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

NOM Prénom

ABDELAIDOUM WAHIBA

Titre :

Analyse de covariance : Théorie et applications

Membres du Comité d'Examen :

Dr.	**BENBRIKA Ghozlane**	UMKB	Président
Dr.	**ROUBI Affef**	UMKB	Encadreur
Dr.	**BERKANE Hassiba**	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce modeste travail

Fuit de mes années d'étude et de patience qui est complété par l'aide de dieu.

A celui qui m'a offert la vie et à ce je dois réussir, source de sagesse, et de tendresse qui m'appris le respect et le sens de devoir et qui a sacrifié le tout pour me voir heureuse.

A toi cher père

A la prunelle de mes yeux celle qui m'a poussé moralement A la femme qui est toujours fiée de moi.

A toi cher mère

A mes chères sœurs

A mon chère frère

A toute ma famille, chacun de son nom.

A mes collègues

A mes chère 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.

2017-2018

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 Généralités 3

1.1.1 Les différents types d'ANOVA 3

1.1.2 Les principes d'ANOVA 4

1.2 ANOVA à un facteur (ANOVA1) 4

1.2.1 Les étapes de l'ANOVA 1 5

1.3 La régression linéaire 8

1.3.1 Analyse du modèle de régression linéaire simple 9

2 Analyse de covariance 13

2.1 Présentation des données d'une ANCOVA 13

2.2 Choix des covariables 14

2.3 Modèle d'analyse de covariance (ANCOVA) 14

2.4 Moyennes ajustées 15

2.5 Contribution des termes du modèle 15
2.6 Construction du modèle et tests d'intérêt 16

3 Application sous R **28**
3.1 Simulation d'analyse de la variance à un facteur 28
3.1.1 Vérification des conditions fondamentales d'ANOVA 31
3.2 Simulation d'analyse de covariance ANCOVA 32

Conclusion **40**

Bibliographie **42**

Annexe A : Logiciel *R* **43**

Annexe B : Abréviations et Notations **44**

Table des figures

2.1 Représentation des modèles 1 (DCH : droite commune horizontale) et 2 (DH : droites horizontales)(facteur à 3 niveau, réponse Y en ordonnée et covariable X en abscisse). 18

2.2 Représentation des modèles 3 (DV : droites verticales : $X_{ij} = \mu_x + \alpha_{x,i} + \varepsilon_{ij}$) et 4 (DC : droite commune oblique : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$). 19

2.3 Représentation des modèles 5 (DD : droites différentes :$Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij}$) et (DCO : droite commune passant par l'origine : $Y_{ij} = \beta_1 X_{ij} + \varepsilon_{ij}$). . 22

2.4 Représentation du modèle 7 (DP : droites parallèles :$Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \varepsilon_{ij}$). 24

2.5 Tableau d'analyse de covariance pour un plan à un facteur et une covariable , ou pour g droites de régression 27

3.1 Les boîtes à moustaches de la variable hauteur en fonction de la variable forêt. 30

3.2 Boites à moustaches de la durée de survie (a) et de l'âge d'apparition du cancer (b) pour chaque traitement. 35

3.3 Nuage de points de la durée de survie en fonction de l'âge. 37

Introduction

La statistique est la science dont le but est de donner un sens aux données. L'application de la statistique peut être utilisée sur les domaines de pharmacologie, psychologie, médecine, science sociales, économétrie,...etc.

Dont l'objectif est de prendre des décisions sur la base de résultats expérimentaux, 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 un problème.
Une hypothèse statistique est alors un énoncé au sujet d'une population qu'on cherche à conserver ou réfuter en s'appuyant sur l'information obtenue à partir de données observées. 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 se situe dans le cadre du modèle linéaire général, elle permet d'expliquer une variable quantitative Y par plusieurs variables explicatives de type à la fois quantitatives et qualitatives. Dans les cas les plus complexes, on peut avoir plusieurs facteurs avec une structure croi-

sée ou hiérarchisée, plusieurs variables quantitatives intervenant de manière linéaire ou polynomiale.
L'objectif de cette méthode alors sera de tenir compte, lors de l'étude, des effets des facteurs sur la variable Y, et des effets possibles de la ou des variables quantitatives auxiliaires ou concomitantes, appelées 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 différentes types et 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 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 va intéresser d'une technique statistique appelée l'analyse de la variance (en abréviation ANOVA).

L'analyse de la variance est un test statistique permettant de vérifier que plusieurs échantillons sont issus d'une même population.

Ce test s'applique lorsque l'on mesure une ou plusieurs variables explicatives catégorielles (appelées alors facteur de variabilité, leurs différentes modalités étant parfois appelées niveaux) qui ont de l'influence sur la distribution d'une variable continue à expliquer. On parle d'analyse de variance à un facteur lorsque l'analyse porte sur un modèle décrit par un seul facteur de variabilité, d'analyse à deux facteurs ou d'analyse multifactorielle sinon.

1.1 Généralités

1.1.1 Les différents types d'ANOVA

• Type I : modèle à effets fixe lorsque les modalités des facteurs sont choisies délibérément par l'expérimentateur. C'est le cas dans la plupart des protocoles expérimentaux, et c'est

le type que nous avons développé dans ce document.

• Type II : modèle à effet aléatoire lorsque les modalités des facteurs sont issues d'un processus d'échantillonnage.

• Type III : modèle à effets mixtes, lorsqu'on dispose des facteurs à effet fixe et des facteurs à effet aléatoire simultanément.

1.1.2 Les principes d'ANOVA

L'objectif d'ANOVA

L'analyse de variance ou ANOVA a pour objectif de tester l'effet d'un ou de plusieurs facteurs sur une variable aléatoire continue. Ceci revient à comparer les moyennes de plusieurs populations normales et de même variance à partir d'échantillons aléatoires et indépendants les uns des autres.

Quant utilises l'ANOVA

• Pour tester l'effet d'une variable indépendante "discrète".

• Chaque variable indépendante est appelée un facteur et chaque facteur peut avoir deux ou plusieurs niveaux ou traitements (ex : niveau d'irrigation ; température d'élevage ; région géographique, etc).

• Une ANOVA teste si toutes les moyennes sont égales, donc

$$\begin{cases} H_0 : \text{égalité}, \\ H_1 : \text{au moins une différence}. \end{cases}$$

• A utiliser quand le nombre de niveaux est supérieur à deux.

1.2 ANOVA à un facteur (ANOVA1)

Définition 1.2.1 *(ANOVA1)*

L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant g modalités (groupes) sur les moyennes d'une variable quantitative Y.

Les problèmes concernés par la technique ANOVA 1 s'écrivent en générale de la manière suivante

N°	groupe 1	groupe 2	$\cdots$	groupe g
1	Y_{11}	Y_{21}	$\cdots$	Y_{g1}
2	Y_{12}	Y_{22}	$\cdots$	Y_{g2}
3	Y_{13}	Y_{23}	$\cdots$	Y_{g3}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n_i	Y_{1n_1}	Y_{2n_2}	$\cdots$	Y_{gn_g}

TAB. 1.1 – Les données d'ANOVA 1.

Et le modèle mathématique leurs associés est donné par

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

où Y_{ij} est la $j^{ème}$ réalisation de la variable quantitative Y dans la $i^{ème}$ échantillon et ε_{ij} sont des erreurs de mesure.

Si on retient ce modèle alors le test à réaliser est défini par

$$H_0 : "\mu_1 = \mu_2 = ... = \mu_g = \mu" \text{ contre } H_1 : "\exists \ i, j \in \{1, 2, ..., g\} \text{ tel que } \mu_i \neq \mu_j". \quad (1.1)$$

Dans ce qui suit, nous allons énumérer les étapes de la mise en œuvre de l'ANOVA1 qui nous permet de réaliser ce test.

1.2.1 Les étapes de l'ANOVA 1

Afin de réaliser le test défini dans (1.1), trois conditions doit être vérifiées préalablement, à savoir

• Les g échantillons comparés sont indépendants.

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

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

Si ces dernières conditions sont vérifiées alors, on peut utiliser la technique ANOVA 1 pour réaliser le test (1.1), et pour se faire nous avons besoin des quantités (statistiques) suivantes

• La moyenne de toutes les observations $\overline{Y} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{ni} Y_{ij}$ avec $n = \sum_{i=1}^{g} n_i$.

• La moyenne de chaque échantillon $\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{ni} Y_{ij}$, pour $i = \overline{1,g}$.

• La variance de toutes les observations $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y})^2$.

• La variance de chaque échantillon $\widehat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y}_i)^2$, pour $i = \overline{1,g}$.

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 g échantillons, c'est-à-dire

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y})^2 = \frac{1}{g}\sum_{i=1}^{g}\sigma_i^2 + \frac{1}{g}\sum_{i=1}^{g}(\overline{Y}_i - \overline{Y})^2, \tag{1.2}$$

ou encore

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y}_i)^2 + \frac{1}{g}\sum_{i=1}^{g}(\overline{Y}_i - \overline{Y})^2. \tag{1.3}$$

On multipliant par n on obtient

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{ni}(Y_{ij} - \overline{Y}_i)^2}_{SCE} + \underbrace{\sum_{i=1}^{g}(\overline{Y}_i - \overline{Y})^2}_{SCF}. \tag{1.4}$$

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 intra-groupes (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, où

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

Si les 3 conditions d'application d'ANOVA (Indépendance, Normalité et Homogénéité) sont vérifiées et H_0 est vraie, alors

$$F = \frac{CMF}{CME} \sim f_{(g-1,n-g)}, \text{ } F \text{ suit une loi de Fisher de degrés de liberté } (g-1) \text{ et } (n-g).$$

Décision

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

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

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

• Si $f \geq f_{\alpha,g-1,n-g} \Longrightarrow$ 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	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio f_{obs}
Inter-groupe(Fac)	SCF	$g-1$	CMF	$\frac{CMF}{CME}$
Intra-groupe(Rés)	SCE	$n-g$	CME	
Total	SCT	$n-1$		

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

1.3 La régression linéaire

La régression est l'une des méthodes les plus connues et les plus appliquées en statistiques pour l'analyse de données quantitatives sous forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant l'une des deux variables en fonction de l'autre. Tandis que, si la relation porte entre une variable et plusieurs autres variables (≥ 2), on parlera de régression multiple.

La mise encœuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle. Cette méthode peut être mise en place sur des données quantitatives observées sur n individus et présentées sous forme

$$Y = f(x_1, x_2, ..., x_p) + \varepsilon, \tag{1.5}$$

où

• Y est une variable quantitative prenant la valeur Y_i pour l'individu i $(i = 1, ..., n)$, appelée variable à expliquer ou variable réponse.

• $x_1, x_2, ..., x_p$ sont p variables quantitatives prenant respectivement les valeurs $x_{1i}, x_{2i}, ..., x_{pi}$ pour le $i^{ème}$ individu, appelées variables explicatives ou prédicteurs.

• ε est une variable aléatoire (résidus).

Considérons un couple de variables quantitatives (X, Y). S'il existe une liaison entre ces deux variables, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y. Si l'on admet qu'il existe une relation de cause à effet entre X et Y , le phénomène aléatoire représenté par X peut donc servir à prédire celui représenté par

Y et la liaison s'écrit sous la forme (1,5) et on dit que l'on fait de la régression de Y sur X (dans le cas d'une régression multiple de Y sur $x_{1,}x_{2,...,}x_p$ la liaison peuvent être écrite sous la forme $Y = f(x_{1,}x_{2,...,}x_p)$).

Dans les cas les plus fréquents, on choisit l'ensemble des fonctions affinées du type

Cas de régression linéaire simple

$$f(x) = ax + b \ . \tag{1.6}$$

Cas de régression linéaire multiple

$$f(x_{1,}x_2, ..., x_p) = a_0 + a_1x_1 + a_2x_2 + ... + a_px_p. \tag{1.7}$$

Dans cette section, nous intéresserons à la régression linéaire simple car elle est la plus simple qui ne considère qu'une seule variable explicative.

1.3.1 Analyse du modèle de régression linéaire simple

Soit le couple (X, Y) de variables aléatoires où X est une variable indépendante et Y la variable dépendante. On cherche une relation du type

$$Y = a + bX + \varepsilon. \tag{1.8}$$

Notons que la mise en œuvre et l'exploitation de ce modèle nécessite une quantification préalable des paramètres inconnus a et b.

Estimation des paramètres du modèle

On suppose que la variable X est contrôlée par l'expérimentateur où il réalise n expériences $y_1, y_{2,}..., y_n$ aux points $x_1, x_2, ..., x_n$ fixés. De plus, on suppose que les Y_i sont manuellement

indépendants. Le modèle s'écrit

$$y_i = a + bx_i \qquad \text{pour } i = \overline{1,n}\ ,$$

tel que $E(\varepsilon_i) = 0$, $cov(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j$ et $var(\varepsilon_i) = \sigma^2$,$\forall i = \overline{1,n}$.
Supposons qu'on opte pour la méthode de moindre carrées pour quantifier a et b, alors les estimateurs des paramètres a et b sont $\widehat{a}$ et $\widehat{b}$ qui minimise la fonction $Q(a,b)$, définie par :

$$Q(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2. \tag{1.9}$$

Cela revient à la détermination d'un optimum minimal de la fonction des erreurs quadratique $Q(a,b)$, qui consiste à résoudre le système des équations suivant

$$\begin{cases} \frac{\partial Q(a,b)}{\partial a} = 0 \\ \frac{\partial Q(a,b)}{\partial b} = 0 \end{cases}, \tag{1.10}$$

c'est-à-dire,

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \end{cases} \Longrightarrow \begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0 \end{cases}. \tag{1.11}$$

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

$$\begin{cases} a(\sum_{i=1}^{n} 1) + b(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} y_i \\ a(\sum_{i=1}^{n} x_i) + b(\sum_{i=1}^{n} x_i^2) = \sum_{i=1}^{n} x_i y_i \end{cases}. \tag{1.12}$$

La résolution du système (1.12), nous fournis la solution suivante

$$\widehat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{XY}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} \quad \text{et} \quad \widehat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{b}\frac{1}{n}\sum_{i=1}^{n} x_i,$$

ou encore

$$\widehat{b} = \frac{s_{xy}}{s_x^2} \text{ et } \widehat{a} = \overline{Y} - \widehat{b}\overline{X},$$

où

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

L'estimation de la fonction de régression s'écrit

$$\widehat{Y} = \widehat{a} + \widehat{b}X$$

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

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

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : 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

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

on peut aussi s'écrire en fonction des résidus

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

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

La statistique du test est la suivante

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

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

Ce chapitre est consacré à l'étude d'une technique statistique appelée l'analyse de la covariance (en abréviation ANCOVA).

Cette 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 qualitatifs avec une ou plusieurs variables quantitatives qui sont reliées à la variable de réponse. Les modèles d'analyse de covariance sont des cas particuliers des modèles d'analyse de régression avec un mélange des variables quantitatives et des variables qualitatives représentées par des variables indicatrices de type 0-1.

2.1 Présentation des données d'une ANCOVA

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 dite covariable et de facteur A à g 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 g 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.

$n = \sum_{i=1}^{g} n_i$ est le nombre d'observations.

2.2 Choix des covariables

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

2.3 Modèle d'analyse de covariance (ANCOVA)

Le modèle d'ANCOVA part d'un modèle d'ANOVA $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ auquel s'ajoute celui de la régression de Y en X : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$. La combinaison des deux modèles conduit à

$$Y_{ij} = \beta_0 + \alpha_i + \beta_1 X_{ij} + \varepsilon_{ij}, \tag{2.1}$$

où β_0 et β_1 sont respectivement l'ordonnée à l'origine et la pente de la droite qui lie Y à X en supposant que la relation soit la même pour g groupes après ajustement sur A. Pour remplacer β_0 par la moyenne générale μ, il suffit de centrer les données de X. Le modèle s'écrit alors

$$Y_{ij} = \mu + \alpha_i + \beta_1 (X_{ij} - \overline{X}) + \varepsilon_{ij}. \tag{2.2}$$

Les conditions d'application de l'ANCOVA sont la normalité, l'équivariance et l'indépendance des résidus.

2.4 Moyennes ajustées

L'ANCOVA compare des moyennes ajustées et non des moyennes arithmétiques observées ($H_0 : \mu_{1aj} = \mu_{2aj} = ... = \mu_{gaj}$). **Ajustées** signifie tenant compte de la relation entre Y et X pour effacer l'effet des différences de moyennes en X. La moyenne ajustée est donc la moyenne à laquelle on s'attend pour le groupe i si les groupes étaient en moyenne parfaitement comparable sur la covariable ($\overline{x}_i = \overline{x}_{i'}$). Admettons que la pente b_1 de la relation entre Y et X soit la même dans tous les groupes comparés, mais que l'ordonnée à l'origine diffère pour au moins deux groupes. L'ANCOVA permet de tenir compte de la non-comparabilité des groupes sur la covariable en apportant une correction proportionnelle à la pente et à l'écart par rapport à la moyenne générale. En effet, si $\overline{x}$ est la moyenne générale de X, $\overline{x}_1$ et $\overline{x}_2$ sont celles des deux groupes, les moyennes ajustées s'écrivent $\overline{y}_{1aj} = \overline{y}_{1/X=\overline{x}} = \overline{y}_1 - b_1(\overline{x}_1 - \overline{x})$ et $\overline{y}_{2aj} = \overline{y}_{2/X=\overline{x}} = \overline{y}_2 - b_1(\overline{x}_2 - \overline{x})$.
Alors la moyenne ajustée du groupe i est égale à

$$\overline{y}_{iaj} = \overline{y}_{i/X=\overline{x}} = \overline{y}_i - b_i(\overline{x}_i - \overline{x}).$$

2.5 Contribution des termes du modèle

Pour tester si un modèle emboîtant explique mieux les données qu'un modèle emboîté, il convient d'utiliser l'approche de l'erreur conditionnelle reposant sur la diminution de l'erreur obtenue par les paramètres supplémentaires du modèle emboîtant. La formule générale s'écrit

$$f = \frac{(SCE_{\text{modèle emboité}} - SCE_{\text{modèle emboitant}})\nu_{\text{modèle emboitant}}}{SCE_{\text{modèle emboitant}}(\nu_{\text{modèle emboité}} - \nu_{\text{modèle emboitant}})}$$

qui suit sous H_0 une loi de Fisher-Snedecor à $v_1 = \nu_{\text{modèle emboité}} - \nu_{\text{modèle emboitant}}$ et $v_2 = \nu_{\text{modèle emboitant}}$ d.d.l($\nu_{\text{ emboité}} > \nu_{\text{ emboitant}}$). on rejetée le profit d'une contribution non nulle

(H_1) si la valeur observée f de la statistique F est plus grande que la valeur critique $f_{\alpha,v_1=1,v_1=n-g}$.

Dans le cas simple d'une ANCOVA à un facteur et une covariable, trois modèles emboîtés peuvent être construits à partir du modèle 2 emboîtant

Modèle 2 : $Y_{ij} = \mu + \alpha_i + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij}$,

Modèle 3 : $Y_{ij} = \mu + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij}$,

Modèle 4 : $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,

Modèle 5 : $Y_{ij} = \mu + \varepsilon_{ij}$.

2.6 Construction du modèle et tests d'intérêt

La méthode présentée ci-dessous s'applique directement pour comparer g droites de régression ou analyser un plan à un facteur et une covariable, ainsi qu'un plan à plusieurs facteurs croisés et une covariable. Dans ce dernier cas, un groupe ne correspond pas à une modalité de A mais à une combinaison de modalités $A_iB_jC_k$.

Les manuels exposent habituellement des approches simplifiées et souvent différentes les unes des autres, car il ne s'agit pas exactement des mêmes hypothèses testées. L'approche retenue s'inspire de celle de **Bruno**. Elle présente la panoplie d'hypothèses pouvant être testées et apportant un élément de réponse aux questions d'intérêt. Les étapes présentées ci-dessous ne correspondent donc pas à une procédure, mais à un ordre logique d'exécution des tests si toutes les hypothèses présentent un intérêt. Rappelons que dans l'approche confirmatoire, le modèle est fixé d'emblée.

Etape1- modèle 0 : $Y_{ij} = \varepsilon_{ij}$ (**DCHO** une seule droite commune horizontale passant par l'origine) et **modèle 1** : $Y_{ij} = \mu + \varepsilon_{ij}$ (**DCH :** une droite commune horizontale).

Il s'agit des modèles les plus simplifiés dans lesquels on considère que seule la moyenne générale décrit adéquatement les données. Ils correspondent à une droite commune horizontale coupant l'axe Y à $\widehat{\mu} = \overline{y}$ pour le modèle 1 (Figure 2.1 **DCH**) et à l'origine pour le

modèle 0. Le carré moyen de l'erreur résiduelle se rapportant au modèle 1 n'est autre que la variance de Y. La dispersion

$$SCE_{\text{modèle 1}} = SCE_{DCH} = (n-1)s_y^2 \qquad (\nu_{e \text{ modèle 1}} = n-1), \tag{2.3}$$

avec $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{Y})^2$, correspond à la dispersion totale SCT. Si les hypothèses

$$\begin{cases} H_{0\,1} : \mu = 0 \\ H_{1\,1} : \mu \neq 0 \end{cases}$$

présentent un intérêt dans le contexte étudié, il suffit d'effectuer le test F de comparaison d'une moyenne à une norme égale à 0 pour le mettre à l'épreuve (lignes 0 et 1 du Tab 2.1). Il suffit de calculer la contribution de μ à l'amélioration de l'explication des données

$$F_{1,(n-1)} = \frac{(SCE_{\text{modèle 0}} - SCT)/1}{SCT/(n-1)}, \tag{2.4}$$

$$SCE_{\text{modèle 0}} = \sum_i \sum_j y_{ij}^2 = (n-1)s_y^2 + n\overline{y}^2 \text{ avec } \nu_{e \text{ modèle 0}} = n. \tag{2.5}$$

Etape2- modèle 2 : $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ (**DH** : g droites horizontales)

Le modèle 2 d'analyse de variance à un facteur peut décrire les données de façon plus précise. Il correspond graphiquement à g droites horizontales coupant l'axe Y à $\overline{Y}_1$et $\overline{Y}_g$ (Figure 2.1 **DH)**. La dispersion (SCA) expliquée par le facteur A est égale à la dispersion intergroupe de l'ANOVA et l'erreur résiduelle est égale à la dispersion intragroupe

$$SCE_{\text{modèle 2}} = SCE_{DH} = \sum_i \sum_j (y_{ij} - \overline{y}_i)^2 \qquad (\nu_{e \text{ modèle 2}} = n-g). \tag{2.6}$$

L'hypothèse d'égalité des moyennes des groupes

$$\begin{cases} H_{0\,2} : \alpha_i = 0 \forall i \Longrightarrow \mu_i = \mu_{i'}, \forall i \text{ et } i' \\ H_{1\,2} : \alpha_i \neq 0 \Longrightarrow \mu_i \neq \mu_{i'}, \text{ pour au moins un } i \text{ et } i' \end{cases}$$

est mise à l'épreuve par le test F (ligne 2 et 3 du Tab 2.1) ou l'approche de l'erreur conditionnelle et donc par la diminution de l'erreur apportée par le modèle 2

$$f = \frac{(SCE_{\text{modèle 1}} - SCE_{\text{modèle 2}})(n-g)}{SCE_{\text{modèle 2}}(g-1)}. \tag{2.7}$$

Ces deux tests de comparaison de moyennes conduisent au même résultat. Si $H_{0\,1}$ est rejetée, la moyenne de Y diffère pour au moins deux groupes, et le modèle 2 s'avère plus approprié que le modèle 1. Si $H_{0\,2}$ est retenue, les moyennes des groupes ne diffèrent pas significativement.

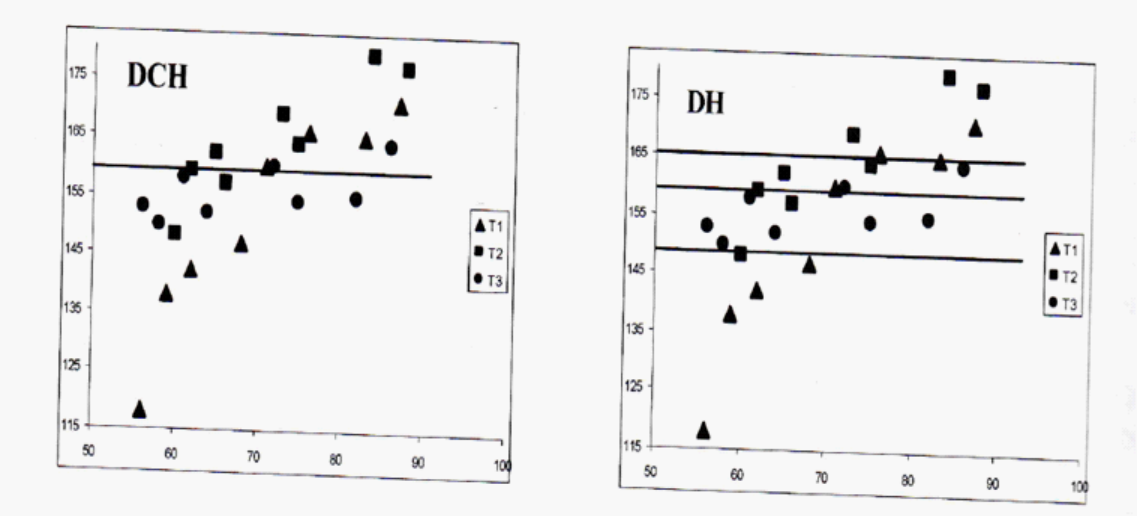

FIG. 2.1 – Représentation des modèles 1 (DCH : droite commune horizontale) et 2 (DH : droites horizontales)(facteur à 3 niveau, réponse Y en ordonnée et covariable X en abscisse).

Etape3- modèle 3 : $X_{ij} = \mu_x + \alpha_{x,i} + \varepsilon_{ij}$ (**DV** : g droites verticales).

Il s'agit du même modèle qu'à l'étape précédente, mais la covariable devient la variable expliquée. Il correspond graphiquement à g droites verticales coupant l'axe des x à $\overline{x}_1, \overline{x}_i$ et $\overline{x}_g$ (Figure 2.2 **DV**).

Le calcul des dispersions inter et intragroupe (SCA_{DV} et SCE_{DV}) et de f (ligne 5 du Tab 2.1) permet de vérifier la comparabilité des groupes sur la covariable

$$\begin{cases} H_{0\,3} : \mu_{x,i} = \mu_{x,i'}, \forall i \text{ et } i' \\ H_{1\,3} : \mu_{x,i} \neq \mu_{x,i'} \text{ pour au moins un } i \text{ et un } i' \end{cases}.$$

Notons que même si les moyennes ne diffèrent pas significativement ($H_{0\,3}$ retenue), l'utilisation de la covariable X reste justifiée par le gain de puissance du test d'égalité des moyennes ajustées si $\rho_{xy} \neq 0$. Le modèle 3 permet d'estimer l'amplitude des écarts entre les moyennes de X.

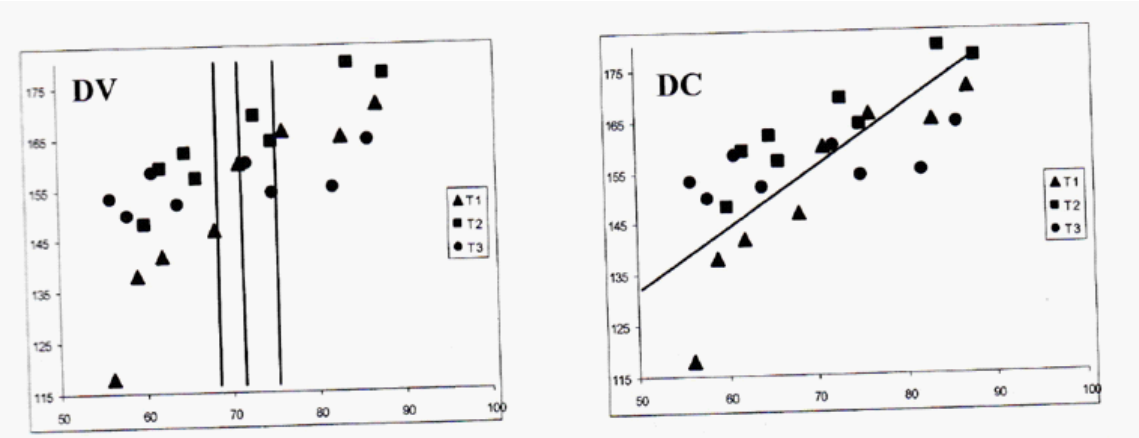

FIG. 2.2 – Représentation des modèles 3 (DV : droites verticales : $X_{ij} = \mu_x + \alpha_{x,i} + \varepsilon_{ij}$) et 4 (DC : droite commune oblique : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$).

Etape4- modèle 4 : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij} = \mu + \beta_1(X_{ij} - \overline{X}) + \varepsilon_{ij}$ (**DC** : 1 droit commune oblique).

Ce modèle suppose que le facteur A n'a aucun effet, mais que X explique Y (Figure 2.2 **DC)**. La dispersion de Y expliquée par X est égale à

$$SCR_{DC} = b_1^2(n-1)s_x^2 \,, \tag{2.8}$$

où $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$ et b_1 représente l'estimation de la pente β_1 de la droite commune . La dispersion SCE_{DC} qui n'est pas expliquée par le modèle 4 s'élève à

$$SCE_{\text{modèle 4}} = SCE_{DC} = SCT - SCR_{DC}. \tag{2.9}$$

Le test de signification de la pente (ligne 6 et 7 du Tab 2.1) met à l'épreuve les hypothèses

$$\begin{cases} H_{0\ 4} : \beta_1 = 0 \\ H_{1\ 4} : \beta_1 \neq 0 \end{cases}.$$

Si $H_{0\ 2} : \alpha_i = 0 \forall i$ est retenue et $H_{0\ 4}$ est rejetée, alors le modèle 4 est plus approprié que les modèles 0, 1 et 2, mais ce n'est pas nécessairement le plus approprié.

Etape 5- modèle 5 : $Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij} = \mu + \alpha_i + \beta_{1i}(X_{ij} - \overline{X}) + \varepsilon_{ij}$ (**DD** : g droites distinctes obliques) et **modèle 6 :** $Y_{ij} = \beta_1 X_{ij} + \varepsilon_{ij}$ (**DCO :** 1 droite commune passant par l'origine).

Ce modèle correspond à g droites de régression distinctes dont les pentes diffèrent, ainsi que les ordonnées à l'origine (Figure 2.3 **DD**). Il s'agit donc d'estimer les paramètres $\widehat{\beta}_i = (b_{0i}\ b_{1i})$ des droites se rapportant à chaque groupe en utilisant les formules (1.13).

L'analyse de variance effectuée sur chaque régression fournit l'erreur attachée à chacune d'elles

$$SCE_{Di} = \sum_j (y_{ij} - \widehat{y}_{ij})^2. \tag{2.10}$$

Chaque ANOVA permet également de mettre à l'épreuve l'hypothèse

$$\begin{cases} H_{0\ 5} : \beta_{1i} = 0 \Longrightarrow \rho_i = 0 \\ H_{1\ 5} : \beta_{1i} \neq 0 \Longrightarrow \rho_i \neq 0 \end{cases}.$$

Pour effectuer le test de signification de la pente β_{1i} (voir lignes 8 et 9 du Tab 2.1), Cette information peut s'avérer utile pour savoir dans quels groupes il existe une relation entre Y_i et la covariable.

Les g droites de régression ne permettant pas d'expliquer toute la variation des données, et la partie non expliquée s'élève à

$$SCE_{\text{modèle } 5} = SCE_{gD} = \sum_i SCE_{Di}. \tag{2.11}$$

Comme les g droites sont distinctes et qu'il faut estimer deux paramètres par droite, $2g$

degrés de liberté sont perdus. On sait que l'objectif général de l'ANCOVA est de comparer les valeurs $\widehat{Y}_{i|X=x_0}$ obtenues sur chaque régression. Il devient maintenant possible de vérifier si la moyenne de Y dépend effectivement de X et donc de mettre à l'épreuve l'hypothèse

$$\begin{cases} H_{0\,6} : \beta_{1i} = 0 \forall i \Longrightarrow E(Y_{ij} \mid X = x_0) = \mu + \alpha_i \\ H_{1\,6} : \beta_{1i} \neq 0 \text{ pour au moins un } i \Longrightarrow E(Y_{ij} \mid X = x_0) = \mu + \alpha_i + \beta_{1i}(x_0 - \overline{X}) \end{cases}.$$

La valeur de la statistique F du test global de signification des pentes apparait à la ligne 10 du Tab 2.1. Notons que $H_{0\,6}$ diffère de $H_{0\,5}$, car au lieu d'effectuer autant de tests que de pentes, un seul test permet de vérifier si au moins une pente diffère de 0. Pour vérifier si au moins deux droites diffèrent dans leur pente ou leur ordonnée à l'origine, on met à l'épreuve les hypothèses

$$\begin{cases} H_{0\,7} : \alpha_i = 0 \text{ et } \beta_{1i} = \beta_1 \forall i \Longrightarrow E(Y_{ij} \mid X = x_0) = \mu + \beta_1(x_0 - \overline{X}) \\ H_{1\,7} : \alpha_i \neq \alpha_{i'} \text{ ou } \beta_{1i} \neq \beta_{1i'} \text{ pour au moins un } i \text{ et un } i' \Longrightarrow E(Y_{ij} \mid X = x_0) = \mu + \alpha_i + \beta_{1i}(x_0 - \overline{X}) \end{cases}.$$

Si $H_{0\,7}$ est vraie, le modèle 5 se ramène au modèle 4. La réduction de l'erreur apportée par le modèle 5 par rapport au 4 s'élève à

$$SCX_{DD} = SCE_{\text{modèle } 4} - SCE_{\text{modèle } 5}. \tag{2.12}$$

Un test F de comparaison des droites de régression (ligne 11 du Tab 2.1) permet de retenir l'une des deux hypothèses et s'il s'agit de $H_{1\,7}$, au moins deux droites de régression diffèrent dans leur équation. Si $H_{1\,6}$ et $H_{1\,7}$ sont retenues, le modèle 5 est plus approprié que le 2 ou le 4 mais pas nécessairement le meilleur.

Si le modèle 4 s'avère approprié et si le modèle 6 : $Y_{ij} = \beta_1 X_{ij} + \varepsilon_{ij}$ s'avère pertinent dans le contexe étudié (droite passant par l'origine : Y directement proportionnelle à X), il convient alors de vérifier s'il est plus approprié que le modèle 4. L'erreur attachée au

modèle 6 est donnée par

$$SCE_{\text{modèle 6}} = SCE_{DCO} = \sum_i \sum_j (y_{ij} - \widehat{y}_{ij})^2. \tag{2.13}$$

Les hypothèses

$$\begin{cases} H_{0\ 8} : \ \beta_0 = 0 \text{ (La droite commune passe par l'origine des axes)} \\ H_{1\ 8} : \beta_0 \neq 0 \end{cases}$$

sont mises à l'épreuve par un test F de signification de l'ordonnée à l'origine (lignes 12 et 13 du Tab 2.1).

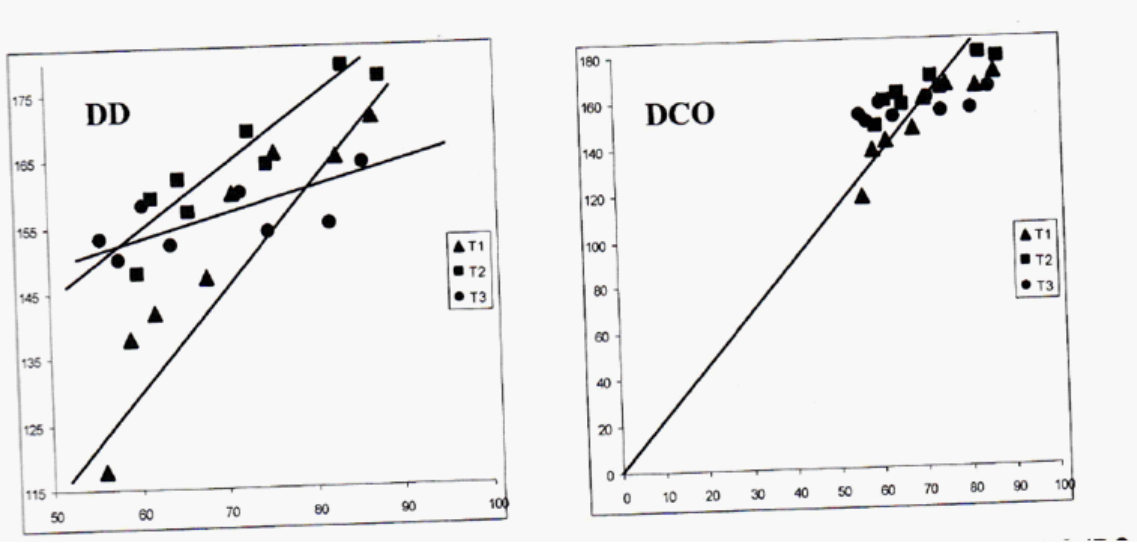

FIG. 2.3 – Représentation des modèles 5 (DD : droites différentes :$Y_{ij} = \beta_{0i} + \beta_{1i}X_{ij} + \varepsilon_{ij}$) et (DCO : droite commune passant par l'origine : $Y_{ij} = \beta_1 X_{ij} + \varepsilon_{ij}$).

Etape6- modèle 7 : $Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \varepsilon_{ij}$ (**DP :** g droites parallèles obliques).

Si l'on suppose que les pentes sont égales (Figure 2.4 **DP**), le modèle 7 nécessite l'estimation de seulement $(g+1)$ paramètres au lieu de $2g$ pour le modèle 5. L'estimation par les

moindres carrés de la pente sous l'hypothèse du parallélisme est fournie par b_{1P}

$$b_{1P} = b_{1DP} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)(y_{ij}-\overline{y}_i)}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2} = \frac{\sum_{i=1}^{g}(n_i-1)s_{xy,i}}{\sum_{i=1}^{g}(n_i-1)s_{x,i}^2} = \frac{\sum_{i=1}^{g} b_{1,i}(n_i-1)s_{x,i}^2}{\sum_{i=1}^{g}(n_i-1)s_{x,i}^2} \tag{2.14}$$

où $s_{x,i}^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2$, $s_{xy,i} = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)(y_{ij}-\overline{y}_i)$.

Notons que cette pente est différente de celle estimée par la droite commune de régression (modèle 4), on sait que $SCR = b_1^2\sum_i(x_i-\overline{x})^2$. Le même principe s'applique avec la pente moyenne intergroupe $b_{1p} = b_{1DP}$. La dispersion expliquée par X dans l'hypothèse du parallélisme est donnée par

$$SCR_{DP} = b_{1DP}^2\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 = b_{1DP}^2 SCE_{DV}. \tag{2.15}$$

La dispersion non expliquée par la modèle 7 est égale à l'erreur résiduelle de l'ANOVA effectuée sur Y (modèle 2) moins la dispersion expliquée par la covariable (SCR_{DP})

$$SCE_{\text{modèle 7}} = SCE_{DP} = SCE_{DH} - SCR_{DP}. \tag{2.16}$$

Pour vérifier si la contribution de X est significative en testant

$$\begin{cases} H_{0\,9}: \ \beta_{1DP} = 0 \\ H_{1\,9}: \ \beta_{1DP} \neq 0 \end{cases}$$

à l'aide d'un test F de signification de la moyenne des pentes (ligne 14 et 15 du tableau). Si $H_{0\,9}$ est retenue alors que $H_{0\,5}: \beta_{1i} = 0$ a été rejetée, le modèle 5 est vraisemblablement plus approprié que le modèle 7, car cette incohérence est probablement liée à une hétérogénéité des pentes avec certaines positives et d'autres négatives.

Etape7- modèle 5 : $Y_{ij} = \beta_{0i}+\beta_{1i}X_{ij}+\varepsilon_{ij} = \mu+\alpha_i+\beta_{1DP}(X_{ij}-\overline{X})+\delta\beta_{1i}(X_{ij}-\overline{X})+\varepsilon_{ij}$ **(DD)**

Le modèle 7 (DP) est plus approprié que le modèle 5 (DD) si le pentes sont égales, c'est-

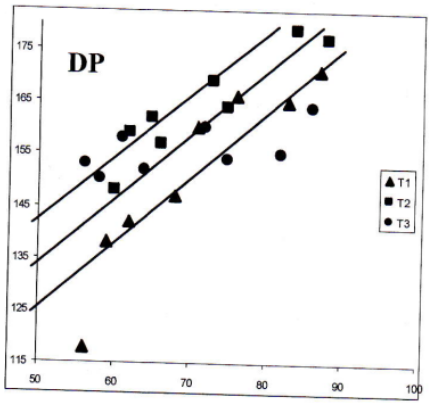

FIG. 2.4 – Représentation du modèle 7 (DP : droites parallèles :$Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \varepsilon_{ij}$).

à-dire en l'absence d'interaction entre la covriable et le traitement. Le modèle 5 peut aussi s'écrire comme ci-dessus où $\delta\beta_{1i} = (\beta_{1i} - \beta_{1DP})$ représente l'interaction entre A et X, à savoir l'écart entre la pente du groupe i et la pente moyenne. La contribution de ce terme est égale à la diminution de l'erreur résiduelle liée à son ajout dans le modèle

$$SCA \times X = SCE_{DP} - SCE_{DD} = SCE_{\text{modèle 7}} - SCE_{\text{modèle 5}}. \tag{2.17}$$

Le test F d'homogénéité des pentes qui lui est associé (voir la ligne 16 du tableau) met à l'épreuve les hypothèses

$$\begin{cases} H_{0\ 10} : \ \delta\beta_{1i} = 0 \ \forall i \\ H_{1\ 10} : \ \delta\beta_{1i} \neq 0 \text{ pour au moins un } i \end{cases}.$$

Si $H_{0\ 10}$ est retenue, le modèle 7 (DP) est plus approprié que le 5 (DD).
Si le modèle 7 (DP) est plus approprié, la contribution de A étant donné X n'est pas égale à la contribution de A (étape 2). La contribution de $A \mid X$ est égale à la diminution des erreurs liée au passage du modèle 4 (X seul) au modèle 7 (X et A)

$$SCA \mid X = SCE_{DC} - SCE_{Dp} = SCE_{\text{modèle 4}} - SCE_{\text{modèle 7}}. \tag{2.18}$$

Le test F de comparaison des moyennes ajustées de la ligne 17 de tableau vérifie si la contribution de $A \mid X$ est significatif ou si au moins deux moyennes ajustées sont significativement différentes ou, ce qui revient au même, si les ordonnées à l'origine des g droites de régression ayant la même pente sont égales. Le test F met donc à l'épreuve les hypothèses

$$\begin{cases} H_{0\,11} : \mu_i \mid (X = \overline{x}) = \mu_{i'} \mid (X = \overline{x}) \text{ ou } \beta_{0i} = \beta_{0i'} \Longrightarrow \mu + \alpha_i = \mu + \alpha_{i'} = \beta_{0i} = \beta_{0i'} \ \forall i \text{ et } i\prime \\ H_{1\,11} : \mu_i \mid (X = \overline{x}) \neq \mu_{i'} \mid (X = \overline{x}) \text{ ou } \beta_{0i} \neq \beta_{0i'} \Longrightarrow \mu + \alpha_i = \beta_{0i} \neq \mu + \alpha_{i'} = \beta_{0i'} \end{cases}$$

Si $H_{1\,11}$ est retenue, alors les moyennes de Y pour une valeur de $X = x_0$ ne sont pas homogènes, et ce quelle que soit x_0 puisque les droites sont parallèles.

Résumé

Les étapes précédemment décrites sont principalement ordonnées en fonction de la séquence des calculs et non de la séquence des tests permettant dans une approche exploratoire de retenir le modèle qui décrit le mieux les données.

Le modèle 7 (**DP**) est habituellement retenu dans le protocole. L'analyse principale vérifie si l'effet du facteur A ajusté à X est significatif. Cette analyse mettant à l'épreuve $H_{0\,11}$ est valide si les droites sont parallèles. L'hypothèse du parallélisme ($H_{0\,10}$) est testée dans la partie diagnostique du modèle, et ce, au même titre que la vérification des autres conditions d'application.

Lignes : Sources de variation	Dispersion et carrés moyens	Nombre de d.d.l.	Représentation graphique et modèle	Test F	Hypothèse testée et valeur critique
0 : Pur bruit de fond (DCHO)	$SCE_{DHCO} = \sum_i \sum_j y_{ij}^2$ $CME_{DHCO} = \sum\sum y_{ij}^2/n = n\overline{y}^2$	n	1 droite horizontale passant par l'origine **Modèle 0 :** $Y_{ij} = \varepsilon_{ij}$		
1 : Bruit de fond autour de $\mu \neq 0$ (DCH)	$SCT = \sum_i \sum_j (y_{ij} - \overline{y})^2$ $CMT = s_Y^2 = SCT/(n-1)$	$n-1$	1 droite horizontale **Modèle 1 :** $Y_{ij} = \mu + \varepsilon_{ij}$	$f = \dfrac{n \times \overline{y}^2}{CMT}$	$\mathbf{H_{0\,1}}$: $\mu = 0$ $\mathbf{H_{1\,1}}$: $\mu \neq 0$ $f_{\alpha,\,1,\,n-1}$
2 : Facteur A : Effet sur Y (non ajusté à X). Bruit de fond autour de $h \geq 2$ moyennes	$SCA = \sum_i n_i (\overline{y}_i - \overline{y})^2$ $CMA = SCA/(g-1)$	$g-1$	g droites horizontales **Modèle 2 :** $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ (ANOVA sur Y)	$f = \dfrac{CMA}{CME_{DH}}$	$\mathbf{H_{0\,2}}$: $\alpha_i = 0 \ \forall \ i$ ou $\mu_i = \mu_{i'} = \mu$ $f_{\alpha,\,g-1,\,n-g}$

(DH)					
3 : Intragroupe *Y* (résiduelle du modèle 2 : DH)	$SCE_{DH} = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ $CME_{DH} = SCE_{DH}/(n-g)$	$n-g$	*g* droites horizontales **Modèle 2** : $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$		
4 : Comparabilité des groupes sur *X* ou effet de A sur *X* (DV)	$SCA_{DV} = \sum_i n_i (\bar{x}_i - \bar{x})^2$ $CMA_{DV} = SCA_{DV}/(g-1)$	$g-1$	*g* droites verticales **Modèle 3 :** $X_{ij} = \mu_X + \alpha_{Xi} + \varepsilon_{ij}$ (ANOVA sur *X*)	$f = \frac{CMA_{DV}}{CME_{DV}}$	$\mathbf{H_{03}}$: $\alpha_{X,i} = 0 \ \forall \ i$ ou $\mu_{X,i} = \mu_{X,i'} = \mu$ $f_{\alpha,\, g-1,\, n-g}$
5 : Intragroupe *X* (résiduelle du modèle 3 : DV)	$SCE_{DV} = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$ $CME_{DV} = SCE_{DV}/(n-g)$	$n-g$	*g* droites verticales **Modèle 3** : $X_{ij} = \mu_X + \alpha_{Xi} + \varepsilon_{ij}$		
6 : Contribution de *X* seule (non ajustée sur A). Corrélation globale entre Y et X (DC)	$SCR_{DC} = b_i^2 (n-1)\, s_X^2$ $CMR_{DC} = SCR_{DC}$	1	1 droite commune **Modèle 4 :** $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$ (Régression de *Y* en *X*)	$f = \frac{CMR_{DC}}{CME_{DC}}$	$\mathbf{H_{04}}$: $E(Y_{ij} \mid x) = \beta_0 \ \forall x \Leftrightarrow \beta_1 = 0$ $\Leftrightarrow \rho_{YX} = 0$ $f_{\alpha,\, 1,\, n-2}$
7 : Erreur résiduelle du modèle 4 : DC). Bruit de fond autour de la droite de régression	$SCE_{DC} = SCT - SCR_{DC}$ $CME_{DC} = SCE_{DC}/(n-2)$	$n-2$	1 droite commune **Modèle 4 :** $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$. $Y_{ij} = \mu + \beta_1 (X_{ij} - \bar{X}) + \varepsilon_{ij}$.		
8 : Effet de *X* et résiduelle dans le groupe *i* . Corrélation (ρ_{YX}) dans le groupe *i* (DG*i*)	$CMR_{Di} = b_{1,i}^2 \sum_j (x_{ij} - \bar{x}_i)^2$ $SCE_{Di} = \sum_j (y_{ij} - \hat{y}_i)^2$ $CME_{Di} = SCE_{Di}/(n_i - 2)$	1 $n_i - 2$	la droite du groupe *i* **Modèle 5 :** $Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij}$ (Régression de Y_i en X_i)	$f = \frac{CMR_{Di}}{CME_{Di}}$	$\mathbf{H_{05}}$: $\beta_{1,i} = 0 \Leftrightarrow \rho_{Yi,Xi} = 0$ $f_{\alpha,\, 1,\, n_i - 2}$
9 : Résiduelle des régressions (gD) (résiduelle du modèle 5). Bruit de fond autour des *g* droites de régression	$SCE_{gD} = \sum_i SCE_{Di}$ $CME_{gD} = SCE_{gD}/(n-2g)$	$n-2g$	*g* droites distinctes **Modèle 5 :** $Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij}$		
10 : Contribution de *X* | A si les pentes β_{1i} sont distinctes (gD)	$SCR_{gD} = SCE_{DH} - SCE_{gD}$ $CMR_{gD} = SCR_{gD}/g$	g	*g* droites distinctes **Modèle 5 :** $Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij}$	$f = \frac{CMR_{gD}}{CME_{gD}}$	$\mathbf{H_{06}}$: $\beta_{1,i} = 0 \ \forall i$ $\mathbf{H_{16}}$: au moins un $\beta_{1i} \neq 0$ $f_{\alpha,\, g,\, n-g}$
11 : Différence en β_0 ou β_1 entre les *g* droites de régression	$SCX_{DD} = SCE_{DC} - SCE_{gD}$ $CMX_{DD} = SCX_{DD}/(2g-2)$	$2g-2$	*g* droites distinctes **Modèle 5** : $Y_{ij} = \mu + \alpha_i + \beta_{1,i}(X_{ij} - \bar{X}) + \varepsilon_{ij}$	$f = \frac{CMX_{DD}}{CME_{gD}}$	$\mathbf{H_{07}}$: $\alpha_i = 0$ et $\beta_{1i} = \beta_1 \ \forall i$ $f_{\alpha,\, 2g-2,\, n-g}$
12 : Résiduelle du modèle 6. Bruit de fond autour de la droite de régression passant par 0 (DCO)	$e_{ij} = y_{ij} - (\sum_i \sum_j x_{ij} y_{ij} / \sum_i \sum_j x_{ij}^2) x_{ij}$ $SCE_{DCO} = \sum_i \sum_j e_{ij}^2$	$n-1$	1 droite commune passant par $X = 0$ **Modèle 6** : $Y_{ij} = \beta_1 X_{ij} + \varepsilon_{ij}$		
13 : Contribution de l'ordonnée à l'origine	$SCOO = SCE_{DCO} - SCE_{DC}$ $CMOO = SCOO$	1	1 droite commune **Modèle 4** : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$	$f = \frac{CMOO}{CME_{DC}}$	$\mathbf{H_{08}}$: $\beta_0 = 0$ $f_{\alpha,\, 1,\, n-2}$

14: Contribution de $X \mid A$ si $\beta_{1i} = \beta_{1DP}\ \forall i$	$SCR_{DP} = b^2_{1DP}\, SCE_{DV}$ $CMR_{DP} = SCR_{DP}$	1	g droites parallèles **Modèle 7 :** $Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \varepsilon_{ij}$	$f = \frac{CMR_{DP}}{CME_{DP}}$	$\mathbf{H_{0\,9}}$: $\beta_{1DP} = 0$ $f_{\alpha,\,1,\,n-g-1}$
15 : Non expliqué par X et A (Résiduelle du modèle 7). Bruit de fond autour de g droites parallèles	$SCE_{DP} = SCE_{DH} - SCR_{DP}$ $CME_{DP} = SCE_{DP} / n-g-1$	$n-g-1$	g droites parallèles **Modèle 7** : $Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \varepsilon_{ij}$		

16 : Interaction entre A et X $(\delta\beta_i = \beta_{1,i} - \beta_{1DP})$ Parallélisme des droites	$SCA{\times}X = SCE_{DP} - SCE_{gD}$ $CMA{\times}X = SCA{\times}X / (g-1)$	$g-1$	g droites différentes **Modèle 5 :** $Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \delta\beta_i(X_{ij} - \overline{X}) + \varepsilon_{ij}$	$f = \frac{CMA{\times}X}{CME_{gD}}$	$\mathbf{H_{0\,11}}$: $\beta_{1i} = \beta_{1DP}$ ou $\delta\beta_i = 0\ \forall i$ égalité des pentes $f_{\alpha,\,g-1,\,n-2g}$
17 : Différence d'ordonnées à l'origine ou effet de $A \mid X$ si $\beta_{1i} = \beta_{1DP}$ (condition : g droites parallèles)	$SCA \mid X = SCE_{DC} - SCE_{DP}$ $CMA \mid X = SCA \mid X / (g-1)$ Indice *OOD* employé au § 18.1.9.3 $SCE_{OOD} = SCA \mid X$ $CME_{OOD} = CMA \mid X$	$g-1$	g droites parallèles **Modèle 7:** $Y_{ij} = \mu + \alpha_i + \beta_{1DP}(X_{ij} - \overline{X}) + \varepsilon_{ij} = \beta_{0i} + \beta_{1DP}X_{ij} + \varepsilon_{ij}$	$f = \frac{CMA \mid X}{CME_{DP}}$	$\mathbf{H_{0\,12}}$: $(\mu_i \mid X = \overline{x}) = \mu$ ou $\alpha_i = 0$ ou $\beta_{0i} = \beta_0\ \forall i$ égalité des moyennes conditionnelles $f_{\alpha,\,g-1,\,n-g-1}$

FIG. 2.5 – Tableau d'analyse de covariance pour un plan à un facteur et une covariable , ou pour g droites de régression .

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 pour simuler la méthode d'analyse de la variance à un facteur.et la méthode d'analyse de la covariance.

3.1 Simulation d'analyse de la variance à un facteur

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).

Exemple 3.1.1 *Plantations d'arbres*

Des forestiers ont réalisé des plantations d'arbres en trois endroits. Plusieurs années plus tard, ils souhaitent savoir si la hauteur moyenne des arbres est identique dans les trois forêts. Chacune des forêts constitue une population et dans chacune d'entre elles, un échantillon d'arbre est tiré au sort. Puis la hauteur de chaque arbre est mesurée en mètres. L'objectif ici est de comparer les moyennes théoriques des hauteurs des arbres dans les trois forêts.

Arbre	**Forêt**	**Hauteur** (en m)	**Arbre**	**Forêt**	**Hauteur** (en m)
1	Forêt 1	23.4	4	Forêt 2	22.1
2	Forêt 1	24.4	5	Forêt 2	22.5
3	Forêt 1	24.6	6	Forêt 2	23.5
4	Forêt 1	24.9	1	Forêt 3	22.5
5	Forêt 1	25.0	2	Forêt 3	22.9
6	Forêt 1	26.2	3	Forêt 3	23.7
1	Forêt 2	18.9	4	Forêt 3	24.0
2	Forêt 2	21.1	5	Forêt 3	24.0
3	Forêt 2	21.1	6	Forêt 3	24.5

TAB. 3.1 – Hauteurs des arbres selon le type de forêt.

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

```
> X<-data.frame(forêt1=c(23.4,24.4,24.6,24.9,25.0,26.2),forêt2=c(18.9,21.1,21.1,22.1
22.5,23.5),forêt3=c(22.5,22.9,23.7,24.0,24.0,24.5))
> hauteur<-stack(X)$values
> forêt<-stack(X)$ind
> tapply(hauteur,forêt,summary)
$forêt1
   Min. 1stQu.  Median    Mean 3rdQu.    Max.
  23.40   24.45   24.75   24.75   24.98   26.20
$forêt2
   Min. 1stQu.  Median    Mean 3rdQu.    Max.
  18.90   21.10   21.60   21.53   22.40   23.50
$forêt3
   Min. 1stQu.  Median    Mean 3rdQu.    Max.
  22.50   23.10   23.85   23.60   24.00   24.50
```

Le test porte sur la comparaison des espérances. À cette étape, il serait bon de tracer les boîtes à moustaches de la variable hauteur en fonction de la variable forêt. Pour cela, tapez la ligne de commande suivant

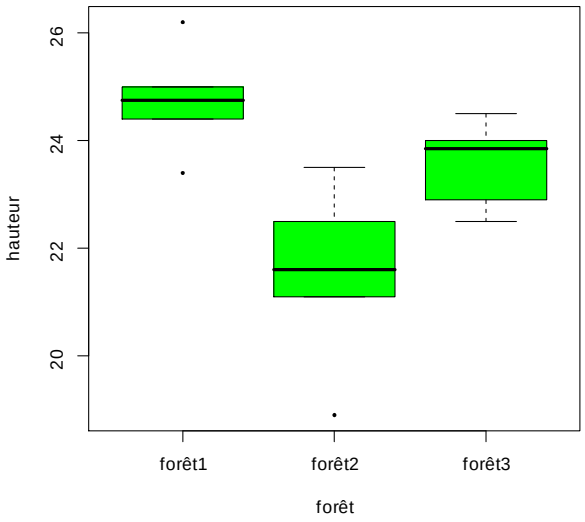

FIG. 3.1 – Les boîtes à moustaches de la variable hauteur en fonction de la variable forêt.

```
> plot(hauteur~forêt,pch=16,cex=0.5,col="green")
```

Le résultat est le graphique suivant

hauteur<-stack(X)$values

forêt<-stack(X)$ind

tapply(hauteur,forêt,summary)

plot(hauteur~forêt,pch=16,cex=0.5,col="green")

model1.aov<-aov(hauteur~forêt,data=arbre)

summary(model1.aov)

• **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(hauteur~forêt)
> summary(model1.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
forêt	2	31.88	15.94	12.31	0.000687	***
Residuals	15	19.43	1.295			

– – –

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(hauteur~forêt,data=arbre)
> 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 $p = 0.0006867$ nous permet de conclure que les hauteurs des arbres d'au moins deux forêts sont différents au risque 5%.

3.1.1 Vérification des conditions fondamentales d'ANOVA

Condition d'indépendance

Il n'existe pas, dans un contexte général, de test statistique simple permettant d'étudier l'indépendance. Ce sont les conditions de l'expérience qui vous permettront d'affirmer que cette condition est remplie.

Condition de normalité

Nous désirons tester la normalité des variables d'erreur ε_{ij} avec le test de **Shapiro-Wilk.** Nous utilisons l'échantillon formé par les résidus pour réaliser ce test. Pour cela, en tapant les deux lignes de commande suivantes

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

Shapiro-Wilk normality test

data : residus

W = 0.9748, p-value = 0.882

Nous utiliserons la p-valeur donnée par R pour conclure. Des études ont montré que, lorsque l'effectif de l'échantillon est supérieur ou égale à 30, la puissance du test de Shapiro-Wilk est acceptable.

La p-valeur étant strictement supérieur à 5% , alors l'hypothèse de normalité des variables d'erreur est statisfaite.

Condition d'homogénéité des variances

Plusieurs tests permettent de tester l'égalité de plusieurs variances. Parmi ceux-ci, le test le plus utilisé est le test de **Bartlett** (ce test nécessite la normalité et l'indépendance des variables dont les variances sont comparées) est obtenu par

```
> bartlett.test(residus~forêt,data=arbre)
```

Bartlett test of homogeneity of variances

data : residus by forêt

Bartlett's K-squared = 2.8279, df = 2, p-value = 0.2432

La p-valeur est supérieur à 0.05, on peut conclure qu'il existe une homogénéité des variances.

3.2 Simulation d'analyse de covariance ANCOVA

L'ANCOVA est également un cas particulier de modèle linéaire, avec cette fois une variable qualitative et une quantitative. Elle s'effectue donc de manière similaire à l'ANOVA et la régression linéaire.

Exemple 3.2.1 *Cancer du sein*

On étudie la durée de survie Y de femmes atteintes de cancer du sein soumises à trois traitements A, B et C. Ces durées figurent dans le tableau suivant dans les colonnes *Survie*, on a aussi indiqué l' *âge* X d'apparition d'un cancer.

Traitement A		Traitement B		Traitement C	
Âge	Survie	Âge	Survie	Âge	Survie
32.7	6.5	33.3	8.5	30.3	11.9
37.2	8.8	40.4	5.6	31.7	9.1
37.3	10.0	41.6	9.1	31.9	7.9
39.8	8.7	43.4	7.4	33.9	9.9
42.6	8.4	44.5	4.1	36.2	8.7
44.2	4.1	46.5	5.9	39.9	9.8
45.4	6.1	47.8	7.7	41.4	9.5
47.0	5.6	47.9	6.4	42.6	7.6
47.4	3.7	49.2	5.8	43.3	7.7
47.6	8.0	52.3	6.3	43.6	5.2
49.3	6.4	52.8	5.7	43.6	8.5
50.2	5.2	52.8	3.3	44.1	7.4
50.4	7.4	53.0	2.7	44.5	5.1
51.4	4.0	55.2	4.0	45.9	5.7
51.8	7.0	56.1	3.2	46.5	7.3
52.0	6.8	56.4	4.3	48.8	4.6
53.5	4.6	56.5	3.8	49.0	6.8
53.6	4.7	56.6	1.5	49.2	5.8
55.8	4.7			50.4	8.6
56.4	4.7			50.7	5.1
58.7	4.3			52.7	6.5
59.4	3.8				
63.3	2.1				

TAB. 3.2 – Durée de survie de femmes atteintes de cancer de sein en fonction d'âge d'apparition d'un cancer et le traitement suivi.

Dans la présente application, l'objectif est de savoir si le facteur traitement influe sur la durée de survie indépendamment de l'âge d'apparition du cancer. Avant de répondre à ce problème, on veut répondre aux questions suivantes

1- Y a-t-il un effet de traitement sur la durée de survie sans tenir compte de l'âge d'apparition de cancer ?

2- Est-ce-que l'âge d'apparition de cancer est différent selon le traitement ?

3- Comment représenter graphiquement la durée de survie en fonction de l'âge pour les différents niveaux du traitement ?

4- Y a t'il une interaction entre la méthode de traitement et le rapport entre la durée de survie et l'âge d'apparition de cancer ?

Réponses

Les données sont entrées dans R au moyen des instructions suivantes

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

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

On calcule la durée moyenne de survie et l'âge moyen d'apparition du cancer dans chaque groupe

```
> tapply(Survie,Traitement,mean)
```

```
A        B        C
5.895652 5.300000 7.557143
```

```
> tapply(CancerSein$Age,CancerSein$Traitement,mean)
```

```
A        B        C
49.00000 49.23889 42.86667
```

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

```
> par(mfrow=c(1,2))
> plot(Survie~Traitement,data=CancerSein,col="green",main="a")
> plot(Age~Traitement,data=CancerSein,col="green",main="b")
```

A partir de ces résultats préliminaires, on remarque que les moyennes de la durée de survie dans les deux groupes A et B sont presque similères, et elles sont plus petites

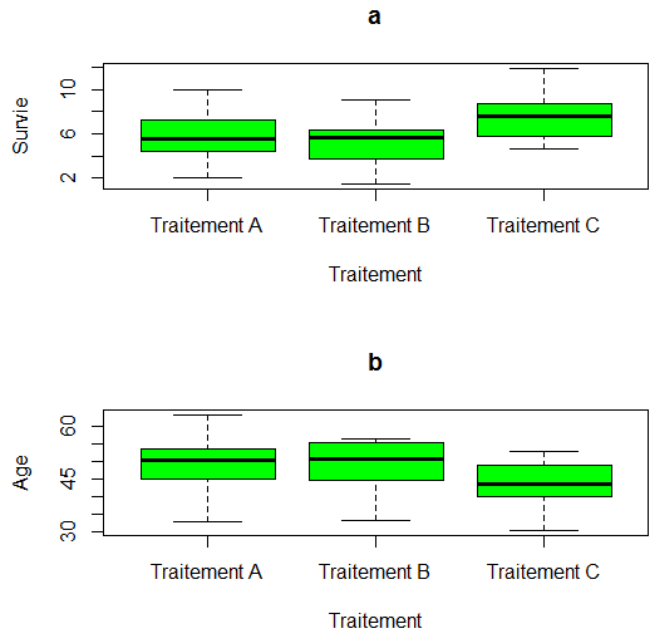

FIG. 3.2 – Boites à moustaches de la durée de survie (a) et de l'âge d'apparition du cancer (b) pour chaque traitement.

que la moyenne de la durée se survie dans le groupe C. Tandis que, les moyennes d'âge d'apparition du cancer dans les deux groupes A et B sont aussi similaires, mais elles sont plus grandes que celle du groupe C.

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

```
> aov1<-aov(Survie~Traitement,data=CancerSein)
summary(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Traitement	2	54.936	27.4680	6.8921	0.002042	**
Residuals	59	235.141	3.9854			

– – –

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

Alors, la p-valeur étant strictement inférieur à 0.05, il y a un effet de traitement sur la durée de survie sans tenir compte de l'âge d'apparition de cancer.

```
> aov2<-aov(Age~Traitement,data=CancerSein)
```

summary(aov2)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Traitement	2	541.0	270.498	5.4475	0.006745	**
Residuals	59	2929.7	49.655			

– – –

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

Donc, les moyennes de la variable âge d'apparition de cancer sont différentes selon le traitement au risque 5%.

Pour continuer notre étude, on suppose qu'il n'y a pas une différence de l'âge moyen d'apparition de cancer entre les niveaux du traitement.

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

```
> plot(Survie~Age,data=CancerSein,type="n")
> points(Survie~Age,data=subset(CancerSein,Traitement=="A"),col="red",pch="A")
> points(Survie~Age,data=subset(CancerSein,Traitement=="B"),col="blue",pch="B")
> points(Survie~Age,data=subset(CancerSein,Traitement=="C"),col="green",pch="C")
> legend(50,12,c("Traitement A","Traitement B","Traitement C"),pch="ABC",
col=c("red","blue","green"),cex=1)
```

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

```
> mod1<-lm(Survie~Age+Traitement+Age :Traitement,data=CancerSein)
> anova(mod1)
```

Analysis of Variance Table

Response : Survie

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Age	1	169.242	169.242	84.7121	$8.458e-13$	***
Traitement	2	7.913	3.956	1.9803	0.1476	

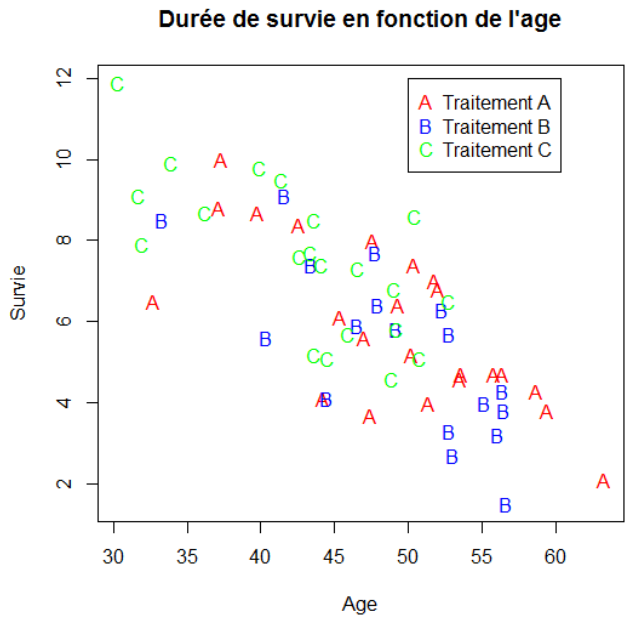

FIG. 3.3 – Nuage de points de la durée de survie en fonction de l'âge.

Age : Traitement	2	1.042	0.521	0.2609	0.7713
Residuals	56	111.880	1.998		

– – –

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

De cette table d'ANOVA, on constate qu'il n'ya pas une interaction entre Age et Traitement car la p valeur est supérieur à 0.05 (p valeur=0.7713), 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 des résidus pour le modèle linéaire (mod1) à l'aide de l'instruction* ***shapiro.test(residuals(mod1))***.

L'application d'ANCOVA, nous permet de répondre à notre objectif (savoir si le facteur traitement influe sur la durée de survie indépendamment de l'âge d'apparition du cancer). Sous R, on l'exécute de la manière

```
> (mod2<-lm(Survie~Age+Traitement,data=CancerSein))
> anova(mod2)
```

Analysis of Variance Table

Response : Survie

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Age	2	169.242	169.242	86.9277	$3.896e-13$	$***$
Traitement	1	7.913	3.956	2.0321	0.1403	
Residuals	58	112.922	1.947			

$---$

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'ANOVA, on remarque que la p valeur associée au facteur traitement est supérieur à 0.05, ce qui veut dire que le facteur traitement n'influe pas sur la durée de survie. Ce résultat montre que si en tenant compte de l'âge d'apparition d'un cancer, les durées de survies moyennes (moyennes ajustées) sont égaux pour les trois traitements. Aussi, on peut dire que la modélisation des données sera par une seule droite.

Cacul des moyennes ajustées

Grâce aux instructions suivantes, on obtient les moyennes ajustées : meanAaj, meanBaj et meanCaj pour les traitements A, B et C respectivement

```
coe2 <- coefficients(mod2)
meanAaj=mean(Survie[Traitement=="A"])-((coe2[2])*
(mean(Age[Traitement=="A"])-mean(Age)))
meanBaj=mean(Survie[Traitement=="B"])-((coe2[2])*
(mean(Age[Traitement=="B"])-mean(Age)))
meanCaj=mean(Survie[Traitement=="C"])-((coe2[2])*
(mean(Age[Traitement=="C"])-mean(Age)))
```

d'où

meanAaj=6.305798, meanBaj=5.758939 et meanCaj=6.714558.

Notons qu'après l'ajustement, les moyennes les plus basses de la durée de survie, c'est-à-dire 5.89 et 5.30 ont été augmentées à 6.30 et 5.76 respectivement, et la moyenne la plus élevée c'est-à-dire 7.66 a été diminuée à 6.71.

Ceci dù au fait que la différence entre les groupes au niveau de l'âge d'apparition d'un cancer a été éliminée.

Le test de signification avait indiqué que ces moyennes ajustées ne sont pas significativement différentes, et que leur différence peut être atttribuable au hazard de l'échantillonnage.

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

En réalisant cet ajout, nous cherchons à réduire la variance du terme d'erreur ϵ présent dans le modèle et rendre ainsi l'analyse plus précise.

Le but générale de l'ANCOVA est soit

- d'étudier l'effet de facteurs explicatifs en tenant compte de l'effet de facteurs quantitatifs, ou d'une autre façon de comparer des moyennes (via un facteur de classification) tout en tenant compte d'une variable auxiliaire quantitative (covariable). Le but est ici d'éliminer l'influence de la covariable, qui peut être une variable différente de la variable à expliquer ou une mesure antérieure (ex : en début d'expérience) de la variable à expliquer.

- de comparer plusieurs droites de régression en vérifiant leur obliquité, parallélisme et position. D'un point de vue mathématique, les modèles d'analyse de la covariance sont en fait simplement un type particulier de modèle de régression linéaire.

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. Analyse de covariance Tp n° 9, Magistère 2ème année-2008/2009.

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

[3] Francour, P. (2015). Analyse de Variance à un ou Plusieurs Facteurs, Regressions, Analyse de Covariance, Modèles Linéaires Généralisés. francour@unice.fr.

[4] Giorgio, R. (2011).Analyse de la covariance, Analyse de donnée-méthodes explicatives (STA102). Département IMATH CNAM, giorgio.russilillo@cnam.fr.

[5] Marie Chavent. Régression linéaire Simple, Chapitre 1 Licence 3 MIASHS-Université de Bordeaux.

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

[7] Michel, Carbon. (2015). Cours d'Analyse de la variance. Département de Mathématiques et Statistique, Université de Laval.

[8] Scherrer, B. (2007). Biostatistique, vol. 1. Gaëtan Morin éditeur (816 pp.).

[9] Scherrer, B. (2009). Biostatistique, volume 2, chapitre 25. 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 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 moyens associés au facteur.
CME	Carrés moyens résiduels.
SMR	Somme carré de régression.
$f_{(n_1,n_2)}$	Une loi de Fisher de degrés de liberté $n_1,\ n_2$.