République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

KHOULOUD AIDAOUI

Titre :

Analyse de covariance

Membres du Comité d'Examen :

Dr.	**BENELMIR Imen**	UMKB	Président
Dr.	**ROUBI Affef**	UMKB	Encadreur
Dr.	**OUANOUGHI Yasmina**	UMKB	Examinateur

2019

Dédicace

Je dédie ce modeste travail

A mes très chers parents **DJAMAL**, **AKILA** qui ont bien élevés, aidés, soutenus et encouragés durant toutes ces années d'étude, qu'**ALLAH** les protèges.

A mon grand père et grande mère pour tendresse, je lui souhaite une langue vie.

A mon chère frère **ABDELHAI** pour son aide morale

A mes très chères sœurs : **HADIL** et **TASNIM** pour leurs précieuses aide et ses merveilleux conseils.

A mes oncles et leurs fils et filles surtout **LINDA** et **NOUR** et **CHEHRAZED** et **HOUDA** et **ILHAM** et **ALA** et toute la famille **AIDAOUI**

A mes chères amies

A ceux et celle que je connais, que me connais, que je serai connais.

A tout la promotion 2eme Master mathématique en particulier statistique.

2018-2019

A Tous ceux que j'ai oublié de mentionner leurs noms.

REMERCIEMENTS

Avant tout choses, je remercie Dieu le tout puissant, pour m'avoir donnée la force et la patience, la santé et la volonté pour réaliser ce modeste travail.

Je tiens à remercier sincèrement Melle Roubi affef mon encadreur, qu'il trouve ici l'expression de ma profonde reconnaissance pour avoir guidées dans mon travail. Ses conseils, ses orientations, sa patience, et sa correction sérieuse de ce travail.

Mes remerciements infinis aux membres des jurys qui nous a fait l'honneur D'accepter de jurer et évaluer ce travail.

Je n'oublie pas de remercier vivement Le chef département et tous mes enseignants, pour les informations et les aides au coures des années de mes études.

Un grand merci particulier à mes collègues et mes amies pour les sympathiques moments qu'on a passés ensemble, on les remercie pour leur confiance, et leur soutien moral au cours de ces années.

Que tous ceux, que je n'ai pas nommés.

Table des matières

Remerciements ii

Table des matières iii

Liste des figures v

Introduction 1

1 Analyse de variance et régression linéaire simple 3

1.1 ANOVA à un facteur (ANOVA1) 3

1.1.1 Généralités 3

1.1.2 Objectif 4

1.1.3 Structure des données 4

1.1.4 Modèle d'ANOVA 1 5

1.1.5 Les étapes de l'ANOVA 1 6

1.2 La régression linéaire 9

1.2.1 Analyse du modèle de régression linéaire simple 9

2 Analyse de covariance à un facteur et une covariable 14

2.1 Choix des covariables 14

2.2 Présentation des données d'une ANCOVA 1 15

2.3 Modèle d'ANCOVA 1 16

2.4 Moyennes ajustées 17
2.5 La réalisation de L'ANCOVA 1 18
2.5.1 Vérification des conditions d'applicabilité 18
2.5.2 Calcul des : moyennes, sommes des carrés et des produits 18
2.5.3 Calcul des sommes des carrés ajustées 19
2.5.4 Tests d'hypothèse 22
2.6 Comparaison des droites de régression 23
2.7 Exemple d'application 25

3 Application sous R **30**
3.1 Exemples sur l'ANOVA 1 et sur la régression linéaire simple 30
3.1.1 Exemple sur l'ANOVA 1 30
3.1.2 Exemple sur la régression linéaire simple 33
3.2 Exemple de l'analyse de covariance ANCOVA 1 36

Conclusion **45**

Bibliographie **45**

Annexe A : Logiciel *R* **46**

Annexe B : Abréviations et Notations **47**

Table des figures

3.1 Les boîtes à moustaches de la variable quantité en fonction de la variable graisse. 32
3.2 Nuage de points du poids des fils en fonction du poids des pères 34
3.3 Représentation de la droite de régression des moindres carrés sur le nuage de 35
3.4 Boites à moustaches de logarithme népérien de taux de leucocytes T4 (a) et de nombre de jours après avoir inoculé à l'animal le virus du sida (b) en fonction du sexe. 38
3.5 Nuage de points de la mesure de LNT4 en fonction de jours. 40

Introduction

La statistique est l'étude d'un phénomène par la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques. La statistique appliquée est utilisée dans presque tous les domaines de l'activité humaine, ingénierie, management, économie, biologie, informatique, etc. La statistique utilise des règles et des méthodes sur la collecte des données, pour que celles-ci puissent être correctement interprétées, souvent comme composante d'une aide à la décision, en étant conscient qu'il y a un risque d'erreur lié à l'incertitude des observations ou des résultats expérimentaux, avant de prendre une telle décision, on testera une hypothèse statistique correspondant à notre problème.

Une hypothèse statistique est un énoncé (une affirmation) concernant les caractéristiques (valeurs des paramètres, forme de la distribution des observations) d'une ou de populations,

Un test statistique est un ensemble de règles par lesquelles on arrive à prendre une décision concernant les hypothèses.

Dans le cadre des tests d'hypothèses, nous avons émis des hypothèses concernant l'effet des variables qualitatives à plusieurs niveaux sur une variable quantitative ou l'effet des variables quantitatives sur une variable quantitative. L'analyse de la variance et la régression sont les méthodes employées pour traiter ces hypothèses respectivement.

Un mélange de ces deux méthodes c'est à-dire d'ANOVA et de la régression linéaire constituée une autre méthode appelée analyse de la covariance ou ANCOVA. Cette dernière est

une méthode statistique visant à tester, par un modèle linéaire général, l'effet sur une variable dépendante continue d'une ou de plusieurs variables indépendantes catégorielles, indépendamment de l'effet des autres facteurs quantitatif continus, dits covariables. L'ANCOVA permet de tester si certains facteurs ont un effet sur la variable dépendante après avoir enlevé la variance due aux covariables.

Dans ce mémoire, composé de trois chapitres, on s'intéresse à cette dernière méthode, et au cas où seulement une variable, parmi les variables explicatives, est quantitative et l'autre est qualitative.

Chapitre 1 : Nous traitons dans ce chapitre, la technique d'analyse de la variance à un facteur (ANOVA 1), leurs principes, ainsi leurs différentes étapes les plus indispensable. Aussi nous parlons sur la régression linéaire simple pour mener à faire une compilation entre les deux techniques.

Chapitre 2 : Ce chapitre est consacré à l'étude en détails de la méthode de l'ANCOVA. Cette méthode qui permet de combiner les éléments des modèles de régression et les modèles d'analyse de la variance a pour but de comparer les moyennes ajustées et non arithmétiques.

Chapitre 3 : Ce dernier chapitre est consacré à l'application de tous ce que nous avons parlés dans les chapitres précédents sur des données réelles sous le logiciel R.

Chapitre 1

Analyse de variance et régression linéaire simple

Ce chapitre est consacré à l'étude de deux techniques statistiques quelles sont l'analyse de la variance à un seul facteur et la régression linéaire simple, ces méthodes ont pour but d'étudier l'effet d'une variable qualitative ou quantitative sur une variable quantitative.

1.1 ANOVA à un facteur (ANOVA1)

1.1.1 Généralités

L'analyse de variance est une technique qui consiste à séparer la variation totale d'un ensemble de données en composantes raisonnées associées à des sources spécifiques de variation, dont le but est de comparer plusieurs moyennes des populations considérées et de conclure à l'égalité ou à la non-égalité globale de toutes les moyennes.

Elle permet également de tester certaines hypothèses concernant les paramètres du modèle, ou d'estimer les composantes de la variance.

Les sources de variation se résument globalement en une composante appelée "erreur" et une autre composante que l'on pourrait désigner par le terme "effet" (bien que, selon les

cas, ce terme prenne des noms plus précis et puisse encore se subdiviser). ([3], [5] ,[6])

1.1.2 Objectif

L'analyse de la variance (ANOVA) est une méthode qui permet d'étudier la modification de la moyenne μ du phénomène étudié Y (variable quantitative) selon l'influence d'un ou de plusieurs facteurs d'expérience qualitatifs (traitements). Dans le cas où la moyenne n'est influencée que par un seul facteur, il s'agit d'une analyse de la variance à un facteur. Un facteur est souvent une variable qualitative présentant un nombre restreint de modalités. Le nombre de modalités (c'est-à-dire de niveaux) du facteur A sera noté p. On suppose que Y suit une loi normale $N(\mu, \sigma^2)$ sur chaque sous-population i définie par les modalités de A.

L'objectif est de tester l'égalité des moyennes de ces p populations, à savoir de tester les hypothèses

$$\begin{cases} H_0 : "\mu_1 = \mu_2 = ... = \mu_p = \mu" \\ H_1 : "\exists\ i, j \in \{1, 2, ..., p\} \text{ tel que } \mu_i \neq \mu_j" (i \neq j) \end{cases} . \tag{1.1}$$

1.1.3 Structure des données

L'analyse de la variance est une technique statistique qui sert à tester l'influence d'un ou plusieurs facteurs qualitatifs sur une variable quantitative.

Notons A le facteur et $A_1, A_2, ..., A_p$ ses p modalités, soit Y la variable étudiée.

Donc pour chaque niveau A_i du facteur on a réalisé n_i mesures de $Y(Y_{i1}, Y_{i2}, ..., Y_{in_i}$, $i = \overline{1, p})$

N°	A_1	A_2		A_i		A_p
1	Y_{11}	Y_{21}	$\cdots\cdots\cdots$	Y_{i1}	$\cdots\cdots\cdots$	Y_{p1}
2	Y_{12}	Y_{22}	$\cdots\cdots\cdots$	Y_{i2}	$\cdots\cdots\cdots$	Y_{p2}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots\cdots$	$\vdots$	$\cdots\cdots\cdots$	$\vdots$
n_i	Y_{1n_1}	Y_{2n_2}	$\cdots\cdots\cdots$	Y_{in_i}	$\cdots\cdots\cdots$	Y_{pn_p}

TAB. 1.1 – Les données d'ANOVA 1.

1.1.4 Modèle d'ANOVA 1

Le modèle mathématique leurs associés est donné par

$$Y_{ij} = \mu_i + \varepsilon_{ij}; i = \overline{1,p},\ j = \overline{1,n_i} \text{ et } \varepsilon_{ij} \sim N(0,\sigma^2). \tag{1.2}$$

On peut aussi écrire μ_i sous la forme suivante

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

Le modèle linéaire ci-dessus peut donc aussi s'écrire sous la forme

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

où

- Y_{ij} représente la $j^{ème}$ observation recevant le traitement i,
- μ la moyenne générale commune à tous les traitements,
- α_i est l'effet sur l'observation du traitement i,
- ε_{ij} est l'erreur expérimentale de l'observation Y_{ij}.

Le test d'hypothèse associé à ce modèle est

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0 \\ H_1 : "\exists\ i \in \{1,2,...,p\} \text{ tel que } \alpha_i \neq 0 \end{cases}.$$

1.1.5 Les étapes de l'ANOVA 1

La mise en oeuvre d'une ANOVA 1, se fait principalement en 4 étapes. Les détails de ces étapes sont comme suit

Étape 1 (conditions)

Afin de réaliser une analyse de variance à deux facteurs, les conditions suivantes doivent être vérifiées préalalement

• Les p échantillons comparés sont indépendants.

• La variable quantitative étudiée suit une loi normale dans les p populations comparées.

• Les p populations comparées ont même variance : homogénéité des variances ou homoscédasticité.

Étape 2 (Moyennes et variances)

Quantifier les différentes statistiques intervenant dans l'ANOVA à un facteur et qui sont

• La moyenne de toutes les observations

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij} \text{ avec } n = \sum_{i=1}^{p} n_i.$$

• La moyenne de chaque échantillon

$$\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \text{ pour } i = \overline{1, p.}$$

• La variance de toutes les observations

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2.$$

• La variance de chaque échantillon

$$\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2, \text{ pour } i = \overline{1,p}.$$

On peut démontrer facilement que la variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances des p échantillons, c'est-à-dire

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i\hat{\sigma}_i^2 + \frac{1}{n}\sum_{i=1}^{p} n_i(\overline{Y}_i - \overline{Y})^2, \tag{1.3}$$

ou encore

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2 + \frac{1}{n}\sum_{i=1}^{p} n_i(\overline{Y}_i - \overline{Y})^2. \tag{1.4}$$

on multipliant(1.4)par n on obtient

$$\underbrace{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2}_{SCE} + \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{Y}_i - \overline{Y})^2}_{SCF}. \tag{1.5}$$

où

SCF : est la variation due au facteur,

SCE : est la variation résiduelle,

SCT : est la variation totale.

Étape 3 (Moyenne des carrés)

L'idée la plus naturelle est que le facteur n'a pas d'impact sur le caractère étudié si la variation totale n'est engendrée que par la variation résiduelle associée au caractère, c'est-à-dire :

• Si H_0 est vraie, alors la variation SCF due au facteur doit être petite par rapport à la variation résiduelle SCE.

• Par contre, si H_1 est vraie alors la variation SCF due au facteur doit être grande par rapport à la quantité SCE.

Pour comparer ces quantités, Fisher a considéré le rapport moyenne des carrés associés au facteur CMF et moyenne des carrés résiduels CME

$$CMF = \frac{SCF}{p-1} \text{ et } CME = \frac{SCE}{n-p}.$$

Si les 3 conditions d'application d'ANOVA sont vérifiées et H_0 est vraie, alors la statistique

$$F = \frac{CMF}{CME}.$$

suit une loi de Fisher de $(p-1)$ et $(n-p)$ degrés de liberté ($F \sim f_{(p-1,n-p)}$)

Étape 4 (Décision)

Pour un seuil de risque donné les tables de Fisher nous fournissent une valeur critique $f_{(\alpha,p-1,n-p)}$ telle que

$$P(\frac{CMF}{CME} < f_{(\alpha,p-1,n-p)}) = 1-\alpha. \tag{1.6}$$

• Si $f < f_{(\alpha,p-1,n-p)} \Longrightarrow$ on ne peut pas rejeter H_0 (il n'y a pas d'influence du facteur).

• Si $f \geq f_{(\alpha,p-1,n-p)} \Longrightarrow$ on rejette H_0 (il y a une influence du facteur), avec f est la réalisation de la variable (statistique) F.

Les résultats d'une ANOVA 1 sont souvent présentés dans un tableau sous la forme suivante

Source de variation	Degrés de libertés ddl	Somme des carrés SC	Carré moyen CM	Ratio f
Facteur.A	$p-1$	SCF	CMF	$\frac{CMF}{CME}$
Résiduelle	$n-p$	SCE	CME	
Total	$n-1$	SCT		

TAB. 1.2 – Tableau d'analyse de variance à un facteur.

1.2 La régression linéaire

En statistiques, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

Parmi les modèles de régression linéaire, le plus simple est l'ajustement affine. Celui-ci consiste à rechercher la droite permettant d'expliquer le comportement d'une variable statistique Y comme étant une fonction affine d'une autre variable statistique X.

En général, le modèle de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de Y sachant X est une transformation affine en les paramètres. ([2], [3])

1.2.1 Analyse du modèle de régression linéaire simple

Modèle de régression linéaire simple

Un modèle de régression linéaire simple est défini par une équation de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{1.7}$$

où

- Y est la variable à expliquer (à valeurs dans $\mathbb{R}$),
- β_0 et β_1 sont deux paramètres à estimer,
- X est la variable indépendante (variable explicative),
- ϵ est une erreur aléatoire.

Pour n observations, on peut écrire le modèle de régression linéaire simple sous la forme

$$y_i = \beta_0 + \beta_1 x_i \ + \epsilon_i \ \text{, pour } i = \overline{1,n}.$$

On fait les trois hypothèses additionnelles suivantes $E(\varepsilon_i) = 0$, $cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$ et $var(\varepsilon_i) = \sigma^2$,$\forall i = \overline{1,n}$.

On peut écrire matriciellement le modèle de la manière suivante

$$Y = X\ \beta + \epsilon, \tag{1.8}$$

tel que

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \text{et } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}.$$

- Y désigne le vecteur à expliquer de taille (n,1),
- X la matrice explicative de taille (n,2),
- ϵ le vecteur d'erreurs de taille (n,1).

Estimation des paramètres du modèle

Supposons qu'on opte pour la méthode de moindre carrées pour quantifier β_0 et β_1, alors les estimateurs des paramètres β_0 et β_1 sont $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimise la fonction $\Psi(\beta_0, \beta_1)$, définie par

$$\Psi(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{1.9}$$

Cela revient à la détermination d'un optimum minimal de la fonction des erreurs quadratique $\Psi(\beta_0, \beta_1)$, qui consiste à résoudre le système des équations suivant

$$\begin{cases} \frac{\partial \Psi(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial \Psi(a_1, \beta_1)}{\partial \beta_1} = 0 \end{cases}, \tag{1.10}$$

c'est-à-dire,

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \Longrightarrow \begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \beta_1\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \beta_0\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2 = 0 \end{cases}.$$

Finalement, le système à résoudre, pour estimer les coefficients de régression β_0 et β_1, ni rien d'autre qu'un système linéaire à deux équations et à deux inconnus, qui est donné par

$$\begin{cases} \beta_0(\sum_{i=1}^{n} 1) + \beta_1(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} y_i \\ \beta_0(\sum_{i=1}^{n} x_i) + \beta_1(\sum_{i=1}^{n} x_i^2) = \sum_{i=1}^{n} x_i y_i \end{cases}. \tag{1.11}$$

La résolution du système(1.11) nous fournis la solution suivante

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}^2} \text{ et } \hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n} x_i.$$

Ou encore

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \tag{1.12}$$

et

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

tel que

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}),\ s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2.$$

L'estimation de la fonction de régression s'écrit

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \tag{1.13}$$

Test sur la validité du modèle

Sous l'hypothèse de normalité des erreurs on peut construire le test de validation du modèle.

En effet, la variation totale de Y se décompose comme suit

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{Y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{SCE}+\underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\bar{Y})^2}_{SCR},$$

où

SCT : Variation de Y ou variation totale.

SCE : Variation des résidus.

SCR : Variation de la régression ou variation expliquée par la régression.

On déduit de cette décomposition une mesure de qualité de l'ajustement appellé coefficient de détermination, défini par

$$R^2=\frac{SCR}{SCT}\text{ où }0\leq R^2\leq 1. \tag{1.14}$$

On peut aussi s'écrire en fonction des résidus

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

Pour valider le modèle, on test

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

La statistique du test est la suivante

$$F=\frac{\sum_{i=1}^{n}(\widehat{y}_i-\bar{Y})^2/1}{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2/(n-2)}\sim f_{(1,n-2)}, \tag{1.15}$$

où $f_{(1,n-2)}$ désigne une loi de Fisher de degrés de liberté $n_1 = 1$ et $n_2 = n - 2$.

Ainsi, pour un risque α on décide que

• si $f > f_{(\alpha,1,n-2)}$ alors le modèle est valide,

• si $f \leq f_{(\alpha,1,n-2)}$ le modèle n'est pas valide,

dont f est la réalisation de la statistique F et $f_{(\alpha,1,n-2)}$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher de degrés de liberté 1 et $(n-2)$.

Chapitre 2

Analyse de covariance à un facteur et une covariable

Dans ce chapitre on s'intéresse à une technique statistique qui combine certaines des caractéristiques de l'analyse de variance et de la régression linéaire appelée l'analyse de la covariance (en abréviation ANCOVA).

L'idée à la base d'analyse de covariance est d'ajouter à un modèle d'analyse de la variance, associe à une ou plusieurs variables qualitatives, une ou plusieurs variables quantitatives qui pourraient être liées à la réponse étudiée.

L' ANCOVA est une style d'estimation et de test des effets des traitements, elle permet de déterminer s'il existe une différence significative entre plusieurs moyennes de traitements en tenant compte des valeurs observées d'une ou des variables quantitatives dites covariables.

2.1 Choix des covariables

Nous appelons une variable quantitative qui est ajoutée à un modèle d'ANOVA une covariable.

Le choix des covariables est un processus très important. S'il s'avère que les variables retenues n'ont aucun lien avec la réponse étudiée, le gain du modèle d'ANCOVA par rapport

à celui du modèle d'ANOVA ne sera inexistant et nous retiendrons vraisemblablement au final ce modèle plus simple.

Les covariables X_1, $X_2, ...$ sur lesquelles la variable Y est ajustée doivent être quantitatives, mesurables et corrélées linéairement avec cette dernière. En effet, en l'absence de corrélation linéaire, l'ajustement par l'ANCOVA ne présente aucun intérêt puisque les résultats avec ou sans ajustement deviennent identiques ou très proches. S'il existe une relation curvilinéaire entre X et Y, une transformation de la covariable ou l'ajout de termes polynomiaux dans le modèle ($\beta_j X + \beta_{j+1} X^2 + ...$) devient nécessaire.

2.2 Présentation des données d'une ANCOVA 1

Sur un échantillon de n individus, on observe deux variables quantitatives X et Y et une variable qualitative A. La variable quantitative Y est la variable réponse que l'on cherche à expliquer en fonction de la variable quantitative X (covariable) et de facteur A à p niveaux.

Chaque individu de l'échantillon est repéré par un double indice (i, j), i représentant le niveau du facteur A auquel appartient l'individu, et j correspondant à l'indice de l'individu dans le niveau i ($j = \overline{1, n_i}$). Pour chaque individu (i, j), on dispose d'une valeur X_{ij} de la variable X et d'une valeur Y_{ij} de la variable Y.

$n = \sum_{i=1}^{p} n_i$ est le nombre d'observations.([3])

2.3 Modèle d'ANCOVA 1

Le modèle linéaire que nous devons considérer est le suivant

$$Y_{ij} = \mu + \alpha_i + \beta\left(X_{ij} - \bar{X}\right) + \varepsilon_{ij}, \text{ tel que } i = \overline{1,p},\ j = \overline{1,n_i},\ \varepsilon_{ij} \sim N(0,\sigma^2), \tag{2.1}$$

où

- Y_{ij} représente la j-ème observation recevant le traitement i,
- μ la moyenne générale commune à tous les traitements,
- α_i l'effet réel sur l'observation du traitement i,
- X_{ij} la valeur de la covariable,
- ε_{ij} l'erreur expérimentale de l'observation Y_{ij},
- $\bar{X}$ la moyenne générale de X.

Les estimateurs des paramètres du modèle sont donnés par les formules suivantes

$$\begin{aligned}
\hat{\mu} &= \bar{Y};\\
\hat{\alpha}_i &= \bar{Y}_i - \hat{\beta}(\bar{X}_i - \bar{X}) - \hat{\mu}, \quad i = \overline{1,p};\\
\hat{\varepsilon}_{ij} &= Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}(X_{ij} - \bar{X}), \quad i = \overline{1,p}, \quad j = \overline{1,n_i};
\end{aligned}$$

où $\hat{\beta}$ est donné par

$$\hat{\beta} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}. \tag{2.2}$$

Selon ce modèle le test à réaliser est le suivant

$$\begin{cases} H_0 : & \alpha_1 = \alpha_2 = ... = \alpha_p = 0 \\ H_1 : & \exists i \in \{1,2,...,p\} \text{ tel que } \alpha_i \neq 0 \end{cases}. \tag{2.3}$$

2.4 Moyennes ajustées

Considérant l'équation(2.1), la variable dépendante peut être ajustée pour la covariable comme suit

$$Y_{ij\ aj} = Y_{ij} - \beta(X_{ij} - \bar{X}) = \mu + \alpha_i + \varepsilon_{ij},\ \forall i = \overline{1,p},\ \forall j = \overline{1,n_i}; \tag{2.4}$$

où Y_{ijaj} est la valeur de la variable dépendante ajustée pour la covariable. Comme dans l'analyse de la variance, le but d'analyse de covariance est de tester l'égalité des moyennes (de la variable dépendante) des groupes, mais ces moyennes sont des moyennes ajustées (ou moyennes conditionnelles), alors le test défini dans (2.3) peut être réécrit de la forme suivante

$$\begin{cases} H_0 : & \mu_i\backslash_{(X=\bar{X})} = \mu_{i'}\backslash_{(X=\bar{X})}, \forall i, \grave{\imath} \in \{1,2,...,p\} \\ H_1 : & \exists i, \grave{\imath} \in \{1,2,...,p\}\,,\ \mu_i\backslash_{(X=\bar{X})} \neq \mu_{\grave{\imath}}\backslash_{(X=\bar{X})} \end{cases}.$$

L'estimation des moyennes ajustées est donnée par

$$\hat{\mu}_i\backslash_{(X=\bar{X})} = \bar{Y}_{i\ aj} = \bar{Y} - \hat{\beta}(\bar{X}_i - \bar{X}), \forall i = \overline{1,p};$$

tel que $\hat{\beta}$ est défini par (2.2).

Remarque 2.4.1 - *L'analyse de covariance élimine les différences entre les moyennes observées relatives à la covariable en ramenant les valeurs de la variable dépendante à une même valeur de référence de la covariable, telle que la moyenne générale de cette variable.*
- Il apparaît de l'équation (2.1) que l'analyse de la covariance peut être considérée comme une analyse de la variance sur les valeurs ajustées de Y à condition que le paramètre β doit être connu.
- Parmi les méthodes alternatives existantes dans la littérature pour répondre à notre objective une méthode qui sera étudiée dans la section suivante consiste à calculer des sommes des carrés et des carrés moyens appropriés. Ces sommes sont ajustées pour la covariable

2.5 La réalisation de L'ANCOVA 1

De même que l'analyse de la variance, exposée dans le chapitre précédent, la réalisation de l'ANCOVA 1 peut être ordonnée comme suit

2.5.1 Vérification des conditions d'applicabilité

Afin de réaliser le test défini dans (2.3), les conditions suivantes doivent être vérifiées préalablement

1. Les conditions d'application de l'ANOVA 1.
2. La liaison entre la covariable X et la variable dépendante Y est linéaire.
3. Absence d'effet du facteur étudié sur la covariable X.
4. Absence d'interaction entre le facteur étudier et le rapport entre la covariable et la variable dépendante.

2.5.2 Calcul des : moyennes, sommes des carrés et des produits

Notons $\bar{X}_i$ et $\bar{Y}_i$ respectivement les moyennes des valeurs de X et de Y pour le traitement i $(i = \overline{1,p})$ et $\bar{X}$ et $\bar{Y}$ respectivement les moyennes de toutes les valeurs de X et de Y. Pour pouvoir effectuer le test de Fisher qui va permettre de déterminer s'il existe une différence significative entre les traitements, nous avons besoin des sommes des carrés et des sommes des produits suivantes**([5])**

1. La somme des carrés totale pour X

$$(SCT)_X = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right)^2 . \tag{2.5}$$

2. La somme des carrés totale pour Y (SCT) elle se calcule de la même manière que $(SCT)_X$ en remplaçant les X par des Y.

3. La somme des produits totale de X et Y

$$SPT = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right) \left(Y_{ij} - \bar{Y}\right). \tag{2.6}$$

4. La somme des carrés des traitements pour X

$$(SCF)_X = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(\bar{X}_i - \bar{X}\right)^2. \tag{2.7}$$

5. La somme des carrés des traitements pour Y (SCF) elle se donne comme $(SCF)_X$ dont en remplaçant les X par des Y.

6. La somme des produits des traitements de X et Y

$$SPF = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(\bar{X}_i - \bar{X}\right) \left(\bar{Y}_i - \bar{Y}\right). \tag{2.8}$$

7. La somme des carrés des erreurs pour X

$$(SCE)_X = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2. \tag{2.9}$$

8. La somme des carrés des erreurs pour Y (SCE) elle se calcule par la formule(2.9) dont en remplaçant les X par des Y.

9. La somme des produits des erreurs de X et Y

$$SPE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right) \left(Y_{ij} - \bar{Y}_i\right). \tag{2.10}$$

2.5.3 Calcul des sommes des carrés ajustées

Si l'on souhaite prédire la valeur d'une observation particulière sur la variable dépendante en connaissant la moyenne de l'échantillon pour la variable dépendante et la valeur de la covariable associée, mais en ignorant le groupe ou le traitement associé à chaque

observation, une approche raisonnable consisterait à utiliser la valeur de la moyenne de l'échantillon ajustée par la valeur de la covariable. Plus précisément, le modèle utilisé pour produire cette estimation serait([11])

$$\hat{Y}_{ij} = \bar{Y} + \hat{\beta}_T \left(X_{ij} - \bar{X}\right), \ \forall i = \overline{1,p}, \ \forall j = \overline{1,n_i}. \tag{2.11}$$

tel que

$\hat{Y}_{ij}$ la valeur prédit pour l'observation j dans le groupe i,

$\hat{\beta}_T$ l'estimateur de coefficient de régression donné par(1.12), ou plus précisément par

$$\hat{\beta}_T = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)\left(Y_{ij} - \bar{Y}\right)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2}. \tag{2.12}$$

Avec cette estimation, il est possible de calculer la somme des carrés totale de Y ajustée pour la covariable (et la moyenne générale), qui est également la somme des carrés résiduels autour de la droite de régression Elle est notée par SCT_{aj} et elle est calculée comme suit

$$\begin{aligned} SCT_{aj} &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y} - \hat{\beta}_T\left(X_{ij} - \bar{X}\right)\right)^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 + \hat{\beta}_T^2\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2 - 2\hat{\beta}_T\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)\left(Y_{ij} - \bar{Y}\right), \end{aligned}$$

mais de l'équation(2.12)

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)\left(Y_{ij} - \bar{Y}\right) = \hat{\beta}_T\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2,$$

donc

$$SCT_{aj} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 - \hat{\beta}_T^2\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2.$$

• En utilisant la somme des produits totale de X et Y (2.6), la somme des carrés totale

ajustée est exprimée de la manière suivante

$$SCT_{aj} = SCT - \hat{\beta}_T^2 \left(SCT\right)_X = SCT - \frac{SPT^2}{\left(SCT\right)_X}. \tag{2.13}$$

La variation de Y associée au terme d'erreur dans le modèle d'analyse de covariance est spécifiée par la somme des carrés des erreurs ajustée (SCE_{aj}). Elle correspond à la somme des carrés des écarts de la valeur observée de la variable dépendante par rapport à la valeur prédite avec connaissance du groupe associé ou du niveau de traitement et de la covariable. Cette valeur prédite est calculée comme suit

$$\hat{Y}_{j\backslash i} = \bar{Y}_i + \hat{\beta}\left(X_{ij} - \bar{X}_i\right),\ \forall i = \overline{1,p},\ \forall j = \overline{1,n_i}.$$

où $\hat{\beta}$ est l'estimateur de β dans le modèle d'ANCOVA 1 donné par (2.2).
Alors SCE_{aj} est calculée comme suit

$$\begin{aligned} SCE_{aj} &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_{j\backslash i}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i - \hat{\beta}\left(X_{ij} - \bar{X}_i\right))^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 + \hat{\beta}^2\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2 - 2\hat{\beta}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)\left(Y_{ij} - \bar{Y}_i\right), \end{aligned}$$

mais de l'équation (2.2)

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)\left(Y_{ij} - \bar{Y}_i\right) = \hat{\beta}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2,$$

donc

$$SCE_{aj} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 - \hat{\beta}^2\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2.$$

• En utilisant la formule (2.2), la somme des carrés des erreurs ajustée peut être écrite de

la forme suivante

$$SCE_{aj} = SCE - \hat{\beta}^2 (SCE)_X = SCE - \frac{SPE^2}{(SCE)_X}. \tag{2.14}$$

• La somme des carrés totale ajustée (SCT_{aj}) représente la variation due au traitement ou à l'effet du groupe plus l'effet résiduel, et la somme des carrés des erreurs ajustée (SCE_{aj}) représente la variation due uniquement à l'effet résiduel. Par conséquent, la somme des carrés des traitements ajustée (SCF_{aj}) qui représente la variation due uniquement au traitement ou à l'effet de groupe, peut être calculée par soustraction

$$SCF_{aj} = SCT_{aj} - SCE_{aj}. \tag{2.15}$$

2.5.4 Tests d'hypothèse

Sous l'hypothèse nulle H_0 (d'absence d'effet de facteur A) et lorsque les conditions de validité du modèle sont respectées, alors la statistique de test peut être définie comme dans une analyse de variance à un facteur par

$$F = \frac{CMF_{aj}}{CME_{aj}}. \tag{2.16}$$

où $CMF_{aj} = \frac{SCF_{aj}}{p-1}$ et $CME_{aj} = \frac{SCE_{aj}}{n-p-1}$.
Cette statistique suit une loi de Fisher de degré de liberté $(p-1)$ et $(n-p-1)$.
Pour un seuil de risque α donné, la table de la loi de Fisher nous fournis la valeur critique du test $f_{(\alpha,p-1,n-p-1)}$ telle que

$$P(\frac{CMF_{aj}}{CME_{aj}} < f_{(\alpha,p-1,n-p-1)}) = 1 - \alpha.$$

• si $f < f_{(\alpha,p-1,n-p-1)}$ on ne peut pas rejeter H_0 (Il n'y a pas d'influence du facteur),

• si $f \geq f_{(\alpha,p-1,n-p-1)}$ on rejette H_0 (Il y a une influence du facteur),

avec f est la réalisation de la statistique F.

Remarque 2.5.1 - *Par rapport au modèle d'ANOVA 1, on remarque que les degrés de liberté associés aux sommes des carrés totale ajustés et sommes des carrés des erreurs ajustés sont diminués par un degré de liberté.*

- Il est clair que, pour faire une analyse de covariance, le coefficient β est supposé différent de zéro. Alors pour tester l'hypothèse ($H_0 : \beta = 0$) on utilise la statistique

$$F = \frac{SPE^2/\left(SCE\right)_X}{CME_{aj}}.$$

qui, sous l'hypothèse nulle, est de loi Fisher avec 1 *et* $(n-p-1)$ *degrés de liberté. L'hypothèse nulle sera rejetée au seuil de signification α si la réalisation de la statistique F est supérieure ou égale à $f_{(\alpha,1,n-p-1)}$ (lue dans la table de Fisher).*

- Les résultats d'ANCOVA 1 sont aussi présentés dans un tableau de la forme suivant

Source de variation	Degrés de libertés *ddl*	Somme des carrés *SC*	Carré Moyen *CM*	Ratio *f*
Taitements	$p-1$	SCF_{aj}	CMF_{aj}	$\frac{CMF_{aj}}{CME_{aj}}$
Erreurs	$n-p-1$	SCE_{aj}	CME_{aj}	
Total	$n-2$	SCT_{aj}		

TAB. 2.1 – Tableau d'analyse de covariance à un facteur et une covariable.

2.6 Comparaison des droites de régression

Par comparaison du modèle d'ANCOVA 1 (2.1) et le modèle de la régression linéaire simple(1.7), il apparaît que les quantités $(\mu + \alpha_i - \beta\bar{X})$ sont les ordonnées à l'origine des p droites de régression relatives, séparément aux populations, ces droites étant supposées de même coefficient de régression β, c'est à dire parallèles. L'hypothèse de nullité des effets principaux α_i est donc une hypothèse d'égalité des p ordonnées à l'origine.([9])

L'équation du groupe i peut s'écrire

$$E(Y) = \beta_{0,i} + \beta_{1,i}X,\ i = 1, ..., p. \tag{2.17}$$

On peut alors se poser trois questions

• La droite de régression est-elle la même pour tous les groupes ? ou encore le modèle $E(Y) = \beta_0 + \beta_1 X$ est-il préférable à l'équation $E(Y) = \beta_{0,i} + \beta_{1,i}X$?

• Si la réponse à la première question est négative, les pentes sont-elles identiques ? Ou encore le modèle $E(Y) = \beta_{0,i} + \beta_1 X$ est-il préférable à $E(Y) = \beta_{0,i} + \beta_{1,i}X$?

• Si la réponce à la question précédente est positive (non-rejet de H_0 : égalité des pentes), les ordonnées à l'origine sont-elles identiques ? Ou encore le modèle $E(Y) = \beta_0 + \beta_1 X$ est-il préférable à $E(Y) = \beta_{0,i} + \beta_1 X$?

En gardant les hypothèses de la régression linéaire simple et l'hypothèse d'égalité des variances des groupes, la réponse à ces questions est ordonnée selon la démarche suivante peut être faite en se basant sur l'approche de l'erreur conditionnelle ou encore sur le test de comparaison des modèles, tel que

• Pour la première question on fait un test d'égalité des droites de régression ou une comparaison des modèles

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \beta_{1,i}(X_{ij} - \overline{X}) + \varepsilon_{ij} \\ Y_{ij} = \mu + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij} \end{cases}.$$

• Concernant la $2^{ème}$ question on fait un test d'égalité des pentes ou encore une comparaison entre les modèles

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \beta_{1,i}(X_{ij} - \overline{X}) + \varepsilon_{ij} \\ Y_{ij} = \mu + \alpha_i + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij} \end{cases}.$$

• Pour la $3^{ème}$ question on applique un test d'égalité des ordonnées à l'origine on une

comparaison entre les modèles

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij} \\ Y_{ij} = \mu + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij} \end{cases}.$$

2.7 Exemple d'application

Considérons l'expérience qui consiste à comparer les effets de trois méthodes de nutrition sur une population de porcs. Les données se présentent sous la forme d'un tableau comprenant trois méthodes de nutrition, chacune soumise à 5 porcs. Les poids initiaux sont notés par la variable concomitante X(en kg), et les gains de poids (après traitement) sont notés par Y.([5])

Méthode de nutrition					
1		2		3	
X	Y	X	Y	X	Y
32	167	26	182	36	158
29	172	33	171	34	191
22	132	22	173	37	140
23	158	28	163	37	192
35	169	22	182	32	162

TAB. 2.2 – Les poids de porcs en fonction de la méthode de nutrition et de poids initial.

Supposons que les conditions d'applicabilité de l'ANCOVA 1 sont vérifiées, et on désir prendre nos décision pour un risque $\alpha = 5\%$, alors les étapes à suivre pour répondre à notre objectif sont résumées comme suit

• **Calcul des moyennes** : les moyennes de X et de Y des différents échantillons sont calmculées comme suit

$\overline{X}_1 = \frac{32+29+22+23+35}{5} = 28.2,$

$\overline{X}_2 = \frac{26+33+22+28+22}{5} = 26.2,$

$\overline{X}_3 = \frac{36+34+37+37+32}{5} = 35.2,$

$\overline{X} = \frac{1}{15}(32 + 29 + ... + 32) = 29.87,$

$\overline{Y}_1 = \frac{167+172+132+158+169}{5} = 1,596,$

$\overline{Y}_2 = \frac{182+171+173+163+182}{5} = 174.2,$

$\overline{Y}_2 = \frac{158+191+140+192+162}{5} = 168.6,$

$\overline{Y} = \frac{1}{15}(167 + 172 + ... + 162) = 167.47.$

• **Calcul des sommes des carrés et des produits**

– La somme des carrés totale pour X

$$\begin{aligned}(SCT)_X &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2 \\ &= (32 - 29.87)^2 + (29 - 29.87)^2 + ... + (32 - 29.87)^2 \\ &= 453.73.\end{aligned}$$

– La somme des carrés totale pour Y

$$\begin{aligned}SCT &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 \\ &= (167 - 167.47)^2 + (172 - 167.47)^2 + ... + (162 - 167.47)^2 \\ &= 542.53.\end{aligned}$$

– La somme des produits totale de X et Y

$$\begin{aligned}SPT &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)\left(Y_{ij} - \bar{Y}\right) \\ &= (32 - 29.87)(167 - 167.47) + (29 - 29.87)(172 - 167.47) + ... + (32 - 29.87)(162 - 167.47) \\ &= 158.93.\end{aligned}$$

– La somme des carrés des traitements pour X

$$\begin{aligned}(SCF)_X &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\bar{X}_i - \bar{X}\right)^2 \\ &= 5(28.2 - 29.84)^2 + 5(26.2 - 29.87)^2 + 5(35.2 - 29.87)^2 \\ &= 223.33.\end{aligned}$$

– La somme des carrés des traitements pour Y

$$\begin{aligned}SCF &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{Y}\right)^2 \\ &= 5(159.6 - 167.47)^2 + 5(174.2 - 167.47)^2 + 5(168.6 - 167.47)^2 \\ &= 542.53.\end{aligned}$$

– La somme des produits des traitements de X et Y

$$\begin{aligned}SPF &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\bar{X}_i - \bar{X}\right)\left(\bar{Y}_i - \bar{Y}\right) \\ &= 5(28.2 - 29.84)(159.6 - 167.47) + 5(26.2 - 29.87)(174.2 - 167.47) \\ &+ 5(35.2 - 29.87)(168.6 - 167.47) \\ &= -27.67.\end{aligned}$$

– La somme des carrés des erreurs pour X

$$\begin{aligned}(SCE)_X &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2 \\ &= (32 - 28.2)^2 + (29 - 28.2)^2 + ... + (32 - 35.2)^2 \\ &= 230.40.\end{aligned}$$

– La somme des carrés des erreurs pour Y

$$\begin{aligned} SCE &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2 \\ &= (167-159.6)^2+(172-159.6)^2+...+(162-168.6)^2 \\ &= 3343.20. \end{aligned}$$

– La somme des produits des erreurs X et Y

$$\begin{aligned} SPE &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\left(Y_{ij}-\bar{Y}_i\right) \\ &= (32-28.2)(167-159.6)+(29-28.2)(172-159.6)+...+(32-35.2)(162-168.6) \\ &= 186.60. \end{aligned}$$

• Calcul des sommes des carrés ajustées

L'ajustement de la variable Y à la variable concomitante X donne les nouvelles sommes des carrés suivantes

1. La somme des carrés totale ajustée

$$\begin{aligned} SCT_{aj} &= SCT-\frac{SPT^2}{(SCT)_X} = 3885.73-\frac{(158.93)^2}{453.73} \\ &= 3830.06. \end{aligned}$$

2. La somme des carrés des erreurs ajustée

$$\begin{aligned} SCE_{aj} &= SCE-\frac{SPE^2}{(SCE)_X} = 3343.20-\frac{(186.60)^2}{230.40} \\ &= 3192.07. \end{aligned}$$

3. La somme des carrés des traitements ajustée

$$SCF_{aj} = SCT_{aj} - SCE_{aj} = 3830.06 - 3192.07$$
$$= 637.99.$$

- **Décision**

La valeur observée de la statistique de test s'élève à

$$f = \frac{CMF_{aj}}{CME_{aj}} = \frac{SCF_{aj}/2}{SCE_{aj}/11} = \frac{318.995}{290.188}$$
$$= 1.099.$$

De la table de Fisher et pour un seuil de signification $\alpha = 5\%$, on trouve que $f_{(0.05,2,11)} = 3.98$.

Comme $f < f_{(0.05,2,11)}$, on ne peut pas rejeter H_0 au seuil de signification α et on conclut qu'il n'y a pas de différences significative entre les trois méthodes de nutrition.

Les résultats d'ANCOVA 1 sont aussi présentés dans un tableau de la forme suivant

Source de variation	Degrés de libertés *ddl*	Somme des carrés *SC*	Carré moyen *CM*	ratio *F*
Taitements	2	637.99	318.995	1.099
Erreurs	11	3192.07	290.188	
Total	13	453.73		

TAB. 2.3 – Tableau d'analyse de covariance à un facteur.

Chapitre 3

Application sous R

Dans ce chapitre, nous traitons en pratiquement sous R les méthodes statistiques que nous avons vu dans les chapitres précédents. Nous donnons quelques exemples conçernant la méthode d'analyse de la variance à un facteur, de la régression linéaire simple et de la méthode d'analyse de la covariance.

3.1 Exemples sur l'ANOVA 1 et sur la régression linéaire simple

3.1.1 Exemple sur l'ANOVA 1

Pendant leur cuisson, les croissants absorbent de la graisse en quantité variable. On veut voir si la quantité absorbée dépend du type de graisse. On prépare donc quatre graisses différentes et on fait cuire six croissants par type de graisse. Les données enregistrées sont résumées dans le tableau (3.1).([3])

• **Inspection graphique :** Tous d'abord, nous allons effectuer une brève analyse descriptive de ces données pour voir si certaines tendances probables se dégagent.

Croissant	**Graisse**	**Quantité**	**Croissant**	**Graisse**	**Quantité**
1	Graisse 1	64	1	Graisse 3	75
2	Graisse 1	72	2	Graisse 3	93
3	Graisse 1	68	3	Graisse 3	78
4	Graisse 1	77	4	Graisse 3	71
5	Graisse 1	56	5	Graisse 3	63
6	Graisse 1	95	6	Graisse 3	76
1	Graisse 2	78	1	Graisse 4	55
2	Graisse 2	91	2	Graisse 4	66
3	Graisse 2	97	3	Graisse 4	49
4	Graisse 2	82	4	Graisse 4	64
5	Graisse 2	85	5	Graisse 4	70
6	Graisse 2	77	6	Graisse 4	68

TAB. 3.1 – Quantités de graisse absorbées par croissant.

> X<-data.frame(Graisse1=c(64,72,68,77,56,95),Graisse2=c(78,91,97,82,85,77),

Graisse3=c(75,93,78,71,63,76),Graisse4=c(55,66,49,64,70,68))

> Quantité<-stack(X)$values

> Graisse<-stack(X)$ind

> tapply(Quantité,Graisse,summary)

$Graisse

Min.	1stQu.	Median	Mean	3rdQu.	Max.
56.00	65.00	70.00	72.75	75.75	95.00

$Graisse2

Min.	1stQu.	Median	Mean	3rdQu.	Max.
77.00	79.00	83.50	85.00	89.50	97.00

$Graisse3

Min.	1stQu.	Median	Mean	3rdQu.	Max.
63.00	72.00	75.50	76.00	77.50	93.00

$Graisse4

Min.	1stQu.	Median	Mean	3rdQu	Max.
49.00	57.25	65.00	62.00	67.50	70.00

Le test porte sur la comparaison des moyennes. À cette étape, il serait bon de tracer les

boîtes à moustaches de la variable Quantité en fonction de la variable Graisse. Pour cela, tapez la ligne de commande suivant

```
> plot(Quantité~Graisse,pch=16,cex=0.5,col="green")
```

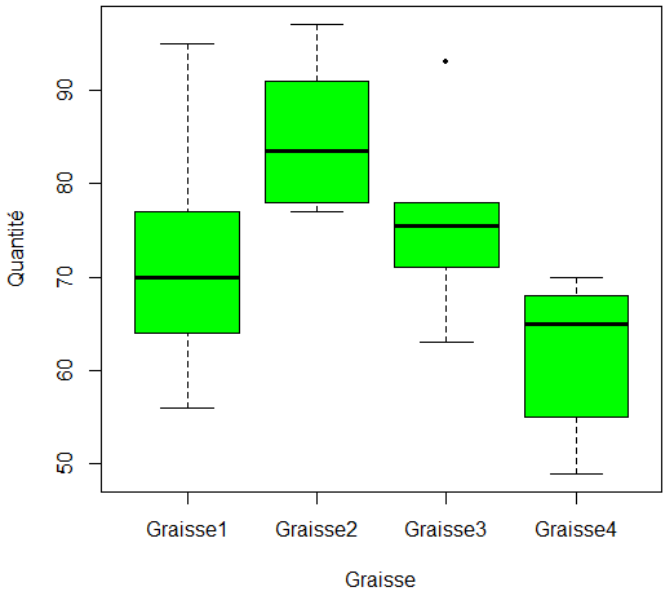

FIG. 3.1 – Les boîtes à moustaches de la variable quantité en fonction de la variable graisse.

• **Instruction R pour la table d'ANOVA :** La fonction à utiliser est aov(). Comme pour le modèle de régression, l'ANOVA fonctionne avec des formules R, il faut donc spécifier le modèle à utiliser.

```
> model1.aov<-aov(Quantité~Graisse,data=croissant)
> summary(model1.aov)
```

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Graisse	3	1636	545.5	5.406	0.000688	***
Residuals	20	2018	100.9			

— — —

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Remarque 3.1.1 *Comme l'ANOVA est en fait un modèle linéaire, notons qu'il aussi possible d'effectuer l'analyse de la variance du modèle linéaire sous-jacent*

```
> model1<-lm(Quantité~Graisse,data=croissant)
> anova(model1)
```

La fonction `anova(model1)`*, nous permet d'obtenir la table d'ANOVA.*

Le tableau d'analyse de la variance renvoie le résultat du test de Fisher associé aux hypothèses : $H_0 : \mu_1 = \mu_2 = ... = \mu_p$ et $H_1 : \exists i \neq i'/\mu_i \neq \mu_{i'}$ (il existe au moins deux moyennes différentes). La valeur $p = 0.0006888$ nous permet de conclure que les quantités de graisse obserbées par croissant d'au moins deux graisses sont différentes au risque 5%

Remarque 3.1.2 *Ce qui concerne les conditions d'application d'ANOVA 1, on peut tester la normalité par le test de* ***shapiro-wilk*** *comme suit*

```
> residus<-residuals(model1)
> shapiro.test(residus)
```

Aussi on peut tester d'égalité des variances par le test de ***Bartlett***

```
> bartlett.test(residus)
```

3.1.2 Exemple sur la régression linéaire simple

L'étude statistique ci-dessous porte sur les poids respectifs des pères (X) et de leur fils aîné (Y).([5])

X	65	63	67	64	68	62	70	66	68	67	69	71
Y	68	66	68	65	69	66	68	65	71	67	68	70

TAB. 3.2 – Les poids des pères et de leur fils.

• Lecture des données :

```
> X <- c(65,63,67,64,68,62,70,66,68,67,69,71)
> Y <- c(68,66,68,65,69,66,68,65,71,67,68,70)
```

• Inspection graphique :

Afin d'etudier la relation entre le poids de fils et le poids de père nous pouvons commencer par tracer le nuage des points grâce à l'instruction

> plot (Y, xlab="poids de père", ylab="poids de fils")

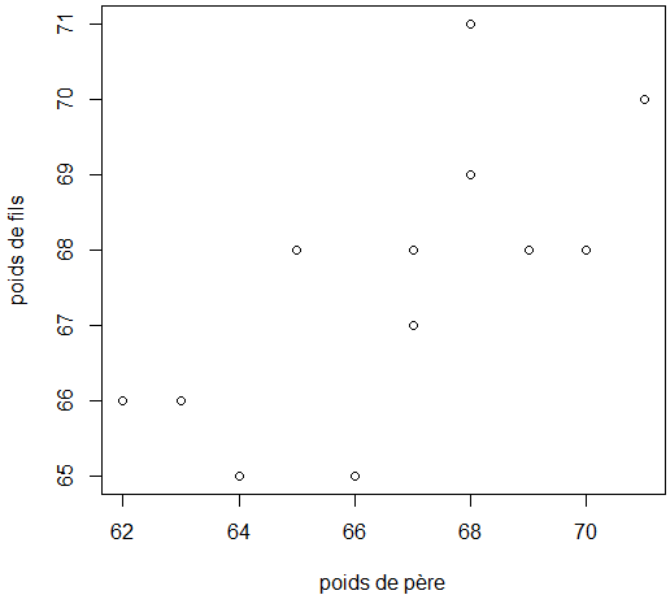

FIG. 3.2 – Nuage de points du poids des fils en fonction du poids des pères

• **Estimation des paramètres :** Nous estimons le modèle par la fonction lm().

> modele1 <- lm(Y~X)

> modele1

Call :

lm(formula = Y~X)

Coefficients :

(Intercept) Y

35.8248 0.4764

Nous pouvons maintenant répresenter la droite de régression sur le nuage de points au moyen de la fonction abline ()

> plot (Y~X, xlab="poids de père", ylab="poids de fils")

> abline(modele1,col="red")

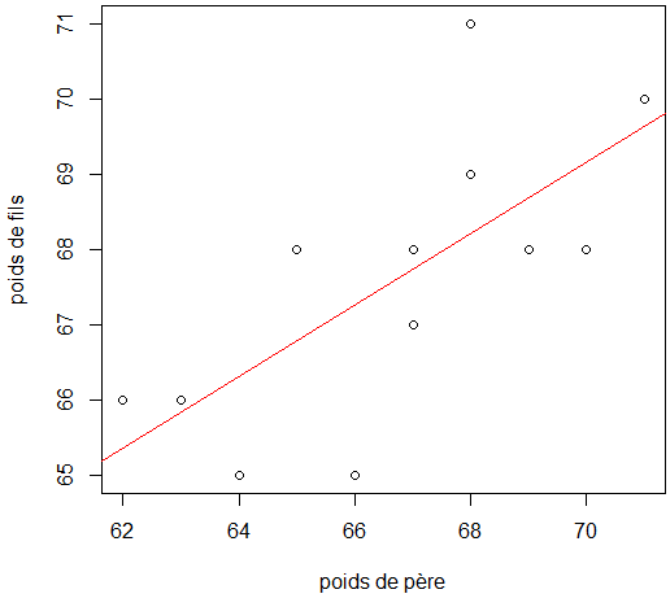

FIG. 3.3 – Représentation de la droite de régression des moindres carrés sur le nuage de

• **Tableau d'analyse de la variance :** Le test de Fisher est souvent associé à une table d'analyse de la variance que vous obtenez en utilisant la fonction anova () .

> anova(modele1)

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
forêt	1	19.214	19.2139	9.7519	0.01082	*
Residuals	10	19.703	1.9703			

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Remarque 3.1.3 *La relation linéaire entre X et Y est démontrée par le résultat du test de Fisher sur le coefficient β_1. La p-valeur<0.05 nous indique une relation linéaire signi-*

ficative entre le poids de fils et le poids de père.

3.2 Exemple de l'analyse de covariance ANCOVA 1

On mesure le taux de leucocytes $T4$ chez le chat X_2 jours après avoir inoculé à l'animal le virus $FeLV$, analogue au HIV. On appelle Y le logarithme népérien de ce taux. Le tableau ci-dessous donne les mesures faites sur 17 chats mâles et 15 chats femelles. Le facteur Sexe est noté X_1.([1])

Mâles		Femelles	
Jours	Ln(Taux de T4)	Jours	Ln(Taux de T4)
44	4.66	84	3.45
317	3.06	47	3.89
292	1.28	20	3.79
179	3.17	209	3.79
39	5.59	106	3.81
257	2.88	343	0.61
354	1.60	325	2.04
349	3.48	346	2.67
195	3.39	151	0.89
245	3.47	267	4.39
270	3.20	80	2.56
166	2.90	249	0.28
57	4.83	341	2.43
198	2.96	189	3.85
20	5.17	50	
187	3.44		
270	3.18		

TAB. 3.3 – Le logarithme népérien de taux de leucocytes T4 chez le chat en fonction du nombre de jours après avoir inoculé à l'animal le virus du sida et de sexe.

Dans la présente application, l'objectif est de déterminer si le facteur sexe affecte le taux de leucocytes T4 chez le chat quelques jours après avoir inoculé à l'animal le virus du sida. Avant de répondre à ce problème, on veut répondre aux questions suivantes

1- Y a-t-il un effet de sexe sur le taux de leucocytes après avoir inoculé à l'animal le virus du sida ?

2- Est-ce-que le nombre de jours après avoir inoculé à l'animal le virus du sida est différent selon le sexe ?

3- Comment représenter graphiquement le logarithme népérien de taux de leucocytes T4 en fonction de jours pour les différents niveaux du sexe ?

4- Y a t'il une interaction entre le sexe et le rapport entre le taux de leucocytes T4 et le nombre de jours après avoir inoculé à l'animal le virus du sida ?

Notation 3.2.1 *Les variables de l'exemple sont codées de la manière suivant*

Le logarithme népérien de taux de leucocytes : LNT4.

Femme : F , Mâle : M

Nombre de jours après avoir inoculé à l'animal le virus du sida : JOURS.

Réponses

Les données sont entrées dans R au moyen des instructions suivantes

```
> options(contrasts=c("contr.sum","contr.poly"))
> SIDACHAT=read.table(("D :/anavar/LNTAUXT.TXT"),header=TRUE,sep="\t")
> attach(SIDACHAT)
> names(SIDACHAT)
```

Pour la 1ère et la 2ème question, l'analyse préliminaire des échantillons dont on dispose nous fournis les résultats suivants

On calcule la moyenne de LNT4 et la moyenne du nombre de jours après avoir inoculé à l'animal le virus du sida dans chaque groupe

```
> tapply(LNT4,SEXE,mean)
F        M
2.590667 3.428235
> tapply(SIDACHAT$JOURS,SIDACHAT$SEXE,mean)
F        M
187.1333 202.2941
```

On représente graphiquement les données à l'aide des boites à moustaches des variables *JOURS* et *LNT4* en fonction de SEXE, pour cela on va taper les lignes de commandes suivantes

```
> par(mfrow=c(1,2))
> plot(LNT4~SEXE,data=SIDACHAT,col="green",main="a")
> plot(JOURS~SEXE,data=SIDACHAT,col="green",main="b")
```

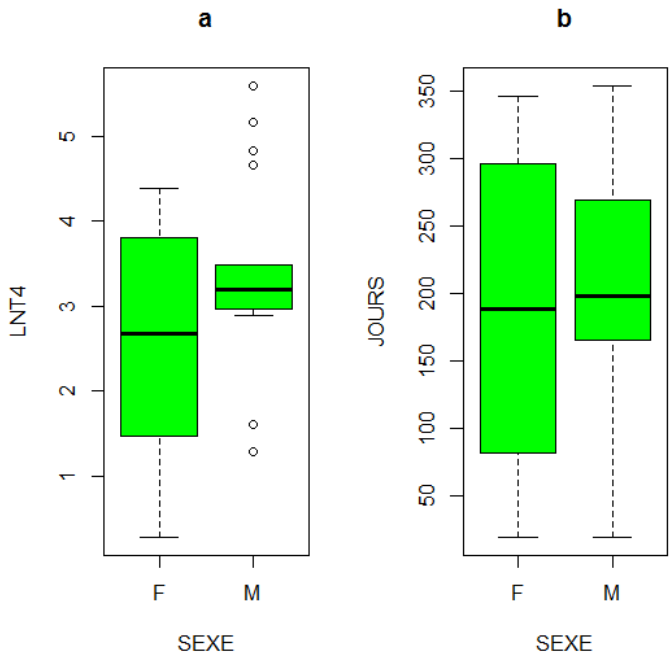

FIG. 3.4 – Boites à moustaches de logarithme népérien de taux de leucocytes T4 (a) et de nombre de jours après avoir inoculé à l'animal le virus du sida (b) en fonction du sexe.

A partir de ces résultats préliminaires, on remarque que la moyenne de LNT4 dans le groupe M est plus grand que la moyenne de LNT4 dans le groupe F. Tandis que, la moyenne du nombre de jours après avoir inoculé à l'animal le virus du sida dans le groupe M est aussi plus grande que la moyenne dans le groupe F.

La fonction aov, nous permet de répondre aux questions 1 et 2, comme suit

```
> aov1<-aov(LNT4~SEXE,data=SIDACHAT)
summary(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
SEXE	1	5.59	5.59	3.415	0.0745	.
Residuals	30	49.11	1.637			

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Alors, au seuil de risque $\alpha = 0.05$, la p-valeur étant strictement supérieur à 0.05, ce qui indique qu'il n' y a pas un effet significatif du facteur sexe sur LNT4 sans tenir compte le nombre de jours après avoir inoculé à l'animal le virus du sida.

```
> aov2<-aov(JOURS~SEXE,data=SIDACHAT)
> summary(aov2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SEXE	1	1832	1832	0.142	0.709
Residuals	30	387237	387237		

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

De la comparaison de la p valeur au seuil de signification α ($\alpha = 0.05$), on résulte qu'il n'existe pas une différence significative entre les moyennes de la variable nombre de jours après avoir inoculé à l'animal le virus du sida des deux sexes.

Pour la question 3, nous pouvons tracer le nuage des points grâce à l'instruction plot(LNT4∼JOURS)

```
> plot(LNT4~JOURS,data=SIDACHAT,type="n")
> points(LNT4~JOURS,data=subset(SIDACHAT,SEXE=="F"),col="blue",pch="F")
> points(LNT4~JOURS,data=subset(SIDACHAT,SEXE=="M"),col="red",pch="M")
> legend(250,5,c("Femelles","Mâles"),pch="FM",col=c("blue","red"),cex=1)
```

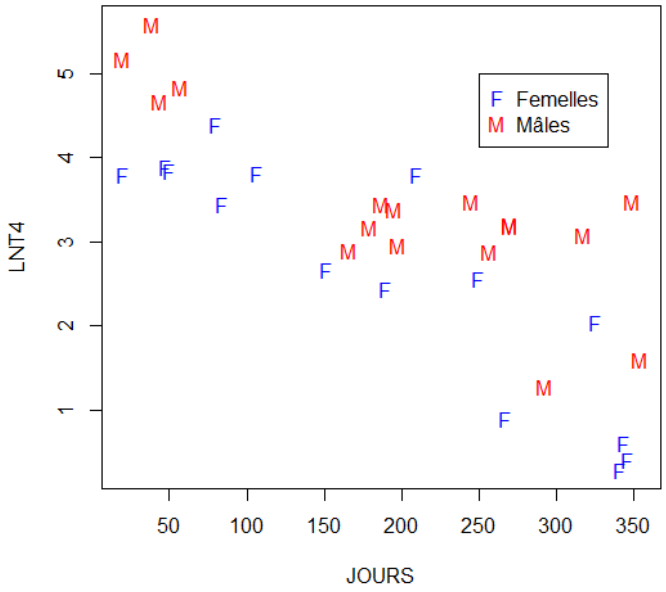

FIG. 3.5 – Nuage de points de la mesure de LNT4 en fonction de jours.

La quatrième question concernant un test sur l'interaction entre JOURS et SEXE ou sur l'égalité des pentes des droites de régressions, qu'il se fait de la manière suivante

```
> mod1<-lm(LNT4~JOURS+SEXE+JOURS :SEXE,data=SIDACHAT)
> anova(mod1)
```

Analysis of Variance Table

Response : LNT4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
JOURS	1	34.062	34.062	76.3023	$1.743e-09$	***
SEXE	1	7.680	7.680	17.2048	0.0002823	***

JOURS : SEXE	1	0.642	0.462	1.0342	0.3178879
Residuals	28	12.499	0.446		

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

De cette table et au seuil de risque $\alpha = 0.05$, on constate qu'il n'existe pas un effet significatif d'interaction entre JOURS et SEXE car la p valeur est supérieure à 0.05 (p-valeur=0.3178879), d'où les droites de régressions sont parallèles.

Remarque 3.2.1 *Comme dans l'analyse de variance à un facteur, nous testons l'hypothèse de normalité et d'égalité des variances des résidus pour le modèle linéaire (mod1) à l'aide des instructions* ***shapiro.test(residuals(mod1))****.et* ***bartlett.test(residuals(mod1))****.*

L'application d'ANCOVA, nous permet de répondre à notre objectif

```
> mod2<-lm(LNT4~JOURS+SEXE,data=SIDACHAT)
> anova(mod2)
```

Analysis of Variance Table

Response : LNT4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
JOURS	1	34.062	34.062	76.213	$1.308e-09$	***
SEXE	1	7.680	7.680	17.185	0.0002692	***
Residuals	29	12.961	0.447			

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Contrairement à ce qu'on a vu à la première question, de la table d'ANCOVA 1, on remarque que la p valeur associée au facteur sexe est inférieure à 0.05, ce qui veut dire que le facteur SEXE influe sigificativement sur LNT4. Ce résultat montre que si en tenant compte le nombre de jours après avoir inoculé à l'animal le virus du sida, les moyennes de LNT4 des deux sexes (moyennes ajustées) sont différentes. Aussi, on peut dire que la modélisation des données ne sera pas par une seule droite.

Cacul des moyennes ajustées

Grâce aux instructions suivantes, on obtient les moyennes ajustées : meanFaj, meanMaj de la variable LNT4 des deux sexes F, M respectivement.

coe2 <- coefficients(mod2)

```
meanFaj=mean(LNT4[SEXE=="F"])-((coe2[2])*(mean(JOURS[SEXE=="F"])-mean(JOURS)))
meanMaj=mean(LNT4[SEXE=="M"])-((coe2[2])*(mean(JOURS[SEXE=="M"])-mean(JOURS)))
```

d'où meanFaj=2.512846, meanMaj=3.496901.

Noton

s qu'après l'ajustement, les moyennes ajustées divergent peu des moyennes non ajustées tel que la moyenne la plus basse de LNT4, c'est-à-dire 2.59 a été diminuée à 2.51, et la moyenne la plus élevée c'est-à-dire 3.42 a été augmentée à 3.49.

Ceci dù au fait que la différence entre les groupes au niveau de nombre de jours après avoir inoculé à l'animal le virus du sida a été éliminée.

Remarque 3.2.2 *Pour comparer deux modèles sous R, on utilise l'instruction* ***anova(Model.1, Model.2).***

Pour choisir le meilleur modèle qui décrit les données, on peut aussi utiliser quelques méthodes de sélection de variables disponibles avec le logiciel R comme la fonction step().

Conclusion

L'analyse de covariance se situe dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. L'idée à la base d'analyse de covariance est d'ajouter à un modèle d'analyse de la variance, associe à une ou plusieurs variables qualitatives, une ou plusieurs variables quantitatives qui pourraient être liées à la réponse étudiée. L'objectif du présent mémoire est d'étudier le cas de l'ANCOVA 1, c'est-à-dire l'étude de l'effet d'un facteur fixe sur une variable quantitative où une autre variable quantitative se présente (une seule covariable). Le modèle d'ANCOVA 1 peut être vu soit comme une comparaison des moyennes ajustées pour la covariable soit comme une comparaison des droites de régression.

La prise en compte de l'effet d'une covariable peut viser plusieurs objectifs

1 - Etudier l'effet d'un facteur en prenant en compte de l'effet de la covariable X sur la variable d'intérêt Y. En effet la relation entre le facteur A et Y peut dépendre de X, d'où l'amplitude de l'effet du facteur ne s'interprète aisément qu'après ajustement sur X. Par exemple (l'effet d'un traitement sur l'intensité Y des symptômes où X est l'état initiale du patient).

2 - Evité les méfaits de la non comparabilité des groupes au niveau d'un covariable. En effet, les différences observées en Y peuvent provenir des différences en X et non des niveaux de A. L'ANCOVA permet alors de rétablir la comparabilité des situations en X.

3 - Accroître la puissance des tests relatifs à l'effet du ou des facteurs étudiés. En effet, plus la covariable est corrélée avec la variable d'intérêt, plus la variance résiduelle décroît,

et plus la puissance des tests est grande.

Des ajouts peuvent être apportés à ce travail afin de le rendre plus riche, par exemple

- De voir la méthode appliquée pour l'analyse détaillée de l'ANCOVA (Tests de comparaisons multiples).
- D'introduire au modèle plusieurs facteurs qualitatifs.
- D'étudier le cas où il y a plus d'une variable dépendante (MANCOVA).

Bibliographie

[1] Bertrand, F. (2008/2009). Analyse de covariance Tp n° 9, Magistère 2ème année.

[2] Chavent, M. (2012-2013). Régression linéaire Simple, Chapitre 1. Licence 3 MIASHS-Université de Bordeaux.

[3] Cherfaoui, M. (2017/2018). Statistiques Appliquées à l'Expérimentation En Sciences biologique, polycopié du cours BIOSTATISTIQUES, Université de biskra.

[4] Chouquet.C. (2009-2010). Modèles Linéaires. Laboratoire de statistique et probabilités-Université Paul Sabatier-Toulouse, M1 IMAT.

[5] Dodge,Y. (2007). Statistique dictionnaire encyclopédique-Université de Neuchâtel-Suisse. yadolah.dodge@unine.ch.

[6] Drouilhet, R. Lafaye, P. Liquet, B. (2011). Le logiciel R. Maîtriser le langage. Effectuer des analyses (bio) statiques, 2 édition, Springer.

[7] Guyader, A. (2012-2013). Régression Linéaire. Université Rennes 2.

[8] Maumy-Bertrand, M. Bertrand, F. (2010). Initiation à la statistique avec R : Cours, exemples, exercices et problèmes corrigés. Dunod.

[9] Scherrer, B. (2007). Biostatistique, volume1. Gaëtan Morin éditeur (816 page).

[10] Scherrer, B. (2009). Biostatistique, volume 2, chapitre 25. Ed. Gaëtan Morin-Chenelière.

[11] Wildt, Albert R, Ahotola, O. (1978). Analysis of covariance. Sage University Papers Series.

Annexe A : Logiciel *R*

Les différentes commandes utilisées tout au long de ce mémoire sont expliquées ci-dessous.

data.frame	Crée un nouveau jeu de données.
tapply(x,y,z)	Applique la fonction z aux groupes constituée à partir du vecteur x grâce aux modalités du facteur y.
plot	Trace le graphe.
aov	Analyse de variance.
summary	Résumé du modèle.
shapiro.test	Permet de réaliser un test de normalité.
bartlett.test	Permet de tester le l'homogénéité des variances.
read.table	Crée un jeu de données à partir un fichier texte.
attach(data)	Attache le tableau de données data en mimoire.
names	Noms de colonnes.
head("data")	Afficher les 6 premières lignes de data.
points	Trace des points sur un graphe.
lm	Modèle linéaire.
coefficients	Récupère les coefficients d'un modèle.
abline	Ajoute une ou plusieurs lignes droites à un graphe en spécifiant leur équation
step	Sélection de modèle par AIC.

Annexe B : Abréviations et Notations

Les différentes abréVariations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

$ANOVA$	:	Analyse de variance.
$ANCOVA$	:	Analyse de covariance.
SCF	:	La variation due au facteur.
SCE	:	La variation résiduelle.
SCT	:	La variation totale.
CMF	:	Carrés des moyennes associés au facteur.
CME	:	Carrés des moyennes résiduels.
SPT	:	Somme des produits totale de X et Y
SPF	:	Somme des produits des traitements de X et Y
SPE	:	Somme des produits des erreurs de X et Y
SCT_{aj}	:	Somme des carrés totale ajustée
SCE_{aj}	:	Somme des carrés des erreurs ajustée
SCF_{aj}	:	Somme des carrés des traitements ajustée
CMF_{aj}	:	Carrés des moyennes des erreurs ajustée
CME_{aj}	:	Carrés des moyennes des erreurs ajustée
$f_{(n_1,n_2)}$	:	Une loi de Fisher de degrés de liberté $n_1,\ n_2$.