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

MAROUA TABET

Titre :

Analyse de covariance

Membres du Comité d'Examen :

Pr.	**DJABRANE Yahia**	UMKB	Président
MAA.	**ROUBI Affef**	UMKB	Encadreur
MAA.	**BENBRAIKA Ghozlane**	UMKB	Examinateur

2020

Dédicace

Je dédie ce modeste travail.

A mes très chers parents est plus proch de mon coeur celles qui ont fait l'impossible pour me donner le bonheur et pour suivre mes études jusqu'à ce jour, qu'**ALLAH** les protèges.

A mes très chères soeurs et frères et tout ma famille.

A mes chères amies

A tous le promotion de mathématique **2019/2020**

A toute personne que prendera de son temps pour lire ce document à parfaire

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

Table des figures

Liste des tableaux

Introduction

La statistique en général est la branche des mathématiques de l'étude de "données", résultats obtenus lors d'expérimentation ou d'observations de phénomènes aléatoires ou mal prévisibles. Les méthodes statistiques sont aujourd'hui utilisées dans presque tous les secteurs de l'activité humain et font partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du biologiste, de l'informaticien, etc. La statistique a plusieurs objets : Descriptif ou exploratoire, décisionnel (tests), modélisation selon que l'on cherche à représenter des structures des données, confirmer ou expliciter un modèle théorique ou encore prévoir.
Dans l'objectif est de prendre ont décisions sur la base de résultats expérimentaux, en étant conscient qu'il y a 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 des populations, alors 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 (ANOVA) 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 utile à la fois dans la recherche expérimentale et non expérimentale. Il s'agit d'une tech-

nique statistique basée sur le modèle linéaire général. L'idée à la base de l'ANCOVA est de considérer une situation plus générale dans la quelle les variables explicatives sont à la fois quantitatives, appelées covariables, et qualitatives (facteurs) ou d'ajouter à un modèle d'analyse de la variance, associé à une ou plusieurs variables qualitatives, une ou plusieurs variables quantitatives qui pourraient être liées à la réponse étudiée.

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 : On traite dans ce chapitre, la technique d'analyse de la variance à un facteur (ANOVA1), leurs principes, ainsi leurs différentes étapes les plus indispensable. Aussi on parle 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 1. L'intérêt de cette méthode qui permet de combiner les éléments des modèles de régression et les modèles d'analyse de la variance est de séparer l'effet spécifique du facteur étudié de l'effet de la covariable.

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

Chapitre 1

Analyse de variance et régression linéaire simple

Dans ce chapitre on étudie deux techniques statistiques appelées l'analyse de la variance à un seul facteur (en abréviation ANOVA 1) 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 (ANOVA 1)

1.1.1 Généralités

L'analyse de la variance (ANOVA) est une méthode statistique qui permet d'étudier la modification de la moyenne μ d'une quantité Y (variable quantitative) selon l'influence éventuelle d'un ou de plusieurs facteurs d'expérience qualitatifs (traitements). Dans le cas où la moyenne n'est influencée que par un seul facteur (noté facteur A), 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. ([14])

1.1.2 Définition d'ANOVA 1

Définition 1.1 *ANOVA 1*

L'analyse de variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités sur les moyennes d'une variable quantitative Y. ([9])

1.1.3 Les données et le modèle d'ANOVA 1

Les données

On cherche à étudier l'effet d'un facteur A, que l'on supposera à p niveaux, sur une variable quantitative Y. On suppose que le facteur A influe uniquement sur les moyennes des distributions de chacun des p groupes et non sur leur variance.

Pour chaque niveau i du facteur A (avec $1 \leq i \leq p$), on dispose de n_i mesures de Y (Y_{i1}, Y_{i2}, ..., Y_{in_i}). Dans la suit, on notera par n le nombre total d'observation, $n = \sum_{i=1}^{p} n_i$.

On présente généralement les données à l'aide du tableau suivant ([10])

Niveau du facteur A	A_1	A_2	...	A_i	...	A_p
	Y_{11}	Y_{21}	...	Y_{i1}	...	Y_{p1}
	Y_{12}	Y_{22}	...	Y_{i2}	...	Y_{p2}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	Y_{1n_1}	Y_{2n_2}	...	Y_{in_i}	...	Y_{pn_p}
Effectifs	n_1	n_2	...	n_i	...	n_p
Moyennes empiriques	$\bar{Y}_1$	$\bar{Y}_2$	...	$\bar{Y}_i$	...	$\bar{Y}_p$

TAB. 1.1 – Les donnés d'ANOVA 1.

Présentation du modèle d'ANOVA 1

On fait les hypothèses de normalité et d'indépendance suivantes

1. Pour tout $(i, j) \in \{1, 2, ..., p\} \times \{1, 2, ..., n_i\}$, la donnée Y_{ij} suit la loi $N(\mu_i, \sigma^2)$.

2. Les variables aléatoires (Y_{ij}) sont globalement indépendantes.

On peut résumer ces hypothèses en écrivant le modèle

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{avec} \quad \varepsilon_{ij} \backsim N(0, \sigma^2), \tag{1.1}$$

on décrit l'effet du facteur A en suppossant

• une espérance spécifique μ_i pour chaque groupe ou chaque niveau du facteur,

• et une variance intra-groupe σ^2 commune à tous les groupes.

L'objet de cette étude sera de savoire si, au vu des données de tableau (1.1), les moyennes des p échantillons sont égale ou différentes à savoir de tester les hypothèses

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

On décompose parfois μ_i en

$$\mu_i = \mu + \alpha_i \ \text{ avec } \ \sum_{i=1}^{p} n_i \alpha_i = 0,$$

le modèle s'écrit alors

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{1.2}$$

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

$$\begin{cases} H_0 : "\alpha_1 = \alpha_2 = ... = \alpha_p = 0" \\ H_1 : "\exists\ i \in \{1, 2, ..., p\}\ \text{ tel que }\ \alpha_i \neq 0" \end{cases}. \tag{1.3}$$

1.1.4 Les étapes d'ANOVA 1

Pour réaliser une analyse de variance à un facteur il faut suivre les étapes suivantes

Vérification des conditions de validité de l'ANOVA 1

Afin de réaliser le test défini dans (1.3), les trois conditions suivantes doivent être vérifiées préalablement

1. Les p échantillons comparés sont indépendants.

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

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

Calcul des moyennes et variances

Quantifier les différentes statistiques intervenant dans l'analyse de la variance à un facteur et qui sont

• La moyenne des observations de chaque échantillon

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

• La moyenne générale des observations

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

• La variance de toutes les observations

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

• La variance des observations 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

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

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

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

où

SCF : est la variation due au facteur,

SCE : est la variation résiduelle,

SCT : est la variation totale.

Calcul des carrés moyens

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 des carrés moyens associés au facteur CMF et les carrés moyens résiduels CME, dont

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

Si les 3 conditions d'application d'ANOVA 1 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 ($F \backsim f_{(p-1,n-p)}$).

Décision

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

$$P(\frac{CMF}{CME} < f_{(\alpha,p-1,n-p)}) = 1-\alpha. \tag{1.5}$$

• 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.

Tableau d'ANOVA 1

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

Variation	Degrés de liberté (ddl)	Somme des carrés (SC)	Carré moyen (CM)	F
Facteur A	$p-1$	SCF	CMF	CMF/CME
Résiduelle	$n-p$	SCE	CME	
Total	$n-1$	SCT		

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

1.2 La régression linéaire

En statistique, la régression correspond à la relation entre la grandeur approximative d'un phénomène et la grandeur certaine d'un autre phénomène.

On veut modéliser une variable Y (variable à expliquer, réponse) en fonction d'une ou de plusieurs variables explicatives $X_1, ..., X_p$. L'objectif est de prédire ou simplement expliquer Y à partir des données disponible $X_1, ..., X_p$.

$(Y, X_1, ..., X_p)$ sont des variables quantitatives.

On suppose un lien linéaire entre les variables de la forme

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon,$$

où les coefficients $\beta_0, \beta_1, ..., \beta_p$ sont des réels inconnus et ε est un bruit correspondant à la part de Y indépendantes des variable explicatives. L'objectif principal est d'estimer les coefficients $\beta_0, \beta_1, ..., \beta_p$.

Selon le nombre des variables explicatives dans le modèle de la régression linéaire on peut distinguer le modèle de régression linéaire simple et le modèle de régression linéaire multiple. Dans cette section, nous intéresserons à la régression linéaire simple qui considère qu'une seule variable explicative.

1.2.1 Définition

La régression linéaire simple est une technique statistique permetant de trouver des liens linéaires entre deux variables X et Y et elle sert à prévoir les valeurs futures de l'une d'elle en fonction de l'autre.

1.2.2 Modèle de régression linéaire simple

Une modèle linéaire simple est de la forme suivante

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

où :

- Y : est une variable aléatoire réelle à expliquée,
- X : est la variable explicative,
- β_0 et β_1 : sont les paramètres du modèle (à estimer),
- ε : bruit ou erreur (aléatoire).

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , pour } i = \overline{1, n.} \tag{1.6}$$

Les hypothèses relatives à ce modèle sont

i) $E(\varepsilon_i) = 0$,

ii) $var(\varepsilon_i) = \sigma_\varepsilon^2 < \infty \qquad \forall i = \overline{1, n}$,

iii) $cov(\varepsilon_i, \varepsilon_j) = 0 \qquad \forall i \neq j$.

Le modèle s'écrit sous la forme matricielle

$$Y = XB + \varepsilon, \tag{1.7}$$

avec

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \quad , \quad B = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad , \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} .$$

- 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 $(n, 1)$.

1.2.3 Estimation des paramètres du modèle

On cherche les estimateurs des paramètres β_0 et β_1 ($\hat{\beta}_0$ et $\hat{\beta}_1$) qui minimise la somme des carrés des résidus

$$\hat{\varepsilon}_i = y_i - \hat{y}_i,$$

où $\hat{y}_i$ la valeur prédite par le modèle (1.6) lorsque $x = x_i$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

On doit donc résoudre le problème d'optimisation suivant

$$(\hat{\beta}_0, \hat{\beta}_1) = Arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2.$$

Le problème d'optimisation est

$$\min_{(\beta_0,\beta_1)} F(\beta_0, \beta_1),$$

avec

$$F(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2.$$

Alors $(\hat{\beta}_0, \hat{\beta}_1)$ est la solution du système des équations suivant

$$\begin{cases} \frac{\partial F(\beta_0,\beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial F(\beta_0,\beta_1)}{\partial \beta_1} = 0 \end{cases}.$$

Soit après quelques calculs

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

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

$$\begin{cases} \beta_0(\sum_{i=1}^{n} 1) + \beta_1(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} y_i \\ \beta_0(\sum_{i=1}^{n} x_i) + \beta_1(\sum_{i=1}^{n} x_i^2) = \sum_{i=1}^{n} x_i y_i \end{cases}. \tag{1.8}$$

La résolution du système (1.8) 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

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \text{ et } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{1.9}$$

tels que

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

et

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

L'estimateur de la fonction de régression s'écrit

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

1.2.4 Validation du modèle

Critère du R^2

Les moyens complémentaires dont on dispose pour valider un modèle sont les critère du R^2 et du R^2 ajusté. Le critère du R^2 est défini de façon général par ([13])

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

où

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

$$SCR = \sum_{i=1}^{n}(\widehat{y_i} - \bar{Y})^2,$$

$$SCE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

On a que

$$SCT = SCE + SCR,$$

tels que

SCT : Variation de Y ou variation totale.

SCE : Variation des résidus.

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

Cette relation est valable pour les modèles linéaires généraux, pas uniquement pour le modèle de régression linéaire simple. On en déduit donc une autre définition du R^2

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

On peut définir également le $R^2_{ajusté}$ par

$$R^2_{ajusté} = 1 - \frac{SCE/(n-2)}{SCT/(n-1)}.$$

Pour valider le modèle, on test $H_0 : "\beta_1 = 0"$ contre $H_1 : "\beta_1 \neq 0"$.

La statistique du test est la suivante

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

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

Dans ce chapitre, on va intéresser d'une technique statistique appelée l'analyse de covariance (en abréviation ANCOVA) qui est un mélange de modèles d'analyse de variance et de régression linéaire. Il est à préciser que nous limitant au cas d'analyse de covariance à un facteur et une covariable (ANCOVA 1).

2.1 Généralités sur l'ANCOVA

2.1.1 Définition et objectif d'ANCOVA

L'analyse de la covariance est une technique qui combine certaines des caractéristiques de l'analyse de la variance et de la régression linéaire. Elle consiste à ajouter au modèle d'analyse de variance associe à une ou plusieurs variables qualitatives une ou plusieurs variables quantitatives annexes appelées covariables (ou variables concomitantes) qui sont, par hypothèse, reliées à la variable dépendante d'intérêt (réponse). ([16])
L'objectif d'ANCOVA sera donc de tenir compte, lors de l'étude, des effets d'une ou des facteurs sur la variable dépendante des effets possibles d'une ou des covariables. ([11])

2.1.2 Covariable

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

Il existe différents types de covariables. Les types courants sont les mesures de base (c'est-à-dire les mesures de pré-test basées sur la même instrumentation que celle utilisée pour mesurer la variable dépendante) et les variables autres que les mesures de base qui sont corrélées avec la variable dépendante, y compris les caractéristiques organismiques (telles que l'âge, la tension artérielle, taille corporelle et sexe,...) et les caractéristiques environnementales (à la fois physiques et sociales).

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 sera inexistant et nous retiendrons vraisemblablement au final ce modèle plus simple. ([2])

2.2 Modélisation d'ANCOVA 1

2.2.1 Structure des données

Le modèle est explicite dans le cas où une variable quantitative Y est expliquée par

• Un facteur A à p niveaux (traitements).

• Une variable quantitative X, appelée covariable.

Pour chaque niveau $i = 1, ..., p$ de A, on observe, n_i mesures de X notées X_{ij}, et n_i mesures de Y notées Y_{ij}. On notera n le nombre d'observations $n = \sum_{i=1}^{p} n_i$.

2.2.2 Le modèle

Pour tout $\{i = 1, 2, ..., p \text{ et } j = 1, 2, ..., n_i\}$, nous supposons que la donnée, Y_{ij} est une réalisation d'une variable aléatoire Y_{ij} liée à X_{ij} par

$$Y_{ij} = \mu_i + \beta_i X_{ij} + \varepsilon_{ij}, \tag{2.1}$$

tels que

• μ_i et β_i sont les paramètres du modèle de régression pour le traitement i,

• ε_{ij} termes résiduels aléatoires sont indépendants identiquements distribués de loi $N(0, \sigma^2)$.

On peut décomposer chaque paramètre de régression en une partie due à l'effet global du traitement et une partie spécifique à chaque niveau du traitement,

avec

$$\mu_i = \mu + \alpha_i,$$

et

$$\beta_i = \beta + \gamma_i,$$

dont

• α_i est l'effet du traitement,

• β est la pente de la droite de régression de Y en X pour le traitement i,

• γ_i est un effet spécifique du traitement,

le modèle s'écrit alors

$$Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \gamma_i X_{ij} + \varepsilon_{ij}.$$

Le dernier terme $\gamma_i X_{ij}$ peut-être considéré comme un terme d'interaction entre le facteur et la variable quantitative X.

En notant

• Y le vecteur aléatoire $(Y_{ij} \setminus i = 1, ..., p; j = 1, ..., n_i)'$,

• ε le vecteur aléatoire $(\varepsilon_{ij} \setminus i = 1, ..., p; j = 1, ..., n_i)'$,

• θ le vecteur des paramètres $(\mu, \alpha_1, ..., \alpha_p, \beta, \gamma_1, ..., \gamma_p)'$,

• X est la matrice de taille $n \times (2p+2)$ suivante

$$X = \begin{pmatrix} 1 & 1 & ... & 0 & X_{11} & X_{11} & ... & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 1 & 1 & ... & 0 & X_{1n_1} & X_{1n_1} & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 1 & 0 & ... & 1 & X_{p1} & 0 & ... & X_{p1} \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 1 & 0 & ... & 1 & X_{pn_p} & 0 & ... & X_{pn_p} \end{pmatrix}$$

le modèle peut se mettre sous la forme matricielle

$$Y = X\theta + \varepsilon, \qquad \text{avec } \varepsilon \backsim N(0, \sigma^2 I_n),$$

où I_n est la matrice identité de taile $n \times n$.

L'application d'ANCOVA nécessite l'absence d'interaction entre le facteur et la covariable (les γ_i sont tous nuls), donc le modèle d'ANCOVA 1 sera donné par

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

qu'on préfère l'écrire sous la forme

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij}, \tag{2.2}$$

où β est la pente de la régression de Y en X et $\bar{X}$ est la moyenne générale de X.

On peut également réécrit ce modèle sous la même forme matricielle ci-dessus en éliminant

les p dernières colonnes de la matrice X et en changeant la colonne des X_{ij} par $(X_{ij}-\bar{X})$ pour tout $i=\overline{1,p}, j=\overline{1,n_i}$ et bien sûr le vecteur des paramètres devient $\theta=(\mu,\alpha_1,...,\alpha_p,\beta)'$.
Le modèle (2.2) devient identifiable en ajoutant la contrainte

$$\sum_{i=1}^{p} n_i\alpha_i = 0.$$

2.2.3 Estimation des paramètres du modèle d'ANCOVA 1

Notons $\bar{X}_i$ et $\bar{Y}_i$ respectivement les moyennes des valeurs de X et de Y pour le traitement i $(i=1,2,...,p)$ et $\bar{X}$ et $\bar{Y}$ respectivement les moyennes de toutes les valeurs de X et de Y.
On estime le vecteur $\theta=(\mu,\alpha_1,...,\alpha_p,\beta)'$ à l'aide de la méthode des moindres carrés.
On cherche donc à minimiser

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i-\beta(X_{ij}-\bar{X}))^2,$$

par l'annulation des dérivées partielles, on obtient le système d'équations suivant

$$\begin{cases} n\hat{\mu}+\sum_{i=1}^{p} n_i\hat{\alpha}_i = n\bar{Y} \\ n_i\hat{\mu}+n_i\hat{\alpha}_i+n_i(\bar{X}_i-\bar{X})\hat{\beta}=n_i\bar{Y}_i, \quad i=\overline{1,p} \\ \sum_{i=1}^{p} n_i(\bar{X}_i-\bar{X})\hat{\alpha}_i+\hat{\beta}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2=\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})Y_{ij} \end{cases},$$

sous la contrainte $\sum_{i=1}^{p} n_i\hat{\alpha}_i=0$, on obtient alors

$$\begin{cases} \hat{\mu}=\bar{Y} \\ \hat{\alpha}_i=\bar{Y}_i-\bar{Y}-\hat{\beta}(\bar{X}_i-\bar{X}), \quad i=\overline{1,p} \end{cases},$$

avec $\hat{\beta}$ est la solution de l'équation

$$\sum_{i=1}^{p} n_i(\bar{X}_i-\bar{X})\hat{\alpha}_i+\hat{\beta}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2=\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})Y_{ij}. \tag{2.3}$$

En remplaçant alors $\hat{\alpha}_i$ par son expression dans (2.3), on obtient

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

Dans le modèle (2.2), on estime alors le vecteur $\theta=(\mu,\alpha_1,...,\alpha_p,\beta)'$ par

$$\begin{cases}\hat{\mu}=\bar{Y}\,;\\ \hat{\alpha}_i=\bar{Y}_i-\bar{Y}-\hat{\beta}(\bar{X}_i-\bar{X})\text{ pour tout }i=1,2,...,p\,;\\ \hat{\beta}=\dfrac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(Y_{ij}-\bar{Y}_i)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2}\,;\end{cases}\tag{2.4}$$

les résidus sont estimés par

$$\hat{\varepsilon}_{ij}=Y_{ij}-(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}(X_{ij}-\bar{X}))=Y_{ij}-\bar{Y}_i-\hat{\beta}(X_{ij}-\bar{X}_i).$$

2.3 Moyennes ajustées

Considérant l'équation (2.2), la variable dépendante peut être ajustée pour la covariable comme suit

$$Y_{ij\ aj}=Y_{ij}-\beta(X_{ij}-\bar{X})=\mu+\alpha_i+\varepsilon_{ij},\ \forall i=\overline{1,p},\ \forall j=\overline{1,n_i}\,;\tag{2.5}$$

où $Y_{ij\ aj}$ est la valeur de la variable dépendante ajustée pour la covariable. Comme dans l'analyse de la variance, dans l'ANCOVA 1, on s'intéresse principalement à tester l'absence d'effet du facteur ajusté pour la covariable sur la variable dépendante ou à 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 par

$$\begin{cases} H_0 : & \alpha_1 = \alpha_2 = ... = \alpha_p = 0 \\ H_1 : & \exists i \in \{1, 2, ..., p\} \text{ tel que } \alpha_i \neq 0 \end{cases}, \tag{2.6}$$

peut être réécrit de la forme suivante

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

L'estimation des moyennes ajustées est donnée par

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

tel que $\hat{\beta}$ est défini par (2.4).

Remarque 2.1 *- 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 (2.5) 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.

2.4 Etapes de la réalisation de l'ANCOVA 1

En utilisant une méthode similaire à celle présentée dans le premier chapitre, on peut résumer les étapes de la réalisation de l'ANCOVA 1 comme suit

2.4.1 Vérification des conditions d'application d'ANCOVA

Les hypothèses implicites de l'ANCOVA 1 recouvrent celles de l'ANOVA et de la régression, et on postule alors que

1. Les résidus sont indépendants, et se distribuent normalement ;
2. La variance des résidus est homogène entre les différentes modalités du facteur ;
3. La relation entre la covariable et la variable dépendante est linéaire ;
4. Absence d'effet du facteur étudié sur la covariable (la pente de cette relation linéaire est la même pour les différentes modalités du facteur) ;
5. Absence d'interaction entre la facteur étudier et le rapport entre la covariable et la variable dépendante.

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

Par analogie avec l'analyse de variance, on peut calculer les sommes suivantes ([7])

1. La somme des carrés totale pour X

$$(SCT)_X = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2. \tag{2.7}$$

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

$$SPT = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})(Y_{ij} - \bar{Y}). \tag{2.8}$$

4. La somme des carrés des traitements pour X

$$(SCF)_X = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2. \tag{2.9}$$

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

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

$$SPF = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})(\bar{Y}_i - \bar{Y}). \tag{2.10}$$

7. La somme des carrés des erreurs pour X

$$(SCE)_X = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2. \tag{2.11}$$

8. La somme des carrés des erreurs pour Y (SCE) elle se calcule par la formule (2.11) dont en remplaçant les X par des Y.

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

$$SPE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i). \tag{2.12}$$

On peut démontrer que la relation entre les sommes des produits se comporte de façon similaire à la relation entre les somme de carrés $(SCT = SCF + SCE)$.

2.4.3 Calcul des sommes des carrés ajustées

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

1- La somme des carrés totale ajustée : Elle est notée par SCT_{aj} et elle est calculée comme suit

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

avec

$$\hat{Y}_{ij} = \bar{Y} + \hat{\beta}_T \left(X_{ij} - \bar{X} \right), \ \forall i = \overline{1,p}, \ \forall j = \overline{1,n_i}, \tag{2.13}$$

tels que

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

$\hat{\beta}_T$ l'estimateur de coefficient de régression donné par (1.9), ou plus précisément par

$$\hat{\beta}_T = \frac{\sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X} \right) \left(Y_{ij} - \bar{Y} \right)}{\sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X} \right)^2}. \tag{2.14}$$

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 résiduelle autour de la droite de régression.

mais de l'équation (2.14)

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

donc la somme des carrés totale ajustée est donnée par

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

• En utilisant la somme des produits totale de X et Y (2.8), la somme des carrés totale ajustée est exprimée de la manière suivante

$$SCT_{aj} = SCT - \hat{\beta}_T^2 \left(SCT \right)_X = SCT - \frac{SPT^2}{\left(SCT \right)_X}. \tag{2.15}$$

2- La somme des carrés des erreurs ajustée

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 (2.4).

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 (2.4)

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

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 (2.4), la somme des carrés des erreurs ajustée peut être écrite de la forme suivante

$$SCE_{aj} = SCE - \hat{\beta}^2\left(SCE\right)_X = SCE - \frac{SPE^2}{\left(SCE\right)_X}. \tag{2.16}$$

• 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 écarts 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 du facteur, peut être calculée par soustraction

$$SCF_{aj} = SCT_{aj} - SCE_{aj}. \tag{2.17}$$

2.4.4 Tests d'hypothèses

La statistique de test (qui va servir à tester l'hypothèse nulle qu'il n'y a pas de différence significative entre les moyennes des traitements une fois la variable Y ajustée) peut être définie comme dans une analyse de variance à un facteur par

$$F = \frac{CMF_{aj}}{CME_{aj}} \tag{2.18}$$

où $CMF_{aj} = \frac{SCF_{aj}}{p-1}$ et $CME_{aj} = \frac{SCE_{aj}}{n-p-1}$.

Sous l'hypothèse H_0, cette statistique suit une loi de Fisher de degrés 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.

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

Source de variation	Degrés de liberté 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.

Remarque 2.2 - *Les réductions appliquées aux sommes des carrés des erreurs et totale de Y se présentent sous la même forme que celle qui concerne, en régression, le passage de la*

somme des carrés totale à la somme des carrés résiduelle. (Une portion de variation qui est prédictible de la connaissance de la covariable à été éliminée).

- Par rapport à l'ANOVA 1, on remarque une réduction d'une unité des nombres de degrés de liberté des deux sommes des carrés ajustées (SCE_{aj}, SCT_{aj}). Cette réduction de degrés de liberté peut être justifiée par l'estimation du paramètre supplémentaire β. En règle générale, on perde un degré de liberté pour chaque covariable utilisée dans le modèle.

- Pour tester la nullité du coefficient β dans le modèle d'ANCOVA 1, on utilise la statistique du test suivante

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

qui sous l'hypothèse nulle ($H_0 : \beta = 0$), est de loi Fisher avec 1 et $(n - p - 1)$ degrés de liberté. On rejette H_0 au risque α, si la réalisation de cette statistique est plus grande que la valeur critique, lue dans la table de Fisher, $f_{(\alpha,1,n-p-1)}$.

2.5 Tests de comparaison des droites de régression

Le modèle d'ANCOVA 1 (2.2) est un cas particulier du modèle (2.1), où ce dernier correspond à p droites de régression distinctes dont les pentes diffèrent, ainsi que les ordonnées à l'origine, alors que le modèle d'ANCOVA 1 correspond à p droites étant supposées de même coefficient de régression β, c'est-à-dire parallèles. Ces droites ont comme ordonnées à l'origine la quantité $(\mu + \alpha_i + \beta\bar{X})$, d'où le test cité en (2.6) est un test d'égalité de ces ordonnés à l'origine.

Dans le cas général, on peut voir l'analyse de covariance à un facteur et une covariable comme une comparaison de plusieurs droites de régression ou de plusieurs modèles emboîtés (réduits) construits à partir d'un modèle emboîtant (complet) comme le modèle (2.2) ou plus généralement le modèle (2.1).

En se basant sur l'approche de l'erreur conditionnelle reposant sur la diminution de l'erreur obtenue par les paramètres supplémentaires du modèle emboîtant [16] ou sur le test de comparaison entre modèles emboîtés donné par le théorème 3.2 de [6], différentes hypothèses sont alors testées en comparant les modèles suivants

$M_1 : Y_{ij} = \mu + \alpha_i + \beta_i(X_{ij} - \bar{X}) + \varepsilon_{ij}$,

$M_2 : Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij}$, (Modèle d'ANCOVA 1)

$M_3 : Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,

$M_4 : Y_{ij} = \mu + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij}$,

ceci revient à considérer les hypothèses suivantes

a) Pas d'interaction, H_0 :"$\beta_1 = \beta_2 = ... = \beta_p = \beta$", les droites partagent la même pente,

b) H_0 :"$\beta = 0$",

c) H_0 :"$\alpha_1 = \alpha_2 = ... = \alpha_p = 0$", les droites partagent la même constante à l'origine.

On peut commencer donc par évaluer **a)**, si le test n'est pas significatif, on regard **b)** qui, s'il n'est pas non plus significatif, conduit à l'absence d'une relation linéaire entre Y et X. De même, toujours si **b)** n'est pas significatif, on s'intéresse à **c)** pour juger de l'effet du facteur. Alors tester l'hypothèse d'absence d'effet du facteur dans l'ANCOVA 1 est équivalent à comparer le modèle (M_2) avec le modèle (M_4). Dans ce cas la statistique qui lui est associé vaut

$$F = \frac{(SCE(M_4) - SCE(M_2))/(n - 2 - (n - p - 1))}{SCE(M_2)/(n - p - 1)}, \tag{2.19}$$

où

$SCE(M_4)$ correspond à la somme des carrés des erreurs du modèle réduit (M_4)

$SCE(M_2)$ correspond à la somme des carrés des erreurs du modèle complet (M_2)

$(n-2)$ et $(n-p-1)$ sont respectivement les nombres de degrés de liberté associé aux sommes des carrés des erreurs du modèle réduit (M_4) et du modèle complet (M_2).

Sous H_0, cette statistique suit une loi de Fisher à $(n - 2 - (n - p - 1))$ et $(n - p - 1)$ degrés de liberté, donc, $F \overset{H_0}{\backsim} f(p - 1, n - p - 1)$.

L'hypothèse H_0 sera rejetée si l'observation de la statistique F est supérieure ou égale à la valeur critique $f_{(\alpha,p-1,n-p-1)}$.

Du développement de l'équation (2.15) il est noté que $SCE(M_4) = SCT_{aj}$, donc la statistique

F (2.19) peut être reformulée comme

$$F = \frac{(SCT_{aj} - SCE_{aj})/(p-1))}{SCE_{aj}/(n-p-1)} = \frac{SCF_{aj}/(p-1)}{SCE_{aj}/(n-p-1)},$$

ce qui équivaut à ce qu'on a donné précédemment comme l'équation (2.18).

De la même manière on peut tester les hypothèses citées en **a** et **b** en faisant respectivement une comparaison entre les modèles (M_1) et (M_2) et entre les modèles (M_3) et (M_4).

2.6 Illustration sur un example

On souhaite étudier la croissance de 3 variétés de Leucaena Leucocephala (Cassie Blanc ou faux mimosa) sur une période de 4 mois.

Chaque variété est cultivée dans 10 pracelles d'une station expérimentale. On dispose des hauteurs moyennes initiales et finales de chaque parcelle (basées sur 40 observations chacunes). Les données sont enregistrées dans le tableau (2.2). ([11])

Variété 1		Variété 2		Variété 3	
H Init (X)	H Finale (Y)	H Init (X)	H Finale (Y)	H Init (X)	H Finale (Y)
18	145	27	161	31	180
22	149	28	164	27	158
26	156	27	172	34	183
19	151	25	160	32	175
15	143	21	166	35	195
25	152	30	175	36	196
16	144	21	156	35	187
28	154	30	175	23	147
23	150	22	158	34	184
24	151	25	165	32	184

TAB. 2.2 – Hauteurs moyennes initiales et finales de chaque parcelle.

On netera

• X_{ij} la hauteur moyenne initiale de la parcelle j $(1 \leq j \leq 10)$ cultivée avec la variété i $(1 \leq i \leq 3)$,

• y_{ij} la hauteur moyenne finale de la parcelle j cultivée avec la variété i,

• n le nombre total des données, $n = 30$.

Nous allons modéliser ces données avec un modèle de covariance à 1 facteur (Variété) et 1 covariable (Hauteur initiale).

On suppose donc que pour tout (i, j) la donnnée y_{ij} est une réalisation de la variable aléatoire Y_{ij} définie par

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij}, \qquad (\varepsilon_{ij}) \overset{iid}{\backsim} N(0, \sigma^2).$$

Pour tester l'hypothèse d'absence de l'effet de variété sur leur croissance (H_0) au risque $\alpha = 5\%$, il faut suivre les étapes suivantes

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

$\bar{X}_1 = 21.6$, $\bar{X}_2 = 25.6$, $\bar{X}_3 = 31.9$, $\bar{X} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3}{3} = 26.36$,

$\bar{Y}_1 = 149.5$, $\bar{Y}_2 = 165.2$, $\bar{Y}_3 = 178.9$, $\bar{Y} = \frac{\bar{Y}_1+\bar{Y}_2+\bar{Y}_3}{3} = 164.53$.

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

- Des équations (2.7) à (2.12), on trouve

$(SCT)_X = (18 - 26.36)^2 + (22 - 26.36)^2 + ... + (32 - 26.36)^2 = 966.966$,

$(SCT)_Y = (145 - 164.53)^2 + (149 - 164.53)^2 + ... + (184 - 164.53)^2 = 7073.467$,

$SPT = (18 - 26.36)(145 - 164.53) + ... + (32 - 26.36)(184 - 164.53) = 2367.13$,

$(SCF)_X = 10(21.6 - 26.36)^2 + 10(25.6 - 26.36)^2 + 10(31.9 - 26.36)^2 = 539.266$,

$(SCF)_Y = 10(149.5 - 164.53)^2 + 10(165.2 - 164.53)^2 + 10(178.9 - 164.53)^2 = 4328.467$,

$SPF = 10(21.6 - 26.36)(149.5 - 164.53) + ... + 10(31.9 - 26.36)(178.9 - 164.53) = 1506.43$,

$(SCE)_X = (18 - 21.6)^2 + (22 - 21.6)^2 + ... + (32 - 31.9)^2 = 427.70$,

$(SCE)_Y = (145 - 149.5)^2 + (149 - 149.5)^2 + ... + (184 - 178.9)^2 = 2745$,

$SPE = (18-21.6)(145-149.5)+(22-21.6)(149-149.5)+...+(32-31.9)(184-178.9) = 860.7$.

On peut voir que

$(SCE)_X = (SCT)_X - (SCF)_X = 966.9667 - 539.266 = 427.70$,

$(SCE)_Y = (SCT)_Y - (SCF)_Y = 7073.467 - 4328.467 = 2745$,

$SPE = SPT - SPF = 2367.13 - 1506.43 = 860.7$.

• Calcul des sommes des carrés ajustées

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

1. De l'équation (2.15) la somme des carrés totale ajustée

$$SCT_{aj} = 7073.467 - \frac{(2367.13)^2}{966.9667} = 1278.744.$$

2. De l'équation (2.16) la somme des carrés des erreurs ajustée

$$SCE_{aj} = 2745 - \frac{(860.7)^2}{427.70} = 1012.934.$$

3. De l'équation (2.17) la somme des carrés des traitements ajustée

$$SCF_{aj} = 1278.744 - 1012.934 = 265.81.$$

• Décision

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

$$f = \frac{CMF_{aj}}{CME_{aj}} = \frac{SCF_{aj}/2}{SCE_{aj}/26} = \frac{265.81/2}{1012.934/26} = 3.411.$$

De la table de Fisher et pour un seuil de signification $\alpha = 5\%$, on trouve que $f_{(0.05,2,26)} = 3.37$. Comme $f > f_{(0.05,2,26)}$, on doit rejeter H_0 au seuil de signification α et on considère que la variété des Leucaena influe de manière significative sur leur croissance.

Les résultats d'ANCOVA 1 sont résumés dans le tableau suivant

Source de variation	Degrés de liberté *ddl*	Somme des carrés *SC*	Carré Moyen *CM*	Ratio *f*
Taitements	2	265.81	132.905	3.411
Erreurs	26	1012.934	38.959	
Total	28	1278.744		

TAB. 2.3 – Résultats obtenus par l'ANCOVA 1.

Chapitre 3

Application sous R

Dans ce chapitre, on traite en pratiquement sous R les méthodes statistiques qu'on a vu dans les chapitres précédents. On donne quelques exemples sur la méthode d'ANOVA 1, de régression linéaire simple et d'ANCOVA 1.

3.1 Exemple sur l'ANOVA 1

Le but de cet exemple est de comparer 4 groupes d'athlètes habilité à sauter en hauteur. Les données sont enregistrées dans le tableau (3.1).

Inspection graphique : Tous d'abord, nous allons effectuer une brève analyse descriptive de ces données pour voir si certaines tendances probables se dégagent. ([5])

Athlètes	Sport	Hauteur saut (cm)	Athlètes	Sport	Hauteur saut (cm)
1	Soccer	38	1	Football	55
2	Soccer	43	2	Football	68
3	Soccer	33	3	Football	43
4	Soccer	40	4	Football	45
5	Soccer	35	5	Football	53
1	Tennis	45	1	Basketball	60
2	Tennis	53	2	Basketball	65
3	Tennis	38	3	Basketball	55
4	Tennis	55	4	Basketball	53
5	Tennis	50	5	Basketball	55

TAB. 3.1 – Hauteur des sauts des athlètes.

```
> X<-data.frame(Soccer=c(38,43,33,40,35),Tennis=c(45,53,38,55,50),
Football=c(55,68,43,45,53),Basketball=c(60,65,55,53,55))
> Hauteur.saut.cm<-stack(X)$values
> Sport<-stack(X)$ind
> tapply(Hauteur.saut.cm,Sport,summary)
$Soccer
   Min. 1stQu. Median Mean 3rdQu. Max.
   33.0 35.0 38.0 37.8 40.0 43.0
$Tennis
   Min. 1stQU. Median Mean 3rdQu. Max.
   38.0 45.0 50.0 48.2 53.0 55.0
$Football
   Min. 1stQu. Median Mean 3rdQu. Max.
   43.0 45.0 53.0 52.8 55.0 68.0
$Basketball
   Min. 1stQu. Median Mean 3rdQu. Max.
   53.0 55.0 55.0 57.6 60.0 65.0
```

Le test porte sur la comparaison des moyennes. À cette étape, il serait bon de tracer les boîtes à moustaches de la variable hauteur saut (cm) en fonction de la variable sport, donnée dans la figure (3.1). Pour cela tapez la ligne de commande suivante

```
> plot(Hauteur.saut.cm~Sport,pch=16,cex=0.5,col="green")
```

Remarque 3.1 *La boîte à moustaches résume seulement quelques caractéristiques de position du caractère étudié (**Médiane**, **Quartille**, **Minimum**, **Maximum**). Ce diagramme est utilisé principalenment pour comparer un même caractère dans deux populations de tailles différentes.*

• **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.

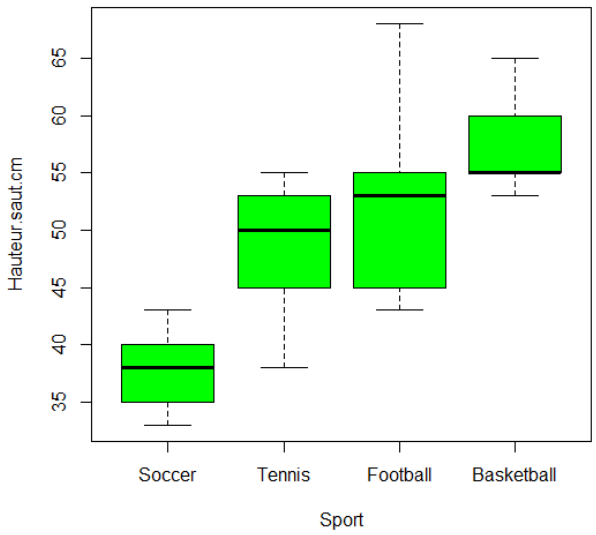

FIG. 3.1 – Les boîtes à moustaches de la variable hauteur saut (cm) en fonction de la variable sport.

> model1.aov<-aov(Hauteur.saut.cm~Sport)

> summary(model1.aov)

	Df	Sum Sq	Mean Sq	F value	Pr(> F)	
Sport	3	1072.2	357.4	7.753	0.00203	**
Residuals	16	737.6	46.1			

- - -

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

Remarque 3.2 *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(Hauteur saut cm~Sport)`

> `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 $p = 0.00203$ nous permet de conclure que les hauteurs des sauts (cm)

d'athlètes d'au moins deux sports sont différentes au risque 5%.

Remarque 3.3 *Ce qui concerne les conditions d'application d'ANOVA 1, on peut tester la normalité par le test de* ***shapiro-wilk*** *comme suit*

> `residus<-residuels(model1)`

> `shapiro.test(residus)`

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

>*bartlett.test(residus)*

3.2 Exemple sur la régression linéaire simple

On considère 5 groupes de femmes âgées. Dans chaque groupe, on a mesuré la tension artérielle en mm de mercure de chaque femme et on a calculé la valeur moyenne pour chaque groupe. On définit donc les variables ([1])

X	Âge du groupe considéré	35	45	55	65	75
Y	Tension moyenne en mm Hg	114	124	143	158	166

TAB. 3.2 – Âge et tension moyenne en mm de mercure de chaque groupe de femmes.

• **Lecture des données :**

> X<-c(35,45,55,65,75)

> Y<-c(114,124,143,158,166)

• **Inspection graphique :**

Afin d'étudier la relation entre l'âge du groupe et la tension moyenne en mm nous pouvons commencer par tracer le nuage des points, donnée dans la figure (3.2) grâce à l'instruction

> plot(Y,xlab="Âge du groupe",ylab="Tension moyenne en mm")

• **Estimation des paramètres :** On estime le modèle par la fonction lm().

> modele1<-lm(Y~X)

> modele1

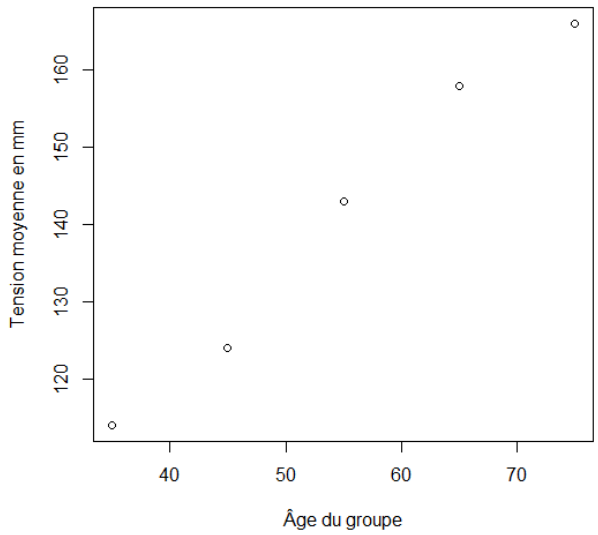

FIG. 3.2 – Nuage de points de tension moyenne en mm selon l'âge du groupe.

Call :

lm(formula = Y˜X)

Coefficients :

(Intercept)	X
65.10	1.38

On peut maintenant réprésenter la droite de régression sur le nuage de points au moyen de la fonction abline (), donnée dans la figure (3.3)

> plot (Y˜X, xlab="Âge du groupe", ylab="Tension moyenne en mm")

> abline(modele1,col="blue")

• **Tableau d'analyse de 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

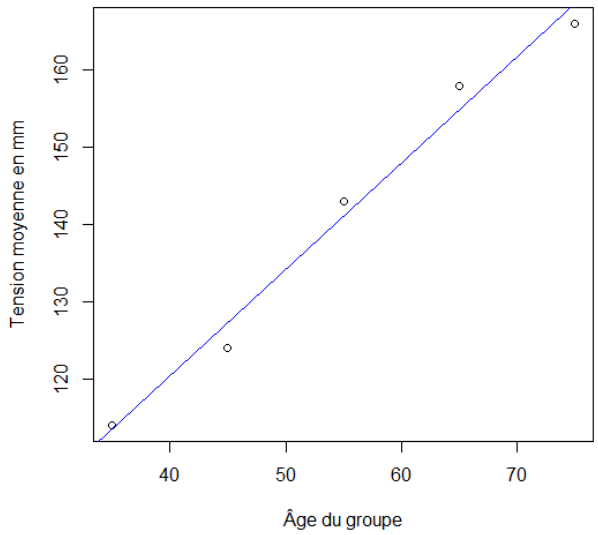

FIG. 3.3 – Représentation de la droite de régression des moindres carrés sur le nuage de points.

Response :Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X	1	1904.4	1904.40	180.8	0.0008894	***
Residuals	3	31.6	10.53			

- - -

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

Remarque 3.4 *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 significative entre l'âge du groupe et la tension moyenne en mm.*

3.3 Exemple sur l'ANCOVA 1

Une entreprise a étudié les effets de trois types de promotions sur les ventes de son produit alimentaire :

1 : Échantillonnage du produit par les clients ;

2 : Éspace additionnel dans les étagères habituelles ;

3 : Étalage additionnel dans les allées.

Quinze magasins ont été sélectionnés d'une façon aléatoire pour l'étude. Chaque magasin s'est vu attribuer au hasard l'un des types de promotion, avec cinq magasins attribué à chaque type de promotion. Autres conditions pertinentes sous le contrôle de l'entreprise, comme le prix et la publicité, ont été conservées les mêmes pour tous les magasins de l'étude. Les données sur le volume des ventes du produit pendant la période promotionnelle, notés Y-vente, sont présentées dans le tableau (3.3), ainsi que les données sur le volume des ventes du produit au cours de la période précédente (avant le période promotionnelle), désigné par X-vente avant. ([5])

Promotion 1		Promotion 2		Promotion 3	
X-vente avant	Y-vente	X-vente avant	Y-vente	X-vente avant	Y-vente
21	38	34	43	23	24
26	39	26	38	29	32
22	36	29	38	30	31
28	45	18	27	16	21
19	33	25	34	29	28

TAB. 3.3 – Volume des ventes du produit pendant et avant la période promotionnelle.

Dans la présente application, l'objectif est de déterminer si le facteur promotion influe sur le volume des ventes du produit pendant la période promotionnelle indépendamment de volume des ventes de la période précédente. Avant de répondre à ce problème, on veut répondre aux questions suivantes

1- Y a-t-il un effet de promotion sur le volume des ventes du produit pendant la période promotionnelle sans tenir compte de son volume des ventes de la période précédente ?

2- Est-ce-que le volume des ventes du produit avant la promotion (X-vente avant) est différent selon la promotion ?

3- Comment représenter graphiquement le volume des ventes du produit pendant la période promotionnelle en fonction de son volume des ventes de la période précédente pour les différents niveaux du promotion ?

4- Y a-t'il une interaction entre la méthode de promotion et le rapport entre le volume des

ventes pendant la période promotionnelle et le volume des ventes de la période précédente ?

Réponses

Les données sont entrées dans R au moyen des instructions suivantes

```
> options(contrasts=c("contr.sum","contr.poly"))
> CampagnePromotionnelle=read.table(("E :/ANCOVA/campagnepromotionnelle.txt"),
header=TRUE,sep="\t")
> attach(CampagnePromotionnelle)
> names(CampagnePromotionnelle)
```

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 Y-vente et de X-vente avant dans chaque groupe

```
> tapply(Y.vente,Promotion,mean)
Promotion 1   Promotion 2   Promotion 3
     38.2          36.0          27.2
> tapply(X.vente.avant,Promotion,mean)
Promotion 1   Promotion 2   Promotion 3
     23.2          26.4          25.4
```

On représente graphiquement les données à l'aide des boîtes à moustaches des variables X-vente avant et Y-vente de promotion donnée dans la figure (3.4), pour cela on va taper les lignes de commandes suivantes

```
> par(mfrow=c(2,1))
> plot(Y.vente~Promotion,col="green",main="a")
> plot(X.vente.avant~Promotion,col="green",main="b")
```

A partir de ces résultats préliminaires, on remarque que la moyenne de Y-vente dans les promotions 1 et 2 est plus grande que la moyenne de Y-vente dans la promotion 3. Tandis que, la moyenne de X-vente avant dans les promotions 2 et 3 est supérieure à sa moyenne dans la promotion 1.

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

```
> aov1<-aov(Y.vente~Promotion)
summary(aov1)
```

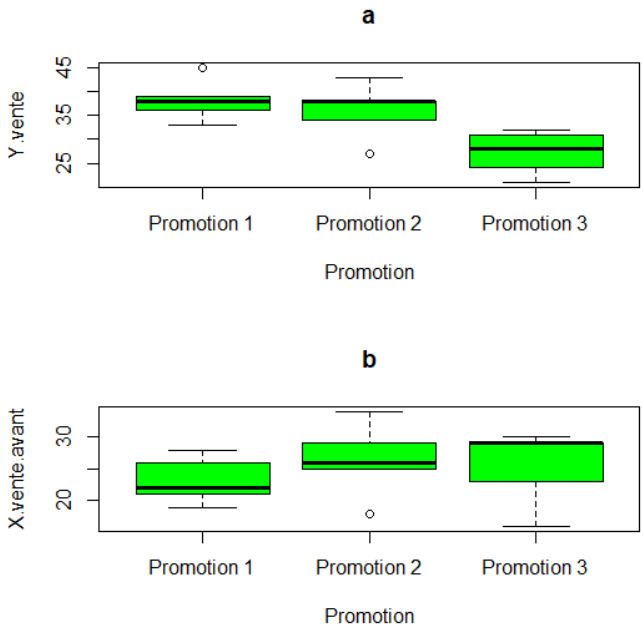

FIG. 3.4 – Boîtes à moustaches de Y-vente (a) et de X-vente avant (b) pour chaque promotion.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Promotion	2	338.8	169.40	6.609	0.0116	*
Residuals	12	307.6	25.63			

- - -

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 inférieure à 0.05, ce qui indique qu'il y a un effet significatif du facteur promotion sur Y-vente sans tenir compte de X-vente avant.

> aov2<-aov(X.vente.avant~Promotion)

summary(aov2)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Promotion	2	26.8	13.40	0.483	0.629
Residuals	12	333.2	27.77		

Alors 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 X-vente avant.

Pour la question 3, on peut tracer le nuage des points grâce à l'instruction plot(Y.vente~X.vente.avant)

```
> plot(Y.vente~X.vente.avant,data=CampagnePromotionnelle,type="n")
> points(Y.vente~X.vente.avant,data=subset(CampagnePromotionnelle,Promotion==
"Promotion 1"),col="blue",pch="1")
> points(Y.vente~X.vente.avant,data=subset(CampagnePromotionnelle,Promotion==
"Promotion 2"),col="green",pch="2")
> points(Y.vente~X.vente.avant,data=subset(CampagnePromotionnelle,Promotion==
"Promotin 3"),col="red",pch="3")
> legend(20,45,c("Promotion.1","Promotion.2","Promotion.3"),pch="123",
col=c("blue","green","red"),cex=1)
```

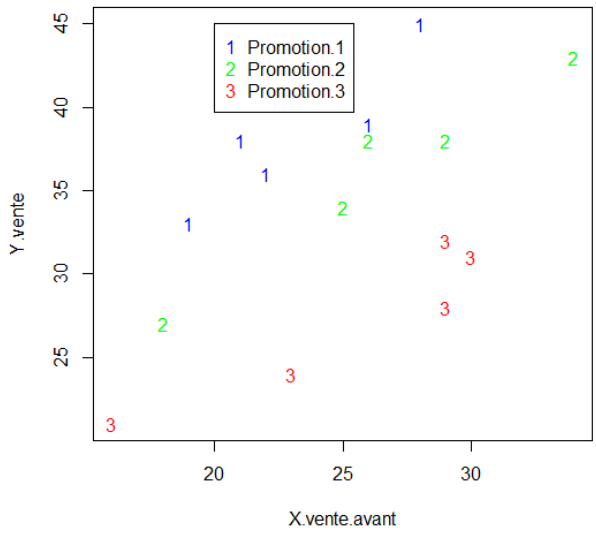

FIG. 3.5 – Nuage de points de Y-vente en fonction de X-vente avant.

La 4ème question concernant un test sur l'interaction entre X-vente avant et Promotion ou sur l'égalité des pentes des droites de régressions, qu'il se fait de la manière suivante

```
> mod1<-lm(Y.vente~X.vente.avant+Promotion+X.vente.avant :Promotion)
> anova(mod1)
Analysis of Variance Table
```

Response : Y.vente

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X.vente.avant	1	190.68	190.678	54.4434	4.198e-05	***
Promotin	2	417.15	208.575	59.5536	6.457e-06	***
X.vente.avant : Promotion	2	7.05	3.525	1.0065	0.4032	
Residuals	9	31.52	3.502			

- - -

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 X-vente avant et Promotion car la p valeur est supérieur à 0.05 (p-valeur=0.4032), d'où les droites de régressions sont parallèles.

Remarque 3.5 *Comme l'analyse de variance à un facteur, on teste 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(Y.vente~X.vente.avant+Promotion))

>anova(mod2)

Analysis of Variance Table

Response :Y.vente

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X.vente.avant	1	190.68	190.678	54.379	1.405e-05	***
Promotion	2	417.15	208.575	59.483	1.264e-06	***
Residuals	11	38.57	3.506			

- - -

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

De la table d'ANCOVA 1, on remarque que

- La p valeur associée au facteur promotion est inférieure à 0.05, ce qui veut dire que le facteur promotion influe significativement sur la variable Y-vente. Ce résultat montre que

si en tenant compte le volume des vantes avant la promotion, les moyennes des ventes sous promotion (moyennes ajustées) sont significativement différentes pour les trois promotions. Aussi, on peut dire que la modélisation des données ne sera pas par une seule droite.

- La somme des carrés des erreurs est plus petite qu'avant lorsqu'on a appliqué l'ANOVA 1 sans la covariable (X-vente avant), de plus l'ANCOVA 1 a montré qu'il y'a un effet hautement significatif du facteur promotion sur les volumes des ventes.

• **Calcul des moyennes ajustées**

Grâce aux instructions suivantes, on obtient les moyennes ajustées : meanPromotion1aj, meanPromotion2aj, meanPromotion3aj de la variable Y-ventes des trois promotions 1, 2, et 3 respectivement.

```
>coe2<-coefficients(mod2)
>meanPromotion1aj=mean(Y.vente[Promotion=="Promotion 1"])-
((coe2[2])*(mean(X.vente.avant[Promotion=="Promotion 1"])-mean(X.vente.avant)))
>meanPromotion2aj=mean(Y.vente[Promotion=="Promotion 2"])-
((coe2[2])*(mean(X.vente.avant[Promotion=="Promotion 2"])-mean(X.vente.avant)))
>meanPromotion3aj=mean(Y.vente[Promotion=="Promotion 3"])-
((coe2[2])*(mean(X.vente.avant[Promotion=="Promotion 3"])-mean(X.vente.avant)))
```

d'où meanPromotion1aj= 39.81741, meanPromotion2aj=34.74202, meanPromotion3aj=26.84058.

Notons qu'après l'ajustement, les moyennes ajustées divergent peu des moyennes non ajustées telles que la moyenne la plus élevée de Y- vente, c'est-à-dire 38.2 a été augmentée à 39.81741, et la moyenne la plus basse c'est-à-dire 36.0 et 27.2 ont été diminuées à 34.74202 et 26.84058 respectivement.

Ceci dû au fait que la différence entre les groupes au niveau de volume des ventes avant la promotion (X-vente avant) a été éliminée.

Remarque 3.6 *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 est une méthode d'analyse combine les éléments des modèles de régression et les modèles d'analyse de la variance. L'idée de base est d'augmenter les modèles d'analyse de variance contenant les effets de facteurs catégoriques (qualitatifs) avec une ou plusieurs variables continues (quantitatives) qui sont reliées à la variable de réponse. Cette augmentation a pour objectif de réduire la variance du terme d'erreur dans le modèle augmentant ainsi la sensibilité de l'analyse à détecter des effets significatifs. Les modèles d'analyse de covariance sont des cas particuliers des modèles d'analyse de régression avec un mélange de variables quantitatives et de variables qualitatives.

L'objectif du présent mémoire est de discuter le cas d'analyse de covariance à un facteur et une covariable (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 un test de 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.

2 - Evité les méfaits de la non comparabilité des groupes au niveau d'une 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.

Alors, on peut dire que l'intérêt du modèle d'ANCOVA, c'est qu'il permet de séparer l'effet spécifique du facteur étudié de l'effet de la covariable (la variable dépendante sera ajustée de façon à enlever l'effet provenant de la covariable, lequel effet nous n'avons pas pu contrôler directement au niveau de schéma expérimental retenu), réduire la variance résiduelle, ce qui augmente la puissance du test de l'effet du facteur étudié.

Le présent mémoire peut être enrichi par ce qui suit :

- D'introduire au modèle plusieurs facteurs et covariables.
- D'étudier en détail le lien entre l'ANCOVA et la régression linéaire multiple.

Bibliographie

[1] Azaïs, J. M, & Bardet, J. M. (2006). Le Modèle linéaire par l'exemple Régression, Analyse de la variance et Plans d'Exprériences Illustrations numériques avec les logiciels R, SAS et Splus.

[2] Bertrand, F. (2016-2017). Analyse de la covariance (ANCOVA).

[3] Besse, P. (2003). Pratique de la modélisation statistique. Publications du laboratoire de statistiques et probabilités. Université Paul Sabatier, Toulouse. Disponible à partir de l'URL http : // www-sv. cict. fr/lsp/Besse.

[4] Chavent, M. Analyse de la Variance, Chapitre 3. Licence 3 MIASHS-Université de Bordeaux.

[5] Clément, B, PHD. (2005). Modèles d'analyse de variance avec statistica, Génistat Conseils inc.

[6] Cornillon, P. A., & Matzner-lober, E. (2007). Régression : théorie et applications (pp. 302-p). Springer.

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

[8] Dress, F. (2007). Les probabilités et la statistique de A à Z : 500 définitions, Formules et tests d'hypothèse. Dunod.

[9] Fanny, M. Morgane, C. Margaux, G. Analyse de la variance, M2 Statistiques et Econométrie.

[10] Godichon-Baggioni, A. Analyse de la variance à 1 facteur. INSA de Rouen-Génie Mathématique-4$^{\text{ème}}$ année.

[11] Godichon-Baggioni, A. (2017-2018). Analyse de covariance, cours 6. INSA-GM4-cours de statistique.

[12] Le digabel, S. (2017). 12. Régression linéaire simple MTH2302D, École Polytechnique de Montréal.

[13] Lévy-Leduc, C. (2017-2018). Notes de cours sur les bases statistiques du modèle linéaire.

[14] Michel, C. (2015). Cours d'Analyse de la Variance. Département de Mathématiques et Statistique, Université de Laval.

[15] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

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

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 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 abréviations 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 moyens associés au facteur.
CME	:	Carrés moyens résiduels.
SCR	:	Variation de la régression.
$f_{(n_1,n_2)}$	:	Une loi de Fisher de degrés de liberté $n_1,\ n_2$.
SPT	:	Somme des produits total 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}	:	Sommes des carrés totale ajustée.
SCF_{aj}	:	Sommes des carrés des traitements ajustée.
SCE_{aj}	:	Sommes des carrés erreurs ajustée.
CMF_{aj}	:	Carrés des moyennes des traitements ajustée.
CME_{aj}	:	Carrés des moyennes des erreurs ajustée.
$E(\varepsilon_i)$	:	Espérance mathématique de la variable ε_i.

$var(\varepsilon_i)$	:	Variance mathématique de la variable ε_i.
$cov(\varepsilon_i, \varepsilon_j)$	:	Covariance entre ε_i et ε_j.
iid	:	indépendant identiquement distribué.