

AgroParisTech

Bases du Modèle Linéaire

J.J. DAUDIN, E. LEBARBIER, C. VUILLET

Table des matières

1	Introduction	3
2	Estimation des paramètres	5
2.1	Estimation des paramètres de l'espérance	5
2.1.1	Méthode des moindres carrés	5
2.1.2	Cas où $r = p + 1$	9
2.1.3	Cas où $r < p + 1$: résolution par inverse généralisé	9
2.1.4	Combinaisons linéaires invariantes	12
2.2	Propriétés des estimateurs de θ	13
2.3	Loi des estimateurs	14
2.3.1	Loi de $\hat{\theta}$	14
2.3.2	Combinaisons linéaires des θ_k	14
2.4	Estimation de la variance	14
2.4.1	Résidus	14
2.4.2	Estimation de σ^2	15
2.4.3	Loi de σ^2	15
2.4.4	Décomposition de la variance et coefficient de détermination	15
2.5	Intervalle de confiance	17
3	Tests d'hypothèses	19
3.1	Test sur un paramètre	19
3.2	Test d'une combinaison linéaire des paramètres	19
3.3	Test de modèles emboîtés	20
3.3.1	Modèles emboîtés	20
3.3.2	Test de modèles emboîtés	20
3.3.3	Vision géométrique du test	21
4	Notion d'orthogonalité	23
4.1	Définition de l'orthogonalité	23
4.1.1	Cas de 2 facteurs	23
4.1.2	Cas général : Orthogonalité d'un plan d'expériences vis à vis d'un modèle	24
4.1.3	Cas de la Régression multiple et de l'analyse de la covariance	26
4.2	Propriétés d'un dispositif orthogonal	26
4.3	Sommes de carrés ajustées : Type I, II et III	27
4.3.1	Réduction	27
4.3.2	Sommes de carrés	27

4.4	Moyennes ajustées	28
5	Limites et extensions du modèle linéaire	30
5.1	Analyse des résidus	30
5.2	Généralisations du modèle linéaire	31
5.2.1	Modèles non linéaires	31
5.2.2	Modèle linéaire général	31
5.2.3	Modèle linéaire généralisé	31
5.3	Les variables explicatives sont aléatoires	32
5.3.1	Cas de la régression	32
5.3.2	Modèles mixtes	32
6	Annexes	33
6.1	Loi de $(\hat{\theta}_1 - a)/\sqrt{\widehat{V}(\hat{\theta}_1)}$	33
6.1.1	Loi sous H_0	33
6.1.2	Loi sous H_1	33
6.2	Loi de $\frac{(SCM_1 - SCM_0)/(p_1 - p_0)}{SCR_1/\nu_1}$	33
6.2.1	Loi sous H_0	33
6.2.2	Loi sous H_1	34
6.3	Théorème de Cochran	34
6.3.1	Lemme 1	34
6.3.2	Lemme 2	34
6.3.3	Démonstration du théorème de Cochran	35
6.4	Espérance et matrice de variance-covariance de $\hat{\theta}$	35
6.4.1	Espérance de $\hat{\theta}$	35
6.4.2	Matrice de variance-covariance de $\hat{\theta}$	36
6.5	Espérance de la somme des carrés résiduelle	37
6.6	Démonstration du théorème de Gauss-Markov	37

Chapitre 1

Introduction

Le modèle linéaire¹ est fondamental pour l'analyse des résultats numériques d'expériences basées sur du matériel expérimental variable. Il suppose des hypothèses mathématiques assez contraignantes, mais il a deux qualités majeures :

- il est capable de bien représenter la plupart des expériences,
- sa simplicité permet de bien connaître ses propriétés.

De ce fait son emploi est très fréquent et il joue un rôle pivot dans toute la modélisation statistique. La façon dont on traite les modèles plus généraux en découle directement. Il est donc essentiel de bien comprendre ce modèle fondamental de la statistique.

Le modèle statistique de base que l'on utilise pour analyser une expérience où l'on étudie sur n unités expérimentales les variations d'une *variable réponse* y en fonction de facteurs qualitatifs ou quantitatifs (appelés aussi *variables explicatives*), peut s'écrire :

$$y_i = m_i + e_i$$

où

- i est le numéro de l'unité expérimentale,

$m_i = f(x_{i1}, x_{i2}, \dots, x_{ip+1})$ est l'espérance de y_i qui dépend donc de la valeur des variables explicatives x_1, x_2, \dots, x_{p+1} . f est une combinaison linéaire de ces variables explicatives.

- e_i est une variable aléatoire résiduelle, appelée erreur, incluant la variabilité du matériel expérimental, celle due aux variables explicatives non incluses dans le modèle, et celle due aux erreurs de mesure.

L'écriture matricielle du modèle est :

$$Y = X \theta + E, \tag{1.1}$$

où

- $Y_{(n,1)}$ contient les valeurs de y pour les n expériences. L'ordre de rangement des y_i est arbitraire. C'est un vecteur aléatoire.
- $E_{(n,1)}$ contient les variables aléatoires résiduelles du modèle, rangées dans le même ordre que Y . Les E_i sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

1. Ce polycopié est fortement inspiré de celui de C. Duby [3] dont nous avons repris le plan et certaines parties.

- $\theta_{(p+1,1)}$ contient $p + 1$ paramètres fixes et inconnus.
- $X_{(n,p+1)}$ est une matrice (fixe et connue) contenant les valeurs des variables explicatives, rangées en ligne dans le même ordre que Y . Dans le cas de facteurs qualitatifs ces valeurs sont des 1 ou des 0. X s'appelle "matrice du plan d'expérience". Le polycopié [6] contient de nombreux exemples de modèles linéaires et des matrices X associées.

Attention : le modèle linéaire n'est pas aussi réducteur qu'on peut le penser. Par exemple $Y = \theta_0 + \theta_1 x + \theta_2 x^2 + E$ est un modèle linéaire. "Modèle linéaire" signifie que $\mathbb{E}(Y)$ est une combinaison linéaire de variables qui peuvent être le résultat de transformations non linéaires séparées.

Chapitre 2

Estimation des paramètres

Une fois le modèle posé, la question qui se pose ensuite est l'estimation des paramètres inconnus du modèle. Les paramètres sont de deux sortes, ceux qui relèvent de l'espérance et sont contenus dans le vecteur θ , et le paramètre σ^2 qui mesure la variabilité qui subsiste lorsque l'on a enlevé à la variabilité totale des observations tout ce qui est expliqué par le modèle. Il met donc en jeu $p + 1$ paramètres pour l'espérance (p pour les p variables et 1 pour la constante) et 1 paramètre pour la variance (σ^2) (cf chapitre 1).

2.1 Estimation des paramètres de l'espérance

Pour estimer θ , nous utilisons la méthode des moindres carrés ordinaires (M.C.O.).

2.1.1 Méthode des moindres carrés

Avant de présenter la méthode dans le cadre général, considérons l'exemple simple de la régression linéaire simple.

Cas de la régression linéaire simple

Rappelons que le modèle de régression linéaire simple s'écrit

$$Y_i = a + b x_i + E_i$$

où les erreurs $\{E_i\}$ sont supposées indépendantes et de même distribution, une gaussienne centrée et de variance σ^2 . Les paramètres de l'espérance sont donc a et b . Ce modèle est représenté Figure 2.1.

L'idée naturelle pour choisir les meilleures valeurs de ces deux paramètres est de déterminer parmi l'ensemble des droites du plan, celle qui est la plus proche de tous les points observés, au sens d'une distance à préciser. On cherche donc à ce que l'écart entre les valeurs observées et les valeurs données par la droite (c'est-à-dire $a + b x_i$) soit la plus petite possible et pour tous les points en même temps. Ainsi, la méthode des moindres carrés ordinaires consiste à chercher les valeurs des deux constantes a et b qui minimisent la somme suivante :

$$S(a, b) = \sum_{i=1}^n (Y_i - (a + b x_i))^2 = \sum_{i=1}^n E_i^2.$$

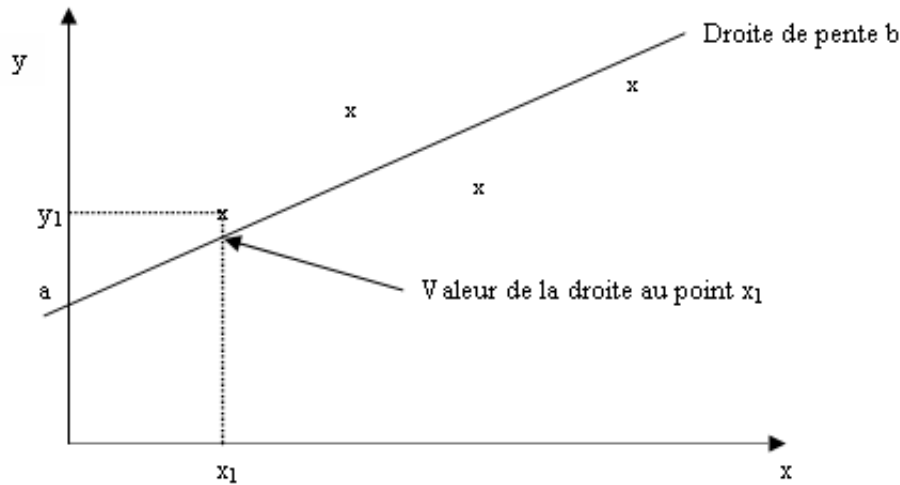


FIGURE 2.1 – Représentation graphique du modèle de régression linéaire simple.

Si on note x , le vecteur colonne contenant les valeurs x_i , cette somme s'écrit aussi

$$S(a, b) = \|Y - (a + b x)\|^2,$$

où $\|\cdot\|^2$ correspond à la norme euclidienne de \mathbb{R}^n : si u est un vecteur de \mathbb{R}^n , $u = (u_1, \dots, u_n)$, la norme de u vaut $\|u\|^2 = \sum_{i=1}^n u_i^2$. D'un point de vue géométrique, cette norme s'interprète comme la longueur du vecteur \vec{ou} qui sera par la suite confondu abusivement avec u . Ainsi S s'interprète comme le carré de la distance entre le vecteur Y et $a + b x$. Le choix de cette distance est dû au fait qu'elle est la plus simple à manipuler.

Pour estimer a et b , on cherche donc à minimiser $S(a, b)$. Pour obtenir ce minimum, il faut résoudre le système suivant :

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

La solution bien connue (cf [2]) fournit les estimateurs suivants :

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

où $\bar{x} = \sum_{i=1}^n x_i / n$ est la moyenne des x_i . Les estimations associées de ces deux paramètres sont notées \hat{a} et \hat{b} et correspondent aux valeurs des estimateurs calculés sur les observations y . Attention, dans toute la suite nous ne ferons pas toujours la distinction entre les estimations et les estimateurs.

Une fois les estimations \hat{a} et \hat{b} calculées, on peut calculer :

- les prédictions par le modèle de $\mathbb{E}(Y)$ au point x_i :

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

- les estimations des erreurs :

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

Remarque. Puisque que l'on a supposé que les erreurs E_i étaient gaussiennes i.i.d., la méthode des moindres carrés revient exactement à la méthode du maximum de vraisemblance. Elle consiste à chercher les valeurs des paramètres qui maximisent la vraisemblance des observations. La vraisemblance de $Y = \{Y_1, \dots, Y_n\}$ s'écrit

$$L(Y_1, \dots, Y_n; a, b, \sigma^2) = \prod_{t=1}^n \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - a - b x_i)\right)^2 \right].$$

On a donc :

$$\log L = -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{t=1}^n (Y_i - a - b x_i)^2.$$

Ainsi

$$\begin{aligned} & \text{maximiser } \log L \text{ en } a \text{ et } b \\ & \Leftrightarrow \\ & \text{minimiser en } a \text{ et } b S(a, b) = \sum_{t=1}^n (Y_i - a - b x_i)^2 \end{aligned}$$

Cas général

Définition 1 L'estimateur des moindres carrés des paramètres θ du modèle 1.1 est noté $\hat{\theta}$ et est tel que

$$S = \|Y - X\theta\|^2 \quad \text{est minimale.}$$

Cet estimateur vérifie le résultat énoncé dans le théorème suivant.

Théorème 1 L'estimateur des moindres carrés des paramètres θ du modèle vérifie le système

$$X'X\hat{\theta} = X'Y. \quad (2.1)$$

Ce système s'appelle traditionnellement **système des équations normales**.

Dans la suite, X' correspond à la transposée de la matrice X .

Preuve

Notons $L(X)$ le sous-espace linéaire de \mathbb{R}^n engendré par les vecteurs colonnes de la matrice X notés $\{X_0, X_1, \dots, X_p\}$, c'est-à-dire que tout élément de $L(X)$ s'écrit comme une combinaison linéaire de ces vecteurs : pour tout $U \in L(X)$, il existe η un vecteur colonne de dimension $p+1$ tel que $U = X\eta$. On cherche donc un élément de $L(X)$, noté \hat{Y} , qui vaut

$$\hat{Y} = X\hat{\theta}, \quad (2.2)$$

et tel que le vecteur \hat{Y} réalise

$$\|Y - \hat{Y}\|^2 = \min_{U \in L(X)} \|Y - U\|^2. \quad (2.3)$$

Avant de donner la solution à ce problème, nous redonnons les définitions et propriétés de la notion de projecteur et projecteur orthogonal.

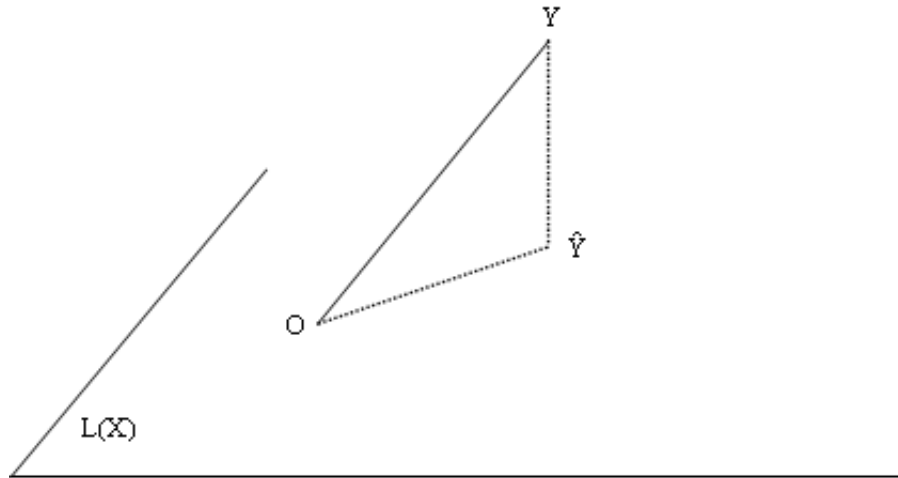


FIGURE 2.2 – Représentation de la projection de Y sur $L(X)$.

Définition 2 On appelle projecteur (noté P), une application linéaire de \mathbb{R}^n dans \mathbb{R}^n qui vérifie $P^2 = P$.

Définition 3 Un projecteur P est orthogonal si dans toute base orthonormée sa matrice est symétrique ($P' = P$).

On note P aussi bien le projecteur que la matrice associée. Un projecteur orthogonal vérifie alors la propriété suivante.

Propriété 1 Si P est un projecteur orthogonal alors pour tout X de \mathbb{R}^n ,

$$\langle P(X), (I_n - P)(X) \rangle = 0$$

où $\langle \cdot, \cdot \rangle$ est le produit scalaire dans \mathbb{R}^n (si u et v sont deux vecteurs de \mathbb{R}^n , alors $\langle u, v \rangle = \sum_{i=1}^n u_i v_i = u'v$)

La solution au problème de minimisation qui nous intéresse (2.3) est bien connue, il s'agit de la projection orthogonale du vecteur Y sur l'espace $L(X)$. En notant P la matrice de projection, on a :

$$\begin{aligned} \widehat{Y} &= \text{projeté orthogonal de } Y \text{ sur } L(X) \\ &= PY \end{aligned}$$

Cette projection est illustré Figure 2.2. Nous verrons un peu plus loin comment calculer la matrice P .

D'après la propriété donnée ci-dessus d'un projecteur orthogonal, on peut écrire que le vecteur $Y - \widehat{Y}$ est orthogonal à l'espace $L(X)$, en particulier à tous les vecteurs de la matrice X . On peut donc écrire que tous les produits scalaires entre $Y - \widehat{Y}$ et les X_k sont nuls :

$$\forall k = 1, \dots, p+1 \quad \langle X_k, Y - \widehat{Y} \rangle = X_k'(Y - \widehat{Y}) = 0$$

ce qui s'écrit plus synthétiquement :

$$X'(Y - \widehat{Y}) = 0$$

où X' correspond à la transposée de la matrice X . En reprenant la définition de \widehat{Y} (2.2), on obtient

$$X'(Y - \widehat{Y}) = X'(Y - X\widehat{\theta}) = 0$$

Ce qui conclut la preuve du théorème.

On ne peut résoudre le système des équations normales (2.1) que si la matrice carrée $X'X$ de dimension $(p+1) \times (p+1)$ admet une inverse $(X'X)^{-1}$. Or cette dernière n'existe que si la matrice $X'X$ est de plein rang, c'est-à-dire que son rang vaut exactement $p+1$. Noter que le rang de la matrice $X'X$ est le même que celui de X et rappelons qu'il est égal au nombre de colonnes de la matrices X qui sont linéairement indépendantes (qui ne sont pas combinaisons linéaires des autres colonnes). Notons

$$r = \text{rang de } X$$

Cela revient à dire que le système (2.1) est un système linéaire de r équations à p inconnues. On voit donc que si r est différent de p , on n'aura pas assez d'équations pour estimer tous les paramètres et donc qu'il existera une infinité de solutions.

2.1.2 Cas où $r = p + 1$

C'est le cas des modèles de régression linéaire simple, de régression polynomiale et de régression linéaire multiple à condition bien sûr que les variables explicatives ne soient pas liées linéairement. Dans ce cas, la matrice $X'X$ est inversible et la solution du système des équations normales est unique. Son expression est :

$$\widehat{\theta} = (X'X)^{-1}X'Y \tag{2.4}$$

On a donc que

$$\widehat{Y} = X\widehat{\theta} = X(X'X)^{-1}X'Y$$

Grâce à cette dernière égalité, on peut retrouver la matrice du projecteur orthogonal P :

$$P = X(X'X)^{-1}X'$$

On peut vérifier qu'il a bien toutes les propriétés d'un projecteur orthogonal, en particulier que $P^2 = P$ et qu'il est symétrique ($P' = P$).

2.1.3 Cas où $r < p + 1$: résolution par inverse généralisé

Ce cas correspond aux situations où les vecteurs colonnes de la matrice X sont liés par $p + 1 - r$ relations linéaires indépendantes. C'est le cas des modèles d'analyse de la variance et d'analyse de la covariance. Dans ce cas, la matrice $X'X$ n'est pas inversible et donc il existe une infinité de solutions $\widehat{\theta}$ vérifiant le système. Nous sommes dans un cas d'indétermination du système. La stratégie consiste alors à choisir, parmi cette infinité de solutions, une solution particulière intéressante pour le problème étudié, en rajoutant des contraintes sur les paramètres

de θ . On devra rajouter autant de contraintes (i.e. de colonnes à la matrice X) qu'il manque d'équations au système. Cependant, il ne faudra pas oublier le sens particulier que les contraintes choisies auront alors donné aux paramètres. En général, on ne s'intéresse pas aux estimations de ces paramètres (qui n'ont donc pas de sens intrinsèque) mais plutôt à des combinaisons invariantes de ces paramètres (qui ne dépendent pas des contraintes choisies, cf 2.1.4).

On introduit donc des contraintes linéaires sur les paramètres de départ :

$$\begin{matrix} H & \theta & = & 0 \\ (h \times (p+1)) & ((p+1) \times 1) & & \end{matrix}$$

H est une matrice de dimension $h \times (p+1)$, où $h = p+1 - r$, telle que la matrice G (définie plus bas) soit de rang $p+1$. En effet, dans ce cas, on a une solution unique pour l'estimation des paramètres liés par la relation $H\hat{\theta} = 0$ en résolvant le système à $n+h$ équations et $p+1$ inconnues :

$$\begin{cases} X\hat{\theta} = \hat{Y} \\ H\hat{\theta} = 0 \end{cases}$$

dont l'écriture condensée est :

$$G\hat{\theta} = \begin{bmatrix} X \\ H \end{bmatrix} \hat{\theta} = \begin{bmatrix} \hat{Y} \\ 0 \end{bmatrix}$$

La méthode des moindres carrés mène alors au système suivant :

$$G'G\hat{\theta} = X'X\hat{\theta} + H'H\hat{\theta} = X'X\hat{\theta} = X'Y$$

Puisque $G'G$ est de plein rang (construite pour cela), elle est inversible. On obtient comme solution :

$$\hat{\theta} = (G'G)^{-1}X'Y \tag{2.5}$$

Il est commode d'écrire la matrice $(G'G)^{-1}$ avec la notation $(X'X)^-$. On appelle $(X'X)^-$ "inverse généralisée" de $X'X$. Le projecteur orthogonal P est dans ce cas égal à :

$$P = X(X'X)^-X'$$

Remarque 1. Si on change la matrice H , c'est-à-dire les contraintes, on obtient des valeurs différentes pour les estimations des paramètres de θ , mais le projecteur P ne change pas, cela parce que le sous-espace $L(X)$ sur lequel on projette ne dépend pas des contraintes.

Remarque 2. Le choix des contraintes supplémentaires peut se faire avec deux objectifs :

- faciliter les calculs numériques,
- donner un sens particulier aux paramètres.

Exemple. Considérons un exemple simple illustratif : on mesure la hauteur de 4 arbres de même âge de deux espèces différentes (deux pins et deux hêtres), et on souhaite savoir si l'espèce a une influence sur la hauteur des arbres. On considère donc un modèle d'analyse de la variance à un facteur à $I = 2$ niveaux avec un même nombre de répétitions par niveau $n_r = 2$ qui s'écrit :

$$Y_{it} = \mu + \alpha_i + E_{it} \quad \text{pour } i = 1, 2 \text{ et } t = 1, \dots, n_r$$

Les données observées sont les suivantes :

$$\left\{ \begin{array}{lll} \text{hêtre} & 30 & y_{11} \\ \text{hêtre} & 34 & y_{12} \\ \text{pin} & 40 & y_{21} \\ \text{pin} & 42 & y_{22} \end{array} \right.$$

Les paramètres de l'espérance du modèle à estimer sont : μ, α_1 et α_2 et le système des équations normales s'écrit :

$$\left\{ \begin{array}{l} n\hat{\mu} + n_r \sum_i \hat{\alpha}_i = \sum_{i,t} y_{it} \\ n_r \hat{\mu} + n_r \hat{\alpha}_1 = \sum_t y_{1t} \\ n_r \hat{\mu} + n_r \hat{\alpha}_2 = \sum_t y_{2t} \end{array} \right.$$

On remarque que l'équation 1 est la somme des deux autres équations. Ainsi cette équation n'apporte pas d'informations supplémentaires pour estimer les 3 paramètres du modèle. On dispose donc d'un système de $I = 2$ équations à 3 inconnues. Du point de vue de la matrice X , cette dernière n'aura pas un rang égal à $I + 1 = 3$ mais à $I = 2$. Et dans ce cas, le sous-espace $L(X)$ sera engendré par $I + 1$ vecteurs (X_0, \dots, X_I) mais sera de dimension I . Pour récupérer une équation, on introduit une contrainte. Nous présentons deux jeux de contraintes classiquement utilisés :

1. **contrainte(a)** : $\sum_{i=1}^I \alpha_i = 0$ (la somme de chaque niveau du facteur est nulle). Dans ce cas, la matrice H est :

$$H = (0, 1, 1),$$

et en résolvant le système des équations normales, on obtient dans un plan équilibré (i.e. quand le nombre de répétitions est le même) :

$$\left\{ \begin{array}{l} \hat{\mu} = y_{\bullet\bullet} \\ \hat{\alpha}_i = y_{i\bullet} - y_{\bullet\bullet} \quad \forall i = 1, \dots, I \end{array} \right.$$

où $y_{\bullet\bullet}$ représente la moyenne générale des x_i et $y_{i\bullet}$ celle du groupe i . Ces estimations s'interprètent comme des écarts à la moyenne générale.

$$\hat{\mu} = 36.5, \quad \hat{\alpha}_1 = -4.5, \quad \hat{\alpha}_2 = 5.$$

Dans le cas d'un dispositif non équilibré, on utilise plutôt la contrainte $\sum_{i=1}^I n_i \alpha_i = 0$.

2. **contrainte(b)** : $\alpha_I = 0$ (modalité de référence). Dans ce cas, la matrice H est :

$$H = (0, 0, 1).$$

Cette contrainte est particulièrement intéressante dans le cas où un des niveaux du facteur sert de témoin de l'expérience. C'est la contrainte utilisée par certains logiciels, comme SAS. On obtient ici comme solution particulière :

$$\begin{cases} \hat{\mu} &= y_{I\bullet} \\ \hat{\alpha}_i &= y_{i\bullet} - y_{I\bullet} \quad \forall i = 1, \dots, I \end{cases}$$

Ces estimations s'interprètent comme des écarts au groupe I de référence.

$$\hat{\mu} = 41, \quad \hat{\alpha}_1 = -9, \quad \hat{\alpha}_2 = 0.$$

On voit donc que les estimations de μ et α_i dépendent des contraintes choisies. Et par conséquent leur interprétation aussi.

2.1.4 Combinaisons linéaires invariantes

Tout d'abord, une combinaison linéaire des paramètres θ s'écrit

$$\begin{array}{rcl} \psi & = & \sum_{j=1}^p c_j \theta_j = \begin{array}{cc} c & \theta \\ 1 \times 1 & 1 \times p \quad p \times 1 \end{array} \end{array}$$

On peut généraliser à une fonction linéaire ψ de dimension q des paramètres de θ :

$$\begin{array}{rcl} \psi & = & c \quad \theta \\ q \times 1 & & q \times p \quad p \times 1 \end{array}$$

Définition 4 Une combinaison linéaire des paramètres est invariante si son estimation est la même quelque soit le système de contraintes.

Par exemple, pour tout i , $\mathbb{E}(Y_i) = (X\theta)_i$ est une combinaison linéaire invariante. En effet, $(X\hat{\theta})_i = (PY)_i = \hat{Y}_i$ et P ne dépend pas des contraintes. Reprenons l'exemple très simple de l'étude de l'influence de l'espèce sur la hauteur des arbres considéré dans le paragraphe précédent et calculons \hat{y}_i pour les deux jeux de contraintes. Par définition,

$$\hat{y}_{it} = \hat{\mu} + \hat{\alpha}_i \quad \text{pour tout } t \text{ appartenant au groupe } i$$

Ainsi on obtient pour les deux contraintes : $\hat{y}_{it} = y_{i\bullet}$, et donc \hat{y} ne dépend pas des contraintes et est donc invariant.

Par contre, comme on a pu le voir dans le paragraphe précédent, $\hat{\mu}$ et les $\hat{\alpha}_i$ dans un modèle d'analyse de la variance à un facteur dépendent des contraintes et donc ne sont pas invariants.

Parmi les combinaisons linéaires possibles des paramètres, certaines sont particulièrement intéressantes, il s'agit de celles que l'on appelle contrastes dont la définition est donnée ci-dessous.

Définition 5 On appelle contraste toute combinaison linéaire des paramètres $\psi = \sum_{j=1}^p c_j \theta_j$ tel que $\sum_{j=1}^p c_j = 0$, c'est-à-dire que la somme de ses coefficients est nulle.

Par exemple, on peut s'intéresser à la comparaison de deux niveaux d'un facteur α_1 et α_2 . Pour cela, on va considérer la combinaison linéaire des paramètres suivante : $\psi = \alpha_1 - \alpha_2$, qui est un contraste car $c_1 = 1$ et $c_2 = -1$.

2.2 Propriétés des estimateurs de θ

Le théorème suivant donne les expressions de l'espérance et la variance de $\hat{\theta}$.

Theorème 2 *L'espérance et la matrice de variance-covariance de l'estimateur des moindres carrés des paramètres θ du modèle (1.1) (dont l'expression est donnée par l'équation (2.4) dans le cas où X est de plein rang et par (2.5) sinon) valent respectivement*

$$\mathbb{E}(\hat{\theta}) = \theta \quad \text{quelque soit le jeu de contraintes.}$$

$\hat{\theta}$ est donc un estimateur sans biais de θ , et

$$\mathbb{V}(\hat{\theta}) = (X'X)^{-1}\sigma^2 \quad \text{si} \quad r = p + 1 \quad (2.6)$$

ou bien

$$\mathbb{V}(\hat{\theta}) = (X'X)^-X'X(X'X)^-\sigma^2 \quad \text{si} \quad r < p + 1$$

On peut détailler ce dernier résultat : en notant x^{lk} le terme général de la matrice $(X'X)^{-1}$ (ou de $(X'X)^-X'X(X'X)^-$ suivant les cas), on a

$$\mathbb{V}(\hat{\theta}_k) = x^{kk}\sigma^2 \quad \text{et} \quad \text{Cov}(\hat{\theta}_k, \hat{\theta}_l) = x^{kl}\sigma^2$$

On trouvera la démonstration du théorème en Annexe 6.4.

Remarque 1. La formule de la variance des estimateurs permet d'apprécier la précision de ces estimateurs. Ici on voit que la variance dépend de σ^2 , la précision des estimations dépendra donc à la fois de la qualité du matériel expérimental et de la qualité du modèle.

Remarque 2. La matrice $\mathbb{V}(\hat{\theta})$ contient sur la diagonale principale les variances des estimateurs des paramètres, et ailleurs les covariances de ces estimateurs deux à deux. Il est donc intéressant de voir si cette matrice est diagonale, car alors les estimateurs des paramètres sont non corrélés.

Optimalité de l'estimateur des moindres carrés ordinaires. Une fois un estimateur défini, on s'intéresse naturellement à savoir si il est optimal au sens où c'est celui qui est le plus précis (i.e. qui a la plus petite variance parmi la classe d'estimateurs considérée). Le théorème de Gauss-Markov donné ci-dessous indique que l'estimateur des moindres carrés est optimal parmi tous les estimateurs linéaires sans biais. Cette propriété est vraie même si la loi de E n'est pas gaussienne.

Theorème 3 (Gauss-Markov) *Dans le cas où $r = p + 1$, $\hat{\theta}$ est le meilleur estimateur linéaire sans biais de θ au sens de la variance, i.e. $\hat{\theta}$ est l'estimateur linéaire sans biais de θ de variance minimale.*

La démonstration de ce résultat est donnée en Annexe 6.6.

2.3 Loi des estimateurs

Dans les deux cas ($r = p+1$ ou $r < p+1$), les estimateurs des paramètres sont des estimateurs linéaires, c'est-à-dire qu'ils sont des combinaisons linéaires des variables observées. En utilisant les suppositions que nous avons faites sur la loi de la matrice des variables aléatoires résiduelles E , nous pouvons déduire la loi de la matrice estimateur $\hat{\theta}$.

2.3.1 Loi de $\hat{\theta}$.

On a fait l'hypothèse que la loi des résidus était gaussienne, ce qui implique que les Y_i le sont aussi. Les estimateurs $\hat{\theta}_k$ le sont donc aussi et

$$\hat{\theta} \sim \mathcal{N}_p(\theta, \mathbb{V}(\hat{\theta})),$$

où la variance de $\hat{\theta}$ est donnée par (2.8). On peut aussi donner ce résultat pour un $\hat{\theta}_k$ particulier :

$$\hat{\theta}_k \sim \mathcal{N}(\theta_k, x^{kk} \sigma^2), \quad (2.7)$$

où x^{kk} est défini dans le paragraphe précédent.

2.3.2 Combinaisons linéaires des θ_k .

On peut aussi s'intéresser à une combinaison linéaire des paramètres ψ (comme les combinaisons linéaires invariantes) :

$$\psi = c\theta \quad \text{d'estimateur} \quad \hat{\psi} = c\hat{\theta},$$

où c est un vecteur de dimension $p \times 1$. On a alors

$$\begin{aligned} \mathbb{E}(\hat{\psi}) &= c\mathbb{E}(\hat{\theta}) = c\theta, \\ \mathbb{V}(\hat{\psi}) &= c\mathbb{V}(\hat{\theta})c', \end{aligned}$$

soit

$$\hat{\psi} \sim \mathcal{N}(\psi, c\mathbb{V}(\hat{\theta})c').$$

2.4 Estimation de la variance

2.4.1 Résidus

Une fois obtenues les valeurs estimées \hat{Y} , on a facilement l'estimation des résidus \hat{E} qui est :

$$\hat{E} = Y - \hat{Y} = (I_n - P) Y,$$

où I_n est la matrice identité de dimension $n \times n$. \hat{E} est la projection orthogonale de Y sur l'espace orthogonal à $L(X)$ dans \mathbb{R}^n (la matrice de projection est $I_n - P$). Puisque \hat{Y} est la projection de Y sur $L(X)$, les vecteurs \hat{E} et \hat{Y} sont orthogonaux, ce qui signifie que les vecteurs aléatoires \hat{E} et \hat{Y} sont non corrélés.

2.4.2 Estimation de σ^2

Nous avons vu au paragraphe précédent qu'on cherche à estimer les paramètres de l'espérance en minimisant la distance euclidienne dans \mathbb{R}^n entre le point des valeurs observées et le point du sous-espace $L(X)$ des valeurs estimées. Soit SCR la somme des carrés résultante :

$$SCR = (Y - \hat{Y})'(Y - \hat{Y}).$$

Cette somme est appelée "**Somme des Carrés Résiduelle**" car c'est la somme des carrés des résidus :

$$SCR = \hat{E}'\hat{E} = \sum_{t=1}^n \hat{E}_t^2.$$

Si on utilise les hypothèses faites sur les variables aléatoires résiduelles, on peut montrer que

$$\mathbb{E}(SCR) = (n - r)\sigma^2,$$

où r est le rang de la matrice X , i.e. la dimension de l'espace $L(X)$. La démonstration de cette égalité est donnée dans l'Annexe 6.5.

On peut comprendre géométriquement la raison du terme $n - r$: Y évolue dans un espace de dimension n alors que \hat{Y} appartient à l'espace $L(X)$ de dimension r , le rang de la matrice X . \hat{E} est donc dans un espace de dimension $n - r$, a donc $n - r$ degrés de liberté.

On définit donc S^2 l'estimateur de σ^2 :

$$S^2 = \frac{SCR}{n - r} = \frac{(Y - \hat{Y})'(Y - \hat{Y})}{n - r}, \quad (2.8)$$

et cet estimateur est sans biais par construction :

$$\mathbb{E}(S^2) = \sigma^2.$$

On notera $\hat{\sigma}^2$ l'estimation de σ^2 .

2.4.3 Loi de σ^2

On peut montrer (voir la démonstration dans l'Annexe 6.3) que

$$\frac{(n - r)S^2}{\sigma^2} \sim \chi_{n-r}^2$$

2.4.4 Décomposition de la variance et coefficient de détermination

Theorème 4 *La variabilité de Y se décompose en la somme de deux termes :*

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|Y - \hat{Y}\|^2$$

ou encore si le modèle comprend un terme constant

$$\begin{aligned} \|Y - \bar{Y}\|^2 &= \|\hat{Y} - \bar{Y}\|^2 + \|Y - \hat{Y}\|^2 \\ SCT &= SCM + SCR \end{aligned}$$

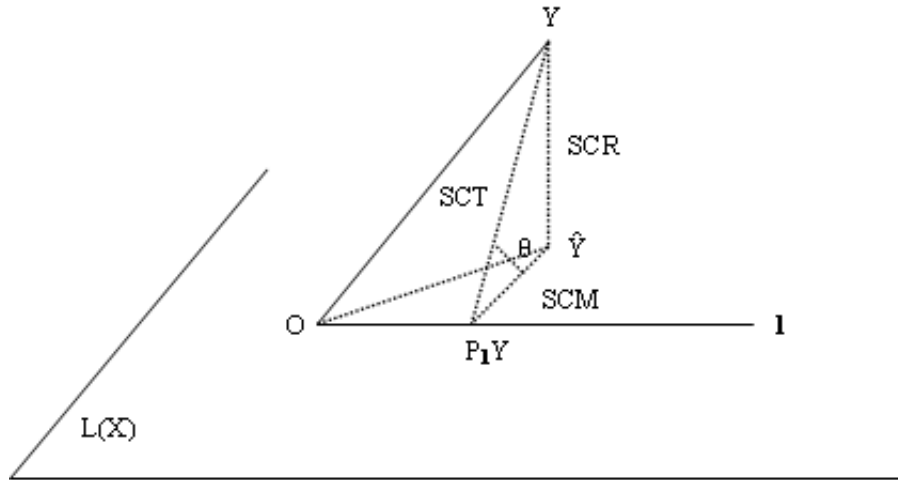


FIGURE 2.3 – Sommes de carrés.

Ceci signifie que la somme des carrés totale (SCT), est égale la somme des carrés due au modèle (SCM) plus la somme des carrés résiduelle (SCR). Ce sont ces sommes de carrés que l'on trouve dans les tables d'analyse de variance. La figure 2.3 est une représentation géométrique de ces sommes de carrés.

Preuve Pour cela, il suffit juste de se rappeler que les vecteurs \hat{Y} et \hat{E} sont orthogonaux et d'utiliser le théorème de Pythagore.

Définition 6 On appelle **Coefficient de détermination** la quantité notée R^2 suivante :

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

Cette quantité est une mesure de la qualité d'ajustement du modèle aux observations et est toujours comprise entre 0 et 1.

Géométriquement ce rapport est égal au carré du cosinus de l'angle du vecteur $Y - \bar{Y}$ avec le sous-espace $L(X)$, (noté θ figure 2.3).

On s'intéresse surtout à cette quantité lorsque l'on cherche un modèle prédictif. Si elle est proche de 1, le modèle est très bien ajusté aux données et les prédictions par ce modèle seront très précises.

Cependant R^2 est un estimateur biaisé qui surévalue la vraie qualité d'ajustement. En particulier, R^2 augmente avec le nombre de variables incluses dans le modèle. Ainsi dans une optique de prédiction, si on doit choisir entre un modèle avec $p+1$ variables et le même modèle mais avec une variable en moins, en se fiant au R^2 on choisira toujours le modèle avec le plus de variables. Or cette variable peut ne pas être intéressante. C'est pourquoi, il est préférable d'utiliser un autre estimateur qui prend en compte le nombre de variables du modèle et qui est un estimateur sans biais du R^2 théorique.

Définition 7 On appelle **Coefficient de détermination ajusté** la quantité notée R_{adj}^2 suivante, définie si $n - p - 1 > 0$:

$$R_{adj}^2 = 1 - \frac{SCR/(n - p - 1)}{SCT/(n - 1)} = R^2 - \frac{p}{n - p - 1}(1 - R^2)$$

où p est le nombre de variables explicatives considérées dans le modèle.

2.5 Intervalle de confiance

Connaissant la loi des paramètres, on peut construire facilement un intervalle de confiance pour chacun des paramètres au niveau de confiance $(1 - \alpha)$.

Intervalle de confiance sur un paramètre $\hat{\theta}_k$. D'après la loi de $\hat{\theta}_k$ donnée par l'équation (2.7), on a que

$$\frac{\hat{\theta}_k - \theta_k}{\sigma \sqrt{x^{kk}}} \sim \mathcal{N}(0, 1)$$

Puisque la variance σ^2 est inconnue, on la remplace par son estimation et on obtient que la quantité obtenue suit une loi de Student à $(n - r)$ degrés de liberté :

$$\frac{\hat{\theta}_k - \theta_k}{\hat{\sigma} \sqrt{x^{kk}}} \sim \mathcal{T}_{n-r}.$$

Le nombre de degrés de liberté de cette variable de Student est celui de l'estimateur de la variance résiduelle. Ceci permet d'obtenir l'intervalle de confiance à un niveau de confiance $1 - \alpha$ donné par

$$\hat{\theta}_k - t_{1-\alpha/2, n-r} \hat{\sigma} \sqrt{x^{kk}} \leq \theta_k \leq \hat{\theta}_k + t_{1-\alpha/2, n-r} \hat{\sigma} \sqrt{x^{kk}}$$

La valeur $t_{1-\alpha/2, n-r}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $(n - r)$ degrés de liberté.

Intervalle de confiance d'une combinaison linéaire. De la même façon, on peut construire un intervalle de confiance pour une combinaison linéaire qui est :

$$\hat{\psi} - t_{1-\alpha/2, n-r} \hat{\sigma} \sqrt{c \hat{\mathbb{V}}(\hat{\psi}) c'} \leq \psi \leq \hat{\psi} + t_{1-\alpha/2, n-r} \hat{\sigma} \sqrt{c \hat{\mathbb{V}}(\hat{\psi}) c'}$$

où $\hat{\mathbb{V}}(\hat{\psi})$ est l'estimation de $\mathbb{V}(\hat{\psi})$.

Exemple. Dans le modèle d'analyse de la variance à un facteur qualitatif considéré dans le paragraphe 2.1.3 (cas d'un dispositif équilibré avec n_r répétitions par niveau), on peut chercher un intervalle de confiance de :

$$\psi = \alpha_1 - \alpha_2$$

qui est un contraste invariant des paramètres. Son estimation est :

$$\widehat{\psi} = \widehat{\alpha}_1 - \widehat{\alpha}_2 = y_{1\bullet} - y_{2\bullet}$$

$$\begin{aligned}\mathbb{V}(\widehat{\psi}) &= \mathbb{V}(\widehat{\alpha}_1) + \mathbb{V}(\widehat{\alpha}_2) + 2\text{Cov}(\widehat{\alpha}_1, \widehat{\alpha}_2) \\ &= 2\sigma^2/n_r\end{aligned}$$

Puisque σ^2 est inconnue, on en prend une estimation et

$$\widehat{\mathbb{V}}(\widehat{\psi}) = 2\widehat{\sigma}^2/n_r$$

Et l'intervalle de confiance de ψ au niveau $1 - \alpha$ est :

$$(y_{1\bullet} - y_{2\bullet}) - t_{1-\alpha/2, n-r}(2\widehat{\sigma}^2/n_r)^{1/2} \leq \alpha_1 - \alpha_2 \leq (y_{1\bullet} - y_{2\bullet}) + t_{1-\alpha/2, n-r}(2\widehat{\sigma}^2/n_r)^{1/2}$$

Intervalle de confiance de σ^2 . Pour la variance, on obtient,

$$\frac{(n-r)\widehat{\sigma}^2}{\chi_{1-\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-r)\widehat{\sigma}^2}{\chi_{\alpha/2}^2}$$

Se référer au livre de première année [2] pour la construction de cet intervalle.

Chapitre 3

Tests d'hypothèses

3.1 Test sur un paramètre

On considère le modèle $Y = X\theta + E$ où $\theta = (\theta_1, \dots, \theta_{p+1})$. On veut tester l'hypothèse $H_0 = \{\theta_1 = a\}$, où a est une valeur définie a priori. On rejette H_0 au niveau α si

$$\frac{|\hat{\theta}_1 - a|}{\sqrt{\widehat{\mathbb{V}}(\hat{\theta}_1)}} \geq t_{1-\alpha/2; \nu} \quad (3.1)$$

où

- $\nu = n - r$ est le nombre de degrés de liberté de la résiduelle du modèle
- $t_{1-\alpha/2; \nu}$ est le quantile $1 - \alpha/2$ de la loi de Student à ν degrés de liberté.

De façon équivalente on rejette H_0 au niveau α si la probabilité critique du test,

$$P \left[|T| > |\hat{\theta}_1 - a| / \sqrt{\widehat{\mathbb{V}}(\hat{\theta}_1)} \right] \quad (3.2)$$

est inférieure à α , où T est distribué selon une loi de Student à ν degrés de liberté.

Pour la démonstration de ce résultat, voir l'Annexe 6.1. En pratique on utilise ce test avec $a = 0$ pour savoir si le paramètre θ_1 peut être considéré comme nul. Ce test est utilisé par exemple dans une régression multiple pour savoir si la variable X_1 est utile pour prédire Y (on suppose ici que θ_1 est le coefficient de régression associé à X_1).

3.2 Test d'une combinaison linéaire des paramètres

En analyse de la variance le test précédent est inutile voire dangereux car souvent mal interprété. En effet les paramètres n'ont pas de sens intrinsèque car leur valeur dépend du système de contraintes. Par contre on s'intéresse à l'hypothèse $H_0 = \{c\theta = a\}$ où $c\theta$ est un contraste invariant (qui ne dépend pas des contraintes) et a est une valeur définie a priori. On rejette l'hypothèse H_0 au niveau α si

$$\frac{|c\hat{\theta} - a|}{\sqrt{c\widehat{\mathbb{V}}(\hat{\theta})c'}} \geq t_{1-\alpha/2; \nu} \quad (3.3)$$

Exemple : Dans le modèle d'analyse de la variance à 2 facteurs sans interaction,

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

($i = 1, 2$ et $j = 1, 2$), on veut savoir s'il y a une différence entre les niveaux 1 et 2 du premier facteur, dont les effets sont notés α_1 et α_2 .

On est donc amené à tester $H_0 = \{\alpha_1 - \alpha_2 = 0\}$. On a $\theta' = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2)$, $c' = (0, 1, -1, 0, 0)$ et $a = 0$.

3.3 Test de modèles emboîtés

3.3.1 Modèles emboîtés

On considère le modèle $Y = X\theta + E$, appelé modèle M_1 et le modèle $Y = X_0\theta_0 + E_0$, appelé modèle M_0 .

Définition 8 M_0 est emboîté dans M_1 si et seulement si le sous-espace vectoriel de R^n engendré par les colonnes de X_0 est contenu dans le sous-espace vectoriel engendré par les colonnes de X .

Dans [6], page 26, on trouvera un exemple de régression linéaire emboîté dans un modèle d'analyse de la variance.

En pratique, la règle suivante est utile pour identifier la plupart des modèles emboîtés usuels : on dit qu'un modèle $Y = X_0\theta_0 + E$, noté M_0 , est emboîté dans M_1 si et seulement si M_0 est un cas particulier de M_1 , obtenu en annulant certains paramètres ou certaines combinaisons linéaires de paramètres de M_1 .

Exemple : le modèle $Y_{ij} = \mu + \alpha_i + E_{ij}$ est emboîté dans le modèle $Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$.

3.3.2 Test de modèles emboîtés

Les tests (3.1) et (3.3) ne permettent de tester qu'un seul paramètre ou un seul contraste. Mais

- en régression on veut souvent savoir si un ensemble de variables peut être éliminé sans perte d'information. Par exemple dans un modèle de régression à p variables, on veut tester $H_0 = \{\theta_q = \theta_{q+1} = \dots = \theta_{p+1} = 0\}$.
- en analyse de la variance on veut savoir si un facteur (qui comporte plusieurs paramètres) peut être éliminé du modèle. Par exemple dans un modèle d'analyse de la variance à 2 facteurs $Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$, on veut tester que le premier facteur (α) n'a pas d'effet, c'est à dire $H_0 = \{\alpha_1 = \alpha_2 = \dots = \alpha_I = 0\}$.

On utilise alors le test donné par le théorème 5.

Le modèle M_1 est plus général que le modèle M_0 , donc il s'ajuste mieux aux données que ce dernier. Soit $\hat{Y} = X\hat{\theta}$ la prédiction de Y et $SCM_1 = \|\hat{Y} - \bar{Y}\|^2$ la somme de carrés "expliquée" par le modèle M_1 et $\hat{Y}_0 = X_0\hat{\theta}_0$ la prédiction de Y et $SCM_0 = \|\hat{Y}_0 - \bar{Y}\|^2$ la somme de carrés "expliquée" par le modèle M_0 . On a $SCM_1 - SCM_0 \geq 0$.

La question qui se pose est la suivante : est-ce que l'augmentation du pouvoir explicatif (sur Y) apportée par le modèle M_1 par rapport au modèle M_0 est suffisamment forte pour justifier le choix de M_1 de préférence à M_0 ?

Comme M_1 a plus de paramètres ($p_1 + 1$) que M_0 ($p_0 + 1$), la quantité utile est $(SCM_1 - SCM_0)/(p_1 - p_0)$, l'amélioration du pouvoir explicatif par paramètre supplémentaire. Cette statistique a une loi de probabilité qui dépend de la variance résiduelle σ^2 qu'il faut éliminer.

Theorème 5 *L'hypothèse $H_0 = \{M = M_0\}$ est rejetée au profit de $H_1 = \{M = M_1\}$ au niveau α si*

$$\frac{\frac{SCM_1 - SCM_0}{p_1 - p_0}}{\frac{SCR_1}{\nu_1}} \geq \mathcal{F}_{1-\alpha; p_1 - p_0; \nu_1} \quad (3.4)$$

où

- $SCR_1 = SCT - SCM_1 = \|Y - \widehat{Y}_1\|^2$ est la somme des carrés résiduelle du modèle M_1 ,
- ν_1 est son nombre de degrés de liberté
- et $\mathcal{F}_{1-\alpha; p_1 - p_0; \nu_1}$ est le quantile $1 - \alpha$ de la loi de Fisher à $p_1 - p_0$ et ν_1 degrés de liberté.

De façon équivalente, on rejette H_0 si la probabilité critique du test,

$$P \left[\mathcal{F} > \frac{\nu_1 (SCM_1 - SCM_0)}{(p_1 - p_0) SCR_1} \right] \quad (3.5)$$

est inférieure à α où \mathcal{F} est distribué selon une loi de Fisher à $p_1 - p_0$ et ν_1 degrés de liberté.

Pour la démonstration de ce résultat, voir l'Annexe 6.2.

Propriétés

1. Invariance : on dit que le test est invariant si son résultat ne dépend pas du système de contraintes (voir 2.1.4). Dans l'exemple de la section 3.3.1, le test de modèles emboîtés est invariant et le test de $H_0 = \{\alpha_1 = 0\}$ ne l'est pas si le facteur testé a plus de 2 niveaux.
2. En analyse de la variance le test de nullité de **tous** les paramètres d'un facteur ou d'une interaction est invariant. En pratique les tests \mathcal{F} emboîtés donnés dans les logiciels de modèle linéaire sont invariants.
3. le test de $H_0 = \{\theta_1 = 0\}$ par test de modèle emboîté est équivalent au test (3.1). La statistique de test donnée par (3.4) est le carré de (3.1). Sa loi de probabilité H_0 est une loi de Fisher à 1 et ν_1 degrés de liberté, c'est à dire la loi du carré d'une variable aléatoire suivant une loi de Student à ν_1 degrés de liberté.

3.3.3 Vision géométrique du test

En reprenant la figure (2.3), nous pouvons représenter le sous-espace $L(X_0)$ engendré par les colonnes de X_0 par une droite passant par le point o , dans le plan qui représente le sous-espace $L(X)$ engendré par les colonnes de X . L'idée du test de l'hypothèse H_0 consiste à voir si le vecteur Y est suffisamment proche du sous-espace $L(X_0)$, auquel cas on peut conclure que c'est le modèle M_0 qui est le bon. C'est le cas si $\|P_1(Y) - P_0(Y)\|^2$ est "petit".

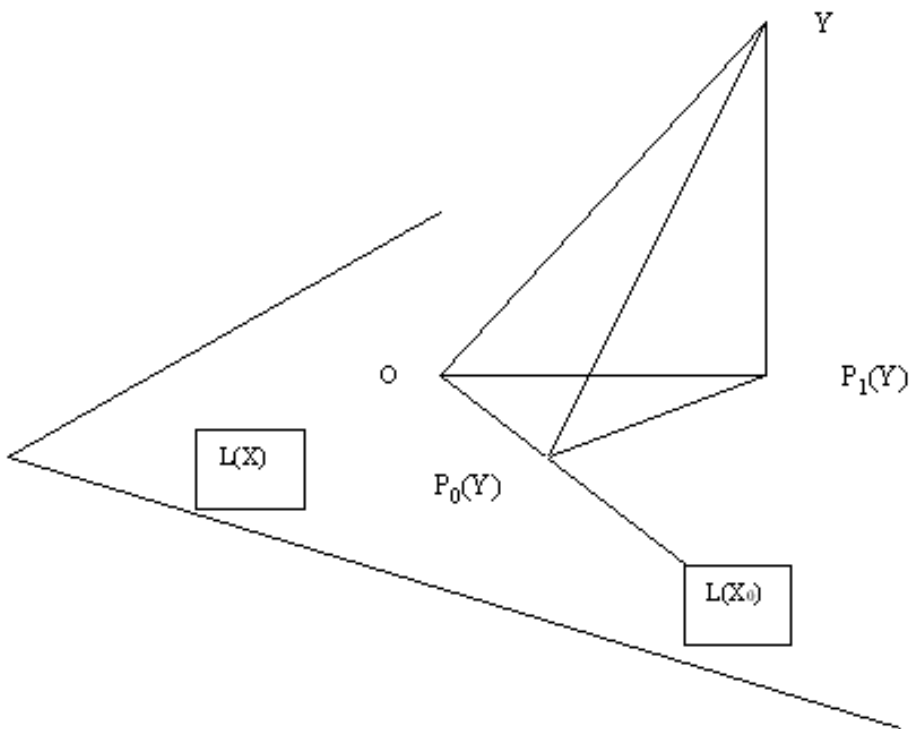


FIGURE 3.1 – Représentation graphique du test emboîté

Chapitre 4

Notion d'orthogonalité

En analyse de la variance, la notion d'orthogonalité entre les facteurs et plus généralement entre les termes d'un modèle joue un rôle important. Ce chapitre est inspiré de [7] dont sont issus en particulier les exemples.

4.1 Définition de l'orthogonalité

4.1.1 Cas de 2 facteurs

On considère les modèles d'analyse de la variance à deux facteurs, additif et avec interaction avec la notation simplifiée suivante :

$$\begin{aligned}\mathbb{E}(Y) &= I + A + B && \text{modele sans interaction} \\ \mathbb{E}(Y) &= I + A + B + AB && \text{modele avec interaction}\end{aligned}$$

Définition 9 *Un plan factoriel à deux facteurs, notés A et B , est orthogonal si il vérifie :*

$$\forall i = 1 \dots I; j = 1 \dots J \quad n_{ij} = \frac{n_{i+} \times n_{+j}}{n} \quad (4.1)$$

où n_{i+} est l'effectif total pour le niveau i du facteur A (respectivement n_{+j} pour B).

Dans la pratique on dresse le tableau des effectifs croisés des modalités de A et B (Table 4.1) et l'on vérifie cette propriété sur toutes les cases.

On note cette orthogonalité :

$$A \perp B$$

La définition (4.1) signifie que si on tire au sort une unité expérimentale (avec tirage équiprobable entre toutes les unités), les événements "l'unité expérimentale a la modalité A_i " et "l'unité expérimentale a la modalité B_j " sont indépendants.

$$P(A_i) = \frac{n_{i+}}{n}, \quad P(A_i \cap B_j) = \frac{n_{ij}}{n}, \quad P(B_j) = \frac{n_{+j}}{n}$$

TABLE 4.1 – tableau croisé (A,B)

	B_1	...	B_j	...	B_J	total
A_1	n_{11}		n_{1j}		n_{1J}	n_{1+}
...						
A_i	n_{i1}		n_{ij}		n_{iJ}	n_{i+}
...						
A_I	n_{I1}		n_{Ij}		n_{IJ}	n_{I+}
total	n_{+1}		n_{+j}		n_{+J}	n

Si la propriété est vraie :

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

On dit alors que les facteurs A et B sont indépendants dans le plan d'expérience.

4.1.2 Cas général : Orthogonalité d'un plan d'expériences vis à vis d'un modèle

Tout d'abord, nous donnons la définition de la notion de couples disjoints d'un modèle.

Définition 10 Deux termes d'un modèle sont disjoints si ils ne partagent pas d'effet commun.

Exemple. Prenons le modèle d'analyse de la variance écrit sous la forme suivante :

$$\mathbb{E}(Y) = I + A + B + C + AB$$

où A , B , C , AB , sont appelés 'termes' du modèle. C et AB sont des termes disjoints car ils n'ont aucune lettre commune ; B et AB ne le sont pas. Ainsi tous les couples de termes disjoints de ce modèle sont : (A, B) , (A, C) , (B, C) , (AB, C) .

Définition 11 Un plan factoriel est orthogonal vis à vis d'un modèle M si pour tout couple (U, V) de termes disjoints du modèle, on a que U est orthogonal à V .

Un plan complet équilibré est donc orthogonal, mais ce n'est pas une condition nécessaire.

Exemple 1. On considère le plan décrit dans la table 4.2, pour le modèle : $\mathbb{E}(Y) = I + A + B + C$.

TABLE 4.2 – Plan d'expérience de l'exemple 1

unité expérimentale	A	B	C
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

Ce plan n'est pas complet car par exemple la combinaison (1,2,1) n'a pas été expérimentée, pourtant il est bien orthogonal vis à vis du modèle choisi comme on le voit sur la table 4.3.

TABLE 4.3 – couples de termes disjoints

Couple de termes disjoints	vérification de l'orthogonalité	Conclusion									
A,B	<table border="1"> <tr><td></td><td>B1</td><td>B2</td></tr> <tr><td>A1</td><td>1</td><td>1</td></tr> <tr><td>A2</td><td>1</td><td>1</td></tr> </table>		B1	B2	A1	1	1	A2	1	1	$A \perp B$
	B1	B2									
A1	1	1									
A2	1	1									
A,C	<table border="1"> <tr><td></td><td>C1</td><td>C2</td></tr> <tr><td>A1</td><td>1</td><td>1</td></tr> <tr><td>A2</td><td>1</td><td>1</td></tr> </table>		C1	C2	A1	1	1	A2	1	1	$A \perp C$
	C1	C2									
A1	1	1									
A2	1	1									
B,C	<table border="1"> <tr><td></td><td>C1</td><td>C2</td></tr> <tr><td>B1</td><td>1</td><td>1</td></tr> <tr><td>B2</td><td>1</td><td>1</td></tr> </table>		C1	C2	B1	1	1	B2	1	1	$B \perp C$
	C1	C2									
B1	1	1									
B2	1	1									

TABLE 4.4 – Plan d'Expérience de l'exemple 2

A	B	C	D
1	1	1	1
1	1	2	2
1	2	1	2
1	2	2	1
2	1	1	1
2	1	2	2
2	2	1	2
2	2	2	1
3	1	1	1
3	1	2	2
3	2	1	2
3	2	2	1

Exemple 2. On considère le plan à quatre facteurs, non complet, décrit dans la table 4.4 pour le modèle $\mathbb{E}(Y) = I + A + B + C + D + AB + CD$. Les couples de termes disjoints sont : (A,B), (A,C), (A,D), (B,C), (B,D), (B,CD), (C,D), (C,AB), (D,AB) et (AB,CD). L'orthogonalité est vérifiée pour tous les couples sauf (B,CD) et (AB,CD). La table 4.5 montre que B n'est pas orthogonal à (C,D) car $0 \neq 3 \times 6/12$. On peut vérifier de même que AB n'est pas orthogonal à CD. On conclut que le plan n'est pas orthogonal pour le modèle considéré.

Confusion d'effets

- On dit que 2 facteurs A et B sont confondus (on dit aussi "totalement confondus") dans un plan d'expériences donné, si dans ce plan la connaissance du niveau de A implique celle de B et réciproquement. On sera, dans ce cas, incapable de séparer l'effet de A sur la variable réponse Y de celui de B .

TABLE 4.5 – tableau croisé (B,CD)

	CD11	CD12	CD21	CD22
B1	3	0	0	3
B2	0	3	3	0

- On dit que 2 facteurs A et B sont partiellement confondus dans un plan d'expériences donné, si dans ce plan A n'est pas orthogonal à B mais les deux facteurs ne sont pas totalement confondus.

Exemple Supposons que le facteur A n'ait aucune influence sur la variable réponse Y et que seul le facteur B ait une action sur Y . Supposons que A soit partiellement confondu avec B . Si A change de valeur, cela fait bouger B qui fait alors bouger Y . On a donc l'impression que A a une action sur Y alors que ce n'est pas vrai :

$$A \rightarrow B \rightarrow Y.$$

La première flèche traduit une influence artificielle qui dépend des choix de l'expérimentateur et peut disparaître si on change le plan d'expérience. La séparation de l'effet de A sur la variable réponse Y de celui de B se fait à l'aide des sommes de carrés de Type III, en regardant l'effet de A sur Y à B fixé, voir paragraphe 4.3. Elle sera plus ou moins délicate selon la force de la confusion.

- On dit que les effets de A et B sont séparés lors que $A \perp B$.

Ces définitions s'étendent de façon immédiate à des termes disjoints plus généraux qu'un facteur comme une interaction. Par exemple on voit dans la table 4.5 que B est partiellement confondu avec CD .

4.1.3 Cas de la Régression multiple et de l'analyse de la covariance

- En régression multiple, le dispositif est orthogonal si les variables explicatives sont non corrélées entre elles. Dans ce cas les colonnes de la matrice X correspondant aux variables explicatives sont orthogonales entre elles. Ce cas est peu fréquent car en général les variables explicatives ne sont pas maîtrisées (à moins que cela soit fixé par l'expérimentateur).
- En analyse de la covariance les covariables ne sont orthogonales entre elles et orthogonales aux termes du modèle d'analyse de la variance que dans des cas particuliers très rarement rencontrés dans la pratique. Dans le cas d'une seule covariable X , les mêmes valeurs de X doivent être testées pour chaque modalité du facteur A pour que $A \perp X$.

4.2 Propriétés d'un dispositif orthogonal

L'orthogonalité est une propriété recherchée des dispositifs expérimentaux, en effet :

- Il n'y a aucune confusion d'effets entre les termes du modèle.

- La matrice de variance covariance des estimateurs est diagonale par blocs, en conséquence les estimateurs associés à des couples de termes disjoints sont non corrélés.
- Les estimations des paramètres concernant un terme ne changent pas lorsque l'on considère un sous-modèle. La somme de carrés due à un terme donné (c'est le numérateur du test 3.4 de nullité de ce terme du modèle d'analyse de la variance) ne dépend pas du modèle "emboîtant".
- Pour l'analyse de la variance, il existe des estimateurs intuitifs obtenus par de simples moyennes.

4.3 Sommes de carrés ajustées : Type I, II et III

On considère ici le modèle d'analyse de la variance à deux facteurs avec interaction :

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

4.3.1 Réduction

Définition 12 La réduction $R(c/\mu, a, b)$ représente la diminution de la somme de carrés résiduelle lorsque l'on passe du modèle comportant les effets a et b au modèle comportant a, b, c .

Exemple : en notant SCR_0 la somme de carrés du modèle :

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

et SCR_1 la somme de carrés résiduels du modèle :

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

on a

$$R(\beta/\mu, \alpha) = SCR_0 - SCR_1$$

Cette notation est cohérente avec la notion de modèles emboîtés. Elle est adaptée aux sommes de Type I et II mais pas totalement pour les sommes de Type III.

4.3.2 Sommes de carrés

On note $L(\alpha)$, l'espace vectoriel engendré par les colonnes de la matrice X associées à l'effet α . $L(\alpha)$ est donc un sous-espace de $L(X)$.

Le tableau suivant définit les sommes de Type I, II et III en terme de réduction dans le cadre d'un modèle d'analyse de la variance à deux facteurs avec interaction. Les résultats pour le modèle d'analyse de la covariance s'obtiennent par analogie.

Tout d'abord, remarquons que l'ordre d'écriture des facteurs dans le modèle a une importance si on considère les tests de Type I, ce qui n'est pas le cas pour les deux autres Types : l'effet entré en premier dans le modèle est aussi testé en premier sans prendre en compte les autres effets. Pour les sommes de Type II, on voit aussi que l'interaction n'est pas prise en

TABLE 4.6 – Sommes de Type I, II, III

	Type I	Type II	Type III
facteur 1 : α	$R(\alpha/\mu)$	$R(\alpha/\mu, \beta)$	$R(\alpha^*/\mu^*, \beta^*, \gamma^*)$
facteur 2 : β	$R(\beta/\mu, \alpha)$	$R(\beta/\mu, \alpha)$	$R(\beta^*/\mu^*, \alpha^*, \gamma^*)$
interaction	$R(\gamma/\mu, \alpha, \beta)$	$R(\gamma/\mu, \alpha, \beta)$	$R(\gamma^*/\mu^*, \alpha^*, \beta^*)$

compte lors du test des effets principaux : on ne peut donc pas de mesurer l'apport uniquement d'un effet principal en présence d'un modèle avec interaction. Cette interaction est par contre prise en compte dans les sommes de Type III. Cependant comme il est indiqué dans le tableau par le symbole "*", chaque effet est testé au travers de versions contraintes. La raison est assez simple : le sous-espace engendré par γ , $L(\gamma)$, contient les deux sous-espaces engendrés par les deux effets principaux, $L(\alpha)$ et $L(\beta)$. Ainsi intéressons-nous, par exemple, au test de l'effet α . Ce test s'effectue au moyen de la somme de carrés $R(\alpha/\mu, \beta, \gamma)$: elle représente la diminution de somme résiduelle lorsque l'on passe du modèle pour lequel on projette sur l'espace engendré par $L(\gamma)$ et $L(\beta)$ au modèle complet (avec α , β et γ). Or comme $L(\alpha)$ est inclu dans le sous-espace $L(\gamma)$, ces deux modèles sont exactement les mêmes (cf [6] p 115). Et dans ce cas, la réduction $R(\alpha/\mu, \beta, \gamma)$ est nulle. On ne peut donc pas utiliser cette somme pour tester α . Pour se convaincre que ces tests sont différents, nous donnons ci-dessous explicitement l'hypothèse testée pour chacun des trois Types dans le cas du test de l'effet principal α :

	Hypothèse testée pour α
Type I	$H : \alpha_i + \sum_j^J \frac{n_{ij}}{n_{i+}} (\beta_j + \gamma_{ij})$ égaux
Type II	$H : \alpha_i + \sum_j^J \frac{n_{+j}}{n} (\beta_j + \gamma_{ij})$ égaux
Type III	$H : \alpha_i + \sum_j^J \frac{1}{J} (\beta_j + \gamma_{ij})$ égaux

On peut remarquer que le test de α basé sur la somme de carrés de Type III correspond au test de l'égalité des moyennes ajustées, alors que celui basé sur le Type I correspond plutôt au test de l'égalité des moyennes classiques (cf paragraphe suivant). Ainsi en Type I, quand le dispositif n'est pas équilibré, on teste un effet sans s'ajuster aux autres effets. La conséquence est que l'on peut conclure qu'il existe un effet significatif d'un des effets principaux alors qu'il peut être uniquement du à l'effet des autres facteurs.

Notons que si le dispositif est équilibré alors les trois tests sont les mêmes.

4.4 Moyennes ajustées

Nous donnons tout d'abord la définition des moyennes ajustées pour un modèle d'analyse de la variance à deux facteurs avec interaction qui s'écrit donc :

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

où n_{ij} est le nombre de répétitions pour la combinaison (i, j) des facteurs 1 et 2.

Définition 13 La moyenne ajustée pour le niveau i du premier facteur dans un modèle d'analyse

de la variance à deux facteurs avec interaction est donnée par

$$\begin{aligned}\tilde{\mu}_i &= \frac{1}{J} \sum_{j=1, J} \mathbb{E}(Y_{ijk}) \\ &= \mu + \alpha_i + \frac{1}{J} \sum_{j=1, J} \beta_j + \frac{1}{J} \sum_{j=1, J} \gamma_{ij}\end{aligned}$$

où J est le nombre de niveaux du second facteur. Son estimation est $\hat{\tilde{\mu}}_i = \hat{\mu} + \hat{\alpha}_i + \frac{1}{J} \sum_{j=1, J} \hat{\beta}_j + \frac{1}{J} \sum_{j=1, J} \hat{\gamma}_{ij}$.

Remarquons que l'estimation de la moyenne simple vaut

$$\begin{aligned}\hat{\mu}_i &= \frac{1}{n_{i+}} \sum_{jk} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}) \\ &= \hat{\mu} + \hat{\alpha}_i + \frac{1}{n_{i+}} \sum_{jk} (\hat{\beta}_j + \hat{\gamma}_{ij}) \\ &= \hat{\mu} + \hat{\alpha}_i + \sum_j \frac{n_{ij}}{n_{i+}} (\hat{\beta}_j + \hat{\gamma}_{ij})\end{aligned}$$

Ainsi on observe que les 2 moyennes ne sont pas calculées avec les mêmes pondérations. La moyenne ajustée représente la moyenne que l'on aurait obtenue si le dispositif avait été équilibré. Ces deux moyennes ne sont d'ailleurs identiques que si le plan est équilibré.

Pour un modèle d'analyse de la covariance, qui s'écrit

$$Y_{ij} = \mu + \alpha_i + bx_{ij} + \gamma_i x_{ij} + E_{ij}$$

la moyenne ajustée s'écrit de la façon suivante

Définition 14 La moyenne ajustée pour le groupe i est donnée par

$$\tilde{\mu}_i = \mu + \alpha_i + bx_{\bullet\bullet} + \gamma_i x_{\bullet\bullet}$$

où $x_{\bullet\bullet}$ représente la moyenne générale de la covariable. Son estimation est obtenue en utilisant les estimations des différents paramètres. Elle représente la prédiction de y pour le groupe i pour une valeur moyenne de la covariable $x_{\bullet\bullet}$.

De la même façon que pour le modèle précédent, si on s'intéresse à la moyenne classique, on a

$$\begin{aligned}\hat{\mu}_i &= \frac{1}{n_{i+}} \sum_j (\hat{\mu} + \hat{\alpha}_i + \hat{b}x_{ij} + \hat{\gamma}_i x_{ij}) \\ &= \hat{\mu} + \hat{\alpha}_i + (\hat{b} + \hat{\gamma}_i) \frac{1}{n_{i+}} \sum_j x_{ij} \\ &= \hat{\mu} + \hat{\alpha}_i + (\hat{b} + \hat{\gamma}_i) x_{i\bullet}\end{aligned}$$

C'est donc la prédiction de y pour le groupe i pour une valeur moyenne de la covariable qui est la moyenne sur le groupe i considéré.

Il ne faut pas oublier que, dans des dispositifs non orthogonaux, les moyennes classiques ne sont pas comparables et qu'il faut utiliser les moyennes ajustées.

Chapitre 5

Limites et extensions du modèle linéaire

Les résultats obtenus avec le modèle linéaire sont valides si les suppositions qui ont été faites le sont :

1. $\mathbb{E}(Y) = X\theta$ qui est équivalent à $\mathbb{E}(E) = 0$,
2. $\mathbb{V}(E_i) = \sigma^2, \forall i = 1, n$,
3. les variables aléatoires E_i sont indépendantes,
4. $E_i \sim \mathcal{N}(0, \sigma^2), \forall i = 1, n$.

Le modèle linéaire tolère plus ou moins bien un écart sur chacun des points énoncés ci-dessus.

- Un écart modéré de E à la loi normale a peu d'impact sur la qualité des résultats, d'autant moins que la “vraie” distribution des variables résiduelles est symétrique.
- En analyse de la variance l'inégalité de la variance résiduelle pour différents niveaux d'un facteur peut avoir des conséquences importantes si le dispositif n'est pas équilibré. Si le dispositif est équilibré, les résultats obtenus seront convenables.
- La supposition d'indépendance des variables aléatoires résiduelles est importante. Les résultats de l'estimation et des tests y sont sensibles, surtout si les corrélations sont élevées.

Le premier problème qui se pose est de savoir diagnostiquer si une ou plusieurs suppositions ne sont pas satisfaites. Le deuxième problème est de savoir quoi faire lorsque le diagnostic a été établi. Le diagnostic peut orienter l'utilisateur vers d'autres types de modèles plus généraux que le modèle linéaire.

5.1 Analyse des résidus

Toutes les suppositions du modèle linéaire concernent les variables aléatoires résiduelles, E_i . Elles ne sont malheureusement pas observables. En revanche, on peut estimer les valeurs qu'elles prennent dans l'expérience étudiée, ce sont les résidus, $\hat{E}_i = Y_i - \hat{Y}_i$. C'est l'analyse graphique ou inférentielle de ces résidus qui permet, a posteriori, de remettre en cause certaines des hypothèses.

1. Pour valider $\mathbb{E}(E_i) = 0, \forall i = 1, n$, on fait les graphiques des résidus (en ordonnée) versus chaque variable explicative (en abscisse) ou a minima versus \hat{Y} . Si les résidus comportent une tendance apparente sur ce graphique, on peut mettre en question cette supposition,

voir [2] page 113. Quelquefois il suffit de transformer une variable explicative par une fonction (\log , $\sqrt{\cdot}$, x^2 ...) ou d'ajouter cette variable transformée pour résoudre le problème. On peut aussi faire un test statistique, voir [5].

2. Pour valider $\mathbb{V}(E_i) = \sigma^2$, $\forall i = 1, n$, on fait les mêmes graphiques que dans le cas précédent. Si la variabilité de l'erreur ne semble pas constante, des transformations de Y (comme $\log(Y)$ ou \sqrt{Y}) peuvent parfois stabiliser la variance, voir [2], page 114. On peut aussi faire des tests d'égalité de variance entre plusieurs traitements ([http : //en.wikipedia.org/wiki/Bartlett's_test](http://en.wikipedia.org/wiki/Bartlett's_test)).
3. Pour tester l'indépendance des E_i , il existe plusieurs tests dont le test de Durbin-Watson, [http : //en.wikipedia.org/wiki/Durbin - Watson_statistic](http://en.wikipedia.org/wiki/Durbin-Watson_statistic).
4. Il existe plusieurs tests de l'hypothèse $E_i \sim \mathcal{N}(0, \sigma^2)$, $\forall i = 1, n$:
 - Shapiro-Wilks ([http : //en.wikipedia.org/wiki/Shapiro - Wilk_test](http://en.wikipedia.org/wiki/Shapiro-Wilk_test)),
 - Kolmogorov-Smirnov ([http : //en.wikipedia.org/wiki/Kolmogorov - Smirnov_test](http://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test)),
 - test du χ^2 ([http : //en.wikipedia.org/wiki/Chi - square_test](http://en.wikipedia.org/wiki/Chi-square_test)).

L'analyse graphique joue un grand rôle dans le diagnostic recherché, car les tests sont souvent ciblés sur une hypothèse alternative particulière et peuvent difficilement prendre en compte la diversité des situations possibles.

5.2 Généralisations du modèle linéaire

5.2.1 Modèles non linéaires

Certains phénomènes sont fondamentalement non linéaires dans un domaine de variation déterminé. On peut toujours restreindre le domaine d'étude suffisamment pour considérer les variations comme linéaires localement. Par exemple la croissance d'un organisme ou d'une population, globalement n'est pas linéaire. On peut la modéliser par un modèle linéaire en approchant la fonction non linéaire par un développement limité de degré suffisant, ce qui revient à adopter un modèle de régression polynomiale, mais ce type de modélisation a l'inconvénient d'introduire un nombre important de paramètres qui, de plus, ne sont généralement pas interprétables biologiquement. Certains modèles non linéaires (fonction exponentielle par exemple) se ramènent à un modèle linéaire par une transformation de Y (\log). Mais ce n'est pas toujours faisable comme on le voit si la fonction non linéaire est une somme de 2 exponentielles.

Il est plus adapté dans ce cas de choisir de modéliser l'espérance de la variable à expliquer par une fonction non linéaire. On a alors un modèle "non linéaire" (*proc nlin* de SAS).

5.2.2 Modèle linéaire général

Dans le modèle $y_i = m_i + e_i$ on peut avoir $\mathbb{V}(e_i) = \lambda_i \sigma^2$ où λ_i est connu pour tout $i = 1, n$. C'est le cas quand y_i est la moyenne de n_i données individuelles. Dans ce cas $\lambda_i = 1/n_i$. L'analyse doit prendre en compte les poids différents accordés à chaque observation (instruction "weight" dans *proc glm* de SAS).

5.2.3 Modèle linéaire généralisé

Le modèle est le suivant : $\mathbb{E}(g(Y)) = X\theta$ et $Y \sim G(X\theta, \sigma^2)$, où g est une fonction monotone appelée fonction de lien et G est une loi de probabilité à choisir parmi les lois normales, de

Poisson, Binomiale ou Gamma. Ce modèle permet de prendre en compte une variable réponse discrète (données de comptage ou données 0/1) et d'autres lois de probabilité que la loi normale, voir [1]. La procédure de SAS utilisée est *proc genmod*. La *régression logistique* (*proc logistic* de SAS) est un cas particulier de modèle linéaire généralisé très utile : c'est l'équivalent de la régression multiple pour le cas où la variable réponse est binaire (0/1).

5.3 Les variables explicatives sont aléatoires

5.3.1 Cas de la régression

Le modèle linéaire peut aussi être utilisé pour expliquer la variable aléatoire Y en fonction de variables X dont les valeurs ne sont pas fixées et qui sont modélisées par des variables aléatoires. Par exemple, on étudie la relation entre la qualité du vin et la température cumulée de l'année, qui n'est pas fixée mais subie par l'expérimentateur. On utilise le modèle linéaire usuel, conditionnellement aux réalisations des variables aléatoires obtenues dans l'expérience. Si le vecteur (Y, X) est gaussien les estimateurs et les risques de première espèce des tests usuels restent valides mais la puissance de ces derniers est modifiée.

5.3.2 Modèles mixtes

En génétique quantitative, par exemple, il n'est pas raisonnable de supposer que les effets de certains facteurs, comme l'effet de la mère sur les performances des descendants, est fixe. Il est supposé aléatoire, car cet effet peut être assimilé à un tirage aléatoire dans le génome de la mère. De façon plus générale, on considère que les effets d'un facteur sont aléatoires si les niveaux de ce facteur ne sont pas choisis, mais tirés aléatoirement dans un ensemble de niveaux.

Si les effets de tous les facteurs sont aléatoires, le modèle s'appelle "modèle à effets aléatoires", si seulement certains le sont, on parle de "modèle mixte" (*proc mixed* de SAS).

Ce modèle est de plus en plus utilisé, bien au delà de la génétique quantitative car il peut aussi prendre en compte des corrélations entre les observations. C'est le cas des mesures répétées sur les mêmes individus à des instants différents. Si le nombre d'instant de mesure sur un individu est faible, on utilise le modèle mixte pour prendre en compte la dépendance entre observations issues du même individu. En revanche si le nombre d'instant de mesure n'est pas faible et si l'évolution dans le temps est le sujet d'intérêt, alors il faudra utiliser une autre catégorie de modèles, les modèles de "séries chronologiques", (*proc autoreg* de SAS).

La corrélation des variables aléatoires résiduelles peut venir aussi d'implantations dans l'espace. C'est le cas par exemple des données mesurées sur des arbres voisins dans une parcelle forestière ou de mesures de pollution en différents points de l'espace.

Chapitre 6

Annexes

6.1 Loi de $(\hat{\theta}_1 - a)/\sqrt{\widehat{\mathbb{V}}(\hat{\theta}_1)}$

6.1.1 Loi sous H_0

Sous l'hypothèse $H_0 = \{\theta_1 = a\}$, $\hat{\theta}_1 - a \sim \mathcal{N}(0, \mathbb{V}(\hat{\theta}_1))$ et donc

$$\frac{\hat{\theta}_1 - a}{\sqrt{\mathbb{V}(\hat{\theta}_1)}} \sim \mathcal{N}(0, 1).$$

Quand on remplace $\mathbb{V}(\hat{\theta}_1)$ par $\widehat{\mathbb{V}}(\hat{\theta}_1)$, la loi $\mathcal{N}(0, 1)$ devient une loi de Student dont le nombre de degrés de liberté, ν_1 , est celui de la somme de carrés qui sert à estimer la variance σ^2 .

6.1.2 Loi sous H_1

Sous l'hypothèse $H_1 = \{\theta_1 \neq a\}$, $\hat{\theta}_1 - a \sim \mathcal{N}(\theta_1 - a, \mathbb{V}(\hat{\theta}_1))$ et donc

$$\frac{\hat{\theta}_1 - a}{\sqrt{\mathbb{V}(\hat{\theta}_1)}} \sim \mathcal{N}(\theta_1 - a, 1).$$

Quand on remplace $\mathbb{V}(\hat{\theta}_1)$ par $\widehat{\mathbb{V}}(\hat{\theta}_1)$, la loi $\mathcal{N}(\theta_1 - a, 1)$ devient une loi de Student **décentrée** voir ([HTTP://MATHWORLD.WOLFRAM.COM/NONCENTRALSTUDENTST-DISTRIBUTION.HTML](http://mathworld.wolfram.com/NoncentralStudentt-Distribution.html)) dont le nombre de degrés de liberté est ν_1 et le paramètre de décentration est $\theta_1 - a$. Ceci permet de calculer la puissance du test, mais la loi de Student décentrée n'est pas facile à manipuler.

Le cas de $(c\hat{\theta} - a)/\sqrt{c\widehat{\mathbb{V}}(\hat{\theta})c'}$ est similaire au cas précédent.

6.2 Loi de $\frac{(SCM_1 - SCM_0)/(p_1 - p_0)}{SCR_1/\nu_1}$

6.2.1 Loi sous H_0

La construction mathématique de ce test est basée sur le théorème de Cochran (6.3) qui montre que sous H_0 , $SCM_1 - SCM_0$ et SCR_1 sont des variables aléatoires indépendantes et

distribuées selon des lois du χ^2 à respectivement $p_1 - p_0$ et ν_1 degrés de liberté. Par suite la statistique de test donnée par (3.4) est distribuée selon une loi de Fisher, voir [2], p 141, $\mathcal{F}(p_1 - p_0, \nu_1)$ sous H_0 .

6.2.2 Loi sous H_1

$SCM_1 - SCM_0$ est une somme de carrés de lois normales non centrées. Elle est distribuée selon une loi du χ^2 décentrée (cf [HTTP://EN.WIKIPEDIA.ORG/WIKI/NONCENTRAL_CHI-SQUARE_DISTRIBUTION](http://en.wikipedia.org/wiki/Noncentral_chi-square_distribution)), $\chi^2(p_1 - p_0, \lambda)$, où $p_1 - p_0$ est le nombre de degrés de liberté et λ , qui est égal à la somme des carrés des paramètres de θ qui ont été annulés dans θ_0 , divisée par σ^2 , est le paramètre de décentrement. Par suite la statistique (3.4) est distribuée selon une loi de Fisher décentrée (cf [HTTP://EN.WIKIPEDIA.ORG/WIKI/NONCENTRAL_F-DISTRIBUTION](http://en.wikipedia.org/wiki/Noncentral_f_distribution)), $F(p_1 - p_0, \nu_1, \lambda)$ sous H_1 . Cette distribution permet de calculer la puissance du test.

6.3 Théorème de Cochran

Theorème 6 *On considère n variables aléatoires, Y_1, Y_2, \dots, Y_n , indépendantes et de même loi normale $\mathcal{N}(0, 1)$. Soit $Y = (Y_1, Y_2, \dots, Y_n)$ et $\|Y\|^2 = \sum_i Y_i^2$. On sait, [2], p 140, que $\|Y\|^2 \sim \chi_n^2$. Soit L un sous espace de \mathbb{R}^n , de dimension r . On note $P_L(Y)$ le projeté orthogonal de Y sur L . On a la décomposition orthogonale suivante :*

$$Y = P_L(Y) \oplus (Y - P_L(Y)).$$

Alors les variables aléatoires $\|P_L(Y)\|^2$ et $\|Y - P_L(Y)\|^2$ sont indépendantes et $\|P_L(Y)\|^2 \sim \chi_r^2$ et $\|Y - P_L(Y)\|^2 \sim \chi_{n-r}^2$.

6.3.1 Lemme 1

Soit $u = (u_1, \dots, u_n)$ et $v = (v_1, \dots, v_n)$ deux vecteurs unitaires de \mathbb{R}^n , on note le produit scalaire $\langle u, v \rangle = \sum_{i=1}^n u_i v_i$. On note $P_u(Y)$ le projeté orthogonal de Y sur la droite vectorielle portée par u . Ce vecteur aléatoire est colinéaire à u , et on note $c_u(Y)$ le coefficient de linéarité : $P_u(Y) = c_u(Y)u$, et $c_u(Y) = \langle Y, u \rangle = \sum_{i=1}^n u_i Y_i$. On a alors

$$c_u(Y) \sim \mathcal{N}(0, 1) \tag{6.1}$$

En effet $c_u(Y)$ est une variable aléatoire qui suit une loi normale (combinaison de lois normales indépendantes) et vérifie $\mathbb{E}(c_u(Y)) = \sum_{i=1}^n u_i \mathbb{E}(Y_i) = 0$ et $\mathbb{V}(c_u(Y)) = \sum_{i=1}^n u_i^2 \mathbb{V}(Y_i) = 1$.

6.3.2 Lemme 2

$$\text{Cov}(c_u(Y), c_v(Y)) = 0 \Leftrightarrow \langle u, v \rangle = 0$$

Preuve

$$\begin{aligned}\text{Cov}(c_u(Y), c_v(Y)) &= \text{Cov}\left(\sum_{i=1}^n u_i Y_i, \sum_{i=1}^n v_i Y_i\right) \\ &= \mathbb{E}\left(\sum_{i=1}^n u_i Y_i\right)\left(\sum_{i=1}^n v_i Y_i\right) \\ &= \sum_{i,j} \mathbb{E}(Y_i Y_j) u_i v_j \\ &= \sum_{i=1}^n u_i v_i \\ &= \langle u, v \rangle\end{aligned}$$

Car les variables Y_i sont i.i.d et suivent des lois normales centrées et réduites.

6.3.3 Démonstration du théorème de Cochran

Première étape

On construit une base orthonormée (l_1, \dots, l_r) du sous-espace L (ceci est toujours possible grâce au procédé d'orthonormalisation de Gram-Schmitt). Le projeté orthogonal de Y sur L est la somme des projetés orthogonaux sur les axes l_i :

$$P_L(Y) = \sum_{i=1}^r P_{l_i}(Y) = \sum_{i=1}^r c_{l_i}(Y) l_i$$

Où $c_{l_i}(Y) \sim N(0, 1)$, $i = 1 \dots r$

Les l_i formant un système orthonormé, les $c_{l_i}(Y)$ sont non corrélés donc indépendants grâce à la normalité du vecteur aléatoire Y . Par conséquent :

$$\|P_L(Y)\|^2 = \sum_{i=1}^r \|P_{l_i}(Y)\|^2 = \sum_{i=1}^r (c_{l_i}(Y))^2$$

Donc $\|P_L(Y)\|^2 \sim \chi^2(r)$.

Deuxième étape

On complète la base orthonormée (l_1, \dots, l_r) du sous-espace L en une base (l_1, \dots, l_n) de \mathbb{R}^n et on projette sur le sous-espace F engendré par (l_{r+1}, \dots, l_n) . On obtient $Y = P_L(Y) + P_F(Y)$, donc $P_F(Y) = Y - P_L(Y)$ d'où $\|Y - P_L(Y)\|^2 \sim \chi^2(n - r)$
L'orthogonalité des sous espaces L et F permet d'obtenir l'indépendance des vecteurs $P_L(Y)$ et $Y - P_L(Y)$ et donc de leurs normes.

6.4 Espérance et matrice de variance-covariance de $\hat{\theta}$

6.4.1 Espérance de $\hat{\theta}$

Cas $r = p + 1$. L'estimateur des paramètres θ est :

$$\hat{\theta} = (X'X)^{-1} X'Y$$

Les seuls termes aléatoires de cette expression étant Y , on a :

$$\mathbb{E}(\hat{\theta}) = (X'X)^{-1}X'\mathbb{E}(Y) = (X'X)^{-1}X'X\theta = \theta$$

Cas $r < p + 1$. L'estimateur s'écrit :

$$\hat{\theta} = (G'G)^{-1}X'Y$$

et son espérance est égale à :

$$\mathbb{E}(\hat{\theta}) = (G'G)^{-1}X'\mathbb{E}(Y) = (G'G)^{-1}(X'X + H'H)\theta = \theta$$

puisque $G'G = X'X + H'H$ et $H\theta = 0$.

6.4.2 Matrice de variance-covariance de $\hat{\theta}$

Rappelons tout d'abord la propriété suivante :

Propriété 2 Soit A et B deux matrices de dimension $p \times n$ de coefficients réels, soit X un vecteur colonne aléatoire de dimension n , $X' = (X_1, \dots, X_n)$, alors la matrice de variance-covariance de AX est :

$$\mathbb{V}(AX) = A\mathbb{V}(X)A' \quad \text{et} \quad \text{Cov}(AX, BX) = A\text{Cov}(X, X)B' = A\mathbb{V}(X)B'$$

Cas $r = p + 1$. D'après la forme de l'estimateur donné ci-dessus et en utilisant la propriété précédente avec $A = (X'X)^{-1}X'$, on obtient que

$$\mathbb{V}(\hat{\theta}) = (X'X)^{-1}X'\mathbb{V}(Y)((X'X)^{-1}X')' = (X'X)^{-1}X'\mathbb{V}(Y)X(X'X)^{-1},$$

Or on a supposé que $\mathbb{V}(Y) = \sigma^2 I_n$, donc

$$\mathbb{V}(\hat{\theta}) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

Cas $r < p + 1$. De la même façon, on montre facilement que

$$\mathbb{V}(\hat{\theta}) = \sigma^2(X'X)^-X'X(X'X)^-.$$

Si on choisit une inverse généralisée réflexive, c'est-à-dire qui satisfait aussi l'égalité :

$$(X'X)^-(X'X)(X'X)^- = (X'X)^-$$

Alors :

$$\mathbb{V}(\hat{\theta}) = \sigma^2(X'X)^-,$$

qui permet d'avoir pour cette matrice de covariance une forme analogue à celle du cas de plein rang.

Attention : les règles de calcul sur les inverses généralisées ne sont pas identiques à celles des inverses. Par définition, $(X'X)^-$ est une inverse généralisée de $X'X$ si :

$$(X'X)(X'X)^-(X'X) = (X'X)$$

En général, $(X'X)^-(X'X)(X'X)^- \neq (X'X)^-$, sauf pour les inverses généralisées réflexives.

6.5 Espérance de la somme des carrés résiduelle

Par définition :

$$SCR = (Y - \hat{Y})'(Y - \hat{Y}) = (Y - PY)'(Y - PY) = Y'(\mathbf{I}_n - P)Y,$$

où P est le projecteur orthogonal sur le sous-espace $L(X)$, et \mathbf{I}_n est la matrice identité d'ordre n . Le résultat a été obtenu en utilisant les propriétés d'idempotence et de symétrie de P et de \mathbf{I}_n ($P^2 = P$ et $P' = P$).

L'espérance de la somme des carrés résiduelle est donc :

$$\mathbb{E}(SCR) = \mathbb{E} [Y'(\mathbf{I}_n - P)Y].$$

La matrice $\mathbb{E}(SCR)$ étant de dimension 1×1 , est égale à sa trace, ce qui permet d'utiliser la propriété de commutativité de la trace d'un produit de matrices. De plus les opérateurs trace et espérance sont commutatifs ce qui donne :

$$\begin{aligned} \mathbb{E}(SCR) &= \text{trace} [\mathbb{E}(SCR)] \\ &= \mathbb{E} [\text{trace} (Y'(\mathbf{I}_n - P)Y)] \\ &= \mathbb{E} [\text{trace} ((\mathbf{I}_n - P)YY')] \\ &= \text{trace} [(\mathbf{I}_n - P)\mathbb{E}(YY')] \\ &= \text{trace} [(\mathbf{I}_n - P)(\sigma^2\mathbf{I}_n + X\theta\theta'X')] \end{aligned}$$

\mathbf{P} étant le projecteur orthogonal sur le sous-espace $L(X)$ de \mathbb{R}^n , le produit $(\mathbf{I}_n - P)X\theta$ est nul puisque $X\theta$ appartient à $L(X)$. On a donc aussi que $PX\theta = X\theta$. On obtient alors

$$\mathbb{E}(SCR) = \sigma^2 \text{trace} [(\mathbf{I}_n - P)] = \sigma^2 [n - \text{trace}P].$$

On sait que la trace d'un projecteur orthogonal est égale à la dimension de l'espace sur lequel il projette : ici, c'est le sous-espace $L(X)$ dont la dimension est égale au rang r de la matrice X . On trouve donc le résultat annoncé :

$$\mathbb{E}(SCR) = \sigma^2 (n - r).$$

Si la matrice X est de plein rang, $r = p + 1$, sinon $r < p + 1$.

Ce résultat peut aussi être obtenu comme une conséquence immédiate du théorème de Cochran 6.3.

6.6 Démonstration du théorème de Gauss-Markov

On pourra se référer à [8].

Plaçons nous dans le cas où $r = p + 1$. L'estimateur des moindres carrés est $\hat{\theta} = (X'X)^{-1}X'Y$. Tout d'abord, on voit que $\hat{\theta}$ est un estimateur linéaire de Y puisqu'il existe C tel que $\hat{\theta} = CY$.

De plus, il est sans biais d'après ce qui précède (cf paragraphe 6.4).

Il nous faut donc montrer maintenant que, pour tout autre estimateur $\tilde{\theta}$ de θ linéaire et sans biais, $\mathbb{V}(\tilde{\theta}) \geq \mathbb{V}(\hat{\theta})$. On écrit la variance de $\tilde{\theta}$ en faisant apparaître la variance de $\hat{\theta}$:

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}(\tilde{\theta} + \hat{\theta} - \hat{\theta}) = \mathbb{V}(\tilde{\theta} - \hat{\theta}) + \mathbb{V}(\hat{\theta}) - 2 \mathbb{Cov}(\tilde{\theta} - \hat{\theta}, \hat{\theta}).$$

Par définition d'une variance, elle est toujours définie positive. Calculons le dernier terme $\mathbb{Cov}(\tilde{\theta} - \hat{\theta}, \hat{\theta})$. On suppose que $\tilde{\theta}$ est linéaire en Y donc qu'il existe A tel que $\tilde{\theta} = AY$.

De plus, on suppose qu'il est sans biais, i.e. $\mathbb{E}(\tilde{\theta}) = \theta$. Or

$$\mathbb{E}(\tilde{\theta}) = \mathbb{E}(AY) = A\mathbb{E}(Y) = AX\theta,$$

Donc $\mathbb{E}(\tilde{\theta}) = \theta$ que si $AX = \mathbf{I}$.

La covariance de $\tilde{\theta} - \hat{\theta}$ avec $\hat{\theta}$ vaut alors

$$\begin{aligned} \mathbb{Cov}(\tilde{\theta} - \hat{\theta}, \hat{\theta}) &= \mathbb{Cov}(AY, (X'X)^{-1}X'Y) - \mathbb{V}(\hat{\theta}) \\ &= A\mathbb{V}(Y)((X'X)^{-1}X')' - \mathbb{V}(\hat{\theta}) \end{aligned}$$

D'après le calcul de la matrice de variance-covariance de $\hat{\theta}$ (cf paragraphe 6.4),

$$\begin{aligned} \mathbb{Cov}(\tilde{\theta} - \hat{\theta}, \hat{\theta}) &= \sigma^2 AX(X'X)^{-1} - \sigma^2(X'X)^{-1} \\ &= 0 \end{aligned}$$

Ce qui conclut la preuve.

Bibliographie

- [1] An Introduction to generalized linear model, **A. J. Dobson**, Chapman and Hall/CRC 2^e ed, 2001
- [2] Statistique inférentielle, **J. J. Daudin, S. Robin et C. Vuillet**, PUR Paris 2001
- [3] Le Modèle linéaire, **C. DUBY**, Polycopié INA P-G Paris 2000
- [4] Advanced Linear Models - Theory and Applications, **Song-Gui Wang & Shein-Chung Chow**, Marcel Dekker, Inc. 1994
- [5] A Simple Goodness-of-Fit Test for Parametric Regression Models, **JJ Daudin & A Barhen**, Communications in Statistics : Theory and Methods, Volume 33, Number 6/2004.
- [6] Exemples d'application du modèle linéaire, **E. Lebarbier & S. Robin**, Polycopié AgroParisTech Paris 2007.
- [7] Pratique des plans d'expériences, **P. Schimmerling, J.-C. Sisson & A. Zaidi**, Lavoisier, Paris, 1998.
- [8] Régression, Théorie et applications, **P-A. Cornillon & E. Matzner-Lober**, Springer, 2007.