intéressons aux performances de* 206 *processeurs (Central Processor Unity, CPU en anglais), en fonction de diverses covariables listées ci-dessous :*

- *syct : cycle time in nanoseconds.*
- *mmin : minimum main memory in kilobytes.*
- *mmax : maximum main memory in kilobytes.*
- *cach : cache size in kilobytes.*
- *chmin : minimum number of channels.*
- *chmax : maximum number of channels.*
- *perf : published performance on a benchmark mix relative to an IBM 370/158-3.*
- *estperf : estimated performance (by Ein-Dor et Feldmesser).*

Nous donnons ci-dessous les trois premières lignes du tableau de données :

```
##                syct mmin  mmax cach chmin chmax perf estperf
## ADVISOR 32/60  125  256  6000  256    16   128  198     199
## AMDAHL 470V/7   29 8000 32000   32     8    32  269     253
## AMDAHL 470/7A   29 8000 32000   32     8    32  220     253
```

On cherche à modéliser la performance `perf` *comme une combinaison linéaire des covariables* `syct`, `mmin`, `mmax`, `cach`, `chmin` *et* `chmax`. *Le modèle est implémenté sous R par le code suivant :*

```
res_lm = lm(perf ~ syct + mmin + mmax + cach + chmin + chmax, data=data.cpu)
```

Les sorties R sont données ci dessous.

1. *Ecrire le modèle correspondant à l'instruction R précédente (on donnera la taille des objets).*
2. *Pourquoi pensez-vous que la variable* `estperf` *n'a pas été incluse dans le modèle. On pourra utiliser les corrélations entre les variables, représentées à la page 72 de l'énoncé ([Graphe 1]).*
3. *Les hypothèses sur les résidus sont-elles respectées ? Indiquer le graphique utilisé et jutifier les réponses.*
4. *Donner les hypothèses du test du modèle global. Rappeler l'expression de la statistique de test, sa loi sous $\mathcal{H}_0$. Donner la valeur de sa réalisation ici. Conclure*
5. *Quel est le test fait sur la ligne* `mmin` *du* `summary` *(page 73, [Instruction 1]).*
6. *Quel est le modèle estimé par l'instruction R suivante ?*

```
res_lm_0 = lm(perf ~ 1, data=data.cpu)
```

Commandes et sorties R pour l'exercice 3.6

- *[Graphe 1] Représentation des corrélations : plus l'ellipse ressemble à un cercle et moins les variables sont corrélées. Plus l'ellipse ressemble à une droite et plus les variables sont corrélées.*
- *[Graphe 2]*

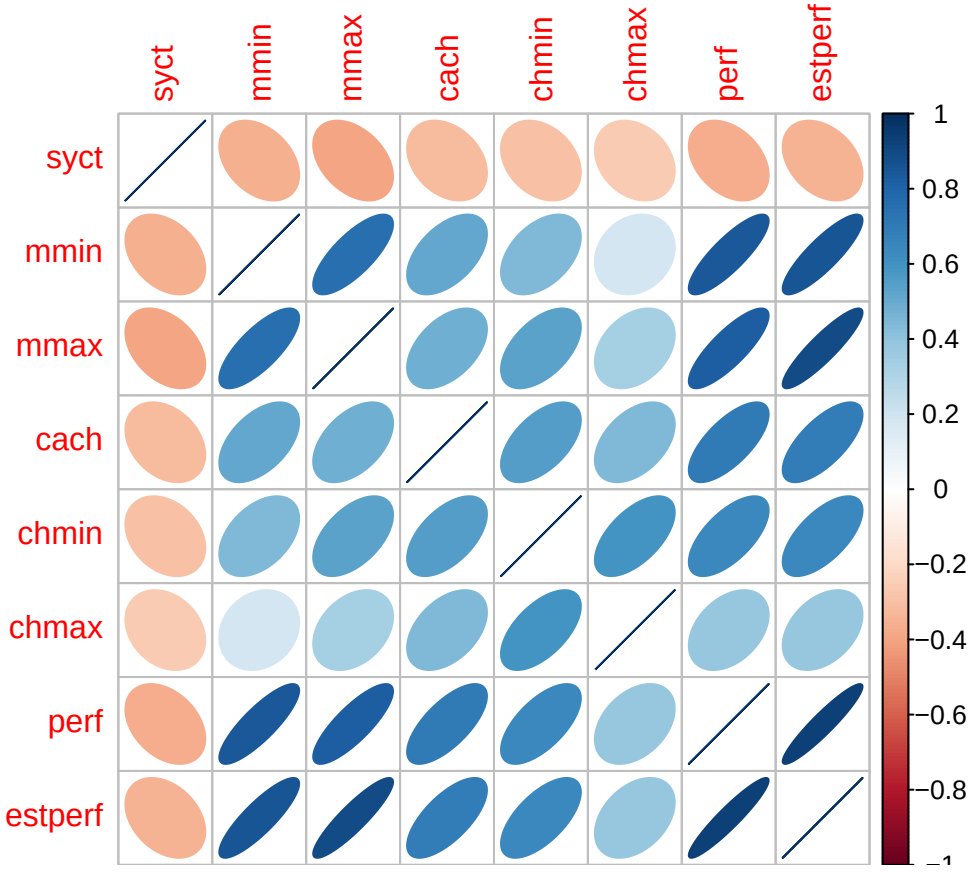


FIGURE 3.3 – Exercice Performances : représentation des corrélations entre variables

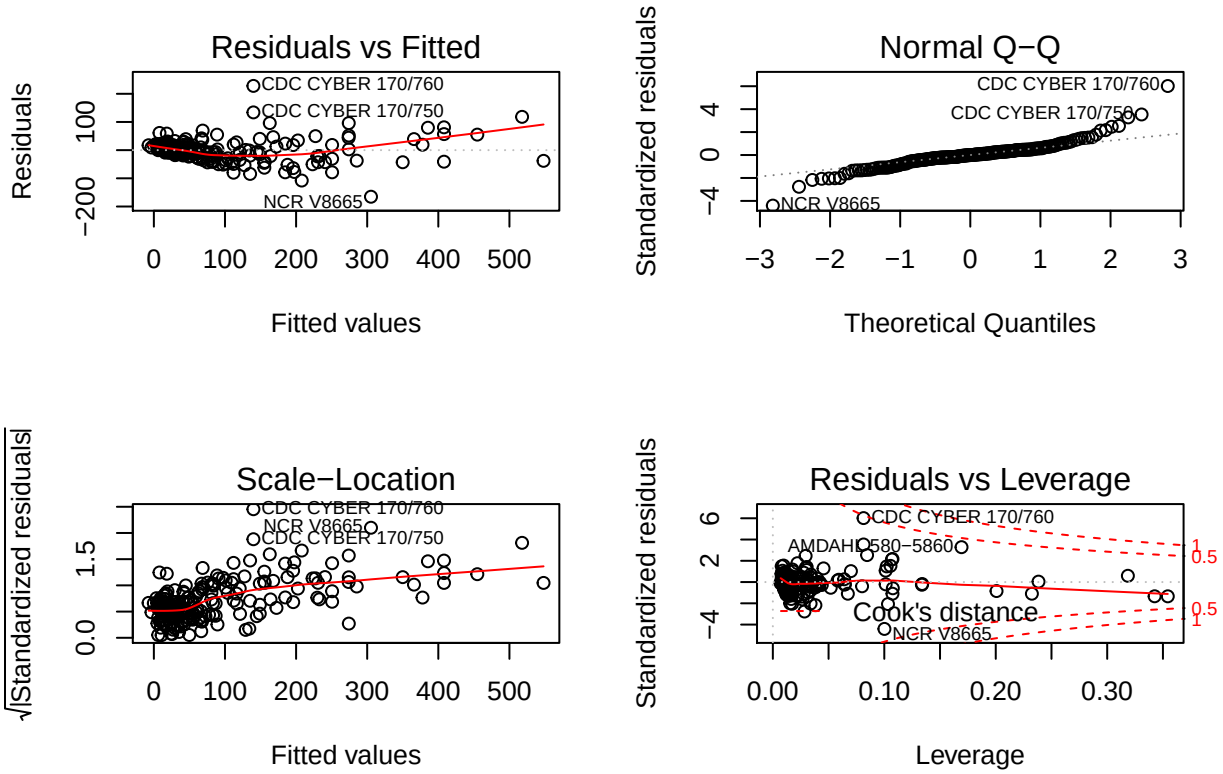


FIGURE 3.4 – Exercice Performances : graphes des résidus

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

- *[Instruction 1]*

```
summary(res_lm)
```

```
##
## Call:
## lm(formula = perf ~ syct + mmin + mmax + cach + chmin + chmax,
##     data = data.cpu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -165.328  -16.471   -0.174   16.642  228.007
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.697e+01  5.855e+00  -2.898  0.00418 **
## syct         1.608e-02  1.175e-02   1.368  0.17276
## mmin         1.394e-02  1.375e-03  10.138  < 2e-16 ***
## mmax         3.431e-03  4.607e-04   7.448 2.82e-12 ***
```

```
## cach         7.285e-01  9.318e-02   7.818 3.06e-13 ***
## chmin        2.828e+00  6.094e-01   4.641 6.27e-06 ***
## chmax       -2.561e-03  1.743e-01  -0.015  0.98829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.54 on 199 degrees of freedom
## Multiple R-squared:  0.8764, Adjusted R-squared:  0.8726
## F-statistic: 235.1 on 6 and 199 DF,  p-value: < 2.2e-16
```

Exercice 3.7. [Evaluation des stocks de carbone]
L'estimation des stocks de carbone en forêt tropicale repose sur l'estimation de la biomasse aérienne des arbres (AGB : Above Ground Biomass). Pour estimer cette grandeur, on utilise des relations allométriques qui lient la valeur AGB au diamètre D, à la hauteur H et à la densité des arbres ρ. Des arguments géométriques simples suggèrent que la grandeur AGB pour un arbre de diamètre D doit être proportionnelle au produit de la densité spécifique ρ de l'arbre par la surface basale du tronc ($\pi * D^2/4$) et la hauteur H, ce qui conduit à une expression générique

$$AGB = F \times \rho \times \left(\frac{\pi D^2}{4} \times H \right), \tag{3.6}$$

où F est un coefficient de proportionnalité inconnu.
Cette relation suppose que le tronc est comparable à un cylindre de densité homogène. Pour alléger cette hypothèse, la relation allométrique la plus couramment suggérée, notamment dans **?**, *est˜ :*

$$AGB = F \times \left(\rho \times \left(\frac{\pi D^2}{4} \times H \right) \right)^{\beta}, \tag{3.7}$$

où β est un coefficient à estimer.
Pour estimer la quantité de biomasse contenue dans une parcelle plantée d'arbres de la même espèce, on mesure pour chaque arbre, son diamètre, sa hauteur et sa densité et on mesure après séchage sa biomasse au-dessus du sol. Nous disposons de 60 mesures d'arbres.
Le diamètre, comme la hauteur sont exprimés en mètres. Nous avons également ajouté les logarithmes de ces variables : `DLog`$= \log(D)$, `HLog`$= \log(H)$, `rhoLog`$= \log(\rho)$, `AGBLog`$= \log(AGB)$.
Un extrait des données est fourni par

```
head(AGBParcelle, n=5)
```

```
##      D  H  rho  AGB        DLog     HLog     rhoLog    AGBLog
## 1 0.90 53 0.46 2.41 -0.10536052 3.970292 -0.7765288 0.8796267
## 2 0.86 53 0.57 2.48 -0.15082289 3.970292 -0.5621189 0.9082586
## 3 0.64 32 0.59 1.15 -0.44628710 3.465736 -0.5276327 0.1397619
## 4 0.77 49 0.33 1.35 -0.26136476 3.891820 -1.1086626 0.3001046
## 5 0.97 66 0.54 3.35 -0.03045921 4.189655 -0.6161861 1.2089603
```

1. On se propose de modéliser le logarithme de AGB (`AGBLog`) en fonction du logarithme du diamètre (`DLog`), du logarithme de la hauteur (`HLog`) et du logarithme de la densité (`rhoLog`). Écrire le modèle linéaire associé $\mathcal{M}_1$ en spécifiant toutes les hypothèses.
2. Écrire ce même modèle sous forme matricielle, en donnant les dimensions des différents objets considérés.
3. Si la relation (3.6) est vérifiée, quelles seront les valeurs attendues pour les paramètres du modèle $\mathcal{M}_1$?
On ajuste le modèle $\mathcal{M}_1$ et on obtient les sorties suivantes˜ :

```
M1 <- lm(AGBLog~DLog+HLog+rhoLog, data = AGBParcelle )
summary(M1)
```

```
##
## Call:
## lm(formula = AGBLog ~ DLog + HLog + rhoLog, data = AGBParcelle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37482 -0.12748 -0.00588  0.13141  0.31588
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16282    1.05758  -0.154   0.8782
## DLog         2.36365    0.33010   7.160 1.88e-09 ***
## HLog         0.43195    0.25319   1.706   0.0935 .
## rhoLog       0.48035    0.07981   6.019 1.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1849 on 56 degrees of freedom
## Multiple R-squared:  0.9354, Adjusted R-squared:  0.9319
## F-statistic: 270.3 on 3 and 56 DF,  p-value: < 2.2e-16
```

```
M0 <- lm(AGBLog~1, data = AGBParcelle )
anova(M0,M1)
```

```
## Analysis of Variance Table
##
## Model 1: AGBLog ~ 1
## Model 2: AGBLog ~ DLog + HLog + rhoLog
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     59 29.6366
## 2     56  1.9146  3    27.722 270.28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

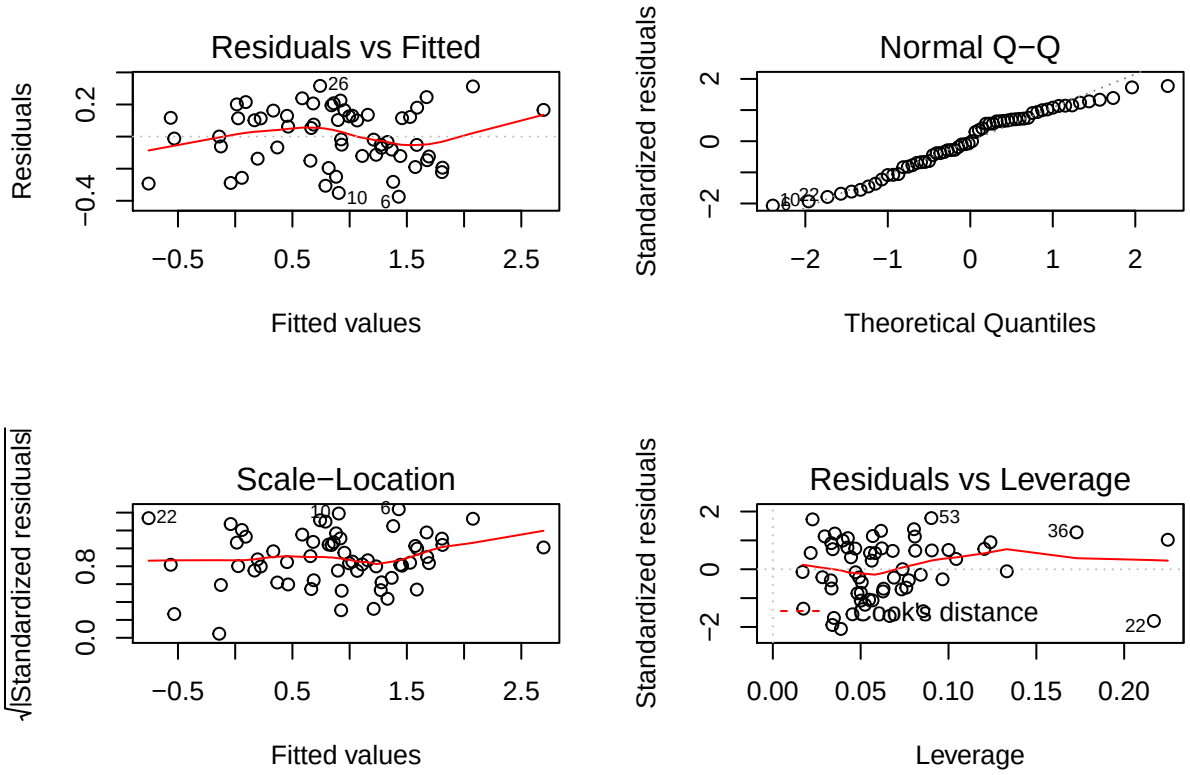


FIGURE 3.5 – Exercice Arbres : graphes des résidus

```
par(mfrow=c(2,2))
plot(M1)
```

4. Les hypothèses du modèle linéaire sont-elles vérifiées ? Préciser quel graphique est utilisé pour chaque hypothèse.
5. Donner les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test du modèle complet. Donner l'expression de la statistique de test et sa loi sous $\mathcal{H}_0$. On spécifiera les degrés de liberté. Donner la réalisation de la statistique de test sur ces donnée ainsi que la p-value. Conclure.
6. Donner l'expression de l'estimateur de la variance résiduelle et la valeur de cette estimation.
7. Quelle est l'hypothèse $\mathcal{H}_0$ du test proposé sur la ligne `Dlog` *de la commande* `summary` *? Comment l'interprète-t-on ? Quelle est la loi de la statistique de test ?*
8. Que pouvez-vous dire de l'ajustement du modèle aux données ?
9. Quel test pourriez-vous proposer pour tester si l'effet du diamètre est bien celui prévu par la relation allométrique de l'équation (3.6) *?*
On pourra utiliser la fonction linearHypothesis *du package* car *(voir page 65).*

Exercice 3.8. [Taux de mortalité] *On s'intéresse au taux de mortalité en fonction de facteurs économiques et environnementaux.*

```
names(death_data)
```

```
##  [1] "Precipitation"            "January_temperature"
##  [3] "July_temperature"         "percent_65_or_older"
##  [5] "household_size"           "schooling_over_22"
##  [7] "full_kitchens"            "urban_population_density"
##  [9] "nonwhite_population"      "office_workers"
## [11] "poor_families"            "hydrocarbons"
## [13] "oxides_of_Nitrogen"       "Sulfur_Dioxide"
## [15] "humidity"                 "death_rate"
```

On cherche à comprendre l'influence des caractéristiques sociologiques et environnementales ci-dessous sur le taux de mortalité. Pour cela on utilise la commande ci-dessous.

```
death_lm = lm(death_rate ~ Precipitation + January_temperature + July_temperature +
    percent_65_or_older + household_size + schooling_over_22 + full_kitchens +
    urban_population_density + nonwhite_population + office_workers + poor_families +
    hydrocarbons + oxides_of_Nitrogen + Sulfur_Dioxide + humidity, data = death_data)
```

1. *Ecrire le modèle correspondant aux instructions R données ci dessous. Rappelez ses hypothèses.*

```
summary(death_lm)
```

```
##
## Call:
## lm(formula = death_rate ~ Precipitation + January_temperature +
##     July_temperature + percent_65_or_older + household_size +
##     schooling_over_22 + full_kitchens + urban_population_density +
##     nonwhite_population + office_workers + poor_families + hydrocarbons +
##     oxides_of_Nitrogen + Sulfur_Dioxide + humidity, data = death_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.285 -14.640   0.694  14.790  75.586
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.863e+03  4.108e+02   4.535  4.4e-05 ***
## Precipitation             2.072e+00  8.418e-01   2.462  0.01781 *
## January_temperature      -2.178e+00  6.752e-01  -3.225  0.00238 **
## July_temperature         -2.834e+00  1.771e+00  -1.600  0.11670
## percent_65_or_older      -1.404e+01  7.746e+00  -1.813  0.07670 .
## household_size           -1.154e+02  6.200e+01  -1.862  0.06933 .
## schooling_over_22        -2.425e+01  1.121e+01  -2.163  0.03605 *
## full_kitchens            -1.146e+00  1.467e+00  -0.781  0.43871
```

```
## urban_population_density  1.004e-02  4.123e-03   2.435  0.01899 *
## nonwhite_population       3.533e+00  1.282e+00   2.755  0.00850 **
## office_workers            5.229e-01  1.551e+00   0.337  0.73760
## poor_families             2.671e-01  2.565e+00   0.104  0.91755
## hydrocarbons             -8.890e-01  4.524e-01  -1.965  0.05574 .
## oxides_of_Nitrogen        1.866e+00  9.345e-01   1.997  0.05201 .
## Sulfur_Dioxide           -3.447e-02  1.423e-01  -0.242  0.80968
## humidity                  5.331e-01  1.052e+00   0.507  0.61474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.33 on 44 degrees of freedom
## Multiple R-squared:  0.7985, Adjusted R-squared:  0.7298
## F-statistic: 11.63 on 15 and 44 DF,  p-value: 9.56e-11
```

2. *On cherche à tester le modèle. Rappelez la définition de ce test. Rappelez l'expression de la statistique de test ainsi que sa loi. Retrouvez les degrés de libertés donnés dans les sorties R précédentes.*
3. *On fournit aussi les sorties suivantes :*

```
death_lm_0 = lm(death_rate ~1,data=death_data)
anova(death_lm_0,death_lm)
```

```
## Analysis of Variance Table
##
## Model 1: death_rate ~ 1
## Model 2: death_rate ~ Precipitation + January_temperature + July_temperature +
##     percent_65_or_older + household_size + schooling_over_22 +
##     full_kitchens + urban_population_density + nonwhite_population +
##     office_workers + poor_families + hydrocarbons + oxides_of_Nitrogen +
##     Sulfur_Dioxide + humidity
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     59 228308
## 2     44  46001 15    182307 11.625 9.56e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir des sorties précédentes, donnez la valeur des sommes des carrées SCT, SCR et SCM, après avoir rappelé leurs définitions mathématiques.

4. *Donnez l'estimation de σ^2 pour le modèle contenant toutes les covariables.*
5. *On ajoute une covariable au modèle appelée* `add` *et on ré-estime les paramètres du modèle.*

```
summary(death_lm_add)
```

```
##
## Call:
## lm(formula = death_rate ~ add + Precipitation + January_temperature +
##     July_temperature + percent_65_or_older + household_size +
##     schooling_over_22 + full_kitchens + urban_population_density +
##     nonwhite_population + office_workers + poor_families + hydrocarbons +
##     oxides_of_Nitrogen + Sulfur_Dioxide + humidity, data = death_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.321 -15.734   1.795  15.682  68.796
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.755e+03  4.162e+02   4.218 0.000125 ***
## add                        6.669e+00  5.159e+00   1.293 0.202991
## Precipitation              2.245e+00  8.460e-01   2.654 0.011112 *
## January_temperature       -2.299e+00  6.766e-01  -3.397 0.001476 **
## July_temperature          -2.565e+00  1.770e+00  -1.449 0.154568
## percent_65_or_older       -1.553e+01  7.774e+00  -1.998 0.052043 .
## household_size            -1.131e+02  6.156e+01  -1.837 0.073139 .
## schooling_over_22         -2.581e+01  1.119e+01  -2.306 0.025999 *
## full_kitchens             -8.305e-01  1.476e+00  -0.563 0.576531
## urban_population_density   1.137e-02  4.218e-03   2.694 0.010019 *
## nonwhite_population        3.128e+00  1.311e+00   2.387 0.021473 *
## office_workers             6.977e-01  1.545e+00   0.451 0.653900
## poor_families              6.688e-01  2.565e+00   0.261 0.795522
## hydrocarbons              -1.022e+00  4.606e-01  -2.219 0.031826 *
## oxides_of_Nitrogen         2.176e+00  9.578e-01   2.271 0.028181 *
## Sulfur_Dioxide            -8.819e-02  1.472e-01  -0.599 0.552231
## humidity                   6.052e-01  1.045e+00   0.579 0.565598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.09 on 43 degrees of freedom
## Multiple R-squared:  0.8061, Adjusted R-squared:  0.7339
## F-statistic: 11.17 on 16 and 43 DF,  p-value: 1.574e-10
```

Comparez les sorties des deux modèles (estimation des paramètres, test de la nullité des paramètres...) Observez maintenant les R^2 du modèle $\mathcal{M}$ et de celui contenant la variable additionnelle. La variable ajoutée a en fait été simulée complètement au hasard. Commentez.

6. *On revient aux tests de paramètres β_k. Quelles sont les variables qui vous semblent pertinentes pour expliquer le taux de mortalité ? En particulier, les comparer aux corrélations représentées ci-desssous ?*

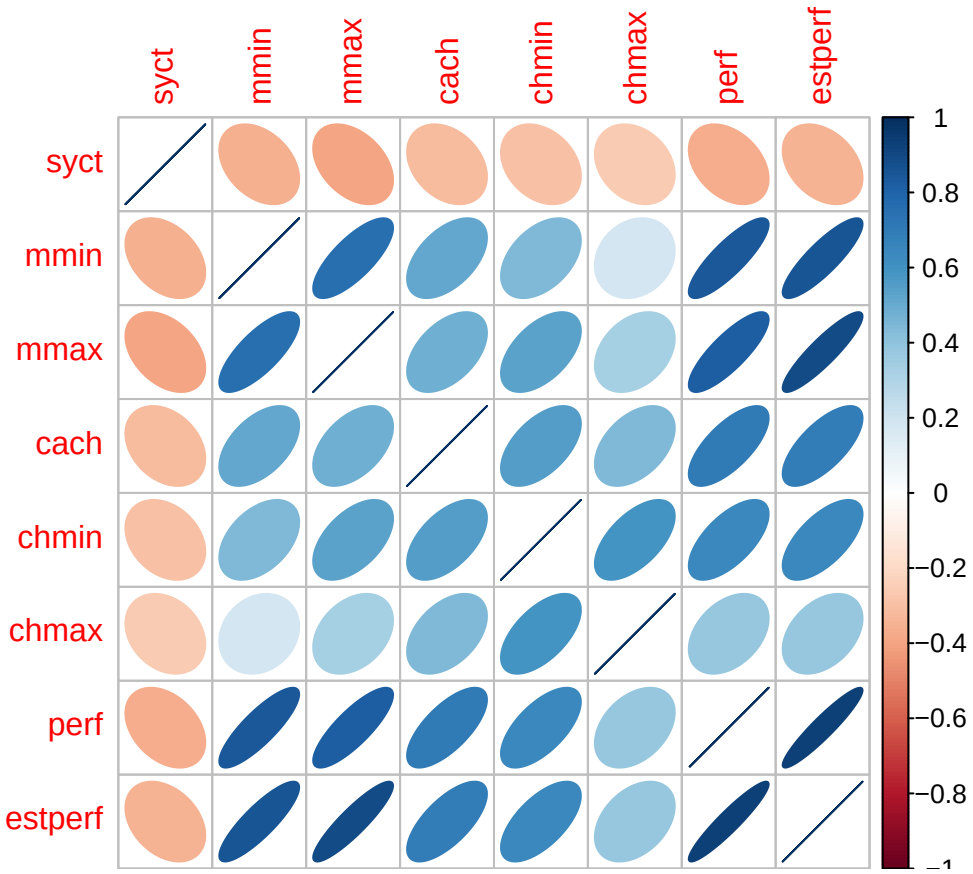


FIGURE 3.6 – Exercice mortalité : Correlations between variables

```
library(corrplot)
corrplot(cor(death_data),method = "ellipse")
```

Chapitre 4

Régression sur variables qualitatives

Jusqu'à présent, les variables explicatives étaient des variables quantitatives continues, or il arrive fréquemment que certaines variables explicatives soient des variables qualitatives. Dans ce cas, pouvons-nous appliquer la méthode des moindres carrés que nous venons de voir ?

Reprenons l'exemple des eucalyptus. Nous avons mesuré 1429 couples circonférence-hauteur. Parmi ces 1429 arbres, 527 proviennent d'une partie du champ notée bloc A1, 586 proviennent d'une autre partie du champ notée bloc A2 et 316 proviennent de la dernière partie du champ notée bloc A3. Le tableau suivant donne les 3 premières mesures effectuées dans chaque bloc :

Individu	`ht`	`circ`	`bloc`
1	18.25	36	A1
2	19.75	42	A1
3	16.5	33	A1
528	17	38	A2
529	18.5	46	A2
530	16.5	37	A2
1114	17.75	36	A3
1115	19.5	45	A3
1116	17.25	36	A3

Nous avons dorénavant deux variables explicatives : la circonférence et la provenance de l'arbre. La première est quantitative et la seconde est qualitative.

De manière générale, une variable qualitative est une variable qui ne prend qu'un nombre fini de valeurs (ces valeurs n'étant pas forcément des nombres, on peut les coder par des nombres cependant). Par exemple, la variable sexe ne prend que deux valeurs (F/M). (En langage R, on parle de facteurs (`factors`).)

Revenons sur notre exemple : la provenance pourrait avoir un effet sur la hauteur mais cela est difficile à observer. Afin d'intégrer la variable `bloc`, il faut commencer par la recoder car les calculs ne peuvent être effectués avec la variable en l'état. La variable `bloc` sera transformée en 3 nouvelles variables binaires (0/1) d'appartenance à chaque bloc. Cela donne, si on se restreint

aux 9 observations du tableau précédent,

$$\texttt{bloc} = \begin{pmatrix} A1 \\ A1 \\ A1 \\ A2 \\ A2 \\ A2 \\ A3 \\ A3 \\ A3 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

On a trois modalités au départ donc 3 "nouvelles" variables binaires.

Remarque 4.1. *On pourrait se poser la question de l'intérêt de ce codage : pourquoi ne pas coder 0/1 pour une variable à 2 modalités, 0/1/2 pour une variable à 3 modalités etc ? En fait, pour une variable à 2 modalités, ça a du sens, mais pour plus de modalités, ça pose problème : par exemple, imaginons que la variable qualitative soit associée à la richesse d'un individu : (riche,moyen,pauvre). Pourquoi coder 0/1/2 ? Peut-être que la différence entre riche et moyen est plus grande que la différence entre moyen et pauvre ! Pourquoi pas 0/1/3 ? ou autre...c'est pour éviter ce type de problème que l'on code l'appartenance à chaque catégorie.*

De manière générale, si la variable qualitative a J modalités, elle est en fait représentée par J variables binaires. Ce processus est fait automatiquement dans les logiciels.
Pour simplifier l'exposé des résultats, nous allons d'abord voir le cas d'une variable explicative qualitative seule. On parle alors d'analyse de la variance à un facteur. Nous verrons ensuite le cas de deux facteurs, puis le cas d'un facteur associé à une variable quantitative, comme l'exemple ci-dessus avec `bloc` et `circ`. Dans ce dernier cas "mixte", on parle d'analyse de la covariance.

4.1 Analyse de la variance à un facteur

Nous modélisons la concentration en ozone en fonction du vent (appelée aussi `vent` dans le fichier `ozone`) : on a les 4 modalités (EST, NORD, OUEST, SUD). Dans le tableau suivant figurent les dix premières lignes du tableau de données.

Individu	`max03`	`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

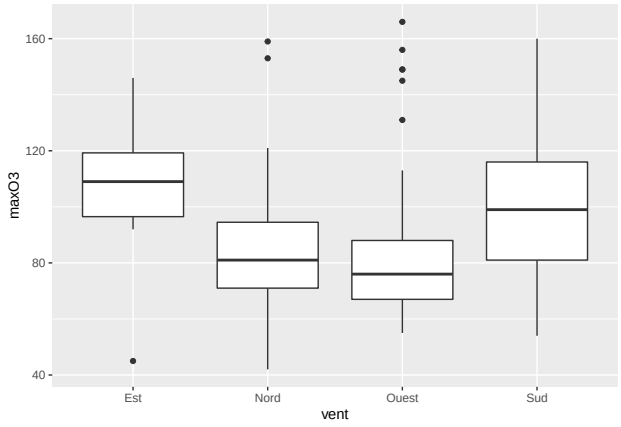


FIGURE 4.1 – boxplot de la variable vent

La première analyse à effectuer est une représentation graphique des données. Les boites à moustaches (boxplots) de la variable y par cellule semblent les plus adaptées à notre analyse, cf Figure 4.1. Cette figure est obtenue par la commande :

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

Au vu de ce graphique, il semblerait que le vent ait une influence sur la valeur de la concentration en ozone. La concentration est plus élevée en moyenne quand le vent vient de l'est ou du sud. Afin de préciser cette hypothèse, nous allons construire une analyse de variance à un facteur explicatif : le vent.

4.1.1 Modélisation du problème et modèle régulier

Supposons que la variable qualitative qui nous intéresse a I modalités. Chaque modalité est prise par n_i observations ($\sum_{i=1}^{I} n_i = n$). Par exemple pour les eucalyptus, on a $n_1 = 527$, $n_2 = 586$, $n_3 = 316$. Les observations sont alors regroupées par modalité, et la valeur de la variable à expliquer est notée avec deux indices : y_{ij} représente la valeur de la variable à expliquer pour le j-ème individu ayant pris la $i$ème modalité. Par exemple, pour l'ozone, si on prend l'ordre alphabétique (est, sud, ouest, nord)=(1,2,3,4), alors $y_{3,2}$ désigne la valeur de la concentration en ozone pour la 2ème journée où il y a eu un vent venant de l'ouest.
NB : dans le dataframe `euca`, les observations sont déjà regroupées par modalités, ce n'est pas le cas de `ozone`, mais ce regroupement n'est nécessaire ici que pour la modélisation mathématique, tout est fait automatiquement dans les logiciels.
On cherche à savoir si le facteur fait varier la variable y. On suppose que y_{ij} est la réalisation de Y_{ij} où

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i \in \{1, \dots I\},\ j \in \{1, \dots n_i\}$$

avec les mêmes hypothèses sur le bruit que précédemment, i.e.

$$\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

σ^2 étant inconnu, les paramètres μ_i, pour $i \in \{1, \dots, I\}$ inconnus aussi et fixes.

Remarque 4.2. Ce modèle est dit régulier *car si on écrit sa version matricielle, la matrice de design X^r est de rang plein (I colonnes et rang I). En effet, définissons la matrice suivante :*

$$X^r = \begin{pmatrix} 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & 1 & \\ & & \vdots & \\ & & 1 & \end{pmatrix}$$

Alors, on a l'écriture matricielle suivante :

$$\boldsymbol{Y} = X^r \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$$

Estimation des paramètres Naturellement, μ_i est estimé par $\widehat{\mu}_i = \frac{1}{n_i} = \sum_{j=1}^{n_i} y_{ij}$, que l'on note $\bar{y}_{i.}$, c'est-à-dire la moyenne des valeurs de la variable dans la modalité i. On obtient ainsi $\hat{y}_{ij} = \hat{\mu}_i$. σ^2 s'estime par $\frac{1}{n-I} \sum_{ij} (y_{ij} - \hat{\mu}_i)^2$

Code R : modèle d'Anova

Reprenons l'exemple de l'ozone.

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

```
##
## Call:
## lm(formula = maxO3 ~ -1 + vent, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.600 -16.807  -7.365  11.478  81.300
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## ventEst    105.600      8.639   12.22   <2e-16 ***
## ventNord    86.129      4.907   17.55   <2e-16 ***
## ventOuest   84.700      3.864   21.92   <2e-16 ***
## ventSud    102.524      5.962   17.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.9195, Adjusted R-squared:  0.9165
## F-statistic: 308.5 on 4 and 108 DF,  p-value: < 2.2e-16
```

On retrouve bien la sortie habituelles : les estimations, les écart-types, le R^2, les p-values des tests de Student et de Fisher.

Exercice 4.1. *Donner la prédiction du taux d'ozone pour une journée où le vent vient du sud.*

Cependant, ici le test de Fisher global effectué ici est de peu d'intérêt en général. En effet, la question principale qu'on se pose ici c'est : la provenance du vent a-t-elle une influence sur la concentration en ozone ? Cette question, avec l'introduction des variables binaires, se traduit par le test suivant :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{contre} \quad H_1 : \text{ les } \mu_j \text{ ne sont pas tous égaux}$$

Or ce n'est pas ce que fait le test global de Fisher, qui, on le rappelle, fait le test suivant :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0 \quad \text{contre} \quad H_1 : \text{ l'un des } \mu_j \text{ n'est pas nul}$$

La p-valeur de la F-statistic n'est donc pas très utile ici. Pour remédier à cela, il suffit d'introduire un effet moyen μ et de regarder, pour chaque modalité, s'il y a une différence avec l'effet moyen.

4.1.2 Version singulière du modèle d'Anova

On suppose que y_{ij} est la réalisation de Y_{ij} avec :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i \in \{1, \dots I\},\ j \in \{1, \dots n_i\}$$

On a en fait introduit un effet moyen μ et, pour chaque i, α_i l'effet de la $i$ème modalité sur la variable à expliquer y par rapport à l'effet moyen : $\mu_i = \mu + \alpha_i$ se lit aussi $\alpha_i = \mu_i - \mu$.
μ est vu comme un niveau de référence et α_i s'entend comme une différence par rapport au niveau de référence. Ce modèle compte $I + 1$ paramètres. Si on écrit la matrice X correspondante à ce modèle on s'aperçoit qu'elle n'est pas de rang plein.

$$X = \begin{pmatrix} 1 & 1 & & \\ 1 & \vdots & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & \vdots & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & \vdots \\ 1 & & & 1 \end{pmatrix}$$

Ce problème de rang est associé à un problème d'identifiabilité. En effet, on a $\mu_i = \mu + \alpha_i$ pour tout $i = 1, \dots, I$. On a I équations mais le modèle singulier contient $I + 1$ paramètres. Il faut en fait introduire une contrainte sur les paramètres, pour rendre le modèle identifiable, c'est-à-dire

pour que l'écriture soit unique. Cette contrainte est aussi associée à la question : qu'appelle-t-on l'effet moyen ?

Exemples de contraintes :

- La contrainte $\mu = 0$ correspond à la première présentation, c'est-à-dire au modèle sans intercept : α_i représente alors l'effet de la modalité i. On n'a pas vraiment introduit d'effet moyen.
- On met souvent la contrainte $\alpha_1 = 0$. Cela revient à prendre en référence la première modalité. Les autres α_i mesurent donc les différences des autres modalités avec cette modalité de référence. L'effet moyen est l'effet de la première modalité. **Cette contrainte est celle utilisée sous R**.
- On trouve aussi souvent la contrainte $\sum_{i=1}^{I} \alpha_i = 0$. L'effet moyen est alors la moyenne des effets de chaque modalité ($\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i.$). En effet, nous avons les équations :

$$\begin{cases} \mu + \alpha_1 &= \mu_1 \\ \quad \vdots & \\ \mu + \alpha_i &= \mu_i \\ \quad \vdots & \\ \mu + \alpha_I &= \mu_I \end{cases}$$

  Comme $\sum_{i=1}^{I} \alpha_i = 0$, on obtient (on sommant les I lignes), $I\mu + \sum_{i=1}^{I} \alpha_i = I\mu = \sum_i \mu_i$ donc

$$\mu = \frac{1}{I}\sum_i \mu_i.$$

- De manière générale, on peut mettre une contrainte linéaire quelconque (cf option `contrast`).

> Les résultats (en terme de prédiction, de tests d'effets...) seront les mêmes que l'on écrive le modèle sous une forme ou une autre, c'est-à-dire que l'on mette une contrainte ou une autre : **c'est l'interprétation des coefficients et donc des tests associés qui différera**.

On peut d'ailleurs retrouver les résultats d'une écriture avec les résultats d'une autre écriture. Considérons la deuxième contrainte, $\alpha_1 = 0$, c'est celle qui est celle faite par défaut dans R. Cette contrainte revient en fait à supposer que l'on met l'intercept, et que l'on code uniquement les trois dernières modalités de la variable `vent`. Cela revient aussi à prendre la première modalité comme référence. Le coefficient de l'intercept correspond donc à l'effet de la première modalité. Les autres coefficients mesurent les différences d'effet avec cette modalité.

Le test de Fisher du modèle global correspond à

$$\mathcal{H}_0 : \alpha_1 = 1, \ldots, \alpha_I = 0.$$

On est bien en train de se demander si le facteur entraine des différences siginificatives sur y en fonction de ses modalités.

Code R : test de l'effet du facteur

Vérifions avec le test global de Fisher l'influence de la provenance du vent sur la concentration en ozone :

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

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

La p-value est relativement petite (0.02074) donc H_0 est rejetée, autrement dit, on va considérer que la provenance du vent a bien une influence sur la concentration en ozone, ce qui est concordant avec le boxplot observé.

Dans le code, nous ne précisons rien : l'intercept est inclus automatiquement par R comme dans une régression quelconque, mais ici, étant donné qu'on a une variable qualitative, R propose de mettre une contrainte supplémentaire (option `contrast`), par défaut c'est la contrainte $\alpha_1 = 0$ (R considère l'ordre alphabétique `Est`=1ère modalité, `Nord`=2, `Ouest`=3ème, `Sud`=4ème).
On remarque aussi que les résultats sont bien les mêmes entre les deux contraintes : on peut retrouver les résultats de la première sortie (qui utilisait la contrainte $\mu = 0$) :

```
ventEst    105.600
ventNord    86.129
ventOuest   84.700
ventSud    102.524
```

à partir des résultats de la seconde (avec la contrainte $\alpha_1 = 0$) :

```
 (Intercept)  105.600
```

```
ventNord     -19.471
ventOuest    -20.900
ventSud       -3.076
```

On peut évidemment faire tous les types de test étudiés précédemment (par exemple $\alpha_1 = 0$, $\alpha_1 = 1$, $\mu = 3$, $\alpha_1 = \alpha_2$ etc).
On peut bien sûr prendre une autre modalité de référence que la première, pour prendre la seconde par exemple on utilise

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

SI on veut utiliser la contrainte $\sum_{i=1}^{I} \alpha_i = 0$, il suffit d'utiliser

```
lm(maxO3~C(vent,sum),data=ozone)
```

Remarque 4.3. *On constate que le R^2 est très petit. C'est normal car notre modèle est très pauvre.*

4.1.3 Validation du modèle

De la même façon, on va chercher à valider les hypothèses sur les résidus. On regarde les mêmes graphes (Figure 4.2).

Code R : validation du modèle

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

On constate qu'on n'a que 4 valeurs de $\hat{y}_{ij}$. C'est normal puisqu'on a vu qu'on prédisait y_{ij} par $\hat{\mu} + \hat{\alpha}_i$ donc on a bien I valeurs possibles.

4.1.4 Comparaison de traitements

Comparaison de deux traitements Supposons qu'on cherche à comparer si il y a une différence significative sur les valeurs de y entre les modalités i et i' du facteur.
On peut le traduire en terme de test : $\mathcal{H}_0 : \mu_i = \mu_{i'}$ versus $\mathcal{H}_1 : \mu_i \neq \mu_{i'}$.
Le test peut être fait en utilisant la statistique suivante :

$$T = \frac{\hat{\mu}_i - \hat{\mu}_{i'}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}} \sim_{\mathcal{H}_0} \mathcal{T}_{n-I}$$

Et ainsi on peut construire UN test de niveau α.

Problème des tests multiples On peut vouloir comparer les autres groupes (appelés *traitements*) entre eux. En effet, on peut chercher à identifier tous les couples (i, i') tels que $\mu_i \neq \mu_{i'}$. Ainsi, on va vouloir faire $I(I-1)/2$ tests. Supposons que l'on décide de faire tous les tests au niveau δ. Pour tous les couples (i, i'), on veut tester

$$\mathcal{H}_0^{ii'} : \mu_i = \mu_{i'} \quad \text{versus} \quad \mathcal{H}_0^{ii'} : \mu_i \neq \mu_{i'}$$

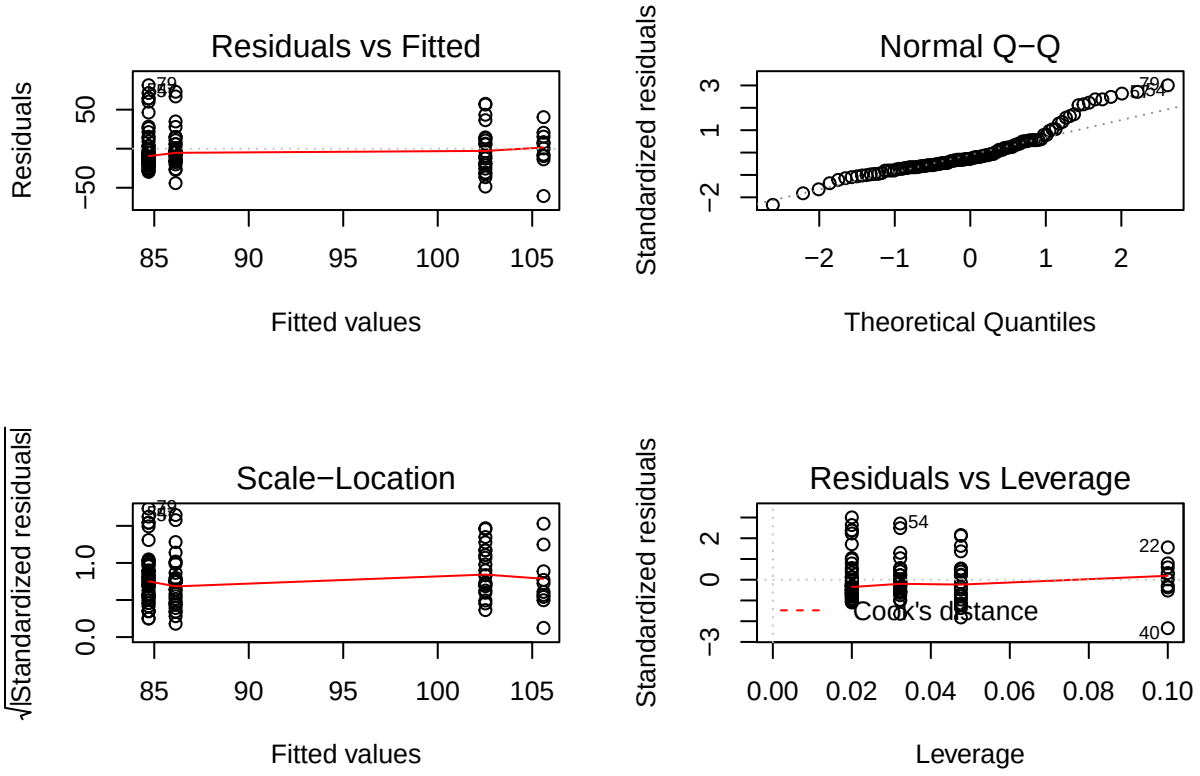


FIGURE 4.2 – Résidus de l'Anova pour le jeu de données Ozone

Pour chaque test, on contrôle l'erreur de première espèce, i.e. la probabilité de rejeter $\mathcal{H}_0^{ii'}$ alors que $\mathcal{H}_0^{ii'}$ est vraie (probabilité $\leq \delta$). Calculons la probabilité de se tromper au moins une fois :

$$\begin{aligned}\mathbb{P}\left(\text{rejeter au moins une } \mathcal{H}_0^{ii'} \text{ alors que } \mathcal{H}_0^{ii'} \text{ vraie}\right) &\leq \sum_{i,i',i<i'} \mathbb{P}_{\mathcal{H}_0^{ii'}}(\text{rejeter } \mathcal{H}_0^{ii'}) \\ &\leq \sum_{i,i',i<i'} \delta \\ &\leq \frac{I(I-1)}{2}\delta\end{aligned}$$

Ainsi, si $I = 7$ et $\delta = 5\%$, on borne la probabilité de se tromper au moins une fois par 1 ! Donc on n'a aucun contrôle. Pour palier à cela, il existe plusieurs méthodes. Une méthode classique est la méthode de Bonferroni qui consiste à recorriger le niveau de chaque test. Chaque test sera fait avec un niveau $2\delta/(I(I-1))$, atteignant ainsi au final un niveau global δ. Attention, de cette façon, il devient plus dur de rejeter les hypothèses nulles.

Code R pour comparer les traitements

Sans correction de Bonferroni

```
comp.statut = pairwise.t.test(ozone$max03,ozone$vent,p.adjust.method = "none")
comp.statut
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ozone$max03 and ozone$vent
##
##       Est   Nord  Ouest
## Nord  0.053 -     -
## Ouest 0.029 0.819 -
## Sud   0.770 0.036 0.014
##
## P value adjustment method: none
```

Avec correction de Bonferroni

```
pairwise.t.test(ozone$max03,ozone$vent,p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ozone$max03 and ozone$vent
##
##       Est   Nord  Ouest
## Nord  0.316 -     -
## Ouest 0.176 1.000 -
```

```
## Sud   1.000 0.216 0.082
##
## P value adjustment method: bonferroni
```

Exercice 4.2. *Introduisons quelques notations*

$$y_{..} = \frac{1}{n}\sum_{i,j} y_{ij}, \qquad y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Donner les estimateurs $\hat{\mu}$ et $(\hat{\alpha}_i)_{1\leq j\leq J}$ pour la contrainte $\sum_{i=1}^{1}\alpha_i = 0$ (utiliser le résultat de l'exercice 5.1).
Regarder le cas particulier où $n_i = J$ constant, i.e. il y a le même nombre d'individus dans chaque cellule. (On parle de plan "équirépété" ou plan "équilibré").

Exercice 4.3.
1. *On considère la régression de `max03` sur `Vent` avec la contrainte par défaut (cf sortie précédente). Que peut-on dire à partir des résultats des tests de Student ?*
2. *En observant les boxplots, on a la sensation que les vents du Nord et de l'Ouest ont le même effet. Faire le test correspondant pour vérifier ces deux hypothèses.*

Remarque 4.4. *En utilisant la décomposition*

$$y_{ij} - y_{..} = y_{ij} - \bar{y}_{i.} + y_{i.} - y_{..}$$

on peut montrer que

$$\sum_{i,j}(y_{ij} - y_{..})^2 = \sum_{i=1}^{I} n_j(y_{i.} - y_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})^2$$

La première somme est appelée la variation totale, la seconde la variation inter-groupes et la troisième est la variation intra-groupes.

Exercice 4.4. *On travaille ici avec le dataframe `euca`.*
1. *Faire les boxplots. Y-a-t-il une influence du bloc sur la hauteur ? Le vérifier par une régression.*
2. *La hauteur semble la même sur les deux premiers blocs. Est-ce le cas ?*
3. *Vérifier graphiquement l'hypothèse d'homoscédasticité.*

4.2 Analyse de la variance à deux facteurs

On considère à présent le cas de deux facteurs. Par exemple, nous pouvons modéliser la concentration en ozone par le `vent` (4 modalités) et la `temps` (2 modalités : pluie/sec).

Remarque 4.5. *On pourra en fait généraliser les résultats à un nombre quelconque de facteurs. Mais afin d'alléger les notations on se restreint à 2 facteurs*

4.2.1 Modèle régulier et singulier

Supposons que le premier facteur ait I modalités et le second J modalités. On note y_{ijk} la k-ème observation dans les modalités i pour le premier facteur et j pour le deuxième facteur. On suppose que y_{ijk} est la réalisation de Y_{ijk} :

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i \in \{k, \dots, n_{ij}\}, j \in \{1, \dots, J\}, i \in \{1, \dots, I\}$$

où comme précédemment

$$\varepsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

et les μ_{ij} sont les paramètres inconnus (et fixes).
Par exemple, pour l'ozone : si on considère que le premier facteur est vent et le second est pluie et que (pluie, sec)=(1,2) et (est, ouest, nord, sud)=(1,2,3,4) alors y_{125} est la 5ère journée où il a à la fois plu et où le vent venait de l'ouest.

Estimation des paramètres du modèle régulier On peut écrire la matrice correspondant au modèle, elle a IJ colonnes et est de rang IJ. On le nomme modèle régulier.
On peut montrer la proposition suivante de la même manière que pour l'anova à un facteur, introduisons une notation

$$\bar{y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$$

c'est donc la moyenne des valeurs de y sur les observations ayant pris la modalité j pour le premier facteur et la modalité k pour le second facteur.

Proposition 4.1.

$$\hat{\mu}_{ij} = \bar{y}_{ij.}$$

Exercice 4.5. *Donner l'estimateur de la variance σ^2.*

Ce modèle régulier n'est pas le modèle utilisé en pratique. Comme précédemment on introduit sa version singulière :
Pour mieux analyser l'influence des facteurs, nous allons considérer la décomposition suivante de μ_{ij} :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{4.1}$$

Remarque 4.6.

Comme dans la section précédente, cette écriture n'est pas unique. En effet, nous avons au départ $I \times J$ paramètres μ_{ij}. Nous les avons décomposés en $1 + I + J + J \times I$ nouveaux paramètres, ce qui fait $1+J+I$ paramètres de plus. Autrement dit, si on écrit les équations (4.1), on a IJ lignes mais $1 + I + J + IJ$ inconnues. Il faut imposer $1 + I + J$ contraintes pour rendre ces nouveaux paramètres identifiables. Ces contraintes doivent être linéairement indépendantes.
Les contraintes classiques sont :

- contrainte de type analyse par cellule, qui revient en fait à ne pas introduire les nouveaux paramètres

$$\mu = 0, \qquad \forall i \quad \alpha_i = 0, \qquad \forall j \quad \beta_j = 0$$

- Contrainte de type cellule de référence

$$\alpha_1 = 0, \qquad \beta_1 = 0, \quad \forall (i,j) \quad \gamma_{i1} = \gamma_{1j} = 0$$

- Contrainte de type somme

$$\sum_{i=1}^{I} \alpha_i = 0 \qquad \sum_{j=1}^{J} \beta_j = 0, \qquad \forall j \qquad \sum_{i} \gamma_{ij} = 0, \qquad \forall i \quad \sum_{j} \gamma_{ij} = 0.$$

Remarque 4.7. *Dans les contraintes de type cellule de référence ou les contraintes de type somme nous avons $2+J+K$ contraintes. Ce n'est pas un problème car ces $2+J+I$ contraintes ne sont pas linéairement indépendantes. Cela signifie qu'une des contraintes est en fait inutile car elle se déduit des autres contraintes. En effet, dans les contraintes de types cellule de référence, l'équation $\gamma_{11} = 0$ est écrite deux fois ci-dessus ($\gamma_{i1} = 0$ pour $i = 1$ et $\gamma_{1j} = 0$ pour $j = 1$). Pour les contraintes de type somme, la dernière équation est vérifiée si les $I + J - 1$ premières équations sont vérifiées. On peut le voir facilement dans le cas où les deux facteurs ont seulement deux modalités. Les équations $\forall j \quad \sum_i \gamma_{ij} = 0$ et $\forall i \ \sum_j \gamma_{ij} = 0$ s'écrivent alors*

$$\begin{aligned} \gamma_{11} + \gamma_{12} &= 0 \\ \gamma_{21} + \gamma_{22} &= 0 \\ \gamma_{11} + \gamma_{21} &= 0 \\ \gamma_{12} + \gamma_{22} &= 0 \end{aligned}$$

On voit qu'on peut déduire la quatrième équation des trois premières : en effet, avec les trois premières, on a

$$\gamma_{12} = -\gamma_{11} = \gamma_{21} = -\gamma_{22}$$

Les nouveaux coefficients s'interprètent de la façon suivante :

- Les coefficients α_i représentent **l'effet principal de la modalité i du premier facteur**
- Les coefficients β_j représentent représentent **l'effet principal de la modalité j du second facteur**
- Les coefficients γ_{ij} représentent **l'interaction entre les modalités i et j du premier et second facteur respectivement**.
- Le coefficient μ est aussi l'effet moyen, ce qu'on appelle ici l'effet moyen dépendant comme précédemment des contraintes imposées.

Remarque 4.8.

— *Si on suppose que le plan est équilibré, c'est-à-dire que le nombre d'observations par couple de modalité est le même : $n_{ij} = K$ pour tout couple (i,j) on a alors $n = IJK$. On parle aussi de plan "équirépété".*

— *On a supposé dans les n_{ij} étaient tous non nuls. Parfois, il est impossible de remplir cette condition. Par exemple, si on s'intéresse à la note d'un champagne mise par des œnologues. Tous les œnologues ne peuvent pas goûter tous les champagnes, et pourtant on voudrait être capable de comparer les champagnes entre eux en retirant l'effet "œnologue". On parle de plan d'expérience incomplet. Il faudra imposer plus de contraintes dans le modèle pour le rendre identifiable.*

Code R : Lecture des coefficients pour l'ANOVA à 2 facteurs

Avant de passer à la partie test/comparaison de modèles, intéressons-nous à la lecture des coefficients. Nous allons utiliser l'exemple lié à l'ozone. Comme nous l'avons indiqué, et comme cela se produit pour l'analyse de la variance à un facteur, R utilise l'écriture $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ et il faut donc préciser les contraintes. Les contraintes par défaut sont là encore les contraintes de type cellule de référence. C'est-à-dire que par défaut, l'effet moyen, c'est-à-dire le coefficient de l'intercept, sera l'effet du couple des premières modalités de chaque facteur. Par exemple, pour l'ozone, ce sera le couple `(Est,Pluie)`.

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

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

Dans cette formule, la contrainte mise par défaut dans R est la contrainte cellule de référence. Nous avons donc ici $\alpha_1 = \beta_1 = \gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14} = \gamma_{21} = 0$. Le reste des coefficients se lit sur la sortie : $\mu = 70.5, \alpha_2 = -1.8, \alpha_3 = 1.4, \alpha_4 = 20.9, \beta_2 = 43.8, \gamma_{22} = -18.1, \gamma_{32} = -17.3, \gamma_{42} = -29.2$. On peut retrouver les coefficients μ_{jk} du départ : par exemple, $\mu_{11} = \mu + \alpha_1 + \beta_1 + \gamma_{11} = 70.5, \mu_{12} = \mu + \alpha_1 + \beta_2 + \gamma_{12} = 70.5 + 0 + 43.8 + 0 = 114,3, \mu_{32} = \mu + \alpha_3 + \beta_2 + \gamma_{32} = 70.5 + 1.4 + 43.8 - 17.3$.

On trouve le même résultat en utilisant la formule

```
lm(maxO3~vent+temps+vent:temps,data=ozone)
```

Les tests de la sortie de `summary` ci-dessus sont difficiles à interpréter et en général ce ne sont pas les tests qui nous intéressent.

Code R : changement de contrainte

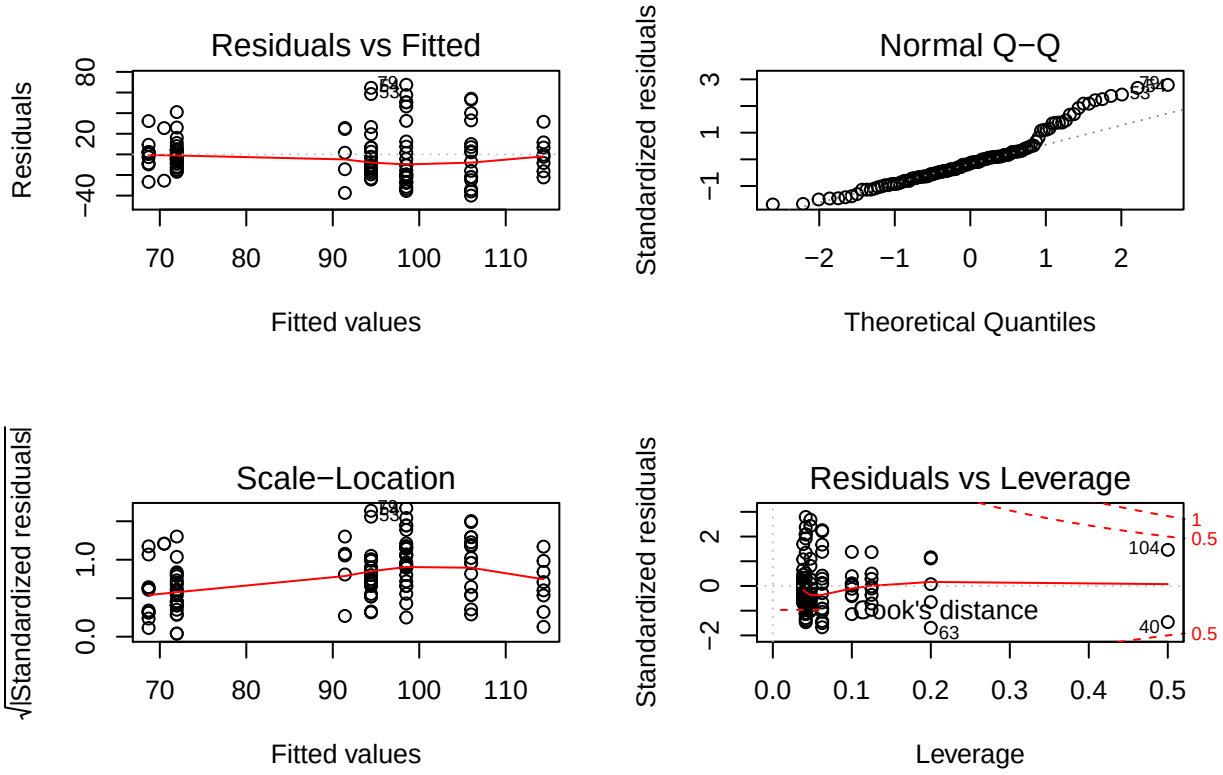
On peut faire apparaitre les résultats avec une autre contrainte que la contrainte par défaut, par exemple la contrainte somme.

```
> options(contrasts=c("contr.sum","contr.sum"))
> modcomplet=lm(maxO3~vent+temps+temps:vent,data=ozone)
```

4.2.2 Validation de modèle et test du modèle

Comme précédemment on essaie de valider le modèle. On lit pour cela les résidus.

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



D'autre part, on fait le test du modèle, i.e. on va tester $\mathcal{H}_0 : Y_{ijk} = \mu + \varepsilon_{ijk}$ versus le modèle complet. Ce test est donné directement dans le summary.

Code R : test du modèle

```
##
## Call:
## lm(formula = maxO3 ~ vent * temps, data = ozone)
...
## Multiple R-squared:  0.2807, Adjusted R-squared:  0.2322
## F-statistic: 5.797 on 7 and 104 DF,  p-value: 1.092e-05
```

La p-valeur est bien $< 5\%$ donc au moins un des facteurs a un effet. Par ailleurs, on remarquera que R^2 est très faible.

4.2.3 Tests des facteurs

On cherche à tester l'effet des facteurs sur y. Plusieurs modèles peuvent être intéressants à comparer. Ces modèles sont les suivants :

$$\begin{array}{rcll} \mathcal{M}_{\mu} & : & Y_{ijk} & = \mu + \varepsilon_{ijk} \\ \mathcal{M}_{\mu,\alpha} & : & Y_{ijk} & = \mu + \alpha_i + \varepsilon_{ijk} \\ \mathcal{M}_{\mu,\beta} & : & Y_{ijk} & = \mu + \beta_j + \varepsilon_{ijk} \\ \mathcal{M}_{\mu,\alpha,\beta} & : & Y_{ijk} & = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ \mathcal{M}_{\mu,\alpha,\beta,\gamma} & : & Y_{ijk} & = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \end{array}$$

$\mathcal{M}_{\mu,\alpha,\beta,\gamma}$ est le modèle complet. Il prend en compte tous les phénomènes mais comporte beaucoup de paramètres. Le modèle $\mathcal{M}_{\mu,\alpha,\beta}$ est appelé modèle *additif*. Il suppose l'absence d'interaction entre les facteurs.

Test des interactions Afin de simplifier le modèle, (avoir moins de paramètres donc moins d'incertitude), on va tester la nullité des interactions. On va tester

$$\mathcal{H}_0 : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \text{versus} \quad \mathcal{H}_1 : Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Dans ce cas, on est dans le cas du test d'un modèle emboîté (ou sous-modèle). On utilise le test de Fisher :

$$F = \frac{(\text{SCR}_{\mu,\alpha,\beta} - \text{SCR}_{\mu,\alpha,\beta,\gamma})/(\mathbf{rg}\left(X^{(\mu,\alpha,\beta,\gamma)}\right) - \mathbf{rg}(X^{(\mu,\alpha,\beta)}))}{SCR^{(\mu,\alpha,\beta,\gamma)}/(n - \mathbf{rg}\left(X^{(\mu,\alpha,\beta,\gamma)}\right))} \sim \mathcal{F}((I-1)(J-1), n-IJ)$$

où les SCR sont les sommes des carrés résiduels.
On rejette $\mathcal{H}_0$ si $F > q_{(I-1)(J-1),n-IJ,\alpha}$ où $q_{(I-1)(J-1),n-IJ,1-\alpha}$ est le $1-\alpha$ quantile d'une loi de Fisher à $((I-1)(J-1), n-IJ)$ degrés de liberté.

Code R : test des interactions

Nous allons illustrer ces tests avec l'exemple de l'ozone. Nous avons alors les deux facteurs `vent` et `temps`. Nous commençons par tester la pertinence de l'interaction. Cela revient à comparer le modèle complet , qu'on va appeler `mod1`,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

avec le modèle sans interaction, qu'on va appeler `mod2`

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

On utilise ensuite la fonction `anova` pour comparer les deux modèles.

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

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

L'hypothèse $\mathcal{H}_0$ est donc conservée, c'est-à-dire qu'on considère qu'il n'y a pas d'interaction. Nous nous sommes donc ramenés au modèle

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

Attention : Si ce test rejette l'hypothèse nulle, alors il n'y a pas de sens à tester les facteurs séparément (il ne peut y avoir d'interaction entre les facteurs sans les facteurs).

Test des effets Si on a pu enlever l'interaction, alors on définit le modèle sans l'interaction et on teste l'effet de chaque facteur restant.
Dans l'exemple Ozone, nous souhaitons savoir si la variable `temps` est pertinente dans le modèle mod2.
Nous souhaitons donc comparer les deux modèles :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

et

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

Code R : test des facteurs

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

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ vent
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    108 80606
```

```
## 2    107 64446  1     16159 26.829 1.052e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette donc l'hypothèse H_0 c'est-à-dire que la pluie a un effet sur le taux d'ozone, c'est-à-dire on choisit le modèle `mod2`.
On peut ensuite tester l'influence de la variable `vent` sur le taux d'ozone dans le modèle `mod2` :

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

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

On est plus mitigé sur le rôle de la provenance du vent. (une fois qu'on a introduit la variable `temps`).

Remarque 4.9. *On a étudié le cas de 2 facteurs. Si on veut faire une Anova à 3 facteurs on procèdera de la même façon. Cependant, pour des raisons d'interprétation et par soucis de parcimonie, on ne définira pas d'interactions entre trois facteurs. On introduira chacun des 3 facteurs et seulement les interactions 2 à 2 :*

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + c_k + \rho_{ij} + \nu_{ik} + \gamma_{jk} + \varepsilon_{ijkl}$$

Exercice 4.6. *Imaginons que nous ne nous intéressions pas à la variable `temps`. Nous souhaitons cependant connaitre l'influence de la variable `vent` sur `maxO3`.*

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

```
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
```

ou plus simplement encore

```
> anova(lm(maxO3~vent,data=ozone))

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
```

N'y-a-t-il pas une contradiction avec la sortie précédente ? Comment expliquer cette contradiction ?
Finalement, quel modèle faut-il retenir ?
Pour info : les données ont été collectées à Rennes.

Exercice 4.7. *1. Charger le fichier nommé "kidiq.txt". Ce fichier contient les variables* `kid_score`, `mom_hs` *et* `mom_work`*. La première est quantitative et les secondes sont qualitatives. On va expliquer le score d'un enfant au test du quotient intellectuel par le fait que la mère ait eu l'équivalent du bac (1 si oui) et par le fait qu'elle travaille plus ou moins (valeurs de 1 à 4).*
Afficher ses variables, vérifier qu'il s'agit bien de variables qualitatives.

2. *Faire les tests précédents sur l'interaction et les effets prinicipaux.*
3. *Donner la prévision du score d'un enfant si sa mère a un diplôme de second degré et travaille beaucoup (i.e. couple de modalités (1,4)).*

4.3 Analyse de la covariance

L'analyse de la covariance concerne le cas où on mélange variables explicatives qualitatives et variables explicatives quantitatives. Nous présenterons le modèle prenant en compte un facteur et une variable quantitative. Le raisonnement sur les modèles plus complexes se déduisent de ce modèle.
En particulier on code à nouveau les variables qualitatives par des variables binaires d'appartenance à une modalités, ou un groupe de modalités s'il y a plusieurs variables qualitatives, et cela ramène le problème à un problème de régression multiple.
Avec deux variables explicatives, la première étant un facteur et la seconde, qu'on note x, étant quantitative, les notations sont du même style que dans les sections précédentes : si on a I

modalités pour le facteur, et si on suppose que n_i observations ont pris la i-ème modalité, alors on rassemble ces observations qu'on note $(y_{ij})_{1\leq j\leq n_i}$ pour la variable y et $(x_{ij})_{1\leq j\leq n_i}$ pour la variable x. On a alors, possiblement, une régression sur la variable quantitative x différente selon les niveaux du facteur : y_{ij} est réalisation de

$$Y_{ij} = b_j + a_j x_{ij} + \varepsilon_{ij} \tag{4.2}$$

C'est ce qui correspond au modèle complet (le plus complexe). Par exemple, pour un facteur à 3 niveaux, cela donnerait 3 droites différentes de régression. Attention, le σ^2 est commun à toutes les données.
L'équation (4.2) correspond au modèle régulier. Sa version singulière s'écrit :

$$Y_{ij} = \mu + \beta_j + (\alpha + \gamma_j)x_{ij} + \varepsilon_{ij} \tag{4.3}$$

où γ_j est l'interaction entre le facteur et la variable quantitative.
La figure 4.3 représente en haut à gauche des données réalisation du modèle précédent pour un facteur à 2 modalités.

A. Les données correspondent au modèle complet
B. Les données correspondent au cas où les interactions sont nulles (modèle additif) :

$$Y_{ij} = \mu + \beta_j + \alpha x_{ij} + \varepsilon_{ij} \tag{4.4}$$

On a alors des droites parallèles.

C. Les données correspondent au cas où les droites ont même ordonnée à l'origine mais pas même pente : $\beta_i = 0$ pour tout i (modèle assez peu utilisé sauf cas particulier en physique). Le modèle est donc :

$$Y_{ij} = \mu + (\alpha + \gamma_j)x_{ij} + \varepsilon_{ij}$$

D. Les données correspondent au modèle où $\beta_i = \gamma_i = 0$ pour tout i, le facteur n'a pas d'effet, c'est un modèle de régression linéaire simple.

$$Y_{ij} = \mu + \alpha\, x_{ij} + \varepsilon_{ij}$$

E. Les données correspondent au modèle où $\alpha = \gamma_i = 0$ pour tout i, la covariable n'a pas d'effet, c'est un modèle d'anova à un facteur :

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

F. Les données correspondent au modèle nul $\beta_i = \alpha = \gamma_i = 0$ pour tout i :

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

Inférence Toute l'inférence se passe exactement comme précédemment.

- On peut montrer que les EMC des paramètres (a_j, b_j) sont obtenus en faisant la régression de y contre x pour les données dans la modalité j.
- Les paramètres du modèle singulier sont non-identifiables. Il faut imposer des contraintes. **R** impose comme attendu $\beta_1 = \gamma_i = 0$. On obtient l'estimation des paramètres du modèle singulier à partir de celle des (a_j, b_j).

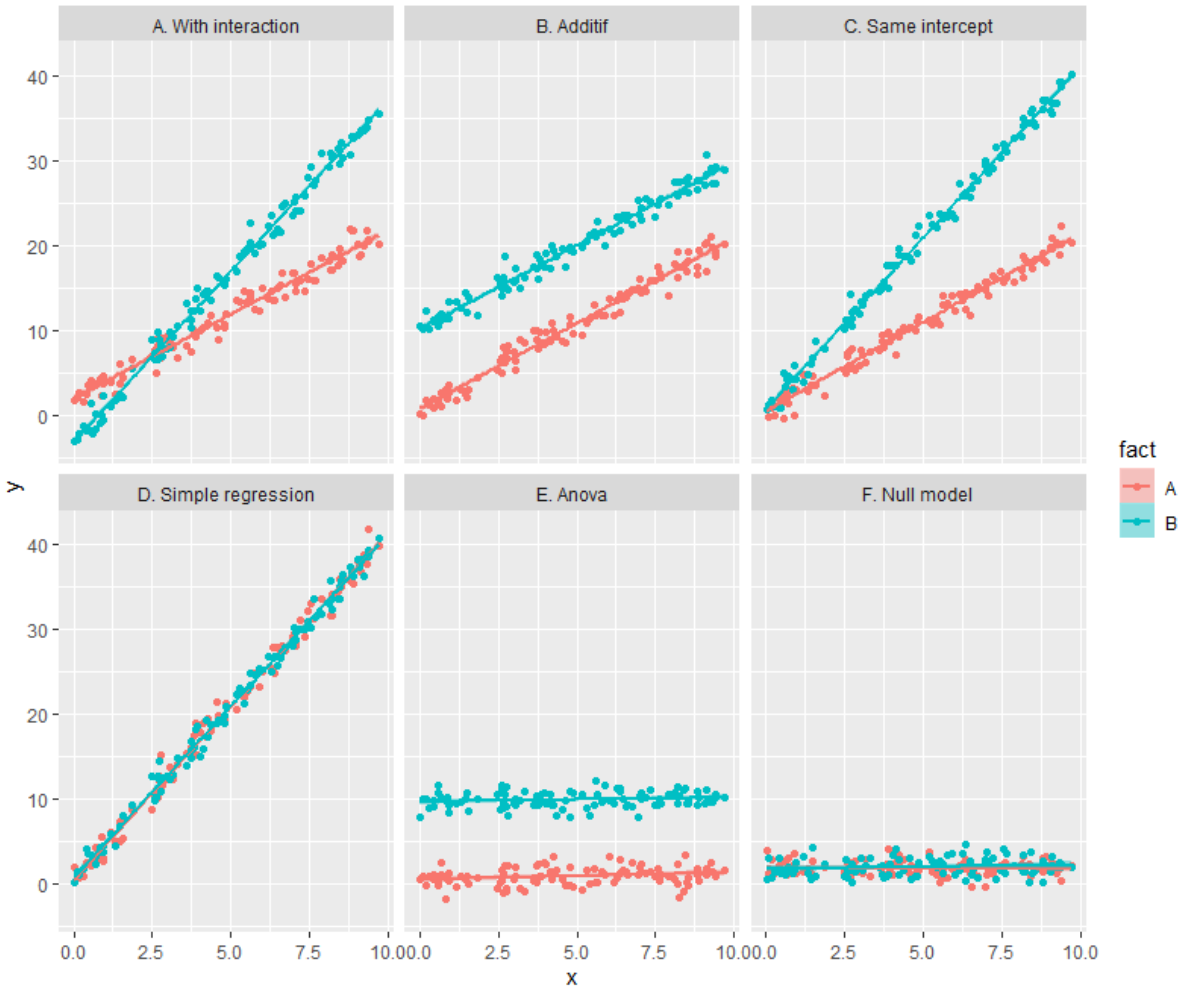


FIGURE 4.3 – Ancova : modèle avec interaction en haut à gauche, modèle additif en haut à droite, modèle avec $b_j = 0$, pour tout j en bas à gauche, régression simple en bas à droite

- Notons $\hat{y}_{ij} = \hat{b}_i + \hat{a}_i x_{ij}$ la prédiction, alors σ^2 est estimé sans biais par

$$\hat{\sigma}^2 = \frac{1}{n-2I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$$

.

- Les tests du facteur et de la covariable seront faits en comparant les SCR des différents modèles en compétition. On prendra en compte le nombre de paramètres dans chaque modèle. On peut utiliser la commande `anova`

Code R : modèle d'Ancova et test de l'interaction

```
mod1 = lm(ht~circ*bloc,data=euca) #le  modèle complet
summary(mod1)
```

```
##
## Call:
## lm(formula = ht ~ circ * bloc, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7723 -0.7037  0.0539  0.8114  3.3255
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.031e+00  2.895e-01  31.199   <2e-16 ***
## circ         2.576e-01  6.043e-03  42.618   <2e-16 ***
## blocA2      -1.850e-01  4.044e-01  -0.457    0.647
## blocA3       6.165e-01  4.766e-01   1.294    0.196
## circ:blocA2 -1.927e-05  8.454e-03  -0.002    0.998
## circ:blocA3 -6.834e-03  9.802e-03  -0.697    0.486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.187 on 1423 degrees of freedom
## Multiple R-squared:  0.7736, Adjusted R-squared:  0.7728
## F-statistic: 972.6 on 5 and 1423 DF,  p-value: < 2.2e-16
```

```
mod2 = lm(ht~bloc+circ,data=euca)# modèle sans interaction
anova(mod2,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: ht ~ bloc + circ
## Model 2: ht ~ circ * bloc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
```

```
## 1   1425 2005.9
## 2   1423 2005.0  2   0.84752 0.3007 0.7403
```

Exercice 4.8. *On utilise le dataframe* `euca`.

1. *A-t-on intérêt à garder le facteur bloc ?*
2. *Vérifier l'hypothèse d'homoscédasticité.*
3. *On se place désormais dans ce modèle. Donner les estimations des paramètres.*
4. *Comment prédit-on une nouvelle valeur ?*

4.4 Exercices récapitulatifs

4.4.1 Exercice sur l'Anova à 1 facteur

Exercice 4.9 (Anova a 1 facteur). *On s'intéresse aux résultats d'une expérience cherchant à expliquer un rendement agricole (mesuré en poids sec de plantes) en fonction du traitement (2 traitements différents et un groupe de contrôle). On dispose de* 30 *observations réparties de façon équilibrées dans les trois modalités. Les données sont représentées sous la forme de boxplot sur la figure suivante :*

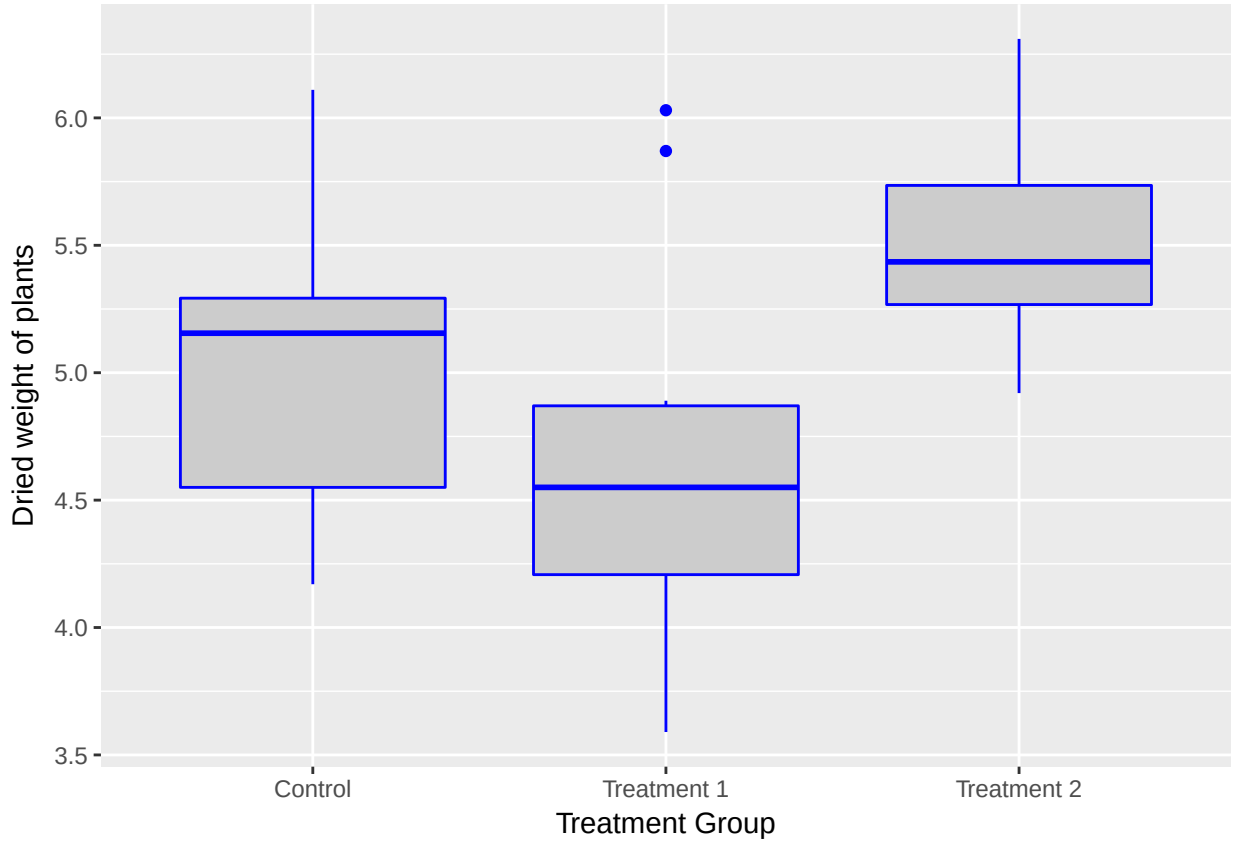


FIGURE 4.4 – Exercice rendement agricole : boxplot

Les instructions sont données ci-dessous.

1. *Quelle contrainte est utilisée dans* `Instruction 1` *?*
2. *Les hypothèses du modèle linéaire sont-elles vérifiées ?*
3. *Que fait-on dans* `Instruction 3` *?*
4. *Rappelez l'expression de la statistique du test du modèle. Rejette-t-on* $\mathcal{H}_0$ *ici ?*
5. *Que pensez-vous du* R^2.
6. *Interpréter les sorties* `Coefficients`.
7. *Que fait-on dans l'*`instruction 5`*. Rappelez la statistique de test. A quoi correspond l'instruction* `Bonferroni` *? Interprétez les résultats.*

- `Instruction 1`

```
plant.mod1 = lm(weight ~ group, data = plant.df)
summary(plant.mod1)
```

```
##
## Call:
## lm(formula = weight ~ group, data = plant.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.0320     0.1971  25.527   <2e-16 ***
## groupTreatment 1  -0.3710     0.2788  -1.331   0.1944
## groupTreatment 2   0.4940     0.2788   1.772   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

- *Instruction 2*

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

- *Instruction 3*

```
contrasts(plant.df$group) <- contr.sum
plant.mod2 = lm(weight ~ group, data = plant.df)
summary(plant.mod2)
```

```
##
## Call:
## lm(formula = weight ~ group, data = plant.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0730     0.1138  44.573   <2e-16 ***
## group1       -0.0410     0.1610  -0.255   0.8009
## group2       -0.4120     0.1610  -2.560   0.0164 *
## ---
```

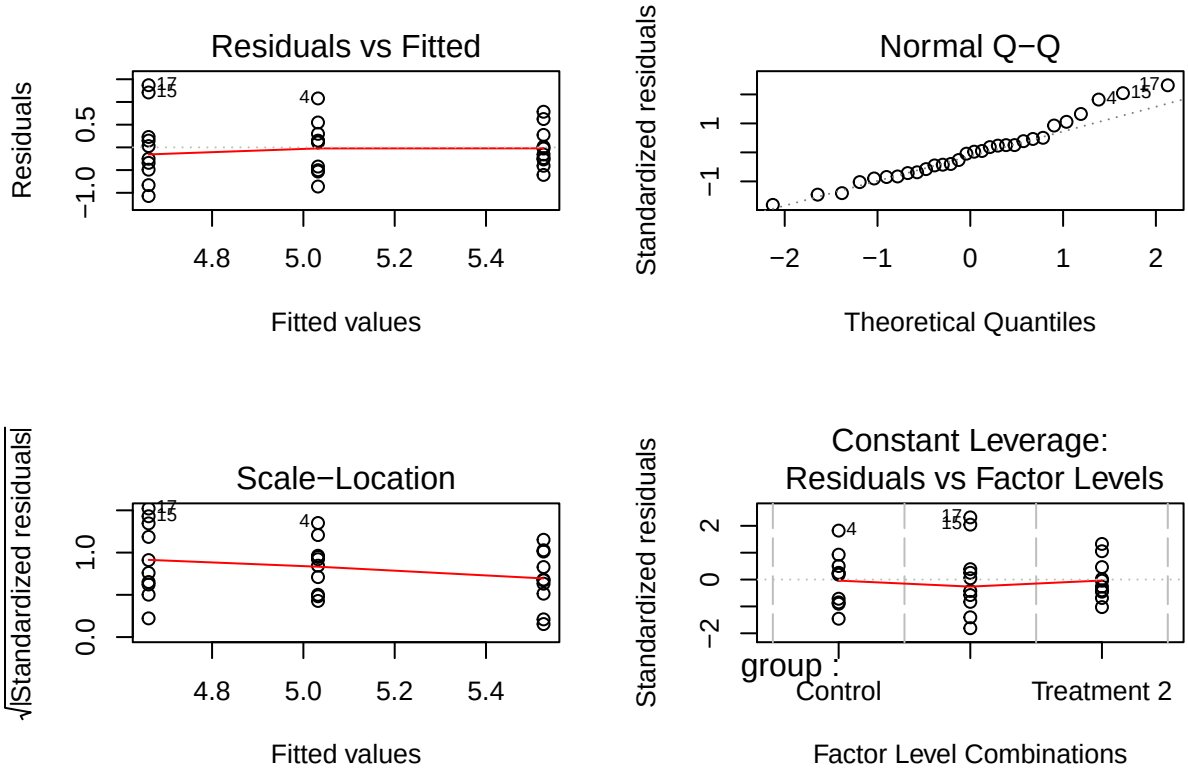


FIGURE 4.5 – Exercice rendement agricole : résidus

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

- *Instruction 4*

```
anova(plant.mod1)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## group      2  3.7663  1.8832  4.8461 0.01591 *
## Residuals 27 10.4921  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Instruction 5*

```
test.treatment = pairwise.t.test(plant.df$weight,plant.df$group,

                p.adjust.method="bonferroni")
test.treatment
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  plant.df$weight and plant.df$group
##
##             Control Treatment 1
## Treatment 1 0.583   -
## Treatment 2 0.263   0.013
##
## P value adjustment method: bonferroni
```

4.4.2 Exercice sur l'Anova à 2 facteurs

Exercice 4.10 (Anova à deux facteurs). *Nous considérons le jeu de données* `ToothGrowth`. *La réponse Y est la longueur des odontoblates (cellules intervenant dans la croissance des dents) chez $n = 60$ cochons de Guinée. Chaque animal a reçu une des trois doses possibles de vitamince C (0.5, 1, and 2 mg/day) par le biais d'une des deux méthodes d'administration (jus d'orange ou acide ascorbique). On s'intéresse à l'influence de ces facteurs sur la croissance dentaire. Le jeu de données est représenté par le box-plot suivant. (Pour l'exercice, nous avons transformé la variable quantitative* `dose` *en un facteur* `doselevel`*).*

```
ToothGrowth$doselevel = as.factor(ToothGrowth$dose)
names(ToothGrowth)=c('len','suppfactor','dose','doselevel')
summary(ToothGrowth)
```

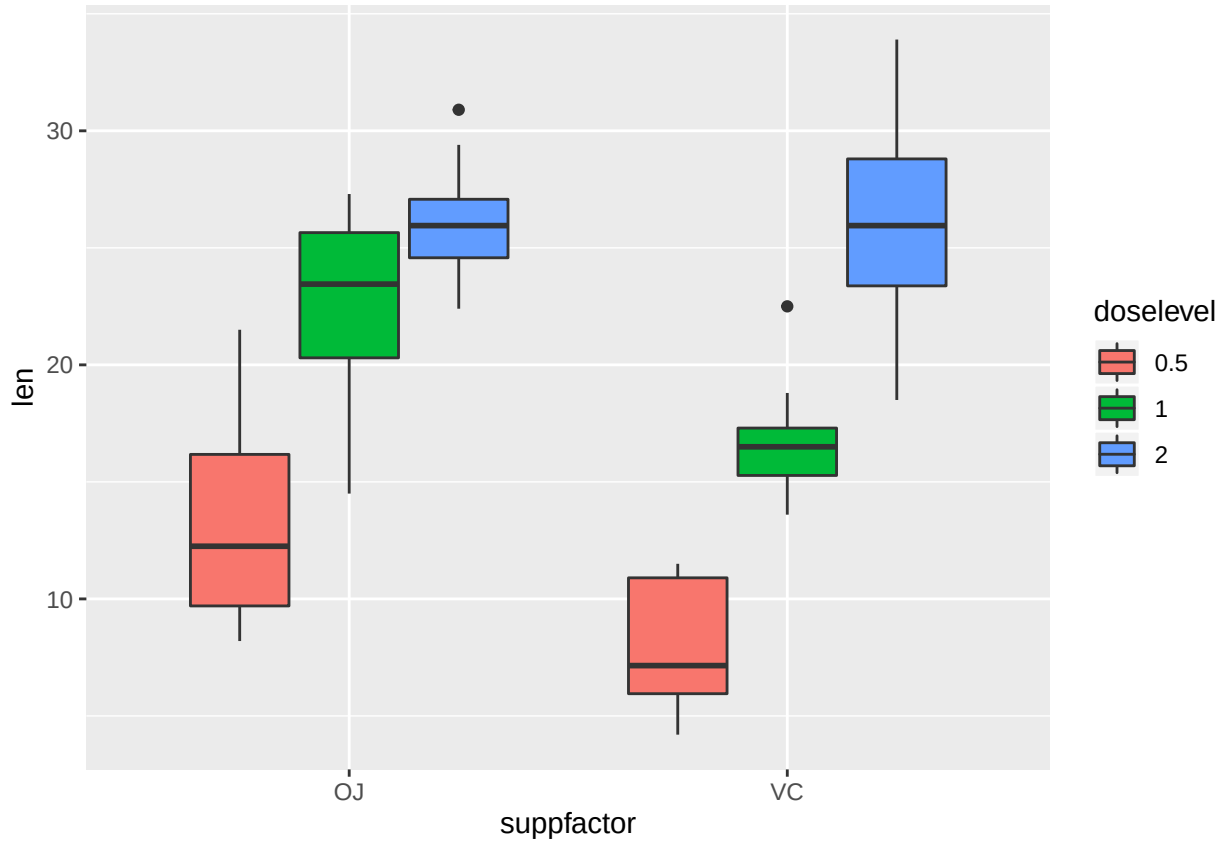
Statistiques descriptives

```
##       len        suppfactor      dose       doselevel
##  Min.   : 4.20   OJ:30      Min.   :0.500   0.5:20
##  1st Qu.:13.07   VC:30      1st Qu.:0.500   1  :20
##  Median :19.25              Median :1.000   2  :20
##  Mean   :18.81              Mean   :1.167
##  3rd Qu.:25.27              3rd Qu.:2.000
##  Max.   :33.90              Max.   :2.000
```

```
table(ToothGrowth$doselevel, ToothGrowth$suppfactor)
```

```
##
##       OJ VC
##   0.5 10 10
##   1   10 10
##   2   10 10
```

```
ggplot(ToothGrowth, aes(y = len,x=suppfactor,fill = doselevel))+ geom_boxplot()
```



Modélisation et inférence

1. *Ecrire le modèle correspondant aux instruction suivantes (sans oublier les hypothèses et les gammes de variation des indices).*

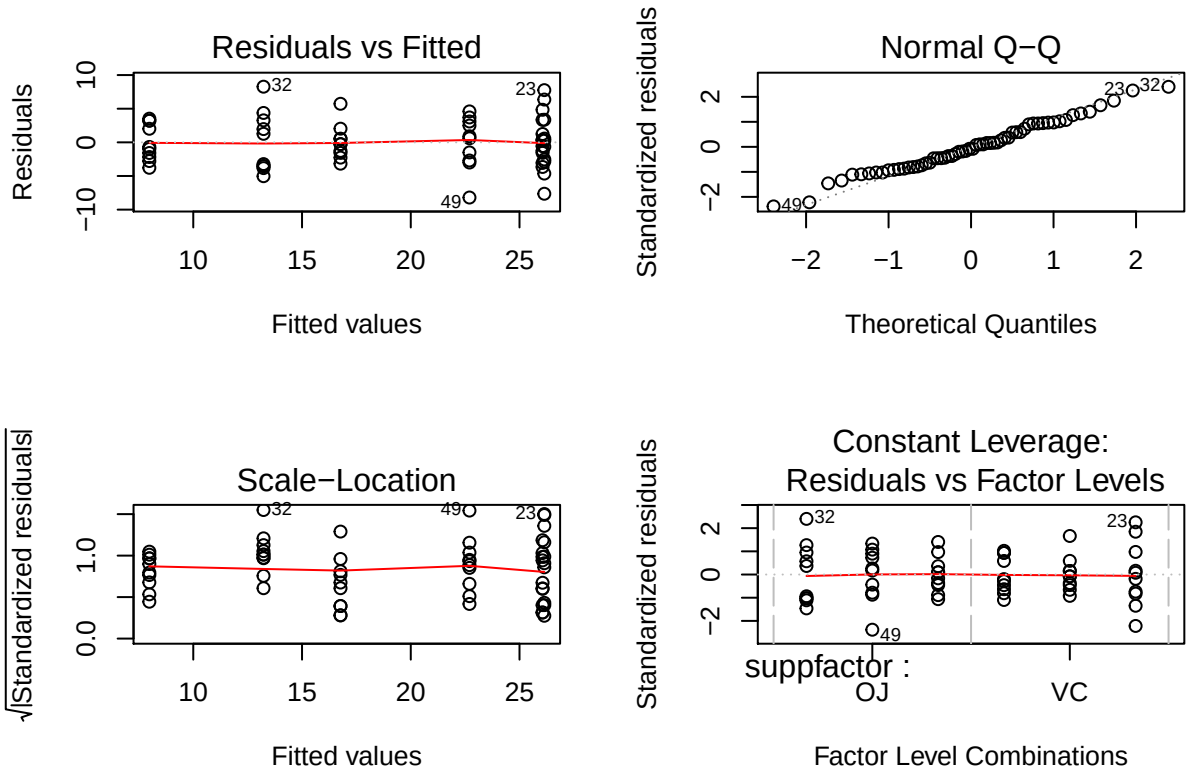
```
mod_compl = lm(len~suppfactor*doselevel,data=ToothGrowth)
summary(mod_compl)
```

```
##
## Call:
## lm(formula = len ~ suppfactor * doselevel, data = ToothGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.230      1.148  11.521 3.60e-16 ***
## suppfactorVC                -5.250      1.624  -3.233  0.00209 **
## doselevel1                   9.470      1.624   5.831 3.18e-07 ***
## doselevel2                  12.830      1.624   7.900 1.43e-10 ***
## suppfactorVC:doselevel1     -0.680      2.297  -0.296  0.76831
## suppfactorVC:doselevel2      5.330      2.297   2.321  0.02411 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

2. *Quelles sont les contraintes utilisées par le logiciel R ? (retrouver ces contraintes dans les sorties précédentes)*
3. *Les hypothèses du modèle linéaire sont elles vérifiées ?*

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



4. *Doit-on garder les interactions ?.*

```
mod_add = lm (len~suppfactor + doselevel,data=ToothGrowth)
anova(mod_add, mod_compl)
```

```
## Analysis of Variance Table
##
## Model 1: len ~ suppfactor + doselevel
## Model 2: len ~ suppfactor * doselevel
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     56 820.43
```

```
## 2     54 712.11  2    108.32 4.107 0.02186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. *A-t-on un effet significatif des facteurs sur la croissance ?*

```
anova(mod_compl)
```

```
## Analysis of Variance Table
##
## Response: len
##                      Df  Sum Sq Mean Sq F value    Pr(>F)
## suppfactor            1  205.35  205.35  15.572 0.0002312 ***
## doselevel             2 2426.43 1213.22  92.000 < 2.2e-16 ***
## suppfactor:doselevel  2  108.32   54.16   4.107 0.0218603 *
## Residuals            54  712.11   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4.3 Exercice sur l'Ancova

Exercice 4.11 (Ancova). *Nous nous intéressons à la consommation (exprimée en en miles par galon) de voitures en fonction de leur puissance et du type de transmission (automatique ou manuelle).*

```
data(mtcars)
attach(mtcars)
don <- mtcars[,c("am","mpg","hp")]
don$am <- as.factor(don$am)
```

Un extrait des données est fourni ci-dessous.

```
head(don)
```

```
##                   am  mpg  hp
## Mazda RX4          1 21.0 110
## Mazda RX4 Wag      1 21.0 110
## Datsun 710         1 22.8  93
## Hornet 4 Drive     0 21.4 110
## Hornet Sportabout  0 18.7 175
## Valiant            0 18.1 105
```

Répondre aux questions en utilisant les instructions R données en fin d'énoncé.

1. *Écrire le modèle* `mod2` *correspondant (sans oublier les hypothèses et les gammes de variation des indices).*
2. *Quelles sont les contraintes utilisées par le logiciel R ?*
3. *Sur la ligne* `am1:hp`*, interpréter la valeur* 0.0004029 *dans la sortie* `summary` *dans l'instruction 1. .*
4. *Justifier le passage au modèle* `mod2b`*.*
5. *Le type de transmission a-t-il une influence sur la consommation (à justifier soigneusement) ? Détailler les hypothèses comparées par le test que vous utilisez.*
6. *Donner une estimation de la consommation moyenne (miles / galon) pour un véhicule manuel de puissance 150 ? Même question pour un véhicule automatique ?*

- *Instruction 1*

```
mod2 = lm(mpg~am*hp,data=don)
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ am * hp, data = don)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3818 -2.2696  0.1344  1.7058  5.8752
```

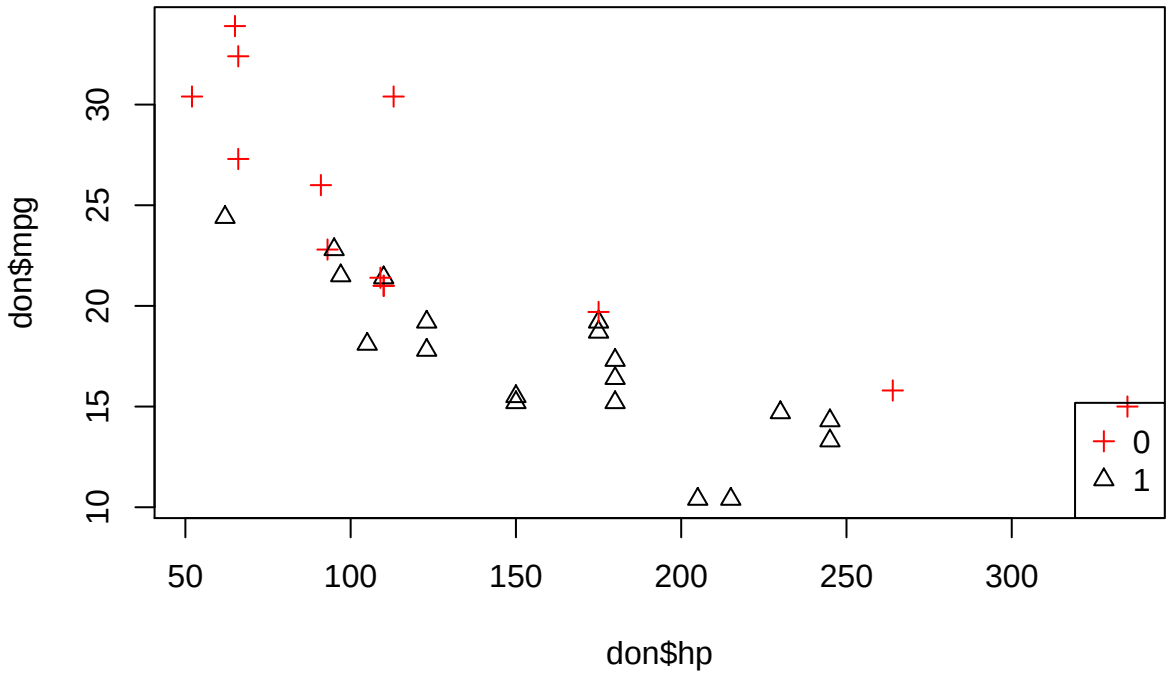


FIGURE 4.6 – Consommation (en miles par galon) en fonction de la puissance des véhicules

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.6248479  2.1829432  12.197 1.01e-12 ***
## am1          5.2176534  2.6650931   1.958   0.0603 .
## hp          -0.0591370  0.0129449  -4.568 9.02e-05 ***
## am1:hp       0.0004029  0.0164602   0.024   0.9806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 28 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7587
## F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.112e-09
```

- *Instruction 2*

```
mod0 = lm(mpg~1,data=don)
anova(mod0,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ am * hp
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     31 1126.05
## 2     28  245.43  3    880.61 33.488 2.112e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Instruction 3*

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15 46.2210 2.197e-07 ***
## hp         1 475.46  475.46 54.2419 5.088e-08 ***
## am:hp      1   0.01    0.01  0.0006    0.9806
## Residuals 28 245.43    8.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Instruction 4*

```
mod2b=lm(mpg~am + hp,data = don)
summary(mod2b)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp, data = don)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
## am1          5.277085   1.079541   4.888 3.46e-05 ***
## hp          -0.058888   0.007857  -7.495 2.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

- *Instruction 5*

```
anova(mod2b)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15  47.871 1.327e-07 ***
## hp         1 475.46  475.46  56.178 2.920e-08 ***
## Residuals 29 245.44    8.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Annexe A

Variance, covariance et corrélation empirique

On introduit ici des quantités qui seront nécessaires dès le chapitre 1.

A.1 Moyenne, variance, covariance empirique

Soit x et y deux vecteurs de $\mathbb{R}^n$:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

On peut voir le vecteur x comme une variable aléatoire réelle $\tilde{x}$, qui prend des valeurs discrètes $(x_1, \ldots, x_n)$ avec une probabilité $\frac{1}{n}$:

$$\mathbf{P}(\tilde{x} = x_i) = \frac{1}{n}.$$

Alors l'espérance de la variable aléatoire réelle $\tilde{x}$ associée au vecteur x est donnée par

$$\mathbb{E}(\tilde{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

De la même manière on peut calculer la variance

$$\mathbf{Var}(\tilde{x}) = \mathbb{E}\Big[(\tilde{x} - \mathbb{E}(\tilde{x}))^2\Big] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathbb{E}(\tilde{x}))^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

On peut également associer un couple de variables aléatoires réelles $(\tilde{x}, \tilde{y})$ au couple de vecteurs $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, en posant

$$\mathbf{P}\Big((\tilde{x}, \tilde{y}) = (x_i, y_i)\Big) = \frac{1}{n}$$

Alors la covariance des variables aléatoires réelles $\tilde{x}$ et $\tilde{y}$ est donnée par

$$\mathbf{Cov}(\tilde{x}, \tilde{y}) = \mathbb{E}\Big[(\tilde{x} - \mathbb{E}(\tilde{x}))(\tilde{y} - \mathbb{E}(\tilde{y})\Big] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Toutes ces quantités associés aux vecteurs x et y sont appelées moyennes, variances et covariance *empiriques*.

Définition A.1. *On appelle moyenne empirique et on note $\bar{x}$ la quantité*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

On appelle variance empirique de x et on note σ_x^2 la quantité

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

On appelle covariance empirique et on note $\sigma_{x,y}$ la quantité

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

On dit qu'un vecteur y est centré si $\bar{y} = 0$.
Pour centrer un vecteur y, il suffit de retrancher $\bar{y}$ à toutes ses composantes. Cela donne le vecteur $z = y - \bar{y}\mathbf{1}$ où $\mathbf{1}$ est le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1. En effet

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - n\bar{y} = 0$$

A.2 Corrélation empirique

Définition A.2. *On appelle coefficient de corrélation empirique et on note $\rho_{x,y}$ la quantité*

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

Propriétés :
Si on suppose que les vecteurs x et y ont été centrés, i.e. $\bar{x} = 0$ et $\bar{y} = 0$ alors la corrélation empirique s'écrit

$$\rho_{x,y} = \frac{\langle x, y\rangle}{\|x\|\|y\|}$$

On interprète le coefficient de corrélation empirique entre deux vecteurs **centrés** x et y de la manière suivante :

- Si $\rho_{x,y} = 0$ alors les vecteurs sont orthogonaux.
- Si $\rho_{x,y}^2 = 1$ alors les vecteurs sont colinéaires.
- Si $\rho_{x,y}^2$ est très proche de 1 alors les vecteurs x et y sont "presque" colinéaires.

$\rho_{x,y}$ est en fait le cosinus de l'angle entre x et y : si $\rho_{x,y}^2$ est proche de 1, l'angle est proche de 0 et donc les vecteurs x et y sont "presque" colinéaires.

Pour des vecteurs **non centrés** maintenant : si $\rho_{x,y}^2$ est proche de 1, alors les vecteurs centrés $x - \bar{x}\mathbf{1}$ et $y - \bar{y}\mathbf{1}$ sont presque colinéaires : Autrement dit il existe un réel a tel que

$$y - \bar{y}\mathbf{1} \approx a(x - \bar{x}\mathbf{1})$$

cela s'écrit aussi

$$y \approx ax + b\mathbf{1}$$

où $b = (\bar{y} - a\bar{x})$. Autrement dit, $\rho_{x,y}^2$ est proche de 1 si y est combinaison linéaire des vecteurs x et $\mathbf{1}$.

$$\rho^2_{x,y} \approx 1 \Leftrightarrow \text{ il existe } a \text{ et } b \text{ t.q. } y \approx ax + b\mathbf{1}$$

NB : Dans la suite de ce cours, on parlera de "variables" pour des vecteurs (constants) x à n coordonnées. Cela vient de cette identification avec des variables aléatoires réelles et discrètes $\tilde{x}$ dont on ne parlera plus. On omettra aussi le plus souvent le terme "empirique". Par exemple, on dira que la "variable" x est centrée si $\bar{x} = 0$. Si y est un autre vecteur, on parlera de la "corrélation" entre les deux "variables" x et y (cela correspondra donc à la corrélation empirique).

Annexe B

Rappels d'algèbre linéaire

Ces rappels sont nécessaires aux quelques démonstrations mathématiques du chapitre suivant (il y en a peu et elles sont **non exigées**) . Les propriétés qui suivent ne sont donc pas exigées.

B.1 Propriétés basiques

Soit une matrice X réelle de dimension $n \times p$.
On note $x^1, \ldots, x^p$ les vecteurs colonnes de X.

1. Si $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p$ alors

$$X\beta = \sum_{j=1}^{p} \beta_j x^j$$

   c'est-à-dire : le vecteur $X\beta$ est une combinaison linéaire des vecteurs colonnes de la matrice X, les coefficients de cette combinaison linéaire étant les composantes du vecteur β.

2. l'image de X, notée $\mathbf{Im}(X)$, est définie par

$$\mathbf{Im}(X) = \{X\beta, \beta \in \mathbb{R}^p\}$$

   Autrement dit c'est le sous espace de $\mathbb{R}^n$ engendré par les vecteurs colonnes de X :

$$\mathbf{Im}(X) = \{\sum_{j=1}^{p} \beta_j x^j, (\beta_j)_{1 \leq j \leq p} \in \mathbb{R}^p\}$$

3. Le noyau de X est $\mathbf{Ker}(X) = \{\beta \in \mathbb{R}^p : X\beta = 0\}$.
4. Un système de vecteur $x^1, \ldots, x^p$ est linéairement indépendant si $\beta_1 x^1 + \ldots + \beta_p x^p = 0$ implique $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. Cela revient à dire que $X\beta = 0$ implique $\beta = 0$, ce qui signifie que le noyau est réduit à 0.
5. On dit que la matrice X est de plein rang en colonnes si les colonnes de X sont indépendantes. Cela revient donc à dire que le noyau est réduit à $\{0\}$.
6. Pour qu'une matrice $n \times p$ soit de plein rang en colonnes, il faut que $n \geq p$.
7. Si $X \in M(n,p)$ est de plein rang en colonnes alors $X^T X$ est inversible.

8. Réciproquement, si X n'est pas de plein rang en colonnes, i.e. si les colonnes de X ne sont pas linéairement indépendantes, alors la matrice carrée X^TX n'est pas inversible.
9. Le produit scalaire usuel $\langle x, y\rangle$ s'écrit matriciellement
$$\langle x, y\rangle = x^Ty = y^Tx$$
10. Système d'équations :
Si X est inversible alors, pour tout $b \in \mathbb{R}^n$, l'équation $X\beta = b$ a une unique solution donnée par $\beta = X^{-1}b$.
Si X n'est pas inversible alors l'équation $Xx = b$ n'a de solution que si $b \in \mathbf{Im}(X)$, i.e. s'il existe x_0 tel que $b = Xx_0$ et alors l'ensemble des solutions est donné par $x_0 + \mathbf{Ker}(X)$.

B.2 Produits

— Soit A, B deux matrices et $\beta = (\beta_1, \ldots, \beta_p)^T$ un vecteur de $\mathbb{R}^p$.
Deux formules de base :
Si $A = (a_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ est de dimension $n \times p$ et $B = (b_{ij})_{1\leq i\leq p, 1\leq j\leq q}$ de dimension $p \times q$ alors le produit AB existe et le terme principal de la matrice $C = AB$ est donné par
$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$$
Le vecteur $y = A\beta \in \mathbb{R}^n$ est donné par
$$y_i = \sum_{j=1}^{p} A_{ij}\beta_j$$
où on a noté y_i (respectivement β_i) la ième coordonnée de y (respectivement de β).

B.3 Projection orthogonale sur un sous-espace

Soit F un sous espace de $\mathbb{R}^n$. On peut écrire
$$\mathbb{R}^n = F \oplus F^\perp$$
où $F^\perp$ est l'espace orthogonal à F (i.e. l'ensemble des vecteurs orthogonaux à tous les vecteurs de F). Cela signifie que tout vecteur x de $\mathbb{R}^n$ se décompose en $x = x_1 + x_2$ avec $x_1 \in F$ et $x_2 \in F^\perp$, et la décomposition est unique.
On dit que P est la matrice de projection orthogonale sur l'espace F si :
$$Px = x_1$$
Autrement dit, la décomposition ci-dessus est donnée par : $x_1 = Px$ et $x_2 = x - Px$.
En particulier, si on projette sur F, on a $Px \in F$ et $x - Px \in F^\perp$. Et donc Px et $x - Px$ sont sont orthogonaux.

— Soit $x \in \mathbb{R}^n$ et F un sous espace, le problème suivant
$$\min_{y\in F} \|y - x\|^2$$
a une unique solution donnée par $y = Px$ où P est la projection orthogonale sur F.

— Si F est engendré par les colonnes d'une matrice X, i.e. si $F = \text{Vect}(x_1, \ldots, x_p\}$, et si les x_i sont linéairement indépendants (autrement dit si la matrice X est de plein rang en colonnes) alors la matrice de projection orthogonale sur F est donnée par

$$P = X(X^T X)^{-1} X^T$$

— (Pythagore) Si P est une matrice de projection orthogonale alors

$$\|x\|^2 = \|Px\|^2 + \|x - Px\|^2$$

— Si F et G sont des sous espaces de $\mathbb{R}^n$ tels que $F \subset G$, alors on a, pour tout vecteur $y \in \mathbb{R}^n$,

$$\|P_G(y) - y\| \leq \|P_F(y) - y\|$$

— si $e_1, \ldots, e_p$ est une base orthonormale de F alors la projection orthogonale sur F s'écrit

$$p(x) = \sum_{i=1}^{p} \langle x, e_i \rangle e_i$$

B.4 Un rappel de probabilité : les vecteurs aléatoires

Soit $Z = (Z_1, \ldots, Z_p)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^p$. Cela signifie que Z a p composantes Z_j qui sont toutes des variables aléatoires réelles. On peut donc considérer leurs variances et leurs covariances. Toutes ces informations sont contenues dans la matrice de variance-covariance de Z, qu'on note ici Σ. On a donc

$$\mathbf{Var}(Z_j) = \Sigma_{jj} \text{ et } \mathbf{Cov}(Z_j, Z_k) = \Sigma_{jk}$$

On rappelle aussi deux formules utiles pour le calcul de l'espérance et de la matrice de covariance d'une transformée linéaire d'un vecteur aléatoire Z.
Soit A une matrice déterministe (i.e. constante, i.e. non aléatoire). Alors l'espérance du vecteur aléatoire AZ est donnée par

$$\mathbb{E}(AZ) = A\mathbb{E}(Z)$$

et la matrice de variance-covariance de AZ est donnée par

$$\Sigma_{AZ} = A\Sigma_Z A^T$$

Si u est un vecteur déterministe, alors $u^T Z = \sum_{i=1}^{p} u_j Z_j$ est une variable aléatoire réelle de variance

$$\mathbf{Var}(u^T Z) = u^T \Sigma_Z u$$

Enfin, un rappel sur les vecteurs gaussiens : un vecteur gaussien $Z = (Z_1, \ldots, Z_n)$ a ses composantes indépendantes si et seulement si la matrice de variance-covariance de Z est diagonale, c'est-à-dire si et seulement si ses composantes sont non corrélées.

Annexe C

Validation de modèle

Nous reprenons ici plus en détail le principe de validation de modèle.
Comme nous l'avons vu dans le cours, tous les résultats (estimation,tests, intervalles de confiance) reposent sur des hypothèses fondamentales liés au terme d'erreur ε qui résume les informations absentes du modèle . Il importe donc que l'on vérifie ces hypothèses afin de pouvoir interpréter les résultats. Rappelons brièvement les hypothèses liées au terme d'erreur :

- Les erreurs sont centrées $\mathbb{E}(\varepsilon) = 0$.
- La distribution est gaussienne
- La variance est constante (σ^2)
- Les erreurs $(\varepsilon_i)_{1 \leq i \leq n}$ sont indépendantes.

Pour inspecter ces hypothèses, nous disposons des erreurs $\hat{\varepsilon}_i$ qui sont des sortes d'approximations des vraies erreurs. Si le modèle est correct, c'est-à-dire si les hypothèses précédentes sur le bruit sont vérifiées, mais aussi si on n'a pas oublié de variables explicatives, si le lien est bien linéaire (etc) alors on attend un certain comportement de la part de ces résidus.
Evidemment, ces résidus étant aléatoires, il faudra faire la différence entre une déviation du comportement attendu qui serait liée au hasard, et une déviation du comportement attendu qui serait liée à une mauvaise modélisation.
En particulier, on pourra utiliser essentiellement des analyses graphiques dans un premier temps pour repérer d'éventuels problèmes. Les graphiques des résidus en ordonnées avec des quantités diverses en abscisses (temps, indice etc) permettent de repérer une inadéquation du modèle. Dans ce type de graphique, on s'inquiètera si on voit apparaitre une forme particulière (vague, tendance, etc).

- **L'hypothèse d'indépendance** sera en général considérée comme vérifiée lorsque chaque donnée correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes (se rappeler que l'indépendance des y_i équivaut à l'indépendance des ε_i). En revanche, **dans des problèmes où le temps joue un rôle important, elle est plus difficilement vérifiée** (une évolution se fait rarement de façon totalement indépendante du passé). La figure C.1 est un exemple de graphique qu'on pourrait obtenir dans le cas non indépendant. Dans cet exemple, avec en ordonnée les résidus et en abscisse le temps, on voit une certaine rémanence ou inertie du phénomène étudié. Cela s'observe par le fait que les données n'oscillent pas en permanence autour de la droite de régression, mais semblent s'attarder une fois qu'elles sont d'un côté.
- **On peut souvent se passer de l'hypothèse de normalité général quand le nombre de données n est important** (en pratique $n > 30$ environ, en fait ça dé-

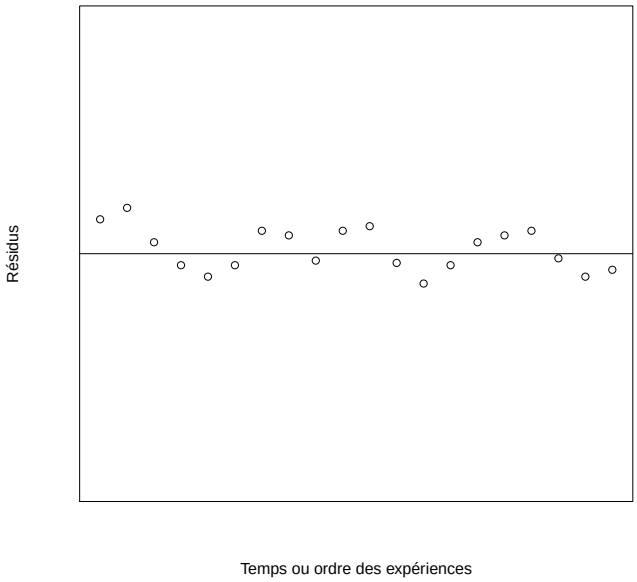


FIGURE C.1 – Les résidus en fonction du temps ou de l'ordre des observations

pend...), sans que cela change réellement les résultats principaux sur l'inférence statistique (IC, tests).

— L'hypothèse d'homoscédasticité, c'est-à-dire l'hypothèse selon laquelle les erreurs sont de variance constante ($\mathbf{Var}(\varepsilon_i) = \sigma^2$ pour tout i) n'est pas toujours vérifiée dans la pratique.

Remarque C.1. *Une étape importante du processus est aussi de choisir le modèle, c'est-à-dire choisir les variables qu'on va utiliser pour la régression. On a uniquement parlé du R^2 ajusté (ou des tests pour modèles emboités) dans ce cours. Il existe beaucoup d'autres méthodes qui seront vus dans les cours suivants.*

Voici un plan de vérification qu'il serait bon de suivre après chaque régression (les hypothèses à vérifier sont classées par ordre d'importance décroissante). A partir des résidus $\hat{\varepsilon}_i$,

1. On vérifie que le lien est bien linéaire.
2. Vérification de l'homoscédasticité
3. Vérification éventuelle de l'indépendance
4. Vérification éventuelle de la normalité

Il est d'abord recommandé de **tracer les résidus en fonction des valeurs ajustées $\hat{y}_i$ c'est-à-dire tracer les couples de points** $(\hat{\varepsilon}_i, \hat{y}_i)$. Concrètement, si on ne voit rien de notable sur le graphique (nuage de points quelconque), c'est très bon signe : les résidus ne semblent avoir aucune propriété particulière et c'est bien ce qu'on demande à l'erreur. La figure C.4 est un exemple de graphe résidus/valeurs ajustées "pathologique". On rappelle qu'on peut obtenir directement ce graphe avec la commande `plot(reg,which=1)` où `reg` est la sortie de `lm`. Ce graphique est assez généraliste et permet de détecter un certain nombre de problèmes (non linéarité, hétéroscédasticité etc). Pour voir ce à quoi ressemble un graphique "normal" des résidus contre les valeurs ajustées, un exemple est donné figure C.5.

Une analyse grossière des résidus peut se faire à l'aide du coefficient de détermination R^2. On rappelle que

$$R^2 = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|y - \bar{y}\mathbf{1}\|^2}$$

et donc que $R^2 \approx 1$ si $\hat{\varepsilon} \approx 0$.
Les raisons pour lesquelles un R^2 est faible sont multiples. Quand on oublie une variable explicative importante, ce coefficient a des chances d'être grand. Cela peut venir aussi du fait que la relation entre y et l'une (ou plusieurs) des variables explicatives n'est pas linéaire.

C.1 Vérification de la linéarité/transformation des données

Dans le cas de la régression simple où il y a une seule variable explicative x, un simple graphique de y contre x donne déjà une bonne indication. Nous pouvons aussi afficher le graphique des résidus contre la variable x. On s'inquiétera si une forme particulière apparait.
L'affaire se corse quand il s'agit de passer à la régression linéaire multiple. En effet, nous sommes en présence de plusieurs variables explicatives. Même si les nuages de points dans le repère (x^j, y) peuvent être intéressants pour analyser le rôle de chaque variable explicative x^j, ils sont faussés parce qu'il y a souvent corrélation entre les variables explicatives. Nous nous tournons alors vers un outil spécifique : **le graphique des résidus partiels**.
Le résidu partiel associé à la variable x^j est le vecteur défini par

$$\hat{\varepsilon}_P^j = \hat{\varepsilon} + \hat{\beta}_j x^j$$

et en utilisant le fait que $\hat{\varepsilon} = y - \hat{y} = y - \sum_{k=1}^{p} \hat{\beta}_k x^k$ on obtient

$$(\hat{\varepsilon}_P^j) = y - \sum_{k \neq j} \hat{\beta}_k x^k$$

L'intérêt des résidus partiels est donc d'enlever l'effet estimé de toutes les variables explicatives autres que x^j sur y.
Penser aussi au théorème de Frish-Waugh : si on écrit le résultat de ce théorème avec $q = 1$ et le découpage $\{x^j\}$ et $\{x^1, \ldots, x^p\} \setminus \{x^j\}$ alors on obtient que le coefficient $\hat{\beta}^j$ est obtenu par la régression de $y - X_2\hat{\beta}^2 = y - \sum_{k \neq j} \hat{\beta}_k x^k$ sur x^j. Et ce vecteur $y - X_2\hat{\beta}^2 = y - \sum_{k \neq j} \hat{\beta}_k x^k$ est justement le résidu partiel. Si on a le bon modèle, alors en faisant le nuage de points $(X_{ij}, (\hat{\varepsilon}_j^P)_i)_{1 \leq i \leq n}$ on devrait voir une droite de coefficient $\hat{\beta}_j$.
Concrètement, pour chaque variable explicative x^j, on fait le graphe des résidus partiels associés à x^j en ordonnée, contre x^j en abscisse. **Si le lien est bien linéaire, alors le graphique montrera une tendance linéaire. Si au contraire le lien n'est pas linéaire, une tendance devrait apparaitre sur le graphique, liée à une fonction f, et il sera bon de remplacer x^j par $f(x^j)$.**

Remarque C.2. *Si on rajoute ou remplace une variable x^j par une transformée, il faut évidemment vérifier qu'on a amélioré le modèle et qu'on n'a pas fait de surajustement : on peut regarder le R^2 ajusté ou utiliser* `anova` *si les modèles sont emboités. On doit aussi regarder tous les graphiques diagnostiques du nouveau modèle.*

Pour obtenir les résidus partiels, utiliser la commande `resid(reg,type="partial")`. Chaque colonne du résultat donne l'ensemble des résidus partiels associé à une variable explicative donnée. Par exemple, pour avoir le graphique des résidus partiels associés à la variable x on écrit

`plot(x,resid(reg,type="partial")[,"x"])`. Pour voir la tendance de façon plus précise, on peut utiliser un lisseur, par exemple `scatter.smooth` dont la syntaxe est la suivante :

```
scatter.smooth(y~x)
```

On peut aussi obtenir directement ces graphiques en utilisant la fonction `crPlots(reg)` du package `car`.

Exercice C.1. *On utilise le fichier ozone. On choisit d'utilises les variables explicatives `max03`, `T12`, `Ne9` ainsi que `Vx9`. Faire les graphiques des résidus partiels associés à ces variables et commenter.*

La non linéarité peut aussi venir d'une **interaction entre les variables explicatives**. Prenons l'exemple du fichier Advertising.cvs. Il contient les données des ventes d'un produit en fonction des budgets de publicité pour la télévision, la radio et les journaux. Après avoir étudié l'influence de chaque variable, nous décidons d'utiliser `radio` et `TV` comme variables explicatives. Nous nous posons la question d'une interaction éventuelle entre ces deux variables. On peut afficher un graphique en 3 dimensions du plan de régression, ainsi que des résidus de la régression, en utilisant par exemple le package `rgl` (à charger). Cf code et figure interactive dans le document rgl_plot.html. On voit qu'il semble y avoir une interaction entre les deux variables explicatives. Il serait bon d'essayer un modèle où un terme d'interaction intervient (par exemple `radio:TV` ou autre).
Enfin **un changement de variable est parfois nécessaire sur la variable à expliquer**. On peut éventuellement détecter le problème à partir des graphiques données par défaut dans `plot(reg,which=c(1,3))`. Quand on voit un graphique comme celui de la figure C.2, où la variance des résidus augmente avec la valeur ajustée, on peut penser à un changement de variable sur y du type log (vérifier avant que la variable est bien positive, sinon ajouter éventuellement une valeur qui la rend positive).
D'autres transformations peuvent être utiles. Une transformation relativement utilisée dans la pratique est la transformation boxcox. Il s'agit de transformer la variable y par

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(y) & \text{si } \lambda = 0 \end{cases}$$

On utilise donc ces transformations quand on pense que $y(\lambda) = X\beta + \varepsilon$, au lieu de $y = X\beta + \varepsilon$, et ce pour un certain $\lambda > 0$.
On ne sait pas toujours à l'aide du graphique quelle transformation utiliser (i.e. quel lambda ?). Pour cela on peut utiliser la fonction `boxcox` du package `MASS`. Par exemple

```
library(MASS)
bc=boxcox(reg,lambda=seq(-3,3))
bestlambda=bc$x[which.max(bc$y)]# le lambda qui maximise la vraisemblance.
```

Remarque C.3. *(tirée de `https://newonlinecourses.science.psu.edu/stat501/node/2/`)*
« You will discover that data transformation definitely requires a "trial and error" approach. In building the model, we try a transformation and then check to see if the transformation eliminated the problems with the model. If it doesn't help, we try another transformation and so on. We continue this cyclical process until we've built a model that is appropriate and we can use. That is,

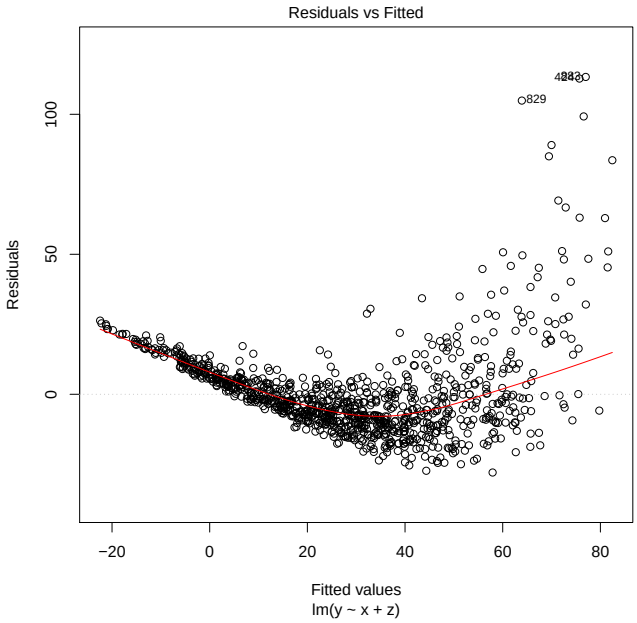


FIGURE C.2 – hétéroscédasticité : changement de variable en y ?

the process of model building includes model formulation, model estimation, and model evaluation [...]. We don't leave the model building process until we've convinced ourselves that the model meets the [...] conditions [...] of the linear regression model. One important thing to remember is that there is often more than one viable model. The model you choose and the model a colleague chooses may be different and yet both equally appropriate. »

C.2 Vérification de l'indépendance

Une hypothèse importante est que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes. Cela signifie par exemple que le fait que ε_i soit positif ne fournit aucune information sur le signe de ε_{i+1}.
L'indépendance des données est en général assurée par le protocole expérimental.
Cependant, des corrélations dans les erreurs se produisent souvent dans le contexte d'observations mesurées dans le temps. Il peut y avoir des quantités aléatoires (comme des facteurs environnementaux auxquels on n'a pas pensé) qui influencent les erreurs dans le temps.
Donnons un exemple de cas où on a non indépendance entre les erreurs. Si les erreurs vérifient

$$\varepsilon_{i+1} = \rho\varepsilon_i + \nu_i, \qquad \text{avec } \nu_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\nu^2),$$

cela signifie que l'erreur ε_{i+1} est liée à l'erreur précédente ε_i, il n'y a donc pas indépendance des ε_i. On parle d'autocorrélation. Pour détecter le problème, on pourra se contenter dans une première approche de regarder le graphique de ε en ordonnée et l'indice en abscisse.
Illustrons ce qui se produit au niveau des résidus quand les erreurs sont auto-corrélées : cf figure C.3. On a simulé un modèle $y_i = 2x_i + z_i + 1 + \varepsilon_i, i = 1, \ldots, n$ avec des erreurs auto-corrélées comme indiqué ci-dessus, en faisant varier le coefficient de corrélation ρ (on a pris $n = 100$, x_i et z_i gaussiennes standard , $\sigma_\nu = 0.5$).
Pour information, on trouve alors, pour le coefficient R^2 pour les valeurs respectives de $\rho = (0, 0.5, 0.9, -0.9)$, $R^2 = (0.94, 0.91, 0.81, 0.80)$.

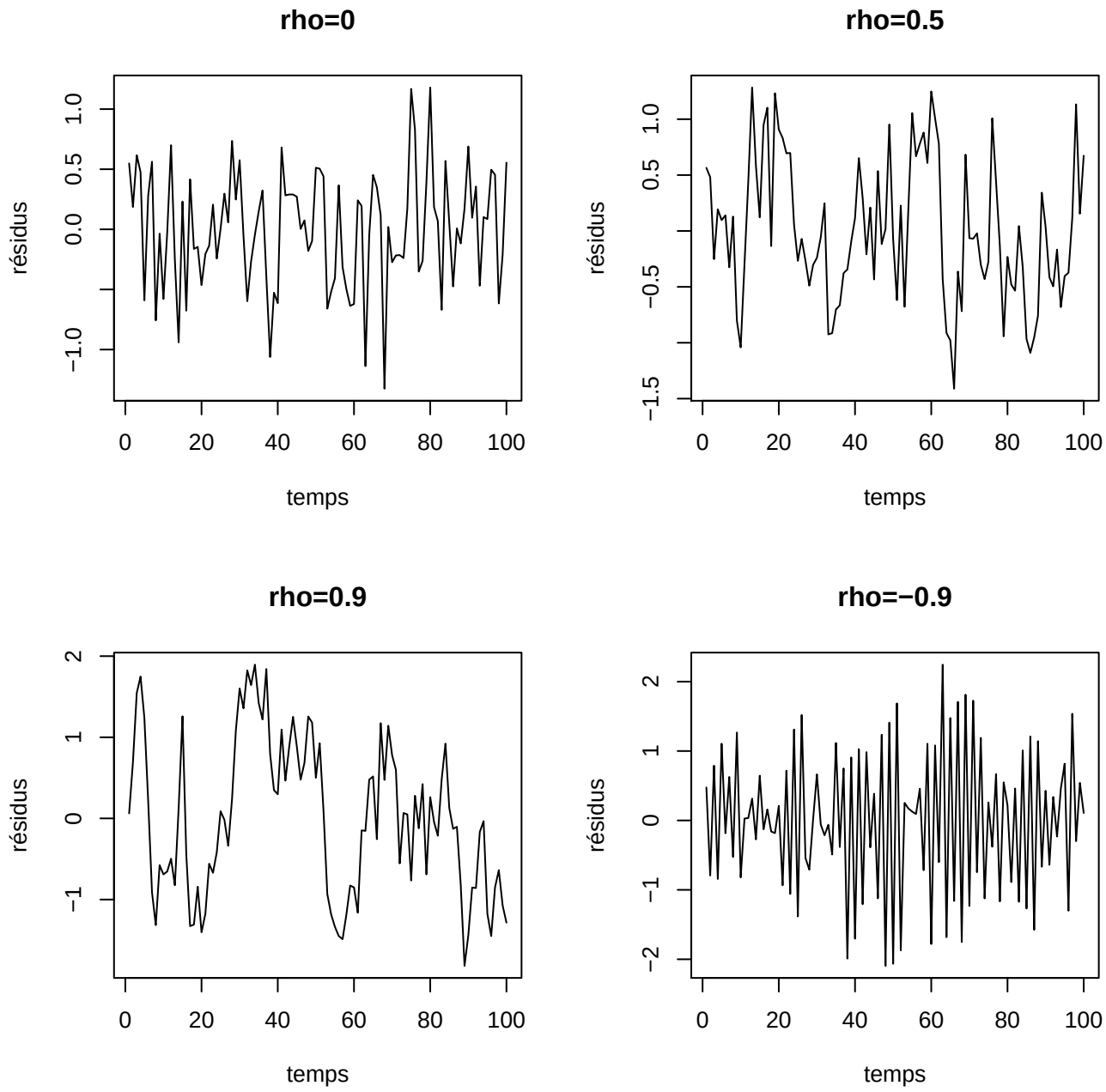


FIGURE C.3 – autocorrélation des résidus

Si on répète cette simulation un grand nombre de fois, on trouvera en moyenne les bons coefficients, mais avec d'autant plus de variance que ρ est grand. On trouve approximativement un écart-type de (0.05,0.06,0.10,0.12) pour le coefficient de x.

Ce comportement s'explique mathématiquement : si les erreurs ne sont pas indépendantes, la matrice de covariance du vecteur ε n'est plus égale à $\sigma^2 I_n$ mais est égale à une matrice Ω qui n'est plus diagonale. L'estimateur $\hat{\beta}^{OLS}$ est toujours sans biais mais son opérateur de covariance n'est plus de la forme $\sigma^2(X^T X)^{-1}$. Si on connait Ω, il existe un estimateur sans biais et qui a une variance plus petite que $\hat{\beta}^{OLS}$: il s'appelle l'estimateur des moindres carrés généralisés ("gls"=generalized least squares") mais nécessite la connaissance de Ω, inconnue en générale, et qu'il faut donc estimer...

De manière générale, **il faut détecter graphiquement si les résidus suivent un processus**

particulier.

Si on suspecte une forme de corrélation entre les erreurs, du type des deux présentées ci-dessus, il existe aussi des tests : par exemple le test de Durbin-Watson pour détecter le problème ci-dessus, le test de runs pour détecter un problème d'indépendance plus générale.
Pour le test de runs : on regarde les signes des résidus, on les transforme en facteurs 0/1, et on donne l'échantillon correspondant à la fonction `runs.test` du package `tseries`. Le test de Durbin-Watson peut se faire à l'aide de la fonction `durbinWatsonTest` du package `car`. On lui donne la sortie de lm en argument.
Remarque : le test de Durbin-Watson fonctionne très bien pour détecter l'auto-corrélation dans les trois cas (p-valeur du test très proche de zéro sur 1000 expériences avec $\rho = (0.5, 0.9, -0.9)$). Mais finalement les problèmes sont surtout importants en cas de forte corrélation, ce qu'on peut voir avec un simple graphique. Il existe des méthodes pour traiter ce cas d'auto-corrélation des erreurs (par exemple la fonction `gls` du package `nlme` avec le résultat de la fonction `lm` comme argument combiné à l'argument `correlation=corAR1`). Les autres formes possibles de corrélation au niveau des résidus seront traitées dans le cours de séries temporelles.

Remarque C.4. *Plus généralement encore, dans des données temporelles, on peut avoir un modèle où la variable y_{i+1} est influencée par la variable y_i (ou d'autres valeurs précédentes). Le modèle linéaire n'est alors pas du tout adapté (on peut éventuellement repérer le problème à l'aide du R^2 qui sera faible dans ce cas ou simplement utiliser des tests du type Durbin-Watson sur les y_i). Ces types de modèles seront étudiés dans le cours de séries temporelles.*

Exercice C.2. *On utilise le fichier ozone et on fait la régression de `maxO3` contre `T12`, `Vx9`, `Ne9` et `maxO3v` (choix de variables obtenu après une procédure de choix de modèles, à partir de la fonction `regsubsets`, cf cours ultérieur). Vérifier l'hypothèse d'indépendance sur ces données. (on rappelle qu'il s'agit de données temporelles donc la vérification de cette hypothèse fait sens).*

C.3 Vérification de l'homoscédasticité

Au sens strict, on a hétéroscédasticité quand les variances des variables de bruit ε_i ne sont pas constantes, mais dépendent d'un cofacteur (le temps, l'espace, une variable explicative etc). Une forte violation de l'hypothèse d'homoscédasticité peut entraîner des conséquences sur :

- les estimations des écart-types des paramètres $\hat{\sigma}_{\hat{\beta}_i}$ ("standard error")
- les risques des tests
- les intervalles de confiance.

Si la variance des ε_i n'est pas constante, alors la variance des y_i n'est pas constante non plus.
Exemple concret d'hétéroscédasticité (issus de l'article wikipedia `https://en.wikipedia.org/wiki/Heteroscedasticity`) :
« A classic example of heteroscedasticity is that of income versus expenditure on meals. As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food ; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption. »

Comme les ε_i sont estimés par les résidus, **on va faire des graphiques associés à ces résidus. On dira plus généralement qu'on a un problème d'hétéroscédasticité quand la variance des y_i n'est pas constante** (ça peut être dû à une transformation nécessaire de la variable à expliquer y).

Nous proposons **plusieurs graphiques possibles pour détecter une hétéroscédasticité.**
Il est recommandé de **tracer les résidus ou les résidus standardisés en fonction des valeurs ajustées** $\hat{y}_i$. Les résidus standardisés sont égaux à $\frac{\hat{\varepsilon}_i}{\hat{\sigma}_{\hat{\varepsilon}_i}}$ où $\hat{\sigma}_{\hat{\varepsilon}_i}$ est l'estimateur de la variance du résidu $\hat{\varepsilon}_i$. On peut obtenir ces résidus standardisés grâce à la commande `rstandard(reg)`. On rappelle que le graphe des résidus non standardisés versus les valeurs ajustées s'obtient directement à l'aide de la commande `plot(reg,which=1)`. Ces deux graphes (standardisés/non standardisés) ne sont pas toujours très différents en allure, ça dépend du design. Un autre type de résidus standardisés différemment existe, il s'agit des résidus studentisés, qu'on obtient par la commande `rstudent(reg)`. Ceux-ci sont censés suivre une loi de Student. Ils sont souvent très proches des résidus standardisés. Ils sont plutôt utilisés pour repérer des outliers, cf section suivante.
Un autre graphe peut aider à détecter une éventuelle hétéroscédasticité, il s'agit du 3ème graphique de sortie de `plot(reg)`, auquel on peut accéder directement par `plot(reg, which=3)` . C'est le graphe de la racine des valeurs absolues des résidus standardisés contre les valeurs ajustées.
Si l'hypothèse d'homoscédasticité est vérifiée, les résidus standardisés devraient être approximativement de même variance.
Si la variance des résidus varie avec la valeur ajustée, une transformation sur la variable y peut être judicieuse. On cherche alors à faire une transformation sur y qui stabilise la variance (boxcox par exemple, cf section précédente).
Pour en savoir plus sur les transformations possibles sur y dans ce cas (variance des résidus qui change avec la valeur ajustée, vous pouvez aussi consulter par exemple `https://en.wikipedia.org/wiki/Variance-stabilizing_transformation` ou bien `https://www.math.univ-toulouse.fr/%7Eazais/styles/other/student/modlin.pdf` page 37. Les transformations Box-Cox semblent les plus utilisées.
Parfois il vaut mieux simplement changer de modèle et passer aux modèles linéaires généralisés (cf cours suivant)

Mais éventuellement l'hétéroscédasticité peut venir d'un autre problème (a priori problème moins fréquent, d'ailleurs non traité dans les graphiques diagnostiques par défaut). En effet, comme évoqué en introduction de cette sous-section et dans l'exemple tiré de wikipedia, le problème peut venir du fait que l'erreur a une variance qui change avec cette variable explicative ou avec le temps... Alors d'autres choix que $\hat{y}$ en abscisse peuvent s'avérer plus pertinents comme l'indice si les données sont temporelles, ou une variable explicative, u n plot3d des résidus dans l'espace peut être pertinent si on suspecte une hétéroscédasticité spatiale etc.
Pour détecter ce type de problèmes, il existe également des tests, comme le test de Breush-Pagan. Il est possible que ces tests rejettent l'hypothèse d'homoscédasticité quand le modèle est mal spécifié (variables manquantes, non linéarité) alors que l'hypothèse d'homoscédasticité est pourtant bien vérifiée. On règle donc d'abord le problème de la spécification du modèle avant de procéder aux tests d'hétéroscédasticité.
Le test de Breush-Pagan s'occupe du problème de test suivant (avec deux variables explicatives pour fixer les idées)

$$H_0 : y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i, \quad \mathbf{Var}(\varepsilon_i) = \sigma^2, \quad i = 1, \dots, n$$

contre

$$H_1 : y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i, \quad \mathbf{Var}(\varepsilon_i) = \eta_1 + \eta_2 x_i + \eta_3 z_i + \omega_i$$

(où $\omega_i \overset{iid}{\sim} N(0,1)$).
Pour utiliser le test de Breush-Pagan dans R, on peut par exemple utiliser la fonction `bptest` du package `lmtest`.
Pour tester une éventuelle hétéroscédasticité liée à la variable x seule ($\mathbf{Var}(\varepsilon_i) = \eta_1 + \eta_2 x_i + \omega_i$) on peut utiliser

```
bptest(reg,varformula=~x)
```

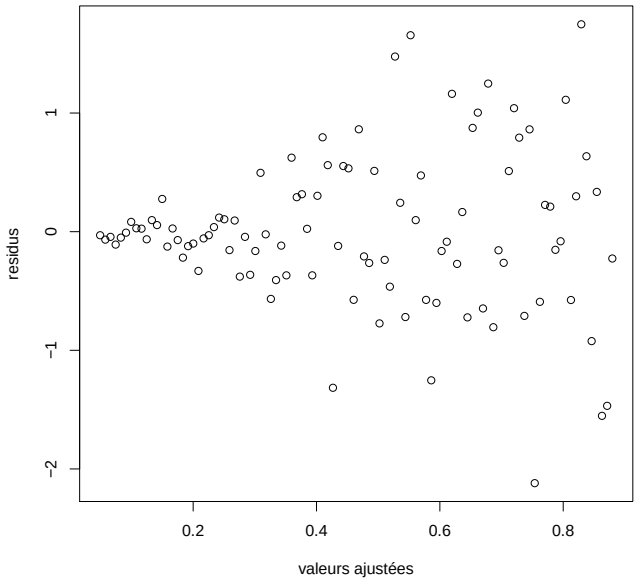


FIGURE C.4 – La variance des résidus augmente avec la valeur ajustée

On peut aussi préciser une autre variable dans l'argument `varformula` (comme l'indice, x^2 etc).
En première intention, on préfère souvent les graphiques.
A nouveau, en cas d'hétéroscédasticité, la matrice de variance-covariance des erreurs Ω n'est plus égale à $\sigma^2 I_n$.On peut là encore utiliser les moindres carrés quasi-généralisés (comme pour le problème des erreurs corrélées). Par exemple, si on pense à une variance qui croit avec une des variables explicatives, on utilise l'argument `weights=varPower())` dans la fonction `gls`. Mais il faut avoir une réelle idée de la forme de l'hétéroscédasticité.
Il existe en réalité plein d'autres méthodes pour traiter ce type de problème d'hétéroscédasticité.
Attention cependant, même en cas d'une déviation par rapport à l'hypothèse d'homoscédasticité, il n'est pas toujours recommandé d'utiliser les MCG (gls). En effet, ils ne sont réellement utiles que quand on a une grande hétéroscédasticité, car sinon ils ont tendance à gonfler les écarts-types et réduisent alors la puissance des tests. Il faut donc s'abstenir de les utiliser quand l'hétéroscédasticité observée n'est pas très forte.

C.4 Vérification de la normalité

On peut vérifier cette normalité, quand la taille d'échantillon est faible, en traçant un Q-Q plot (ou diagramme quantile-quantile) des résidus standardisés. Si l'hypothèse de normalité est satisfaite, le graphique est proche de la première bissectrice.
On peut afficher ce graphe à l'aide de la commande `qqnorm(rstudent(reg))`. On peut obtenir ce graphe directement à l'aide la commande `plot(reg,which=2)`.
Cela donne la figure C.6 pour l'exemple des eucalyptus, pour lequel on accepte l'hypothèse de normalité.
Cependant, quand la taille d'échantillon est grande et sous des conditions assez larges, cette hypothèse n'est pas importante : en effet on peut montrer que, sous certaines conditions, le coefficient $\hat{\beta}$ "devient" gaussien et donc que les tests de pertinence, qui sont eux même basés sur cette hypothèse de normalité, demeurent valables. (Donc on peut avoir un qqplot qui est loin de la bissectrice, tout en ayant des résultats quand même valables.)

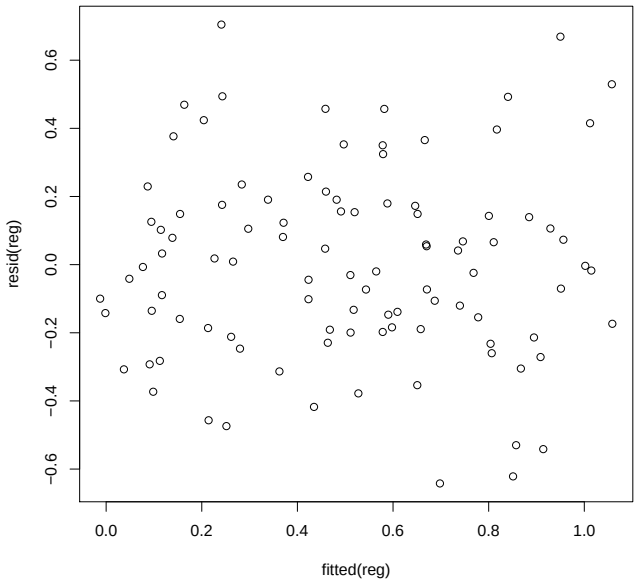


FIGURE C.5 – Un exemple de graphique où il n'y a pas de problèmes

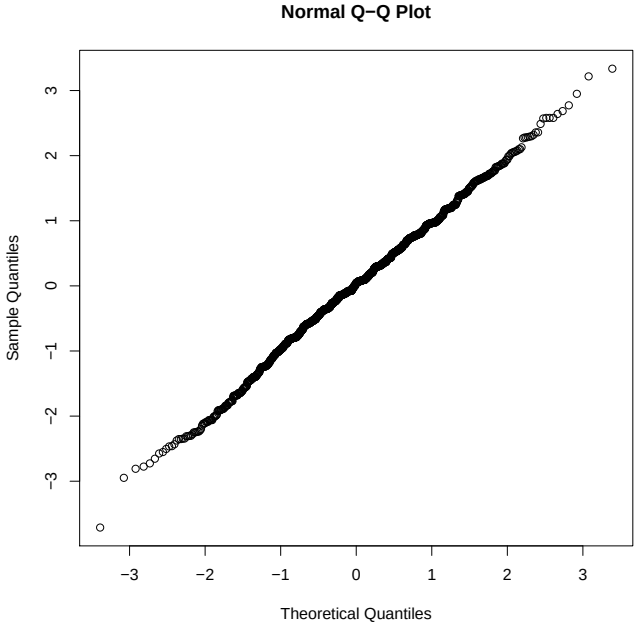


FIGURE C.6 – Q-Q plot

Evidemment la formulation "n est assez grand" est vague. En fait cela dépend de la distribution du bruit (et donc de la distribution de y). Pour un bruit avec des queues légères on aura besoin de moins de données que pour un bruit avec des queues lourdes. Il existe aussi des distributions du bruit pour lesquelles ces propriétés asymptotiques n'ont pas lieu (ex : si le bruit suit une loi de Cauchy). Mais les conditions sous lesquelles ce comportement asymptotique gaussien se produit sont relativement larges.

Exercice C.3. *On utilise le dataset* `cars` *qui est un dataframe avec 2 colonnes : la vitesse de la voiture et la distance de freinage. Faire la régression de* `distance` *contre* `speed`*. Faites les vérifications préconisées. Identifier un problème possible et essayez d'y remédier. Vérifier que vous avez bien amélioré les résultats.*

C.5 Points aberrants et points leviers

C.5.1 Points leviers

On appelle point levier un point (x_i, y_i) tel que x_i est éloigné du centre de gravité du nuage de points $(x_j)_{1\leq i\leq n}$. Autrement dit, il s'agit d'une observation telle que l'une (ou plusieurs) des variables explicatives a des valeurs très différentes de l'ensemble de valeurs prises par les autres observations. Un exemple en dimension 2 est donné dans la figure C.7. On voit qu'un des points est éloigné des autres : c'est la valeur de la variable x^2 qui est atypique pour ce point.
NB : une distance particulière est utilisée dans la mesure de l'éloignement d'une observation par rapport au centre de gravité : cette distance tient compte de la forme du nuage.

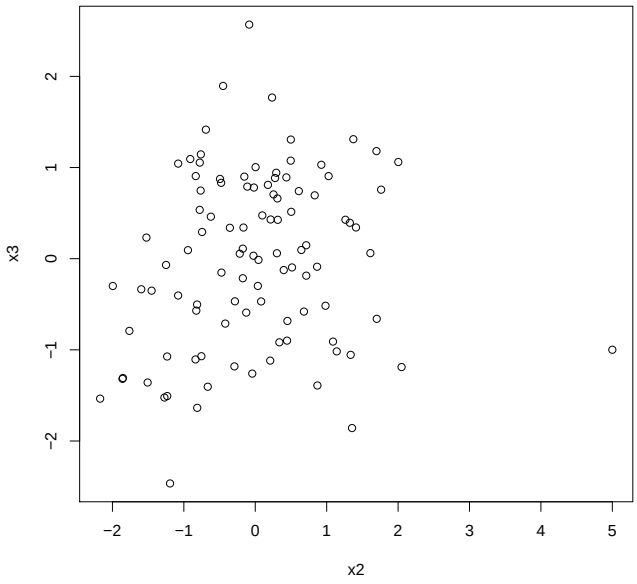


FIGURE C.7 – Exemple d'un point levier en dimension 2

Evidemment, si on a plus de trois variables, on ne pourra pas visualiser ce phénomène sur un graphique. On dispose cependant de la quantité suivante

$$h_{ii} = x_i^T (X^T X)^{-1} x_i = (x_i - \bar{x})^T (\tilde{X}^T \tilde{X})^{-1} (x_i - \bar{x})$$

(on rappelle la notation : x_i désigne la $i$ème observation, autrement dit aussi la $i$ème ligne de X, ici $\tilde{X}$ est la matrice des variables explicatives centrées). On admet que h_{ii} est entre 0 et 1 et que, **plus h_{ii} est élevé, plus l'observation est éloignée des autres observations (en terme des variables explicatives** seulement, on ne se préoccupe pas de la valeur de y_i). (ici l'éloignement est mesuré avec une distance qui dépend du design, plus précisément de la dispersion du nuage dans les différentes directions).
Voyons maintenant ce qu'un tel point provoque au niveau d'une régression.
On rappelle que le vecteur des valeurs ajustées $\hat{y}$ est la projection orthogonale de y sur l'espace engendré par les colonnes de la matrice de design X :

$$\hat{y} = Py$$

où P est la matrice de la projection orthogonale $P = X(X^TX)^{-1}X^T$. Le terme h_{ii} est en fait le $i$ème terme diagonal de la matrice de projection P. On a donc, si on note h_{ij} le terme général de la matrice P,

$$\hat{y}_i = \sum_j h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$$

On admet également que, quand h_{ii} est grand, les autres termes h_{ij} pour $j \neq i$, sont petits.
En utilisant des propriétés d'une matrice de projection, on peut montrer que (on admet) :

- pour tout i on a : $0 \leq h_{ii} \leq 1$.
- si $h_{ii} = 1$, on a $h_{ij} = 0$ pour tout $i \neq j$.
- $\sum_i h_{ii} = p$ donc la valeur moyenne des h_{ii} est $\frac{p}{n}$.

Ainsi si h_{ii} est « grand », y_i influe fortement sur $\hat{y}_i$.
Un point (x_i, y_i) est un point levier si la valeur h_{ii} pour cette observation est "grande".
En pratique, certains auteurs préconisent le seuil $2p/n$ si $p > 6$ et $n - p > 12$, d'autres le seuil $3p/n$ et d'autres encore le seuil 0.5.

Les valeurs h_{ii} sont données par la commande `hatvalues(reg)`.

C.5.2 Points aberrants

On rappelle que l'on considère les résidus $\hat{\varepsilon}_i$ comme des estimateurs des vraies erreurs. Cependant ils n'ont pas les mêmes caractéristiques que les erreurs ε_i. En effet ils sont bien gaussiens et centrés mais leur matrice de covariance, $\sigma^2(I - P)$, n'est pas diagonale, ce qui signifie que les résidus ne sont pas indépendants. Les éléments diagonaux de la matrice de covariance sont égaux à $\sigma^2(1 - h_{ii})$, donc les résidus ne sont pas non plus de variance égale (contrairement aux vraies erreurs). Pour les rendre de variance égale à 1, c'est-à-dire pour les standardiser, il suffirait de les diviser par $\sqrt{\sigma(1 - h_{ii})}$. Or h_{ii} est calculable (il ne dépend que du design), mais pas σ^2. On remplace donc σ^2 par son estimateur $\hat{\sigma}^2$. Cette standardisation donne ce qu'on appelle des **résidus standardisés**.
On note t_i les résidus standardisés. On a donc

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Si les hypothèses sur le bruit ε sont vérifiées alors **les résidus standardisés doivent suivre approximativement une loi $\mathcal{N}(0, 1)$** (mais ils ne sont pas indépendants).

Ainsi, en théorie, 95% de ces résidus doivent se trouver dans l'intervalle $[-2, 2]$ et l'immense majorite (99,7%) de ces résidus se trouvent dans l'intervalle [-3,3].
L'introduction des résidus standardisés conduit naturellement à la notion de valeur aberrante.
Une observation qui admet un résidu standardisé anormalement élevé est appelée observation aberrante. En théorie, comme naturellement 5% de ces résidus se trouvent en dehors de la bande de confiance ± 2, on s'intéresse surtout à ceux qui sont nettement en dehors de cette bande de confiance.
Un point est donc aberrant si :

- il est mal prédit, c'est-à-dire son résidu $\hat{\varepsilon}_i = y_i - \hat{y}_i$ est élevé.
- le point est levier : en effet si h_{ii} est grand ($h_{ii} \approx 1$), alors $1 - h_{ii} \approx 0$ et donc le rapport est élevé.

L'explication de ces valeurs aberrantes peut être ardue. Elles peuvent être engendrées par des erreurs de mesure ou être issues d'un changement de population. Pour afficher, par ordre croissant, les résidus standardisés qui se trouvent en dehors de la bande de confiance $[-2; 2]$, on peut faire comme suit

```
indices=which(rstandard(reg)>2|rstandard(reg)<(-2))
sort(rstandard(reg)[indices])
```

Quel problème peut-on avoir avec des valeurs aberrantes ?
Un point aberrant peut fausser l'estimation : on dit que l'estimation par moindres carrés n'est pas robuste.
En effet, si on note $\hat{y}_j^{(-i)}$ la prédiction de la $j$ème observation lorsque l'analyse a été faite en retirant l'observation i, on a

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_j^{(-i)})^2}{p\hat{\sigma}^2} = \frac{1}{p} \frac{h_{ii}}{1 - h_{ii}} t_i^2$$

On parle de point "influent" quand cette distance est grande. On voit donc que, quand le résidu $\hat{\varepsilon}_i$ est élevé (mauvaise prédiction) et que le point est levier (grand h_{ii}), ce qui se produit si l'observation est aberrante, alors l'estimation du coefficient β est très différente quand on enlève l'observation i.
Une observation aberrante peut donc fausser nos calculs. Ce n'est cependant pas toujours le cas, puisque le poids de $\hat{\varepsilon}_i$ et de h_{ii} dans le calcul de D_i comparé à leur poids dans le calcul de t_i n'est pas le même : le fait d'être un point aberrant est plus lié à la valeur du résidu $\hat{\varepsilon}_i$ qu'à la valeur du h_{ii}, donc plus lié à une mauvaise prévision $\hat{y}_i$.

Cette distance est appelée distance de Cook, Le calcul de la distance de Cook est fait dans R par la fonction `cooks.distance`. On lui donne le résultat de la fonction `lm`, appelons le reg, et la distance de Cook D_i associée à l'observation i est donnée par `cooks.distance(reg)[i]`.
Il nous faut définir la valeur seuil à partir de laquelle nous pouvons dire que l'influence est exagérée. La règle la plus simple est $D_i > 1$. Mais cette règle est jugée un peu trop permissive par certains et on lui préfère parfois la disposition plus exigeante suivante

$$D_i > \frac{4}{n - p - 1}$$

Que faire des points aberrants ?

Tous les auteurs s'accordent à dire que la suppression automatique des observations atypiques n'est pas la solution. Il faut comprendre pourquoi l'observation se démarque autant et proposer des solutions appropriées, voici quelques idées :

- Premier réflexe : vérifier les données, y-a-t-il des erreurs de saisie ou des erreurs de transcription ? Dans ce cas, il suffit de corriger les valeurs recensées.
- Si la distribution est très asymétrique, il est plus indiqué de tenter de symétriser la distribution avec une transformation de variables adéquate : par exemple, avec des salaires, on peut avoir un problème d'échelle, et on peut donc penser à utiliser la fonction log.
- Les observations atypiques peuvent provenir, dans le cas de données qui ont été recueillies dans le temps (ex : le taux d'ozone), d'un phénomène particulier circonscrit dans le temps (ex guerre, famine). Dans ce cas, on peut introduire une variable explicative qualitative binaire (un facteur à deux modalités 0/1, cf chapitre 5) qui code ce phénomène . On refait la régression en ajoutant cette variable explicative.
- S'il apparait que les observations incriminées ne correspondent pas à la population étudiée (ex : 2 ou 3 citadins dans un échantillon de gens vivant à la campagne) alors on supprime ces observations.

Quant aux points leviers, il est bon de les repérer et de les noter aussi, puis de comprendre pourquoi ces points sont différents : erreur de mesure, erreur d'enregistrement, ou appartenance à une autre population. Même quand ils ne sont pas influents, i.e. sans ces points les estimations ne changent pas ou très peu, on peut se poser la question de la validité du modèle jusqu'à ces points extrêmes. Peut-être aurait-on, avec plus de mesures autour de ces points, un modèle qui changerait, annonçant un modèle différent pour cette population ? Après mûre réflexion ces valeurs pourront être éliminées ou conservées. Dans le premier cas aucun risque n'est pris au bord du domaine, quitte à sacrifier quelques points. Dans le second cas le modèle est étendu de manière implicite jusqu'à ces points.

Un point peut être aberrant sans être influent (c'est le cas s'il a une valeur anormale pour y_i mais une valeur de x_i proche du centre de gravité)
exemple simulé :

```
> x2=runif(99)
> x3=runif(99,min=-1,max=1)
> eps=0.3*rnorm(99)
> y=1+2*x2+x3+eps # des données qui ne devraient pas poser de problèmes
> x2bar=0.5 #une valeur de x2 proche de la moyenne des autres valeurs de x2
> x3bar=0# idem
> x2=c(x2,x2bar) #on a donc rajouté une observation proche du centre de gravité
> x3=c(x3,x3bar)
> yaberr= 3 #observation aberrante : la bonne valeur devrait se trouver
#autour de 1+2*0.5+0=2 avec une erreur max de 2*sigma=2*0.3=0.6
> y=c(y,yaberr)
> reg=lm(y~x2+x3)


> rstandard(reg)[100]#on a bien un point aberrant
```

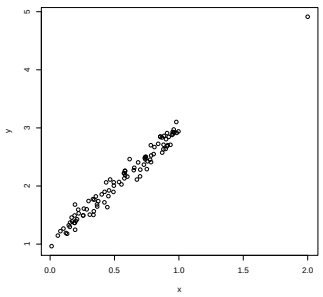


FIGURE C.8 – Exemple d'un point levier non influent

```
     100
3.115163

> hatvalues(reg)[100]# comme prévu, petite valeur pour hii (<2p/n=0.06)
       100
0.01034966

> cooks.distance(reg)[100]#  distance de cooks <4/(100-3-1)=0.04, point non influent
     100
0.031043
```

Un point peut être levier sans être influent : un exemple de point levier non influent est donné sur la figure C.8. En effet, la valeur prise par x est atypique mais le point se situe dans le prolongement de la droite de régression, et donc son résidu est donc petit.
Un point levier et aberrant sera à coup sûr un point influent.

Exercice C.4. *On travaille avec le fichier "conso.txt". Il s'agit de données concernant des voitures : on cherche à expliquer la consommation (`conso`) par 3 caractéristiques : `prix`, `puissance` et `poids`*

1. *Repérer les points aberrants. Pourquoi sont-ils aberrants ?*
2. *Quels sont les points dont la suppression modifierait franchement l'estimation de β ?*

Mentionnons enfin que dans les cas où l'erreur est très loin d'être gaussienne (queues lourdes) ou en présence d'un trop grand nombre d'outliers, il existe des méthodes dites robustes, qui sont donc différentes de l'estimation par moindres carrés (utilisation d'une autre fonction de perte, par exemple on utilise "perte de Huber" plutôt que la perte quadratique, ou on utilise des méthodes faisant appel à la médiane ou plus généralement aux quantiles). Ces méthodes sont appelées méthodes de régression robustes.

Annexe D

Colinéarité des variables explicatives

Dans cette annexe on se concentre sur l'leffet de la présence de variables colinéaires dans un modèle linéaire gaussien.

Définition D.1.
— *On dit que deux variables x^2 et x^3 sont corrélées quand leur coefficient de corrélation empirique ρ_{x^2,x^3} est élevé (i.e. proche de 1, typiquement > 0.8).*
— *On peut généraliser à un ensemble de variables $x^1, x^2, \ldots, x^p$, en disant que ces variables sont multicorrélées si une des variables x^j est corrélée avec une combinaison linéaire des autres variables.*

D.1 Colinéarité et estimation

D'après le chapitre 1, deux variables x^2 et x^3 sont corrélées si il existe un λ tel que

$$x^3 - \bar{x^3}\mathbf{1} \approx \lambda(x^2 - \bar{x^2}\mathbf{1})$$

Cela se réécrit

$$x^3 \approx \lambda x^2 + (\bar{x^3} - \lambda\bar{x^2})\mathbf{1}$$

Cela signifie que x^3 est proche d'une combinaison linéaire de x^2 et $\mathbf{1}$ ou que les variables centrées sont presque collinéaires.
Illustrons le mécanisme de la colinéarité avec deux variables explicatives. On suppose pour simplifier que l'on a deux variables explicatives x^1 et x^2 centrées et pas d'intercept. On peut montrer que le terme $\frac{1}{\rho^2_{x^1,x^2}-1}$ intervient dans la matrice $(X^TX)^{-1}$.

— Si la corrélation entre les deux variables x^1 et x^2 est parfaite, c'est-à-dire le coefficient de corrélation entre les deux variables est égal à ± 1, X^TX est non inversible (le terme $\frac{1}{\rho^2_{x,y}-1}$ est infini). La variance de l'EMC est donc infinie : l'EMC n'est pas unique ! (la formule (3.3) n'a en fait pas de sens).
— Si les variables sont corrélées, on a par définition $\rho^2_{x^1,x^2} \approx 1$. Alors le terme $\frac{1}{\rho^2_{x,y}-1} \approx 0$. Et donc les termes de la matrice $(X^TX)^{-1}$ sont très grands.

Ce problème de variabilité se pose de manière générale avec un nombre quelconque de variables (centrées ou pas). Il se pose quasiment systématiquement en grande dimension (pas l'objet de ce cours). En cas de trop grande colinéarité/multicolinéarité des variables explicatives, notre estimateur a une variance trop grande et est inutilisable.

De manière générale, dire que x^2, $x^3 \dots x^p$ sont corrélées signifie que les variables centrées correspondantes $x^2 - \bar{x^2}, \dots, x^p - \bar{x}^p$ sont telles qu'il existe j tel que la variable centrée $x^j - \bar{x^j}$ est combinaison linéaire des autres variables centrées :

$$\exists j : \exists \lambda_2 \dots, \lambda_p : x^j - \bar{x^j}\mathbf{1} \approx \sum_{k \neq j, k=2 \dots p} \lambda_k (x^k - \bar{x^k}\mathbf{1})$$

ce qui est équivalent à dire que

$$\exists j : \exists \lambda_2 \dots, \lambda_p : x^j \approx \bar{x^j}\mathbf{1} + \sum_{k \neq j, k=2 \dots p} \lambda_k (x^k - \bar{x^k}\mathbf{1})$$

Cela signifie aussi (c'est équivalent) qu'il existe j et des λ_k tels que

$$x^j \approx \lambda_1 \mathbf{1} + \sum_{k \neq j, k \neq 1} \lambda_k x^k$$

c'est-à-dire

$$x^j \approx \sum_{k \neq j} \lambda_k x^k$$

Exemple D.1. *Avec $p = 4$. Supposons que la variable x^2 soit très corrélée aux autres variables. Cela signifie que x^2 est très proche d'une combinaison linéaire des autres variables. Par exemple*

$$x^4 \approx x^1 + x^2 + x^3$$

Nous postulons un modèle avec ces quatre variables

$$y = \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \varepsilon$$

Supposons qu'en réalité on ait

$$y = x^1 + x^2 + x^3 + \varepsilon$$

La présence d'un petit bruit ε fait qu'on a

$$y \approx x^1 + x^2 + x^3$$

Mais, en utilisant le fait que $x^4 \approx x^1 + x^2 + x^3$, on peut aussi faire l' approximation suivante

$$y \approx x^4$$

et aussi $y \approx \frac{1}{2}(x^1 + x^2 + x^3 + x^4)$ et une infinité d'autres combinaisons...
Cette incertitude sur l'écriture de y en fonction des x^j en cas de corrélation se reflète dans l'augmentation de la variance de $\hat{\beta}$.

De manière générale, à cause de la colinéarité des variables, plusieurs problèmes peuvent surgir :

- les valeurs/signes des coefficients sont contradictoires, elles ne concordent pas avec les connaissances du domaine.
- Les variances des estimateurs sont fortes.

— Les valeurs des variances sont si fortes, que les tests de significativité sur ces coefficients (cf sections suivantes) feront accepter l'hypothèse que le coefficient est nul. On rejettera donc parfois à tort une variable significative.
— Les résultats sont instables, l'adjonction ou la suppression de quelques observations modifie du tout au tout les valeurs et les signes des coefficients.

En conclusion, la variance de l'EMC est très liée au design.
Elle est aussi liée au rapport n/p.

— Si n est plus petit que p, on perd l'unicité de l'EMC (et la formule !). La méthode n'est alors pas utilisable.
— Si $n \geq p$ mais p grand, alors on a de très grandes chances d'avoir des variables corrélées, et donc une grande variabilité : l'EMC est là aussi sans intérêt. On utilise alors d'autres méthodes (non traitées ici, cf cours "choix de modèles", ex : Lasso, ridge etc).
— Mais même avec n est beaucoup plus grand que p, on peut avoir des des problèmes de multicolinéarité (mais moins systématique).

Exemple D.2. *(problème de variabilité liée à la colinéarité)*
Introduisons un exemple simulé pour faire apparaître ce problème de variabilité. Prenons deux variables x^1 et x^2 qui sont presque colinéaires :

$$x^2 \approx 2x^1$$

Prenons par exemple $x^2 = 2x^1 + c$ avec un très petit c.

```
> x1=rnorm(100)
> c=0.001*rnorm(100)
> x2=2*x1+c
> eps=rnorm(100)
> y=2*x1+3*x2+eps
> summary(lm(y~x1+x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7085 -0.7321 -0.0025  0.7044  2.2553

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1377     0.1022   1.347    0.181
x1           28.8434   219.2668   0.132    0.896
x2          -10.4476   109.6400  -0.095    0.924
```

On voit donc que les variances sont énormes : de l'ordre de 10000, alors que la variance du bruit ε n'est que de 1. De plus, les coefficients sont estimés à $(29, -10)$, ce qui est très loin de $(2, 3)$.

Remarque D.1. *Pour les personnes qui connaissent le conditionnement d'une matrice (cf wikipedia par exemple), le comportement des coefficients quand il y a corrélation des variables est naturel.*

Pour y voir encore un peu plus clair sur les effets de la colinéarité sur l'EMC, on donne le théorème suivant. Imaginons qu'on ait un ensemble de variables explicatives $x^1, \ldots, x^p$ à notre disposition, et qu'on sépare en deux l'ensemble des variables, disons $\{x^1, \ldots, x^q\}$ et $\{x^{q+1}, \ldots x^p\}$. On note $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{complet}}$ le vecteur des coefficients associé aux q premières variables dans le modèle complet où on utilise toutes les p variables explicatives. On note $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{seules}}$ le vecteur de coefficients de ces q mêmes variables explicatives, mais dans le modèle où elles seraient les seules variables explicatives.
Ajoutons encore quelques notations : On note X_1 la matrice associée aux q premières variables explicatives (c'est donc la matrice X privée de ses $p-q$ dernières colonnes) et X_2 la matrice associée aux $p-q$ dernières variables explicatives. On considère donc les deux modèles suivants :

$$y = X_1 \boldsymbol{\beta_1}^{complet} + X_2 \boldsymbol{\beta_2}^{complet} + \varepsilon$$

$$y = X_1 \boldsymbol{\beta_1}^{seule} + \varepsilon$$

On note également $M_2 = I - P_2$ la projection sur l'orthogonal de l'espace engendré par les $p-q$ dernières variables.
Si on fait la régression de y sur les q premières variables seulement, en "oubliant" les $p-q$ dernières, on obtient $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{seules}} = (X_1^T X_1)^{-1} X_1^T y$. Mais attention ce vecteur $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{seules}}$ ne coincide pas avec le vecteur $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{complet}}$ des coefficients de ces mêmes q premières variables explicatives quand on ajoute les $p-q$ dernières variables au modèle.

Théorème D.1. *(Théorème de Frish-Waugh)*

1. *Le vecteur* $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{complet}}$ *est donné par la régression de* y *sur les* $M_2 x^j$ *pour* $j = 1 \ldots, q$ *:*

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathbf{complet}} = \left[(M_2 X_1)^T M_2 X_1 \right]^{-1} (M_2 X_1)^T y$$

2. $$\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{complet} = (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{complet})$$

   Autrement dit $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{complet}$ *peut être obtenu en régressant* $y - X_2 \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{complet}$*, la partie de* y *inexpliquée par les* $p-q$ *dernières variables explicatives, sur les* q *premières variables explicatives.*

Commentaires sur le second item

— Examinons le cas particulier où $q = 1$ pour simplifier, c'est-à-dire une seule variable explicative est mise de côté (dans cet exemple, x^1 n'est pas forcément l'intercept). Comparons le calcul du coefficient $\hat{\beta}_1^{complet}$ de cette variable dans le modèle complet par rapport à son coefficient $\hat{\beta}_1^{seule}$ dans un modèle où cette variable est prise comme seule variable explicative. On considère donc les deux modèles suivants

$$y = \beta_1^{complet} x^1 + \sum_{j \neq 1} x^j \beta_j + \varepsilon$$

et

$$y = \beta_1^{seule} x^1 + \varepsilon.$$

On a alors

$$\hat{\beta}_1^{seule} = \langle x^1, y \rangle \frac{x^1}{\|x^1\|_2^2}$$

tandis que

$$\hat{\beta}_1^{complet} = \langle x^1, (y - X_2\hat{\boldsymbol{\beta}}_{\mathbf{2}}^{complet})\rangle \frac{x^1}{\|x^1\|_2^2}$$

On voit que le coefficient d'une variable donnée quand elle est utilisée comme seule variable explicative n'est pas le même que le coefficient de cette même variable quand elle est utilisée dans un modèle où elle est en présence d'autres variables explicatives !

— On peut remarquer que le seul cas où le vecteur $\hat{\beta}^1$ dans le modèle complet est le même que dans le modèle avec seulement les q variables explicatives est le cas orthogonal : $X_1^T X_2 = 0$.
— Par exemple si x^1 est orthogonal à toutes les autres variables dans le modèle complet alors $\hat{\beta}_1^{complet} = \hat{\beta}_1^{seule}$.
— Dans le cas contraire, le coefficient de x^j différera selon le modèle dans lequel il est calculé : il sera en général différent selon que l'inclut certaines variables explicatives ou pas, surtout selon que l'on inclut des variables corrélées à x^j ou pas.
— L'ordre des tests de pertinence n'a pas d'importance quand on a des variables orthogonales (cf aussi chapitre 5 par exemple).

Commentaires sur le premier item Examinons le cas particulier où $q = 1$. On isole donc la première variable explicative. Alors le premier item dit que, pour trouver le coefficient de cette variable explicative dans le modèle complet, on peut projeter cette variable sur l'orthogonal des autres variables, puis faire la régression sur cette seule variable :

$$z^1 = x^1 - P_{X_2} x^1$$

$$\hat{\beta}_1^{complet} = \langle \frac{z^1}{\|z^1\|_2^2}, y\rangle$$

On peut évidemment faire cela avec n'importe quelle variable explicative.
On peut donc traduire cela en les termes suivants : le coefficient de régression multiple $\hat{\beta}_j$ représente la contribution additionnelle de x^j sur y, après que x^j a été ajusté par rapport aux autres variables explicatives. Grâce à la formule ci-dessus, on trouve immédiatement

$$\mathbf{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\|z^j\|_2^2}$$

On voit donc que, si $\|z^j\|_2$ est très petite, alors la variance du coefficient $\hat{\beta}_j$ est très grande. Ce cas se produit si la variable x^j est très corrélée aux autres variables explicatives. .

Exercice D.1. *Reprenons le fichier ozone. Afficher les coefficients des régression de* `maxO3` *sur* `T9` *puis les coefficients de la régression de* `maxO3` *sur* `T12` *et* `T9`*. Commentaires.*

D.2 Colinéarité et tests statistiques

Nous allons illustrer les problèmes que pose la collinéarité des variables explicatives pour les tests de nullité d'un coefficient.
Nous simulons le modèle linéaire suivant

$$y = x^2 + x^3 + x^4 + \varepsilon$$

Nous supposons que nous ne connaissions pas le vrai modèle et que nous ayons à notre disposition les variables explicatives x^2, x^3, x^4 et x^5, plus l'intercept. Au départ nous ne savons donc pas que x^5 n'influence pas réellement y. Supposons que x^5 est très corrélée aux autres variables

$$x^5 \approx 2x^2 + x^3$$

```
x2=rnorm(100)
x3=rnorm(100)
x4=rnorm(100)
c=c(1,1,rep(0,98))
x5=2*x2+x3+c
eps=rnorm(100)
y=x2+x3+x4+eps
summary(lm(y~x2+x3+x4+x5))

Call:
lm(formula = y ~ x2 + x3 + x4 + x5)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49243 -0.67679  0.05223  0.76233  2.25481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01772    0.11376  -0.156    0.877
x2           1.29706    1.62490   0.798    0.427
x3           1.28589    0.80412   1.599    0.113
x4           0.87609    0.10718   8.174 1.29e-12 ***
x5          -0.17162    0.80303  -0.214    0.831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.104 on 95 degrees of freedom
Multiple R-squared:  0.7518,Adjusted R-squared:  0.7414
F-statistic: 71.95 on 4 and 95 DF,  p-value: < 2.2e-16
```

On voit dans le résultat des tests de Student sur la pertinence des différentes variables explicatives que, au niveau 5%,

- Les seule variable déclarée pertinente est la variable x^4 (p-valeur <1.29e-12).
- Les variables pertinentes x^3 et surtout x^2 ont de p-valeurs relativement grandes : on les éliminerait donc à tort. Elles ont été "cachées" par la variable x^5.

Regardons ce qui se passe maintenant quand on fait la régression de y sur les trois premières variables :

```
Call:
lm(formula = y ~ x2 + x3 + x4)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.49003 -0.67292  0.06052  0.76603  2.09584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02175    0.11162  -0.195    0.846
x2           0.95071    0.11754   8.088 1.85e-12 ***
x3           1.11572    0.11183   9.977  < 2e-16 ***
x4           0.87928    0.10560   8.326 5.77e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.098 on 96 degrees of freedom
Multiple R-squared:  0.7517,Adjusted R-squared:  0.7439
F-statistic: 96.88 on 3 and 96 DF,  p-value: < 2.2e-16
```

Sans la présence intempestive de la variable x^5, les trois variables x^2, x^3 et x^4 sont déclarées pertinentes.
Si on veut éliminer plusieurs variables en même temps, il vaut mieux faire le test de la section suivante.
Pourquoi ne vaut-il mieux pas utiliser le test de Student de pertinence d'une variables explicative pour supprimer plusieurs variables en même temps ? Pour des problèmes éventuels de colinéarité : en effet, reprenons l'exemple simulé précédemment. On a simulé un modèle avec 4 variables pertinentes et on a introduit dans le modèle une cinquième variable qui n'a rien à voir avec y mais qui est très corrélée avec les variables x^2 et x^3. Si on utilise bêtement les tests de Student pour éliminer toutes les variables déclarées non pertinentes en même temps, on prend le risque d'en éliminer certaines qui sont pourtant pertinentes. Si au contraire, on élimine une à une les variables, **en refaisant la régression à chaque fois**, on prend moins de risque : dans l'exemple, on a enlevé d'abord x^5 (c'est la variable dont la p-value du test de Student est la plus grande, à l'exception de l'intercept). Alors en refaisant un régression sans cette variable, les variables x^2 et x^3 "redeviennent" pertinentes.

Remarque D.2. *La présence de multicollinéarité possible doit inciter à la prudence dans l'interprétation des résultats mais n'est pas toujours dommageable pour la prédiction et peut aussi être utile dans certains cas, en particulier en cas de données manquantes.*

D.3 Détection de colinéarité

Pour savoir si deux variables explicatives sont corrélées entre elles, on peut calculer la corrélation entre ces variables, donnée par la fonction `cor` de R. Par exemple, si on veut voir la corrélation entre les variables `T9` et `T12` dans `ozone`, on peut utiliser

```
> attach(ozone)
> cor(T9,T12)
[1] 0.8829672
```

Donc on a bien une grande corrélation entre la température à 9h et la température à 12H.
Pour plus de deux variables explicatives, on peut encore calculer la corrélation entre tous les couples de variables. Ceci est aussi fait par la fonction `cor` à qui il faut donner une matrice, et pas un dataframe. Pour transformer un dataframe en matrice de numérique, utiliser la fonction `data.matrix`). Cette fonction calcule la matrice de corrélation des variables (on lui donne une matrice dont les colonnes sont les variables explicatives). On a donc accès à tous les ρ_{x^j,x^k}. Par exemple, pour le fichier `ozone`

```
> ozonemat=data.matrix(ozone)
>  cor(ozonemat)
             maxO3         T9        T12        T15        Ne9       Ne12       Ne15
maxO3   1.00000000  0.6993865  0.7842623  0.7745700 -0.6217042 -0.6407513 -0.4783021
T9      0.69938654  1.0000000  0.8829672  0.8464460 -0.4838636 -0.4722475 -0.3251386
T12     0.78426233  0.8829672  1.0000000  0.9461930 -0.5842709 -0.6601002 -0.4580991
T15     0.77456998  0.8464460  0.9461930  1.0000000 -0.5861683 -0.6492261 -0.5746817
Ne9    -0.62170423 -0.4838636 -0.5842709 -0.5861683  1.0000000  0.7883411  0.5502490
Ne12   -0.64075130 -0.4722475 -0.6601002 -0.6492261  0.7883411  1.0000000  0.7098668
Ne15   -0.47830212 -0.3251386 -0.4580991 -0.5746817  0.5502490  0.7098668  1.0000000
Vx9     0.52762341  0.2506896  0.4301045  0.4530892 -0.4976361 -0.4926581 -0.4014717
Vx12    0.43079585  0.2223857  0.3126291  0.3437506 -0.5287752 -0.5103204 -0.4318633
Vx15    0.39189889  0.1703215  0.2706802  0.2866028 -0.4939010 -0.4322695 -0.3782896
maxO3v  0.68451598  0.5822451  0.5636289  0.5678887 -0.2765500 -0.3619227 -0.3084755
vent    0.02943812  0.1433694  0.1248019  0.0947173  0.2070029  0.1885488  0.2616428
temps   0.47566681  0.3833287  0.4409282  0.4151132 -0.3913738 -0.4221216 -0.2948889
              Vx9       Vx12       Vx15     maxO3v        vent       temps
maxO3   0.5276234  0.4307959  0.3918989  0.6845160  0.02943812  0.47566681
T9      0.2506896  0.2223857  0.1703215  0.5822451  0.14336941  0.38332867
T12     0.4301045  0.3126291  0.2706802  0.5636289  0.12480191  0.44092818
T15     0.4530892  0.3437506  0.2866028  0.5678887  0.09471730  0.41511318
Ne9    -0.4976361 -0.5287752 -0.4939010 -0.2765500  0.20700290 -0.39137379
Ne12   -0.4926581 -0.5103204 -0.4322695 -0.3619227  0.18854885 -0.42212160
Ne15   -0.4014717 -0.4318633 -0.3782896 -0.3084755  0.26164280 -0.29488894
Vx9     1.0000000  0.7501775  0.6822608  0.3403172 -0.18934673  0.42493778
Vx12    0.7501775  1.0000000  0.8371720  0.2236755 -0.45302996  0.29785100
Vx15    0.6822608  0.8371720  1.0000000  0.1899220 -0.44568427  0.20591727
maxO3v  0.3403172  0.2236755  0.1899220  1.0000000  0.07991000  0.37536770
vent   -0.1893467 -0.4530300 -0.4456843  0.0799100  1.00000000 -0.05338682
temps   0.4249378  0.2978510  0.2059173  0.3753677 -0.05338682  1.00000000
```

Pour plus de lisibilité, on peut préférer utiliser la fonction `corrplot` de la bibliothèque du même nom (à installer et charger) qui donne ces corrélations sous forme de graphique couleur.

```
>  library(corrplot)
> mcor=cor(ozonemat)
> corrplot(mcor, type="upper", order="hclust", tl.col="black", tl.srt=45)
```

Les corrélations positives sont affichées en bleu et les corrélations négatives en rouge. L'intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation.

Cependant, ceci ne permet pas de détecter la multicollinéarité. En effet une variable x^j peut ne pas être très corrélée à une variable particulière x^k mais être corrélée à une combinaison linéaire des autres variables. Par exemple avec $p = 4$:

$$x^4 \approx 1 + \lambda_2 x^2 + \lambda_3 x^3$$

Pour détecter cela, il suffit en fait de faire une régression linéaire de x^4 sur x^2 et x^3 (i.e. le rôle de y est jouée par x^4). En effet, si x^4 est proche d'une combinaison linéaire des autres variables (intercept inclus) alors en effectuant la régression linéaire de x^4 sur x^2 et x^3, le R^2 qu'on obtient est grand.
Le calcul du R^2 associé à la régression de x^j sur toutes les autres variables explicatives, noté R_j^2, est alors calculé : plus R_j^2 est grand, plus la corrélation est forte.
Ce qui est utilisé en fait est une fonction simple de ce coefficient R_j^2, fonction appelée "variance inflation factors" (VIF) :

$$VIF_j = \frac{1}{1 - R_j^2}$$

Un VIF peut donc être calculé pour chaque variable j et est compris entre 1 et $+\infty$. Comme VIF_j est une fonction croissante de R_j^2, on a :

$$VIF \in [1, +\infty[\text{ et}$$

Plus VIF_j est grand et plus la variable j est corrélée aux autres variables

Dans la pratique, on considère qu'une valeur de VIF supérieure à 5 est considéré comme relativement grande et supérieure à 10 comme très grande. (évidemment le seuil dépend en fait du nombre total de variables explicatives, puisqu'il est lié à un R^2, cf les 2 sections suivantes).
Comme son nom l'indique, VIF_j est le facteur d'augmentation de la variance de $\hat{\beta}_j$: on peut en effet montrer que, si $\sigma_j^{complet}$ est la variance de $\hat{\beta}_j^{complet}$ et σ_j^{seul} la variance de $\hat{\beta}_j^{seul}$ alors $\sigma_j^{complet} = \sigma_j^{seul} \times VIF_j$.

Code

On peut utiliser la fonction `VIF` du package `faraway` (package qu'il faut installer, puis charger). Exemple avec ozone :

```
> library(faraway)
> vif(reg)
    Ne12     Vx12      T12
2.165102 1.354037 1.774706
```

On n'a donc pas de problème de collinéarité entre ces trois variables explicatives.

Exercice D.2. *On reprend le dernier exercice concernant le fichier* `ozone`*. On a fait la régression de* `maxO3` *sur* `T9`*,* `T12`*,* `Ne9`*,* `Ne12`*,* `Vx9` *et* `Vx12`*. Les résultats des tests nous ont fait soupçonner une corrélation des variables explicatives. Vérifier cette hypothèse.*