AgroParisTech

Exemples d'application du modèle linéaire

E. Lebarbier, S. Robin

Table des matières

Chapitre 1

Introduction

1.1 Avertissement

Ce polycopié présente des exemples d'applications du modèle linéaire. Les différents chapitres reprennent les modèles les plus classiques appliqués à des données réelles :

Chapitre 2 : Régression linéaire simple,

Chapitre 3 : Régression linéaire multiple,

Chapitre 4 : Analyse de la variance à un facteur,

Chapitre 5 : Analyse de la variance à deux facteurs équilibrés,

Chapitre 6 : Analyse de la variance à deux facteurs déséquilibrés,

Chapitre 7 : Analyse de la covariance,

Ce polycopié *ne présente pas un cours de statistique sur le modèle linéaire* : il s'appuie sur des notions abordées dans le cours de statistiques de 1ère année de l'INA PG et complète la présentation théorique du modèle linéaire faite dans le cours de 2ème année. Au long de ce document, le lecteur sera fréquemment renvoyé

- au livre de 1ère année de Daudin *et al.* (1999)
- et au polycopié de 2ème année de Duby (2000).

Le dit lecteur est donc fermement invité par les auteurs à lire (relire ?) ces deux ouvrages avec attention.

1.2 Notations

Typographie

- Les lettres Majuscule designent des variables aléatoires (Y).
- Les caractères gras désignent des vecteurs ($\boldsymbol{\theta}$) ou des matrices ($\mathbf{X}$).
- Les lettres grecques représentent des paramètres d'un modèle ($\mu, \sigma, \boldsymbol{\theta}$).
  Certains paramètres sont cependant représentés par des caractères latins (comme la constante a et la pente b en régression) conformément à l'usage.

De plus, on notera

0 le vecteur nul dimensions $n \times 1$ (sauf précision),

I la matrice identité de dimensions $n \times n$ (sauf précision).

Sommes

Le symbole + à l'indice signifie que la variable est sommée sur l'indice qu'il remplace. Ainsi, dans une modèle d'analyse de la variance à deux facteurs (cf chapitre 5), si n_{ij} désigne le nombre de répétitions dans le niveau i du premier facteur et le niveau j du second facteur,

$$n_{i+} = \sum_{j=1}^{J} n_{ij}$$

désigne le nombre total d'observations dans le niveau i du premier facteur ;

$$n_{+j} = \sum_{i=1}^{I} n_{ij}$$

désigne le nombre total d'observations dans le niveau j du second facteur ;

$$n = n_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

désigne le nombre total d'observations dans l'ensemble de l'expérience.

Moyennes

Le symbole • à l'indice signifie que la variable est moyennée sur l'indice qu'il remplace. Ainsi, en reprenant l'exemple de l'analyse de la variance à deux facteurs,

$$Y_{ij\bullet} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

représente la réponse moyenne des observations recueillies avec le niveau i du premier facteur et le niveau j du second ;

$$Y_{\bullet j\bullet} = \frac{1}{n_{+j}} \sum_{i=1}^{I} \sum_{k=1}^{K} Y_{ijk}$$

représente la réponse moyenne des observations recueillies avec le niveau j du second facteur ;

$$Y_{\bullet\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk}$$

représente la moyenne générale de l'ensemble des observations.

Lois de probabilité

$\mathcal{N}(\mu, \sigma^2)$: désigne la loi normal univariée d'espérance μ et de variance σ^2.

$\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: désigne la loi normal multivariée de dimension n, de vecteur d'espérance $\boldsymbol{\mu}$ et de matrice de variance $\boldsymbol{\Sigma}$.

χ^2_m : désigne la loi du khi deux (centré) à m degrés de libertés.

$\mathcal{T}_m$: désigne la loi de student (centrée) à m degrés de libertés.

$\mathcal{F}_{m_1,m_2}$: désigne la loi de Fisher (centrée) à m_1 et m_2 degrés de libertés.

Chapitre 2

Régression linéaire simple

2.1 Présentation

2.1.1 Objectif et dispositif expérimental

On cherche à décrire la relation entre le Taux de DDT d'un brochet (variable à expliquer y) et l'âge du brochet (variable explicative x).

Données

On dispose d'un échantillon de $n = 15$ brochets. Pour chaque brochet, on a
- son âge (variable `Age`),
- la mesure de son Taux de DDT (variable `TxDDT`).

Les données sont présentées dans la table 2.1. On remarque que l'on dispose de 3 mesures de Taux de DDT par âge. Ces mesures peuvent être vues comme des données répétées par âge. On verra dans le paragraphe 2.2.4 à quoi peut servir cette information.

2.1.2 Description des données

La figure 2.1 représente le Taux de DDT en fonction de l'âge des 15 brochets. On peut observer que

1. plus le brochet est agé, plus le Taux de DDT est élevé,
2. plus le brochet est agé, plus le Taux de DDT est variable par âge.

Ce graphique permet d'émettre des hypothèses quant à la relation qui peut exister entre le Taux de DDT et l'âge des brochets mais ne permet en aucun cas de conclure d'une part sur le fait qu'elle existe et d'autre part sur le type de cette relation. Il faut que ces hypothèses soient statistiquement testées, et c'est donc à partir des résultats des tests que l'on pourra conclure. Ici il semble qu'il existe une relation entre le Taux de DDT et l'âge mais que cette relation n'est pas "linéaire".

La table 2.2 donne les statistiques élémentaires sur l'échantillon entier (la moyenne, l'écart-type, la valeur maximale et la valeur minimale du Taux de DDT) puis par âge.

Obs	Age	TxDDT
1	2	0.20
2	2	0.25
3	2	0.18
4	3	0.19
5	3	0.29
6	3	0.28
7	4	0.31
8	4	0.33
9	4	0.36
10	5	0.71
11	5	0.38
12	5	0.47
13	6	1.10
14	6	0.87
15	6	0.83

TABLE 2.1 – Table des données.

Quelques remarques.

Moyennes. On observe que les Taux de DDT moyens par âge sont très différents. Ils augmentent avec l'âge mais pas "linéairement".

Variances. La variabilité du Taux de DDT par âge est de plus en plus forte avec l'âge. En effet, l'écart-type (donné par "`Std Dev`") est de 0.036, 0.055, 0.025 pour les trois premiers âges et 0.17, 0.13 pour les deux derniers, ce qui confirme la remarque 2 précédente. On remarque que cette variabilité est nettement atténuée quand on considère le log du Taux de DDT puisqu'à part pour l'âge 5, les écarts-types sont à peu près les mêmes.

2.2 Régression linéaire simple

2.2.1 Modèle

On cherche à modéliser la relation entre la variable `TxDDT` et la variable `Age`. Le modèle le plus simple est la régression linéaire simple qui s'écrit :

$$Y_i = a + bx_i + E_i, \qquad \{E_i\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n = 15. \tag{2.1}$$

où

- l'indice i représente le numéro du brochet,
- la variable Y_i désigne le Taux de DDT du i-ème brochet,
- la variable x_i est l'âge du i-ème brochet,

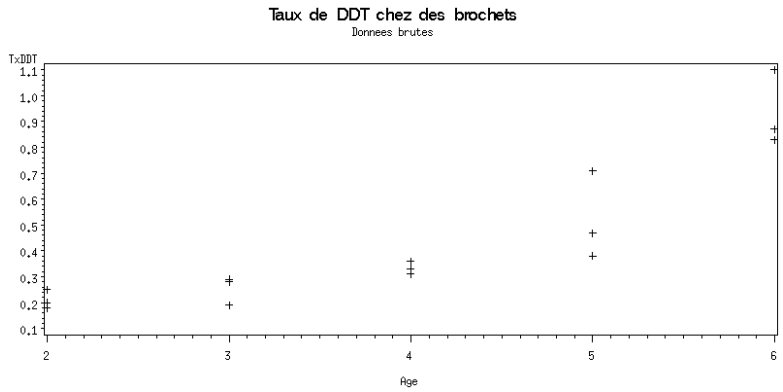

FIGURE 2.1 – Taux de DDT en fonction de l'âge des 15 brochets.

– la variable E_i est un terme résiduel aléatoire,
– σ^2 est la variance résiduelle,
– a et b sont des paramètres inconnus :

$$\begin{cases} b & \text{représente la pente de la droite : l'accroissement de l'espérance de } Y \ (\mathbb{E}[Y]) \\ & \text{entrainé par l'accroissement d'une unité de } x, \\ a & \text{est la valeur de l'ordonnée à l'origine.} \end{cases}$$

Rappelons que i.i.d. signifie indépendant et identiquement distribués. On suppose donc ici, comme dans tous les modèles linéaires que nous étudierons, que les variables aléatoires E_i sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

Le modèle (2.1) signifie que l'on décompose Y en

$$\begin{cases} a + bx_i & \text{partie expliquée par } x \text{ (partie fixe),} \\ E_i & \text{partie inexpliquée (partie aléatoire).} \end{cases}$$

Ecriture en terme de loi des Y_i.

Le modèle (2.1) est équivalent au modèle

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \qquad \{Y_i\} \text{ indépendants} \tag{2.2}$$

en notant

$$\mu_i = a + bx_i.$$

Ecriture matricielle.

Le modèle (2.1) peut s'écrire sous la forme matricielle générale à tous les modèles linéaires :

$$\mathbf{Y} = \mathbf{X\Theta} + \mathbf{E}, \qquad \mathbf{E} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}). \tag{2.3}$$

Statistiques elementaires

Variable	N	Mean	Std Dev	Minimum	Maximum
TxDDT	15	0.4500000	0.2873276	0.1800000	1.1000000
LogDDT	15	-0.4192093	0.2521350	-0.7447275	0.0413927

Statistiques elementaires par Age

Age=2

Variable	N	Mean	Std Dev	Minimum	Maximum
TxDDT	3	0.2100000	0.0360555	0.1800000	0.2500000
LogDDT	3	-0.6819192	0.0728461	-0.7447275	-0.6020600

Age=3

Variable	N	Mean	Std Dev	Minimum	Maximum
TxDDT	3	0.2533333	0.0550757	0.1900000	0.2900000
LogDDT	3	-0.6038968	0.1019130	-0.7212464	-0.5376020

Age=4

Variable	N	Mean	Std Dev	Minimum	Maximum
TxDDT	3	0.3333333	0.0251661	0.3100000	0.3600000
LogDDT	3	-0.4779406	0.0326153	-0.5086383	-0.4436975

Age=5

Variable	N	Mean	Std Dev	Minimum	Maximum
TxDDT	3	0.5200000	0.1705872	0.3800000	0.7100000
LogDDT	3	-0.2989534	0.1380332	-0.4202164	-0.1487417

Age=6

Variable	N	Mean	Std Dev	Minimum	Maximum
TxDDT	3	0.9333333	0.1457166	0.8300000	1.1000000
LogDDT	3	-0.0333367	0.0655196	-0.0809219	0.0413927

TABLE 2.2 – Statistiques élémentaire sur l'échantillon entier et par âge.

où $\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\underset{n\times 1}{\mathbf{E}} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix}$, $\underset{n\times 2}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\underset{2\times 1}{\Theta} = \begin{bmatrix} a \\ b \end{bmatrix}$.

Notion de linéarité pour un modèle

Quand on parle de modèle linéaire, cela signifie linéaire par rapport aux paramètres. Une relation entre x et y qui n'est pas linéaire ne dit pas que le modèle proposé n'est pas linéaire. Voici quelques exemples :

- *Exemple* 1 *:* $Y_i = a + bx_i + cx_i^2 + E_i$ est un modèle linéaire mais la relation entre x et y n'est pas linéaire mais de type polynomial.
- *Exemple* 2 *:* $Y_i = a + b\cos(x_i) + E_i$ est un modèle linéaire.
- *Exemple* 3 *:* $Y_i = ae^{bx_i} + E_i$ n'est pas un modèle linéaire. En effet, $ae^{b_1+b_2x} \neq ae^{b_1} + ae^{b_2x}$.
- *Exemple* 4 *:* $Y_i = \frac{a}{b+x_i} + E_i$ n'est pas un modèle linéaire.

2.2.2 Validation du modèle

Le modèle de régression linéaire simple (2.1) suppose que la régression est linéaire, les termes d'erreurs ont même variance, qu'ils sont indépendants et enfin qu'ils sont issues d'une loi gaussienne. Avant d'effectuer l'analyse de la régression, il est **indispensable** de vérifier ces hypothèses. Cette étude se généralise à tous les modèles que nous verrons dans la suite sauf l'hypothèse de la linéarité qui est spécifique au modèle de régression linéaire simple.
L'examen de la validité des hypothèses du modèle se fait à partir du graphe des résidus $\hat{E}_i$ en fonction des prédictions $\hat{Y}_i$. Rappelons que la prédiction $\hat{Y}_i$ est

$$\hat{Y}_i = \hat{a} + \hat{b}x_i, \tag{2.4}$$

où $\hat{a}$ et $\hat{b}$ sont respectivement les estimations des paramètres a et b (cf Daudin *et al.* (1999)), et les résidus estimés sont définis par

$$\hat{E}_i = Y_i - \hat{Y}_i.$$

D'une **façon générale**, une structure "particulière" du nuage de points du graphe des résidus indique que le modèle proposé n'est pas adapté aux données.

Linéarité de la relation

On ne vérifie cette hypothèse que dans le cas d'un modèle de régression linéaire simple. On peut juger de la linéarité de la relation entre x et Y en visualisant le graphique des couples (x_i, y_i). Par exemple, prenons la figure 2.2, elle montre clairement que la relation entre x et Y n'est pas linéaire. Le graphique des résidus associés est représenté figure 2.3 et montre bien ce phénomène. On voit qu'il existe une tendance quadratique. Cela sous-entend qu'une partie de la moyenne se trouve dans les résidus et il faut l'enlever.
Remarque : on peut aussi faire cette étude pour des modèles simples permettant une étude visuelle comme par exemple un modèle du type $Y_i = a + bx_i + cx_i^2 + E_i$. On ne parle plus dans ce cas de relation linéaire.

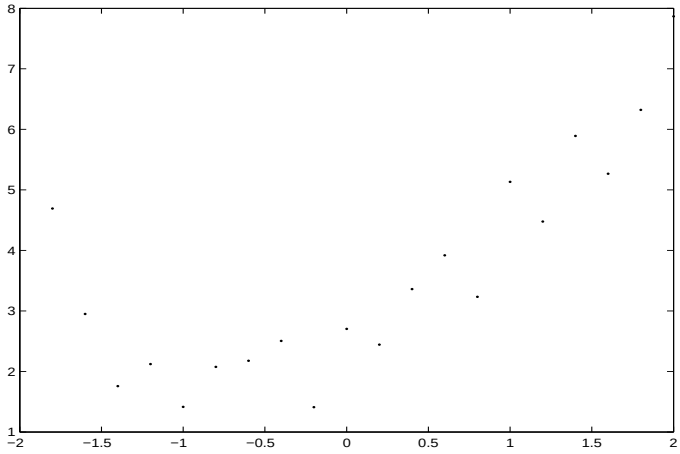

FIGURE 2.2 – Graphe (x_i, y_i).

Homoscédasticité/hétéroscédasticité

$$\begin{cases} \text{homoscédasticité :} & \text{la variance des résidus } \sigma^2 \text{ est constante.} \\ \text{hétéroscédasticité :} & \text{la variance des résidus n'est pas constante.} \end{cases}$$

L'hypothèse faite sur le modèle est celle de l'homoscédasticité. Sur la figure 2.4, les résidus estimés augmentent en valeur absolue avec $\hat{Y}$; ce qui signifie que la variance des erreurs augmente avec $\hat{Y}$. Si au contraire les résidus avaient tendance à se rapprocher de 0, ceci signifierait que la variance diminue avec $\hat{Y}$. Mais dans les deux cas, l'hypothèse d'homoscédasticité n'est pas vérifiée.

Si l'hypothèse d'homoscédasticité n'est pas vérifiée, on peut effectuer une transformation pour stabiliser la variance. Les deux transformations les plus courantes sont :

$$\begin{cases} \log Y & \text{si } \sigma \text{ est proportionnel à } \mathbb{E}[Y], \\ \sqrt{Y} & \text{si } \sigma^2 \text{ est proportionnel à } \mathbb{E}[Y]. \end{cases}$$

Indépendance

Il est difficile de vérifier cette hypothèse à partir du graphe des résidus. Elle se déduit du plan d'expérience mené : on dira que cette hypothèse est vérifiée si chaque observation correspond à une expérience menée dans des conditions indépendantes. Par exemple, si un individu est utilisé deux fois dans l'expérience, on perd l'hypothèse d'indépendance.

Normalité

C'est l'hypothèse la moins importante car d'une part le modèle linéaire est robuste à la normalité et d'autre part les résidus suivent asymptotiquement une loi normale (i.e. pour des grands échantillons). Il existe néanmoins des tests de normalité, tels que Kolmogorov-Smirnov.

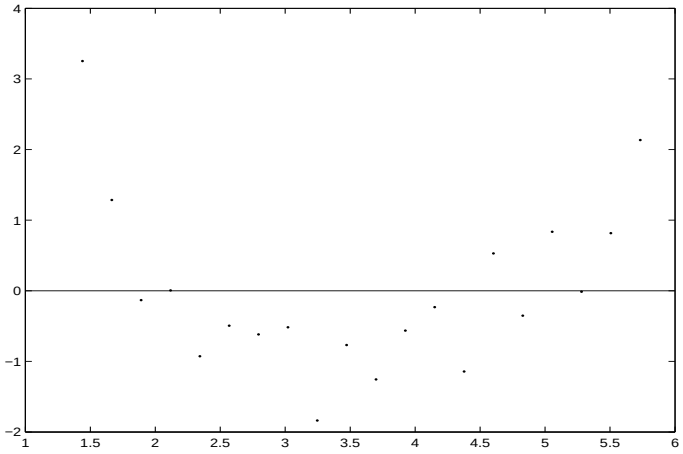

FIGURE 2.3 – Graphe des résidus associés : $(\hat{y}_i, \hat{e}_i)$.

Points aberrants

Il se peut que certaines observations soient suspectes. Par exemple sur le graphe des données figure 2.5, on peut suspecter le premier point. Dans ce cas, on peut s'interroger sur la validité de cette mesure, et il peut être conseillé de refaire la régression sans ce point. Mais attention la suppression d'un point atypique doit toujours être éventuellement justifiée par des considérations non-statistiques.

Bilan : les étapes de l'analyse
- on estime $\mathbb{E}[Y_i]$ soit $\hat{Y}_i$ et E_i soit $\hat{E}_i$,
- on regarde le graphe des résidus,
- si le graphe des résidus valide les hypothèses, on peut continuer l'analyse mais si ce n'est pas le cas, on cherche un nouveau modèle.

2.2.3 Retour chez les brochets : analyse des résidus

La figure 2.6 représente le graphe des résidus sur l'exemple des brochets pour le modèle (2.1). On observe que

1. l'hypothèse d'homoscédasticité n'est pas vérifiée.
2. il y a une tendance, ce qui signifie qu'un terme pertinent a pu être oublié dans la modélisation de $\mathbb{E}[Y]$.

L'hypothèse d'homoscédasticité n'étant pas vérifiée, on se propose de transformer les données en passant au log. Le modèle devient :

$$\log\left(\texttt{TxDDT}_i\right) = a + b\ \texttt{Age}_i + F_i,$$

où $\{F_i\}_i$, comme $\{E_i\}_i$ pour le modèle (2.1), sont des variables aléatoires supposées indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$. Le nouveau graphe des résidus est donné figure

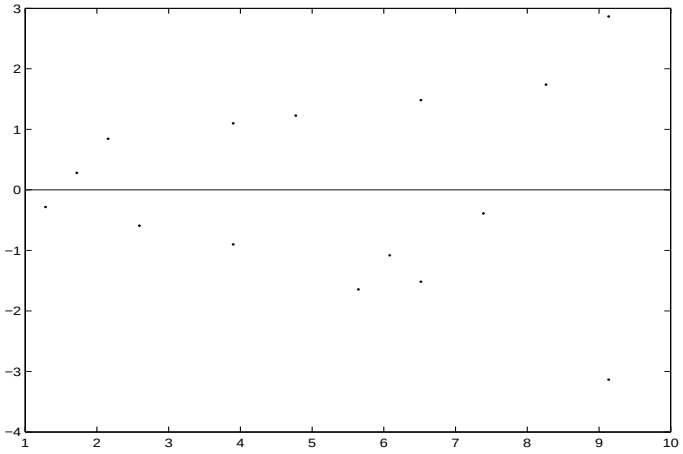

FIGURE 2.4 – Graphe de résidus : $(\hat{y}_i, \hat{e}_i)$.

2.7. Le log a bien joué son rôle puisque la variance est stabilisée. Cependant, on peut remarquer qu'il subsiste une tendance qui semble quadratique. Ainsi pour mieux traduire la liaison entre `Age` et `TxDDT`, on pourrait introduire un terme supplémentaire au modèle précédent dans la modélisation de l'espréance en $\texttt{Age}^2$. De cette façon, nous proposerions un ajustement polynomial plutôt que linéaire et le modèle deviendrait :

$$\log(\texttt{TxDDT}_i) = a + b\ \texttt{Age}_i + c\ \texttt{Age}_i^2 + G_i.$$

Le graphe des résidus donné figure 2.8 montre que la tendance a disparu. Rappelons que ce modèle est bien un modèle linéaire mais c'est la relation entre l'âge et le Taux de DDT qui ne l'est pas (dans ce cas, elle est polynomiale). Comme ici nous n'étudions que le modèle de régression linéaire simple et que le modèle précédent n'entre pas dans ce cadre mais plutôt dans le cadre de la régression linéaire multiple, qui fait l'objet du prochain chapitre, nous ne nous étendrons pas dans ce chapitre sur ce modèle.

Pour les modèles qui présentent des graphes de résidus corrects, il faut privilégier le modèle le plus simple et surtout qui permet une interprétation facile des résultats.

2.2.4 Table d'analyse de la variance

Dans cette section, nous décrivons les résultats donnés par SAS pour le modèle

$$\log(\texttt{TxDDT}_i) = a + b\ \texttt{Age}_i + F_i,$$

et en donnons leurs interprétations. Pour plus de simplicité et de généralité, nous noterons

$$Y = \log(\texttt{TxDDT}), \quad x = \texttt{Age} \ \text{ et } \ \texttt{E} = \texttt{F}.$$

Un des principaux objectifs est de déterminer si la relation entre la variable Y et la variable x est linéaire.

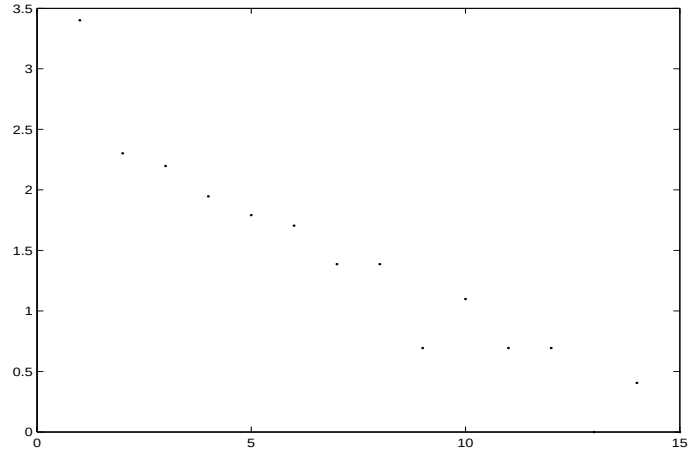

FIGURE 2.5 – Graphe (x_i, y_i).

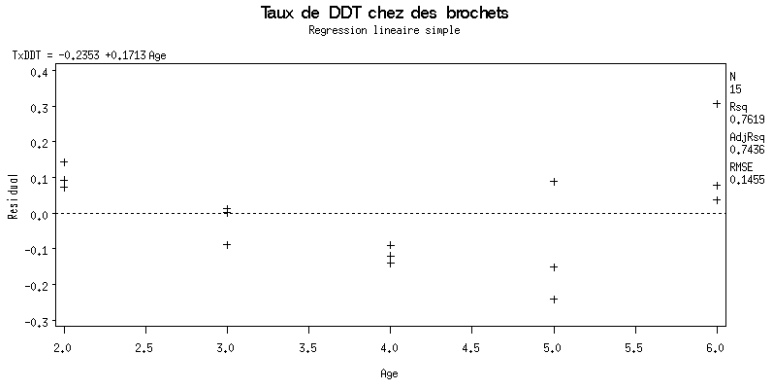

FIGURE 2.6 – Graphe des résidus pour le modèle $\texttt{TxDDT}_i = a + b\ \texttt{Age}_i + E_i$.

Test de la signification de la régression

Pour généraliser les différentes définitions que nous allons donner ici, nous nous plaçons dans un modèle de régression multiple qui fait l'objet de l'étude du chapitre suivant. En régression multiple, on cherche à expliquer Y à partir de p variables explicatives, $x_1, \ldots, x_p$. Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + E_i \quad i = 1, \ldots, n,$$

Notons que pour la régression linéaire simple, $p = 1$.

La première table d'analyse de la variance donne les résultats du test des hypothèses :

$$H_0 = \{Y_i = \beta_0 + E_i\} \quad \text{contre} \quad H_1 = \{Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + E_i\}, \quad (2.5)$$

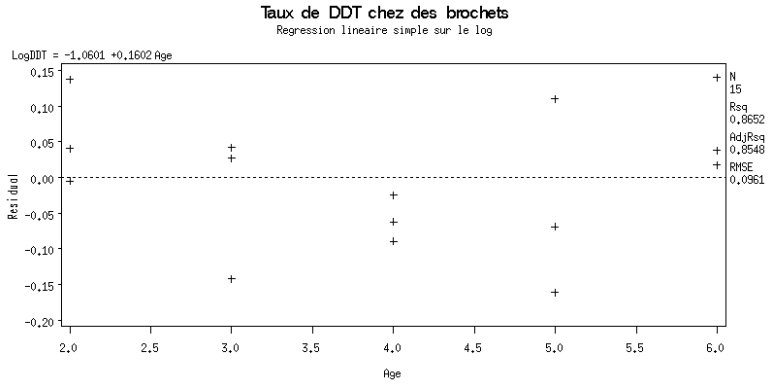

FIGURE 2.7 – Graphe des résidus pour le modèle $\log(\texttt{TxDDT}_i) = a + b\ \texttt{Age}_i + F_i$.

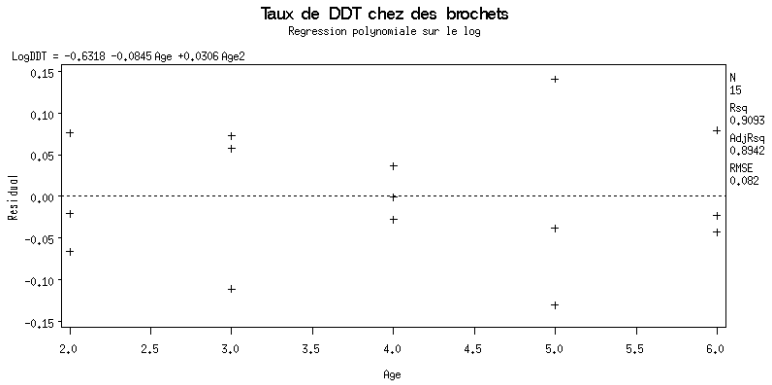

FIGURE 2.8 – Graphe des résidus pour le modèle $\log(\texttt{TxDDT}_i) = a + b\ \texttt{Age}_i + c\texttt{Age}_i^2 + G_i$.

ainsi que d'autres informations que nous allons préciser. En régression linéaire simple, ce test s'écrit

$$H_0 = \{Y_i = a + E_i\} \quad \text{contre} \quad H_1 = \{Y_i = a + bx_i + E_i\}, \tag{2.6}$$

qui s'exprime littérallement par $H_0 = \{$*il n'existe pas de liaison entre* Y *et* $x\}$ et $H_1 = \{$*il existe une liaison entre ces deux variables qui est linéaire*$\}$.

La table 2.3 rappelle, dans le cadre de la régression linéaire multiple, les définitions des éléments de la table 2.4 obtenues sur l'exemple des brochets dans le cadre de la régression linéaire simple.

où les différentes sommes des carrés sont définies par

- $SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$=Somme des Carrés Totale,
- $SCM = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$= Somme des Carrés due au Modèle,
- $SCR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$=Somme des Carrés des Erreurs,

avec $\bar{Y}$ définie par (2.8). Il est facile de voir que SCT se décompose en la somme des deux

Source	Degrés de liberté	Somme de carrés	Carré moyen	Statistique de test	Probabilité critique
Modèle	p	SCM	SCM/p	$F = \frac{SCM/p}{SCR/n-p-1}$	$P(\mathcal{F}_{p,n-p-1} > F)$
Résidu	$n-p-1$	SCR	$SCR/(n-p-1)$		
Total	$n-1$	SCT	$SCT/(n-1)$		

TABLE 2.3 – Définition des quantités de la table d'analyse de la variance dans le cas d'une régression linéaire multiple ($p = 1$ pour une régression simple).

autres termes :

$$SCT = SCM + SCR.$$

Cela signifie que la variabilité totale de Y est décomposée en variabilité due au modèle et en variabilité résiduelle. Dans le cadre de la régression simple, cela se traduit par une décomposition de la variabilité de Y sans tenir compte de x en la variabilité de Y expliquée par x et la variablité de Y autour de la droite de régression.

Source	DF	Sum of Squares	Mean Square	F Value	Pr> F
Model	1	0.77003	0.77003	83.43	<.0001
Error	13	0.11998	0.00923		
Corrected Total	14	0.89001			

TABLE 2.4 – Table d'analyse de la variance testant la linéarité de la relation entre le Taux de DDT et l'âge des brochets.

Statistique de test et probabilité critique. Le test des hypothèses (2.5) (ou (2.6)) se fait au moyen de la statistique de test de Fischer F qui vaut :

$$F = \frac{SCM/p}{SCR/n-p-1},$$

et qui s'interprète comme un rapport de variance :

$$F = \frac{\text{variance expliquée par } x}{\text{variance résiduelle}}.$$

La statistique de test F suit sous l'hypothèse H_0 une loi de Fischer à p et $n-p-1$ degrés de libertés $\mathcal{F}_{p,n-p-1}$. On utilise cette propriété pour calculer la probabilité critique $Pr > F$ qui est définie par :

$$Pr(\mathcal{F}_{p,n-p-1} > F).$$

C'est une mesure de l'accord entre l'hypothèse testée et le résultat obtenu. Plus elle est proche de 0, plus forte est la contradiction entre H_0 et le résultat de l'échantillon.

Règle de décision du test. La règle de décision du test est la suivante : on rejette l'hypothèse H_0 au niveau α si

$$\begin{cases} & F > f_{p,n-p-1,1-\alpha} \\ \text{ou} & P(\mathcal{F}_{p,n-p-1} > F) < \alpha \end{cases}$$

où $f_{p,n-p-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de $\mathcal{F}_{p,n-p-1}$. L'exemple donné figure 2.9, qui représente une loi Fischer à 6 et 10 degrés de libertés, illustre bien cette équivalence.

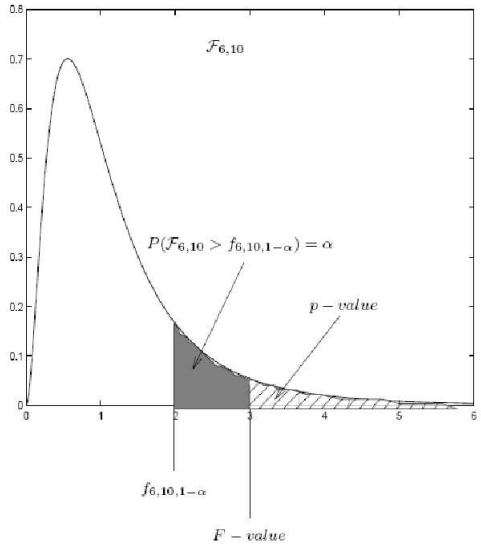

FIGURE 2.9 – Loi de Fisher à 6 et 10 degrés de libertés.

Conclusion. Sur l'exemple des brochets, la probabilité critique est inférieure à 0.05 ($(Pr > F) < 0.0001$), on rejette l'hypothèse H_0 : la part de variabilité expliquée par le modèle est significative, le modèle complet est conservé. La table 2.4 donne une valeur de $F = 83.43$, ce qui signifie que la variabilité due au modèle est largement supérieure à la variabilité résiduelle. On conclut de ce test qu'il existe une relation linéaire entre l'âge et le log du Taux de DDT.

Notion de réduction. La somme des carrés due au modèle (SCM) peut s'exprimer en termes de sommes de carrés résiduelles. Notons M_0 (resp. M_1) le modèle de l'hypothèse H_0 (resp. H_1). Comme on l'a vu dans la table 2.3, la somme des carrés résiduelle associée au modèle M_1 est

$$SCR_1 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

où $\hat{Y}_i$ est rappelé par (2.4), et la somme des carrés résiduelle associée au modèle M_0 vaut

$$SCR_0 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

puisque $\bar{Y}$ correspond à l'estimation de $\mathbb{E}[Y]$ dans le modèle M_0. Alors SCM s'écrit

$$SCM = SCR_0 - SCR_1.$$

Par cette décomposition, on voit bien que SCM quantifie la réduction d'erreurs lorsqu'on passe du modèle M_0 au modèle M_1. C'est pourquoi, on utilise souvent la notion de réduction :

$$SCM = R(b/a). \tag{2.7}$$

Cela se lit :

$$R(b/a) \text{ : } \textit{réduction due à } b \textit{ ajusté à } a.$$

On retrouve cette notation dans Daudin *et al.* (2007). Ici cette quantité correspond à la réduction due au modèle mais elle est aussi souvent utilisée, comme on le verra dans les chapitres suivants, pour écrire le test de modèles emboîtés (on teste un modèle restreint d'un modèle initial contre ce modèle initial).
Intuitivement, on sent bien que la diminution du nombre de variables explicatives contribue à augmenter les résidus. Il est clair que $SCR_0 \geq SCR_1$. Donc si la différence SCM est très grande, la réduction est importante et le modèle M_1 contribue à expliquer Y. Par le test, on obtient la significativité de cette réduction et donc de la contribution.

2.2.5 Ajustement du modèle

Il est important de vérifier si le modèle est bien ajusté aux données. En effet, l'un des objectifs est de pouvoir prédire la valeur de Y connaissant une valeur de la variable x. Or si l'ajustement est mauvais, on ne peut espérer obtenir une bonne prédiction.

Root MSE	0.09607	R-Square	0.8652
Dependent Mean	-0.41921	Adj R-Sq	0.8548
Coeff Var	-22.91699		

TABLE 2.5 – Résultats de l'ajustement du modèle sur l'exemple des brochets.

La table 2.5 présente les valeurs calculées pour les données des brochets des quantités données dans la table 2.6.

On a

– $\hat{\sigma}$ est l'écart-type résiduel estimé (on peut aussi lire la variance résiduelle estimée $\hat{\sigma}^2$, définie par (2.9), dans la table 2.4 à la ligne `Error` et la colonne `Mean Square`.

Ecart-type résiduel estimé	$\hat{\sigma}$	R^2	$\frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}$
Moyenne générale	$\bar{y}$	R^2ajusté	$1 - \frac{SCR/n-p-1}{SCT/n-1}$
Coefficient de variation (en %)	$100\hat{\sigma}/\bar{y}$		

TABLE 2.6 – Définition des quantités donnant l'ajustement du modèle.

– $\bar{y}$ est la moyenne générale des observations :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2.8}$$

– $\hat{\sigma}/\bar{y}$ est le coefficient de variation. C'est une mesure de la dispersion relative correspondant au rapport de l'écart-type à la moyenne.
– le $R^2 = R-square$ s'appelle le coefficient de détermination. Notons que $0 \leq R^2 \leq 1$. Il permet d'évaluer la qualité de l'ajustement. Plus cette valeur sera proche de 1 meilleur sera l'ajustement. Il s'interprète comme la proportion de variabilité de Y expliquée par le modèle et s'exprime en pourcentage.

– le R^2 ajusté est l'ajustement du R^2 au nombre p de variables explicatives. En effet, il est clair que plus on ajoute de variables explicatives, meilleur sera l'ajustement du modèle aux données (R^2 augmente avec le nombre de variables explicatives). Mais en considérant trop de variables, on risque d'obtenir un modèle qui est sur-ajusté et ce n'est pas ce que l'on cherche.

Conclusion. La table 2.5 donne la valeur $R^2 = 0.865$ (ici $p-1=1$), ce qui signifie que 86.5% de la variabilité du log du Taux de DDT est expliquée par l'âge. L'ajustement du modèle aux données est donc bon.

2.2.6 Estimation des paramètres et tests sur les paramètres

Estimation des paramètres

Paramètres de l'espérance. Les estimations des paramètres de l'espérance du modèle (2.1), soient a et b, notées $\hat{a}$ et $\hat{b}$ (cf table 2.8) sont données dans la table 2.7. Ils sont obtenus par la méthode des Moindres Carrés (cf Daudin *et al.* (2007) pour leurs définitions).

Paramètre de la variance. L'estimateur de la variance résiduelle σ^2 est :

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{2.9}$$

D'après, la définition de SCR, elle s'écrit $\frac{SCR}{n-p}$. Elle suit une loi du χ^2 à $n-p$ degrés de liberté (cf Daudin *et al.* (2007)).

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.06005	0.07442	-14.25	<.0001
Age	1	0.16021	0.01754	9.13	<.0001

TABLE 2.7 – Estimation des paramètres et tests sur les paramètres sur l'exemple des brochets.

paramètres	Degré de liberté	Paramètres estimés	Ecart Type	Statistique de test	Probabilité critique
constante	1	$\hat{a}$	$\hat{\sigma}_a$	$T_a = \frac{\hat{a}}{\hat{\sigma}_a}$	$P(\lvert\mathcal{T}_{n-2}\rvert > \lvert T_a\rvert)$
Age	1	$\hat{b}$	$\hat{\sigma}_b$	$T_b = \frac{\hat{b}}{\hat{\sigma}_b}$	$P(\lvert\mathcal{T}_{n-2}\rvert > \lvert T_b\rvert)$

TABLE 2.8 – Estimation des paramètres et tests sur les paramètres.

Discussion sur les estimations des paramètres. `Age`. Le paramètre b est estimé à $\hat{b} = 0.16 > 0$. Cette estimation est cohérente avec le graphique des données puisque quand le brochet grandit, le log du Taux de DDT croit, et donc b doit être positif.

Equation de la droite de régression. Les estimations des deux paramètres nous fournissent l'équation de la droite de régression estimée :

$$y = \hat{a} + \hat{b}x = 0.16x - 1.06.$$

Ceci signifie qu'en un an le log du Taux de DDT augmente de 0.16. Cette droite est représentée figure 2.10.

Estimation de la droite de régression par intervalle de niveau $1-\alpha = 0.95$. On donne souvent une idée de la précision de l'estimation à l'aide d'intervalle de confiance (cf Daudin *et al.* (1999)). Soit x_0 une valeur de x fixée, l'intervalle de confiance pour $\mathbb{E}[Y_{x_0}] = a + b\ x_0$ de niveau $1-\alpha$ est

$$\left[\hat{Y}_{x_0} - t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}, \hat{Y}_{x_0} + t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right],$$

où $t_{n-2,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une Student à $n-2$ degrés de liberté. Cet intervalle a une probabilité de 0.95 de contenir la vraie valeur de $a+bx_0$, qui rappelons-le est inconnue et que nous ne connaitrons jamais. En faisant varier x_0, les intervalles de confiance définissent deux hyperboles qui sont l'intervalle de confiance de la droite de

régression. Elle est représentée figure 2.10. Plus on s'éloigne du point moyen $(\bar{x}, \bar{y})$, moins l'estimation sera précise.

Prédiction d'une valeur de Y pour un âge x_0.
`Estimation ponctuelle de la prédiction du Taux de DDT pour un brochet âgé de 7 ans.` L'équation de la droite de régression va nous servir à faire de la prédiction. Si l'on cherche à prédire le Taux de DDT pour un brochet agé de $x_0 = 7$ ans par exemple, il suffit de calculer :

$$\log\left(\widehat{\texttt{TxDDT}}_{7ans}\right) = 0.16 * 7 - 1.06,$$

i.e.

$$\widehat{\texttt{TxDDT}}_{7ans} = \exp\left(0.16 * 7 - 1.06\right) = 1.0633.$$

`Estimation par intervalle de prédiction.` On veut estimer la valeur que l'on peut observer pour Y sachant que x est égal à x_0, i.e. $a + b\, x_0 + e_{x_0}$. Son intervalle de confiance est

$$\left[\hat{Y}_{x_0} - t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{Y}_{x_0} + t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

De la même façon que dans le paragraphe 2.2.6, en faisant varier x_0, on obtient l'intervalle de prédiction (le plus extérieur sur la figure 2.10). Plus cet intervalle est petit, meilleure sera la prédiction. Si on écrit l'intervalle précédent comme $[\hat{y}_{x_0} - e, \hat{y}_{x_0} + e]$, $1/e$ correspond à la précision absolue et plus e est grand, plus on sera précis. Donc il faut que la variance estimée $\hat{\sigma}^2$ soit la plus petite possible.

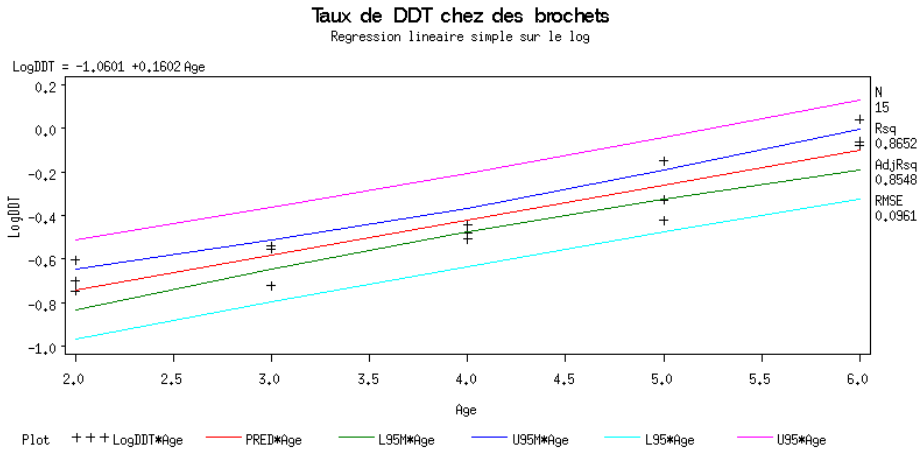

FIGURE 2.10 – Intervalles de confiance et de prédiction, et droite de régression.

Tests sur les paramètres

Pour chaque paramètre, la table 2.7 donne la statistique observée ainsi que la probabilité critique associées au test des l'hypothèses :

$$H_0\{\text{le paramètre est nul}\} \quad \text{contre} \quad H_1\{\text{le paramètre n'est pas nul}\} \tag{2.10}$$

La définition des statistiques et des probabilités critiques sont fournis par la table 2.8 où l'on rappelle que

- $\hat{\sigma}_a^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$ est l'estimation de la variance de l'estimateur de a. C'est un indicateur du caractère de variabilité de l'estimateur.
- $\hat{\sigma}_b^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ est l'estimation de la variance de l'estimateur de b.

Les deux statistiques de tests suivent sous l'hypothèse H_0 une loi de Student $\mathcal{T}_{n-2}$ à $n-2$ degrés de liberté.

Remarque : Le test sur le paramètre b s'écrit :

$$H_0 = \{b = 0\} \quad \text{contre} \quad H_1 = \{b \neq 0\}.$$

On remarque que c'est la traduction du test des hypothèses 2.6 en termes de paramètres. De ce fait, les deux tests sont équivalents et on peut vérifier que les deux statistiques de tests associées sont liées puisque ${T_b}^2 = F$.

Interprétation des tests.

`Intercept`. On appelle intercept le terme constant dans la modélisation de la moyenne de Y, ici le paramètre a. D'après le résultat donné dans la table 2.7, l'hypothèse $H_0 = \{a = 0\}$ est rejetée (la probabilité critique < 0.0001). On conclut que le log du Taux de DDT pour un brochet qui vient de naître est significativement non nul.

`Age`. L'hypothèse $H_0 = \{b = 0\}$ est rejetée (la probabilité critique < 0.0001). On conclut que le paramètre b est significativement non nul, et donc que le log du Taux de DDT est lié linéairement à l'âge d'un brochet (comme on l'avait déjà vu dans le paragraphe 2.2.4).

2.3 Modèle ANOVA et test de linéarité

Puisqu'il y a des répétitions des mesures du taux de DDT pour différents âges des brochets, le modèle de régression linéaire simple peut s'écrire

$$Y_{ij} = a + b\ \texttt{Age}_i + E_{ij}, \qquad \{E_{ij}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \tag{2.11}$$

où

- $Y_{ij} = \log(\texttt{TxDDT}_{\texttt{ij}})$,
- i est l'indice de l'âge : $i = 1, \ldots, 5$,

– j est l'indice de répétitions pour chaque âge : $j = 1, 2, 3$.
Par exemple, y_{42} correspond au log du Taux de DDT du 2ème brochet agé de 5 ans.

On peut effectuer deux études supplémentaires.

2.3.1 Test de l'effet du *facteur* `Age`

On peut considérer la *variable* `Age` qui est quantitative comme un *facteur* qualitatif qui possède 5 *niveaux* : "2 ans", "3 ans","4 ans", "5 ans" et "6 ans". On peut alors étudier l'influence du *facteur* `Age` sur la *variable* log `TxDDT`. Cette étude se fait au moyen du modèle d'analyse de la variance (ANOVA). Pour plus de détails, se référer au chapitre 4. Donnons le modèle d'anlayse de la variance :

$$Y_{ij} = \mu + \alpha_i + E_{ij}, \qquad \{E_{ij}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \tag{2.12}$$

où
– μ est le terme constant,
– α_i est l'effet de l'âge i.
La différence avec la régression se situe dans la question. L'analyse de la variance permet de répondre à la question *Est-ce qu'il y a un effet âge sur le* log *du Taux de DDT ?* et la régression simple permet de préciser si la relation est linéaire.

Les résultats de l'analyse de la variance pour l'ANOVA sont donnés dans la table 2.9.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.80980346	0.20245087	25.24	<.0001
Error	10	0.08020509	0.00802051		
Corrected Total	14	0.89000855			

TABLE 2.9 – Analyse de la variance.

L'hypothèse $H_0 = \{$*l'âge n'a pas d'effet sur le Taux de DDT*$\}$ est rejetée (probabilité critique < 0.0001). On conclut que l'âge contribue significativement à expliquer la variabilité du Taux de DDT. Ce qui conforte le résultat du test sur le paramètre b qui était significativement différent de 0.

2.3.2 Test de la linéarité des moyennes par âge

L'écriture matricielle du "nouveau" modèle de régression (2.11) est $\mathbf{Y} = \mathbf{X}\ \boldsymbol{\Theta}\ + \mathbf{E}$ où les matrices $\mathbf{Y}$, $\mathbf{E}$ et Θ ne changent pas (cf paragraphe 2.2.1) mais par contre la matrice $\mathbf{X}$ devient :

$$\underset{(n\times 2)}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_1 \\ 1 & x_1 \\ 1 & x_2 \\ 1 & x_2 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_5 \\ 1 & x_5 \\ 1 & x_5 \end{bmatrix}.$$

Dans le modèle d'analyse de la variance (2.12), les matrices $\mathbf{X}$ et Θ s'écrivent :

$$\underset{(n\times 6)}{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}, \quad \underset{(6\times 1)}{\Theta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_5 \end{bmatrix}.$$

Modèles emboîtés

La notion de modèles emboîtés est donnée dans Daudin *et al.* (2007) chapitre 3. En notant $\mathbf{X}_{(i)}$ la $i$ème colonne de la matrice $\mathbf{X}$, d'après les matrices $\mathbf{X}$ des modèles de régression 2.11) et d'analyse de la variance (2.12), on peut écrire

$$\mathbf{X}^{Reg}_{(1)} = \mathbf{X}^{ANO}_{(1)},$$

et

$$\mathbf{X}^{Reg}_{(2)} = x_1\mathbf{X}^{ANO}_{(2)} + x_2\mathbf{X}^{ANO}_{(3)} + \ldots + x_5\mathbf{X}^{ANO}_{(6)}.$$

Le sous-espace engendré par les colonnes de $\mathbf{X}^{Reg}$ est contenu dans le sous-espace engendré par les colonnes de $\mathbf{X}^{ANO}$: le modèle de régression est un modèle emboîté du modèle d'analyse de la variance. On peut tester les hypothèses :

$$H_0 = \{Y_{ij} = a + b\ \texttt{Age}_i + E_{ij}\} \quad \text{contre} \quad H_1 = \{Y_{ij} = \mu + \alpha_i + E_{ij}\}. \tag{2.13}$$

Ainsi on suppose que le "vrai" modèle est le modèle d'analyse de la variance et que l'on cherche à voir si ce n'est pas un modèle de régression simple. En termes de paramètres, le test (2.13) s'écrit :

$$H_0 = \{\mu + \alpha_i = a + bx_i\} \quad \text{contre} \quad H_1 = \{\mu + \alpha_i \neq a + bx_i\}.$$

En indiçant par 1 le modèle d'ANOVA et 0 le modèle de régression, la statistique de test des modèles emboîtés s'écrit classiquement

$$F = \frac{(SCR_0 - SCR_1)/(DF_0 - DF_1)}{SCR_1/DF_1},$$

La table 2.9 donne $SCR_1 = 0.80205$ et $DF_1 = 10$, et la table 2.4 donne $SCR_0 = 0.11998$ et $SCR_0 = 13$. La statistique de test vaut donc

$$F = 1.6530.$$

D'après la table des quantiles de la loi de Fisher, pour un niveau de test de $\alpha = 0.05$,

$$f_{3,10,1-\alpha} = 3.71.$$

Conclusion

L'hypothèse H_0 est acceptée ($F - value < f_{3,10,1-\alpha}$). On conclut que la contamination moyenne croit linéairement avec l'âge.

2.4 Programme SAS

```
title h=1.3 'Taux de DDT chez des brochets';
options linesize=89 pagesize=69 pageno=1 nodate;
goptions reset;

title2 'Donnees brutes';
data BROCHET;
     infile 'Brochet.don' firstobs=2;
     input Age TxDDT;
     Age2 = Age*Age;
     LogDDT = log10(TxDDT);
proc Print data=BROCHET;
symbol;
proc GPlot data=BROCHET;
     plot TxDDT * Age;
run;

title2 'Statistiques elementaires';
```

```
proc Means data=BROCHET;
     var TxDDT LogDDT;
proc Sort data=BROCHET;
     by Age;
proc Means data=BROCHET;
     var TxDDT LogDDT;
     by Age;
run;

title2 'régression lineaire simple';
proc Reg data=BROCHET;
     model TxDDT = Age / covB;
     plot TxDDT * Age / conf pred;
     plot residual. * Age / overlay vref=0;
run;

title2 'régression lineaire simple sur le log';
proc Reg data=BROCHET;
     model LogDDT = Age;
     plot LogDDT * Age / conf pred;
     plot residual. * Age / overlay vref=0;
run;

title2 'régression polynomiale sur le log';
proc Reg data=BROCHET;
     model LogDDT = Age Age2;
     plot residual. * Age / overlay vref=0;
     output out=REG p=Predite R=Residu;
proc GPlot data=BROCHET;
     symbol i=RQcli95 v=plus;
     plot TxDDT*Age;
proc Print data=REG;
proc Univariate data=REG normal plot;
     var Residu;
run;

title2 'Analyse de la variance sur l''age';
proc Anova data=BROCHET;
     class Age;
     model LogDDT = Age;
run;
```

Chapitre 3

Régression linéaire multiple

3.1 Présentation

3.1.1 Objectif et dispositif expérimental

Objectif

La processionnaire du pin est un papillon nocturne de la famille des Notodontidés. La chenille se développe de préférence sur des pins et peut causer des dégats considérables. On souhaite connaître l'influence de certaines caractéristiques de peuplements forestiers sur leurs développements.

Données

On dispose d'un échantillon de $n = 33$ parcelles forestières d'une surface de 10 hectares. Chaque parcelle est alors échantillonnée en placettes de 5 ares et on a calculé les moyennes (sur ces placettes) les différentes mesures suivantes :

- le nombre de nids de processionnaires par arbre (variable `NbNids`),
- l'altitude (en mètre) (variable `Altitude`),
- la pente (en °) (variable `Pente`),
- le nombre de pins dans une placette (variable `NbPins`),
- la hauteur de l'arbre échantillonné au centre de la placette (variable `Hauteur`),
- le diamètre de cet arbre (variable `Diametre`),
- la note de densité de peuplement (variable `Densite`),
- l'orientation de la placette (variable `Orient`), allant de 1 (sud) à 2 (autre),
- la hauteur des arbres dominants (variable `HautMax`),
- le nombre de strates de végétation (variable `NbStrat`),
- le mélange du peuplement (variable `Melange`), allant de 1 (pas mélangé) à 2 (mélangé).

Un extrait des données est présenté dans la table 3.1. Une remarque immédiate et surprenante face à ces données est que pour les parcelles 13 et 14, il n'y a aucun pin alors

que l'on observe des valeurs pour les autres variables comme le nombre de nids qui est d'ailleurs très élevé. Une explication possible est la suivante : pour chaque parcelle, les valeurs obtenues sont des moyennes de mesures faites sur un échantillon de placettes de 5 ares. Si le nombre de pins moyen par parcelle était par exemple de 0.3, il se peut que cette valeur ait été arrondie à 0. On peut alors s'interroger sur la pertinence de ces deux observations et sur leurs sensibilités face aux résultats.

Obs	Altitude	Pente	Nb Pins	Hauteur	Diametre	Densite	Orient	Haut Max	Nb Strat	Melange	Nb Nids
1	1200	22	1	4.0	14.8	1.0	1.1	5.9	1.4	1.4	2.37
2	1342	28	8	4.4	18.0	1.5	1.5	6.4	1.7	1.7	1.47
3	1231	28	5	2.4	7.8	1.3	1.6	4.3	1.5	1.7	1.13
4	1254	28	18	3.0	9.2	2.3	1.7	6.9	2.3	1.6	0.85
5	1357	32	7	3.7	10.7	1.4	1.7	6.6	1.8	1.3	0.24
.../...											
12	1182	41	32	5.4	21.6	3.3	1.4	11.3	2.8	2.0	0.70
13	1179	15	0	3.2	10.5	1.0	1.7	4.0	1.1	1.6	2.64
14	1256	21	0	5.1	19.5	1.0	1.8	5.8	1.1	1.4	2.05
15	1251	26	2	4.2	16.4	1.1	1.7	6.2	1.3	1.8	1.75
.../...											
29	1208	23	2	3.5	11.5	1.1	1.7	5.4	1.3	2.0	1.09
30	1198	28	15	3.9	11.3	2.0	1.6	7.4	2.8	2.0	0.18
31	1228	31	6	5.4	21.8	1.3	1.7	7.0	1.5	1.9	0.35
32	1229	21	11	5.8	16.7	1.7	1.8	10.0	2.3	2.0	0.21
33	1310	36	17	5.2	17.8	2.3	1.9	10.3	2.6	2.0	0.03

TABLE 3.1 – Extrait des données portant sur les processionnaires de pins.

3.1.2 Description des données

Statistiques élémentaires

La table 3.2 donnent les statistiques élémentaires. Ces statistiques ne sont pas ici comparables puisque les variables ne sont pas toutes de même nature (il faudrait pour cela travailler sur des données centrées réduites). Seule la comparaison de la hauteur de l'arbre échantillonné au centre de la placette et de la hauteur maximale moyenne des arbres dominants est pertinente. On observe que la moyenne de Hauteur, qui vaut 4.45 est inférieure à la moyenne de HautMax, qui vaut 7.54, ce qui est rassurant !.

Coefficient de corrélation

Le coefficient de corrélation entre deux variables X et Y se note $\rho(X,Y)$ et est estimé par

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Altitude	33	1315	129.03698	43406	1075	1575
Pente	33	29.03030	7.30362	958.00000	15.00000	46.00000
NbPins	33	11.45455	9.53641	378.00000	0	32.00000
Hauteur	33	4.45152	1.04077	146.90000	2.40000	6.50000
Diametre	33	15.25152	4.30255	503.30000	5.80000	21.80000
Densite	33	1.79091	0.71736	59.10000	1.00000	3.30000
Orient	33	1.65758	0.18713	54.70000	1.10000	1.90000
HautMax	33	7.53939	2.35185	248.80000	3.60000	13.70000
NbStrat	33	1.98182	0.56649	65.40000	1.10000	2.90000
Melange	33	1.76061	0.24867	58.10000	1.30000	2.00000
NbNids	33	0.81121	0.80623	26.77000	0.03000	3.00000
logNids	33	-0.81328	1.24494	-26.83825	-3.50656	1.09861

TABLE 3.2 – Statistiques élémentaires.

Il est utilisé pour mesurer le degré de liaison (linéaire) qui peut exister entre deux variables.

La table 3.3 représente les coefficients de corrélation entre toutes les variables décrites précédemment. La valeur qui se trouve au dessous du coefficient de corrélation est la probabilité critique du test de la nullité de ce coefficient, i.e. des hypothèses

$$H_0 = \{\rho = 0\} \quad \text{contre} \quad H_1 = \{\rho \neq 0\}.$$

Certaines liaisons simples auxquelles on pouvait s'attendre se retrouvent dans cette table : une forte liaison significative entre les variables `Hauteur` et `Diametre` (coeff de corrélation à 0.905) qui sont généralement très liées, ou encore entre `NbPins` et `Densité` (coeff de corrélation à 0.98) qui sont des mesures différentes d'un même phénomène. Mais d'autres résultats moins évidents apparaissent aussi comme la forte liaison entre `NbPins` et `NbStrat` (coeff de corrélation à 0.88), ou `NbStrat` et `Densité` (coeff de corrélation à 0.9). Ainsi deux groupes de variables corrélées se dégagent : {Diametre, Hauteur} et {NbStrat, Densité, NbPins, HautMax}.

Remarque : un coefficient de corrélation négatif mais proche de 1 en valeur absolue ne signifie pas qu'il n'existe pas de liaison entre les deux variables. Cela indique au contraire qu'elles sont liées et qu'elles ont tendance à évoluer en sens contraire.

La table précédente indique aussi que les liaisons entre la variable à expliquer (`NbNids`) et toutes les autres variables ne sont pas très élevées puisque le plus grand des coefficients de corrélation vaut -0.594, c'est celui avec la variable `NbStrat`, et le second vaut -0.528 avec la variable `Densite`. Ces corrélations négatives signifient que plus il y a de végétation, moins il y a de nids. Cela se retrouve dans la table des données (3.1). En effet, on peut observer que le nombre de nids moyen par arbre est très élevé pour des parcelles où le nombre de pins est très faible, donc pour une densité faible.

	Altitude	Pente	NbPins	Hauteur	Diametre	Densite
Altitude	1.00000	0.12052	0.53756	0.32105	0.28377	0.51467
		0.5041	0.0013	0.0685	0.1095	0.0022
Pente	0.12052	1.00000	0.32194	0.13669	0.11342	0.30067
	0.5041		0.0677	0.4481	0.5297	0.0891
NbPins	0.53756	0.32194	1.00000	0.41443	0.29492	0.97955
	0.0013	0.0677		0.0165	0.0957	<.0001
Hauteur	0.32105	0.13669	0.41443	1.00000	0.90466	0.43930
	0.0685	0.4481	0.0165		<.0001	0.0105
Diametre	0.28377	0.11342	0.29492	0.90466	1.00000	0.30623
	0.1095	0.5297	0.0957	<.0001		0.0831
Densite	0.51467	0.30067	0.97955	0.43930	0.30623	1.00000
	0.0022	0.0891	<.0001	0.0105	0.0831	
Orient	0.26849	-0.15222	0.12847	0.05810	-0.07871	0.15068
	0.1308	0.3977	0.4762	0.7481	0.6633	0.4026
HautMax	0.36015	0.26191	0.75896	0.77193	0.59620	0.81022
	0.0395	0.1409	<.0001	<.0001	0.0003	<.0001
NbStrat	0.36372	0.32567	0.87679	0.45959	0.26746	0.90853
	0.0375	0.0644	<.0001	0.0071	0.1324	<.0001
Melange	-0.12764	0.12800	0.18700	-0.12111	-0.09063	0.10829
	0.4790	0.4778	0.2974	0.5020	0.6159	0.5486
NbNids	-0.53022	-0.45546	-0.56390	-0.35790	-0.15777	-0.57024
	0.0015	0.0077	0.0006	0.0409	0.3805	0.0005

	Orient	HautMax	NbStrat	Melange	NbNids
Altitude	0.26849	0.36015	0.36372	-0.12764	-0.53022
	0.1308	0.0395	0.0375	0.4790	0.0015
Pente	-0.15222	0.26191	0.32567	0.12800	-0.45546
	0.3977	0.1409	0.0644	0.4778	0.0077
NbPins	0.12847	0.75896	0.87679	0.18700	-0.56390
	0.4762	<.0001	<.0001	0.2974	0.0006
Hauteur	0.05810	0.77193	0.45959	-0.12111	-0.35790
	0.7481	<.0001	0.0071	0.5020	0.0409
Diametre	-0.07871	0.59620	0.26746	-0.09063	-0.15777
	0.6633	0.0003	0.1324	0.6159	0.3805
Densite	0.15068	0.81022	0.90853	0.10829	-0.57024
	0.4026	<.0001	<.0001	0.5486	0.0005
Orient	1.00000	0.06001	0.06325	0.13085	-0.21175
		0.7401	0.7266	0.4680	0.2368
HautMax	0.06001	1.00000	0.85364	0.00327	-0.55113
	0.7401		<.0001	0.9856	0.0009
NbStrat	0.06325	0.85364	1.00000	0.14782	-0.63587
	0.7266	<.0001		0.4117	<.0001
Melange	0.13085	0.00327	0.14782	1.00000	-0.11276
	0.4680	0.9856	0.4117		0.5321
NbNids	-0.21175	-0.55113	-0.63587	-0.11276	1.00000
	0.2368	0.0009	<.0001	0.5321	

TABLE 3.3 – Matrice de corrélations entre les 11 variables.

Coefficient de corrélation partielle

L'utilisation directe du coefficient de corrélation pour interpréter les liens entre les variables n'est pas toujours pertinente. En effet, le lien entre deux variables peut venir du fait que ces deux variables sont liées à une troisième. Par exemple, on a vu que les deux variables les plus corrélées à `NbNids` étaient `NbStrat` et `Densité` mais ces deux variables sont elles mêmes très fortement corrélées. On peut se demander si ce n'est pas cette forte liaison qui rend la liaison entre `NbNids` et `Densité` forte.
On s'intéresse alors au coefficient de corrélation partielle. Il permet en effet de mesurer la relation qui existe entre les deux variables X et Y corrigée pour enlever l'influence de la variable Z, i.e. "conditionnellement" à cette variable Z". Il se note $\rho(X, Y|Z)$ et est estimé par

$$\hat{\rho}(X, Y|Z) = \frac{\hat{\rho}(X, Y) - \hat{\rho}(X, Z)\hat{\rho}(Y, Z)}{\sqrt{(1 - \hat{\rho}(X, Z)^2)(1 - \hat{\rho}(Y, Z)^2)}}.$$

On peut le voir d'une autre façon : c'est la proportion de variance du résidu de X par rapport à Z expliquée par le résidu de Y par rapport à Z. En d'autres termes, si on régresse X et Y par Z,

$$X = a_1 + b_1 Z + E \quad , \quad Y = a_2 + b_2 Z + F$$

on obtient que

$$\hat{\rho}(X, Y|Z) = \hat{\rho}(E, F).$$

La table 3.4 donne les corrélations partielles entre la variable `NbNids` et les autres variables conditionnellement à la variable `NbStrat`. On observe que la corrélation partielle entre `NbNids` et `Densité`, qui vaut 0.023, est très faible et non significative alors que la forte corrélation entre ces deux variables, qui indique une forte redondance d'information, est la deuxième plus élevée.

	Altitude	Pente	NbPins	Hauteur	Diametre	Densite
NbNids	-0.41581	-0.34037	-0.01718	-0.09579	0.01654	0.02316
	0.0179	0.0566	0.9257	0.6020	0.9284	0.8999

	Orient	HautMax	Melange
NbNids	-0.22270	-0.02071	-0.02458
	0.2205	0.9104	0.8938

TABLE 3.4 – Corrélations partielles entre la variable `NbNids` et les autres variables conditionnellement à la variable `NbStrat`.

3.2 Modèle de régression linéaire multiple

3.2.1 Modèle

On cherche à expliquer le nombre de nids de processionnaires à partir de $p = 10$ variables explicatives décrites précédemment. De façon générale, il faut qu'on dispose

de plus d'observations que de paramètres ($n >> p$), sinon on ne pourra pas estimer les coefficients de régression. On considère le modèle de régression linéaire multiple qui s'écrit de la façon suivante :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + E_i \qquad \{E_i\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \tag{3.1}$$

où

- l'indice i représente le numéro de la parcelle,
- la variable Y_i désigne le nombre de nids moyen de la $i$ème parcelle,
- les variables x_e, $e = 1, \ldots, p-1$, sont les différentes variables explicatives. Elles sont généralement supposées linéairement indépendantes (aucune n'est combinaison linéaire des autres), ce qui ne veut pas dire qu'elles sont statistiquement indépendantes. Cette hypothèse signifie que chaque variable doit apporter une information nouvelle par rapport aux autres. Plus les variables sont corrélées, moins bien les paramètres associés seront estimés.
- les β_i sont appelés "coefficients de régression". Un β_i s'interprète comme l'accroissement de Y_i correspondant à l'accroissement d'une unité de la variable x_i associée quand les autres sont maintenues constantes,
- σ^2 est la variance résiduelle.

Ce modèle suppose donc qu'il existe une relation linéaire entre les variables explicatives (les x_i) et la variable à expliquer y. Mais ce n'est pas toujours le cas. Un diagnostic graphique peut-etre fait à l'aide des graphes de y en fonction de chaque x_i. Si la relation ne semble pas linéaire, des transformations peuvent envisagées comme considérer plutôt le $\log x_i$.

Ecriture matricielle.

Le modèle (3.1) s'écrit sous la forme matricielle suivante

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}, \tag{3.2}$$

où
$$\underset{(n\times 1)}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{(n\times 1)}{\mathbf{E}} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix},$$

$$\underset{(n\times p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix}, \quad \underset{(p+1\times 1)}{\boldsymbol{\Theta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

3.3 Régression linéaire multiple

Les définitions des différentes quantités données dans les tables de cette partie ont déjà été décrites dans le chapitre précédent.

3.3.1 Régression linéaire multiple sur la variable `NbNids`

Comme pour toutes les analyses de modèle linéaire, avant de regarder les résultats, il faut vérifier que les hypothèses du modèle sont satisfaites. Pour cela, nous avons recours à des méthodes graphiques déjà évoquées dans le chapitre précédent. Le graphique des résidus en fonction des valeurs prédites, représenté par la figure 3.1, montre une augmentation de la variance. Le modèle (3.1) ne semble pas être adapté aux données. Nous proposons alors d'effectuer une transformation logarithmique de la variable à expliquer `NbNids` dont l'analyse fait l'objet de la partie suivante.

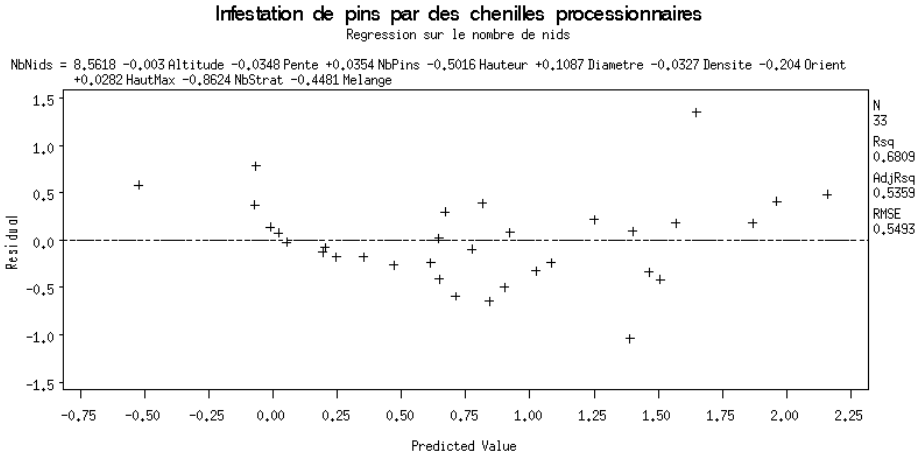

FIGURE 3.1 – Graphe des résidus pour la régression multiple sur la variable `NbNids`.

3.3.2 Régression linéaire multiple sur la variable log `NbNids`

Le modèle (3.1) reste inchangé sauf que Y_i représente maintenant le logarithme du nombre de nids moyen par arbre pour la $i$ème parcelle.

Table d'analyse de la variance

La table 3.5 fournit la synthèse des résultats de l'analyse de la variance : les sommes des carrés dues au modèle (SCM), aux résidus (SCR) et totale (SCT) ainsi que la statistique et la probabilité critique du test du modèle $Y_i = \beta_0 + E_i$ contre le modèle complet décrit par (3.1). En termes de paramètres, cela revient à effectuer le test des hypothèses :

$$H_0 = \{\beta_1 = \ldots = \beta_p = 0\} \quad \text{contre} \quad H_1 = \{\exists i / \beta_i \neq 0\}.$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	34.55715	3.45571	5.06	0.0007
Error	22	15.03892	0.68359		
Corrected Total	32	49.59607			

Root MSE	0.82679	R-Square	0.6968
Dependent Mean	-0.81328	Adj R-Sq	0.5589
Coeff Var	-101.66156		

TABLE 3.5 – Table d'analyse de la variance sur le log du nombre de Nids.

Conclusion. La probabilité critique vaut 0.0007 et est inférieure à 0.05. On conclut que le test est significatif, on rejette l'hypothèse H_0. Ce résultat montre qu'au moins une des variables contribue à expliquer le nombre de nids. Ce résultat est global et ne nous indique pas si plusieurs variables y contribuent et lesquelles.

Ajustement du modèle. Comme pour la régression linéaire simple, un des usages de la régression multiple consiste à prédire la valeur d'un y pour un ensemble de valeurs $x_1, \ldots, x_p$ donné. La mesure de l'ajustement du modèle aux données est donc importante. La proportion de variabilité expliquée par les 10 régresseurs est de

$$R^2 = R - Square = 69.5\%,$$

ce qui montre un ajustement moyen. Comme il a déjà été évoqué dans le chapitre précédent, ce coefficient ne prend pas en compte le nombre de variables explicatives. C'est pourquoi, il est nécessaire de s'intéresser au $R^2 - Adj$, qui lui représente une mesure de l'ajustement corrigée par le nombre de variables du modèle. Ici, ce coefficient vaut

$$R^2 - Adj = 53.6\%,$$

qui reste un ajustement moyen.

Variance résiduelle. La table 3.5 fournit aussi l'estimation de la variance résiduelle notée $\hat{\sigma}^2$ soit sous une forme directe donnée au croisement de la ligne `Error` et de la colonne `Mean Square`, soit par son écart-type qui se retrouve à la ligne `Root MSE`. On obtient $\hat{\sigma}^2 = 0.68359$, ce qui montre une faible variance résiduelle.

Analyse des résidus. Le graphique des résidus en fonction des valeurs prédites, donné figure 3.1, ne montre plus de structure particulière.

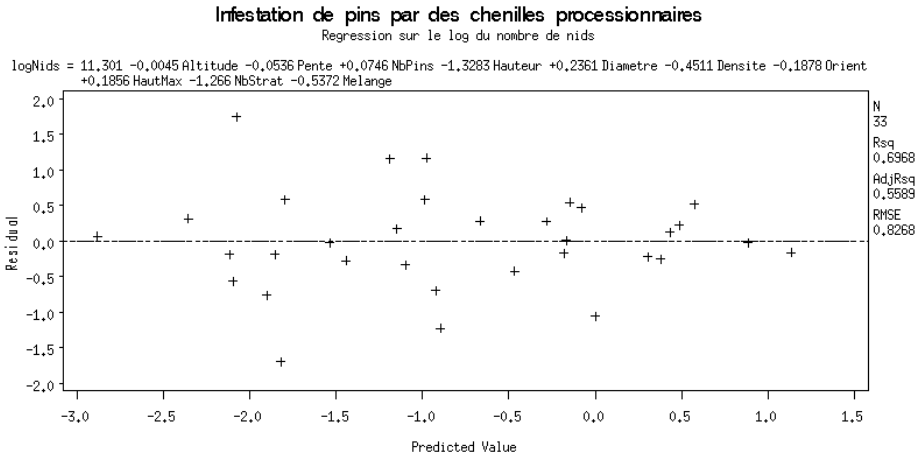

FIGURE 3.2 – Graphe des résidus pour la régression multiple sur la variable `NbNids`.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.30091	3.15655	3.58	0.0017
Altitude	1	-0.00451	0.00156	-2.88	0.0086
Pente	1	-0.05361	0.02184	-2.45	0.0225
NbPins	1	0.07458	0.10023	0.74	0.4647
Hauteur	1	-1.32828	0.57006	-2.33	0.0294
Diametre	1	0.23610	0.10461	2.26	0.0343
Densite	1	-0.45111	1.57292	-0.29	0.7770
Orient	1	-0.18781	1.00795	-0.19	0.8539
HautMax	1	0.18564	0.23634	0.79	0.4406
NbStrat	1	-1.26603	0.86124	-1.47	0.1557
Melange	1	-0.53720	0.77337	-0.69	0.4946

TABLE 3.6 – Estimations et tests sur les paramètres.

3.3.3 Estimation des paramètres et tests sur les paramètres

La table 3.6 fournit les estimations des β_i ainsi que la statistique et la probabilité critique des tests des hypothèses :

$$H_0 = \{\beta_i = 0\} \quad \text{contre} \quad H_1 = \{\beta_i \neq 0\}.$$

Avec un niveau de test de $\alpha = 0.05$, le résultat des tests est cohérent avec le test global précédent puisqu'au moins un des coefficients de régression est non significativement nul. Mais ce n'est pas toujours le cas, il se peut que les coefficients soient significativement nuls alors que leur somme conduit à un effet global significatif. On peut remarquer que les coefficients associés aux variables `Altitude`, `Pente`, `Diametre` et `Hauteur` sont significativement différents de 0. Ce résultat peut paraître supprenant au vu des corrélations

données dans la table 3.3 qui montraient que la variable `NbStrat` était la plus corrélée avec la variable à expliquer.
Dans une situation simple où tous les coefficients de régression sont significatifs, l'utilisateur conservera toutes les variables dans la régression, qui est globalement significative. Si ce n'est pas le cas, comme ici, on serait tenté d'éliminer des variables du modèle sur la base de ces tests (les variables dont les coefficients sont significativement nuls). Mais cette procédure est incorrecte. Il ne faut pas oublier que le test d'un coefficient est effectué alors que les autres variables sont fixées. Donc si deux variables sont très corrélées, le test d'un des deux coefficients peut être non significatif puisque l'information apportée par la variable testée existe déjà par dans l'autre.
On ne peut donc rien conclure sur l'estimation de ces coefficients et de leurs significativités, et il est préférable de s'en tenir à l'analyse des corrélations. Mais ceci nous amène à la question de la sélection de variables qui est l'objet du paragraphe suivant.

3.3.4 Sélection de variables explicatives

La sélection de variables est une perspective naturelle à plus d'un titre. La principale raison est que soit certaines variables ne contribuent pas à l'explication de la variable à expliquer, soit des variables sont très corrélées et apportent donc une redondance d'information. Dans ces deux situations, on a envie de les éliminer du modèle. Il faut bien noter que l'on cherche toujours à privilégier le modèle le plus simple possible permettant ainsi une interprétation facile et pouvant éviter à l'expérimentateur des coûts d'acquisition de certaines données s'il souhaite d'autres données pour prédire Y. De plus, un trop grand nombre de variables conduit d'une part à une imprécision dans l'estimation des coefficients de régression et d'autre part peut mener à une augmentation de la variance résiduelle puisque le nombre de degrés de libertés lui diminue.

L'objectif est donc de déterminer à partir de toutes les variables explicatives un sous-ensemble de variables suffisamment explicatif. Une première possibilité brutale consiste à évaluer toutes les régressions possibles. Malheureusement, cette solution est souvent très longue, voire impossible temporellement, dès lors que le nombre de variables est grand (le nombre de régressions étant de 2^p). Les méthodes les plus utilisées sont dites "pas à pas" dans lesquelles les variables sont introduites ou supprimées dans la régression l'une après l'autre. Ces méthodes sont des heuristiques construites à partir d'approches numériques mais aucune ne garantit un résultat optimal.

Différentes méthodes de sélection

Il existe trois variantes de ces méthodes que nous décrivons ici.

Ascendante. La procédure commence avec le terme constant β_0, soit le modèle nul : $Y_i = \beta_0 + E_i$. Ensuite, elle s'effectue en plusieurs étapes :

Etape 1 *:* On choisit la variable x_{k_1} parmi l'ensemble des variables de départ qui contribue le plus à expliquer Y, i.e. celle qui fait le plus augmenter le R^2 ou encore telle que

$$\hat{\rho}(Y, X_{k_1}) \text{ est maximal.}$$

Ensuite, on test la nullité du coefficient de régression associé et la variable est retenue en cas de significativité du test.

Etape 2 *:* On choisit la variable x_{k_2} parmi l'ensemble des variables auquel on a retiré x_{k_1} telle que

$$\hat{\rho}(Y, X_{k_2}|X_{k_1}) \text{ est maximal.}$$

C'est n'est donc pas la seconde variable la plus corrélée à Y mais c'est celle qui apporte le plus d'information en plus de X_{k_1}. De la même façon que précédemment, le coefficient de régression est testé.

Il existe plusieurs tests d'arrêts de la procédure : en choisissant un nombre a priori de variables ou une valeur finale de R^2, ou encore dès que le test de nullité de la dernière variable introduite n'est pas significatif. C'est cette dernière solution qui est faite dans SAS.

Descendante. C'est la procédure symétrique de la précédente qui part du modèle complet et élimine à chaque étape la variable correspondante au plus petit coefficient de corrélation partielle.

Stepwise. Cette procédure est semblable à l'ascendante avec remise en cause à chaque étape des variables déjà introduites. En effet, il arrive souvent que des variables introduites en tête, par le biais de leur liaison avec une ou plusieurs variables introduites ultérieurement, ne soient plus significatives.

Sélection de variables sur l'exemple par Stepwise

La stratégie utilisée sur l'exemple du nombre de nids est la procédure Stepwise. La table 3.7 donne les différentes étapes de cette stratégie.

Etape 1 *:* le plus fort coefficient de corrélation avec la variable `NbNids` est obtenu par la variable `NbStrat` (cf table 3.3), comme on avait déjà pu le constater. A partir du modèle nul, le modèle devient :

$$\log NbNids_i = \beta_0 + \beta_1 NbStrat_i + E_i. \tag{3.3}$$

La table 3.7 montre une probabilité critique de 0.003 pour le test du modèle nul contre ce nouveau modèle ou de façon équivalente pour le test de la nullité du coefficient β_1. Remarquons que ce dernier test est effectué au moyen de la statistique de Fischer alors qu'avant c'était par la statistique de Student. Par leur relation $T^2 = F$, on a complète équivalence. L'hypothèse nulle est rejetée, il existe une relation linéaire significative entre

Stepwise Selection: Step 1
Variable NbStrat Entered: R-Square = 0.3528 and C(p) = 17.9543

Source	DF	Sum of Squares	Mean Square	F Value	Pr> F
Model	1	17.49867	17.49867	16.90	0.0003
Error	31	32.09740	1.03540		
Corrected Total	32	49.59607			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.77374	0.65375	7.62205	7.36	0.0108
NbStrat	-1.30538	0.31753	17.49867	16.90	0.0003

Stepwise Selection: Step 2
Variable Altitude Entered: R-Square = 0.4690 and C(p) = 11.5220

Source	DF	Sum of Squares	Mean Square	F Value	Pr> F
Model	2	23.26290	11.63145	13.25	<.0001
Error	30	26.33317	0.87777		
Corrected Total	32	49.59607			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.83839	1.69652	10.39562	11.84	0.0017
Altitude	-0.00353	0.00138	5.76423	6.57	0.0156
NbStrat	-1.01283	0.31386	9.14073	10.41	0.0030

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	NbStrat		1	0.3528	0.3528	17.9543	16.90	0.0003
2	Altitude		2	0.1162	0.4690	11.5220	6.57	0.0156

TABLE 3.7 – Procédure Stepwise.

`NbNids` et `NbStrat`. La procédure se poursuit.

Etape 2 *:* à partir de la table 3.4 qui donne les corrélations partielles entre la variable `NbNids` et les autres variables conditionnellement à la variable `NbStrat`, on observe que c'est la variable `Altitude` qui a le plus fort coefficient à -0.416. Le nouveau modèle est alors :

$$\log NbNids_i = \beta_0 + \beta_1 NbStrat_i + \beta_2 Altitude_i + E_i. \tag{3.4}$$

Le test du modèle nul contre ce modèle est fortement significatif avec une probabilité critique de $< .0001$ et l'hypothèse de nullité du coefficient β_2 est rejetée avec une probabilité critique de 0.0156 (le test est moins significatif). On remet alors en cause la variable `NbStrat` en testant la nullité de son paramètre. La probabilité critique est de 0.003, H_0 est rejetée. Le modèle complet est conservé.

Etape 3 *:* SAS s'arrête là en précisant que le résultat du test de Fisher de la variable suivante n'est pas significatif au niveau 5%.

Un résumé de la procédure est donné en fin de sortie. Dans la colonne `Model R-Square`, on retrouve les valeurs du coefficient de détermination R^2 du premier modèle (3.3) dans la ligne `NbStrat` et du second modèle (3.4) dans la ligne `Altitude`. On voit bien que l'introduction d'une variable dans un modèle fait augmenter le R^2. Le R^2 partiel donné dans la colonne correspond à l'apport de variabilité expliquée par la variable introduite en terme de R^2. On obtient que le $R^2 - partial$ de la variable `NbStrat` est logiquement le même que le R^2 puisqu'elle a été introduite au modèle nul et le $R^2 - partial$ de la variable `Altitude` vaut $R_2^2 - R_1^2$.

Modèle final et conclusion

Le modèle retenu est le suivant :

$$\log NbNids_i = \beta_0 + \beta_1 NbStrat_i + \beta_2 Altitude_i + E_i.$$

La figure 3.3 représente le graphe des résidus en fonction des valeurs prédites et ne montre pas de structure particulière.

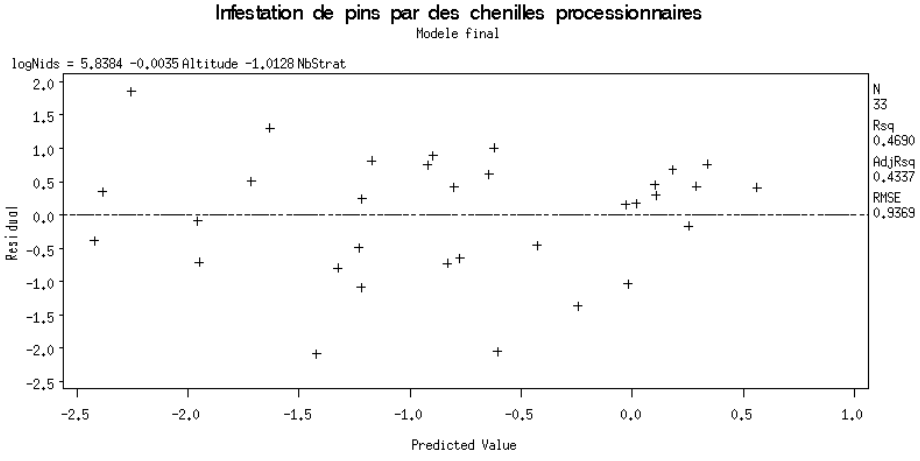

FIGURE 3.3 – Graphe des résidus pour la régression multiple du modèle $\log NbNids_i = \beta_0 + \beta_1 NbStrat_i + \beta_2 Altitude_i + E_i$.

D'après les estimations des coefficients de régressions données dans la table 3.7, la régression estimée est :

$$\log \widehat{NbNids}_i = 5.84 - 0.0035 NbStrat_i - 1.013 Altitude_i.$$

On conclut que pour avoir le moins de parasites il faut que la forêt soit en haute altitude et qu'il y ait beaucoup de végétation.

3.4 Programme SAS

Saisie des Données. Les lignes de commandes suivantes servent à saisir les données en créeant un tableau PINS et le donnent en sortie (**proc Print**).

```
data PINS;
     infile 'Chenille.don' firstobs=2;
     input Altitude Pente NbPins Hauteur Diametre Densite Orient HautMax
NbStrat Melange NbNids logNids;
proc Print data=PINS;
run;
```

Statistiques élémentaires et corrélations. La procédure **proc Corr** donne les statistiques élémentaires de chaque variable ainsi que la matrice des corrélations entre les variables. Dans la seconde procédure, des options ont été posées pour donner les corrélations partielles entre variables conditionnellement à la variable NbStrat.

```
title2 'Correlations entre les variables';
proc Corr data=PINS;
run;

proc Corr data=PINS;
var NbNids Altitude Pente NbPins Hauteur Diametre Densite Orient HautMax
Melange;
partial NbStrat;
run;
```

Régression multiple sur la variable `NbNids` et graphe des résidus en fonction des valeurs prédites.

```
title2 'régression sur le nombre de nids';
proc Reg data=PINS;
     model NbNids = Altitude Pente NbPins Hauteur Diametre Densite Orient
                    HautMax NbStrat Melange;
     plot residual. * predicted. / vref=0;
run;
```

Régression multiple sur la variable log `NbNids`**.**

```
title2 'régression sur le log du nombre de nids';
proc Reg data=PINS;
     model LogNids = Altitude Pente NbPins Hauteur Diametre Densite Orient
                     HautMax NbStrat Melange / covb corrb;
     plot residual. * predicted. / vref=0;
run;
```

Sélection de variables par la procédure Stepwise. L'option slentry donne le niveau de seuil d'entrée des variables et slstay le niveau du test de significativité des coefficients des variables qui sont entrées.

```
title2 'Selection de variables';
proc Reg data=PINS;
     model LogNids = Altitude Pente NbPins Hauteur Diametre Densite Orient
                     HautMax NbStrat Melange
                     / selection=stepwise slentry=0.05 slstay=0.05;
     output out=REG p=Predite R=Residu;
run;
```

Chapitre 4

Analyse de la variance à un facteur

4.1 Présentation

4.1.1 Objectif et dispositif expérimental

On cherche à savoir si le statut de domination d'un arbre a une influence sur son diamètre. On s'intéresse ici à des alisiers pour lesquels 3 statuts sont définis.

Co-dominant : de la même hauteur qu'un ou plusieurs arbres avoisinants ;

Dominant : plus haut que les arbres avoisinants ;

Dominés : plus bas qu'un ou plusieurs arbres avoisinants.

Données

On dispose d'un échantillon de $n = 104$ arbres choisis aléatoirement en forêt de Rambouillet (Yvelines). Sur chacun d'entre eux on a mesuré :

- ses coordonnées géographiques (variable `Xpos` = longitude et `Ypos` = latitude) ;
- son diamètre en cm (variable `Diamètre`) ;
- son statut (variable `Statut`).

Un extrait des données recueillies est présenté dans le tableau 4.1.

4.1.2 Description des données

Effectif. Le tableau 4.2 donnent quelques statistiques élémentaires sur l'ensemble de l'échantillon puis le détail en fonction des statuts. On observe un déséquilibre entre les effectifs des différents statuts : en notant n_i l'effectif du groupe i, il y a $n_1 = 25$ arbres co-dominants, $n_2 = 15$ arbres dominants et $n_3 = 64$ arbres dominés. L'échantillon ayant été tiré aléatoirement, on peut supposer que ces effectifs reflètent les proportions réelles des différents statuts dans la forêt considérée.

Obs	Xpos	Ypos	Diametre	Statut
1	558476.00	2417501.00	8.0	codomina
2	558512.04	2417405.15	15.0	codomina
3	558488.99	2417370.83	22.0	codomina
4	558489.47	2417456.32	20.0	codomina
5	558443.26	2417252.11	17.0	codomina
../..				
100	559158.47	2415602.59	16	domine
101	559157.00	2415602.60	10	domine
102	558888.96	2415683.50	52	domine
103	559441.56	2415244.25	9	domine
104	559857.00	2415238.00	8	domine

TABLE 4.1 – Extrait du tableau données portant sur le diamètre d'alisiers.

Moyennes. On observe également des différences nettes entre les diamètres moyens. En notant $Y_{i\bullet}$ le diamètre moyen du groupe i :

$$Y_{i\bullet} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik},$$

on observe $y_{2\bullet} = 35.1$ cm pour les dominants, $y_{3\bullet} = 29.2$ cm pour les co-dominants, et $y_{1\bullet} = 28.7$ cm pour les dominés. La moyenne générale vaut

$$y_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_i} y_{ik} = 29.8 \text{ cm},$$

en notant $I = 3$ le nombre de groupes (statuts) et n l'effectif global $n = \sum_{i=1}^{I} n_i$.

Variances et distributions. Il existe cependant une forte variabilité [1] à l'intérieur des groupes. L'écart type ("`Std Dev`" pour *standard deviation*) est à peu près le même dans les différents groupes (16.3 cm pour les co-dominants, 18.7 pour les dominants et 13.2 cm pour les dominés). Ces observations se retrouvent dans les "boîtes à moustaches" (*box-plot*) présentées à la figure 4.1. On rappelle que les limites inférieure et supérieure de la boîte correspondent respectivement aux premier et troisième quartiles de la distribution et que les extrémités des moustaches indiquent le minimum et le maximum.

4.2 Analyse de la variance à un facteur

4.2.1 Modèle

On veut étudier l'influence du *facteur* `Statut` sur la *variable* `Diamètre`. Le facteur `Statut` a $I = 3$ *niveaux* : "co-dominant", "dominant" et "dominé". Cette influence peut

1. On parle ici de variabilité au sens large, et non de variance qui a un sens précis en probabilités.

N	Mean	Std Dev	Minimum	Maximum
104	29.7740385	14.8438278	5.0000000	72.0000000

Statut=codomina

N	Mean	Std Dev	Minimum	Maximum
25	29.2200000	16.2980571	5.0000000	55.0000000

Statut=dominant

N	Mean	Std Dev	Minimum	Maximum
15	35.1333333	18.7230441	11.0000000	72.0000000

Statut=domine

N	Mean	Std Dev	Minimum	Maximum
64	28.7343750	13.1562553	8.0000000	56.0000000

TABLE 4.2 – Statistiques élémentaires pour l'ensemble des arbres et par statut.

être décrite au moyen d'un modèle d'analyse de la variance :

$$Y_{ik} = \mu + \alpha_i + E_{ik}, \qquad \{E_{ik}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2) \tag{4.1}$$

où

- l'indice i représente le statut (1 = co-dominant, 2 = dominant, 3 = dominé),
- l'indice k est le numéro de l'arbre au sein de son groupe,
- la variable Y_{ik} désigne le diamètre (supposé aléatoire) du k-ème arbre du i-ème groupe,
- le paramètre μ est un terme constant,
- le paramètre α_i est l'effet (additif) du statut i,
- la variable E_{ik} est un terme résiduel aléatoire,
- σ^2 est la variance résiduelle.

Comme dans tout le modèle linéaire, on suppose que les variables aléatoires $\{E_{ik}\}$ sont indépendantes, et de même loi $\mathcal{N}(0, \sigma^2)$.

Discussion des hypothèses.

1. L'hypothèse de normalité peut être vérifiée par une analyse des résidus analogue à celle présentée pour la régression simple au chapitre 2.

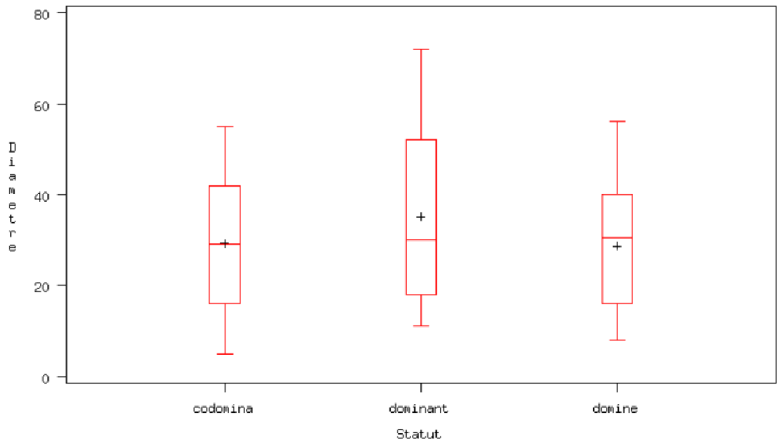

FIGURE 4.1 – Répartition du diamètre en fonction du statut de dominance des alisiers.

2. L'hypothèse d'homogénéité des variances (*homoscédasticité*) entre les groupes est cohérente avec les observations faites au paragraphe 4.1.2.
3. L'indépendance des observations est difficile (voire impossible) à vérifier à posteriori. C'est au moment du recueil des données (*i.e.* en forêt) que toutes les précautions pour garantir cette hypothèse ont dû être prises en évitant, par exemple, de choisir un arbre dominé au voisinage d'un arbre dominant.

Écriture en terme de loi des Y_{ik}.

Le modèle (4.1) est équivalent au modèle

$$Y_{ik} \sim \mathcal{N}(\mu_i, \sigma^2), \qquad \{Y_{ik}\} \text{ indépendants} \tag{4.2}$$

en notant

$$\mu_i = \mu + \alpha_i.$$

Ce modèle est analogue au modèle posé pour le test de student visant à comparer deux populations indépendantes (cf Daudin *et al.* (1999), p. 87–88).

Il est important de remarquer que cette version du modèle n'utilise que 3 paramètres pour l'espérance (μ_1, μ_2, μ_3) alors que le modèle (4.1) en utilise 4 $(\mu, \alpha_1, \alpha_2, \alpha_3)$. Ceci nous rappelle que le modèle (4.1) n'est pas identifiable (cf Daudin *et al.* (2007), subsection 2.1.3) et qu'il faudra appliquer des contraintes sur les paramètres pour pouvoir les estimer.

Écriture matricielle.

Le modèle (4.1) peut également s'écrire sous la forme matricielle générale à tout le modèle linéaire :

$$\mathbf{Y} = \mathbf{X\Theta} + \mathbf{E}, \qquad \mathbf{E} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

avec

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{3n_3} \end{bmatrix}, \quad \underset{n\times(I+1)}{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & 1 & 0 & \vdots \\ \vdots & 0 & 1 & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & 0 \\ \vdots & \vdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \underset{(I+1)\times 1}{\mathbf{\Theta}} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}, \quad \underset{n\times 1}{\mathbf{E}} = \begin{bmatrix} E_{11} \\ \vdots \\ E_{1n_1} \\ E_{21} \\ \vdots \\ E_{2n_2} \\ E_{31} \\ \vdots \\ E_{3n_3} \end{bmatrix} \tag{4.3}$$

en notant n_i l'effectif du groupe i : $n_1 = 25$, $n_2 = 15$ et $n_3 = 64$.

La non-identifiabilité du modèle se retrouve ici dans le fait que la matrice $\mathbf{X}$ n'est pas de rang 4 mais seulement 3 puisque sa première colonne est égale à la somme des trois suivantes.

4.2.2 Test de l'effet du statut

Un des principaux objectifs de l'analyse de la variance est de déterminer si le facteur (ici le statut) à un effet (une influence) sur la variabilité de la variable étudiée (ici le diamètre). Il est important de noter que le test de l'effet statut étudié ici est une notion *globale*. On étudie l'effet du statut *en général* sur le diamètre et non, par exemple, l'effet du statut "dominant".
On a vu au paragraphe précédent que l'effet le statut "dominant" dépend du choix de la contrainte, ce qui rend sont interprétation plus difficile. L'effet d'un niveau est toujours, d'une façon ou d'une autre, défini par rapport aux effets des autres niveaux. La comparaison des différents statuts est l'objet du paragraphe 4.2.4.

Table d'analyse de la variance

Pour mesurer cette influence, on décompose la variabilité totale en variabilité expliquée par le facteur et variabilité résiduelle. On associe une somme de carrés associée à chacune de ces variabilités : la somme des carrés totaux (SCT), la somme des carrés due au facteur (ou au modèle, SCM) et la somme des carrés résiduelle (SCR). Le tableau 4.3 rappelle la définition des éléments des colonnes `df`, `Sum of Squares` et `Mean Square` de la table 4.4 dans le cas de l'analyse de la variance à un facteur à I niveaux pour n données.

La formule d'analyse de la variance nous assure que $SCT = SCM + SCR$. Ces définitions sont cohérentes avec les définitions des sommes de carrés données au chapitre 2, paragraphe 2.2.4. Il suffit pour s'en convaincre de remarquer que la moyenne du groupe i $Y_{i\bullet}$ est le diamètre prédit par le modèle (4.1) ou (4.2) pour tout arbre de ce groupe : $\widehat{Y}_{ik} = Y_{i\bullet}$.

Source	Degrés de liberté	Somme de carrés	Carré moyen
Modèle	$I-1$	$SCM = \sum_i \sum_k (Y_{i\bullet} - Y_{\bullet\bullet})^2$	$SCM/(I-1)$
Résidu	$n-I$	$SCR = \sum_i \sum_k (Y_{ik} - Y_{i\bullet})^2$	$SCR/(n-I)$
Total	$n-1$	$SCT = \sum_i \sum_k (Y_{ik} - Y_{\bullet\bullet})^2$	$SCT/(n-1)$

TABLE 4.3 – Définition des sommes de carrés et carrés moyen dans le modèle d'analyse de la variance à un facteur.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	507.68220	253.84110	1.16	0.3190
Error	101	22187.25771	219.67582		
Corrected Total	103	22694.93990			

R-Square	Coeff Var	Root MSE	Diametre Mean
0.022370	49.77983	14.82146	29.77404

TABLE 4.4 – Table d'analyse de la variance décrivant l'effet du statut sur le diamètre des alisiers.

Notion de réduction. On peut exprimer la somme de carrés dus au modèle en terme de réduction. On a :

$$SCM = R(\alpha/\mu)$$

puisque SCM représente la diminution de la sommes des carrés résiduels quand on introduit l'effet statut α en plus de la constante μ.

Ajustement du modèle. On observe que la somme des carrés due au modèle représente 2% de la variabilité totale :

$$\texttt{R} - \texttt{square} = R^2 = SCM/SCT = 0.02.$$

L'ajustement du modèle est donc très mauvais, ce qui ne signifie pas que le statut n'a pas d'effet sur le diamètre, mais seulement qu'on ne peut espérer fonder une prédiction du diamètre d'un arbre sur la connaissance de son seul statut. En d'autres termes, le diamètre moyen du groupe est un très mauvais prédicteur du diamètre d'un arbre du groupe.

Test de l'effet du statut

Le test de l'effet du statut sur le diamètre s'effectue au moyen de la statistique de Fisher. On veut tester l'hypothèse $\mathbf{H}_0$ selon laquelle *le statut n'a pas d'effet sur le diamètre*, soit

$$\begin{aligned} \mathbf{H}_0 &= \{\alpha_1 = \alpha_2 = \alpha_3 = 0\} = \{Y_{ik} = \mu + E_{ik}\} \\ \text{contre} \qquad \mathbf{H}_1 &= \{\exists i : \alpha_i \neq 0\} = \{Y_{ik} = \mu + \alpha_i + E_{ik}\} \end{aligned} \tag{4.4}$$

La première écriture de cette hypothèse porte sur les valeurs des paramètres : "les effets des différents statuts sont égaux, et donc nuls" alors que la seconde s'exprime en terme de modèle : "le terme α_i dans le modèle (4.1) peut être omis".

Le test l'effet du facteur statut se fait au moyen de la statistique de Fisher qui s'interprète comme un rapport de variance :

$$F = \frac{SCM/(I-1)}{SCR/(n-I)} = \frac{\text{variance expliquée par le statut}}{\text{variance résiduelle}}.$$

La table 4.4 donne une valeur de $F = 1.16$, ce qui indique que la variabilité due au statut est du même ordre de grandeur que la variabilité résiduelle.

Sous l'hypothèse $\mathbf{H}_0$ (4.4), la statistique de Fisher suit une loi de Fisher à $I-1$ et $n-I$ degrés de libertés :

$$F \underset{\mathbf{H}_0}{\sim} \mathcal{F}_{I-1,n-I}.$$

On utilise cette propriété pour calculer la probabilité critique `Pr > F` :

$$\Pr\{\mathcal{F}_{2,101} > 1.16\} = 31.9\%.$$

Conclusion. On obtient un résultat non significatif : le rapport des variances est trop proche de 1 pour qu'on puisse conclure que le statut contribue significativement à expliquer les différences de diamètre entre les alisiers.
Ce résultat ne prouve pas l'absence d'effet du statut sur le diamètre, il signifie que la variabilité individuelle du diamètre des arbres peut tout à fait produire "par hasard" les différences de moyennes observées entre les groupes. Il peut exister un effet statut faible que cette expérience ne permet pas de détecter.

4.2.3 Estimation des paramètres

Paramètres de l'espérance

Les paramètres μ_i du modèle (4.2) ont des estimateurs naturels évidents (qui sont aussi ceux des moindres carrés ou du maximum de vraisemblance) : on estime l'espérance du diamètre dans chaque groupe par le diamètre moyen observé dans ce groupe, soit

$$\hat{\mu}_i = Y_{i\bullet}.$$

L'estimation des paramètres μ et α_i du modèle (4.1) est plus problématique à cause de la non-identifiabilité de ce modèle. On sait (cf Daudin *et al.* (2007)) qu'il faut appliquer une contrainte à ces paramètres pour pouvoir les estimer. On sait aussi que l'interprétation des valeurs estimées dépend fortement de ces contraintes.

Différentes contraintes possibles.

$\alpha_I = 0$ est la contrainte utilisée par le logiciel SAS pour des raisons de simplicité numérique, puisqu'elle revient à supprimer la dernière colonne de la matrice $\mathbf{X}$ qui devient, de ce fait, de plein rang. Elle aboutit aux estimateurs

$$\hat{\mu} = Y_{I\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{I\bullet}.$$

Par construction, on a donc $\hat{\alpha}_I = 0$. Les estimations de α_i s'interprètent comme des écarts à un groupe de référence qui est choisi arbitrairement comme étant le dernier. Le tableau 4.6 donne les estimations obtenues avec cette contrainte, ainsi que leurs écarts types. La note au bas de cette table nous rappelle que ces estimations ne sont pas les seules possibles.

$\sum_i \alpha_i = 0$ est sans doute la contrainte la plus naturelle, puisqu'elle suppose que les effets des différents niveaux se compensent globalement. Si le dispositif est *équilibré* ($n_1 = n_2 = \cdots = n_I$), on aboutit aux estimateurs, naturels eux aussi :

$$\hat{\mu} = Y_{\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{\bullet\bullet}.$$

μ est alors estimé par la moyenne générale et α_i par l'écart entre la moyenne du groupe i et la moyenne générale.
Cependant, dans le cas *déséquilibré* (comme dans l'exemple de ce chapitre), cette contrainte donne

$$\hat{\mu} = \frac{1}{I}\sum_i Y_{i\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - \hat{\mu}.$$

Dans ce cas, μ n'est plus estimé par la moyenne générale, mais par une moyenne qui donne un poids différent aux individus des différents groupes.
L'interprétation des $\hat{\alpha}_i$ est alors plus délicate : ce sont des écarts à une valeur moyenne, qui n'est pas la moyenne générale.

$\sum_i n_i\alpha_i = 0$ donne toujours les estimateurs naturels :

$$\hat{\mu} = Y_{\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{\bullet\bullet}.$$

Les $\hat{\alpha}_i$ sont alors les écarts à la moyenne générale.
La définition de la contrainte montre cependant que ces estimateurs "naturels" donnent en fait plus de poids aux observations issues des petits groupes.

Le tableau 4.5 rappelle les effectifs et moyennes par groupe et donne les estimations des paramètres obtenues avec ces différentes contraintes.

		global	groupe 1	groupe 2	groupe 3
effectif		104	25	15	64
moyenne		29.77	29.22	35.13	28.73
modèle	contrainte	$\hat{\mu}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$
(4.2)	–	–	29.22	35.13	28.73
(4.1)	$\alpha_I = 0$	28.73	+0.49	+6.40	+0.00
(4.1)	$\sum_i \alpha_i = 0$	31.03	−1.81	+4.10	−2.29
(4.1)	$\sum_i n_i \alpha_i = 0$	29.77	−0.55	+5.36	−1.04

TABLE 4.5 – Comparaison des estimations des paramètres μ et α_i avec différentes contraintes. Pour le modèle (4.2), les colonnes $\hat{\alpha_i}$ contiennent les estimations $\hat{\mu}_i$.

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		28.73437500 B	1.85268310	15.51	<.0001
Statut	codomina	0.48562500 B	3.49563548	0.14	0.8898
Statut	dominant	6.39895833 B	4.25176308	1.51	0.1354
Statut	domine	0.00000000 B	.	.	.

```
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used
      to solve the normal equations.  Terms whose estimates are followed by the letter
      'B' are not uniquely estimable.
```

TABLE 4.6 – Estimation des paramètres.

Combinaisons linéaires estimables. On sait (voir Daudin *et al.* (2007), subsection 2.1.4) que, si les estimateurs des paramètres dépendent de la contrainte, certaines combinaisons linéaires des paramètres sont, elles, invariantes. Ces combinaisons sont dites *estimables*. On vérifie facilement que les combinaisons suivantes sont estimables.

$\mathbb{E}(Y_{ik}) = \mu + \alpha_i$ représente l'espérance pour une observation du groupe i. Quelque soit la contrainte, on a

$$\widehat{Y}_{ik} = \widehat{\mu} + \widehat{\alpha}_i = Y_{i\bullet}.$$

Les prédictions fournies par le modèle ne dépendent donc pas de la contrainte, ce qui est rassurant.

$\alpha_1 - \alpha_2$ représente l'écart entre les effets des groupes 1 et 2. Pour toute contrainte, on a

$$\widehat{\alpha}_1 - \widehat{\alpha}_2 = Y_{1\bullet} - Y_{2\bullet}.$$

La comparaison des groupes est donc indépendante de la contrainte.

Paramètre de variance

L'estimateur de la variance résiduelle σ^2 est (cf. Daudin *et al.* (2007), section 2.4)

$$\hat{\sigma}^2 = \frac{SCR}{n-I} = \frac{\sum_i \sum_k (Y_{ik} - Y_{i\bullet})^2}{n-I}. \tag{4.5}$$

Il se retrouve donc directement dans le tableau 4.4, p. 48, à la ligne `Error` et dans la colonne `Mean Square`. On obtient $\hat{\sigma}^2 = 219.7\text{cm}^2$. L'écart type résiduel estimé est donné par le `Root MSE` (pour *Root Mean Square Error*) : on a $\hat{\sigma} = 14.8\text{cm}$.

Ainsi l'écart type intragroupe est d'environ 15cm, ce qui est cohérent avec les valeurs par groupe données au tableau 4.2, p. 45. Le diamètre moyen des arbres est donné dans le tableau 4.2, p. 45 et dans le tableau 4.4, p. 48, à la rubrique `Dependent mean` ; il vaut 29.7cm. La rubrique `C.V` (pour *coefficient de variation*) donne le rapport (en %) entre l'écart type intragroupe et le diamètre moyen : `C.V.` $= 100\hat{\sigma}/Y_{\bullet\bullet}$. Ce rapport vaut 49.8%, ce qui montre encore la forte variabilité du diamètre au sein des groupes et l'impossibilité de prédire avec précision le diamètre d'un arbre seulement à partir de son statut de dominance.

Test sur les paramètres

Le tableau 4.6, p. 51 fournit pour chaque paramètre du modèle (4.1), p. 45 la statistique et la probabilité critique associées au test de l'hypothèse $\mathbf{H}_0 = \{\text{le paramètre est nul}\}$. Cette hypothèse doit être interprétée *en fonction des contraintes* choisies par SAS.
Pour éviter toute confusion, il est utile d'exprimer ces hypothèses en utilisant les paramètres du modèle (4.2), p. 46 :

	paramètre	$\mathbf{H}_0$ (4.1)	$\mathbf{H}_0$ (4.2)
`Intercept`	μ	$\{\mu = 0\}$	$\{\mu_3 = 0\}$
`Statut codomina`	α_1	$\{\alpha_1 = 0\}$	$\{\mu_1 = \mu_3\}$
`Statut dominant`	α_2	$\{\alpha_2 = 0\}$	$\{\mu_2 = \mu_3\}$
`Statut domine`	α_3	$\{\alpha_3 = 0\}$	$\{0 = 0\}$

Le sens biologique des hypothèses exprimées dans le modèle (4.1) dépend de la contrainte choisie alors qu'il n'en dépend pas si elles sont exprimées dans le modèle (4.2).

Interprétation du tableau 4.6, p. 51.

`Intercept :` l'hypothèse $\mathbf{H}_0 = \{\mu = 0\}$ est rejetée (probabilité critique < 0.0001), ce qui signifie que le diamètre moyen des arbres dominés est significativement non nul ! Dans cet exemple, comme dans beaucoup d'autres, ce test ne présente aucun intérêt.

`Statut codomina :` l'hypothèse $\mathbf{H}_0 = \{\alpha_1 = 0\}$ est acceptée (probabilité critique $= 0.89$), ce qui signifie que le diamètre moyen des arbres co-dominants n'est pas significativement différent de celui des arbres dominés.

`Statut dominant :` l'hypothèse $\mathbf{H}_0 = \{\alpha_2 = 0\}$ est acceptée (probabilité critique = 0.14), ce qui signifie que le diamètre moyen des arbres dominants n'est pas significativement différents (au niveau 5%) de celui des arbres dominés.

`Statut dominant :` l'hypothèse $\mathbf{H}_0 = \{\alpha_3 = 0\}$ est acceptée par définition de la contrainte.

Les tests sur les α_i donnés dans le tableau 4.6, p. 51 sont donc des tests de comparaison des différents groupes à un groupe de référence (le dernier). Ce tableau ne donne pas de résultat sur la comparaison des groupes dominants et co-dominants.

4.2.4 Comparaison des groupes de statuts

Comparaison de deux groupes.

On sait (cf Daudin *et al.* (1999), p. 70) que dans un modèle gaussien avec variance homogène, le test de l'hypothèse $\mathbf{H}_0 = \{\mu_i = \mu_j\}$ contre $\mathbf{H}_1 = \{\mu_i \neq \mu_j\}$ est fondé sur la statistique de test

$$T_{ij} = (Y_{i\bullet} - Y_{j\bullet}) \Big/ \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

où $\hat{\sigma}^2$ est la variance (supposées commune) estimée sur les mesures effectuées dans les deux groupes.
On sait aussi que, sous $\mathbf{H}_0$, cette statistique de test suit une loi de student à $n_i + n_j - 2$ degrés de liberté :

$$T_{ij} \underset{\mathbf{H}_0}{\sim} \mathcal{T}_{n_i+n_j-2}.$$

On sait enfin que la puissance de ce test croît avec le nombre de degrés de liberté qui rend compte de la précision avec laquelle sont estimées les espérances μ_i et μ_j.

Modèle d'analyse de la variance. Dans le cadre du modèle d'analyse de la variance, on peut améliorer la puissance de ces tests en utilisant une estimation de la variance fondée sur l'ensemble des groupes. Cette amélioration n'a de sens qu'à cause de l'hypothèse d'homoscédasticité. Pour comparer les groupes i et j, on utilise la même statistique de test T_{ij} mais en utilisant la variance estimée sur l'ensemble des données donnée par l'équation (4.5), p. 52). Ainsi, les données issues des autres groupes contribuent à la définition de la statistique de test.
Dans le cadre du modèle d'analyse de la variance, T_{ij} suit une loi de student à $n-I$ degrés de libertés :

$$T_{ij} \underset{\mathbf{H}_0}{\sim} \mathcal{T}_{n-I}.$$

Application numérique. Dans cet exemple, on a $n = 104$ arbres répartis en $I = 3$ groupes. L'écart type résiduel est estimé par $\hat{\sigma} = 14.8$. On choisit de faire des tests au niveau $\alpha = 5\%$. Le quantile de la loi de student à $n - I = 101$ degrés de liberté vaut $t_{101}(1 - \alpha/2) = 1.98$. On rejette donc l'hypothèse $\mathbf{H}_0$ si $T > 1.98$.

Comparaisons multiples.

On veut maintenant comparer tous les groupes (statuts) deux à deux. En comparant I groupes, on effectue $I(I-1)/2 = 3$ comparaisons. Pour chaque comparaison, on prend une décision (acceptation ou rejet de $\mathbf{H}_0$) aléatoire qui peut correspondre à une erreur. On se trouve confronté ici à un problème de tests multiples : en effectuant chacune des comparaisons à un niveau $\alpha = 5\%$ on encourt globalement un risque α^* de se tromper au moins une fois qui est supérieur à α.

Bonferroni (Dunn) t Tests for Diametre

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	101
Error Mean Square	219.6758
Critical Value of t	2.43451

Comparisons significant at the 0.05 level are indicated by ***.

Statut Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
dominant - codomina	5.913	-5.871	17.698
dominant - domine	6.399	-3.952	16.750
codomina - dominant	-5.913	-17.698	5.871
codomina - domine	0.486	-8.025	8.996
domine - dominant	-6.399	-16.750	3.952
domine - codomina	-0.486	-8.996	8.025

TABLE 4.7 – Comparaison des diamètres moyens des trois groupes d'arbres.

Le tableau 4.7 présente les comparaisons des moyennes des trois groupes. Ce tableau fait référence à la méthode de Bonferroni qui est fondée sur l'inégalité (non démontrée ici) :

$$\alpha^* \leq \alpha \times I(I-1)/2.$$

Cette inégalité dit simplement que le risque de se tromper au moins une fois au total est inférieur au risque de se tromper lors d'une comparaison (α) multiplié par le nombre de comparaisons ($I(I-1)/2$).

Ainsi, si on veut limiter le risque de faire une erreur (de première espèce) à un niveau $\alpha^* = 5\%$, il faut effectuer chacun des tests au niveau

$$\alpha = \frac{\alpha^*}{I(I-1)/2}.$$

Ici, cette règle conduit à effectuer chaque comparaison au niveau $\alpha = 0.05/3 = 1.67\%$. Le quantile de niveau $1 - \alpha/2 = 99.17\%$ de la loi de student est donnée à la ligne "`Critical Value of t`" et vaut 2.43. Seules les statistiques dépassant cette valeur (en valeur absolue) sont significatives.

Le tableau 4.7 fournit également les estimations des différences d'espérances entre les différents groupes, ainsi que les intervalles de confiances pour ces différences. La mention "`Simultaneous 95% Confidence Limits`" signifie que le niveau de confiance pour l'ensemble des 3 intervalles (les 3 autres étant obtenus par symétrie) est de $1 - \alpha^* = 95\%$, ce qui signifie que chacun des intervalles a, en fait, une confiance de $1 - \alpha = 98.33\%$.

Conclusion. Ici, aucune comparaison ne donne lieu à un test significatif. Les diamètres moyens des trois groupes de statuts ne sont pas significativement différents. Même la différence de 6.4 cm observée entre les diamètres moyens des arbres dominés et dominants peut provenir par la variabilité individuelle.

4.2.5 Analyse des résidus

L'analyse des résidus doit être menée pour chaque modèle. Elle a été présentée de façon détaillée au chapitre 2. Des analyses analogues à celles présentées à la section 2.2.2 doivent être menées ici aussi. Elles aboutissent à des résultats satisfaisants.

Répartition spatiale

L'analyse des résidus doit cependant prendre en compte toute l'information disponible afin de traquer tout biais ou effet non pris en compte dans la modélisation.

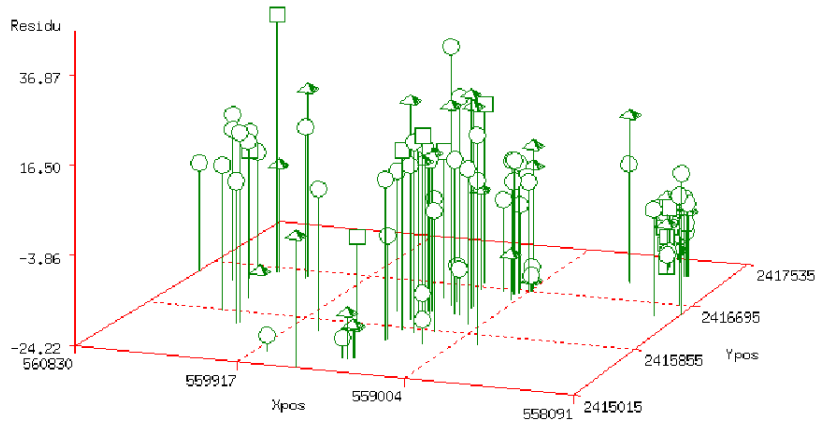

FIGURE 4.2 – Diamètres d'alisiers : répartition spatiale des résidus.

Ainsi, dans cet exemple, nous disposons de la position géographique de chaque arbre étudié au travers des variables `Xpos` et `Ypos`. Le modèle (4.1) suppose que la position

géographique n'a aucun effet (ni en espérance, ni en variance) sur la distribution des résidus. La figure 4.2 présente la répartition des résidus du modèle en fonction de la position géographique de l'arbre. On s'attend, si les hypothèses sont satisfaites, à ce que ce modèle ne présente aucune structure particulière *en ordonnée* (c'est-à-dire le long de l'axe des résidus).

La structure géographique observée sur ce graphe n'est pas contraire aux hypothèses du modèle. On n'observe par contre aucune structure notable en ordonnée (croissance le long d'un axe, écarts de variance *etc.*), ce qui est conforme aux hypothèses du modèle. L'analyse des résidus ne met donc pas en évidence un effet géographique.

4.3 Programme SAS

Lecture et affichage des données. La table 4.1, p. 44 est produite par les instructions suivantes.

```
data ARBRES;
     infile 'Arbre.don' firstobs=2;
     input Xpos Ypos Diametre Statut$;
proc Print data=ARBRES;
run;
```

data ARBRES permet de définir le tableau SAS `ARBRES` à partir des données lues dans le fichier `Arbre.don'`.

proc Print affiche le contenu du tableau `ARBRES`.

Graphiques et statistiques élémentaires. La table 4.2, p. 45 et la figure 4.1, p. 46 sont produites par les instructions suivantes :

```
proc BoxPlot data=ARBRES;
     plot Diametre * Statut;
proc Means data=ARBRES;
     var Diametre;
proc Sort data=ARBRES;
     by Statut;
proc Means data=ARBRES;
     var Diametre;
     by Statut;
run;
```

proc BoxPlot affiche les boîtes à moustaches par groupe (statut).

proc Means donne l'effectif, la moyenne, l'écart type, le minimum et le maximum des diamètres pour l'ensemble de l'échantillon.

proc Sort trie le tableau ARBRES par statut en vue de la procédure suivante.

proc Means / by Statut donne les statistiques élémentaires par groupe de statut.

Analyse de la variance. Les tables 4.4, p. 48, 4.6, p. 51 sont produites par les instructions suivantes :

```
proc GLM data=ARBRES;
     class Statut;
     model Diametre = Statut / solution;
     means Statut / bon;
     output out=ANOVA p=Predite r=Residu;
run;
```

proc GLM est la procédure générale pour le modèle linéaire.

model est l'instruction qui définit le modèle (4.1), p. 45.

solution permet d'obtenir les estimations des paramètres du modèle.

means effectue la comparaison des moyennes des différents statuts et produit le tableau 4.7, p. 54. L'option `bon` permet de choisir la méthode de Bonferroni pour les tests multiples.

output permet de récupérer dans le tableau `ANOVA` les prédictions $\hat{Y}_{ik}$ (**p**) et les résidus $\hat{E}_{ik}$ (**r**).

Analyse des résidus. Le graphique présenté à la figure 4.2, p. 55 est produit par les instructions suivantes :

```
data ANOVA;
     set ANOVA;
     if Statut = 'domine' then Forme = 'balloon';
     if Statut = 'codominant' then Forme = 'pyramid';
     if Statut = 'dominant' then Forme = 'square';
proc G3D data=ANOVA;
     scatter Xpos * Ypos = Residu / shape=Forme;
run;
```

data ANOVA : cette étape permet de créer une variable `Forme` qui gouverne les symboles de la figure.

Chapitre 5

Analyse de la variance à deux facteurs : cas équilibré

5.0.1 Présentation

Objectif et dispositif expérimental

Objectif. On souhaite étudier l'effet du niveau de fertilisation et de la rotation de culture sur le poids des grains de colza. On compare pour cela $I = 2$ niveaux de fertilisations notées 1 pour faible et 2 pour fort et $J = 3$ types de rotation maïs / blé / colza / blé :

A : sans enfouissement de paille,

B : avec enfouissement de paille,

C : avec 4 années de prairie temporaire entre chaque succession sans enfouissement de paille.

Dans cette subsection, on désignera par *traitement* la combinaison Fertilisation*Rotation. Le traitement "1B" désigne la combinaison de la fertilisation "1" et de la rotation "B". On compare ici $2 \times 3 = 6$ traitements.

Dispositif. On dispose de mesures effectuées sur 60 parcelles. Chacune des 6 combinaisons Rotation*Fertilisation a été appliquée sur $K = 10$ parcelles. Un tel dispositif est appelé "plan factoriel *complet*" car il permet de croiser tous les niveaux des deux facteurs.

Description des données

Le poids moyen des grains a été mesuré sur chaque parcelle ; le tableau 5.1 présente un extrait des données ainsi obtenues.

Dispositif équilibré. Le dispositif est ici *équilibré* ce qui signifie que chaque traitement a été reçu par le même nombre de parcelles K :

$$\forall (i,j), \quad n_{ij} \equiv K.$$

Obs	Fertilisation	Rotation	Pds Grains
1	1	A	27.6
2	1	A	16.3
3	1	A	11.4
4	1	A	38.2
5	1	A	38.1
../..			
56	2	C	23.4
57	2	C	44.1
58	2	C	35.3
59	2	C	31.6
60	2	C	25.8

TABLE 5.1 – Extrait du tableau données portant sur le poids des grains.

On a donc au total $n = I \times J \times K = 60$ observations. Chaque fertilisation i a été appliquée à $n_{i+} = J \times K = 30$ parcelles, chaque rotation a été appliquée à $n_{+j} = I \times K = 20$ parcelles.

Un dispositif équilibré est un dispositif *orthogonal*, c'est à dire un dispositif dans lequel on peut décomposer de façon unique les différents effets. Il vérifie en effet la condition (cf. Daudin *et al.* (2007), chapitre 4)

$$\forall i, j : \quad n_{ij} = \frac{n_{i+} n_{+j}}{n}.$$

La propriété d'orthogonalité sera notamment utilisée à la subsection 5.0.2.

Variabilité du poids des grains. Les tableaux 5.2 et 5.3 donnent les moyennes et écarts types du poids des grains en fonction de la rotation et de la fertilisation et la figure 5.1 représente leur distribution sous forme de boîtes à moustache.

Obs	Fertilisation	Rotation	_TYPE_	_FREQ_	Pds Grains	Ecart Type
1	1	A	0	10	24.11	8.61529
2	1	B	0	10	24.00	7.36870
3	1	C	0	10	28.64	5.86273
4	2	A	0	10	15.81	7.43811
5	2	B	0	10	19.84	8.27341
6	2	C	0	10	31.75	7.24542

TABLE 5.2 – Moyennes et écarts types des poids de grains par traitement.

On observe des différences de moyennes assez fortes (de 15.8 pour le traitement 2A à 31.75 pour le traitement 2C). La figure 5.1 et les écarts types du tableau 5.2 montrent que la variabilité est sensiblement la même dans les différents groupes. Cette observation

		rotation			
		$j=1$	$j=2$	$j=3$	total
fertilisation	$i=1$	$y_{11\bullet}=24,11$	$y_{12\bullet}=24,00$	$y_{13\bullet}=28,64$	$y_{1\bullet\bullet}=25,58$
	$i=2$	$y_{21\bullet}=15,81$	$y_{22\bullet}=19,84$	$y_{23\bullet}=31,75$	$y_{2\bullet\bullet}=22,47$
	total	$y_{\bullet 1\bullet}=19,96$	$y_{\bullet 2\bullet}=21,92$	$y_{\bullet 3\bullet}=30,20$	$y_{\bullet\bullet\bullet}=24,03$

TABLE 5.3 – Tableau des moyennes des poids de grains de colza en fonction de la fertilisation et de la rotation.

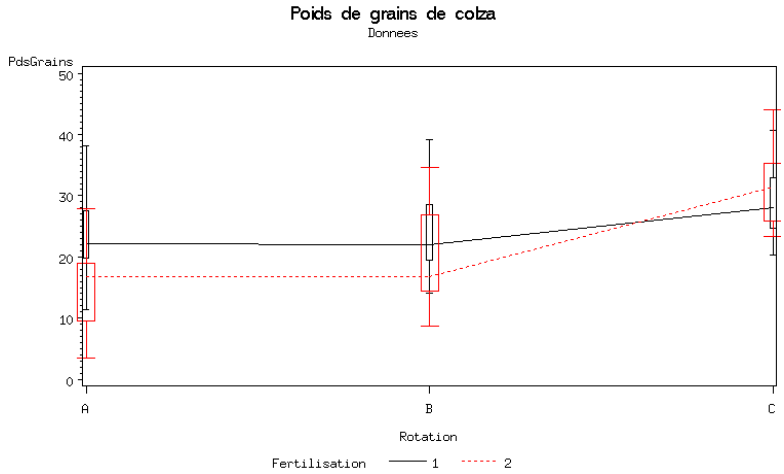

FIGURE 5.1 – Répartition du poids des grains en fonction de la fertilisation et de la rotation.

conforte l'hypothèse d'homoscédasticité qui est faite dans les différents modèles présentés dans cette subsection.

5.0.2 Analyse de la variance à 2 facteurs avec interaction

Écriture du modèle

Modèle sur le traitement. L'objectif étant de décrire l'effet conjoint des deux facteurs (fertilisation et rotation) sur le poids des grains, il nous faut poser un modèle faisant apparaître ces deux effets. Pour cela, on peut généraliser le modèle 4.2, p. 46 au traitement, c'est à dire à la combinaison fertilisation*rotation. On obtient ainsi le modèle

$$Y_{ijk} \sim \mathcal{N}(\mu_{ij}, \sigma^2), \quad \{Y_{ijk}\} \text{ indépendants} \tag{5.1}$$

qui prévoit

- une espérance spécifique μ_{ij} pour chaque traitement (ij)
- et une variance intra-traitement σ^2 commune à tous les traitements.

Décomposition des effets. Le modèle 5.1 est tout à fait général mais ne fait pas apparaître explicitement les effets des différents facteurs. L'écriture traditionnelle du modèle d'analyse de la variance est donc la suivante :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk} \qquad \{E_{ik}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2) \tag{5.2}$$

où

- α_i représente l'*effet principal* de la fertilisation,
- β_j représente l'*effet principal* de la rotation,
- γ_{ij} est le terme d'*interaction*,
- σ^2 est la variance résiduelle.

Le modèle 5.2 est plus explicite mais, comme le modèle d'analyse de la variance à un facteur 4.1, p. 45, il n'est pas identifiable, c'est à dire que l'estimation de ses paramètres nécessite le recours à un système de contraintes.

Graphe d'interaction Le terme d'interaction est un des apports important de l'analyse de la variance à deux facteurs par rapport à l'analyse à un facteur. Il permet de supposer que les effets des deux facteurs ne sont pas seulement *additifs* et de prendre en compte l'effet spécifique du traitement (ij), au-delà de la somme $\alpha_i + \beta_j$ des effets principaux.

En l'absence $(\forall (i,j), \gamma_{ij} = 0)$, l'écart entre les deux niveaux de fertilisation doit être le même, quelle que soit la rotation :

$$\mu_{1j} - \mu_{2j} = \text{constante}.$$

La figure 5.1, p. 60 présente le graphe d'interaction pour l'exemple du poids des grains de colza. En l'absence d'interaction ces lignes joignant les moyennes par traitement doivent être "parallèles". On observe ici qu'elles ne le sont pas, elles se croisent même entre les rotations 2 et 3. On doit donc, *a priori*, prévoir un terme d'interaction γ_{ij} dans le modèle. Le graphe d'interaction ne permet cependant pas de se prononcer quant à la significativité de cet effet.

Écriture matricielle On peut écrire le modèle 5.2 sous la forme matricielle :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}, \qquad \mathbf{E} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

avec

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_{111} \\ \vdots \\ Y_{11K} \\ Y_{121} \\ \vdots \\ Y_{12K} \\ Y_{131} \\ \vdots \\ Y_{13K} \\ Y_{211} \\ \vdots \\ Y_{21K} \\ Y_{221} \\ \vdots \\ Y_{22K} \\ Y_{231} \\ \vdots \\ Y_{23K} \end{bmatrix}, \quad \underset{n\times(I+1)(J+1)}{\mathbf{X}} = \left[\begin{array}{c|cc|ccc|ccccc} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right], \tag{5.3}$$

et

$$\underset{(I+1)(J+1)\times 1}{\boldsymbol{\Theta}} = \left[\begin{array}{c} \mu \\ \hline \alpha_1 \\ \alpha_2 \\ \hline \beta_1 \\ \beta_2 \\ \beta_3 \\ \hline \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \end{array}\right], \quad \underset{n\times 1}{\mathbf{E}} = \begin{bmatrix} E_{111} \\ \vdots \\ E_{11K} \\ E_{121} \\ \vdots \\ E_{12K} \\ E_{131} \\ \vdots \\ E_{13K} \\ E_{211} \\ \vdots \\ E_{21K} \\ E_{221} \\ \vdots \\ E_{22K} \\ E_{231} \\ \vdots \\ E_{23K} \end{bmatrix}. \tag{5.4}$$

Les lignes verticales dans $\mathbf{X}$ correspondent aux lignes horizontales dans $\boldsymbol{\Theta}$: elles séparent les colonnes associées respectivement à la constante μ, aux effets fertilisation α_i, aux effets rotation β_j et aux effets de l'interaction γ_{ij}.

Analyse de la variance

Table d'analyse de la variance La table d'analyse de la variance 5.4 constitue le premier outil pour évaluer l'effet des deux facteurs et de leur interaction sur le poids des grains.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1659.821500	331.964300	5.87	0.0002
Error	54	3052.471000	56.527241		
Corrected Total	59	4712.292500			

R-Square	Coeff Var	Root MSE	PdsGrains Mean
0.352232	31.29432	7.518460	24.02500

TABLE 5.4 – Table d'analyse de la variance du modèle à deux facteurs avec interaction.

Les sommes de carrés qui y figurent sont données dans le tableau 5.5 ; elles sont définies de façon analogue à celles présentées dans le tableau 4.3, p. 48 puisque, ici, $\widehat{Y}_{ijk} = Y_{ij\bullet}$. La définition de ces sommes de carrés montre que la table d'analyse de la variance permet de mesurer l'effet du traitement, c'est-à-dire de la combinaison de la fertilisation et de la rotation. L'hypothèse $\mathbf{H}_0$ testée peut ici est

$$\mathbf{H}_0 = \{Y_{ijk} = \mu + E_{ijk}\}.$$

Elle s'exprime également en termes de paramètres sous la forme

$$\begin{array}{rcl} \mathbf{H}_0 & = & \{\mu_{11} = \cdots = \mu_{23} = \mu\} \\ \text{contre } \mathbf{H}_1 & = & \{\exists (i,j,i',j') : \mu_{ij} \neq \mu_{i'j'}\} \end{array} \quad \text{pour le modèle (5.1),}$$

$$\begin{array}{rcl} \mathbf{H}_0 & = & \{\forall i : \alpha_i = 0, \forall j : \beta_j = 0, \forall (i,j) : \gamma_{ij} = 0\} \\ \text{contre } \mathbf{H}_1 & = & \{\exists (i,j) : \alpha_i \neq 0 \text{ ou } \beta_j \neq 0 \text{ ou } \gamma_{ij} \neq 0\} \end{array} \quad \text{pour le modèle (5.2).}$$

Cette analyse est donc exactement équivalente à une analyse de la variance *à un facteur* à $IJ = 6$ niveaux.

Réduction. La somme des carrés due au modèle mesure la diminution des résidus quand on ajoute les effets fertilisation (α), rotation (β) et l'interaction (γ) à la constante μ. On a donc :

$$SCM = R(\alpha, \beta, \gamma/\mu).$$

Source	Degrés de liberté	Somme de carrés	Carré moyen
Modèle	$IJ-1$	$SCM=\sum_i\sum_j\sum_k (Y_{ij\bullet}-Y_{\bullet\bullet\bullet})^2$	$SCM/(IJ-1)$
Résidu	$n-IJ$ $=IJ(K-1)$	$SCR=\sum_i\sum_j\sum_k (Y_{ijk}-Y_{ij\bullet})^2$	$SCR/[IJ(K-1)]$
Total	$IJK-1$	$SCT=\sum_i\sum_i\sum_k (Y_{ijk}-Y_{\bullet\bullet\bullet})^2$	$SCT/(IJK-1)$

TABLE 5.5 – Définition des sommes de carrés et carrés moyen dans le modèle d'analyse de la variance à deux facteurs avec interaction.

Interprétation. La statistique de Fisher vaut 5.87, ce qui signifie que la variabilité expliquée par le traitement est presque 6 fois supérieure à la variabilité résiduelle. L'effet conjoint de la fertilisation et de la rotation est significatif puisque la probabilité critique vaut $2\ 10^{-4}$. La fertilisation, la rotation ou leur interaction ont un effet significatif sur le poids des grains.

On remarque que l'ajustement du modèle n'est pas bon ($R^2 = 35.2\%$). Ceci n'est pas contradictoire avec l'effet très significatif du traitement. Ceci signifie seulement que, si la combinaison de la rotation et de la fertilisation a un effet très fort sur le poids des grains de colza, on ne peut pas espérer fonder une prédiction précise de ce poids en se fondant seulement sur le traitement. La variabilité résiduelle, c'est-à-dire intra-traitement ($\hat{\sigma^2} = 56.5$, `C.V. = 31.3%`) est trop forte pour permettre une telle prédiction.

Décomposition des sommes de carrés La table d'analyse de la variance 5.4, p. 63 ne permet pas de décomposer l'effet du traitement en effets respectifs des deux facteurs et de leur interaction. Pour obtenir évaluer séparément ces différents effets, il faut décomposer la somme des carrés du modèle en sommes de carrés dues des effets principaux et de l'interaction.
Le dispositif étudié ici étant orthogonal, cette décomposition est unique. En notant A le facteur Fertilisation, B le facteur Rotation et I leur interaction, on a

$$\underbrace{\sum_i\sum_j n_{ij}(y_{ij\bullet}-y_{\bullet\bullet\bullet})^2}_{SCM=1659.8} =$$
$$\underbrace{\sum_i n_{i+}(y_{i\bullet\bullet}-y_{\bullet\bullet\bullet})^2}_{SCA=145.7} + \underbrace{\sum_j n_{+j}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2}_{SCB=1180.5} + \underbrace{\sum_i\sum_j n_{ij}(y_{ij\bullet}-y_{i\bullet\bullet}-y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2}_{SCI=333.6}$$

Le tableau 5.6 donne les valeurs des sommes de carrés ainsi que les tests des effets de chacun de ces facteurs. On retrouve bien ici que grâce à l'orthogonalité du dispositif, la somme des sommes de carrés des différents effets donne la somme des carrés du modèle.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fertilisation	1	145.704167	145.704167	2.58	0.1142
Rotation	2	1180.483000	590.241500	10.44	0.0001
Fertilisati*Rotation	2	333.634333	166.817167	2.95	0.0608

TABLE 5.6 – Effets des différents facteurs dans à deux facteurs avec interaction.

Réductions. Les sommes de carrés associés à chaque effet peuvent s'exprimer en terme de réduction, puisqu'elles mesurent la diminution des résidus lors de l'introduction de l'effet dans le modèle. Du fait de l'orthogonalité du dispositif, les valeurs des réductions ne dépendent que de l'effet considéré et pas du modèle de référence. On a donc :

$$\begin{aligned} SCA &= R(\alpha/\beta,\gamma,\mu) = R(\alpha/\beta,\mu) = R(\alpha/\mu), \\ SCB &= R(\beta/\alpha,\gamma,\mu) = R(\beta/\alpha,\mu) = R(\beta/\mu), \\ SCI &= R(\gamma/\alpha,\beta,\mu). \end{aligned} \tag{5.5}$$

Tests des différents effets La décomposition des sommes de carrés permet de tester des hypothèses plus précises que l'hypothèse générale testée dans la table d'analyse de la variance du tableau 5.4, p. 63. Elle permet notamment de tester de façon spécifique l'effet de chacun des facteurs. Le tableau 5.7 rappelle les hypothèses et les formules des statistiques de tests utilisées dans le tableau 5.6.

	Hypothèse	Statistique de test F	Loi de F sous $\mathbf{H}_0$
$A =$ Fertilisation	$\mathbf{H}_0(A) = \{\forall i : \alpha_i = 0\}$	$F_A = \dfrac{SCA/(I-1)}{SCR/IJ(K-1)}$	$\mathcal{F}_{I-1,IJ(K-1)}$
$B =$ Rotation	$\mathbf{H}_0(B) = \{\forall j : \beta_j = 0\}$	$F_B = \dfrac{SCB/(J-1)}{SCR/IJ(K-1)}$	$\mathcal{F}_{J-1,IJ(K-1)}$
$I =$ Fertilisation *Rotation	$\mathbf{H}_0(I) = \{\forall (i,j) : \gamma_{ij} = 0\}$	$F_I = \dfrac{SCI/(I-1)(J-1)}{SCR/IJ(K-1)}$	$\mathcal{F}_{(I-1)(J-1),IJ(K-1)}$

TABLE 5.7 – Hypothèses, statistiques de test et lois pour le test des effets des différents facteurs sur le poids des grains de colza.

Écriture des hypothèses en termes de modèles. On peut vouloir traduire les hypothèses du tableau 5.7 en termes de modèles. Ainsi, l'hypothèse concernant l'interaction est équivalente à

$$\mathbf{H}_0(I) = \{Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}\},$$

c'est-à-dire à un modèle sans interaction.

L'hypothèse concernant l'effet principal de la fertilisation ne se traduit pas de façon aussi simple. En effet le modèle

$$Y_{ijk} = \mu + \beta_j + \gamma_{ij} + E_{ijk} \tag{5.6}$$

est équivalent au modèle complet (5.2), p. 61. Il suffit pour s'en convaincre de remarquer que la matrice $\mathbf{X}$ correspondant au modèle (5.6) à $1(\mu) + 3\ (\beta_j) + 6\ (\gamma_{ij})$ =10 colonnes (au lieu de 12 pour le modèle complet), mais qu'elle engendre le même sous-espace puisque les colonnes associées à l'effet principal (α_i) de la fertilisation sont engendrées par les colonnes de l'interaction. Ainsi le modèle 5.6 n'est pas un sous modèle du modèle complet, il est égal au modèle complet.
Cette remarque justifie théoriquement la pratique qui consiste à ne jamais supprimer un effet principal (ici α_i) d'un modèle dans lequel subsiste une interaction à laquelle il participe (ici γ_{ij}). Nous donnerons une interprétation de cette règle dans le paragraphe suivant.

On peut par contre exprimer l'hypothèse d'absence d'effet de la fertilisation en reprenant le modèle (5.1), p. 61 sous la forme

$$\mathbf{H}_0(A) = \{\mu_{1\bullet} = \mu_{2\bullet}\}, \qquad \text{avec } \mu_{i\bullet} = \frac{1}{J}\sum_j \mu_{ij},$$

soit, littéralement,

$$\mathbf{H}_0(A) = \left\{ \begin{array}{l} \text{En moyenne sur l'ensemble des rotations étudiées dans cette expé-} \\ \text{rience, le poids moyen des grains ne varie pas entre les fertilisations.} \end{array} \right\}$$

Cette définition est cohérente avec la définition de la somme de carrés SCA qui mesure la dispersion des moyennes par fertilisation $y_{i\bullet\bullet}$ (toutes rotations confondues) autour de la moyenne générale $y_{\bullet\bullet\bullet}$.

Interprétation.

Effet de l'interaction : Le carré moyen de l'interaction : $CMI = SCI/IJ(K-1)$ vaut 166.8 et est $F_I = 2.95$ fois supérieur à la variance résiduelle $\hat{\sigma}^2 = SCR/IJ(K-1) =$ 56.5.

La probabilité critique associée à cette statistique vaut 6%. L'effet de l'interaction n'est donc pas significatif au niveau 5% et on peut accepter un modèle sans interaction.

De façon plus nuancée on peut dire que cet effet est faiblement significatif. Une expérience avec plus de répétitions donnerait un test plus puissant qui permettrait sans doute de détecter un effet significatif.

Effet de la fertilisation : La statistique de Fisher F_A vaut 2.58, ce qui n'est pas significatif au niveau 5% (probabilité critique) 11.4%). On peut donc conclure à l'absence d'effet moyen de l'interaction.

On a vu plus haut qu'on ne peut pour autant pas supprimer l'effet principal du modèle avec interaction. En terme d'interprétation, cela reviendrait à dire que la fertilisation n'a pas d'effet sur le poids des grains, mais qu'elle interagit avec la rotation, ce qui n'a pas de sens.
A ce stade du raisonnement on ne prend donc pas de décision quant à cet effet. Si on choisit de supprimer l'interaction, l'analyse du modèle sans interaction amènera sans doute à le supprimer. Si on choisit de maintenir l'interaction, aucun effet ne peut être retranché du modèle.

Effet de la rotation : L'effet de la rotation est nettement significatif (probabilité critique $= 10^{-4}$). La comparaison des statistiques de Fisher nous montre qu'il est près de 4 fois plus fort que l'effet de la fertilisation ou de l'interaction. C'est clairement l'effet majoritaire de ce modèle.

Estimation des paramètres

Comme en analyse de la variance à un facteur, l'estimation des paramètres du modèle 5.2, p. 61 suppose le recours à un système de contraintes puisque ce modèle n'est pas identifiable. Il faut ici appliquer $I + J + 1$ contraintes indépendantes. Là encore le choix des contraintes est arbitraire. On rappelle ici les estimateurs obtenus avec 2 systèmes de contraintes usuelles.

Contraintes naturelles. On impose à la somme des paramètres associés à chaque effet d'être nul, soit

$$\sum_i \alpha_i = 0, \qquad \sum_j \beta_j = 0, \qquad \forall i : \sum_j \gamma_{ij} = 0 \qquad \forall j : \sum_i \gamma_{ij} = 0.$$

On obtient alors les estimateurs naturels

$$\hat{\mu} = Y_{\bullet\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet}, \qquad \hat{\beta}_j = Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet}, \qquad \hat{\gamma}_{ij} = Y_{ij\bullet} - Y_{i\bullet\bullet} - Y_{\bullet j\bullet} + Y_{\bullet\bullet\bullet},$$

qui s'interprètent comme des écarts entre moyennes. On donnera une explication plus intuitive de l'estimateur $\hat{\gamma}$ dans la subsection 5.0.3.
Les valeurs des estimations se déduisent simplement du tableau 5.3, p. 60 :

		rotation			
		$j = 1$	$j = 2$	$j = 3$	total
fertilisation	$i = 1$	$\hat{\gamma}_{11} = +2.59$	$\hat{\gamma}_{12} = +0.52$	$\hat{\gamma}_{13} = -3.11$	$\hat{\alpha}_1 = +1.56$
	$i = 2$	$\hat{\gamma}_{21} = -2.59$	$\hat{\gamma}_{22} = -0.52$	$\hat{\gamma}_{23} = +3.11$	$\hat{\alpha}_2 = -1.56$
	total	$\hat{\beta}_1 = -4.07$	$\hat{\beta}_2 = -2.11$	$\hat{\beta}_3 = +6.17$	$\hat{\mu} = +24,03$

On peut d'ailleurs ré-exprimer les sommes de carrés SCA, SCB et SCI en fonction des estimations :

$$SCA = J \times K \sum_i \hat{\alpha}_i^2, \qquad SCB = I \times K \sum_j \hat{\beta}_j^2, \qquad SCI = K \sum_i \sum_j \hat{\gamma}_{ij}^2$$

ce qui montre bien les sommes de carrés donnent une mesure de l'amplitude de chaque effet.

Contraintes SAS. Elles annulent tous les paramètres associés au dernier niveau de chaque facteur, soit

$$\alpha_I = 0, \qquad \beta_J = 0, \qquad \forall i : \gamma_{iJ} = 0, \qquad \forall j : \gamma_{Ij} = 0.$$

On rappelle que ce système de contraintes est avantageux du point de vue des calculs. En effet, elles reviennent à supprimer des colonnes de la matrice $\mathbf{X}$ afin qu'elles soient de plein rang. Les matrices $\mathbf{X}$ et $\mathbf{\Theta}$ données par les équations (5.3), p. 62 et (5.4), p. 62 sont remplacées par :

$$\underset{n\times IJ}{\mathbf{X}} = \left[\begin{array}{c|c|cc|cc} 1 & 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \qquad \underset{IJ\times 1}{\mathbf{\Theta}} = \left[\begin{array}{c} \mu \\ \hline \alpha_1 \\ \hline \beta_1 \\ \beta_2 \\ \hline \gamma_{11} \\ \gamma_{12} \end{array}\right].$$

Ces contraintes ne sont cependant pas forcément avantageuses du point de vue de l'interprétation. On obtient les estimateurs suivants :

$$\hat{\mu} = Y_{IJ\bullet}, \qquad \hat{\alpha}_i = Y_{iJ\bullet} - Y_{IJ\bullet}, \qquad \hat{\beta}_j = Y_{Ij\bullet} - Y_{IJ\bullet}, \qquad \hat{\gamma}_{ij} = Y_{ij\bullet} - Y_{iJ\bullet} - Y_{Ij\bullet} + Y_{IJ\bullet}.$$

L'essentiel est de bien noter que les indices finaux I et J interviennent dans tous les estimateurs. Ainsi la comparaison des fertilisations se fait dans la dernière rotation, de même que la comparaison des rotations se fait dans la dernière fertilisation. La combinaison fertilisation 2 * rotation C joue ici le rôle de référence. Ici, compte tenu du dispositif, cette contrainte n'a pas vraiment de sens.
Le tableau 5.8 donne les valeurs des estimations correspondantes. Il faut interpréter les tests associés à chacun de ces paramètres en fonction des contraintes choisies par SAS. Ainsi, l'hypothèse $\mathbf{H}_0 = \{\beta_2 = 0\}$ testée sur la ligne `Rotation B` s'exprime littéralement "*Le poids moyen des grains obtenu avec la rotation B est le même que celui obtenu avec la rotation C pour la fertilisation 2*".

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		31.75000000 B	2.37754581	13.35	<.0001
Fertilisation	1	-3.11000000 B	3.36235753	-0.92	0.3591
Fertilisation	2	0.00000000 B	.	.	.
Rotation	A	-15.94000000 B	3.36235753	-4.74	<.0001
Rotation	B	-11.91000000 B	3.36235753	-3.54	0.0008
Rotation	C	0.00000000 B	.	.	.
Fertilisati*Rotation	1 A	11.41000000 B	4.75509162	2.40	0.0199
Fertilisati*Rotation	1 B	7.27000000 B	4.75509162	1.53	0.1321
Fertilisati*Rotation	1 C	0.00000000 B	.	.	.
Fertilisati*Rotation	2 A	0.00000000 B	.	.	.
Fertilisati*Rotation	2 B	0.00000000 B	.	.	.
Fertilisati*Rotation	2 C	0.00000000 B	.	.	.

TABLE 5.8 – Estimations de paramètres du modèle d'analyse de la variance à 2 facteurs avec interaction pour le poids des grains de colza.

Prédictions. La théorie du modèle linéaire montre que les prédictions $\hat{Y}_{ijk}$ ne dépendent pas du système de contraintes choisies (cf. Daudin *et al.* (2007), subsection 2.1.4). On vérifie facilement que pour les deux systèmes de contraintes présentés ici, on a

$$\hat{Y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = Y_{ij\bullet},$$

ce qui signifie que le poids des grains prédit pour la fertilisation i et la rotation j est simplement le poids moyen observé dans cette combinaison. Il n'y a donc que 6 valeurs prédites possibles (2 fertilisations $\times$ 3 rotations) ; ces valeurs donnent les abscisses des 6 colonnes observées dans la figure 5.2, p. 73.

Comparaison des traitements La comparaison des traitements est un des objectifs classiques de l'analyse de la variance. On vient de voir qu'il peut être difficile de faire cette comparaison en se fondant sur les estimations des paramètres à cause du choix arbitraire de la contrainte.

Effets principaux. Les tableaux 5.9, p. 70 et 5.10, p. 71 présentent des comparaisons de moyennes analogues à celle présentées 4.7, p. 54 sur l'exemple des alisiers.

Fertilisation : On lit dans le premier que les effets moyens des deux fertilisations ne sont pas significativement différents ce qui est cohérent avec l'absence d'effet principal de la fertilisation.

Rotation : Le second tableau nous montre que l'effet principal de la rotation est dû à une différence significative entre la rotation C, d'une part, et les rotations A et B d'autre part. La présence de 4 années de prairie temporaire serait donc la source principale de l'augmentation des poids des grains.

Bonferroni (Dunn) t Tests for PdsGrains

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	54
Error Mean Square	56.52724
Critical Value of t	2.00488
Minimum Significant Difference	3.892

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	Fertilisation
A	25.583	30	1
A			
A	22.467	30	2

TABLE 5.9 – Comparaison des moyennes de poids des grains par fertilisation.

Remarque. On rappelle que ces comparaisons se font dans le cadre d'un modèle avec interaction. Celle-ci est ici faiblement significative, mais si elle était plus forte, il faudrait interpréter ces comparaisons d'effets moyens avec prudence.
Notamment, la "supériorité" de la rotation C peut ne pas être générale : cette rotation peut donner un poids moyen des grains supérieurs en moyenne sur les fertilisations étudiées, et donner de moins bons résultats pour certaines fertilisations spécifiques.

Comparaison des 6 traitements (combinaisons). Le tableau 5.11, p. 72 présente la comparaison des 6 combinaisons sous une forme légèrement différente.

Moyennes : La partie supérieure donne les moyennes $Y_{ij\bullet} = \hat{\mu}_{ij}$ pour chaque combinaison. Ces moyennes sont intitulées "LSMEAN" (pour *"least square means" = moyennes ajustées*, cf. chapitre 6), mais ce sont, en fait, les moyennes usuelles puisque le dispositif est orthogonal.

Probabilités critiques : La partie inférieure donne le tableau des probabilités critiques des tests de comparaisons de moyennes fondés sur les hypothèses nulles de la forme $\mathbf{H}_0 = \{\mu_{ij} = \mu_{i'j'}\}$. Comme on l'a vu à la subsection 4.2.4 du chapitre 4, ces tests sont des tests de student effectués avec la variance résiduelle σ^2 du modèle (5.2), p. 61.

Comparaisons multiples : Les tests présentés dans ce tableau sont des tests de comparaison *deux à deux*. Aucune correction pour la multiplicité des tests n'est faite.

Bonferroni (Dunn) t Tests for PdsGrains

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	54
Error Mean Square	56.52724
Critical Value of t	2.47085
Minimum Significant Difference	5.8746

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	Rotation
A	30.195	20	C
B	21.920	20	B
B			
B	19.960	20	A

TABLE 5.10 – Comparaison des moyennes de poids des grains par rotation.

Si on veut appliquer la méthode de Bonferroni, il faut diviser le niveau choisi (par exemple $\alpha^* = 5\%$) par le nombre de comparaison $IJ(IJ-1)/2 = 15$, et donc comparer chacune des probabilités critiques au seuil

$$\alpha = \alpha^*/[IJ(IJ-1)/] = 0.33\%.$$

En appliquant la règle de Bonferroni, on conclut que les seuls couples de combinaisons significativement différents sont les couples (2A, 1C), (2A, 2C) et (2B, 2C). On peut ainsi reconstruire un système d'accolades analogues à ceux des tableaux 5.9, p. 70 et 5.10, p. 71 de la forme

Groupes			Traitement	Moyenne
A			2A	15.81
A	B		2B	19.84
A	B	C	1B	24.00
A	B	C	1A	24.11
	B	C	1C	28.64
		C	2C	31.75

Analyse des résidus

La figure 5.2, p. 73 présente le graphe des résidus usuels ou les valeurs prédites $\hat{Y}_{ijk}$ apparaissent en abscisse et les résidus $\hat{E}_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$ en ordonnée. L'intérêt de ce

Least Squares Means

Fertilisation	Rotation	PdsGrains LSMEAN	LSMEAN Number
1	A	24.1100000	1
1	B	24.0000000	2
1	C	28.6400000	3
2	A	15.8100000	4
2	B	19.8400000	5
2	C	31.7500000	6

Least Squares Means for Effect Fertilisati*Rotation
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: PdsGrains

i/j	1	2	3	4	5	6
1		0.032715	-1.34727	2.468506	1.269942	-2.27222
		0.9740	0.1835	0.0168	0.2096	0.0271
2	-0.03272		-1.37998	2.435791	1.237227	-2.30493
	0.9740		0.1733	0.0182	0.2214	0.0250
3	1.347269	1.379984		3.815775	2.617211	-0.92495
	0.1835	0.1733		0.0004	0.0115	0.3591
4	-2.46851	-2.43579	-3.81578		-1.19856	-4.74072
	0.0168	0.0182	0.0004		0.2359	<.0001
5	-1.26994	-1.23723	-2.61721	1.198564		-3.54216
	0.2096	0.2214	0.0115	0.2359		0.0008
6	2.272215	2.30493	0.924946	4.740721	3.542158	
	0.0271	0.0250	0.3591	<.0001	0.0008	

TABLE 5.11 – Comparaison des moyennes de poids des grains par combinaison Rotation*Fertilisation. En haut : moyennes de chaque combinaison, en bas : probabilités critiques des comparaisons deux à deux.

graphe vient de la propriété d'indépendance entre les valeurs prédites et les résidus. Si les hypothèses du modèle linéaire sont satisfaites, le nuage de points ainsi obtenu ne doit donc présenter aucune structure particulière.

Cette notion d'*absence de structure* doit cependant être précisée. Ainsi, le nuage présenté dans la figure 5.2 présente clairement une structure horizontale, les points étant répartis en 6 colonnes. Ces 6 colonnes correspondent simplement aux 6 valeurs prédites $\hat{Y}_{ijk}$ possibles, elles-mêmes égales aux moyennes des 6 combinaisons (on le vérifie facilement en se reportant au tableau 5.3, p. 60). Cette structure n'est donc absolument pas contraire aux hypothèses du modèle linéaire.

Les hypothèses portant sur la distribution des résidus seraient remises en cause si on observait une structure verticale (c'est-à-dire le long de l'axe des résidus). Il serait, par

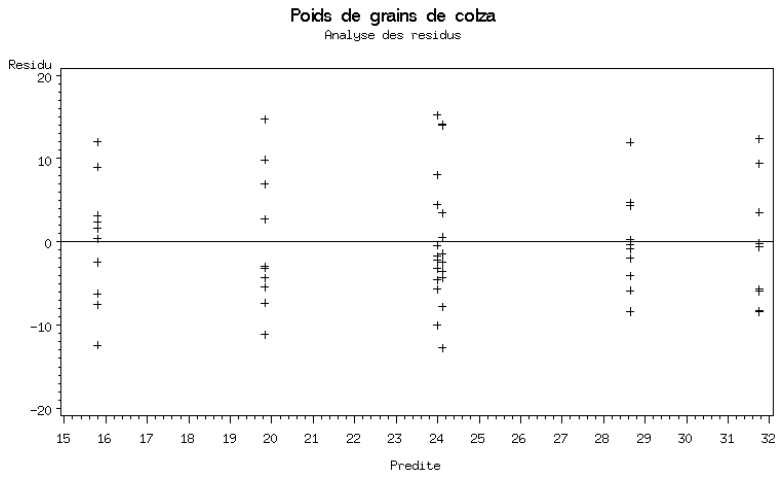

FIGURE 5.2 – Graphe des résidus de l'analyse de la variance à deux facteurs avec interaction pour le poids des grains de colza.

exemple, inquiétant, de voir la dispersion des résidus augmenter avec la valeur prédite car l'hypothèse d'homoscédasticité serait alors violée. Ce n'est pas le cas ici : les 6 colonnes présentent sensiblement la même variabilité, ce qui est cohérent avec les écarts types donnés au tableau 5.3, p. 60.

5.0.3 Études de sous modèles

Analyse de la variance à 2 facteurs sans interaction

Le tableau 5.6, p. 65 montre que l'interaction est faiblement significative (probabilité critique = 6%). On peut donc choisir de se ramener au modèle sans interaction

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk} \tag{5.7}$$

On obtient alors pour la table d'analyse de la variance et les tests des effets des facteurs les résultats présentés dans le tableau 5.12, p. 74.

The ANOVA Procedure

Dependent Variable: PdsGrains

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1326.187167	442.062389	7.31	0.0003
Error	56	3386.105333	60.466167		
Corrected Total	59	4712.292500			

R-Square	Coeff Var	Root MSE	PdsGrains Mean
0.281431	32.36628	7.775999	24.02500

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Fertilisation	1	145.704167	145.704167	2.41	0.1262
Rotation	2	1180.483000	590.241500	9.76	0.0002

TABLE 5.12 – Table d'analyse de la variance et tests des effets du modèle d'analyse de la variance sans interaction du poids des grains de colza.

Sommes de carrés et tests Dans cette subsection, on distingue les sommes de carrés, degrés de libertés et estimations obtenus avec les deux modèles avec les indices suivants :

AB : modèle sans interaction,

ABI : modèle avec interaction.

Comparaison des sommes de carrés. Le dispositif étant orthogonal, les sommes de carrés des différents effets ne changent pas. En notant A la fertilisation et B la rotation,

on a bien :

$$SCA_{AB} = SCA_{ABI} = 145.7, \qquad SCB_{AB} = SCB_{ABI} = 1180.5,$$

Cependant, l'interaction étant supprimée de ce nouveau modèle, sa somme de carrés vient s'ajouter à la somme des carrés résiduelle :

$$SCR_{AB} = SCR_{ABI} + SCI_{ABI} = 3052.5 + 333.6 = 3383.1.$$

L'estimation de la variance résiduelle est donc également différente :

$$\hat{\sigma}^2_{AB} = \frac{SCR_{AB}}{IJK - I - J + 1} = \frac{3386.1}{56} = 60.5$$

contre $\hat{\sigma}^2_{ABI} = 56.6$.

Ce transfert de la somme de carrés de l'interaction vers la résiduelle affecte aussi la somme des carrés du modèle :

$$SCM_{AB} = SCM_{ABI} - SCI_{ABI} = 1659.8 - 333.6 = 1326.2.$$

La sommes des carrés totaux reste, elle, constante ($SCT_{AB} = ACT_{ABI} = 4712.3$, ce qui normal puisqu'elle ne dépend pas du modèle.

Test des effets des facteurs. Bien que les sommes de carrés SCA et SCB soient conservées, les statistiques de Fisher diffèrent du fait de changement du $\hat{\sigma}^2$. On obtient ainsi

- pour la fertilisation : $F = [SCA/(I-1)]/\hat{\sigma}^2_{AB} = 2.41$
- et pour la rotation : $F = [SCB/(J-1)]/\hat{\sigma}^2_{AB} = 9.76$

Ces deux statistiques sont plus faibles que celles obtenues dans le modèle avec interaction (respectivement 2.58 et 10.44, cf. tableau 5.6, p. 65). On obtient donc des résultats similaires à ceux obtenus dans le modèle avec interaction : la rotation a un très fort effet (probabilité critique = $2\ 10^{-4}$) alors que la fertilisation n'en a pas (probabilité critique = 12.6%).

Il est cependant important de remarquer que fait de négliger le terme d'interaction rend les tests des effets principaux légèrement moins significatifs. Les conséquences ne sont ici pas très grandes car l'effet de l'interaction est faible mais on imagine bien que le fait de négliger une interaction forte peut conduire, par le même mécanisme, à des conclusions erronées sur les effets principaux. Nous reviendrons sur les dangers des modèles ne prenant pas tous les effets en compte dans la subsection 5.0.3.

Estimation des paramètres Les estimateurs des paramètres du modèle (5.7), p. 74 s'obtiennent en minimisant la somme des carrés des résiduels

$$\sum_i \sum_j \sum_k (Y_{ijk} - \hat{Y}_{ijk})^2, \qquad \text{avec } \hat{Y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j.$$

Pour un dispositif équilibré comme celui-ci, le minimum est atteint pour

$$\hat{Y}_{ijk} = Y_{i\bullet\bullet} + Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet}.$$

Différentes contraintes. Comme dans le modèle avec interaction, on peut utiliser différents jeux de contraintes pour estimer les paramètres du modèle. Les contraintes naturelles ($\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$) donnent

$$\hat{\mu} = Y_{\bullet\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet}, \qquad \hat{\beta}_j = Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet}$$

alors que les contraintes SAS ($\alpha_I = 0$, $\beta_J = 0$) donnent

$$\hat{\mu} = Y_{I\bullet\bullet} + Y_{\bullet J\bullet} - Y_{\bullet\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet\bullet} - Y_{I\bullet\bullet}, \qquad \hat{\beta}_j = Y_{\bullet j\bullet} - Y_{\bullet J\bullet}.$$

Interprétation du terme d'interaction. L'étude du modèle sans interaction permet de mieux comprendre l'estimateur naturel du paramètre d'interaction γ_{ij} du modèle (5.2), p. 61

$$\hat{\gamma}_{ij} = Y_{ij\bullet} - Y_{i\bullet\bullet} - Y_{\bullet j\bullet} + Y_{\bullet\bullet\bullet}.$$

En comparant les valeurs prédites par les modèles avec (ABI) et sans (AB) interaction :

$$ABI : \hat{Y}_{ijk} = Y_{ij\bullet}, \qquad AB : \hat{Y}_{ijk} = Y_{i\bullet\bullet} + Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet},$$

on voit que $\hat{\gamma}_{ij}$ est la simple différence entre ces deux prédictions. Ce terme comble donc l'écart entre la prédiction du modèle (AB) qui prévoit des effets simplement additifs des deux facteurs avec la moyenne empirique calculée pour chaque combinaison (ij).

Danger d'une analyse séparée des deux facteurs

Pour connaître l'effet de chacun deux facteurs sur le poids des grains, on aurait pu être tenté d'effectuer une analyse de la variance séparée pour chacun des facteurs. L'objectif de cette subsection est de montrer le danger d'une telle approche.

Analyses de la variance à un facteur Pour mesurer l'effet de chacun des deux facteurs, on peut donc écrire des modèles analogues à celui étudié au chapitre 4.

Effet rotation : on note α_i l'effet de la rotation i et on obtient

$$Y_{ijk} = \mu + \alpha_i + E_{ijk}, \qquad \text{avec } \{E_{ijk}\} \text{ i.i.d.} \sim \mathcal{N}(0, \sigma_1^2). \tag{5.8}$$

Effet fertilisation : on note β_j l'effet de la rotation j et on obtient

$$Y_{ijk} = \mu + \beta_j + F_{ijk}, \qquad \text{avec } \{F_{ijk}\} \text{ i.i.d.} \sim \mathcal{N}(0, \sigma_2^2). \tag{5.9}$$

On utilise des notations différentes pour les variables aléatoires résiduelles et leur variances car il n'y a aucune raison pour qu'elles soient les mêmes entre les deux modèles.

Notion de répétition. Il est essentiel de remarquer ici que dans chacun de ces deux modèle, l'indice qui représente le facteur non pris en compte (j dans le premier et i dans le second) a exactement la même fonction que l'indice de répétition k.
Ceci signifie que, dans le premier modèle par exemple, des parcelles ayant reçu le même niveau de fertilisation mais ayant subi des rotations différentes, sont considérées comme des répétitions. Cette hypothèse n'a évidemment de sens que si la rotation n'a pas d'effet ; si la rotation a un effet, ces deux parcelles ne constituent pas des répétitions puisqu'elles n'ont pas été cultivées dans des conditions équivalentes.

Tables d'analyse de la variance et sommes de carrés Les tableaux 5.13 et 5.14 présentent les tables d'analyse de la variance associées respectivement aux modèles 5.8 et 5.9.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	145.704167	145.704167	1.85	0.1790
Error	58	4566.588333	78.734282		
Corrected Total	59	4712.292500			

R-Square	Coeff Var	Root MSE	PdsGrains Mean
0.030920	36.93334	8.873234	24.02500

TABLE 5.13 – Table d'analyse de la variance du poids de grains en fonction du niveau de fertilisation.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1180.483000	590.241500	9.53	0.0003
Error	57	3531.809500	61.961570		
Corrected Total	59	4712.292500			

R-Square	Coeff Var	Root MSE	PdsGrains Mean
0.250511	32.76407	7.871567	24.02500

TABLE 5.14 – Table d'analyse de la variance du poids de grains en fonction de la rotation.

Le dispositif étant orthogonal, la somme de carrés de chacun de ces modèles ainsi que leur sommes de carrés résiduels se recalculent directement à partir des sommes de carrés ($SCA, SCB, SCI, SCM, SCR, SCT$) du modèle complet (5.2), p. 61 données aux tableaux 5.4, p. 63 et 5.6, p. 65 :

Source	Modèle	Somme de carrés	Degrés de liberté
Modèle	(5.8)	$SCA = 145.7$	$I - 1 = 1$
	(5.9)	$SCB = 1180.5$	$J - 1 = 2$
Résidu	(5.8)	$SCR + SCB + SCI$ $= 3052.5 + 1180.5 + 333.6 = 4566.6$	$IJK - I = 58$
	(5.9)	$SCR + SCA + SCI$ $= 3052.5 + 145.7 + 333.6 = 3531.8$	$IJK - J = 57$
Total	(5.8)	$SCT = 4712.3$	$IJK - 1 = 59$
	(5.9)	$SCT = 4712.3$	$IJK - 1 = 59$

La somme des carrés totaux ne dépend pas du modèle et ne varie donc jamais. Pour la somme de carrés du modèle, on retrouve bien un définition analogue à celle de l'analyse de la variance à un facteur.

La différence principale se situe au niveau de la somme des carrés résiduelle dont la définition varie d'un modèle à l'autre. Tous les effets non pris en compte dans le modèle passent dans les résiduelle. Plus ces effets sont forts, plus la résiduelle est importante et donc plus les tests sont sévères. Ainsi, la résiduelle du modèle (5.8) est supérieure à celle du modèle (5.9) parce que l'effet de la rotation est supérieur à celui de la fertilisation.
L'effet de la fertilisation, déjà faible, est en plus jugé avec plus de sévérité dans un modèle où il apparaît seul puisqu'il est comparé à une résiduelle qui contient des effets très forts.

Conclusion en termes de modélisation. Les conclusions tirées des modèles d'analyse de la variance à un facteur sont ici analogues à celle obtenues avec le modèle complet (5.2), p. 61 (fort effet de la rotation, pas d'effet de la fertilisation), mais il est clair que dans des situations moins tranchées, elles pourraient être sensiblement différentes.

Il est important de retenir que les analyses facteur par facteur sont dangereuses dans la mesure où elles conduisent à tester les effets de chaque facteur dans des modèles qui ne sont pas valides, les répétitions n'étant pas vraiment des répétitions. C'est pourtant une pratique courante, que de commencer par tester les facteurs un par un pour définir ceux qui seront retenu dans le modèle complet. On risque ainsi d'écarter à tort certains facteurs.

Dans le cas présent, cette stratégie conduirait à ne retenir que la rotation, et donc à ne jamais considérer le modèle complet avec interaction. C'est pourtant le modèle complet qui nous a permis de mettre en évidence une interaction (faible) entre la rotation et la fertilisation.

5.0.4 Programme SAS

Données. Le tableau 5.1, p. 59 et la figure 5.1, p. 60 sont produits par les instructions suivantes.

```
data COLZA;
     infile 'Colza.don' firstobs=2;
     input Fertilisation$ Rotation$ PdsGrains;
proc Print data=COLZA;
symbol1 i=boxJT l=1 c=black bwidth=1 co=black;
symbol2 i=boxJT l=2 c=red bwidth=3;
proc GPlot data=COLZA;
     plot PdsGrains*Rotation = Fertilisation;
run;
```

data Colza permet de définir le tableau de données.

proc Print permet l'affichage de son contenu.

proc GPlot affiche les boîtes à moustaches pour chacune des combinaisons grâce à l'option `i=boxJT` de l'instruction `symbol`.

Moyennes et écarts types dans les différentes combinaisons. Les tableaux 5.2, p. 59 et 5.3, p. 60 sont obtenus avec les instructions suivantes.

```
proc Means data=COLZA noprint;
     by Fertilisation Rotation;
     output out=INTER mean=PdsGrains std=EcartType;
proc Print data=INTER;
run;
```

proc Means calcule les moyennes et autres statistiques sans les afficher (option `noprint`).

proc Print affiche les valeurs sous un format plus compact.

Analyse de la variance avec interaction. Les tableaux 5.4, p. 63, 5.6, p. 65 et 5.8, p. 69 sont obtenus avec les instructions suivantes.

```
proc GLM data=Colza;
     class Fertilisation Rotation;
     model PdsGrains = Fertilisation Rotation Fertilisation*Rotation / solution;
     means Fertilisation Rotation / bon;
     lsmeans Fertilisation Rotation Fertilisation*Rotation / tdiff pdiff;
     output out=ANOVA2 p=Predite r=Residu;
run;
```

model spécifie le modèle. `Fertilisation*Rotation` représente l'interaction Fertilisation*Rotation.

means commande le calcul des moyennes par fertilisation et par rotation, ainsi que leur comparaison avec la méthode de Bonferroni (option `bon`).

lsmeans commande le calcul des moyennes ajustées qui sont ici égales aux moyenne simples. Cette instruction est utilisée ici pour des convenances de présentation des résultats. `tdiff` commande le calcul des statistiques de tests de student et `pdiff` celui des probabilités critiques associées.
Dans les version plus récentes de SAS, on peut obtenir la correction de Bonferroni pour les moyennes ajustées avec l'option `adjust = bon`.

Analyse des résidus. La figure 5.2, p. 73 est obtenue avec les instructions suivantes.

```
symbol i=none v=plus;
proc GPlot data=ANOVA2;
     plot Residu * Predite / vref=0;
run;
```

symbol précise la forme des symboles dans le graphe (`plus` = "+").

Analyse de la variance sans interaction. Le tableau 5.4, p. 63 est obtenu avec les instructions suivantes.

```
proc Anova data=Colza;
     class Fertilisation Rotation;
     model PdsGrains = Fertilisation Rotation;
run;
```

Analyse de la variance sur la fertilisation et sur la rotation. Les tableaux 5.13, p. 77 et 5.14, p. 77 sont obtenus avec les instructions suivantes.

```
proc Anova data=Colza;
     class Fertilisation;
     model PdsGrains = Fertilisation;
proc Anova data=Colza;
     class Rotation;
     model PdsGrains = Rotation;
run;
```

Chapitre 6

Analyse de la variance à deux facteurs : plan en blocs incomplets

6.1 Présentation

6.1.1 Objectif

On veut comparer les sensations de "piquant" de $I = 7$ champagnes (A, B, C, D, E, F, G). Cette sensation est mesurée par un indice compris entre 1 et 50 attribuée par un juge.

Il s'agit d'un problème d'*analyse sensorielle* : la notation fait appel à une évaluation subjective fournie par un juge entraîné.

Modélisation d'une note. La gamme assez large de notes (de 1 à 50) rend raisonnable le recours au modèle linéaire qui suppose la mesure continue. Une notation sur une gamme plus étroite (de 1 à 5, par exemple) rendrait cette hypothèse plus discutable.

6.1.2 Dispositif

$J = 14$ juges (numérotés de 1 à 14) interviennent dans l'évaluation des champagnes. Pour certains tests, comme c'est le cas ici, on ne demande pas à tous les juges d'évaluer tous les champagnes pour éviter que leurs facultés sensorielles ne s'émoussent.

On utilise la notations n_{ij}, n_{i+} et n_{+j} définies à la section 1.2, p. 4. Ici chaque juge note $n_{+j} = 3$ champagnes parmi les 7 et chaque champagne est noté par $n_{i+} = 6$ juges. On a donc au total

$$n = I \times n_{i+} = J \times n_{+j} = 42$$

observations. Le tableau 6.1, p. 83 donne les valeurs de tous ces effectifs.

Plan en blocs incomplets

D'un point de vue de planification, ce dispositif est un dispositif en blocs. On peut en effet considérer chaque juge comme un appareil de mesure ; toutes les mesures ne sont pas faites avec le même appareil et il faut tenir compte de l'hétérogénéité de ce matériel expérimental (particulièrement forte dans les problèmes d'analyse sensorielle). Chaque juge constitue donc un bloc. Chaque juge ne notant qu'une partie des champagnes mais tous les juges en notant le même nombre, on parle de plan en *Blocs Incomplets Équilibrés* (BIE).

On peut remarquer que, pour que le plan puisse être équilibré, il faut pouvoir trouver deux entiers n_{i+} et n_{+j} tels que $I \times n_{i+} = J \times n_{+j}$. Le nombre $J = 14$ de juges n'a donc pas été choisi au hasard.

Non-orthogonalité

La caractéristique principale de ce dispositif est qu'il n'est pas orthogonal. On rappelle (voir Daudin *et al.* (2007), chapitre 4) qu'un dispositif à deux facteurs est orthogonal si et seulement si on a

$$\forall i, j : n_{ij} = \frac{n_{i+} n_{+j}}{n}.$$

Cette condition n'est pas vérifiée ici puisque $n_{i+} = 6$ pour tous les champagnes, $n_{+j} = 3$ pour tous les juges et donc $n_{i+} n_{+j}/n = 18/42 = 3/7$. Or l'effectif n_{ij} ne peut prendre que deux valeurs :

$$\begin{aligned} n_{ij} &= 1 \quad \text{si le champagne } i \text{ est noté par le juge } j, \\ n_{ij} &= 0 \quad \text{si le champagne } i \text{ n'est pas noté par le juge } j. \end{aligned}$$

Le dispositif n'est donc pas orthogonal, ce qui signifie qu'il n'existe pas de décomposition unique des sommes de carrés associées à chaque effet et qu'*on ne pourra donc jamais complètement séparer les effets des deux facteurs*. Du fait du dispositif, les deux effets sont en partie confondus.

Couples de champagnes. Il existe un grand nombre de répartitions possibles des 7 champagnes entre les 14 juges. Le tableau 6.1, p. 83 présente celle qui a été utilisée dans cette expérience. Dans les plans BIE, on veille à ce que chaque couple de champagne soit noté le même nombre de fois par le même juge.

On observe ici que chaque couple est noté 2 fois par le même juge : le champagne A est noté en même temps que le champagne B par les juges 6 et 13, le champagne C est noté en même temps que le G par les juges 1 et 9, *etc.*

Cet équilibre dans la répartition des couples fait que toutes les comparaisons des champagnes 2 à 2 (contrastes) auront la même variance et donc que les tests les concernant auront la même puissance (cf. Dagnélie (1981) ou Bergonzini et Duby (1995)).

Table of Juge by Champagne

Juge / Champagne	A	B	C	D	E	F	G	Total
1	0	0	1	1	0	0	1	3
2	1	0	0	1	1	0	0	3
3	0	1	0	1	0	1	0	3
4	1	0	0	1	1	0	0	3
5	0	1	0	0	1	0	1	3
6	1	1	1	0	0	0	0	3
7	1	0	0	0	0	1	1	3
8	0	1	0	1	0	1	0	3
9	0	0	1	1	0	0	1	3
10	1	0	0	0	0	1	1	3
11	0	1	0	0	1	0	1	3
12	0	0	1	0	1	1	0	3
13	1	1	1	0	0	0	0	3
14	0	0	1	0	1	1	0	3
Total	6	6	6	6	6	6	6	42

TABLE 6.1 – Dispositif en BIE : répartition des 7 champagnes entre les 14 juges.

6.1.3 Données

Le tableau 6.2, p. 84 présente les résultats de l'expérience. On observe qu'il n'y a aucune donnée manquante, ce qui est une condition *sine qua non* pour conserver les bonnes propriétés statistiques de ce dispositif.

Obs	Juge	Champagne	Note	Obs	Juge	Champagne	Note
1	1	C	36	22	8	B	28
2	1	D	16	23	8	D	21
3	1	G	19	24	8	F	33
4	2	A	22	25	9	C	26
5	2	D	25	26	9	D	16
6	2	E	29	27	9	G	33
7	3	B	24	28	10	A	30
8	3	D	19	29	10	F	25
9	3	F	26	30	10	G	20
10	4	A	27	31	11	B	32
11	4	D	22	32	11	E	37
12	4	E	19	33	11	G	27
13	5	B	29	34	12	C	36
14	5	E	33	35	12	E	19
15	5	G	26	36	12	F	38
16	6	A	29	37	13	A	33
17	6	B	34	38	13	B	25
18	6	C	35	39	13	C	35
19	7	A	37	40	14	C	29
20	7	F	32	41	14	E	20
21	7	G	28	42	14	F	31

TABLE 6.2 – Note de "piquant" de 7 champagnes mesurés par 14 juges.

6.2 Analyse de la variance à 2 facteurs

6.2.1 Danger des analyses de variance séparées

Comme on l'a vu au chapitre 5, il est dangereux d'effectuer une analyse de la variance sur chacun des facteurs. Ici, une analyse de la variance sur le champagne seul serait fondée sur le modèle

$$Y_{ij} = \mu + \alpha_i + E_{ij}, \qquad \{E_{ij}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2) \tag{6.1}$$

en notant Y_{ij} la note obtenue par le i-ème champagne avec le j-ème juge.

Moyennes par champagne

Comme on l'a vu au chapitre 4, les estimateurs des paramètres du modèle (6.1) sont essentiellement fondés sur les notes moyennes obtenues par chaque champagne :

$$Y_{i\bullet} = \frac{1}{n_{i+}} \sum_{j=1}^{n_{i+}} Y_{ij} \qquad \{E_{ij}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2).$$

Dans le dispositif présent, *ces moyennes ne sont pas comparables* puisque les champagnes n'ont pas été notés par les mêmes juges (les indices $j = 1, \ldots, n_{i+}$ ne correspondent pas aux mêmes juges selon les champagnes). Les champagnes notés par des juges sévères

sont donc défavorisés par rapport aux autres champagnes si on ne tient pas compte de l'effet du juge.

Nous retrouvons ici l'importance de la notion de répétition déjà discutée à la section 5.0.3 : le modèle (6.1) considère les différentes notes obtenues par un même champagne comme des répétitions alors qu'elles ont été obtenues dans des conditions expérimentales différentes. Ce modèle n'est donc pas valide, du moins tant que l'absence d'effet juge n'a pas été démontrée.

6.2.2 Analyse de la variance à deux facteurs

Le modèle permettant d'analyser les sources de variabilité de la note Y doit donc prendre en compte l'effet champagne et l'effet juge, même s'il est bien clair que la comparaison des juges entre eux ne nous intéresse pas directement. L'effet juge doit être pris en compte pour analyser précisément l'effet champagne. Nous devons donc considérer un modèle d'analyse de la variance à deux facteurs :

$$\begin{aligned} A &: \text{facteur Champagne à } I = 7 \text{ niveaux,} \\ B &: \text{facteur Juge à } J = 12 \text{ niveaux.} \end{aligned}$$

Interaction Champagne*Juge. Il n'y a *a priori* aucune raison de penser que le champagne et le juge n'interagissent pas sur la note. On devrait donc considérer le modèle d'analyse de la variance à deux facteurs avec interaction :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ij}.$$

en notant α_i l'effet champagne, β_j l'effet juge et γ_{ij} l'interaction.

On voit cependant que, dans ce modèle, le terme d'interaction γ_{ij} est confondu avec le terme résiduel E_{ij} puisque ces deux termes portent exactement les mêmes indices. Ceci vient de l'absence de répétition : aucun champagne n'est noté plusieurs fois par le même juge (et pour cause : il n'est même pas noté par chacun des juges).

Un ajustement par les moindres carrés donnera un ajustement parfait et des résidus nuls, donc une estimation de la variance résiduelle nulle, ce qui interdirait tout test ou calcul d'intervalle de confiance. Dans ce dispositif, on est donc amené à *négliger le terme d'interaction* à cause de l'absence de répétition.

Cela ne signifie absolument pas que cette interaction n'existe pas, cela veut seulement dire que nous n'avons pas assez de données pour l'estimer. Cette interaction n'apparaîtra donc pas dans le modèle d'analyse et sera rejetée dans la résiduelle. Si cette interaction existe effectivement, les tests des effets des facteurs seront moins puissants.

Analyse de la variance à deux facteurs sans interaction

Le modèle d'analyse doit donc prendre en compte le plus d'effets possibles, soit ici
– l'effet champagne qui est l'effet d'intérêt,
– l'effet juge qui est un effet bloc,

– mais pas l'interaction à cause de l'absence de répétition.

On aboutit donc au modèle n'incluant que les deux effets principaux :

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}, \qquad \{E_{ij}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2). \tag{6.2}$$

Ce modèle implique

– 1 constante (μ), I paramètres pour l'effet champagne (α_i) et J paramètres pour l'effet juge (β_j), soient $(1 + I + J)$ paramètres d'espérance au total,

– et 1 paramètre de variance : σ^2.

Écriture matricielle

Le modèle (6.2) peut s'écrire sous la forme matricielle habituelle

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}$$

avec

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_{A2} \\ Y_{A4} \\ Y_{A6} \\ Y_{A7} \\ Y_{A10} \\ Y_{A13} \\ \vdots \\ Y_{G1} \\ Y_{G5} \\ Y_{G7} \\ Y_{G9} \\ Y_{G10} \\ Y_{G11} \end{bmatrix}, \quad \underset{(1+I+J)\times 1}{\boldsymbol{\Theta}} = \left[\begin{array}{c} \mu \\ \hline \alpha_1 \\ \vdots \\ \alpha_I \\ \hline \beta_1 \\ \vdots \\ \beta_J \end{array}\right], \quad \underset{n\times 1}{\mathbf{E}} = \begin{bmatrix} E_{A2} \\ E_{A4} \\ E_{A6} \\ E_{A7} \\ E_{A10} \\ E_{A13} \\ \vdots \\ E_{G1} \\ E_{G5} \\ E_{G7} \\ E_{G9} \\ E_{G10} \\ E_{G11} \end{bmatrix},$$

et [1]

$$\underset{n\times(1+I+J)}{\mathbf{X}} = \left[\begin{array}{c|cccccc|cccccccccccc} 1 & 1 & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . & 1 & . & . \\ \vdots & & & & & & & & & & & & & & & & & & \\ 1 & . & . & . & . & . & . & 1 & 1 & . & . & . & . & . & . & . & . & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & 1 & . & . & . & . & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & . & . & 1 & . & . & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & 1 & . & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & . & 1 & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & . & . & . & . & . & 1 & . \end{array}\right]$$

1. Les 0 sont remplacés par des points dans la matrice $\mathbf{X}$ pour en faciliter la lecture.

Estimation des paramètres

Comme dans les modèles d'analyse de la variance et de la covariance vus aux chapitres 4 et 5, on doit avoir recours à des contraintes pour obtenir des estimateurs des paramètres puisque la matrice $\mathbf{X}$ n'est que de rang $I + J - 1$ (alors qu'elle a $I + J + 1$ colonnes).

De façon générale, dans les plans non orthogonaux, les estimateurs de paramètres n'ont pas toujours une forme explicite, ce qui rend leur interprétation difficile. Nous ne nous attardons donc pas ici sur ce point.

6.3 Tests des effets et comparaison des champagnes

6.3.1 Notations

Les différents modèles que nous aurons à considérer dans cette section sont présentés dans le tableau 6.3. La notation (BA) (et non (AB)) pour le modèle à deux facteurs est volontaire : nous verrons dans la suite que l'ordre des facteurs dans le modèle à une importance.

Nom	Modèle	Interprétation	Tableau
(0) :	$Y_{ij} = \mu + E_{ij}$	aucun effet	–
(A) :	$Y_{ij} = \mu + \alpha_i + E_{ij}$	effet champagne seul	6.5, p. 90
(B) :	$Y_{ij} = \mu + \beta_j + E_{ij}$	effet juge seul	6.6, p. 90
(BA) :	$Y_{ij} = \mu + \beta_j + \alpha_i + E_{ij}$	effets juge et champagne	6.4, p. 88

TABLE 6.3 – Différents modèles d'analyse de la variance pour la comparaison des champagnes.

Les sommes de carrés obtenues dans chacun de ces modèles seront repérées par le nom du modèle. On notera, par exemple, $SCR(A)$ la somme des carrés résiduelle du modèle (A) ou $SCA(BA)$ la somme des carrés associée au facteur champagne (noté A) dans le modèle (BA).

6.3.2 Décomposition des sommes de carrés

Modèle complet

Le tableau 6.4 présente la table d'analyse de la variance du modèle (BA). On y lit que la somme des carrés due au modèle vaut $SCM(BA) = 1014.6$ et représente $R^2 = 64\%$ de la somme des carrés totaux $SCT(BA) = 1585.6$, ce qui montre un ajustement moyen. La significativité du modèle n'est pas excellente non plus : $F(BA) = 5.3\%$.

$SCM(BA)$ représente la réduction due aux effets champagne (α) et juge (β) par rapport à la constante (μ) :

$$R(\alpha, \beta/\mu) = SCM(BA) = 1014.6.$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	1014.595238	53.399749	2.06	0.0530
Error	22	571.047619	25.956710		
Corrected Total	41	1585.642857			

R-Square	Coeff Var	Root MSE	Couleur Mean
0.639864	18.43070	5.094773	27.64286

TABLE 6.4 – Table d'analyse de la variance des notes des champagnes pour le modèle (BA) à deux facteurs.

La difficulté ici va résider dans la décomposition de la somme des carrés $SCM(BA)$ en sommes de carrés dus à chacun des facteurs champagne et juge. Nous savons d'ores et déjà que, à cause de la non-orthogonalité du dispositif, cette décomposition n'est pas unique.

Variance résiduelle. Le tableau 6.4 fournit également l'estimation de la variance résiduelle :

$$\hat{\sigma}^2 = \frac{SCR(BA)}{n - I - J + 1} = 25.96$$

soit un écart type $\hat{\sigma} = 5.1$.
Cette variance résiduelle servira de dénominateur aux statistiques de Fisher permettant de tester les différents effets.

Analyse des résidus. Comme toujours dans le modèle linéaire, l'analyse des résidus est nécessaire pour s'assurer que les hypothèses portant sur l'indépendance, la normalité et l'homoscédasticité des résidus sont vérifiées. On ne s'étend pas ici sur cette analyse. Le graphique croisant les résidus et les valeurs prédites (cf. figure 6.1) ne montre pas de structure particulière, notamment concernant les variances des différents champagnes.

Analyses de la variance à 1 facteur

Nous avons vu au chapitre 5 et à la section 6.2.1 les dangers des analyses de la variance effectuées séparément sur chacun des facteurs. Nous introduisons cependant ces modélisations pour mieux expliquer les différentes méthodes de décomposition des sommes de carrés.

Les tableaux 6.5 et 6.6 présentent les tables d'analyse de la variance associées respectivement aux modèles (A) et (B).

La somme de carrés dus au modèle (A) représente la réduction due à l'introduction

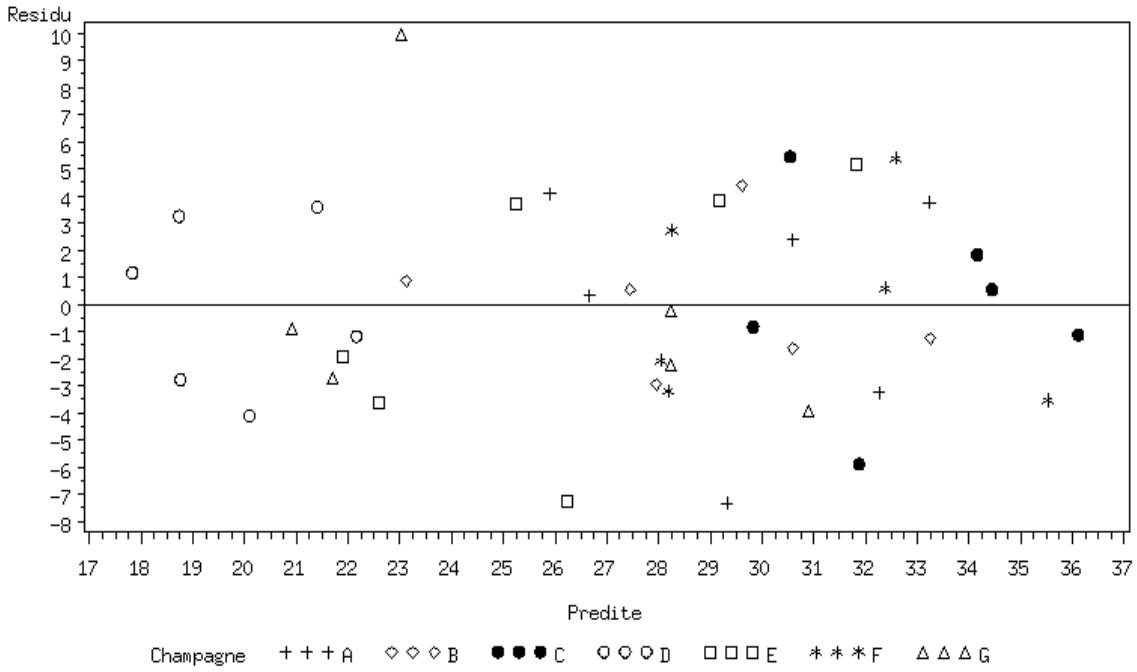

FIGURE 6.1 – Graphe des résidus pour l'analyse de la variance à deux facteurs.

de l'effet champagne (α) par rapport à la constante μ :

$$R(\alpha/\mu) = J\sum_i (Y_{i\bullet} - Y_{\bullet\bullet})^2 = SCM(A) = 660.1.$$

De même, la somme de carrés dus au modèle (B) représente la réduction due à l'introduction de l'effet juge (β) par rapport à la constante μ :

$$R(\beta/\mu) = I\sum_j (Y_{\bullet j} - Y_{\bullet\bullet})^2 = SCM(B) = 522.3.$$

Sommes de carrés de type I

On obtient une première décomposition des sommes de carrés en introduisant les facteurs *dans un ordre donné* dans le modèle. On obtient ainsi les sommes de carrés de type I, présentées dans le tableau 6.7.

Les résultats présentés se lisent donc pas à pas.

Étape 1 : on introduit l'effet juge (β) en plus de la constante μ. La somme de carrés associée à cet effet vaut

$$SCB_I(BA) = R(\beta/\mu) = 522.3$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	660.142857	110.023810	4.16	0.0029
Error	35	925.500000	26.442857		
Corrected Total	41	1585.642857			

TABLE 6.5 – Table d'analyse de la variance des notes des champagnes pour le modèle (A) : effet champagne seul.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	522.309524	40.177656	1.06	0.4302
Error	28	1063.333333	37.976190		
Corrected Total	41	1585.642857			

TABLE 6.6 – Table d'analyse de la variance des notes des champagnes pour le modèle (B) : effet juge seul.

et on retrouve le résultat du tableau 6.6.
Le test de l'effet juge correspond donc ici au test de

$$\mathbf{H}_0 = \{Y_{ij} = \mu + E_{ij}\} \qquad \text{contre} \qquad \mathbf{H}_1 = \{Y_{ij} = \mu + \beta_j + E_{ij}\}.$$

En terme de paramètres, cela revient à tester l'hypothèse

$$\mathbf{H}_0 = \{\mu_{\bullet 1} = \cdots = \mu_{\bullet J}\}.$$

Étape 2 : on introduit l'effet champagne (α). La somme de carrés associée vaut

$$SCA_I(BA) = R(\alpha/\mu, \beta) = 492.3$$

Le test de l'effet champagne correspond ici au test de

$$\mathbf{H}_0 = \{Y_{ij} = \mu + \beta_j + E_{ij}\} \qquad \text{contre} \qquad \mathbf{H}_1 = \{Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}\}.$$

L'expression de cette hypothèse en terme de paramètres est moins évidente.
Par construction, les sommes de carrés de type I sont additives puisque

$$R(\alpha, \beta/\mu) = R(\beta/\mu) + R(\alpha/\mu, \beta).$$

Attention, les tests de Fisher de toutes les hypothèses sont effectués en utilisant la variance estimée $\widehat{\sigma}^2$ dans le modèle complet, c'est à dire le modèle comprenant les deux effets.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Juge	13	522.3095238	40.1776557	1.55	0.1772
Champagne	6	492.2857143	82.0476190	3.16	0.0217

TABLE 6.7 – Sommes de carrés de type I et tests des effets des facteurs pour le modèle à deux facteurs (BA).

Ordre d'introduction des facteurs. Les résultats obtenus avec les sommes de carrés de type I dépendent donc de l'ordre dans lequel les facteurs sont introduits. Cette caractéristique n'est pas toujours souhaitable dans la mesure où il n'existe pas toujours un ordre naturel entre les facteurs.
Ici, on peut cependant considérer qu'il est préférable d'introduire l'effet juge (effet bloc) en premier afin de prendre en compte, avant tout, l'hétérogénéité du matériel expérimental. Le test de l'effet champagne (effet d'intérêt) se fait ainsi après correction de l'effet bloc.

Interprétation des tests. Les tests de type I sont fondés sur les sommes de carrés du même type. La statistique de test pour l'effet juge est

$$F = \frac{SCB_I(BA)/(J-1)}{\hat{\sigma}^2}.$$

Effet juge : le test fondé sur cette dernière statistique indique une absence d'effet 'absolu' du juge (probabilité critique = 17.7%) ? c'est-à-dire que les juges rendent, *en moyenne* des notes homogènes.
Il faut cependant se souvenir que ces moyennes sont obtenues en confondant les champagnes puisque le test est fondé sur la réduction $R(\beta/\mu) = I\sum_j (Y_{\bullet j} - Y_{\bullet\bullet})^2$.

Effet champagne : ce test montre, lui, un effet champagne significatif (probabilité critique = 2.2%) *correction faite de l'effet juge*. Ceci montre l'existence de différences entre les champagnes, au-delà de l'hétérogénéité potentielle introduite par les juges (puisque le test est fondé sur la réduction $R(\alpha/\mu, \beta)$).

Sommes de carrés de type II

On préfère souvent utiliser des sommes de carrés qui ne soient pas sensibles à l'ordre d'introduction des effets dans le modèle. Les sommes de carrés de types II répondent à ce critère.

Les sommes de carrés de types II associées à chaque facteur sont égales à la réduction associée à ce facteur *quand il est introduit en dernier dans le modèle.* On a donc :

$$SCA_{II} = R(\alpha/\mu, \beta) = 492.3, \qquad SCB_{II} = R(\beta/\mu, \alpha) = 354.4.$$

Pour information, si il existait une intéraction dans le modèle, la somme des carrés associés à cette intéraction en type II est égale à celle en type I : $SCI_{II} = SCI_I = R(\gamma/\mu, \alpha, \beta)$.

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Juge	13	354.4523810	27.2655678	1.05	0.4440
Champagne	6	492.2857143	82.0476190	3.16	0.0217

TABLE 6.8 – Sommes de carrés de type II et tests des effets des facteurs pour le modèle à deux facteurs (BA).

Interprétation des tests. Les tests de type II sont fondés sur les sommes de carrés du même type. La statistique de test pour l'effet juge est

$$F = \frac{SCB_{II}(BA)/(J-1)}{\hat{\sigma}^2}.$$

Effet juge : ce test montre une absence d'effet juge *correction faite de l'effet champagne* (probabilité critique = 44.4%). Il montre donc une homogénéité des juges, déduction faite des différences entre champagne (puisqu'il est fondé sur la réduction $R(\beta/\mu, \alpha)$).
Ce résultat montre que le jury est composé de personnes bien entraînées, fournissant des notes cohérentes entre elles.

Effet champagne : ce test est le même que le test de type I puisque l'effet champagne est introduit en dernier dans le modèle. L'effet champagne est significatif.

Comparaison des sommes de carrés

La figure 6.2 montre l'articulation entre les différentes réductions et les différents modèles.

Les sommes de carrés dus aux différents modèles et associées aux différents facteurs sont données par ces différentes réductions, comme le montre le tableau 6.9. On peut notamment reconstituer les sommes de carrés du modèle (AB).

	Modèle			
Réduction	(A)	(B)	(BA)	(AB)
$R(\alpha/\mu)$	SCM	$-$	$-$	SCA_I
$R(\beta/\mu)$	$-$	SCM	SCB_I	$-$
$R(\beta/\mu, \alpha)$	$-$	$-$	SCB_{II}	$SCB_I = SCB_{II}$
$R(\alpha/\mu, \beta)$	$-$	$-$	$SCA_I = SCA_{II}$	SCA_{II}
$R(\alpha, \beta/\mu)$	$-$	$-$	SCM	SCM

TABLE 6.9 – BIE sur les champagnes : comparaison des réductions et des sommes de carrés.

.

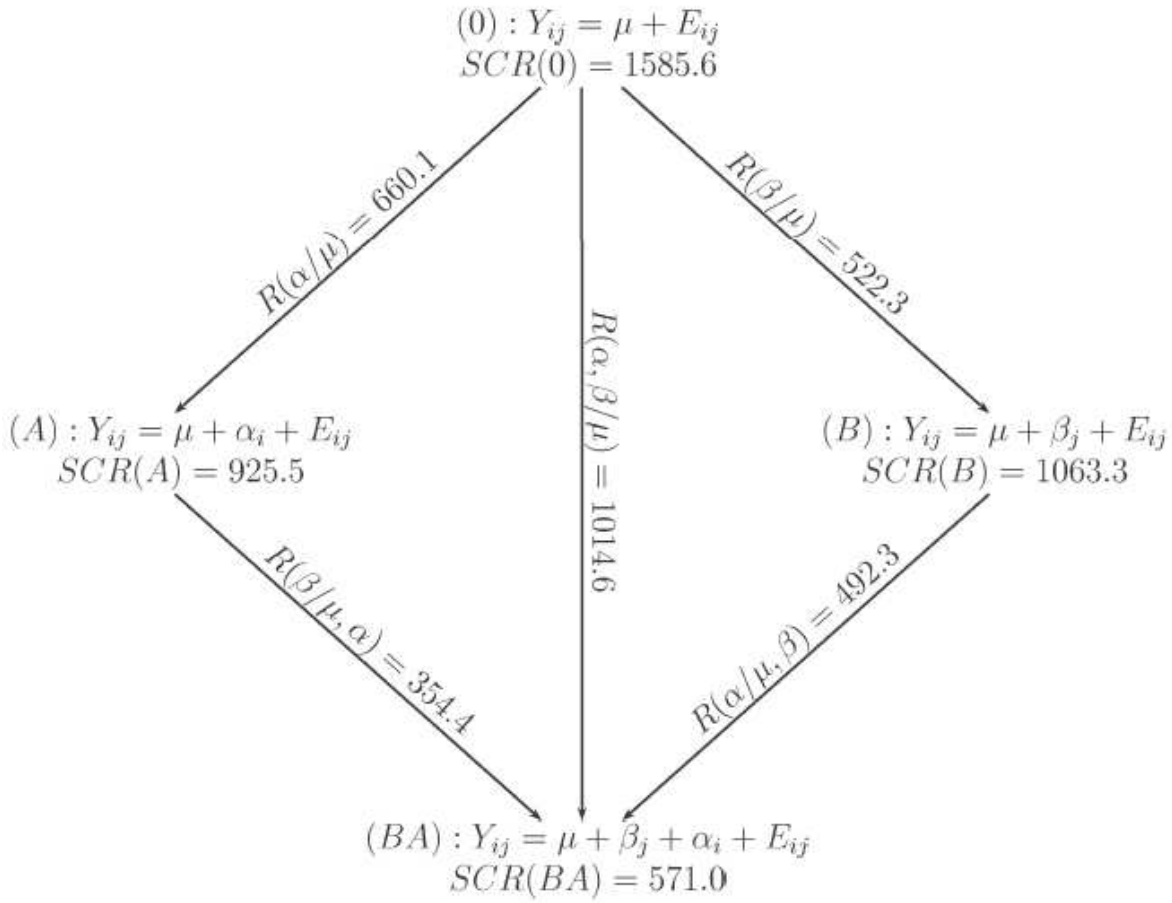

FIGURE 6.2 – Dispositif BIE : décomposition des sommes de carrés et comparaison des différents modèles.

Choix du modèle final

Les résultats en types I et II du modèle (BA) montre l'absence d'effet juge. On serait donc tenté finalement de retenir le modèle (A) dans lequel l'effet champagne apparaît seul. Ce modèle présente l'avantage d'accorder plus de degrés de liberté aux résidus et offre donc des tests plus puissants.

La pratique consiste pourtant, dans un dispositif comme celui-ci, à maintenir l'effet bloc (juge) dans le modèle pour s'assurer que les notes des champagnes sont bien comparables. Cette pratique est commandée par la prudence : la puissance du test de l'effet juge n'est pas parfaite et un écart opposant 1 seul juge aux 13 autres pourrait ne pas être détecté par le test de Fisher. Les champagnes notés par ce juge se trouveraient systématiquement avantagés (ou pénalisés) si on effectuait les comparaisons dans le modèle (A).

Les tests portant sur l'effet juge permettent malgré tout de montrer que le matériel expérimental (*i.e.* l'ensemble des juges) est globalement homogène, ce qui constitue un gage de qualité de l'expérience.

6.3.3 Comparaisons des champagnes

Moyennes classiques

Une première comparaison des champagnes peut se fonder sur les moyennes empiriques

$$\hat{\mu}_{i\bullet} = Y_{i\bullet}$$

données dans le tableau 6.10. Le champagne obtenant la meilleure note moyenne est le champagne C.

Level of Champagne	N	Couleur Mean	Couleur Std Dev
A	6	29.6666667	5.12510163
B	6	28.6666667	3.88158043
C	6	32.8333333	4.26223728
D	6	19.8333333	3.54494946
E	6	26.1666667	7.90990940
F	6	30.8333333	4.79235502
G	6	25.5000000	5.24404424

TABLE 6.10 – Moyennes classiques des notes par champagne.

On a vu cependant à la section 6.2.1 que ces moyennes n'étaient pas comparables à cause de l'hétérogénéité potentielle des juges. Cette approche consiste exactement à se placer dans le modèle (A) pour effectuer les comparaisons.

Moyennes ajustées

Les moyennes ajustées permettent de rendre comparables les moyennes par juge. Le principe consiste à estimer la note moyenne qu'aurait obtenu un champagne donné *s'il avait été noté par tous les juges* : on parle donc de moyenne ajustée sur l'effet juge (ou corrigée de l'effet juge).

L'espérance de la note obtenue par le champagne i avec le juge j est, selon le modèle (BA) :

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j.$$

Il est important de noter que cette espérance est définie dans le cadre du modèle *même si le champagne i n'a pas été noté par le juge j*. La moyenne ajustée est définie par

$$\tilde{\mu}_{i\bullet} = \frac{1}{J}\sum_j \mathbb{E}(Y_{ij}) = \mu + \alpha_i + \frac{1}{J}\sum_j \beta_j \tag{6.3}$$

et estimée par

$$\hat{\tilde{\mu}}_{i\bullet} = \hat{\mu} + \hat{\alpha}_i + \frac{1}{J}\sum_j \hat{\beta}_j.$$

Les valeurs de ces moyennes ajustées estimées sont données dans le tableau 6.11. On remarque que l'ordre des champagnes est le même qu'on considère les moyennes empiriques ou les moyennes ajustées :

$$\mathrm{D} < \mathrm{G} < \mathrm{E} < \mathrm{B} < \mathrm{A} < \mathrm{F} < \mathrm{C}.$$

Ceci est cohérent avec l'absence d'effet juge démontré à la section 6.3.2. En présence d'un effet juge fort, on pourrait tout à fait observer des inversions de classement.
Les écarts entre les champagnes sont cependant modifiés : l'écart de note entre les champagnes E et B diminue de 2.50 à 1.43 quand on passe des moyennes empiriques aux ajustées, alors que l'écart entre A et B augmente de 1.00 à 2.64. Ces changements modifient évidemment la conclusion des tests de comparaison des champagnes entre eux.

Champagne	Couleur LSMEAN	LSMEAN Number
A	29.5714286	1
B	26.9285714	2
C	33.4285714	3
D	21.6428571	4
E	25.5000000	5
F	31.8571429	6
G	24.5714286	7

TABLE 6.11 – Moyennes des notes par champagne ajustées sur l'effet juge.

Les moyennes ajustées définies en (6.3) sont des combinaisons linéaires (estimables) des paramètres μ, α_i et β_j. On peut donc calculer leur variances et covariances et ainsi

effectuer des tests d'hypothèses de la forme

$$\mathbf{H}_0 = \{\tilde{\mu}_{i\bullet} = \tilde{\mu}_{j\bullet}\} \qquad \text{contre} \qquad \mathbf{H}_1 = \{\tilde{\mu}_{i\bullet} \neq \tilde{\mu}_{j\bullet}\}.$$

Le tableau 6.12 donne les valeurs des statistiques de test et les probabilités critiques associées pour toutes les comparaisons [2]. On observe des différences de notes significatives (au niveau $\alpha = 5\%$) entre les couples de champagnes (A, D), (B, C), (C, D), (C, E), (C, G), (D, F) et (F, G).

i/j	1	2	3	4	5	6	7
1		0.792387	-1.15646	2.377161	1.220704	-0.68531	1.49911
		0.4366	0.2599	0.0266	0.2351	0.5003	0.1481
2	-0.79239		-1.94884	1.584774	0.428317	-1.47769	0.706723
	0.4366		0.0642	0.1273	0.6726	0.1537	0.4872
3	1.156456	1.948843		3.533617	2.377161	0.471149	2.655567
	0.2599	0.0642		0.0019	0.0266	0.6422	0.0144
4	-2.37716	-1.58477	-3.53362		-1.15646	-3.06247	-0.87805
	0.0266	0.1273	0.0019		0.2599	0.0057	0.3894
5	-1.2207	-0.42832	-2.37716	1.156456		-1.90601	0.278406
	0.2351	0.6726	0.0266	0.2599		0.0698	0.7833
6	0.685308	1.477694	-0.47115	3.062468	1.906012		2.184418
	0.5003	0.1537	0.6422	0.0057	0.0698		0.0399
7	-1.49911	-0.70672	-2.65557	0.87805	-0.27841	-2.18442	
	0.1481	0.4872	0.0144	0.3894	0.7833	0.0399	

TABLE 6.12 – Comparaison des moyennes ajustées des 7 champagnes. Pour chaque comparaison : valeur supérieure = statistique de test, valeur inférieure = probabilité critique.

Correction pour les tests multiples. Le niveau $\alpha = 5\%$ choisi pour chaque test ne prend pas en compte l'effet des tests multiples déjà vu à la section 5.0.2, p. 70. On effectue ici $7 \times 6/2 = 21$ comparaisons. La correction de Bonferroni amène donc à tester chaque comparaison au niveau $\alpha = 5\%/21 = 0.24\%$.
A ce niveau, seul les champagnes C (le meilleur) et D (le plus mauvais) sont significativement différents. On rappelle que la méthode de Bonferroni est une méthode conservative, c'est-à-dire que les tests qui en résultent sont peu puissants.

6.4 Programme SAS

Données. Le tableau 6.2, p. 84 est produit par les instructions suivantes.

```
data CHAMPAGNE;
```

2. ATTENTION : la correspondance des numéros de 1 à 7 avec les champagnes de A à G est donnée dans le tableau 6.11, dans la colonne `LSMEAN Number`. Dans certains cas, l'ordre de niveaux peut être modifié.

```
     infile 'champagne/Champagne.don' firstobs=2;
     input Juge Champagne$ Couleur;
proc Print data=CHAMPAGNE;
run;
```

Dispositif en Blocs Incomplets Equilibres (BIE). Le tableau 6.1, p. 83 est produit par les instructions suivantes.

```
proc Freq data=CHAMPAGNE;
     tables Juge * Champagne / nocol norow nopct;
run;
```

proc Freq permet d'obtenir les fréquences croisées des champagnes et des juges. Les options `nocol`, `norow` et `nopct` suppriment les calculs des pourcentages en ligne, en colonne et totaux, inutiles ici.

Anova sur le juge. Le tableau 6.6, p. 90 est produit par les instructions suivantes.

```
proc Anova data=CHAMPAGNE;
     class Juge;
     model Couleur = Juge;
run;
```

Anova sur le champagne. Les tableaux 6.5, p. 90 et 6.10, p. 94 sont produits par les instructions suivantes.

```
proc Anova data=CHAMPAGNE;
     class Champagne;
     model Couleur = Champagne;
     means Champagne;
run;
```

means requiert le calcul des notes moyennes par champagne.

Anova globale. Les tableaux 6.4, p. 88, 6.7, p. 91, 6.8, p. 92 6.11, p. 95, 6.12, p. 96 et la figure 6.1 sont produits par les instructions suivantes.

```
proc GLM data=CHAMPAGNE;
     class Juge Champagne;
     model Couleur = Juge Champagne;
     lsmeans Champagne / tdiff pdiff cov;
     output out=GLM p=Predite r=Residu;
     symbol1 v=plus c=black;
     symbol2 v=diamond c=black;
     symbol3 v=dot c=black;
```

```
     symbol4 v=circle c=black;
     symbol5 v=square c=black;
     symbol6 v=star c=black;
     symbol7 v=triangle c=black;
proc GPlot data=GLM;
     plot Residu *  Predite = Champagne / vref=0;
run;
```

model spécifie le modèle analyse (cf (6.2), p. 86). Les sommes de carrés de type I sont calculées par défaut. Pour obtenir les sommes de carrés de type II, il faut ajouter l'option SS2 à la ligne model : "model Couleur = Juge Champagne / SS1 SS2 ;".

lsmeans requiert les calcul des moyennes ajustées. Les options `tdiff` et `pdiff` permettent d'obtenir les valeurs des statistique de test et les probabilités critiques associées.

proc GPlot permet d'obtenir le graphe des résidus. Les instructions `symbol` précise le symbole représentant chaque champagne.

Chapitre 7

Analyse de la covariance

7.1 Présentation

7.1.1 Objectif et dispositif expérimental

Objectif. On cherche à savoir si des conditions de température et d'oxygénation ont une influence sur l'évolution du poids des huîtres.

Expérience et données. On dispose de $n = 20$ sacs de 10 huîtres. On place, pendant un mois, ces 20 sacs de façon aléatoire dans $I = 5$ emplacements différents d'un canal de refroidissement d'une centrale électrique à raison de $K = 4$ sacs par emplacement. Ces emplacements se différencient par leurs température et oxygénation. Pour chaque sac, on a

- son poids avant l'expérience (variable `Pds Init`),
- son poids après l'expérience (variable `Pds Final`),
- l'emplacement (variable `Traitement`) codé de 1 à 5.

Les données sont présentées dans la table 7.1.

Remarque : On peut remarquer qu'ici on cherche à expliquer la variable `Pds Final` (variable quantitative) à partir d'une variable quantitative `Pds Init` et une variable qualitative `Traitement`. L'analyse de la covariance est le premier modèle que l'on voit à mêler variables qualitative et quantitative.

7.1.2 Description des données

Notion d'orthogonalité.

Cette notion a été introduite pour l'analyse de la variance à 2 facteurs (cf chapitre 5). Dans le cadre de l'analyse de la covariance, on dit que le dispositif est orthogonal si la variable quantitative prend les mêmes valeurs pour chaque niveau de la variable qualitative. D'après les données (cf Table 7.1), les poids initiaux par traitement sont différents donc le dispositif n'est ici pas orthogonal.

Obs	Traitement	Repetition	Pds Init	Pds Final
1	1	1	27.2	32.6
2	1	2	32.0	36.6
3	1	3	33.0	37.7
4	1	4	26.8	31.0
5	2	1	28.6	33.8
6	2	2	26.8	31.7
7	2	3	26.5	30.7
8	2	4	26.8	30.4
9	3	1	28.6	35.2
10	3	2	22.4	29.1
11	3	3	23.2	28.9
12	3	4	24.4	30.2
13	4	1	29.3	35.0
14	4	2	21.8	27.0
15	4	3	30.3	36.4
16	4	4	24.3	30.5
17	5	1	20.4	24.6
18	5	2	19.6	23.4
19	5	3	25.1	30.3
20	5	4	18.1	21.8

TABLE 7.1 – Table des données.

Répartition des sacs dans les traitements avant l'expérience.

La table 7.2 donne les moyennes et les écart-types du poids intial (et du poids final) en fonction du traitement. On observe des différences fortes de moyennes de ces poids qui vont de 29.75 pour le traitement 1 à 20.8 pour le traitement 5 avec un poids initial moyen nettement plus petit pour ce dernier traitement. Ces observations se retrouvent dans la table 7.1 des données ou sur le graphe 7.1. En effet, on voit que les poids initaux des sacs des différents traitements ne sont pas dans les mêmes catégories de poids.

Variabilité du poids des sacs.

Le poids final moyen est à peu près le même pour les différents traitements sauf pour le traitement 5 pour lequel le poids est plus faible. On observe également dans la table 7.2 que la variabilité du poids final à l'intérieur de chaque traitement est à peu près la même, seul le traitement 2 sort du lot. Pour ce traitement, cette variabilité est plus faible. Cela peut s'expliquer par le fait que la variabilité du poids initial pour ce traitement est elle aussi très faible, ce qui signifie que les sacs mis dans l'emplacement associé ont presque le même poids. Et dans une même condition, la croissance des huîtres est sensiblement la même.

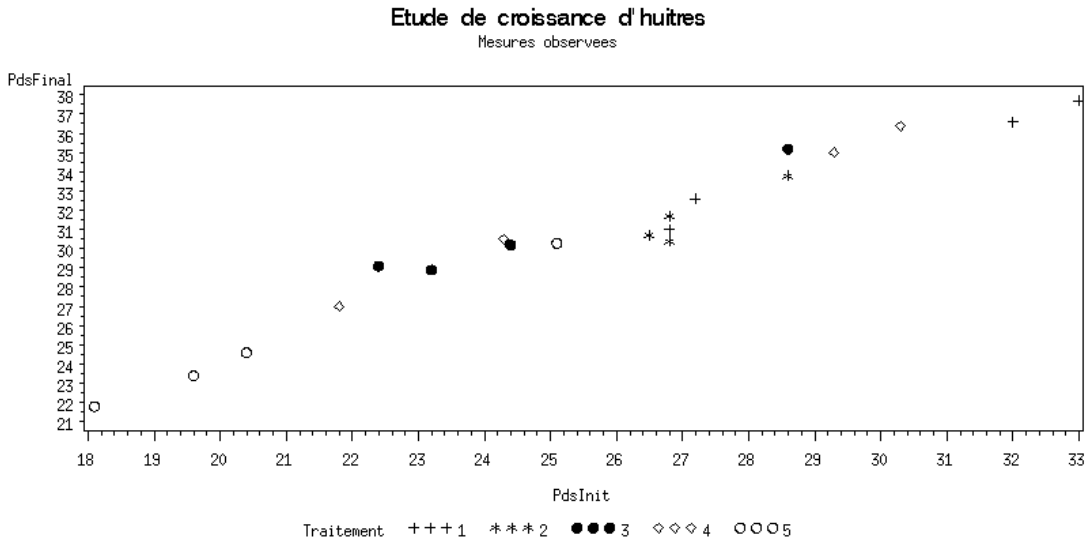

FIGURE 7.1 – Graphe des données portant sur les huîtres.

7.2 Vers l'analyse de la covariance

A partir des données disponibles, on pourrait envisager plusieurs études et donc poser plusieurs modèles. Il est important de resouligner que chaque modélisation permet de répondre à une question précise, comme on a pu le voir durant les précédents chapitres. Par exemple, au vu des données (cf figure 7.1), le poids final semble proportionnel au poids initial à terme constant près. Pour étudier l'évolution de poids chez les huîtres, on pourrait considérer un modèle de régression simple. Le graphe des résidus contre les prédictions donné par la figure 7.2 est correct, les hypothèses du modèle sont validées et les résultats peuvent être analysés. Cependant, en considérant ce modèle, on ne prend pas du tout en compte le fait que la relation entre les variables `Pds Init` et `Pds Final` peut être différente selon le traitement. Or d'après le graphe des données (codées par traitement), il semble que le poids des huîtres n'évolue pas de la même façon selon la condition dans laquelle elles se trouvent. Le graphe des résidus de la régression simple (cf figure 7.2) illustre aussi ce phénomène puisque l'on peut observer que les résidus sont regroupés par traitement et se situent de part et d'autres de l'axe des abscisses.
L'idée alors naturelle est d'effectuer 5 régressions, une régression par traitement, qui permet d'étudier pour chaque traitement la relation entre les variables `Pds Init` et `Pds Final`. Les modèles seraient les suivants :

$$\begin{aligned} Y_{1k} &= a_1 + b_1 x_{1k} + E_{1k}, \qquad k = 1, \ldots, K = 4 \\ Y_{2k} &= a_2 + b_2 x_{2k} + E_{2k}, \\ &\vdots \\ Y_{5k} &= a_5 + b_5 x_{5k} + E_{5k}. \end{aligned}$$

Variable	N	Mean	Std Dev	Minimum	Maximum
PdsInit	20	25.7600000	4.0378994	18.1000000	33.0000000
PdsFinal	20	30.8450000	4.3448063	21.8000000	37.7000000

Traitement=1

Variable	N	Mean	Std Dev	Minimum	Maximum
PdsInit	4	29.7500000	3.2057240	26.8000000	33.0000000
PdsFinal	4	34.4750000	3.1889131	31.0000000	37.7000000

Traitement=2

Variable	N	Mean	Std Dev	Minimum	Maximum
PdsInit	4	27.1750000	0.9604686	26.5000000	28.6000000
PdsFinal	4	31.6500000	1.5373137	30.4000000	33.8000000

Traitement=3

Variable	N	Mean	Std Dev	Minimum	Maximum
PdsInit	4	24.6500000	2.7586228	22.4000000	28.6000000
PdsFinal	4	30.8500000	2.9557853	28.9000000	35.2000000

Traitement=4

Variable	N	Mean	Std Dev	Minimum	Maximum
PdsInit	4	26.4250000	4.0491769	21.8000000	30.3000000
PdsFinal	4	32.2250000	4.2975768	27.0000000	36.4000000

Traitement=5

Variable	N	Mean	Std Dev	Minimum	Maximum
PdsInit	4	20.8000000	3.0210373	18.1000000	25.1000000
PdsFinal	4	25.0250000	3.6989863	21.8000000	30.3000000

TABLE 7.2 – Statistiques élémentaires sur l'échantillon entier et par traitement.

Les hypothèses de ces 5 modèles sont que les $\{E_{ik}\}$ sont indépendants, de loi gaussienne et de même variance au sein de chaque traitement (pour un i fixé). Pour répondre à

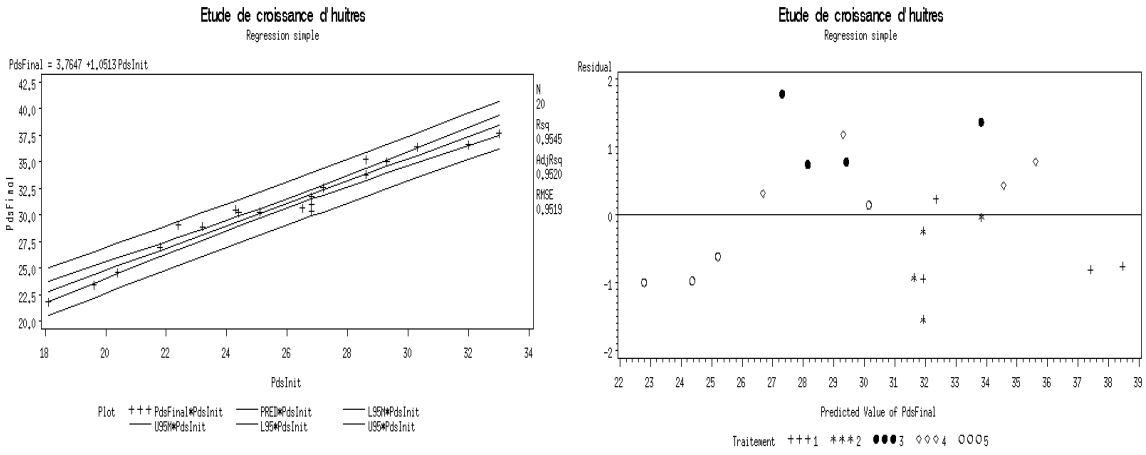

FIGURE 7.2 – Résultat de la régression simple sur l'évolution du poids des huîtres pour tous les traitements en même temps. A gauche : droite de régression estimée - à droite : graphe des résidus associés.

la question que l'on se pose ici qui est : "est-ce que l'évolution du poids des huîtres est différente entre les traitements ?", il s'agit de comparer ces droites de régression. Malheureusement, il n'existe pas de tests pour comparer des droites. La solution est l'analyse de la covariance, qui par le fait qu'elle correspond à la réunion de ces 5 régressions, permet cette comparaison.

7.3 Analyse de la Covariance

7.3.1 Modèle

Modèle général

Le modèle de covariance s'écrit :

$$Y_{ik} = a_i + b_i x_{ik} + E_{ik} \qquad \{E_{ik}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \tag{7.1}$$

où

- i représente l'indice du niveau du traitement, $i = 1, \ldots, I = 5$,
- k représente l'indice de répétition, i.e., le numéro du sac d'huître pour son traitement, $k = 1, \ldots, K = 4$,
- x_{ik} représente le $k$ème poids initial du traitement i,
- a_i représente la valeur du poids final pour un sac de poids initial nul pour le traitement i,
- b_i représente la pente de la régression pour le traitement i,
- σ^2 est la variance résiduelle.

L'hypothèse supplémentaire induite par l'analyse de la covariance par rapport à l'ensemble des 5 modèles de régression est que la variance des variables aléatoires $\{E_{ik}\}$ est la même pour les 5 régressions.

Décomposition des effets

Comme pour l'analyse de la variance à deux facteurs, l'écriture du modèle précédent (7.1) est plus compacte et confond plusieurs effets. Pour les faire apparaître, on décompose chaque paramètre de régression (a et b) comme la somme d'un effet global du traitement et un effet spécifique au niveau de traitement en question : on pose

- $a_i = \mu + \alpha_i$ où μ représente le terme constant et α_i l'effet spécifique du $i$ème traitement,
- $b_i = \beta + \gamma_i$, où β représente l'effet global du traitement et γ_i représente l'effet spécifique du $i$ème traitement. Ce dernier effet est un terme correctif à β en fonction du traitement i considéré. Comme dans le modèle d'analyse de la variance à deux facteurs, c'est le terme qui représente l'effet conjoint du facteur traitement et de la variable `Pds Init`. On l'appelle donc aussi terme d'*interaction*.

Le modèle (7.1) s'écrit alors :

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + E_{ik} \qquad \{E_{ik}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \tag{7.2}$$

Le nombre de paramètres à estimer dans ce modèle est

- 1 pour la constante (μ), I pour l'effet traitement (α_i), 1 pour l'effet global du traitement (β) et I pour l'interaction (γ_i), soient $2(I+1)$ pour l'espérance au total,
- 1 pour la variance résiduelle (σ^2).

Ecriture en terme de loi des Y_{ik}.

Le modèle (7.2) est équivalent au modèle

$$Y_{ik} \sim \mathcal{N}(\mu_{ik}, \sigma^2), \qquad \{Y_{ik}\} \text{ indépendants} \tag{7.3}$$

en notant

$$\mu_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik}.$$

On a vu sur l'exemple de l'analyse de la variance à 2 facteurs du chapitre 5 que l'absence de répétitions ne permet pas d'estimer le terme d'interaction. Ici, ce paramètre pourra être estimé puisqu'on dispose de $K = 4$ répétitions par traitement.

Ecriture matricielle

Le modèle (7.2) s'écrit sous la forme matricielle suivante :

$$\mathbf{Y} = \mathbf{X\Theta} + \mathbf{E}, \qquad \mathbf{E} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}). \tag{7.4}$$

où

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \\ Y_{41} \\ Y_{42} \\ Y_{43} \\ Y_{44} \\ Y_{51} \\ Y_{52} \\ Y_{53} \\ Y_{54} \end{bmatrix}, \quad \underset{n\times 2(I+1)}{\mathbf{X}} = \left[\begin{array}{c|cccccc|c|cccccc} 1 & 1 & 0 & 0 & 0 & 0 & x_{11} & x_{11} & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & x_{12} & x_{12} & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & x_{13} & x_{13} & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & x_{14} & x_{14} & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & x_{21} & 0 & x_{21} & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & x_{22} & 0 & x_{22} & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & x_{23} & 0 & x_{23} & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & x_{24} & 0 & x_{24} & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{31} & 0 & 0 & x_{31} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{32} & 0 & 0 & x_{32} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{33} & 0 & 0 & x_{33} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{34} & 0 & 0 & x_{34} & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & x_{41} & 0 & 0 & 0 & x_{41} & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & x_{42} & 0 & 0 & 0 & x_{42} & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & x_{43} & 0 & 0 & 0 & x_{43} & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & x_{44} & 0 & 0 & 0 & x_{44} & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & x_{51} & 0 & 0 & 0 & 0 & x_{51} \\ 1 & 0 & 0 & 0 & 0 & 1 & x_{52} & 0 & 0 & 0 & 0 & x_{52} \\ 1 & 0 & 0 & 0 & 0 & 1 & x_{53} & 0 & 0 & 0 & 0 & x_{53} \\ 1 & 0 & 0 & 0 & 0 & 1 & x_{54} & 0 & 0 & 0 & 0 & x_{54} \end{array}\right],$$

$$\underset{2(I+1)\times 1}{\Theta} = \left[\begin{array}{c} \mu \\ \hline \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \hline \beta \\ \hline \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{array}\right], \quad \underset{n\times 1}{\mathbf{E}} = \begin{bmatrix} E_{11} \\ E_{12} \\ E_{13} \\ E_{14} \\ E_{21} \\ E_{22} \\ E_{23} \\ E_{24} \\ E_{31} \\ E_{32} \\ E_{33} \\ E_{34} \\ E_{41} \\ E_{42} \\ E_{43} \\ E_{44} \\ E_{51} \\ E_{52} \\ E_{53} \\ E_{54} \end{bmatrix}.$$

Les lignes verticales dans $\mathbf{X}$ correspondent aux lignes horizontales de θ. On note aussi que le rang de la matrice $\mathbf{X}$ est égal à $2(I+1)-2$ puisque la première colonne est égale à la somme des 5 suivantes comme la septième alors que le nombre de paramètres à estimer est $2(I+1)$. La conséquence est que le modèle est non identifiable et il sera nécessaire d'introduire 2 contraintes sur les paramètres pour ne pas avoir une infinité de solutions mais une solution particulière qui tiendra naturellement compte de ces contraintes (cf Daudin *et al.* (2007), subsection 2.1.3).

7.3.2 Test du modèle

Table d'analyse de la variance. Comme dans tous les modèles linéaires, le premier test effectué est le test du modèle *constant* contre le modèle *complet*. Ici, ce test correspond au test des hypothèses :

$$H_0 = \{Y_{ik} = \mu + E_{ik}\} \quad \text{contre} \quad H_1 = \{Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + E_{ik}\}.$$

Les résultats de ce test se trouvent (comme dans l'ANOVA) dans la table d'analyse de la variance (table 7.3). Les sommes de carrés qui y figurent sont écrites de façon générale dans la table 7.4 où $\hat{Y}_{ik}$ est la prédiction de Y_{ik} dans le modèle complet et qui est défini (7.6).
La statistique du test sur le modèle s'écrit :

$$F = \frac{SCM/(2I-1)}{\hat{\sigma}^2},$$

où $\hat{\sigma}^2$ est défini par (7.5). Cette statistique suit, sous l'hypothèse H_0, une loi de Fisher à $2I-1$ et $n-2I$ degrés de libertés. Elle vaut 139.51, ce qui signifie que la variabilité expliquée par le traitement et le poids initial est 140 fois supérieure à la variabilité résiduelle. La probabilité critique est très faible puisque qu'elle est inférieure à 0.0001. On conclut que l'effet conjoint du traitement et du poids initial est significatif, le modèle complet est conservé.
En ce qui concerne l'ajustement, il est très bon. On lit en effet dans la table 7.3 que la somme des carrés du modèle ($SCM = 355.8$) est presque égale à la somme des carrés totale ($SCT = 358.6$), la variabilité du modèle explique $R^2 = 99\%$ de la variabilité du poids final.

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	9	355.8354908	39.5372768	139.51	<.0001
Error	10	2.8340092	0.2834009		
Corrected Total	19	358.6695000			

R-Square	Coeff Var	Root MSE	PdsFinal Mean
0.992099	1.725901	0.532354	30.84500

TABLE 7.3 – Table d'analyse de la variance du modèle de covariance avec interaction.

Analyse des résidus. Comme pour toutes les analyses de modèles linéaires, il faut s'assurer que les hypothèses du modèles sont vérifiées. On s'interesse donc au graphe des

Source	Degrés de liberté	Somme de carrés	Carré moyen
Modèle	$2I-1$	$SCM = \sum_{i=1}^{I}\sum_{k=1}^{K}(\hat{Y}_{ik} - Y_{\bullet\bullet})^2$	$SCM/(2I-1)$
Résidu	$n-2I$	$SCR = \sum_{i=1}^{I}\sum_{k=1}^{K}(Y_{ik} - \hat{Y}_{ik})^2$	$SCR/(n-2I)$
Total	$n-1$	$SCT = \sum_{i=1}^{I}\sum_{k=1}^{K}(Y_{ik} - Y_{\bullet\bullet})^2$	$SCT/(n-1)$

TABLE 7.4 – Définition des sommes de carrés et carrés moyen dans le modèle d'analyse de la covariance à deux facteurs.

résidus contre les prédictions. Pour plus de détails sur l'analyse de ce graphe, on pourra se reporter au paragraphe 2.2.2. Ce graphe obtenu dans le cas de l'analyse de la covariance est représenté figure (7.3). Il ne présente pas de structure particulière.

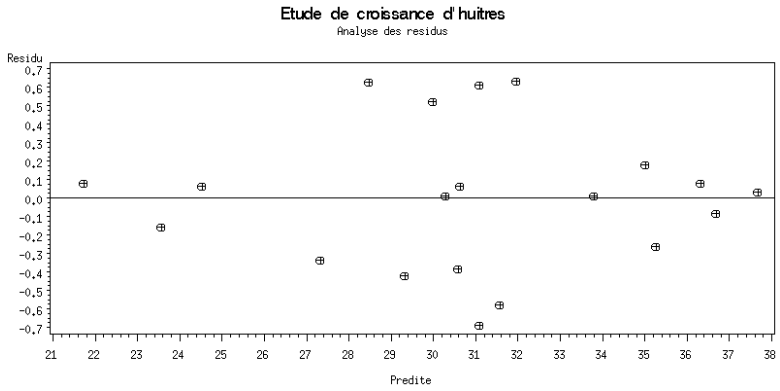

FIGURE 7.3 – Graphe des résidus contre les prédictions pour le modèle d'analyse de covariance avec interaction.

7.3.3 Estimation des paramètres et interprétation

Estimation des paramètres de l'espérance (μ, α_i, β et γ_i). Comme il a déjà été évoqué dans la partie précédente, puisque le rang de la matrice $\mathbf{X}$ est 10 alors que le nombre de paramètres à estimer est 12, il est nécessaire de poser 2 contraintes sur les paramètres. Celles utilisées par SAS sont

$$\alpha_I = 0 \quad \text{et} \quad \gamma_I = 0,$$

comme il est précisé dans la table 7.5, qui donne les estimations des paramètres du modèle (7.2), par le "point". Comme dans l'analyse de la variance, l'interprétation des paramètres

dépend de ces contraintes.
Reprenons la première écriture du modèle, à savoir (7.1). Tout d'abord, on peut montrer que les estimations de a_i et b_i sont exactement les mêmes que quand on fait une régression simple groupe par groupe. Ensuite en se rappelant que $a_i = \mu + \alpha_i$ et $b_i = \beta + \gamma_i$ et des contraintes SAS, il est facile de voir que

- $\hat{\mu} = \hat{a}_I$, donc $\hat{\mu}$ représente l'estimation de l'ordonnée à l'origine de la droite de régression du groupe I.
- $\hat{\alpha}_i = \hat{a}_i - \hat{a}_I$ correspond à la différence à l'origine entre le groupe i et le groupe I.
- $\hat{\beta} = \hat{b}_I$ représente la pente de la droite de régression du groupe I.
- $\hat{\gamma}_i = \hat{b}_i - \hat{b}_I$ représente la différence de pentes entre les régressions des groupes i et I. Ainsi par exemple, si $\hat{\gamma}_i > 0$ cela signifiera que l'évolution (linéaire entre x et y) est plus importante dans le groupe i que le groupe I.

Parameter		Estimate	Error	t Value	Pr >|t|
Intercept		-0.431829803 B	2.13282848	-0.20	0.8436
PdsInit		1.223886048 B	0.10173817	12.03	<.0001
Traitement	1	5.673088317 B	3.57150301	1.59	0.1433
Traitement	2	-8.717492690 B	8.95782354	-0.97	0.3534
Traitement	3	5.249788629 B	3.48748453	1.51	0.1632
Traitement	4	4.727585838 B	2.93990691	1.61	0.1389
Traitement	5	0.000000000 B	.	.	.
PdsInit*Traitement	1	-0.241239276 B	0.13979638	-1.73	0.1151
PdsInit*Traitement	2	0.277468965 B	0.33578844	0.83	0.4279
PdsInit*Traitement	3	-0.167819469 B	0.15087805	-1.11	0.2920
PdsInit*Traitement	4	-0.166961016 B	0.12693423	-1.32	0.2178
PdsInit*Traitement	5	0.000000000 B	.	.	.

TABLE 7.5 – Estimation des paramètres du modèle d'analyse de la covariance.

Variance résiduelle. La table 7.3 donne également une estimation de la variance résiduelle :

$$\hat{\sigma}^2 = \frac{SCR}{n - 2I} = 0.28, \tag{7.5}$$

ou une estimation de son écart-type par `Root MSE`, $\hat{\sigma} = 0.532$.

Prédiction. La prédiction du poids final pour un sac de poids initial x_0 placé dans le traitement i est donnée par :

$$\hat{Y}_{i,x_0} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta} x_0 + \hat{\gamma}_i x_0. \tag{7.6}$$

7.4 Tests de l'effet des différents facteurs et variables

7.4.1 Notations

Pour présenter tous les tests, nous aurons besoin de considérer 5 sous-modèles du modèle d'analyse de la covariance complet noté (C). Ces modèles, codés de (0) à (4), sont présentés dans la table 7.6 et le modèle complet est rappelé.

Nom	Modèle
(0)	$Y_{ik} = \mu + E_{ik}$
(1)	$Y_{ik} = \mu + \alpha_i + E_{ik}$
(2)	$Y_{ik} = \mu + \beta x_{ik} + E_{ik}$
(3)	$Y_{ik} = \mu + \beta x_{ik} + \gamma_i x_{ik} + E_{ik}$
(4)	$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + E_{ik}$
(5)	$Y_{ik} = \mu + \alpha_i + \gamma_i x_{ik} + E_{ik}$
(C)	$Y_{ik} = \mu + \beta x_{ik} + \alpha_i + \gamma_i x_{ik} + E_{ik}$

TABLE 7.6 – Différents sous-modèles du modèle de covariance avec interaction.

On peut remarquer que nous avons écrit le modèle complet dans un ordre différent que l'ordre classique (cf (7.2)). On verra dans la suite en quoi l'ordre d'écriture du modèle peut être importante pour les tests effectués dans un cas où le dispositif est non-orthogonal.

Les sommes des carrés du modèle associées à chaque modèle sont indicées par leur numéro de modèle. Par exemple, pour le modèle (2), $SCM = SCM_{(2)}$.

Illustration. Pour faciliter l'interprétation des différents tests effectués dans la suite, les modèles de la table 7.6 sont représentés par leurs espérances avec $I = 2$ en figure 7.4.

Interprétation. Nous donnons ici l'interprétation des différents modèles de la table 7.6.

- **Modèle** (0) **:** aucun effet n'est présent. Il est par conséquent appelé *modèle nul* ou *modèle constant*.
- **Modèle** (1) **:** il n'y a que l'effet traitement. C'est un modèle d'analyse de la variance à 1 facteur à $I = 5$ niveaux, le traitement.
- **Modèle** (2) **:** c'est un modèle de régression simple : la relation entre le poids initial et le poids final est linéaire.
- **Modèle** (3) **:** l'évolution du poids final en fonction du poids initial est différente selon le traitement. La différence avec le modèle complet est que les I régressions ont la même origine.
- **Modèle** (4) **:** c'est un modèle d'analyse de la covariance sans interaction ($\gamma_i = 0, \forall i$). Il revient à avoir I droites de régression parallèles de pente β : l'évolution du poids des huîtres en fonction du poids initial est la même quelque soit le traitement. Mais

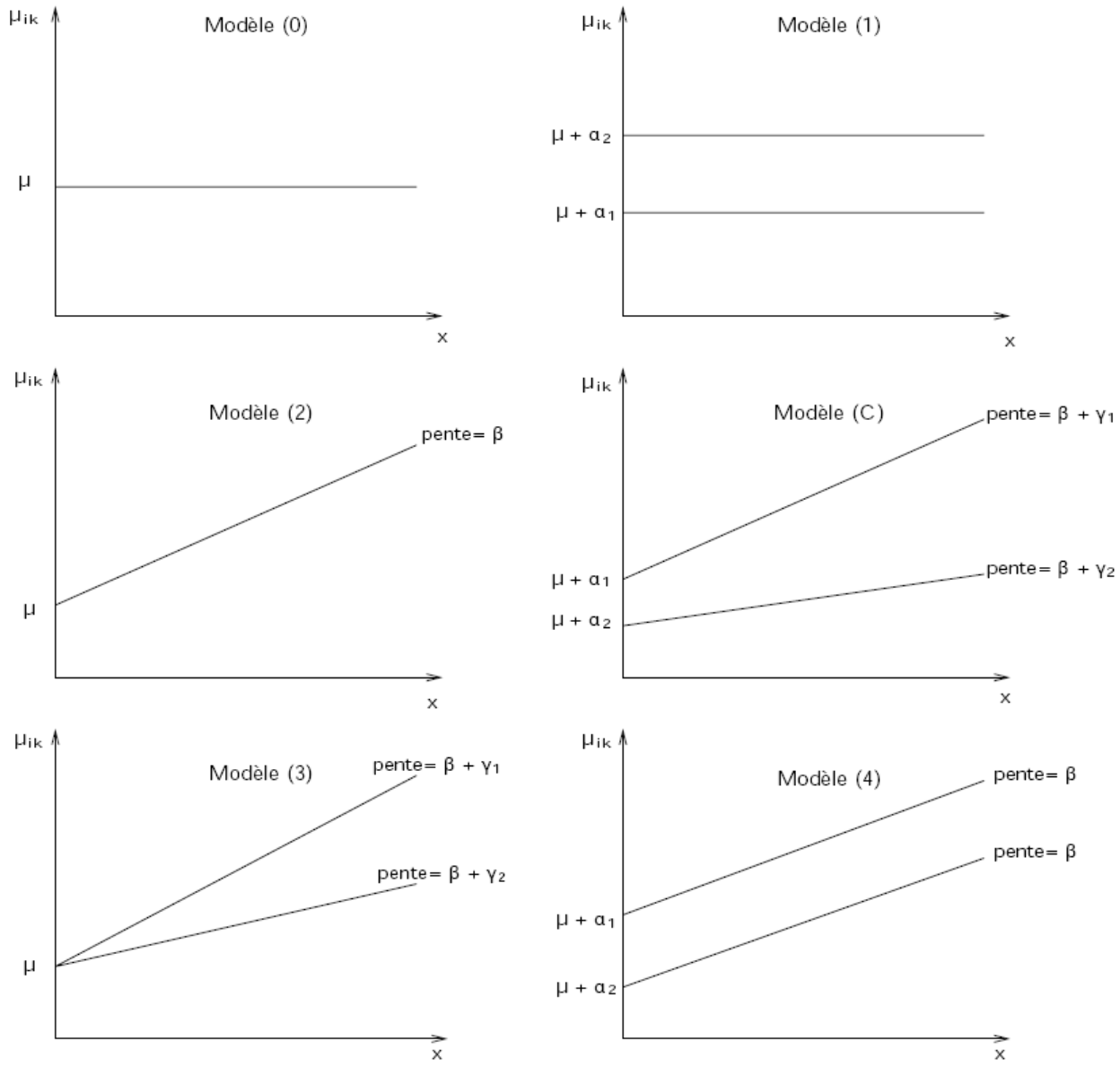

FIGURE 7.4 – Espérance de Y_{ik} dans les différents modèles avec $I = 2$.

par rapport au modèle (2), quelque soit le poids initial, il existe une différence de poids final entre les deux traitements qui est :

$$\mu_{1k} - \mu_{2k} = \alpha_1 - \alpha_2.$$

– **Modèle** (C) **:** s'il y a maintenant interaction entre la variable `PdsInit` et `Traitement`, les pentes de régressions sont égales à $(\beta + \gamma_1)$ pour le traitement 1, $(\beta + \gamma_2)$ pour

le traitement 2, ce qui signifie que les évolutions de poids sont différentes selon le traitement.

7.4.2 Enchaînement des tests des différents effets.

Les tests se déclinent exactement comme dans l'ANOVA : les tests des effets principaux (α et β) n'ont de sens qu'en l'absence d'interaction. On commence donc par tester l'interaction dans le modèle complet (C).

Test de l'interaction

On teste l'effet interaction, ce qui correspond au test du modèle (4) contre le modèle (C). La somme de carrés associée se lit aussi bien en type I qu'en type II puisque les tests de l'interaction sont les mêmes (comme dans l'ANOVA,cf chapitre 5). Le résultat de ce test est donné dans la table (7.7). La statistique de test pour cet effet est

$$F = \frac{R(\gamma/\mu, \alpha, \beta)/(I-1)}{\hat{\sigma}^2_{(C)}}.$$

Elle vaut ici $F = 1.22$. La probabilité critique associée vaut 0.36. On n e rejette pas l'hypothèses H_0 au niveau de 0.05, l'effet de l'interaction n'est pas significative. On conclut que la relation entre le poids final et le poids initial des sacs d'huîtres ne dépend pas des conditions dans lesquels on les a placés.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PdsInit	1	342.3578175	342.3578175	1208.03	<.0001
Traitement	4	12.0893593	3.0223398	10.66	0.0012
PdsInit*Traitement	4	1.3883141	0.3470785	1.22	0.3602

TABLE 7.7 – Sommes de carrés de Type I.

Test des α et β en l'absence d'intéraction

On se place donc dans le modèle (4) :

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + E_{ik}$$

Commençons par la validation des hypothèses et le test sur le modèle. La figure 7.5 représente le graphe des résidus contre les prédites. Il ne présente toujours aucune structure particulière, l'hypothèse d'homoscédasticité est validée. Le test du modèle se trouve dans la table d'analyse de la variance donnée dans la table 7.8. L'hypothèse H_0 est logiquement rejetée. On s'intéresse maintenant aux tests du facteur α et de la variable PoidsInitial. Les résultats de ces tests sont donnés dans cette même table, et les définitions

des sommes de carrés dues aux effets α et β en terme de réduction de types I et II sont présentées respectivement dans les tables 7.9 et 7.10.
Remarquons que les hypothèses du test sur β en type I ne sont pas exprimées en termes de test de modèles (emboîtés). En effet, il ne faut pas oublier que la résiduelle des différents tests est toujours celle du modèle complet $\sigma^2_{(4)}$. Or dans le test du modèle (0) contre le modèle (2), la résiduelle serait $\sigma^2_{(2)}$.

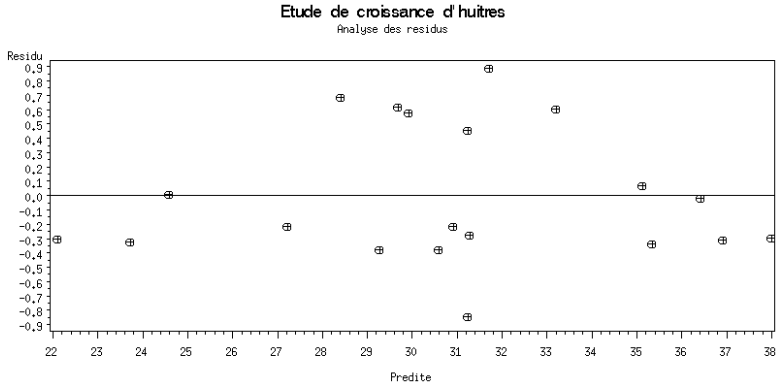

FIGURE 7.5 – Graphe des résidus contre les prédites dans le modèle sans interaction.

Comme dans l'ANOVA, lorsque le dispositif est non-orthogonal, il existe des différences entre les tests de type I et II. Ainsi, les deux tests sur β en type I et en type II sont différents. Pour s'en convaincre il suffit de regarder les modèles testés : $R(\beta/\mu)$ est la différence de sommes de carrés résiduelles entre le modèle (0) et le modèle (2), et $R(\beta/\mu, \alpha)$, celle entre le modèle (1) et le modèle (4).

Test de l'effet poids initial (ou pente moyenne, paramètre β). On s'intéresse donc au test de type II. La statistique et la probabilité critique se lisent dans la table 7.8 à la ligne `PdsInit` et la colonne `Type II SS`. La statistique de test s'écrit :

$$F = \frac{R(\beta/\mu, \alpha)}{\hat{\sigma}^2_{(4)}} = 517.38.$$

On rejette H_0 : la relation entre le poids final et le poids final est linéaire.

Remarque : La somme de carrés d'intérêt, ici $R(\beta/\mu, \alpha)$, peut aussi se lire dans la table correspondant au modèle complet écrit dans l'ordre suivant : $Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + E_{ik}$. La différence fondamentale pour le test sur β dans ce modèle avec le test précédent est la valeur de la statistique de test F, puisque elle dépend de la variance résiduelle $\hat{\sigma}^2$ qui n'est pas la même pour les deux tests. Néanmoins le fait que l'interaction soit très faiblement significative ne change pas la significativité de l'effet poids initial mais on imagine bien qu'une interaction plus forte pourrait amener à des conclusions différentes.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	354.4471767	70.8894353	235.05	<.0001
Error	14	4.2223233	0.3015945		
Corrected Total	19	358.6695000			

R-Square	Coeff Var	Root MSE	PdsFinal Mean
0.988228	1.780438	0.549176	30.84500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PdsInit	1	342.3578175	342.3578175	1135.16	<.0001
Traitement	4	12.0893593	3.0223398	10.02	0.0005

Source	DF	Type II SS	Mean Square	F Value	Pr > F
PdsInit	1	156.0401767	156.0401767	517.38	<.0001
Traitement	4	12.0893593	3.0223398	10.02	0.0005

TABLE 7.8 – Table d'analyse de la variance du modèle de covariance sans interaction.

effet	Somme de carrés	Test des Modèles	Degré de liberté
PdsInit (β)	$R(\beta/\mu)$	...	1
Traitement (α)	$R(\alpha/\mu, \beta)$	(2) contre (4)	$I-1$

TABLE 7.9 – Définitions des sommes de carrés de type I dans le modèle d'analyse de la covariance sans interaction.

effet	Somme de carrés	Test des Modèles	Degré de liberté
PdsInit (β)	$R(\beta/\mu, \alpha)$	(1) contre (4)	1
Traitement (α)	$R(\alpha/\mu, \beta)$	(2) contre (4)	$I-1$

TABLE 7.10 – Définitions des sommes de carrés de type II dans le modèle d'analyse de la covariance sans interaction.

Test de l'effet traitement (α). Comme le test sur β est significatif, on va tester α à l'aide de la somme des carrés $R(\alpha/\mu, \beta)$ qui peut donc se lire dans la colonne `type I` ou `type II` de la table 7.8. Remarque : si α avait été introduit dans le modèle en premier,

on aurait eu accès au test sur α à l'aide de la somme des carrés $R(\alpha/\mu)$, ce qui n'aurait pas eu de sens puisque l'on a déjà rejeté l'hypothèse $H_0 = \{\beta = 0\}$. En type II, on teste si la distance entre les droites est significativement différente de 0 à l'origine. Comme les pentes sont ici égales (pas d'interaction), la distance à l'origine est la même qu'en n'importe quel valeur du poids initial x (donc ce test à un sens). Nous renvoyons à la section 7.5 (comparaison des traitements).
La statistique du test est :

$$F = \frac{R(\alpha/\mu, \beta)/(I-1)}{\hat{\sigma}^2_{(4)}},$$

et vaut 517.36. La p.value est inférieure à 0.05, on conclut que quelque soit le poids initial, il existe une différence significative de le poids final d'au moins deux traitements.

Les droites de régression estimées sont représentées figure 7.6.

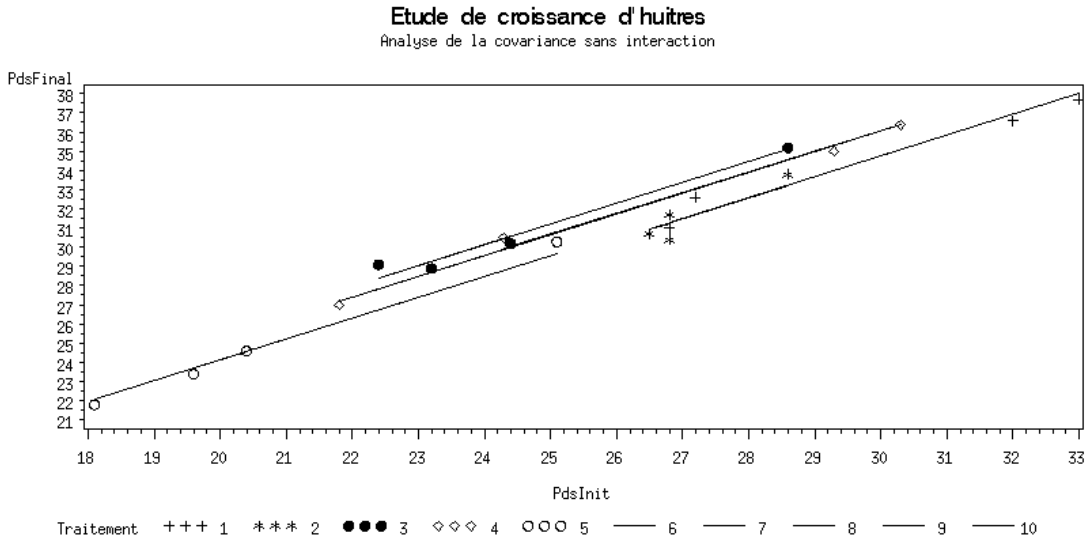

FIGURE 7.6 – Résultat de l'analyse de la covariance sans interaction.

7.5 Comparaison des traitements

Comme pour l'exemple de l'analyse de variance à 2 facteurs dans un plan en blocs incomplets (cf chapitre 6), on va voir que la comparaison des traitements à partir des moyennes classiques du poids final au sein de chaque traitement n'est pas pertinente et qu'il faut se fonder sur les moyennes ajustées.

Moyennes classiques. La moyenne classique du poids final pour le traitement i est $\mu_{i\bullet} = \mathbb{E}[Y_{i\bullet}]$ et s'obtient en moyennant les $\mu_{ik} = \mathbb{E}[Y_{ik}]$ sur K, le nombre de répétitions au

sein du traitement i :

$$\begin{aligned}\mu_{i\bullet} &= \frac{1}{K}\sum_{k=1}^{K}\mu_{ik} \\ &= \mu+\alpha_i+\beta\frac{1}{K}\sum_{k=1}^{K}x_{ik} = \mu+\alpha_i+\beta x_{i\bullet},\end{aligned}$$

et est estimée par

$$\hat{\mu}_{i\bullet} = \hat{\mu}_i = \hat{\mu}+\hat{\alpha}_i+\hat{\beta}x_{i\bullet}.$$

C'est donc le poids que l'on obtiendrait (ou la prédiction du poids final) dans le traitement i pour un sac de poids initial correspondant au poids initial moyen dans ce traitement $x_{i\bullet}$.

Les valeurs des moyennes des poids finaux et initiaux par traitement sont données dans la table 7.11. On remarque c'est dans l'emplacement correspondant au traitement 1 que le poids moyen des sacs est le plus important. En regardant les poids initiaux moyens par traitement, on observe que c'est aussi dans cet emplacement que les sacs étaient au départ en moyenne les plus lourds. On remarque effectivement que les poids moyens intitiaux par traitement sont différents. La figure 7.7 illustre ce que l'on comparerait en comparant deux traitements à partir de ces poids moyens. On imagine bien que pour comparer les poids finaux moyens, il faut que les poids initiaux moyens des sacs soient les mêmes. Il va donc falloir ajuster les poids finaux moyens au poids initial.

Level of		PdsFinal		PdsInit	
Traitement	N	Mean	Std Dev	Mean	Std Dev
1	4	34.4750000	3.18891309	29.7500000	3.20572405
2	4	31.6500000	1.53731367	27.1750000	0.96046864
3	4	30.8500000	2.95578529	24.6500000	2.75862284
4	4	32.2250000	4.29757684	26.4250000	4.04917687
5	4	25.0250000	3.69898635	20.8000000	3.02103735

TABLE 7.11 – Moyennes empiriques des poids finaux par traitement.

Moyennes ajustées. Comme on vient de le voir, la solution consiste à comparer les traitements à travers leurs poids finaux obtenus pour un même poids initial. Il s'agit de choisir une valeur de poids initial x_0 pertinente. Le plus naturel est de considérer le poids initial moyen à tous les traitements, i.e. le poids initial moyen à toutes les observations, noté $x_{\bullet\bullet}$ et définit par

$$x_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{5}\sum_{k=1}^{K}x_{ik}.$$

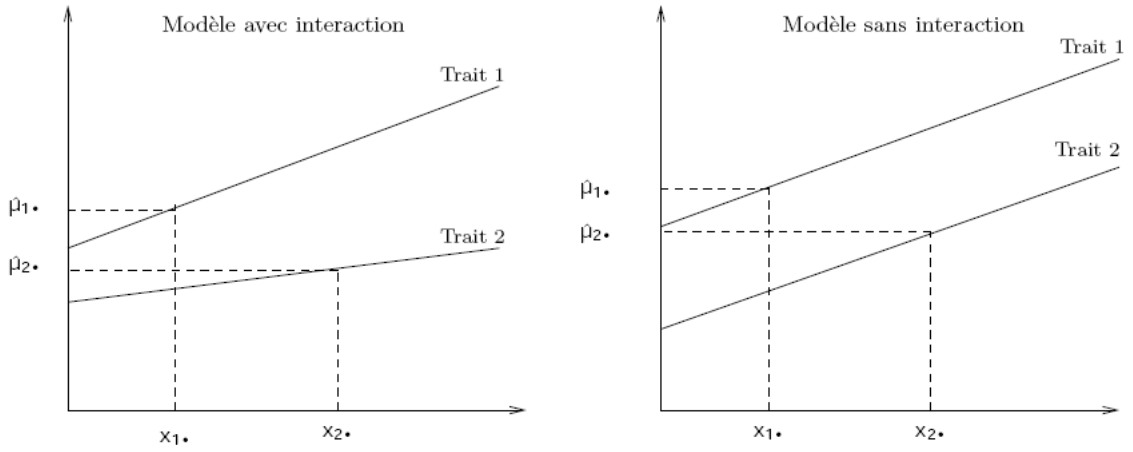

FIGURE 7.7 – Moyennes empiriques avec $I = 2$.

De ce point de vue, on définit la moyenne ajustée comme étant

$$\tilde{\mu}_{i\bullet} = \mu + \alpha_i + \beta x_{\bullet\bullet},$$

qui est estimée par

$$\hat{\tilde{\mu}}_{i\bullet} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta} x_{\bullet\bullet}.$$

La figure 7.8 illustre cette comparaison des 2 traitements pour le modèle sans interaction et le modèle avec interaction. On observe que pour le modèle sans interaction, quelque soit le poids initial x_0 choisi, on compare toujours la même différence entre les deux traitements qui est :

$$\alpha_1 - \alpha_2.$$

Pour le modèle avec interaction, cette valeur correspond à la différence de poids final entre les deux traitements pour un poids initial nul mais on voit que la comparaison dépend complètement de x_0.

Les moyennes ajustées sont données dans la table 7.12. On remarque que ce n'est plus le traitement 1 pour lequel le poids final moyen est le plus fort mais le traitement 3. En triant les traitements par ordre croissant de leurs poids moyens, on obtient

$$2 < 1 < 5 < 4 < 3,$$

ce qui ne coincide pas du tout avec celui que l'on obtiendrait à partir des moyennes classiques.

La table 7.13 donne les statistiques et les probabilités critiques de toutes les comparaisons deux à deux des tests des hypothèses où pour chaque test le niveau est de 0.05 :

$$H_0 = \{\tilde{\mu}_{i\bullet} = \tilde{\mu}_{j\bullet}\} \text{ contre } H_1 = \{\tilde{\mu}_{i\bullet} \neq \tilde{\mu}_{j\bullet}\}$$

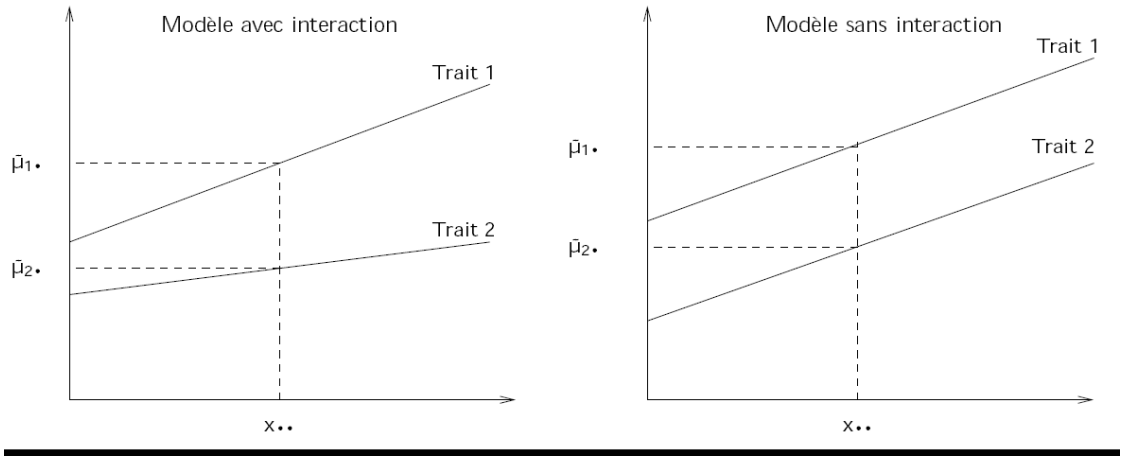

FIGURE 7.8 – Moyennes ajustées avec $I = 2$.

Traitement	PdsFinal LSMEAN	LSMEAN Number
1	30.1531125	1
2	30.1173006	2
3	32.0523296	3
4	31.5046854	4
5	30.3975719	5

TABLE 7.12 – Moyennes des poids finaux ajustées au poids initial par traitement.

Chaque test est effectué au niveau 0.05. On note des différences significatives entre les couples de traitements suivants : $(1, 3)$, $(1, 4)$, $(2, 3)$, $(2, 4)$ et $(3, 5)$.
En corrigeant par Bonferroni pour une comparaison par tests multiples (cf chapitre 6), le niveau de chaque test devient $0.05/10 = 0.005$, soit 0.5%, on fait ici 10 comparaisons. A ce niveau, seuls les couples $(1, 3)$, $(2, 3)$, $(2, 4)$ et $(3, 5)$ sont significatifs.

Test de l'effet du traitement sur le poids initial. La table 7.14 fournit le résultat du test de l'effet du facteur traitement sur le poids initial. La probabilité critique vaut 0.0093, on conclut que le test est significatif et donc qu'il existe un effet traitement sur le poids initial. On avait déjà pu observer une différence nette des poids initiaux entre les traitements à partir des données (cf table 7.1).

Critique sur la comparaison des traitements. Par le résultat précédent, on peut s'interroger sur la pertinence des comparaisons des moyennes ajustées. En effet, prenons par exemple les deux traitements 1 et 5. Comme on peut l'observer sur la figure 7.1, les poids initiaux au sein de ces deux traitements ne se chevauchent pas et la moyenne

i/j	1	2	3	4	5
1		0.087941	-4.1466	-3.22289	-0.42398
		0.9312	0.0010	0.0061	0.6780
2	-0.08794		-4.76003	-3.55771	-0.56861
	0.9312		0.0003	0.0032	0.5786
3	4.146599	4.76003		1.378002	3.853378
	0.0010	0.0003		0.1898	0.0018
4	3.222892	3.557715	-1.378		2.346817
	0.0061	0.0032	0.1898		0.0342
5	0.42398	0.568608	-3.85338	-2.34682	
	0.6780	0.5786	0.0018	0.0342	

TABLE 7.13 – Comparaisons des moyennes ajustées des 5 traitements.

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	4	176.7930000	44.1982500	4.98	0.0093
Error	15	132.9950000	8.8663333		
Corrected Total	19	309.7880000			

TABLE 7.14 – Table d'analyse de la variance décrivant l'effet du traitement sur le poids initial.

du poids initial générale $x_{\bullet\bullet} = 25.76$ n'entre pas dans les deux grilles de poids. Ainsi la moyenne ajustée associée, qui représente la prédiction du poids final à partir d'un poids initial de $x_{\bullet\bullet}$, n'est qu'une interpolation de ce qui pourrait se passer à ce poids à partir de ce qui se passe pour des poids initiaux bien différents. C'est donc un résultat à prendre avec précaution et la comparaison de ces deux traitements s'en trouve peu pertinente. On peut noter qu'un dispositif orthogonal ne poserait pas ce problème, puisque, rappelons le, les poids initiaux seraient les mêmes au sein des traitements.

7.6 Perspectives

L'objectif principal de l'étude est de comparer l'effet de différents traitements sur l'évolution du poids. Pour répondre à cette question, la connaissance du poids initial et du poids final nous a conduit à étudier l'existence de différences entre les relations entre le poids final et le poids initial selon les traitements. Mais d'autres modèles pourraient être envisagés.

Effet traitement sur l'accroissement de poids. Considérons la nouvelle variable suivante

$$Y_{ik} = PdsFinal_{ik} - PoidsInit_{ik},$$

qui est l'accroissement de poids du $k$ème sac dans le traitement i. Pour décrire l'influence du traitement sur l'accroissement (donc sur l'évolution de poids), on utilise alors un modèle d'analyse de la variance à un facteur (le traitement) :

$$Y_{ik} = \mu + \alpha_i + F_{ik} \qquad \{F_{ik}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2),$$

où α_i correspond à l'effet du traitement i.

Effet traitement sur l'accroissement de poids relatif. Notons maintenant

$$Y_{ik} = \frac{PdsFinal_{ik} - PoidsInit_{ik}}{PoidsInit_{ik}}$$

l'accroissement de poids relatif du $k$ème sac dans le traitement i. Le modèle s'écrit

$$Y_{ik} = \mu + \beta_i + G_{ik} \qquad \{G_{ik}\} \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2),$$

où β_i correspond à l'effet du traitement i.

7.7 Programme SAS

Données. Les instructions suivantes permettent d'importer les données dans une table qui s'appelle **huître**, d'imprimer (**proc Print**) cette table en sortie (table 7.1) et faire le graphique (**proc GPlot**) du poids final en fonction du poids initial codé par traitement (figure 7.1). Les données ont été préalablement triées par la procédure **proc Sort**.

```
data huitre;
     infile 'huitre.don' firstobs=2;
     input Traitement Repetition PdsInit PdsFinal;
proc Sort data=huitre;          % trie la table huître par traitement
     by Traitement;
proc Print data=huitre;
proc GPlot data=huitre;
     symbol1 i=none v=plus c=black;
     symbol2 i=none v=star c=black;
     symbol3 i=none v=dot c=black;
     symbol4 i=none v=diamond c=black;
     symbol5 i=none v=circle c=black;
     plot PdsFinal*PdsInit=Traitement;
run;
```

Statistiques élémentaires. Les statistiques élémentaires, générales puis par traitement, données dans la table 7.2 sont obtenues à partir des instructions suivantes :

```
title2 'Statistiques elementaires';
proc Means data=huitre;
     var PdsInit PdsFinal;
proc Sort data=huitre;
     by Traitement;
proc Means data=huitre;
     var PdsInit PdsFinal;
     by Traitement;
run;
```

Régressions simples par traitement. Cette étude n'a pas été analysée dans ce chapitre.

```
title2 'régressions separees';
proc Reg data=huitre;
     model PdsFinal = PdsInit;
     by Traitement;
run;
```

Analyse de la covariance avec interaction.

```
title2 'Analyse de la covariance avec interaction';
proc GLM data=huitre;
     class Traitement;     %class précise les variables qualitatives
     model PdsFinal = PdsInit Traitement PdsInit*Traitement / solution;
run;

title2 'Analyse des residus';
proc GPlot data=GLM;
     plot Residu*Predite='+' / vref=0;    %graphe des résidus
run;
```

L'instruction **model** permet de définir le modèle. Attention l'ordre d'écriture est important pour les types I, comme on a pu le voir.

Les tables de sorties de ces instructions sont les tables 7.3, 7.5, 7.7 et **??**.

Analyse de la covariance sans interaction.

```
title2 'Analyse de la covariance sans interaction';
proc GLM data=huitre;
```

```
     class Traitement;
     model PdsFinal = PdsInit Traitement / solution;
     means Traitement;
     lsmean Traitement / tdiff pdiff;
     output out=GLM p=Predite r=Residu;
```

L'instruction **means Traitement ;** donne les moyennes des poids finaux et initiaux par traitement.

L'instruction **lsmean Traitement / tdiff pdiff ;** donne les moyennes des poids initiaux par traitement ajustées au poids initial. Les options **tdiff pdiff** permettent de disposer de la statistique de test et de la probabilité critique de chaque test de comparaison des moyennes ajustées deux à deux.

Les tables de sorties de ces instructions sont les tables 7.8, 7.11, 7.13.

```
title2 'Analyse des residus';
proc GPlot data=GLM;
     plot Residu*Predite='+' / vref=0;
run;

data GLM;
     set GLM;
     Traitement=Traitement+5;
     PdsFinal=Predite;
data GRAPH;
     set huitre GLM;
proc GPlot data=GRAPH;
     symbol7 i=join v=none c=black;
     symbol8 i=join v=none c=black;
     symbol9 i=join v=none c=black;
     symbol10 i=join v=none c=black;
     symbol11 i=join v=none c=black;
     plot PdsFinal*PdsInit=Traitement;
run;
```

La table appellée **GLM** créée par la procédure **Data**, correspond à l'ancienne table GLM qui contient les résultats de l'ANCOVA, dans laquelle on change la variable Traitement par Traitement+5 pour qu'ils correspondent aux symboles utilisés dans la procédure **Gplot** et la variable PdsFinal (les poids finaux) par Predite (les valeurs prédites).

L'instruction **set huitre GLM** donne l'ordre de coller les deux tables huitre et GLM.

Analyse de la variance sur le poids initial. La table 7.14 est obtenue par les instructions suivantes :

```
proc GLM data=huitre;
     class Traitement;
     model PdsInit = Traitement;
run;
```

Bibliographie

Bergonzini, J.-C. et Duby, C. (1995). *Analyse et planification des expériences*. Masson.

Dagnélie, P. (1981). *Principes d'expérimentation*. Presse Agronomiques de Gembloux.

Daudin, J.-J., Robin, S. et Vuillet, C. (1999). *Statistique inférentielle : idées, démarches, exemples*. Presses Universitaires de Rennes / Société Française de Statistiques.

Daudin, J., Lebarbier, E. et Vuillet, C. (2007). *Bases du Modèle Linéaire*. AgroParisTech.

Duby, C., (2000). *Le modèle linéaire*. Institut National Agronomique Paris-Grignon.