UNIVERSITE
PAUL
SABATIER
TOULOUSE III

Publications du Laboratoire
de
Statistique et Probabilités

Le Modèle Linéaire par l'exemple :

Régression, Analyse de la Variance,...

Jean-Marc Azaïs et Jean-Marc Bardet

Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Université Paul Sabatier — 31062 – Toulouse cedex 4.

Chapitre 1

Introduction

Ce texte est issu d'un texte plus réduit d'une trentaine de pages rédigé à l'intention de la formation continue de l'INRA. La version actuelle a été considérablement étoffée à partir de mes textes d'enseignement à l'Université Paul Sabatier. Pour garder la concision du premier exposé, nous avons ajouté beaucoup de compléments sous forme de problèmes résolus. Ces compléments sont parfois d'un caractère mathématique plus marqué.

Ce document veut être une pratique du modèle linéaire plutôt qu'une théorie, ceci nous a conduit à privilégier certains aspects. Notre principal souci a été, à travers des exemples et par l'étude des graphiques de résidus et des transformations de modèles, d'essayer de décrire les problèmes courants qui peuvent ou ne peuvent pas être analysés par un modèle linéaire.

Pour simplifier les calculs et limiter le nombre de formules au maximum nous avons toujours privilégié l'approche géométrique. Nous n'avons pas cherché à détailler toutes les décompositions possibles. Notre philosophie est de laisser les calculs à l'ordinateur : peu importe donc l'expression explicite de telle ou telle somme de carrés, ce qui importe par contre c'est de posséder les principes généraux des tests et des estimations pour pouvoir en faire une interprétation correcte.

Un document aussi court imposait des choix. Nous avons donc laissé en retrait les notions d'orthogonalité et de modèles non réguliers. Ceci pour deux raisons différentes :

- l'orthogonalité nous paraît une belle notion théorique. Cependant, du fait de la présence de données manquantes, il est bien rare qu'une expérience soit réellement orthogonale. On ne peut donc se limiter au cas orthogonal et de ce fait il est plus simple et plus court de ne pas l'étudier du tout.
- À l'exception du modèle additif d'analyse de la variance à deux facteurs qui ne doit être utilisé que dans certain cas, il est possible de traiter les autres situations à l'aide de modèles réguliers. Nous avons donc très peu abordé la non régularité.

Enfin ce texte s'inscrit en faux contre une idée relativement répandue selon laquelle le modèle linéaire serait une méthode gaussienne. A notre sens, l'hypothèse de normalité ne sert qu'à simplifier la démonstration. Nous avons donc inclus des résultats asymptotiques qui montrent, que pour les grands échantillons, l'hypothèse de normalité n'est pas nécessaire. Ces résultats plus délicats à appréhender nous semblent d'une conséquence pratique importante et donc tout à fait indispensables dans un cours de modèle linéaire.

Chapitre 2

Exemples Simples

Dans ce chapitre nous rappelons brièvement les formules de la régression simple et de l'analyse de la variance à un facteur. Notre but et de faire ressortir la similitude des hypothèses et des méthodes. Il est donc nécessaire de les traiter en détail dans un même cadre. C'est ce qui sera fait au chapitre 4. En attendant cette étude globale nous allons donner de nombreux résultats sans justification.

1 Régression linéaire simple

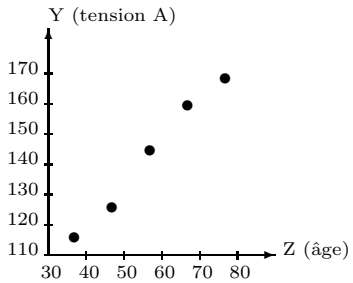
1.1 Exemple

On considére 5 groupes de femmes âgées respectivement de 35, 45, 55, 65 et 75 ans. Dans chaque groupe, on a mesuré la tension artérielle en mm de mercure de chaque femme et on a calculé la valeur moyenne pour chaque groupe. On définit donc les variables :

Y : tension moyenne en mm Hg	114	124	143	158	166
Z : âge du groupe considéré	35	45	55	65	75

(source de ses données Cramer)

Afin de visualiser ces données, on fait une représentation cartésienne :



Commentaires : Sur le graphique on constate que

- la tension artérielle augmente avec l'âge (résultat bien classique).
- mais surtout cette augmentation semble linéaire puisque les points du graphique sont presque alignés.

1.2 Modèle et estimation

Notons Y_i la tension artérielle du $i$ème groupe et Z_i son âge, nous pouvons alors proposer le modèle suivant :

$$Y_i = \mu + \beta Z_i + \varepsilon_i. \tag{2.1}$$

C'est un modèle de dépendance linéaire. Il y a deux paramètres, μ est la constante, β est la pente, et ils sont tous les deux inconnus. Le vecteur aléatoire ε formé par les variables aléatoires ε_i, est appelé l'erreur du modèle.

Plus généralement, supposons que l'on a n observations connue (dans l'exemple, $n = 5$) d'une variable Y appelée variable à expliquer (dans l'exemple, la tension artérielle) et d'une variable Z dite explicative (dans l'exemple, l'âge). On supposera de plus que pour $i = 1, \cdots, n$, les variables Y_i et Z_i suivent le modèle (2.1) et que

- pour $i = 1, \cdots, n$, $\mathrm{E}(\varepsilon_i) = 0$ (les erreurs sont centrées) ;
- pour $i = 1, \cdots, n$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (ne dépend pas de i) et nous supposerons σ^2 inconnue ;
- pour $i = 1, \cdots, n$, les variables ε_i sont indépendantes et de loi gaussienne (dite encore loi normale).

Ces postulats seront commentés un plus loin.

Remarque : Par abus de notation, on désignera aussi bien par Y la variable statistique que le vecteur $(Y_i)_{1 \leq i \leq n}$. Le contexte permettra en général de distinguer entre les deux cas.

Pour déterminer les inconnues μ et β une méthode possible est la méthode dite des moindres carrés. On cherche des valeurs m et b minimisant :

$$SC(m, b) := \sum_{i=1}^{n} [Y_i - (m + bZ_i)]^2.$$

Cela revient à minimiser les carrés des écarts pris verticalement entre la droite de paramètres m et b et les différents points observés. La solution bien connue de ce problème (on peut la retrouver en dérivant $SC(m, b)$ en m et en b) est $(\widehat{\mu}, \widehat{\beta})$ avec :

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum_{i=1}^{n} (Z_i - \overline{Z})^2}$$

$$\widehat{\mu} = \overline{Y} - b\overline{Z}$$

où $\overline{Z} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Z_i$ est la moyenne des Z_i et $\overline{Y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_i$ la moyenne des Y_i. On définit également :

- le vecteur des valeurs estimées $\widehat{Y} = \widehat{\mu} + \widehat{\beta} Z = (\widehat{\mu} + \widehat{\beta} Z_i)_{1 \leq i \leq n}$;
- le vecteur des résidus $\widehat{\varepsilon} = Y - \widehat{Y} = (Y_i - \widehat{\mu} - \widehat{\beta} Z_i)_{1 \leq i \leq n}$;
- l'estimateur de la variance $s^2 = \widehat{\sigma}^2 = \dfrac{1}{n-2} \displaystyle\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$. Le coefficient $n - 2$ peut s'expliquer par la règle : nombre de données (ici n) moins le nombre de paramètres du modèle (ici 2).

Remarque : Par la suite et pour aider à la lecture des résultats, si θ est un paramètre inconnu réel ou vectoriel (par exemple σ^2 ou (μ, β)), nous adopterons la convention de noter $\widehat{\theta}$ un estimateur de θ.

En se plaçant dans le cadre du modèle (2.1) et en considérant les X_i comme des données déterministes, les Y_i sont des variables aléatoires gaussiennes. On peut alors apprécier la précision des estimateurs $\widehat{\mu}$ et $\widehat{\beta}$ à l'aide des formules complémentaires suivantes :

- $\mathrm{E}(\widehat{\mu}) = \mu$ et $\mathrm{E}(\widehat{\beta}) = \beta$;
- $\mathrm{Var}(\widehat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$ et $\mathrm{Var}(\widehat{\mu}) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{Z}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \right) = \dfrac{\sigma^2}{n} \dfrac{\sum_{i=1}^n Z_i^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$;
- $\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = -\dfrac{\sigma^2 \, \overline{Z}}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$.

De la même manière, on définit la matrice de variance du vecteur $(\widehat{\mu}, \widehat{\beta})$ et ainsi :

$$\mathrm{Var} \begin{pmatrix} \widehat{\mu} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(\widehat{\mu}) & \mathrm{Cov}(\widehat{\mu}, \widehat{\beta}) \\ \mathrm{Cov}(\widehat{\mu}, \widehat{\beta}) & \mathrm{Var}(\widehat{\beta}) \end{pmatrix} = \frac{\sigma^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n Z_i^2 & -\overline{Z} \\ -\overline{Z} & 1 \end{pmatrix}.$$

1.3 Table d'analyse de la variance

On complète l'étude précédente en construisant la table suivante, encore appelée table d'analyse de la variance :

Source	Somme de carrés	Degrés de liberté	Carré moyen	$\widehat{F}$
régression	$\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$	1	$\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$	
résiduelle	$\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	$n-2$	$\dfrac{1}{n-2} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	$(n-2) \dfrac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}$
totale	$\sum_{i=1}^n (Y_i - \overline{Y})^2$	$n-1$	$\dfrac{1}{n-1} \sum_{i=1}^n (Y_i - \overline{Y})^2$	

Commentaires :

- la statistique $\widehat{F}$ qui permet de tester la nullité de la pente, à savoir $\beta = 0$, est égale au rapport des deux carrés moyens. Pour un test de niveau α (en général $\alpha = 5\%$), on compare la statistique $\widehat{F}$ à la valeur dépassée avec une probabilité α par la loi de Fisher à $(1, n-2)$ degrés de liberté. Cette quantité, notée $F_{1,n-2,\alpha}$ est le quantile d'ordre $(1-\alpha)$ de cette loi de Fisher à $(1, n-2)$ degrés de liberté.

- la somme des carrés résiduelle est le minimum de $SC(m, b)$ minimisée, soit $SC(\widehat{\mu}, \widehat{\beta})$.

- la somme des carrés expliquée par la régression est la quantité expliquée par la droite de régression par rapport au modèle où on n'ajuste qu'une simple moyenne $\overline{Y}$.

- la somme des carrés totale est normalement utilisée pour le calcul de la variance empirique.

1.4 Test de Student

D'après ce qui précède $\widehat{\beta}$ est une variable gaussienne centrée et comme $\mathrm{Var}(\widehat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$, un estimateur naturel de cette variance est $\dfrac{\widehat{\sigma}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$. On montre alors que

$$\widehat{T} = \widehat{\beta} \sqrt{\frac{\sum_{i=1}^n (Z_i - \overline{Z})^2}{\widehat{\sigma}^2}} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 \sum_{i=1}^n (Z_i - \overline{Z})^2}}$$

est une variable aléatoire distribuée selon une loi de Student à $(n-2)$ degrés de liberté (loi notée T_{n-2}). On déduit de ceci qu'un intervalle de confiance de niveau $(1-\alpha)$ sur la valeur de β est

$$\left[\widehat{\beta} - T_{n-2,\alpha/2} \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}} \ , \ \widehat{\beta} + T_{n-2,\alpha/2} \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}} \right]$$

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

Il est également possible de réaliser un test de Student de niveau α sur la nullité de la pente. Plus précisèment, on testera à l'aide de la statistique $\widehat{T}$, l'hypothèse H_0 : $\beta = 0$ contre l'hypothèse H_1 : $\beta \neq 0$. Ce test revient à regarder si la valeur zéro appartient à l'intervalle de confiance ci-dessus. Donc si

$$|\widehat{\beta}| \sqrt{\frac{\sum_{i=1}^n (Z_i - \overline{Z})^2}{\widehat{\sigma}^2}} > T_{n-2,\alpha/2}$$

l'hypothèse H_0 sera rejeté (ce qui est significatif), sinon l'hypothèse H_0 sera acceptée (ce qui n'est pas significatif). Or pour p quelconque, le carré d'une variable de Student à p degrés de liberté est une variable de Fisher à $(1, p)$ degrés de liberté :
Le test de Student est donc strictement le même que le test issu de la table d'analyse de la variance.

2 Analyse de la variance à un facteur

2.1 Exemple

Un forestier s'intéresse aux hauteurs moyennes de trois forêts. Pour les estimer, il échantillonne un certain nombre d'arbres et mesure leurs hauteurs :

forêt1	forêt2	forêt3
23,4	22,5	18,9
24,4	22,9	21,1
24,6	23,7	21,1
24,9	24,0	22,1
25,0	24,0	22,5
26,2		23,5
		24,5
$n_1 = 6$	$n_2 = 5$	$n_3 = 7$

Ces données peuvent être présentées de deux manières équivalentes.

i. On dispose de trois échantillons indépendants et on désire comparer leurs moyennes. C'est la présentation élémentaire dite de "comparaison de moyennes".

ii. On dispose d'un seul échantillon de longueur 18 et d'une variable explicative qualitative, ou facteur, le numéro de la forêt. En prenant ce second point de vue, on parle d'analyse de la variance à 1 facteur. C'est également le point de vue adopté par la plupart des logiciels et il offre l'avantage de s'adapter à des cas compliqués : 2 facteurs, 3 facteurs, etc... Et c'est également le point de vue que nous adopterons le plus souvent.

2.2 Modèle statistique

La méthode de collecte des données, à savoir un échantillonnage indépendant dans chacune des forêts, nous permet de proposer le modèle suivant :

- on note Y_{ij} la hauteur du $j^{\text{ème}}$ arbre de la forêt i ;
- on note μ_i la moyenne de la forêt i (que l'on pourrait théoriquement calculer en recensant tous les arbres qui s'y trouvent).

Dans ce cadre, un modèle possible est le suivant :

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{2.2}$$

où ε_{ij} est la variabilité de l'arbre j par rapport à la moyenne de la forêt i. comme précédemment, nous allons faire quelques hypothèses sur les ε_i :

- on a par définition $\mathrm{E}(\varepsilon_{ij}) = 0$;
- on suppose que la variance est la même dans chaque forêt, soit $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$;
- l'échantillonnage implique que les mesures et donc les ε_{ij} sont indépendants ;

– enfin, si les effectifs sont petits (ce qui est le cas dans cet exemple forestier), on supposera de plus que les Y_{ij} (et donc les ε_{ij}) sont des variables gaussiennes.

La question cruciale (mais ce n'est pas forcément la seule) à laquelle nous voudrions répondre est :

"les forêts sont-elles équivalentes ?"

Cela ce traduit dans notre modèle par le fait que :

$$\mu_1 = \mu_2 = \mu_3.$$

Cette égalité permet de définir un sous-modèle du modèle (2.2). En notant μ la valeur commune de μ_1, μ_2 et μ_3, ce sous-modèle s'écrit

$$Y_{ij} = \mu + \varepsilon_{ij} \tag{2.3}$$

Dans le grand modèle (2.2), on estimera la hauteur de la forêt i par la moyenne empirique de l'échantillon. Ainsi, si n_i est le nombre d'arbres de la forêt i, alors :

$$\widehat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = Y_{i.}$$

Notation : dans tout ce qui suit, nous adopterons la notation suivante : un point à la place d'un indice veut dire la moyenne sur l'indice considéré.

On déduit de ceci que compte tenu du modèle (2.2),

– pour chaque (i, j), la valeur prédite de Y_{ij} est $\widehat{Y}_{ij} = Y_{i.}$

– les résidus (estimation des erreurs ε_{ij}) sont $\widehat{\varepsilon}_{ij} = Y_{ij} - Y_{i.}$

– la somme des carrés résiduelle, c'est-à-dire non prédite par le grand modèle (2.2), est donc

$$\mathrm{SC}_1 = \sum_{i,j} \widehat{\varepsilon}_{ij}^2 = \sum_{i,j} (Y_{ij} - Y_{i.})^2.$$

Dans le sous-modèle (2.3) correspondant à l'hypothèse d'équivalence des forêts, on estime la moyenne commune μ par la moyenne empirique, soit :

$$\widehat{\mu} = \overline{Y} = Y_{..} = \frac{1}{n_1 + n_2 + n_3}\sum_{i,j} Y_{ij}$$

Remarquons que dans le cas d'effectifs inégaux, cette quantité n'est pas égale à la moyenne des $Y_{i.}$. La somme des carrés non prédite par le sous-modèle (2.3), encore appelée somme des carrés totale, est donnée par :

$$\mathrm{SC}_0 = \sum_{i,j} (Y_{ij} - \overline{Y})^2$$

Par une utilisation de l'identité $(a+b)^2 = a^2+2ab+b^2$ ou par le théorème de Huygens, on obtient :

$$\mathrm{SC}_0 - \mathrm{SC}_1 = \sum_{i,j} (Y_{i.} - \overline{Y})^2.$$

En se paçant dans le cadre général de I groupes, d'un effectif n_i pour le groupe i, et d'un nombre total de données $n = n_1 + \cdots + n_I$, on construit la table d'analyse de la variance :

Source	Somme de carrés	Degrés de liberté	Carré moyen	$\widehat{F}$
modèle	$\mathrm{SC}_0 - \mathrm{SC}_1 = \sum_{i,j}(Y_{i.} - Y_{..})^2$	I-1	$\frac{1}{I-1}\sum_{i,j}(Y_{i.} - Y_{..})^2$	
résiduelle	$\mathrm{SC}_1 = \sum_{i,j}(Y_{ij} - Y_{i.})^2$	$n - I$	$\frac{1}{n-I}\sum_{i,j}(Y_{ij} - Y_{i.})^2$	$\frac{n-I}{I-1}\frac{\sum_{i,j}(Y_{i.} - Y_{..})^2}{\sum_{i,j}(Y_{ij} - Y_{i.})^2}$
totale	$\mathrm{SC}_0 = \sum_{i,j}(Y_{ij} - \overline{Y})^2$	$n - 1$	$\frac{1}{n-1}\sum_{i,j}(Y_{ij} - \overline{Y})^2$	

Remarque : remarquons que certaines des doubles sommes présentées pourraient être écrite comme somme simple. Ainsi, $\sum_{i,j}(Y_{i.} - Y_{..})^2 = J.\sum_i(Y_{i.} - Y_{..})^2$. Cependant, on préférera l'écriture avec doubles, triples,..., sommes, qui permet, d'une part, de visuliser une distance dans $\mathbb{R}^n$, et d'autre part, de ne pas retenir les différents nombres de modalités des différents facteurs.

Rappelons que dans l'exemple des forêts, $I = 3$ est le nombre de forêts et $n = 18$ le nombre total de données. Le test de l'hypothèse d'égalité des moyennes (de niveau α) se fait en comparant la valeur du rapport $\widehat{F}$ au quantile $F_{I-1,n-I,\alpha}$ d'une loi de Fisher à $(I-1),(n-I)$ degrés de liberté.

La statistique F peut s'interpréter comme le rapport de la variabilité intergroupe sur la variabilité intragroupe. En effet le carré moyen du modèle mesure l'écart des moyennes des groupes (forêts) à la moyenne générale ; c'est une mesure de variabilité entre les groupes (d'où la dénomination intergroupe). Le carré moyen résiduel mesure l'écart de chaque individu (arbre) à la moyenne du groupe (forêt) auquel il appartient ; c'est une mesure de la variabilité à l'intérieur de chaque groupe(d'où la dénomination intragroupe).

2.3 Intervalle de confiance et test de Student

Si on choisit a priori (c'est-à-dire avant le résultat de l'expérience) deux populations à comparer, par exemple la 1 et la 2, on peut obtenir un intervalle de confiance pour la différence des moyennes théoriques $\mu_1 - \mu_2$. En effet, un estimateur de cette différence est $\widehat{\mu}_1 - \widehat{\mu}_2 = Y_{1.} - Y_{2.}$ et on montre facilement qu'il a pour variance

$$\sigma^2(\frac{1}{n_1} + \frac{1}{n_2}).$$

Par le même raisonnement vu précédemment, notamment en utilisant le carré moyen résiduel $\widehat{\sigma^2} = \frac{1}{n-I}SC_1 = \frac{1}{n-I}\sum_{i,j}(Y_{i.} - Y_{..})^2$ pour estimateur de σ^2, on montre qu'un intervalle de confiance de niveau $1-\alpha$ pour $\mu_1 - \mu_2$ est donné par

$$\left[Y_{1.} - Y_{2.} - T_{(n-I),\alpha/2}\sqrt{\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\ ,\ Y_{1.} - Y_{2.} + T_{(n-I),\alpha/2}\sqrt{\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right].$$

Comme précédemment, on en déduit également un test de Student de comparaison de ces 2 moyennes en regardant si la valeur zéro appartient à cet intervalle. La statistique de test, $\widehat{T}$, est définie par

$$\widehat{T} = |Y_{1.} - Y_{2.}|\left(\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)^{-1/2}$$

que l'on comparera à $T_{(n-I),\alpha/2}$ pour accepter ou rejeter l'hypothèse H_0 : $\mu_1 = \mu_2$. Ce test est différent (et sous nos hypothèses, plus puissant) que le test de comparaison de deux moyennes

basé sur les deux seuls échantillons des groupes 1 et 2 car la variance résiduelle σ^2 est estimée sur l'ensemble des groupes.

Exercice

i. Soient $y_1, ..., y_n$ des réels et soit $\overline{y}$ leur moyenne. Montrer que cette dernière quantité minimise $SC(a) = \sum_{i=1}^{n} |y_i - a|^2$ (Indication : on peut dériver la fonction $a \mapsto SC(a)$).

ii. On rappelle que dans le cadre de variables aléatoires $(X_1, \cdots, X_n)$ indépendantes, et identiquement distribuées, de loi $\mathcal{N}(m, \sigma^2)$, l'estimateur du maximum de vraisemblance de (m, σ^2) est l'argument maximal de la densité de $(X_1, \cdots, X_n)$.

   (a) Déterminer l'estimateur du maximum de vraisemblance de (μ, β, σ^2) que dans le cadre du modèle (2.1). Comparer avec les estimateurs obtenus par moindres carrés.

   (b) Déterminer l'estimateur du maximum de vraisemblance de $(\mu_1, \cdots, \mu_I, \sigma^2)$ dans le cadre du modèle (2.2). Comparer avec les estimateurs obtenus par moindres carrés.

3 Conclusion

Dans les deux problèmes évoqués dans ce chapitre, à savoir la régression linéaire simple et l'analyse de variance à un facteur, nous avons utilisé :

- le même type d'hypothèses sur les erreurs ;
- l'estimateur des moindres carrés ;
- des tests de Fischer et de Student.

En fait, ces deux problèmes ne sont pas si éloignés qu'ils le paraissent a priori car les deux modèles utilisés font partie d'une même famille de modèles : le modèle linéaire statistique.

Chapitre 3

Introduction au modèle linéaire statistique

1 Écriture matricielle de modèles simples

Dans cette partie nous donnons une présentation unifiée des modèles de la partie précédente et nous présentons de nouveaux modèles.

1.1 Régression linéaire simple

Nous reprenons le modèle de régression (2.1) du chapitre précédent,

$$Y_i = \alpha + \beta Z_i + \epsilon_i \quad i = 1, \ldots, n = 5.$$

On considère les vecteurs $Y = (Y_i)_{1 \leq i \leq 5}$ et $\varepsilon = (\varepsilon_i)_{1 \leq i \leq 5}$. Le modèle (2.1) s'écrit alors :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ 1 & Z_3 \\ 1 & Z_4 \\ 1 & Z_5 \end{pmatrix} \begin{pmatrix} \mu \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

ou encore

$$Y = X.\theta + \varepsilon \text{ avec } \theta = \begin{pmatrix} \mu \\ \beta \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ 1 & Z_3 \\ 1 & Z_4 \\ 1 & Z_5 \end{pmatrix}. \tag{3.1}$$

Attention ! pour conserver les notations usuelles, nous utiliserons la même typographie pour les matrices et les vecteurs, à savoir une lettre majuscule. Cependant, X sera en général une matrice, quant Y et Z seront des vecteurs.

1.2 Analyse de la variance à un facteur

On reprend l'exemple des trois forêts, et de la même manière que précédemment, on utilise la notation matricielle : formé par les trois moyennes μ_1, μ_2 et μ_3

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{25} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \\ Y_{35} \\ Y_{36} \\ Y_{37} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{34} \\ \varepsilon_{35} \\ \varepsilon_{36} \\ \varepsilon_{37} \end{pmatrix}$$

ou encore

$$Y = X.\theta + \varepsilon \quad \text{et} \quad \theta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}. \tag{3.2}$$

1.3 Régression linéaire multiple

Les deux exemples précédents nous ont montré que l'on pouvait écrire synthétiquement les deux modèles sous une même forme matricielle. Mais cette écriture se généralise également à d'autres modèles. Considérons par exemple l'observation de

- Y, un vecteur formé de n rendements Y_i d'une réaction chimique (exprimé en pourcentage) ;
- $Z^{(1)}$, un vecteur formé des n mesures $Z_i^{(1)}$ des températures de la réaction ;
- $Z^{(2)}$, un vecteur formé des n mesures $Z_i^{(2)}$ des pH du bain de la réaction.

On suppose que le rendement de la réaction chimique (variable à expliquer) dépend linéairement des deux variables explicatives, la température et le pH. On écrit donc le modèle de régression multiple suivant :

$$Y_i = \mu + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \varepsilon_i \tag{3.3}$$

pour $i = 1, \cdots, n$. Il s'en suit l'écriture matricielle suivante :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1^{(1)} & Z_1^{(2)} \\ 1 & Z_2^{(1)} & Z_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & Z_n^{(2)} \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ou encore :

$$Y = X.\theta + \varepsilon \quad \text{avec} \quad \theta = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & Z_1^{(1)} & Z_1^{(2)} \\ 1 & Z_2^{(1)} & Z_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & Z_n^{(2)} \end{pmatrix}. \tag{3.4}$$

2 Le modèle linéaire : définition et hypothèses

Nous avons vu que la régression linéaire simple et l'analyse de la variance à un facteur, qui sont des problèmes relativement différents, conduisaient à des outils statistiques proches : tests de Fisher et de Student, et table d'analyse de la variance. Nous avons vu que matriciellement ces 2 modèles s'écrivent de la même manière, ainsi qu'un modèle plus général de régression multiple. Nous allons également montrer que les principaux outils statistiques de ces modèles sont les mêmes.

Définition fondamentale : nous dirons qu'une variable Y constituée de n observations Y_i suit un modèle linéaire statistique si on peut écrire que :

$$Y = X.\theta + \varepsilon \tag{3.5}$$

où

- X est une matrice réelle connue à n lignes et un certain nombre de colonnes que l'on notera k, avec $k < n$. Pour simplifier, on supposera la plupart du temps dans le cadre de ce cours que X est régulière (c'est-à-dire de rang k). Ceci implique notamment que pour un vecteur θ de longueur k, $X.\theta = 0 \implies \theta = 0$.

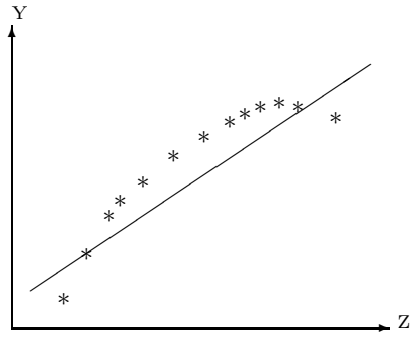
- θ est un vecteur inconnu constitué de k réels qui sont des paramètres du modèle ;

- le vecteur aléatoire ε appelé erreur du modèle est tel que $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ vérifie les 4 postulats (hypothèses invérifiables mais testables) suivants :

• **P1 : les erreurs sont centrés (espérances nulles) soit**

$$\mathrm{E}(\varepsilon) = 0.$$

En clair cela veut dire que le modèle posé [3.5) est correct, que l'on a pas oublié un terme pertinent. Illustrons ceci par le contre-exemple (en régression simple) suivant :



$\Longrightarrow$ il semble qu'un terme quadratique ait été oublié et le modèle devrait plutôt s'écrire : $Y_i = \mu + \beta_1 Z_i + \beta_2 Z_i^2 + \varepsilon_i$. Si on considère à la place un modèle linéaire simple (comme cela est fait sur la représentation graphique), le vecteur d'erreur ne sera pas centré.

- **P2 : la variance des erreurs est constante. C'est le postulat dit d'homoscedasticité, soit :**

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 \quad \text{ne dépend pas de } i.$$

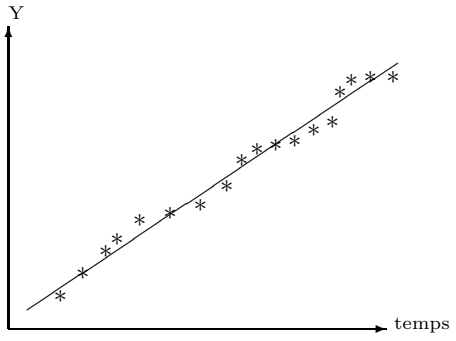
Ce postulat n'est évidemment pas toujours vérifié. Étudions le contre-exemple suivant (issu de l'analyse de variance) :
on étudie le taux de suivie des insectes à deux insecticides A et B. On fait différentes répétitions de l'expérience et on obtient le tableau de données suivant

	taux de survie	
	produit A	produit B
rep1	0,01	0,50
rep2	0.02	0.56
rep3	0.02	0.60
rep4	0.04	0.44
$\vdots$	$\vdots$	$\vdots$

A première vue, l'insecticide A paraît plus efficace que l'insecticide B : le taux de survie à A est plus proche de zéro et semble avoir une amplitude de variations beaucoup plus faible que celle de B (cette dernière propriété contredit l'hypothèse d'homoscedasticité).

- **P3 : les variables ε_i sont indépendantes.**

On considérera en général que ce postulat est vérifié lorsque chaque donnée correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes. En revanche, dans des problèmes où le temps joue un rôle important, il est plus difficilement vérifié (une évolution se fait rarement de façon totalement indépendante du passé). Voici un contre-exemple (cas de la régression simple) à ce postulat visible sur la représentation graphique suivante :



Dans ce contre-exemple, la variable explicative est le temps et il y a une certaine rémanence ou inertie du phénomène étudié. Cela s'observe par le fait que les données n'oscillent pas en

permanence autour de la droite de régression, mais semblent s'attarder une fois qu'elles sont d'un côté.

- **P4 : Les données suivent des lois gaussiennes, soit :**

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ pour tout } i.$$

C'est le postulat le moins important, comme nous le verrons plus tard, puisque l'on peut s'en passer quand le nombre de données est important.

3 Formules fondamentales

3.1 Le modèle linéaire en 4 formules

A partir des postulats du modèle linéaire statistique, on peut associer 4 formules fondamentales (notées F1-F4 par la suite). Ces formules sont très facilement implémentables et donc calculables par un ordinateur. Elles sont communes à l'analyse de la variance et à la régression. Elles sont issues de la minimisation en θ de la somme des carrés des résidus (SCR), somme qui peut s'écrire matriciellement sous la forme :

$$\text{SCR}(\theta) = \| Y - X.\theta \|^2 = (Y - X.\theta)'.(Y - X.\theta),$$

où, dans toute la suite, M' désigne la matrice transposée d'une matrice M quelconque. On a alors :

- **F1 :** $\widehat{\theta} = (X'X)^{-1}X'Y$.

Cette formule fournit l'expression de l'estimateur $\widehat{\theta}$ de θ par moindres carrés (attention, en général θ est un vecteur de paramètres). Un tel calcul est réalisable par un ordinateur qui sait faire du calcul matriciel (on peut traiter sans problème des problèmes comprenant plusieurs centaines de paramètres). Le vecteur Y est gaussien et donc par linéarité $\widehat{\theta}$ l'est également.

- **F2 :** $\text{E}(\widehat{\theta}) = \theta$.

Cette formule se traduit par le fait que l'estimateur est sans biais.

- **F3 :** $\text{Var}(\widehat{\theta}) = \sigma^2(X'.X)^{-1}$.

Cette formule fournit donc l'expression de la matrice de variance-covariance de l'estimateur $\widehat{\theta}$. Elle permet d'apprécier la précision de cet estimateur.

- **F4 :** la somme des carrés résiduels $SCR(\widehat{\theta}) = (Y - X.\widehat{\theta})'.(Y - X.\widehat{\theta})$ suit une loi $\sigma^2\chi^2(n-k)$.

La loi ci-dessus est une loi du khi-deux à $n - k$ degrés de liberté, multipliée par le coefficient σ^2 (on rappelle que X est une matrice de taille (n, k)). Cela permet d'estimer σ^2 par l'intermédiaire du carré moyen résiduel

$$\widehat{\sigma^2} = \text{CMR} = \frac{\text{SCR}(\widehat{\theta})}{n-k},$$

qui est proche de σ^2 quand $n-k$ est grand. De plus $\mathrm{E}\widehat{\sigma^2} = \sigma^2$, cet estimateur est sans biais. Enfin, $\widehat{\sigma^2}$ est une variable aléatoire indépendante de $\widehat{\theta}$.

Démonstration : nous allons utiliser la notion de projection orthogonale pour montrer les différentes formules F1-4. Rappelons auparavant, que si u est un vecteur et E un sous-espace vectoriel de $\mathbb{R}^n$, alors le projeté orthogonal de u sur E, noté $P_E u$, est le vecteur de E qui minimise $\| u - v \|^2$ où $v \in E$, soit encore

$$\| u - P_E u \|^2 = \min_{v \in E} \| u - v \|^2$$

(en effet, si v est un vecteur quelconque de E, d'après le Théorème de Pythagore, $\| u - v \|^2 = \| u - P_E u \|^2 + \| P_E u - v \|^2 \geq \| u - P_E u \|^2$). Dans toute la suite, on notera $[X]$ le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les vecteurs colonnes constituant la matrice X, ou encore $[X] = \{X.\theta,\ \theta \in \mathbb{R}^k\}$.

F1 : $X.\widehat{\theta}$ est donc le projeté orthogonal de Y sur $[X]$. Soit maintenant $X^{(i)}$ le ième vecteur colonne de X. Par définition de la projection orthogonale, pour tout $i = 1, \cdots, k$,

$$< X^{(i)}, Y > = (X^{(i)})'.Y = < X^{(i)}, X.\widehat{\theta} >,$$

($< ., . >$ note le produit scalaire euclidien usuel dans $\mathbb{R}^n$). En mettant les équations l'une au-dessus des autres pour tout $i = 1, \cdots, k$, on obtient

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

car X est supposée être une matrice régulière. Les équations $X'.Y = (X'.X)\widehat{\theta}$ sont appelées les **équations normales**.

F2 : En utilisant la linéarité de l'espérance, on a

$$\mathrm{E}(\widehat{\theta}) = \mathrm{E}\left[(X'.X)^{-1} X'.Y\right] = (X'.X)^{-1} X'.\mathrm{E}(Y) = (X'.X)^{-1} X'(X.\theta) = \theta.$$

F3 : comme $Y = X.\theta + \varepsilon$, d'après les postulats sur ε, $\mathrm{Var}(Y) = \sigma^2 Id$, où Id désigne la matrice identité sur $\mathbb{R}^n$. Ainsi,

$$\mathrm{Var}(\widehat{\theta}) = (X'.X)^{-1} X'.(\mathrm{Var}(Y)).X.(X'.X)^{-1} = \sigma^2 (X'.X)^{-1}.$$

F4 : par linéarité de la projection orthogonale, $P_{[X]} Y = X.\theta + P_{[X]}\varepsilon$. Ainsi,

$$\mathrm{SCR}(\widehat{\theta}) = \| Y - P_{[X]} Y \|^2 = \| \varepsilon - P_{[X]}\varepsilon \|^2 .$$

Or, on peut encore écrire que $\varepsilon - P_{[X]}\varepsilon = P_{[X]^\perp}\varepsilon$, où $[X]^\perp$ désigne le sous-espace vectoriel orthogonal de $[X]$. La dimension de $[X]^\perp$ étant $n-k$, d'après les résultats sur les variables gaussiennes indépendantes (Théorème de Cochran), la variable aléatoire $SCR(\widehat{\theta})$ suit un $\sigma^2 \chi^2(n-k)$.
Une variable aléatoire distribuée suivant la loi $\chi^2(n-k)$ pouvant encore s'écrire comme la somme de $(n-k)$ variables gaussiennes centrées réduites indépendantes, il est clair que $\mathrm{E}(CM) = \sigma^2$, et d'après la loi des grands nombres, $CM \underset{(n-k)\to+\infty}{\xrightarrow{\mathcal{P}}} \sigma^2$.
Enfin, comme $X.\widehat{\theta} = X.\theta + P_{[X]}\varepsilon$, toujours en utilisant le Théorème de Cochran, comme $P_{[X]}\varepsilon$ et $P_{[X]^\perp}\varepsilon$ sont des projections sur des espaces orthogonaux, alors $P_{[X]}\varepsilon$ est indépendant de $SCR(\widehat{\theta})$, donc $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont également indépendants. ■

L'estimateur $\widehat{\theta}$ ainsi défini a des propriétés d'optimalité. C'est un estimateur optimal parmi les

estimateurs sans biais. Soit $\tilde{\theta}$ un autre estimateur sans biais et soit $C \in \mathbb{R}^k$. Alors $C'.\theta$ est une combinaison linéaire des différents θ_i et on peut montrer (nous l'admettrons) que pour tout $C \in \mathbb{R}^k$,

$$\mathrm{Var}(C'\tilde{\theta}) \geq \mathrm{Var}(C'\widehat{\theta}).$$

En absence du postulat **P4**, l'estimateur $\widehat{\theta}$ reste optimal parmi les estimateurs linéaires sans biais.

3.2 Un exemple : les équations explicites dans le cas de la régression linéaire simple.

On considère le cas de la régression linéaire simple et le modèle (2.1). Nous allons retrouver tous les résultats du chapitre 2 à partir de leurs écritures matricielles.

Soit $\overline{Z}$ la moyenne des Z ; on pose $Z_i = Z_i - \overline{Z}$ (régresseur centré). On écrit alors le modèle :

$$Y_i = \mu + \beta\overline{Z} + \beta(Z_i - \overline{Z}) + \varepsilon_i,$$

donc, en posant $\mu' = \mu + \beta\overline{Z}$, on obtient,

$$Y_i = \mu' + \beta Z_i^o + \varepsilon_i.$$

Nous allons voir que travailler avec les Z_i^o au lieu des Z_i permet de se placer dans le cadre de l'orthogonalité et de simplifier les calculs. Le modèle s'écrit alors matriciellement sous la forme

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1^o \\ \vdots & \vdots \\ 1 & Z_1^o \end{pmatrix} \begin{pmatrix} \mu' \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Ainsi, en posant $X = \begin{pmatrix} 1 & Z_1^o \\ \vdots & \vdots \\ 1 & Z_1^o \end{pmatrix}$ et $\theta = \begin{pmatrix} \mu' \\ \beta \end{pmatrix}$, on peut utiliser les formules F1-4 précédentes.

Or,

$$X'.X = \begin{pmatrix} n & \sum Z_i^o \\ \sum Z_i^o & \sum (Z_i^o)^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum (Z_i^o)^2 \end{pmatrix} \text{ d'où } (X'.X)^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}.$$

De plus,

$$X'.Y = \begin{pmatrix} 1 & \dots & 1 \\ Z_1^o & \dots & Z_n^o \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i Z_i^o \end{pmatrix}$$

et donc

$$\widehat{\theta} = \begin{pmatrix} \widehat{\mu'} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} n^{-1} & 0 \\ 0 & \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum Y_i Z_i^o \end{pmatrix} = \begin{pmatrix} n^{-1} \sum Y_i \\ \left(\sum Y_i Z_i^o\right) \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}.$$

On a donc également :

$$\mathrm{Var} \begin{pmatrix} \widehat{\mu'} \\ \widehat{\beta} \end{pmatrix} = \sigma^2 (X'.X)^{-1} = \begin{pmatrix} \sigma^2 n^{-1} & 0 \\ 0 & \sigma^2 \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}$$

Or, $\widehat{\mu'} = \widehat{\beta}\overline{Z} - \bar{Y}$, et donc :

$$\text{Var}\begin{pmatrix} \widehat{\mu} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} \text{Var}(\widehat{\mu'}) + (\overline{Z})^2\text{Var}(\widehat{\beta}) - 2\overline{Z}\text{Cov}(\widehat{\mu'}, \widehat{\beta}) & \text{Cov}((\widehat{\mu'} - \widehat{\beta}\overline{Z}), \widehat{\beta}) \\ \text{Cov}((\widehat{\mu'} - \widehat{\beta}\overline{Z}), \widehat{\beta}) & \sigma^2 \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 n^{-1} + \sigma^2.\overline{Z}^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} & -\overline{Z}\sigma^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} \\ -\overline{Z}\sigma^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} & \sigma^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} \end{pmatrix}$$

Nous avons donc retrouvé toutes les formules que nous avons admises au chapitre 2.

4 Tests fondamentaux

4.1 Tests de Fisher d'un sous-modèle

Ces tests sont communs à l'analyse de la variance et à la régression. Au cours du chapitre précédent, nous avons vu deux exemples de sous-modèles :

Modèle général de la régression linéaire simple :	$Y_i = \mu + \beta Z_i + \varepsilon_i$.
Sous-modèle avec nullité de la pente :	$Y_i = \mu + \varepsilon_i$.
Modèle général de l'analyse de la variance à 1 facteur :	$Y_{ij} = \mu_i + \varepsilon_{ij}$.
Sous-modèle avec égalité des groupes :	$Y_{ij} = \mu + \varepsilon_{ij}$.

Plaçons nous maintenant dans le cadre général du modèle linéaire. Soit le modèle (3.5), où X est une matrice de rang $k_1 < n$. On note SCR_1 la somme des carrés résiduelle de ce modèle, associée à $n - k_1$ degrés de liberté, soit

$$\text{SCR}_1 = \| Y - X.\widehat{\theta} \|^2 .$$

On considère le sous-modèle issu du modèle (3.5) et tel que

$$Y = \underline{X}.\underline{\theta} + \varepsilon, \tag{3.6}$$

où $\underline{X}$ est la matrice constituée de k_0 vecteurs colonnes de X (avec $k_0 < k_1$) et $\underline{\theta}$ est un vecteur de longueur k_0. On note alors SCR_0 la somme des carrés résiduelle de ce sous-modèle, associée donc à $n - k_0$ degrés de liberté, soit

$$\text{SCR}_0 = \| Y - \underline{X}.\widehat{\underline{\theta}} \|^2 .$$

On veut tester que dans le "grand" modèle (3.5), il y a en fait $(k_1 - k_0)$ coefficients qui sont nulles. Cela revient à tester, dans le cadre du modèle linéaire (3.5) et ses différents postulats :

H_0 : le modèle est $Y = \underline{X}.\underline{\theta} + \varepsilon$

contre

H_1 : le modèle n'est pas $Y = \underline{X}.\underline{\theta} + \varepsilon$

Proposition 3.1 *Sous l'hypothèse nulle H_0 (si le sous-modèle (3.6) est vrai), alors :*

$$\widehat{F} = \frac{(SCR_0 - SCR_1)/(k_1 - k_0)}{SCR_1/(n - k_1)}$$

suit une loi de Fisher de paramètres $((k_1 - k_0), (n - k_1))$. De plus $\widehat{F}$ est indépendante de $\underline{X}.\widehat{\underline{\theta}}$.

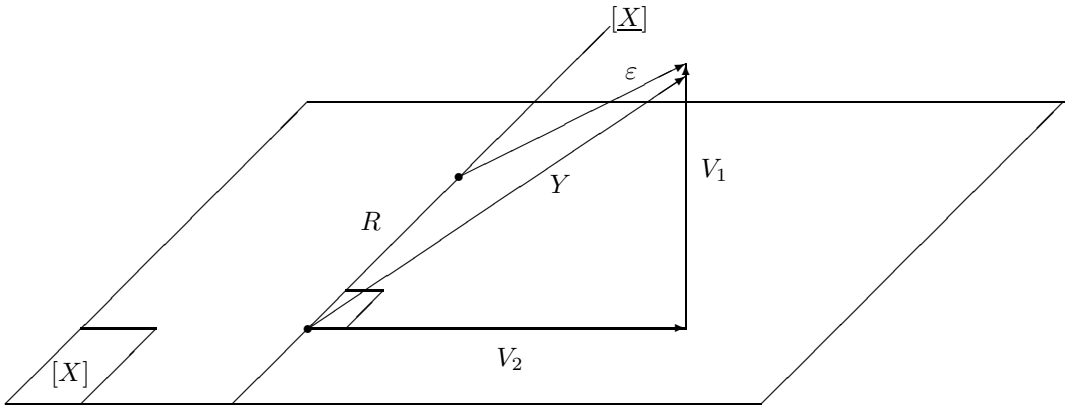
Démonstration : On se place sous les hypothèses et postulats du modèle (3.5). Alors, $Y = R + \varepsilon$, où les ε vérifient les postulats du modèle linéaire et $R = X.\theta,\ \theta \in \mathbb{R}^{k_1}$. On a donc

$$\mathrm{SCR}_1 = \|Y - P_{[X]}Y\|^2.$$

Sous H_0, on a $R = \underline{X}.\underline{\theta}$ avec $\underline{\theta} \in \mathbb{R}^{k_0}$, et

$$\mathrm{SCR}_0 = \|Y - P_{[\underline{X}]}Y\|^2,$$

et on réalise la figure suivante :



Par le théorème des trois perpendiculaires,

$$P_{[\underline{X}]}Y = P_{[\underline{X}]}P_{[X]}Y,$$

et les vecteurs

$$Y - P_{[X]}Y = V_1 \text{ et } P_{[X]}Y - P_{[\underline{X}]}Y = V_2$$

sont orthogonaux. On a aussi $V_1 = P_{[X]^\perp}Y$ et $V_2 = P_{([\underline{X}]^\perp \cap [X])}Y$. On note V_3 leur somme :$V_3 = V_1 + V_2 = Y - P_{[\underline{X}]}Y$ et donc $V_3 = P_{[\underline{X}]^\perp}Y$. Par le Théorème de Pythagore, on a $\|V_3\|^2 = \|V_1\|^2 + \|V_2\|^2$ Par linéarité, et du fait que $R \in [\underline{X}]$, on a encore

$$V_1 = P_{[X]^\perp}\varepsilon \ \text{ et } \ V_2 = P_{([\underline{X}]^\perp \cap [X])}\varepsilon.$$

Comme ε est un vecteur composé de variables gaussiennes centrées de vairance σ^2 indépendantes, d'après le Théorème de Cochran, ses projections sur deux sous-espaces orthogonaux sont indépendantes et leurs normes suivent des $\sigma^2.\chi^2$. On en déduit ainsi que

$$\begin{aligned} \mathrm{SCR}_1 &= \|V_1\|^2 \text{ suit un } \sigma^2\chi^2(n - k_1), \\ \mathrm{SCR}_0 - \mathrm{SCR}_1 &= \|V_3\|^2 - \|V_1\|^2 = \|V_2\|^2 \text{ suit un } \sigma^2\chi^2(k_1 - k_0). \end{aligned}$$

et ces deux quantités sont indépendantes. Donc

$$\widehat{F} = \frac{\|V_2\|^2/(k_1 - k_0)}{\|V_1\|^2/(n - k_1)}$$

suit bien la loi de Fisher de paramètre $((k_1 - k_0), n - k_1)$. De plus, comme $[\underline{X}]$ est orthogonal à $[\underline{X}]^\perp$ et à $[\underline{X}]^\perp \cap [X]$, on en déduit que $\widehat{F}$ est indépendant de $P_{[\underline{X}]}Y = \underline{X}.\underline{\widehat{\theta}}$. ■

4.2 Test de Student de la nullité d'une combinaison linéaire

On se place à nouveau dans le cadre du modèle (3.5). Soit une combinaison linéaire $C'.\theta$ du paramètre θ, avec $C \in \mathbb{R}^k$ (à titre d'exemple, une telle combinaison linéaire pourrait être $\mu_1 - \mu_2$ en analyse de la variance, ou β en régression). On veut tester la nullité de $C'.\theta$, donc tester :

$$H_0 \ : \quad C'.\theta = 0$$

contre

$$H_1 \ : \quad C'.\theta \neq 0$$

Proposition 3.2 *Sous l'hypothèse nulle H_0, alors :*

$$\widehat{T} = \frac{C'.\widehat{\theta}}{\sqrt{\widehat{\sigma}^2.C'(X'.X)^{-1}C}}$$

suit une loi de Student de paramètre $(n - k)$.

Démonstration : Dans le cadre du modèle (3.5) et de ses différents postulats, alors d'après **F2** et **F3**, $\mathrm{Var}(C'.\widehat{\theta}) = \sigma^2 C'.(X'.X)^{-1}C$. Comme vu précédemment, on estime $\mathrm{Var}(C'.\widehat{\theta})$ par $\widehat{\sigma^2}.C'.(X'.X)^{-1}C$. On divise alors $C'.\widehat{\theta}$ par son écart-type estimé. Une telle démarche est classique en statistique ; on dit encore que l'on a Studentisé la statistique. On obtient ainsi la statistique $\widehat{T}$ qui s'écrit encore,

$$\widehat{T} = \frac{C'.\widehat{\theta}}{\sqrt{\sigma^2 C'.(X'.X)^{-1}C}} \times \frac{\sqrt{\sigma^2}}{\sqrt{\widehat{\sigma}^2}}.$$

Or, sous l'hypothèse H_0, comme $C'.\theta = 0$, $C'.\widehat{\theta} = C'.P_{[X]}\varepsilon$ est une variable gaussienne centrée, et il est clair que

$$\frac{C'.\widehat{\theta}}{\sqrt{\sigma^2 C'.(X'.X)^{-1}C}} \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1).$$

De plus, $\dfrac{n-k}{\sigma^2}\widehat{\sigma}^2 = \dfrac{n-k}{\sigma^2}\|P_{[X]^bot}\varepsilon\|^2$ qui suit une loi $\chi^2(n-k)$ et est indépendant de $C'.P_{[X]}\varepsilon$ puisque les deux sous-epsaces sont orthogonaux. Par définition d'une loi de Student, on en déduit que $\widehat{T}$ suit bien une loi de Student à $(n - k)$ degrés de liberté. ■

Cette proposition permet de réaliser un test de Student des hypothèses précédentes. En effet, si $|\widehat{T}| > T_{n-k,\alpha/2}$, où $T_{n-k,\alpha/2}$, alors on rejette l'hypothèse H_0 et donc la nullité de $C'.\theta$. Dans le cas contraire, on acceptera H_0.

Remarque : $C'.\theta = 0$ définit un sous-modèle (pas toujours facile à écrire) du modèle (3.5) et le test de Fisher associé est exactement le même que le test de Student ci-dessus. Il s'agit simplement d'une présentation différente.

4.3 Test de Fisher de la nullité jointe de plusieurs combinaisons linéaires

On suppose que l'on réalise une expérience de type médicale, avec un seul effet traitement (facteur) à 5 niveaux. On peut par exemple écrire le modèle sous la forme $Y_{ij} = \theta_i + \varepsilon_{ij}$ pour $i = 1, \cdots, 5$, ou bien sous la forme générale (3.5) avec $\theta = (\theta_i)_{1 \leq i \leq 5}$.

Suppposons que l'on veuille tester l'hypothèse $H_0 : \theta_1 = \theta_2 = \theta_3$ et $\theta_4 = \theta_5$.
Notons

$$C' = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \text{ , alors } C'\theta = \begin{pmatrix} \theta_1 - \theta_2 \\ \theta_2 - \theta_3 \\ \theta_4 - \theta_5 \end{pmatrix}.$$

Ainsi, cela revient à tester l'hypothèse nulle $C'.\theta = 0$.

Plus généralement, on se place dans le cadre du modèle linéaire (3.5) et ses postulats, et on considère C une matrice réelle connue de taille (k, p), que l'on supposera de rang $p \leq k$. On désire tester :

$$H_0 \ : \quad C'.\theta = 0$$

contre

$$H_1 \ : \quad C'.\theta \neq 0$$

Proposition 3.3 *Sous l'hypothèse nulle H_0, alors :*

$$\widehat{F} = \frac{\widehat{\theta}.(C'.(X'.X)^{-1}C)^{-1}C'.\widehat{\theta}}{p.\widehat{\sigma^2}}$$

suit une loi de Fisher de paramètre $(p, (n-k))$.

<u>*Démonstration*</u> : Comme dans la démonstration précédente, sous l'hypothèse H_0, $C'.\widehat{\theta} = C'.P_{[X]}\varepsilon$ est un vecteur gaussien centré de matrice de variance $V = \sigma^2.C'.(X'.X)^{-1}C$. Or la matrice V ainsi définie est symétrique et inversible, car C est supposée être de rang p. Comme V est diagonalisable avec des valeurs propres positives, il est donc évident qu'il existe une "racine carrée de son inverse" : il existe R symétrique telle que

$$RRV = RVR = I_d.$$

En conséquence, $R.C'.\widehat{\theta}$ est un vecteur gaussien centré de taille p et de matrice de variance identité. On en déduit alors que

$$\|R.C'.\widehat{\theta}\|^2 = \widehat{\theta}'.C.R'.R.C'.\widehat{\theta} = \widehat{\theta}'.C.(\sigma^2 C'(Z'Z)^{-1}C)^{-1}C'.\widehat{\theta} \text{ suit une loi } \chi^2(p).$$

De plus, on l'a déjà vu, $(n-k)\widehat{\sigma^2} = P_{[X]^\perp}\varepsilon$ et suit ainsi une loi $\sigma^2\chi^2(n-k)$, en étant indépendant de $C'.\widehat{\theta}$ puisque $[X]$ et $[X]^\perp$ sont orthogonaux. On en déduit alors que $\dfrac{\|R.C'.\widehat{\theta}\|^2}{p} \times \dfrac{\sigma^2}{\widehat{\sigma}^2}$ suit une loi de Fisher de paramètre $(p,(n-k))$. ■

Comme précédemment, on en déduit un test (dit test de Fisher ou test F)sur les hypothèses H_0 et H_1 : on rejettera notamment H_0 lorsque $\widehat{F} > F_{p,(n-k),1-\alpha}$.

Ce test est une généralisation assez naturelle du test précédent au cas où plusieurs combinaisons linéaires sont nulles conjointement. On montre (voir par exemple Coursol, 1980) que ce test exactement le même que le test de Ficher défini par des diffèrences de somme des carrés résiduelles avec le sous-modèle linéaire défini par l'hypothèse $C'.\theta = 0$.

5 Modèles linéaires et non linéaires

Nous avons vu que la régression multiple est un modèle linéaire. Parmi les variables explicatives (on dit aussi régresseurs dans ce cas), certaines peuvent être fonction l'une de l'autre : la régression polynomiale est un tel exemple de modèle linéaire. En voici un exemple :

Soit Y la variable à expliquer et Z une variable explicative pertinente mais dont l'effet n'est pas linéaire. On propose le modèle polynomial suivant, pour $i = 1, \cdots, n$,

$$Y_i = \mu + \beta_1 Z_i + \beta_2 Z_i^2 + \beta_3 Z_i^3 + \beta_4 Z_i^4 + \varepsilon_i.$$

C'est un modèle de régression multiple avec 4 régresseurs qui sont les puissances sucessives de la variable explicative. Il s'écrit encore sous la fomre d'un modèle linéaire général (3.5) avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1 & Z_1^2 & Z_1^3 & Z_1^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n & Z_n^2 & Z_n^3 & Z_n^4 \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = X.\theta + \varepsilon$$

Ce modèle est polynomial en Z_i, mais linéaire en les paramètres inconnus μ, β_1, β_2, β_3, β_4.

De la même façon, le modèle suivant est un modèle linéaire statistique :

$$Y_i = \mu + \beta_1 Z_i + \beta_2 Z_i^2 + \gamma_1 \sin Z_i + \gamma_2 \sin 2Z_i + \gamma_3 \cos Z_i + \gamma_4 \cos 2Z_i + \tau_1 \log Z_i + \varepsilon_i$$

Il existe aussi des modèles qui ne sont pas linéaires (voir en particulier Tomassone *et al.* 1983, chap 5), par exemple :

- les modèles exponentiels, issus des modèles à compartiments, soit par exemple, pour $i = 1, \cdots, n$,
$$Y_i = \beta_1 \exp(-\alpha_1 t_i) + \beta_2 \exp(-\alpha_2 t_i) + \varepsilon_i.$$

Les paramètres inconnus sont $\alpha_1, \alpha_2, \beta_1, \beta_2$ et la dépendance en α_1, α_2 est non linéaire à cause de l'exponentielle ;

- les modèles logistiques, soit, par exemple, pour $i = 1, \cdots, n$,

$$Y_i = \frac{\beta_1 + \beta_2 \exp(\beta_3 x_i)}{1 + \beta_4 \exp(\beta_3 x_i)} + \varepsilon_i.$$

Les paramètres inconnus sont $\beta_1, \beta_2, \beta_3, \beta_4$ et la dépendance est encore non linéaire.

Le traitement numérique et statistique de tels modèles non linéaires est considérablement plus délicat que celui des modèles linéaires.

6 Comportement asymptotique des statistiques

Il est intéressant maintenant de connaître les performances asymptotiques des statistiques précédemment définies. En premier lieu, examinons si les estimateurs des paramètres du modèle linéaire sous sa forme générale (3.5), à savoir $\widehat{\theta}$ et $\widehat{\sigma^2}$, sont convergents vers θ et σ^2 lorsque le nombre de données (c'est-à-dire le n du modèle (3.5) devient très grand. Dans un deuxième temps, nous étudierons ce qu'il advient des statistiques de test, $\widehat{F}$ ou $\widehat{T}$ lorsque pareillement les effectifs deviennent très grands.
A cet effet, commençons par rappeler les deux théorèmes limites usuels :

6.1 Loi Forte des Grands Nombres et Théorème de la Limite Centrale

Quelle est la fréquence des piles lorsque l'on jette un grand nombre de fois une pièce de monnaie ? Comment un sondage peut-il permettre d'estimer le score d'un candidat à une élection ? C'est à ce genre de question que répondent les deux théorèmes limite suivants. Le premier, appelé Loi, car on a longtemps cru qu'il ressortait des lois de la nature (du même ordre que la gravitation), dit en substance que "la moyenne empirique se rapproche de plus en plus de la moyenne théorique, ou espérance, lorsque le nombre de données croît". Plus précisément, son énoncé est le suivant :

Théorème 3.1 (Loi Forte des Grands Nombres)
Soit $Z_1 \dots Z_n$, n variables aléatoires indépendantes distribuées suivant une même loi d'espérance μ (donc $E|Z| < +\infty$). Soit $\overline{Z}_n = \frac{1}{n}(Z_1 + \cdots + Z_n)$ la moyenne empirique. Alors quand n est grand

$$\overline{Z}_n \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \mu.$$

Notons que les hypothèses peuvent être étendues à des variables qui ne sont plus indépendantes entre elles, ou qui n'ont pas forcément la même espérance.
Le second théorème précise en quelque sorte la manière dont la moyenne empirique se rapproche de l'espérance :

Théorème 3.2 (Théorème de la Limite Centrale)
Soit $Z_1 \dots Z_n$, n variables aléatoires indépendantes distribuées suivant une même loi d'espérance μ et de variance σ^2. Soit $\overline{Z}_n = \frac{1}{n}(Z_1 + \cdots + Z_n)$ la moyenne empirique. Alors quand n est grand

$$\sqrt{n}\left(\frac{\overline{Z}_n - \mu}{\sigma}\right) = \frac{\overline{Z}_n - \mu}{\sqrt{\sigma^2/n}} \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1).$$

La traduction en français de ce résultat est que "la moyenne empirique tend à être gaussienne" quand le nombre de données devient grand. Concrètement, cela veut dire que lorsque n devient grand, $\overline{Z}_n$ a une loi très proche de la loi gaussienne de moyenne μ et de variance σ^2/n. Une applica tion de ce résultat est par exemple la simulation informatique de variables aléatoires gaussiennes que l'on obtient en simulant la moyenne de plusieurs dizaines de variables uniformes sur $[0,1]$ (que les calculatrices ou ordinateurs savent tous (à peu près) simuler : c'est notamment la touche RAND).

6.2 Convergence vers les paramètres

Les théorèmes précédents, ainsi que leurs extensions vont nous permettre de préciser le comportement asymptotique des estimateurs des paramètres du modèle linéaire (3.5). En prenant le cas particulier de la régression simple, on conçoit bien que sous certaines conditions plus l'on dispose de données, plus la droite régression va être "se caler" sur la vraie droite : $\widehat{\theta}$, qui défini entièrement cette droite, va converger vers la "vraie" valeur θ. Et il peut être particulièrement important, de préciser l'erreur faite sur cette estimation ; on pense notamment à des expériences en physique dans lesquelles l'erreur ε joue le rôle d'erreur de mesure, ou bien du mesure de la volatilité (représentée par σ^2) pour des données financières. La conséquence d'une "bonne" estimation des paramètres est de permettre d'obtenir des prédictions avec une marge d'erreur connue. Revenons donc au cas général.

Théorème 3.3 *On se place dans le cadre du modèle linéaire (3.5) et des postulats P1-4 associés. On suppose que k est fixé. Alors :*

i. On a $\widehat{\sigma^2} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \sigma^2$ *et* $\sqrt{n}(\widehat{\sigma^2}-\sigma^2) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,2\sigma^4)$.

ii. Si $(X'.X)^{-1} \underset{n\to+\infty}{\longrightarrow} 0$, *alors* $\widehat{\theta} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \theta$.

iii. On suppose maintenant que $\frac{1}{n}(X'.X) \underset{n\to+\infty}{\longrightarrow} Q$, *où Q est une matrice d'ordre k définie positive et que* $Tr(X.X') = \mathcal{O}(n)$. *Alors*

$$\sqrt{n}(\widehat{\theta}-\theta) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,\sigma^2.Q^{-1}.$$

<u>*Démonstration*</u> : Pour La démonstration de i. on sait que $(n-k)\widehat{\sigma^2}$ suit une loi $\sigma^2.\chi^2(n-k)$, donc peut encore s'écrire comme la somme de $(n-k)$ carrés de variables gaussiennes centrées de variance σ^2. La Loi des Grands Nombres et le Théorème de la Limite Centrale impliquent alors la convergence de $\widehat{\sigma^2}$ vers σ^2.
La démonstration de ii. est issue de l'Inégalité de Bienaymé-Tchebitchev appliquée à $\|X\|$.
Pour iii. on utilise un Théorème de la Limite Centrale un plus général que celui proposé un peu plus haut. Nous admettrons donc ce résultat (voir par exemple Guyon, 2001). ■

Les conditions $(X'.X)^{-1} \underset{n\to+\infty}{\longrightarrow} 0$ et $\frac{1}{n}(X'.X) \underset{n\to+\infty}{\longrightarrow} Q$ sont des conditions que l'on peut souvent vérifier. Par exemple, dans le cadre de l'analyse de la variance à 1 facteur sur k classes, la matrice $X'.X$ est une matrice diagonale d'ordre k, dont les termes diagonaux sont les effectifs de chaque classes (la somme de tous ces effectifs valant n). Dès que la fréquence de chaque classe tend vers une constante, laors on a bien $\frac{1}{n}(X'.X) \underset{n\to+\infty}{\longrightarrow} Q$ (c'est notamment le cas lorsque les classes ont toutes le même effectif).

Un autre exemple parlant est celui de la régression. Lorsque chaque colonne j de X est constitué de réalisations indépendantes d'une variable aléatoire Z^j de carrés intégrable, et les différentes Z^j sont indépendantes entre elles, alors la Loi des Grands Nombres montre que $\frac{1}{n}(X'.X) \underset{n\to+\infty}{\longrightarrow} D$, où D est une matrice diagonale dont les termes diagonaux sont les $\mathrm{E}((Z^j)^2)$.
Cependant, les conditions demandées sur $X'.X$ ne sont pas toujours respectées. A titre d'exemple, considérons le cas de la régression simple de modèle (2.1). Voyons deux cas particuliers significatifs :

- si $x_i = i$ pour $i = 1, \cdots, n$, alors $X'.X = \begin{pmatrix} n & n(n+1)/2 \\ n(n+1)/2 & n(n+1)(2n+1)/6 \end{pmatrix}$ et dans ce cas on a bien $(X'.X)^{-1} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} 0$, donc $\widehat{\theta}$ converge vers θ, mais on n'a pas $\frac{1}{n}(X'.X) \underset{n\to+\infty}{\longrightarrow} Q$. En fait la convergence de $\widehat{\theta}$ vers θ est plus rapide que celle proposée dans iii.du Théorème.

- si $x_i = 1/i$ pour $i = 1, \cdots, n$, alors $X'.X = \begin{pmatrix} n & \sum 1/i \\ \sum 1/i & \sum 1/i^2 \end{pmatrix}$, donc en utilisant le fait que $\sum 1/i^2$ converge, on montre que $(X'.X)^{-1}$ ne converge pas vers 0. Dans ce cas, il n'y a pas convergence de $\widehat{\theta}$ vers θ.

6.3 Convergence des statistiques de test

On se place donc toujours dans le cadre du modèle linéaire statistique (3.5) associé au postulats P1-4. Que se passe-t-il quant aux différentes statistiques de test précédentes lorsque les effectifs deviennent grands ?
Pour répondre à une telle question, il suffit de connaître le comportement asymptotique des lois de Fisher et de Student. C'est ce à quoi répond la proposition suivante :

Théorème 3.4 *Soit p et k deux entiers fixés tels que $0 < p < k$. Alors :*

i. si $\widehat{F}$ suit une loi de Fisher de paramètres $(p, n-k)$, alors

$$\widehat{F} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \frac{1}{p}\chi^2(p).$$

ii. si $\widehat{T}$ suit une loi de Student de paramètre $(n-k)$, alors

$$\widehat{T} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,1).$$

<u>*Démonstration*</u> : pour montrer i., il suffit d'écrire $\widehat{F}$ comme le rapport de deux variables aléatoires indépendantes qui suivent respectivement une loi $\frac{1}{p}\chi^2(p)$ (numérateur) et une loi $\frac{1}{n-k}\chi^2(n-k)$ (dénominateur). Or lorque $n \to \infty$, ce dénominateur tend en probabilité vers 1 (voir plus haut). Par suite, on a un produit de variables donc une converge en probabilité vers une constante (1) : ce produit converge en loi vers la loi du numérateur.
pour montrer ii., il suffit là-encore d'écrire $\widehat{T}$ comme le rapport de deux variables aléatoires indépendantes qui suivent respectivement une loi $\mathcal{N}(0,1)$ (numérateur) et une loi $\sqrt{\frac{1}{n-k}\chi^2(n-k)}$ (dénominateur). Le dénominateur tend en probabilité vers 1 quand $n \to \infty$, d'où la convergence en loi de $\widehat{T}$ vers la loi $\mathcal{N}(0,1)$. ■

De ces résultats, on en déduit que les trois tests proposés (test de Fisher du'un sous-modèle, test de Student d'une combinaison linéaire nulle, test de Fisher de plusieurs combinaisons linéaires nulles) peuvent être utilisés en toute généralité pour de très grands effectifs.

7 Quand les postulats ne sont pas respectés...

7.1 Postulat de non-gaussianité des données

Le postulat de gaussianité des erreurs est le plus difficile à vérifier en pratique. Les tests classiques de normalité (test de Kolmogorov-Smirnov, Cramer-Von Mises, Anderson-Darling ou de Shapiro-Wilks) demanderaient l'observation des erreurs ε_i elles-mêmes ; ils perdent beaucoup de puissance quand ils sont appliqués sur les résidus $\widehat{\varepsilon_i} = (Y - P_{[X]}Y)_i$, notamment en raison du fait que ces résidus ne sont pas indépendants en général. Le postulat de gaussianité sera donc un credo que l'on ne pourra pas vraiment vérifier expérimentalement. Fort heureusement, comme nous allons le décrire, il existe une théorie asymptotique (donc pour les grands échantillons) du modèle linéaire qui n'a pas besoin de cette hypothèse. Comme il est dit dans l'introduction, c'est dans cette optique là qu'il faut réellement penser le modèle linéaire. Avant d'exposer cette théorie, nous allons vérifier que certaines propriétés restent vraies même si la loi n'est pas gaussienne.

Propriétés des estimateurs $\widehat{\theta}$ et $\widehat{\sigma^2}$

On se place dans le cadre du modèle linéaire (3.5) et on suppose donc ici que les postulats P1-3 sont vérifiés, mais pas forcément P4 (gaussianité). Examinons les changements induits sur les estimateurs $\widehat{\theta}$ et $\widehat{\sigma^2}$.

a/ En premier lieu, si on considère l'estimateur du vecteur des paramètres θ, on a encore $\widehat{\theta} = (X'.X)^{-1}X'.Y$ et :

- $\widehat{\theta}$ reste sans biais : $\mathrm{E}(\widehat{\theta}) = \theta$;
- la matrice de variance de $\widehat{\theta}$ reste égale à $\sigma^2(X'.X)^{-1}$;
- $\widehat{\theta}$ n'est plus de manière exacte un vecteur gaussien. Il l'est cependant asymptotiquement sous certaines conditions (voir un peu plus bas) ;
- $\widehat{\theta}$ n'est plus un estimateur optimal parmi les estimateurs sans biais, mais seulement parmi les estimateurs linéaires sans biais.

En ce qui concerne les propriétés asymptotiques de $\widehat{\theta}$, on peut montrer (voir par exemple Huber, 1980, p. 157) la propriété suivante :

Théorème 3.5 *On suppose le modèle linéaire (3.5), ainsi que les postulats P1-3. On suppose également que les ε_i suivent tous la même loi de carré intégrable et on note $h_n = \max_{1\leq i,j\leq k} P_{[X]}$ le terme maximal de la matrice du projecteur sur $[X]$, soit $P_{[X]} = X.(X'.X)^{-1}X'$. Alors :*

$$\textit{si } \; h_n \underset{n\to+\infty}{\longrightarrow} 0, \quad \widehat{\theta} \textit{ est asymtotiquement gaussien.}$$

En particulier, si $\dfrac{1}{n^\alpha}X'.X \underset{n\to+\infty}{\longrightarrow} Q$ *où Q est une matrice définie positive et $\alpha > 0$, alors :*

$$\textit{si } \; h_n \underset{n\to+\infty}{\longrightarrow} 0, \quad n^{\alpha/2}(\widehat{\theta} - \theta) \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, \sigma^2.Q^{-1}).$$

Ainsi, sous des hypothèses assez générales, l'estimateur des paramètres du modèle est asymptotiquement gaussien, tout en convergeant à une vitesse pouvant être plus rapide que $\sqrt{n}$. Notons que ce théorème s'applique bien-sûr au cas gaussien et généralise en quelque sorte celui vu dans le cas uniquement gaussien.

b/ Venons-en maintenant aux propriétés de l'estimateur de σ^2. On a encore $\widehat{\sigma^2} = \dfrac{1}{n-k}\|Y - X.\widehat{\theta}\|^2$ et :

- $\widehat{\sigma^2}$ reste un estimateur sans biais : $\mathrm{E}(\widehat{\sigma^2}) = \sigma^2$;
- il est bien clair en revanche que $\|Y - X.\widehat{\theta}\|^2$ ne suit plus une loi $\sigma^2.\chi^2(n-k)$;
- dès que la loi de ε admet un moment d'ordre 2, on montre facilement (Inégalité de Bienaymé-Tchebitchev) que $\widehat{\sigma^2}$ converge en probabilité vers σ^2 ;
- enfin, si dans le cas gaussien $\widehat{\sigma^2}$ converge à la vitesse $\sqrt{n}$ vers σ^2 (voir plus haut), ce n'est plus forcément vrai dans le cas d'une loi quelconque : la vitesse de convergence vers σ^2 dépend fortement du type de loi des ε_i et de son nombre de moments.

Propriétés des statistiques de test F et T

Un résultat important à retenir est la validité asymptotique (pour de grands effectifs) des estimateurs et des tests évoqués précédemment même si les erreurs ne sont pas gaussiennes (sous certaines conditions peu restrictives).

Pour illustrer ce résultat dans un cas simple, prenons l'exemple de l'analyse de la variance à un facteur. Nous avons vu que l'estimateur $\widehat{\mu}_i$ de la valeur d'une classe était une simple moyenne : $\widehat{\mu}_i = Y_{i.}$. Pour "mesurer" la vitesse de convergence de $\widehat{\mu}_i$ vers μ_i, on utilise le Théorème de la Limite Centrale. Pour revenir au problème précédent d'analyse de la variance à un facteur, une conséquence de ce théorème est que pour n grand, $\widehat{\mu}_i = Y_{i.}$ suit approximativement une loi gaussienne et que $\widehat{\sigma^2}$ est très proche de σ^2. Ceci implique par exemple que si l'on veut tester l'hypothèse nulle "$\mu_i = m$", pour un i donné et avec m un réel connu, alors comme précédemment on considérera le test de Student dont la statistique est

$$\widehat{T} = \frac{\widehat{\mu}_i - m}{\widehat{\sigma^2}}.$$

Pour n grand, $\widehat{T}$ suit approximativement une loi gaussienne centrée réduite, qui n'est autre que la limite d'une loi de Student $T(n)$ dont le nombre n de degrès de liberté tend vers l'infini (voir plus haut).

Revenons au cas général, Sous certaines conditions, un tel exemple pourra se généraliser au traitement général du modèle linéaire général (3.5), notamment aux statistiques $\widehat{F}$ et $\widehat{T}$. C'est ce qu'atteste le théorème général suivant (que nous admettrons) :

Théorème 3.6 *On se place dans le cadre du modèle linéaire (3.5) avec les postulats P1, P2 et P3. On suppose de plus que les ε_i suivent tous une même loi et on note $h_n = \max_{1\leq i,j\leq k} P_{[X]}$ le terme maximal de la matrice du projecteur sur $[X]$, soit $P_{[X]} = X.(X'.X)^{-1}X'$. On teste l'hypothèse H_0 : $C'.\theta = 0$, contre l'hypothèse H_1 : $C'.\theta \neq 0$, où C est une matrice d'ordre (k,p) et de rang $p \leq k$,*

où k est fixé. Enfin, on considère la statistique de test $\widehat{F} = \dfrac{\widehat{\theta}.(C'.(X'.X)^{-1}C)^{-1}C'.\widehat{\theta}}{p.\widehat{\sigma^2}}$. *Alors :*

$$si \;\; h_n \underset{n\to+\infty}{\longrightarrow} 0, \quad \widehat{F} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(p).$$

Remarquons que cette loi limite est la même que celle obtenue dans le cas gaussien et que ce résultat est également valable pour la statistique de Student (car particulier de ce résultat lorsque $p = 1$). En conclusion ce résultat permet de justifier l'utilisation des tests F et T en absence de normalité pour de grands effectifs.

Ce résultat théorique peut être complété par une étude par simulation. Dans un mémoire, Bonnet et Lansiaux (1992) ont étudié le comportement du test de Fisher en analyse de la variance à 1 facteur à 2, 5 ou 10 niveaux, avec des indices de répétition de 2, 4 ou 8 ; on a donc de 4 à 80 données dans chaque expérience. La validité du test est appréciée par le niveau réel du test pour un niveau nominal de 10 %, 5 % ou 1 %.
Divers types de loi non normales sont utilisées. On apercoit un écart au comportement nominal du test seulement dans le cas où tous les éléments suivants sont réunis :

- dispositifs déséquilibrés,
- petits échantillons,
- loi dissymétrique.

Dans les autres cas tout se passe comme si les données étaient gaussiennes. Ainsi cette étude confirme, que sauf cas extrêmes, le test de Fisher n'a pas besoin de l'hypothèse de gaussianité pour être approximativement exact.

7.2 Si les autres postulats ne sont pas vérifiés...

Lorsque un des 3 autres postulats, soit **P1**, **P2** ou **P3**, n'est plus vérifié, beaucoup des résultats présentés ne sont plus vrais. En particulier, si **P2** (postulat d'hétérodsticité) ou **P3** (postulat d'indépendance) ne sont plus vérifiés, alors :

- la formule $\mathrm{Var}(\widehat{\theta}) = \sigma^2(X'.X)^{-1}$ n'est plus vraie ;
- les tests ne sont plus de niveau exact, car les différentes statistiques ne suivent plus en général les lois de Fisher ou de Student déterminés précédemment, et ceci même pour de grands échantillons.

Il est cependant possible de trouver des conditions remplaçant **P2** ou **P3** et permettant d'obtenir des résultats asymptotiques pour les estimateurs ou les statistiques de test du modèle linéaire ; par exemple, on peut travailler avec des erreurs modélisées par un processus autoregressif de type ARMA. Nous préfèrerons ici supposer que les postulats**P1**, **P2** et **P3** sont bien vérifiés.

8 Cas de modèles non réguliers

Certains modèles ne peuvent être paramétrés de façon régulière : ils sont naturellement surparamétrés. Un exemple simple est celui du modèle additif en analyse de la variance à 2 facteurs.

Considérons le cas où les 2 facteurs ont chacun 2 niveaux et que les 4 combinaisons sont observées une fois et une seule. On a donc, avec les notations vues précedemment :

$$\begin{aligned} Y_{11} &= \mu + a_1 + b_1 + \varepsilon_{11} \\ Y_{12} &= \mu + a_1 + b_2 + \varepsilon_{12} \\ Y_{21} &= \mu + a_2 + b_1 + \varepsilon_{21} \\ Y_{22} &= \mu + a_2 + b_2 + \varepsilon_{22}. \end{aligned}$$

La matrice X du modèle vaut :

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

On voit donc que tout vecteur de la forme $(\alpha+\beta, -\alpha, -\alpha, -\beta, -\beta)$ donne la valeur zéro lorsqu'il est multiplié par la matrice X. Les valeurs $\mu, a_i, b_i,\ (i = 1, 2)$ ne sont donc pas identifiables de manière unique.

Définition 3.1 *Le modèle est dit non régulier quand la matrice X est non injective c'est-à-dire s'il existe $\theta \neq 0$ tel que $X.\theta = 0$.*

On note K le noyau de X : $K = \{z \in \mathbb{R}^n / X.z = 0\}$. Commençons par deux remarques :

- $X.\widehat{\theta}$ reste unique, puisque ce vecteur est la projection de Y sur $[X]$,
- $\widehat{\theta}$ ne peut être unique puisque si $\widehat{\theta}$ est solution et si $z \in K$, $\widehat{\theta} + z$ est encore solution. Compte tenu de (i), si $\widehat{\theta}$ est une solution particulière, l'ensemble des solutions s'écrit $\{\widehat{\theta} + z,\ z \in K\}$.

Ceci nous amène à la définition suivante :

Définition 3.2 *Soit M une matrice, M^- est une matrice pseudo-inverse de M si $M.M^-M = M$.*

Proposition 3.4 *Si $(X'.X)^-$ est une matrice pseudo-inverse de $X'.X$, alors $\widehat{\theta} = (X'.X)^- X'.Y$ est une solution des équations normales*

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

Démonstration : On sait que $P_{[X]}Y$ existe de manière unique, donc il existe $u \in \mathbb{R}^k$ tel que $X'.Y = X'.P_{[X]}Y = X'.X.u$. Soit $\widehat{\theta} = (X'.X)^- X'.Y$. Alors, $\widehat{\theta}$ vérifie les équations normales puisque :

$$\begin{aligned} X'.X.\widehat{\theta} &= (X'.X)(X'.X)^- X'.Y \\ &= (X'.X)(X'.X)^- X'.X.z \\ &= X'.X.z = X'.Y \end{aligned}$$

d'après la définition de la matrice pseudo-inverse de $X'.X$. ■

Parmi toutes les matrices pseudo-inverses qui conduisent à une solution particulière des équations normales, certaines sont plus intéressantes que d'autres. C'est ce que nous allons maintenant étudier.

Contraintes d'identifiabilité

Proposition 3.5 *On suppose que* $rg(X) = dim[X] = h < k$ *de sorte qu'il y ait* $(k-h)$ *paramètres redondants. On définit une matrice* H *carrée d'ordre* k *telle que pour l'ensemble des* θ *qui vérifient* $H.\theta = 0$*,* X *est injective, c'est-à-dire telle que :*

$$Ker([H]) \cap Ker([X]) = \{0\}.$$

Alors

- *La matrice* $(X'.X + H'.H)$ *est inversible et son inverse est une matrice pseudo-inverse de* $(X'.X)$*.*
- *Le vecteur* $\widehat{\theta} = (X'.X + H'.H)^{-1}X'.Y$ *est l'unique solution des équations normales qui vérifie* $H.\widehat{\theta} = 0$*.*

Démonstration : Les propriétés de dimension d'espaces vectoriels impliquent qu'il existe un seul $\widehat{\theta}$ tel que

$$X\widehat{\theta} = P_{[X]}.Y \quad \text{avec} \quad H.\widehat{\theta} = 0.$$

Considérons le problème de la minimisation en θ de $\|Y - X.\theta\|^2 + \|H.\theta\|^2$. La valeur $\widehat{\theta}$ précédente est bien la solution de ce problème de minimum car elle minimise séparément les deux termes positifs. Le problème précédent peut alors se mettre sous la forme

$$\left\| \left(\frac{Y}{0}\right) - \left(\frac{X}{H}\right)\theta \right\|^2 \text{ est minimum,}$$

la barre horizontale désignant la concaténation de matrices, c'est-à-dire la mise l'une sous l'autre de chacune des colonnes des matrices. La matrice de droite est de plein rang car

$$\begin{aligned} \left(\frac{X}{H}\right).\theta = 0 \quad &\Rightarrow \quad X.\theta = 0 = H.\theta \\ &\Rightarrow \quad \theta \in \mathrm{Ker}([H]) \cap \mathrm{Ker}([X]) \Rightarrow \theta = 0. \end{aligned}$$

Nous savons alors que la solution des moindres carrés est donnée par

$$\widehat{\theta} = \left(\left(\frac{X}{H}\right)' . \left(\frac{X}{H}\right) \right)^{-1} \left(\frac{X}{H}\right)' . \left(\frac{Y}{0}\right) = (X'.X + H'.H)^{-1}X'.Y.$$

Il reste à montrer que $(X'.X + H'.H)^{-1}$ est une matrice pseudo-inverse de $(X'.X)$, ce qui se déduit facilement, puisque :

$$(X'.X)(X'.X + H'.H)^{-1}(X'.X) = X'.P_{[X]}.X = X'.X$$

car par définition $P_{[X]}.X = X$. ■

Exemples de contrainte : l'opérateur de balayage ou "sweep operator"

Ce type de contraintes est très important en analyse de la variance car il permet de comprendre le "pourquoi" de certaines sorties de programmes informatiques. Lorsque l'on introduit informatiquement un modèle linéaire, le programme doit "traquer" les colinéarités entre les colonnes de la matrice X. S'il détecte une ligne colinéaire aux précédentes, elle est supprimée. Cela permet de diminuer informatiquement la taille du modèle. Mais cela revient aussi en se plaçant dans le cadre précédent à imposer des contraintes de nullité de certaines coordonnées du vecteur θ. Il s'agit d'un

balayage des colonnes de la matrice X dans l'ordre d'entrée. Ceci explique pourquoi l'ordre des termes peut avoir son importance (bien que l'addition soit commutative) et pourquoi on trouve en analyse de la variance le dernier niveau d'un facteur fixé à zéro. Ainsi, dans note exemple introductif d'analyse de la variance à deux facteurs, "le sweep operator" introduira les contraintes $a_2 + b_2 = 0$.

Fonctions estimables et contrastes

Il existe des fonctions de θ qui ne dépendent pas de la solution particulière des équations normales, c'est-à-dire du type de contrainte d'identificabilité choisi. Ces fonctions sont appelées estimables car elles sont intrinsèques.

Définition 3.3 *Une combinaison linéaire $C'.\theta$ est dite estimable si elle ne dépend pas du choix particulier d'une solution des équations normales. On caractérise ces fonctions comme étant celles qui s'écrivent $C'.\theta = D'.X.\theta$, où D est une matrice de rang plein.*

Définition 3.4 *En analyse de la variance, on définit les "contrastes" comme les combinaisons linéaires de poids nul : $C'.\theta$ avec $C'.\mathbb{1} = 0$.*

Si on reprend encore notre exemple introductif d'analyse de la variance à deux facteurs, $a_1 - a_2$ est un contraste. Les contrastes sont le plus souvent des fonctions estimables.

Chapitre 4

Problèmes spécifiques à la régression

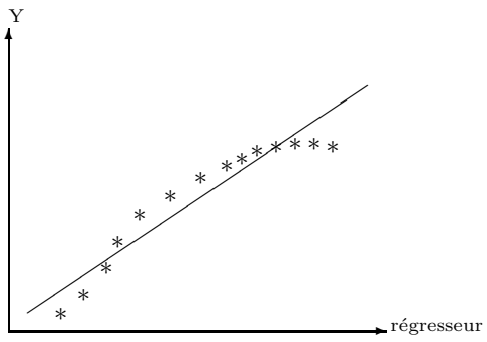
Dans cette partie nous allons essayer principalement de répondre à deux questions :

– quel genre de phénomène peut être modèlisé par un modèle de régression ?

– comment modifier un modèle de régression pour qu'il s'adapte mieux aux données ?

1 Contrôle graphique à posteriori.

Lorsque l'on a posé un modèle de régression il est indispensable de commencer par s'entourer de "protections" graphiques pour vérifier empiriquement les 4 postulats de base. En régression linéaire simple, la confrontation graphique entre le nuage de points (x_i, y_i) et la droite de régression de Y par X par moindres carrés ordinaires donne une information quasi exhaustive. En voici un exemple :



Sur ce graphique, on voit une courbure de la "vraie" courbe de régression de Y et on peut penser que le modèle est inadéquat et que le premier postulat **P1** n'est pas vrai.

Dans le cas de la régression multiple, ce type de graphique n'est pas utilisable car il y a plusieurs régresseurs. Les différents postulats sont à vérifier sur les termes d'erreur ε_i qui sont malheureusement inobservables. On utilise leurs estimateurs naturels, les résidus : $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Par exemple, pour le modèle général de régression

$$Y_i = \mu + \beta_1 Z_i^{(1)} + ... + \beta_p Z_i^{(p)} + \varepsilon,$$

le résidu $\widehat{\varepsilon}_i$ est défini pour $i = 1, \cdots, n$ par

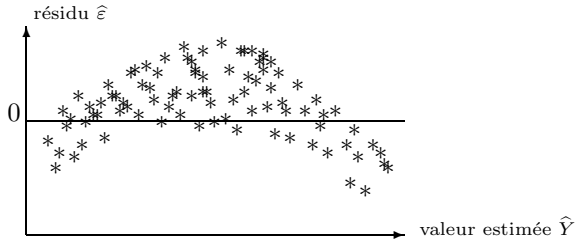
$$\widehat{\varepsilon}_i = Y_i - \widehat{\mu} - \widehat{\beta}_1 Z_i^{(1)} - \widehat{\beta}_2 Z_i^{(2)} - \ldots - \widehat{\beta}_p Z_i^{(p)}.$$

Cela revient encore à tracer les coordonnées du vecteur $P_{[X]^perp}.Y$ en fonction de celles de $P_{[X]}.Y$. L'intérêt d'un tel graphe réside dans le fait que si les quatre postulats **P1-4** sont bien respectés, d'après le Théorème de Cochran, il y a indépendance entre ces deux vecteurs qui sont centrés et gaussiens. Remarquons enfin que certaines options sophistiquées de SAS ou d'autres logiciels statistiques utilisent plutôt des résidus réduits (Studentised residuals) qui sont ces mêmes résidus divisés par un estimateur de leur écart-type (généralement l'écart-type empirique) : cela donne une information supplémentaire sur la distribution des résidus qui doit suivre alors (toujours sous les hypothèses **P1-4**) une loi de student.

Pour vérifier les postulats P1 et P2 : adéquation et homoscedasticité

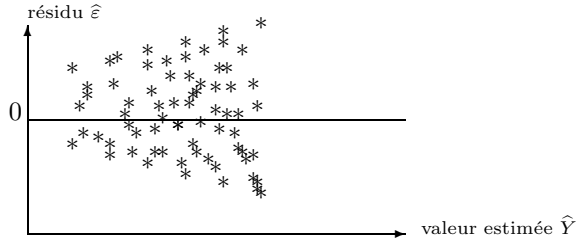
On utilise le graphique des résidus $(\varepsilon_i)_i$ en fonction des valeurs estimées $(\widehat{Y}_i)_i$. Ce graphique doit être fait pratiquement systématiquement. Si on ne voit rien de notable sur le graphique (c'est-à-dire que l'on observe un nuage de points centré et aligné quelconque), c'est très bon signe : les résidus ne semblent alors n'avoir aucune propriété intéressante et c'est bien ce que l'on demande à l'erreur. Voyons justement maintenant deux types de graphes résidus/valeurs estimées "pathologiques" :

Type 1



Dans ce premier cas, on peut penser que le modèle n'est pas adapté aux données. En effet, il ne semble pas y avoir indépendance entre les $\widehat{\varepsilon}_i$ et les $\widehat{Y}_i$ (puisque, par exemple, les $\widehat{\varepsilon}_i$ ont tendance à croître lorsque les $\widehat{Y}_i$ sont dans un certain intervalle et croissent). Il faut donc améliorer l'analyse du problème pour proposer d'autres régresseurs pertinents, ou transformer les régresseurs $Z^{(i)}$ par une fonction de type $(\log, \sin)$, ce que l'on peut faire sans précautions particulières.

Type 2



Dans ce cas la variance des résidus semble inhomogène, puisque les $\widehat{\varepsilon_i}$ ont une dispersion de plus en plus importante au fur et à mesure que les $\widehat{Y_i}$ croissent. Un changement de variable pour Y pourrait être une solution envisageable pour "rendre" constante la variance du bruit (voir un peu plus bas)..

Modifications possibles à apporter au modèle :

• On peut transformer les régresseurs $Z^{(1)}, \cdots, Z^{(p)}$ par toutes les transformations algébriques ou analytiques connues (fonctions puissances, exponentielles, logarithmiques,...), pourvu que le nouveau modèle reste interprétable. Cela peut permettre d'améliorer l'adéquation du modèle ou de diminuer son nombre de termes.

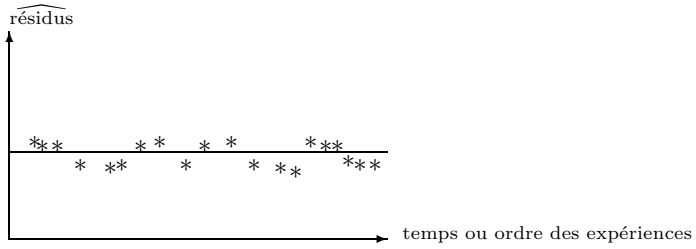
• En revanche, on ne peut envisager de transformer Y, que si les graphiques font suspecter une hétéroscedasticité. Dans ce cas, cette transformation doit obéir à des règles précises basées sur la relation suspectée entre l'écart type résiduel σ et la réponse Y. C'est ce que précise le tableau ci-dessous :

Nature de la relation	Domaine pour Y	Transformation
$\sigma = (\text{cte})Y^k$, $k \neq 1$	$\mathbb{R}_+^*$	$Y \mapsto Y^{1-k}$
$\sigma = (\text{cte})\sqrt{Y}$	$\mathbb{R}_+^*$	$Y \mapsto \sqrt{Y}$
$\sigma = (\text{cte})Y$	$\mathbb{R}_+^*$	$Y \mapsto \log Y$
$\sigma = (\text{cte})Y^2$	$\mathbb{R}_+^*$	$Y \mapsto Y^{-1}$
$\sigma = (\text{cte})\sqrt{Y(1-Y)}$	$[0,1]$	$Y \mapsto \arcsin\left(\sqrt{Y}\right)$
$\sigma = (\text{cte})\sqrt{1-Y}.Y^{-1}$	$[0,1]$	$Y \mapsto (1-Y)^{1/2} - 1/3(1-Y)^{3/2}$
$\sigma = (\text{cte})(1-Y^2)^{-2}$	$[-1,1]$	$Y \mapsto \log(1+Y) - \log(1-Y)$

Souvent ces situations correspondent à des modèles précis. Par exemple, la cinquième transformation correspond le plus souvent à des données de comptage. Dans le cas où les effectifs observés sont faibles (de l'ordre de la dizaine), on aura plutôt interêt à utiliser un modèle plus précis basé sur des lois binomiales. Il s'agit alors d'un modèle linéaire généralisé. D'ailleurs toutes les situations issues d'une des transformations ci-dessus peuvent être traitées par modèle linéaire généralisé. Il n'entre pas dans le champ de ce cours de préciser ces modèles. Notons cependant que pour des grands échantillons la transformation de Y peut suffit à transformer le modèle en un modèle linéaire classique et est beaucoup plus simple à mettre en œuvre. Par exemple, dans une étude bactériologique sur des désinfectants dentaires, on mesure le degré d'infection d'une racine dentaire en comptant les germes au microscope électronique. Sur les dents infectées, le nombre de germes est élevé et variable. L'écart-type est proportionnel à la racine carrée de la réponse. Une loi ayant cette propriété est la loi de Poisson, qui donne alors lieu à un modèle linéaire généralisé. Toutefois, si les comptages sont en nombre important, travailler directement avec pour donnée la racine carrée du nombre de germe peut répondre tout aussi bien à la question.

Pour vérifier le postulat P3 : indépendance

Un graphe pertinent pour s'assurer de l'indépendance des résidus entre eux et celui des résidus estimés $\widehat{\varepsilon}_i$ en fonction de l'ordre des données (lorsque celui-ci a un sens, en particulier s'il représente le temps). Par exemple, on peut obtenir le graphe suivant :



Un graphique comme celui ci-dessus est potentiellement suspect car les résidus ont tendance à rester par paquets lorqu'ils se trouvent d'un côté ou de l'autre de 0. On pourra confirmer ces doutes en faisant un test de runs. Ce test est basé sur le nombre de runs, c'est-à-dire sur le nombre de paquets de résidus consécutifs de même signe. Sur le graphique ci-dessus, il y a 8 runs. On trouve les références de ce test dans tout ouvrage de tests non-paramètriques ou dans un livre comme celui de Draper et Smith, (p. 157).

Par ailleurs, si les erreurs sont correlées suivant certaines conditions (par exemple si les résidus sont des processus ARMA), il est tout d'abord possible d'obtenir encore des résultats quand à l'estimation des paramètres, mais il existe également des méthodes de correction (on peut penser par exemple à des estimations par moindres carrés généralisés ou pseudo-généralisés; voir par exemple Guyon, 2002). Ceci dépassant le cadre de ce cours, nous n'entrerons pas plus dans les détails.

2 Trouver la bonne régression

2.1 Erreur sur les régresseurs

La théorie du modèle de régression telle que nous l'avons vue dans le chapitre précédent, suppose que Y est aléatoire et que les régresseurs $Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}$ sont connus et non aléatoires. Les résultats peuvent s'étendre sans difficulté au cas où les régresseurs sont aléatoires, à condition qu'ils soient indépendants des erreurs ε. Il y a alors deux situations à considérer. Pour en donner une illustration, considérons l'exemple d'un phénomène chimique dépendant de la température T d'un bain suivant une fonction connue. Deux cas sont alors possibles :

- **Cas 1 :** On ne sait pas bien fixer la température du bain mais on sait très bien la mesurer. On avait prévu une expérience à 10^0C, 20^0C et 30^0C mais, en fait, on a effectivement des températures de 11^0C, 20^0C et 28^0C. Il est clair que la vraie variable pertinente est la température réelle et non pas celle planifiée. Si on introduit cette dernière dans la régression, on obtiendra un modèle parfaitement correct, bien que la température réelle ait un certain caractère aléatoire.

- **Cas 2 :** On ne sait ni fixer ni mesurer précisément la température. Dans ce cas, on ne possède pas un régresseur vraiment pertinent (la vraie température) et en toute rigueur, on n'a pas

le droit de poser un modèle de régression. Cependant, en pratique, on acceptera de le faire si l'erreur de mesure est plus faible que l'erreur d'ajustement du modèle.

2.2 Un cas particulier de régression : l'étalonnage ou calibration

Soit le problème suivant :

On considère la variable Z représentant l'âge d'un fœtus et la variable T, la longueur du fémur mesurée par échographie. On veut prédire Z à partir de T. Pour cela, on réalise une première expérience dite d'étalonnage : on constitue un échantillon de femmes enceintes à cycle menstruel régulier, ce qui permet de connaître précisément la date de conception du et donc l'âge du fœtus. On mesure alors à l'aide d'une échographie la longueur des fémurs de leur fœtus. Dans la deuxième expérience dite de prédiction, on utilise l'expérience précédente pour prédire l'âge de d'autres fœtus à partir de la mesure de la longueur fémorale.

Quel modèle utiliser ? Une première réponse, trop naïve, est la suivante :

La variable à expliquer est l'âge Z et la variable explicative est la longueur fémorale T.

Bien que l'on puisse adopter ce point de vue sous certaines conditions, il faut bien être conscient que ce n'est pas la démarche normale. En effet dans l'expérience d'étalonnage l'âge est parfaitement connu. Par contre pour un âge donné, les longueurs fémorales se répartissent autour d'une valeur moyenne, à cause de la variabilité génétique de la population humaine considérée. Pour respecter les postulats **P1-4** du modèle linéaire il vaut mieux écrire

$$\text{longueur du fémur} = T = \alpha + \beta.Z + \varepsilon.$$

Dans la deuxième phase, l'équation de régression sera inversée. En effet

$$T = \alpha + \beta.Z \Longleftrightarrow Z = \frac{T - \alpha}{\beta}$$

(lorsque $\beta \neq 0$, mais c'est ce que l'on suppose intrinséquement), ce qui permet d'estimer l'âge Z à partir de la longueur T du fémur.

L'intérêt du cas particulier de l'étalonnage et de bien illustrer ce que veut dire un modèle de régression.

2.3 Choisir parmi les régresseurs

Dans de nombreux problèmes concrets, on ne désire pas conserver tous les régresseurs, mais plutôt éliminer tous ceux qui n'apportent pas d'explication supplémentaire pour la variable à expliquer Y.

La régression avec l'ensemble des régresseurs donne un test de student portant sur la nullité des coefficients de chaque variables. Ce test correspond à l'hypothèse d'enlever une variable en conservant toutes les autres. Cependant, on se heurte alors au problème suivant : dans le cas général, dès que l'on enlève une variable, tous les tests de student des autres variables sont modifiés.

La difficulté pour éliminer certaines variables explicatives régresseurs vient souvent de la possible colinéarité entre les différents régresseurs. Imaginons que deux variables $Z^{(1)}$ et $Z^{(2)}$ soient presque colinéaires et très liées avec la variable à expliquer Y.

- dans le modèle avec $Z^{(1)}$ seul, $Z^{(1)}$ est significatif ;
- dans le modèle avec $Z^{(2)}$ seul, $Z^{(2)}$ est significatif ;
- dans le modèle avec les deux variables, aucune des variables explicatives n'est significative : on peut en enlever une du moment que l'autre reste.

Trouver le meilleur ensemble de régresseurs sera difficile pour deux raisons. Tout d'abord, si l'on considère un total de k regresseurs, il y 2^k sous-modèles possibles. Dans Tomassone *et al.* (1983), on pourra consulter l'exemple traité sur la chenille processionnaire du pin, dans lequel on dispose initialement de 10 régresseurs, soit 1024 sous-modèles possibles. Ensuite, s'il y a une colinéarité entre différents régresseurs, plusieurs sous-modèles peuvent donner des résultats identiques. Aussi, va-t-on commencé par mesurer la colinéarité entre les régresseurs, ce qui peut être fait à partir des deux étapes suivantes :

i. en premier lieu, qualitativement, on peut observer si les tests de student de nullité des coefficients des différentes variables sont modifiés suivant que l'on change de modèle, auquel cas, il peut effectivement y avoir un problème de colinéarité entre les régresseurs ;

ii. ensuite, et si l'étape précédente conduit au soupçon de colinéarité, on mesure quantitativement la colinéarité d'un régresseur par rapport aux autres par le VIF (Variance Inflation Factor, facteur d'augmentation de la variance). Soit $Z^{(i)}$ le ième régresseur parmi k autres régresseurs. On peut alors effectuer une régression de $Z^{(i)}$ par les autres régresseurs. On calcule alors le coefficient de détermination R_i^2 (pourcentage de variance expliquée) de $Z^{(i)}$ régressé sur les autres régresseurs. Soit alors $\widehat{Z}^{(i)}$ l'estimation de $Z^{(i)}$ par une telle régression. On a donc :

$$R_i^2 = \frac{||\widehat{Z}^{(i)} - \overline{Z}^{(i)}||^2}{||Z^{(i)} - \overline{Z}^{(i)}||^2} \text{ obtenu par régression de } Z^{(i)} \text{ par les } Z^{(j)},\ j \neq i.$$

Ce coefficient est aussi le rapport de la variance résiduelle par la variance totale. On définit ainsi le VIF associe à $Z^{(i)}$ par l'expression :

$$VIF_i = \frac{1}{1 - R_i^2}.$$

C'est une quantité toujours supérieure à 1. Elle vaut 1 quand le régresseur n'est pas du tout colinéaire aux autres régresseurs (il est orthogonal) car alors $R_i^2 = 0$. Une valeur supérieure à 10 est considérée comme un signe de colinéarité importante (ceci ne donne qu'une idée imprécise de la situation : cette valeur de 10 est à moduler en fonction du nombre de données et du niveau de confiance qu'un test sur le VIF demanderait. Mais cela nous emmenerait plus loin que l'objectif avoué de ce cours...).

Remarque : Certains cours sur le modèle linéaire définissent également le coefficient TOL (tolerance) qui se définit comme l'inverse du VIF : $TOL_i = \dfrac{1}{VIF_i} = 1 - R_i^2$. On conçoit bien la similitude des résultats obtenus avec l'un ou l'autre des indices.

Nous finirons ce chapitre par des remarques générales concernant l'heuristique d'une modélisation. Pour commencer, nous dirons qu'un modèle de régression peut être explicatif ou prédictif :

- un modèle est explicatif quand il y a une vraie liaison causale (par exemple issue d'une loi physique ou chimique) entre la variable à expliquer et les régresseurs.

- un modèle est seulement prédictif quand il n'a pas la propriété précédente, mais prédit bien la variable à expliquer. Un autre modèle pourrait tout aussi bien convenir.

Toute personne ayant l'habitude de la statistique sait que l'on n'est jamais sûr d'obtenir un modèle explicatif. Prenons pour exemple le cas de la liaison négative existant entre le rendement et le taux de traitement d'une culture (voir Box). Cette liaison est négative car l'on ne traite que quand la parcelle est infectée et donc quand le rendement est bas. La liaison n'est pas causale, car arrêter le traitement n'augmentera pas le rendement, bien au contraire. Dans la recherche d'un modèle, on distingue deux solutions extrêmes entre lesquelles la réalité se trouvera le plus souvent :

- le bon cas : les VIF sont faibles et le test de student d'une variable dans n'importe quel sous-modèle est qualitativement le même. Dans ce cas, on aura l'espoir de trouver un modèle explicatif ;

- le mauvais cas : certains VIF sont élevés et les tests de student d'une variable changent suivant les variables associées. Dans ce cas, on ne pourra qu'escompter trouver un modèle prédictif.

3 Stratégies de sélection d'un modèle explicatif

Les études précédentes du modèle linéaire, nous offre un premier choix de stratégies pour essayer de sélectionner un modèle explicatif par rapport à un autre, ou par rapport à tous les autres. En effet, nous avons vu des techniques pour tester la validité d'un sous-modèle par rapport à un modèle plus "grand". Les critères qui vont guider notre choix sont au nombre de trois :

– le pourcentage de variance expliquée par un sous-modèle, appelé encore le coefficient de détermination :
$$R^2 = \frac{\text{SC totale} - \text{SCR}}{\text{SC totale}} = \frac{||\overline{Y} - \widehat{Y}||^2}{||Y - \overline{Y}||^2},$$
où l'estimation $\widehat{Y}$ est obtenue pour le sous-modèle ;

– le carré moyen résiduel : $\widehat{\sigma}_n = \dfrac{1}{n-k}||Y - \widehat{Y}||^2$, où l'estimation $\widehat{Y}$ est obtenue pour le sous-modèle et avec k est la dimension du sous-modèle ;

– la comparaison de deux modèles emboîtés par la statistique du test F.

Notons que le premier critère privilégie toujours les grands modèles, dans le sens où plus on ajoute de régresseurs, plus la somme des carrés résiduelle $||Y - \widehat{Y}||^2$ diminue, donc plus le coefficient R^2 se rapproche de 1. Il ne permet de comparer que des modèles de même taille.

On suppose donc maintenant que l'on dispose d'un certain nombre de variables explicatives et que seules certaines d'entre elles interviennent dans le vrai modèle. On suppose de plus que ces variables sont peu corrélées entre elles, ou plus précisément que les différents coefficients VIF sont proches de 1. Voyons alors les stratégies possibles pour tenter d'identifier le vrai modèle.

- **Stratégie 1** : régression descendante (conseillée au débutant). On pose le grand modèle (celui avec tous les régresseurs possibles) et ensuite, à chaque étape, on calcule la statistique F correspondant au retrait de chaque variable : (nullité du cœfficient associée). On enlève du modèle la variable associée au F le plus faible. On arrête cette procédure quand le F de la variable enlevée dépasse un certain seuil fixé par l'utilisateur (par exemple la valeur critique au niveau 5%).

- **Stratégie 2** : régression ascendante (déconseillée). On part d'un petit modèle censé représenter relativement bien les données, et à chaque étape on rajoute parmi les régresseurs non utilisés celui qui a le F d'introduction le plus élévé. En d'autres termes, on ajoute le régresseur le plus pertinent, celui qui améliore par exemple le coefficient R^2. On s'arrête quand le F de la variable introduite est inférieur à une valeur donnée par l'utilisateur (qui peut être la valeur critique à 5%). L'inconvénient de cette stratégie est d'être mal fondée théoriquement : en effet si on rajoute des régresseurs au différents modèles, cela implique que tous les modèles sauf le dernier sont faux et donc que tous les calculs faits dans la stratégie ascendante ne sont pas strictement corrects.

- **Stratégie 3** : il existe une stratégie dite régression hiérarchique qui mélange les deux ci-dessus : à chaque étape, on peut ajouter ou enlever un régresseur. Nous ne détaillerons pas.

Ceci concerne donc la recherche d'un modèle explicatif. Cependant, en pratique, ce sera surtout en vue de nouvelles prédictions que l'on cherchera à établir un modèle. Dans ce cadre, on peut se contenter de sélectionner un "bon" modèle prédictif : le chapitre suivant est ainsi consacré à définir des critère de sélection de modèles prédictifs.

Chapitre 5

Critères de sélection de modèles prédictifs

Dans de nombreux problèmes empiriques se pose la question du choix du modèle. En effet, au cours d'une expérience quelconque, on dispose en général de données issues de variables potentiellement explicatives, dont on ne sait pas a priori si elles ont une influence réelle sur la ou les variables d'intérêt. On peut rajouter ainsi toutes les variables intervenant plus ou moins lors de l'expérience et "bricoler" un modèle de plus en plus précis pour s'adapter aux données obtenues. Cependant, que ce soit pour comprendre, expliquer, ou prédire, il est important de ne sélectionner que des variables intervenant vraiment. Si aucune information de nature causale, par exemple une loi de la physique, n'est connue quant au problème considéré, il n'y a aucun a priori à avoir vis-à-vis de toutes les variables intervenant dans l'expérience (la seule limite étant qu'il ne faut pas plus de variables explicatives que de données à étudier). Nous avons vu au chapitre précédent que le critère VIF permettait d'obtenir une information quant aux liens existants entre les variables et ainsi de savoir si l'on dispose vraiment d'un modèle explicatif. Mais le plus souvent, lorsque le nombre de variables potentiellement explicatives devient important, on devra se contenter de déterminer un modèle permettant d'obtenir de bonnes performances en terme de prédiction. Dans le droit fil du chapitre précédent, nous ne considérerons que le cadre du modèle linéaire, qui de toute manière nous paraît incontournable lors d'une première approche, mais il existe un grand nombre d'autres modélisations possibles : des modèles non-linéaires, à cartes, des réseaux de neurones, des modèles additifs, des séries chronologiques,...

1 Sélection d'un modèle prédictif en régression linéaire paramétrique

1.1 Présentation et définition

Hypothèses et notations : On suppose que l'on dispose des n résultats d'une expérience concernant une variable d'intérêt quantitative Y ayant pris n valeurs, $(y_1, y_2, \cdots, y_n)$, et que p variables quantitatives Z_i, $i = 1, \cdots, p$, prenant les valeurs $z_{1,i}, \cdots, z_{n,i}$, peuvent être utilisée pour expliquer Y. Par soucis de concision dans les notations, on supposera que $Z^1 = (1, \cdots, 1)'$, qui correspond à la partie constante de la régression. Enfin, on définit $\mathcal{M}$ un ensemble de modèles, qui est une famille de sous-ensemble de $\{1, \cdots, p\}$. Deux exemples sont à retenir :

- $\mathcal{M} = \mathcal{P}(\{1, \cdots, p\})$: famille exhaustive de modèles ;
- $\mathcal{M} = \{1, \cdots, j\}_{1 \leq j \leq p}$: famille hiérarchique de modèles.

Par la suite, pour $m \in \mathcal{M}$, on notera $|m| = \text{Card}(m)$ et $X^{(m)}$ la matrice d'ordre $(n, |m|)$ telle que $[X_{i_1}, X_{i_2}, \cdots, X_{i_{|m|}}]$, avec $m = \{i_1, \cdots, i_{|m|}\}$.

Hypothèses sur le vrai modèle : On suppose qu'il existe $m^* \in \mathcal{M}$ tel que le vrai modèle s'écrive :

$$Y = \mu^* + \varepsilon^* = X^{(m^*)}.\beta_{(m^*)} + \varepsilon^*, \quad \text{où} \quad \varepsilon^* \sim \mathcal{N}(0, \sigma_*^2 I_n), \tag{5.1}$$

et avec m^*, $\mu^* \in \mathbb{R}^n$, $\beta_{(m^*)} \in (\mathbb{R} \setminus \{0\})^{|m^*|}$, ε^* et σ_* inconnus.

Exemple : Le vrai modèle peut par exemple être :

$$Y = \beta_1 + \beta_3 X_3 + \beta_7 X_7 + \varepsilon^*,$$

ce qui signifie que $m^* = \{1, 3, 7\}$.

Remarque : Observons que :

i. Le bruit est un bruit blanc gaussien (les ε_i^* sont indépendants) ;

ii. On suppose que $\beta_{(m^*)} \in (\mathbb{R} \setminus \{0\})^{|m^*|}$; ceci permet l'écriture unique du modèle. En effet, si $m \in \mathcal{M}$ contient strictement $m*$, alors le vrai modèle pourra également s'écrire $Y = X^{(m)}.\beta_{(m)} + \varepsilon^*$, mais au moins une des coordonnées de $\beta_{(m)}$ sera nulle.

Modèles utilisés : Pour modéliser l'expérience, on utilise la famille de modèles suivante, qui est également en correspondance avec $\mathcal{M}$, soit

$$Y = \mu + \varepsilon = X^{(m)}.\beta_{(m)} + \varepsilon, \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{5.2}$$

avec $m \in \mathcal{M}$, $\mu \in \mathbb{R}^n$, $\beta_{(m)} \in (\mathbb{R} \setminus \{0\})^{|m|}$.

Pour préciser la modélisation, nous utiliserons les définitions suivantes :

Définition : On suppose que m_1, m_2 et m_3 sont trois ensembles non vides et disjoints de $\{1, \cdots, p\}$ et que $m^* = m_1 \cup m_2$. Alors :

- si le modèle utilisé est m_1, on dit que le modèle est sous-ajusté ;
- si le modèle utilisé est $m_1 \cup m_2 \cup m_3$, on dit que le modèle est sur-ajusté.

Exemple : En reprenant l'exemple précédent, le modèle $Y = \beta_1 + \beta_7 X_7 + \varepsilon$ est sous-ajusté, quand le modèle $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$ est lui sur-ajusté.

Rappels : On suppose que l'on choisit le modèle $m \in \mathcal{M}$. Alors :

i. L'estimateur de $\beta_{(m)}$ par moindres carrés ordinaires est :

$$\widehat{\beta}_{(m)} = ((X^{(m)})'.X^{(m)})^{-1}(X^{(m)})'.Y \quad \text{d'où} \quad \widehat{Y}_{(m)} = X^{(m)}.\widehat{\beta}_{(m)}.$$

ii. On utilise également l'estimateur de σ^2 (non biaisé lorsque $m^* \subset m$)

$$\widehat{s}^2 = \frac{1}{(n-|m|)}||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{(n-|m|)}(Y - \widehat{Y}_{(m)})'.(Y - \widehat{Y}_{(m)}).$$

1.2 Distances entre deux modèles

Pour mesurer l'écart entre le vrai modèle (avec m^*) et un modèle candidat (soit $m \in \mathcal{M}$), nous utiliserons les trois distances suivantes.

Distance quadratique

Définition : La distance quadratique entre les modèles M et M^* est :

$$L_2(M, M^*) = \frac{1}{n}\,||X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\beta_{(m)}||^2 = \frac{1}{n}\,(X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\beta_{(m)})'.(X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\beta_{(m)}).$$

Remarquons que le vecteur de paramètres $\beta_{(m)}$ n'est a priori pas plus connu que le vecteur de paramètres $\beta_{(m^*)}$ du vrai modèle. Aussi va-t-on plutôt considérer $L_2(\widehat{M}, M^*)$, où $\widehat{M}$ est obtenu par moindres carrés. Dans le cas où $m = m*$, on a plus précisément

$$\mathbb{E}L_2(\widehat{M}, M^*) = \mathbb{E}\left(\frac{1}{n}\,||X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\widehat{\beta}_{(m)}||^2\right) = \frac{1}{n}\mathrm{Var}\left(\widehat{Y}_{(m)}\right),$$

ce qui, a une constante près, correspond au risque quadratique de l'estimateur $\widehat{Y}_{(m)}$ de μ^*.

Dissemblance de Kullback

Définition : Soit P et P^* deux mesures de probabilité dominées par une même mesure ν. La dissemblance de Kullback entre P et P^* est :

$$K(P, P^*) = \mathbb{E}_{P^*}\left(\log\frac{dP^*}{dP}\right).$$

En particulier, si $p = \dfrac{dP}{d\nu}$ et $p^* = \dfrac{dP^*}{d\nu}$, alors, $K(P, P^*) = \begin{cases} \displaystyle\int p^* \log\frac{p^*}{p} d\nu & \text{si } P^* << P; \\ 0 & \text{sinon.} \end{cases}$

On montre par des arguments de convexité que $K(P, P^*) \geq 0$ pour toutes mesures P et P^*, et que $K(P, P^*) = 0$ si et seulement si $P = P^*$ P presque sûrement.

Remarque : Pour des raisons de simplicité d'écriture, nous utiliserons plutôt la dissemblance entre les modèles M et M^* définie par $K(M, M^*) = \dfrac{2}{n}K(P, P^*)$, où P et P^* sont les mesures de probabilité respectives de Y dans le cadre des modèles M et M^*.

Dans le cadre gaussien initialement proposé, on a considère des densités par rapport à la mesure

de Lebesgue et :

$$\begin{aligned} f_Y^*(y) &= \frac{1}{(2\pi\sigma_*^2)^{n/2}} \exp\left(\frac{1}{2\sigma_*^2}(y-\mu^*)'.(y-\mu^*)\right) \\ f_Y(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{1}{2\sigma^2}(y-\mu)'.(y-\mu)\right) \end{aligned}$$

Ainsi, en considérant que Y suit la densité f^*, on a :

$$K(M, M^*) = \log\left(\frac{\sigma_*^2)}{\sigma^2}\right) - \frac{1}{n}\mathbb{E}\left[\frac{1}{\sigma_*^2}||Y-\mu^*||^2 - \frac{1}{\sigma^2}||Y-\mu||^2\right]$$

Mais comme $\mathbb{E}(Y) = \mu^*$, alors $\mathbb{E}\left[(Y-\mu^*)'.(Y-\mu^*) - (Y-\mu)'\right] = \mathbb{E}(||\varepsilon^*||^2) = n\sigma_*^2$.

De plus, $||Y-\mu||^2 = ||Y-\mu^*||^2 + 2(Y-\mu^*)'.(\mu^*-\mu) + ||\mu^*-\mu||^2$, donc avec ce qui précède

$$K(M, M^*) = \log\left(\frac{\sigma^2}{\sigma_*^2}\right) + \frac{\sigma_*^2}{\sigma^2} - 1 + \frac{1}{\sigma^2}L_2(M, M^*).$$

Par la suite, et contrairement à ce que nous avons pu faire avec la distance quadratique, nous utiliserons la dissemblance de Kullback pour $M = \widehat{M}$. Pour ce faire, on remplacera μ et σ^2 par leurs estimations respectives, $\widehat{\mu}$ et $\widehat{\sigma}^2$ qui dépendent alors de Y et en prenant l'espérance (sous la loi du vrai modèle) de cette formule.

1.3 Trois critères pour sélectionner un modèle

Rappels : 1/ **Loi de Fisher :** Soit U_1 et U_2 deux variables indépendantes suivant respectivement des lois $\chi^2(n_1)$ et $\chi^2(n_2)$ avec $(n_1, n_2)(\mathbb{N}^*)^2$. Alors

$$F = \frac{U_1/n_1}{U_2/n_2} \quad \text{suit une loi de Fisher} \quad F(n_1, n_2),$$

telle que $\mathbb{E}F = \dfrac{n_2}{n_2-2}$ si $n_2 > 2$ et $\mathrm{Var}F = \dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}$ si $n_2 > 4$.

2/ **Théorème de Cochran :** Soit $X = (X_1, \cdots, X_n)$ un n-échantillon de loi $\mathcal{N}(0, \sigma^2)$ et soit $||.||$ la norme euclidienne classique. Alors :

i. Les composantes de X dans toute base orthonormée de $\mathbb{R}^n$ forment un n-échantillon de loi $\mathcal{N}(0, \sigma^2)$.

ii. Si $\mathbb{R}^n = E_1 + \cdots + E_p$ où les E_i sont p sous-espaces vectoriels deux à deux orthogonaux et de dimensions respectives $d_1, \cdots, d_p$, alors si on note $P_i(X)$ le projeté orthogonal de X sur E_i pour $i = 1, \cdots, p$, les $P_i(X)$ sont indépendants, ainsi que les $||P_i(X)||^2$, dont chacun a pour loi $\sigma^2.\chi^2(d_i)$.

Minimisation de la distance quadratique

Une idée naturelle pour sélectionner le "meilleur modèle" est de minimiser le risque quadratique.

Cela revient aussi à minimiser l'espérance de la distance L_2 définie un peu plus haut. Dans le cadre dans lequel nous nous sommes placés, pour un modèle $m \in \mathcal{M}$, le risque quadratique s'écrit :

$$\begin{aligned} \mathbb{E}\,||\widehat{Y}_{(m)} - \mu^*||^2 &= n\mathbb{E}\,L_2(\widehat{M}, M^*) = \mathbb{E}\,||X^{(m)}.\widehat{\beta}_{(m)} - \mu^*||^2 \\ &= \mathbb{E}\,||X^{(m)}.\widehat{\beta}_{(m)} - \mu^*_{(m)}||^2 + \mathbb{E}\,||\mu^* - \mu^*_{(m)}||^2, \end{aligned}$$

en appelant $\mu^*_{(m)}$ le projeté orthogonal de μ^* sur le sous-espace vectoriel engendré par les X^i où $i \in m$ (sous-espace que nous noterons $< X^{(m)} >$ par la suite), la dernière égalité découlant du Théorème de Pythagore. Par linéarité de la projection et de l'espérance, il est clair que $\mathbb{E}\,\widehat{Y}_{(m)} = \mu^*_{(m)}$. En conséquence, comme $\widehat{Y}_{(m)}$ est la projection orthogonale de Y sur $< X^{(m)} >$, alors :

$$\begin{aligned} \mathbb{E}\,||\widehat{Y}_{(m)} - \mu^*||^2 &= \mathbb{E}\,||\varepsilon^*_{(m)}||^2 + \mathbb{E}\,||\mu^* - \mu^*_{(m)}||^2 \\ &= |m|\sigma_*^2 + ||\mu^* - \mu^*_{(m)}||^2, \end{aligned}$$

le membre de droite de cette égalité ne nécessitant pas d'espérance puisqu'il est déterministe. Notons que $k\sigma_*^2$ aurait aussi été obtenu dans un cadre non gaussien (ici le Théorème de Cochran peut être utilisé).

Remarque : On obtient ainsi l'expression générale du risque quadratique pour un modèle quelconque. Remarquons qu'il est composé de deux parties comportant des paramètres inconnus. Il n'est donc pas possible de minimiser directement ce risque. On observe également que la première partie est un terme de variance qui augmente avec la dimension du modèle choisi, quant la deuxième partie est un terme de biais, qui diminue quand augmente la dimension du modèle, jusqu'à s'annuler quand celle-ci est supérieure ou égale à la vraie dimension du modèle (soit $|m^*|$). On a donc affaire à un compromis biais-variance, qui heuristiquement s'explique par le fait que plus la dimension du modèle augmente, meilleur est l'ajustement du modèle aux données, mais plus importante est la variabilité du modèle (donc la possible erreur lors d'une prédiction ou d'un lissage).

Pour minimiser la distance quadratique L_2 et ainsi sélectionner un "meilleur modèle" suivant ce critère, on est amené à utiliser des estimations des parties biais et variance. En effet, on a :

$$\begin{aligned} \mathbb{E}\,||Y - \widehat{Y}_{(m)}||^2 &= \mathbb{E}\,||Y - \mu^*_{(m)}||^2 - \mathbb{E}\,||\widehat{Y}_{(m)} - \mu^*_{(m)}||^2 \quad \text{(Pythagore)} \\ &= \mathbb{E}\,||Y - \mu^* + \mu^* - \mu^*_{(m)}||^2 - \mathbb{E}\,||\widehat{Y}_{(m)} - \mu^*_{(m)}||^2 \\ &= \mathbb{E}\,||Y - \mu^*||^2 + ||\mu^* - \mu^*_{(m)}||^2 - |m|\sigma_*^2 \\ &= (n - |m|)\sigma_*^2 + ||\mu^* - \mu^*_{(m)}||^2 \end{aligned}$$

(ici encore cette expression est vraie dans un cadre non gaussien). Par suite, la partie biais $||\mu^* - \mu^*_{(m)}||^2$ du risque quadratique est estimée puisque par substitution :

$$\begin{aligned} n\mathbb{E}\,L_2(\widehat{M}, M^*) &= |m|\sigma_*^2 + \mathbb{E}\,||Y - \widehat{Y}_{(m)}||^2 - (n - |m|)\sigma_*^2 \\ \text{soit } \frac{\mathbb{E}\,L_2(\widehat{M}, M^*)}{\sigma_*^2} &= 2\,\frac{|m|}{n} - 1 + \frac{\widehat{\sigma}^2_{(m)}}{\sigma_*^2} \end{aligned}$$

en estimant $\mathbb{E}\,||Y - \widehat{Y}_{(m)}||^2$ par $||Y - \widehat{Y}_{(m)}||^2$, ce qui n'est pas illégitime car, par exemple, lorsque $m* \in m$, alors $\widehat{\sigma}^2_{(m)}$ est un estimateur convergent de $\dfrac{\mathbb{E}\,||Y - \widehat{Y}_{(m)}||^2}{n}$. On conserve cependant un terme inconnu dans cette estimation de la distance quadratique, à savoir σ_*, issu de la partie

variance du risque quadratique. Or on sait que $m^* \in \mathcal{M}$, donc en considérant $m_p = \{1, \cdots, p\}$, $\widehat{\sigma}_{(m_p)}$ est un estimateur convergent de σ_* tant que p ne varie pas avec n, ce que nous avons supposé. Ainsi nous utiliserons pour minimiser notre distance quadratique une minimisation du critère appelé Cp de Mallows et valant pour $m \in \mathcal{M}$, :

$$Cp(m) = \frac{\widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m|}{n}.$$

En selectionnera donc le modèle $\widehat{m}$ tel que :

$$\widehat{m} = \min_{m \in \mathcal{M}} \{Cp(m)\}.$$

Ce critère a été introduit par Mallows (voir Mallows 1967 et 1995). Nous étudierons ses propriétés un peu après.

Minimisation de la dissemblance de Kullback

Comme précédemment, la sélection du modèle peut se faire en minimisant une distance, ici la dissemblance de Kullback, entre un modèle candidat et le vrai modèle. Mais comme précédemment, cela ne pourra être fait qu'après certaines approximations. Soit $m \in \mathcal{M}$, la dissemblance de Kullback entre la loi de l'estimation issue de ce modèle et le vrai modèle :

$$K(\widehat{M}, M^*) = \mathbb{E}_{M^*}\left[\log\left(\frac{\widehat{\sigma}^2_{(m)}}{\sigma^2_*}\right) + \frac{\sigma^2_*}{\widehat{\sigma}^2_{(m)}} - 1 + \frac{1}{\widehat{\sigma}^2_{(m)}} L_2(\widehat{M}, M^*)\right].$$

Le calcul de cette espérance étant dans le cas général non obtensible, on va faire l'hypothèse que $m^* \in m$. Alors, sous l'hypothèse gaussienne :

$$\sigma^2_{(m)} = \frac{1}{n}||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{n}||P_{(m)^\perp}\varepsilon^*||^2 \sim \frac{\sigma^2_*}{n}\chi^2(n - |m|)$$

d'après le Théorème de Cochran et en appellant $P_{(m)^\perp}$ l'opérateur de projection orthogonale sur $< X^{(m)} >^\perp$. De même,

$$L_2(\widehat{M}, M^*) = \frac{1}{n}||\widehat{Y}_{(m)} - \mu^*||^2 = \frac{1}{n}||P_{(m)}\varepsilon^*||^2 \sim \frac{\sigma^2_*}{n}\chi^2(|m|)$$

et d'après ce même Théorème de Cochran,

$$L_2(\widehat{M}, M^*) \text{ est indépendant de } \sigma^2_{(m)}.$$

Par suite,

$$K(\widehat{M}, M^*) = \mathbb{E}_{M^*}\left[\log\left(\widehat{\sigma}^2_{(m)}\right)\right] - \log\left(\sigma^2_*\right) - 1 + \frac{n}{n - |m| - 2} + \frac{|m|}{n - |m| - 2},$$

d'après les expressions de l'espérance d'une variable suivant une loi de Fisher. Il nous reste à déterminer l'expression de $\mathbb{E}_{M^*}\left[\log\left(\widehat{\sigma}^2_{(m)}\right)\right]$. Comme on a supposé que $m^* \in m$, cela est possible, mais de ne sera pas forcément probant car dépendant de σ^2_*. On va donc préférer utiliser un

estimateur de cette espérance, à savoir $\log\left(\widehat{\sigma}^2_{(m)}\right)$, qui converge. En enlevant les parties constantes, on voit que minimiser $K(\widehat{M}, M^*)$ revient approximativement à minimiser le critère suivant, noté AIC_c (pour AIC corrigé), établi en 1989, et qui s'écrit pour $m \in \mathcal{M}$,

$$AIC_c(m) = \log\left(\widehat{\sigma}^2_{(m)}\right) + \frac{n + |m|}{n - |m| - 2}.$$

Le modèle sélectionné sera donc le modèle $\widehat{m}$ tel que :

$$\widehat{m} = \min_{m \in \mathcal{M}} \{AIC_c(m)\}.$$

Lorsque $n \to \infty$, on peut trouver un équivalent au critère AIC_c, puisque :

$$\frac{n + |m|}{n - |m| - 2} \sim 1 + 2\frac{|m| + 1}{n}.$$

Il conviendra donc asymptotiquement et approximativement à minimiser le critère noté AIC, tel que :

$$AIC(m) = \log\left(\widehat{\sigma}^2_{(m)}\right) + 2\frac{|m| + 1}{n}.$$

Ce critère, à la constante n près, à été introduit par Akaiké en 1973 (AIC pour Akaiké Information Criterium) qui l'a généralisé sous la forme $AIC(m) = -2 \times \log(\text{Vraisemblance}) + 2 \times$ Nombre de paramètres, ce qui ne le limite plus au cas gaussien. Cependant, ce critère donne de très mauvais résultats pour n petit, et on préférera donc utiliser AIC_c.

Plus généralement, la sélection de modèle se fait en utilisant un critère de type vraisemblance pénalisée, c'est-à-dire que pour un modèle $m \in \mathcal{M}$, il s'écrira sous la forme :

$$\text{Crit}(m) = \text{Vraisemblance} + \text{Pen}(|m|),$$

où la vraisemblance dépendra en général de $\widehat{\sigma}^2_{(m)}$ et décroîtra lorsque $|m|$ croîtra, et la pénalisation Pen est une fonction croissante de $|m|$. Le critère BIC (Bayesian Information Criterium) par exemple (voir Akaike, 1978, ou Schwarz, 1978), s'écrira sur le même modèle, puisque

$$BIC(m) = \log\left(\widehat{\sigma}^2_{(m)}\right) + \frac{\log n}{n}|m|.$$

Ce critère (que l'on va également minimiser) est obtenu à partir d'a priori sur le modèle et à partir d'unedécomposition de sa vraisemblance. Nous n'entrerons pas plus dans les détails, car, si ce critère donne de très bons résultats, son écriture n'est pas forcément très bien justifiée. Le terme en $\log n$ a été remplacé par $2\log\log n$ par Hannan et Quine (1979), et par une suite (c_n) telle que $(\log\log n)^{-1}.c_n \to \infty$ et $n^{-1}.c_n \to 0$ quand $n \to \infty$ par Rao et Wu (1991).

Conclusion : Au final, nous disposons de 3 critères à minimiser pour sélectionner un "bon" modèle prédictif :

$$\begin{aligned}
Cp(m) &= \frac{\widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m|}{n} \\
AIC_c(m) &= \log\left(\widehat{\sigma}^2_{(m)}\right) + \frac{n + |m|}{n - |m| - 2} \\
BIC(m) &= \log\left(\widehat{\sigma}^2_{(m)}\right) + \frac{\log n}{n}|m|
\end{aligned}$$

1.4 Probabilités de sur-ajustement par un critère

Sur-ajuster, on l'a déjà évoqué, c'est choisir un modèle "trop gros", ou plus précisément et en notant Crit le critère utilisé, c'est choisir par ce critère le modèle $m^* \cup m$, où $m^* \cup m \subset \mathcal{M}$, $m \neq \emptyset$ et $m \cap m^* = \emptyset$. La probabilité de sur-ajustement par le modèle $m^* \cup m$ sera donc :

$$P\left[\text{Crit}(m^* \cup m) \leq \text{Crit}(m^*)\right].$$

Nous allons calculer cette probabilité pour deux des critères définis précédemment.

Cp de Mallows

$$\begin{aligned} P\left[\text{Cp}(m^* \cup m) < \text{Cp}(m^*)\right] &= P\left[\frac{\widehat{\sigma}^2_{(m^* \cup m')}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m^*| + |m'|}{n} < \frac{\widehat{\sigma}^2_{(m^*)}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m^*|}{n}\right] \\ &= P\left[\frac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m^* \cup m)}}{\widehat{\sigma}^2_{(m_p)}} > 2\,\frac{|m|}{n}\right] \end{aligned}$$

Or $\dfrac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m^* \cup m)}}{\widehat{\sigma}^2_{(m_p)}} = \dfrac{||Y - \widehat{Y}_{(m^*)}||^2 - ||Y - \widehat{Y}_{(m^* \cup m)}||^2}{||Y - \widehat{Y}_{(m_p)}||^2}$. D'après le Théorème de Pythagore, comme $Y - \widehat{Y}_{(m^*)} = P_{(m^*)^\perp}(\varepsilon^*)$ et $Y - \widehat{Y}_{(m^* \cup m)} = P_{(m^* \cup m)^\perp}(\varepsilon^*)$, et $< X_{(m^* \cup m)} >^\perp \subset < X_{(m^*)} >^\perp$,

$$\begin{aligned} P_{(m^*)^\perp}(\varepsilon^*) &= P_{(m^* \cup m)^\perp}(\varepsilon^*) + P_{(m^*)^\perp \cap (m)}(\varepsilon^*) \\ \Longrightarrow \quad ||P_{(m^*)^\perp}(\varepsilon^*)||^2 &= ||P_{(m^* \cup m)^\perp}(\varepsilon^*)||^2 + ||P_{(m)}(\varepsilon^*)||^2 \end{aligned}$$

car $(m^*)^\perp \cap (m) = (m)$. De plus, comme on a également $\widehat{\sigma}^2_{(m_p)} = \dfrac{1}{n} P_{(m_p)^\perp}(\varepsilon^*)$ et comme $m \subset m_p$ alors m est orthogonal à $(m_p)^\perp$ et donc d'après la définition de la loi de Fisher

$$F = \left(\frac{n - |m_p|}{|m|}\right) \frac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m^* \cup m)}}{\widehat{\sigma}^2_{(m_p)}} \sim F(|m|, n - |m_p|)$$

donc la probabilité de sur-ajustement par le modèle $m^* \cup m$ est :

$$P\left[\text{Cp}(m^* \cup m) < \text{Cp}(m^*)\right] = P\left[F > 2\,\frac{n - |m_p|}{n}\right].$$

Notons que le résultat eu encore été plus simple si on avait considéré l'estimateur $s^2_{(m_p)}$ sans biais de σ^2_* au lieu de $\widehat{\sigma}^2_{(m_p)}$. Par suite, lorsque $n \to \infty$ et avec m et m^* fixés, on obtient que :

$$P\left[\text{Cp}(m^* \cup m) < \text{Cp}(m^*)\right] \underset{n \to +\infty}{\longrightarrow} P\left[C_m > 2|m|\right] \quad \text{où} \quad C_m \sim \chi^2(|m|)$$

d'après la Loi des Grands Nombres.
La probabilité générale de sur-ajustement qui est supérieure à celle par un sur-modèle quelconque est donc strictement positive. Il est donc possible que le modèle sélectionné soit plus "gros" que le vrai modèle.

AIC et AIC corrigé

En utilisant les mêmes hypothèses et arguments que pour le Cp de Mallows, on a :

$$\begin{aligned} P\left[\mathrm{AIC}(m^* \cup m) < \mathrm{AIC}(m^*)\right] &= P\left[\log\left(\frac{\widehat{\sigma}^2_{(m^* \cup m')}}{\widehat{\sigma}^2_{(m^*)}}\right) > 2\,\frac{|m|}{n}\right] \\ &= P\left[\frac{\widehat{\sigma}^2_{(m^* \cup m')} - \widehat{\sigma}^2_{(m^*)}}{\widehat{\sigma}^2_{(m^*)}} > \exp\left(2\,\frac{|m|}{n}\right) - 1\right] \\ &= P\left[F > \frac{n - |m^*| - |m|}{|m|}\exp\left(2\,\frac{|m|}{n}\right) - 1\right] \\ &\quad \text{où} \;\; F \sim F(|m|, n - |m^*| - |m|). \end{aligned}$$

Les calculs pour le critère AIC corrigé sont exactement du même ordre. Ceci nous permet d'évaluer la probabilité asymptotique de sur-ajustement par le modèle $m \cup m$, puisque comme précédemment, lorsque $n \to \infty$ et avec m et m^* fixés,

$$P\left[\mathrm{AIC}(m^* \cup m) < \mathrm{AIC}(m^*)\right] \underset{n \to +\infty}{\longrightarrow} P\left[C_m > 2|m|\right] \;\; \text{où} \;\; C_m \sim \chi^2(|m|)$$

On obtient ainsi le même type de comportement asymptotique de la probabilité de sur-ajustement par un modèle donné, et également de la probabilité générale de sur-ajustement que pour le Cp de Mallows.

BIC

Par les mêmes calculs que ceux utilisés précédemment, on montre que :

$$\begin{aligned} P\left[\mathrm{BIC}(m^* \cup m) < \mathrm{BIC}(m^*)\right] &= P\left[F > \frac{n - |m^*| - |m|}{|m|}\exp\left(\frac{|m|.\log n}{n}\right) - 1\right] \\ &\quad \text{où} \;\; F \sim F(|m|, n - |m^*| - |m|). \end{aligned}$$

Le comportement asymptotique n'est ici plus du tout le même, puisque l'on obtient maintenant que :

$$\begin{aligned} P\left[\mathrm{BIC}(m^* \cup m) < \mathrm{BIC}(m^*)\right] &\underset{n \to +\infty}{\longrightarrow} P\left[C_m > 2|m|.\log n\right] \;\; \text{où} \;\; C_m \sim \chi^2(|m|) \\ &\underset{n \to +\infty}{\longrightarrow} 0, \end{aligned}$$

et donc la probabilité de sur-ajuster par un modèle donné tend vers 0. On vient de montrer assez simplement qu'asymptotiquement le modèle sélectionné a une probabilité positive d'être trop "grand" pour les critères AIC et Cp et nulle pour le crièrte BIC..Il y a donc ici une différence qui se révèlera déterminante, entre du part, les critère AIC et Cp de Mallows, et d'autre part, le critère BIC.

Notons que les résultats que nous venons d'établir l'ont été Sans aucune condition sur les variables explicatives X^j. Les probabilités de sous-ajustement sont elles beaucoup plus délicates à obtenir, puisque $||\mu^* - \mu^*_{(m)}||$ va intervenir dans les calculs. Voyons maintenant plus en détail ce qu'il en est.

1.5 Convergence asymptotique du modèle sélectionné

Nous allons étudier si le modèle choisi par un des critères précédemment définis converge bien vers le "vrai" modèle. Le premier à avoir montrer un tel résultat est Nishii (1984) dont nous allons reprendre l'essentiel des hypothèses et preuves.

Pour commencer, on généralise les définitions des critères en considérant pour $m \in \mathcal{M}$,

$$\begin{aligned} Cp^{(a)}(m) &= n.\frac{\widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + a\,|m| \\ AIC^{(a)}(m) &= n.\log\left(\widehat{\sigma}^2_{(m)}\right) + a|m| \\ GIC(m) &= n.\log\left(\widehat{\sigma}^2_{(m)}\right) + c_n|m|, \end{aligned}$$

avec $a > 0$ et (c_n) une suite de nombres positifs telle que $\lim c_n = +\infty$ et $\lim n^{-1}c_n = 0$ quand $n \to \infty$. Le critère GIC est donc une généralisation du critère BIC. Pour des raisons de simplicité d'écriture, on considérera AIC et non AIC corrigé, et pour permettre l'établissement de lois limites, on a multiplié les critères Cp, AIC et GIC par n).

On définit $\widehat{m}$ le modèle minimisant un critère et les deux ensembles :

$$\begin{aligned} \mathcal{M}_1 &= \{m \in \mathcal{M} \setminus \{m^*\} \;\mid\; m^* \not\subset m\} \\ \text{et } \mathcal{M}_2 &= \{m \in \mathcal{M} \setminus \{m^*\} \;\mid\; m^* \subset m\} \end{aligned}$$

On a le premier résultat suivant :

Lemme 5.1 *Pour $m \in \mathcal{M}_2$, et avec $\xi^{(a)}(k)$ la variable aléatoire telle que $\xi^{(a)}(k) + ak \sim \chi^2(k)$,*

i. $Cp^{(a)}(m^*) - Cp^{(a)}(m) \xrightarrow[n\to+\infty]{\mathcal{L}} \xi^{(a)}(|m| - |m^*|)$;

ii. $AIC^{(a)}(m^*) - AIC^{(a)}(m) \xrightarrow[n\to+\infty]{\mathcal{L}} \xi^{(a)}(|m| - |m^*|)$.

<u>*Démonstration*</u> : 1/ On a pour $m \in \mathcal{M}_2$:

$$\begin{aligned} Cp^{(a)}(m^*) - Cp^{(a)}(m) &= n\,\frac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + a(|m^*| - |m|) \\ &\overset{\mathcal{L}}{\sim} n\,\frac{\chi^2(|m| - |m^*|)}{\chi^2(n - |m_p|)} + a(|m^*| - |m|) \;\text{ (les deux } \chi^2 \text{ étant indépendants),} \\ &\underset{n\to+\infty}{\overset{\mathcal{L}}{\sim}} \chi^2(|m| - |m^*|) + a(|m^*| - |m|) \end{aligned}$$

d'après les résultats déjà vus précédemment.

2/ De même, pour $m \in \mathcal{M}_2$:

$$\begin{aligned} AIC^{(a)}(m^*) - AIC^{(a)}(m) &= n\,\log\left(\frac{\widehat{\sigma}^2_{(m^*)}}{\widehat{\sigma}^2_{(m)}}\right) + a(|m^*| - |m|) \\ &= n\,\log\left(1 + \frac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m)}}\right) + a(|m^*| - |m|) \end{aligned}$$

$$\begin{aligned} &\overset{\mathcal{L}}{\sim} && n\,\log\left(1+\frac{\chi^2(|m|-|m^*|)}{\chi^2(n-|m|)}\right)+a(|m^*|-|m|) \quad (\chi^2 \text{ indépendants}),\\ &\underset{n\to+\infty}{\overset{\mathcal{L}}{\sim}} && \chi^2(|m|-|m^*|)+a(|m^*|-|m|). \end{aligned}$$ ■

Pour continuer, nous avons besoin d'hypothèses sur les variables explicatives.

Hypothèse 1 : *On suppose qu'il existe une matrice M définie positive telle que :*

$$\lim_{n\to\infty}\frac{1}{n}(X^{(m_p)})'.(X^{(m_p)})=M.$$

En pratique, cette condition est assez souvent vérifiée. Ainsi, lorsque chaque X^j est constitué de réalisations indépendantes d'une variable aléatoire X_j de carrés integrable, et les différentes X^j étant indépendantes entre elles, alors la Loi des Grands Nombres montre que $\frac{1}{n}(X^{(m_p)})'.(X^{(m_p)}) \underset{n\to+\infty}{\longrightarrow} D$, où D est une matrice diagonale dont les termes diagonaux sont les $\mathrm{E}((X_j)^2)$.

Considérons maintenant la probabilité p_n qu'un ensemble soit sélectionné par un critère, soit pour $m\in\mathcal{M}$,

$$p_n(m)=P(\widehat{m}=m).$$

Enfin, on peut définir le risque quadratique que le modèle $m\in\mathcal{M}$ soit le modèle sélectionné, soit

$$R_n(m)=\mathbb{E}_{m^*}\left(||\mu^*-\widehat{Y}_{(m)}||^2\,\mathbb{I}_{\{\widehat{m}=m\}}\right),$$

et le risque quadratique du modèle sélectionné, soit :

$$R_n=\mathbb{E}_{m^*}\left(||\mu^*-\widehat{Y}_{(\widehat{m})}||^2\right).$$

Il est clair que :

$$R_n=\sum_{m\in\mathcal{M}}R_n(m). \tag{5.3}$$

On peut alors montrer la double propriété suivante :

Propriété 5.1 *i. Pour $m\in\mathcal{M}_1$, si $\lim_{n\to\infty} n\,p_n(m)=0$ alors $\lim_{n\to\infty}R_n(m)=0$.*
ii. Pour $m\in\mathcal{M}_2$, si $\lim_{n\to\infty}p_n(m)=0$ alors $\lim_{n\to\infty}R_n(m)=0$.

<u>*Démonstration*</u> : On peut encore écrire que $R_n=\mathbb{E}_{m^*}\left(||\mu^*-\widehat{Y}_{(m)}||^2\,\mathbb{I}_{\{\widehat{m}=m\}}\right)$, soit :

$$\begin{aligned} R_n(m) &= \mathbb{E}_{m^*}\left(||\mu^*-\mu^*_{(m)}||^2\,\mathbb{I}_{\{\widehat{m}=m\}}+||\mu^*_{(m)}-\widehat{Y}_{(m)}||^2\,\mathbb{I}_{\{\widehat{m}=m\}}\right)\\ &= ||\mu^*-\mu^*_{(m)}||^2\,p_n(m)+\mathbb{E}_{m^*}\left(||\mu^*_{(m)}-\widehat{Y}_{(m)}||^2\,\mathbb{I}_{\{\widehat{m}=m\}}\right)\\ &= I_1+I_2. \end{aligned}$$

i/ En utilisant l'Hypothèse 1 sur les variables explicatives, on peut écrire que :

$$\begin{aligned} ||\mu^*-\mu^*_{(m)}||^2\,p_n(m) &= \frac{1}{n}\,(\mu^*-\mu^*_{(m)})'.(\mu^*-\mu^*_{(m)})'\,n\,p_n(m)\\ &= \frac{1}{n}\,\beta'_{m_p}.(X^{(m_p)})'.(I_n-Q_{(m)}).X^{(m_p)}.\beta_{m_p}\,n\,p_n(m), \end{aligned}$$

avec I_n la matrice identité d'ordre n et $Q_{(m)}$ la matrice du projecteur orthogonal sur (m). En fait,

$$Q_{(m)} = X^{(m)}.((X^{(m)})'.X^{(m)})^{-1}(X^{(m)})' = X^{(m_p)}D_{(m)}.(D'_{(m)}(X^{(m_p)})'.X^{(m_p)}D_{(m)})^{-1}.D'_{(m)}(X^{(m_p)})',$$

où $D_{(m)}$ est la matrice d'ordre $(p, |m|)$ composée de 0 et de 1 permettant de sélectionner les vecteurs X^j pour $j \in m$ parmi $X^{(m_p)}$. Par exemple, si $p = 5$ et $m = \{1, 3, 4\}$, alors :

$$D_{(m)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Par suite, en utilisant l'Hypothèse 1, on montre que :

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \beta'_{m_p}.(X^{(m_p)})'.X^{(m_p)}.\beta_{m_p} &= \beta'_{m_p}.M.\beta_{m_p} \\ \lim_{n\to\infty} \frac{1}{n} \beta'_{m_p}.(X^{(m_p)})'.Q_{(m)}.X^{(m_p)}.\beta_{m_p} &= \beta'_{m_p}.M.D_{(m)}.(D'_{(m)}.M.D_{(m)})^{-1}D'_{(m)}.M.\beta_{m_p}, \end{aligned}$$

les deux limites étant des réels positifs ne dépendant plus de n. En conséquence,

$$I_1 \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{car} \quad \lim_{n\to\infty} n.p_n(m) = 0.$$

Par ailleurs, on peut facilement majorer I_2 en utilisant l'inégalité de Cauchy-Schwartz, puisque :

$$I_2 \leq \left[\mathbb{E}_{m^*}\left((||\mu^*_{(m)} - \widehat{Y}_{(m)}||^2)^2\right)\right]^{1/2} [p_n(m)]^{1/2},$$

d'après la définition de $p_n(m)$. Or, on a vu que $||\mu^*_{(m)} - \widehat{Y}_{(m)}||^2 \overset{\mathcal{L}}{\sim} \sigma^2_* \chi^2(|m|)$, donc en utilisant la décomposition d'un χ^2 en somme de carrés de gaussiennes, $\mathbb{E}_{m^*}\left((||\mu^*_{(m)} - \widehat{Y}_{(m)}||^4\right) = (|m|^2 + 2|m|)\sigma^2_*$. on en déduit donc que comme $p_n(m) \to 0$ quand $n \to \infty$, alors $I_2 \to 0$ quand $n \to \infty$.

ii/ En utilisant ce qui précède, dans le cas où $m \in \mathcal{M}_2$, alors on a directement $I_1 = 0$ car alors $\mu^*_{(m)} = \mu^*$. En utilisant le même raisonnement pour I_2, on voit bien qu'il suffit d'avoir la condition $p_n(m) \to 0$ pour obtenir que $R_n(m) \to 0$. ■

Nous allons montrer maintenant le résultat le plus important de ce chapitre, à savoir le fait qu'asymptotiquement on n'a pas tendance à sous-ajuster, ce qui implique que pour les critères $C_p^{(a)}$ et $AIC^{(a)}$ le modèle sélectionné pourra être plus "gros" (au sens de l'inclusion) que le vrai modèle, mais pas plus petit, alors que le modèle sélectionné par le critère GIC convergera vers le vrai modèle (sous l'hypothèse émise quant au comportement asymptotiques des régresseurs).

Théorème 5.1 *i. Pour les critères $C_p^{(a)}$, $AIC^{(a)}$ et GIC, pour tout modèle $m \in \mathcal{M}_1$, alors* $\forall h > 0$,

$$\lim_{n\to\infty} n^h.p_n(m) = 0.$$

ii. Pour les critères $C_p^{(a)}$ et $AIC^{(a)}$, pour tout modèle $m \in \mathcal{M}_2$, alors

$$\lim_{n\to\infty} p_n(m) = p(m) = P\left(\xi^{(a)}(|m^*| - |m|) \geq \xi^{(a)}(|m^*| - |l|)\ \forall l \in \mathcal{M}_2 \cup \{m^*\}\right).$$

iii. Pour le critère GIC, pour tout modèle $m \in \mathcal{M}_2$, alors

$$\lim_{n\to\infty} p_n(m) = 0.$$

Démonstration : Nous allons choisir de travailler avec le critère AIC mais comme son comportement asymptotique est le même que celui du Cp de Mallows, on en déduira les résultats pour ce dernier. Pour le critère GIC, les résultats seront obtenus *mutatis mutandis*. Comme de nombreuses étapes ont déjà été effectuées au cours des preuves précédentes, nous ne donnons que les lignes principales de la démonstration.

i. Soit $m \in \mathcal{M}_1$. Alors :

$$\begin{aligned}
p_n(m) &\leq P\left(AIC(m) \leq AIC(m^*)\right) \\
&\leq P\left(n(\widehat{\sigma}^2_{(m)} - \widehat{\sigma}^2_{(m^*)}) \leq n\left[\exp\left(\frac{a(|m^*| - |m|)}{n}\right) - 1\right]\widehat{\sigma}^2_{(m^*)}\right) \\
&\leq P\left(2\left((Y - \widehat{Y}_{(m)}) - (\mu^* - \mu^*_{(m)})\right)'.(\mu^* - \mu^*_{(m)})\right) + \left(||\widehat{Y}_{(m^*)} - \mu^*||^2 - ||\widehat{Y}_{(m)} - \mu^*_{(m)}||^2\right) \\
&\qquad\qquad + ||\mu^* - \mu^*_{(m)}||^2 \leq b_n\,\widehat{\sigma}^2_{(m^*)}\Big) \\
&\leq P\left(X_n + Y_n + c_n \leq b_n\,\widehat{\sigma}^2_{(m^*)}\right),
\end{aligned}$$

où

- le réel b_n est tel que $b_n = n\left[\exp\left(\frac{a(|m^*| - |m|)}{n}\right) - 1\right] \underset{n\to+\infty}{\longrightarrow} b = a(|m^*| - |m|)$;
- la variable aléatoire $X_n = 2\left((Y - \widehat{Y}_{(m)}) - (\mu^* - \mu^*_{(m)})\right)'.(\mu^* - \mu^*_{(m)})\right)$;
- la variable aléatoire $Y_n = \left(||\widehat{Y}_{(m^*)} - \mu^*||^2 - ||\widehat{Y}_{(m)} - \mu^*_{(m)}||^2\right)$;
- et le réel $c_n = ||\mu^* - \mu^*_{(m)}||^2$.

La variable aléatoire X_n est clairement une variable gaussienne car, avec les notations précédentes, $X_n = 2.\left(Q_{(m^\perp)}.\varepsilon^*\right)'.Q_{(m^\perp)}.\mu^* = 2.\left(Q_{(m^\perp)}.\mu^*\right)'.Q_{(m^\perp)}.\varepsilon^* \overset{\mathcal{L}}{\sim} \mathcal{N}(0, \gamma_n^2)$. On peut déterminer le comportement asymptotique de $\gamma_n^2 = 4\sigma_*^2.(\mu^*)'.Q_{(m^\perp)}.\mu^*$, puisque l'on a vu qu'avec l'Hypothèse 1,

$$\frac{1}{n}\,\gamma_n^2 \underset{n\to+\infty}{\longrightarrow} \gamma^2, \quad \text{où} \quad \gamma > 0.$$

Maintenant, on a bien-sûr $\dfrac{X_n}{\gamma_n} \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$ et on peut écrire que

$$\begin{aligned}
p_n &\leq P\left(\frac{X_n}{\gamma_n} \leq -\frac{c_n}{\gamma_n} - \frac{Y_n}{\gamma_n} + \frac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)}\right) \\
\text{ou encore } p_n &\leq P\left(\frac{X_n}{\gamma_n} \leq \left[2n^{1/4} - \frac{c_n}{\gamma_n}\right] + \left[-n^{1/4} - \frac{Y_n}{\gamma_n}\right] + \left[\frac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)} - n^{1/4}\right]\right).
\end{aligned}$$

Or, pour des événements quelconques F, G et H, on a

$$P(F) \leq P(F \cap G \cap H) + P(G^c) + P(H^c)$$

puisque $P(F) = P(F\cap G\cap H) + P(F\cap G\cap H^c) + P(F\cap G^c)$. Donc en considérant les événements :

- $F = " \dfrac{X_n}{\gamma_n} \leq \left[2n^{1/4} - \dfrac{c_n}{\gamma_n}\right] + \left[-n^{1/4} - \dfrac{Y_n}{\gamma_n}\right] + \left[\dfrac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)} - n^{1/4}\right] "$;
- $G = " \left[-n^{1/4} - \dfrac{Y_n}{\gamma_n} \leq 0\right] "$;
- $H = " \left[\dfrac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)} - n^{1/4} \leq 0\right] "$;

on obtient que

$$p_n \leq P\left(\frac{X_n}{\gamma_n} \leq 2n^{1/4} - \frac{c_n}{\gamma_n}\right) + P\left(-n^{1/4} - \frac{Y_n}{\gamma_n} \geq 0\right) + P\left(\frac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)} - n^{1/4} \geq 0\right).$$

Il ne reste plus maintenant qu'à étudier le comportement asymptotique de chacune de ces trois probabilités. En premier lieu, et toujours d'après l'Hypothèse 1, $\dfrac{1}{n}\,||\mu^* - \mu^*_{(m)}||^2 \underset{n\to+\infty}{\longrightarrow} c > 0$, donc $\dfrac{c_n}{\gamma_n} \underset{n\to+\infty}{\sim} \dfrac{c}{\gamma}\,n^{1/2}$ et ainsi $2n^{1/4} - \dfrac{c_n}{\gamma_n} \underset{n\to+\infty}{\sim} -\dfrac{c}{\gamma}\,n^{1/2}$. Comme $\dfrac{X_n}{\gamma_n}$ suit une loi gaussienne centrée réduite, en appliquant l'inégalité de Bienaymé-Tchébitchev,

$$\begin{aligned} P\left(\frac{X_n}{\gamma_n} \leq 2n^{1/4} - \frac{c_n}{\gamma_n}\right) &= P\left(\exp\left(\frac{X_n}{\gamma_n}\right) \leq \exp\left(2n^{1/4} - \frac{c_n}{\gamma_n}\right)\right) \\ &= o\left(e^{-n^{1/4}}\right), \end{aligned}$$

l'espérance de l'exponentielle d'une loi gaussienne centrée réduite étant une constante.
Quand à la seconde probabilité, de la même manière, on a :

$$\begin{aligned} P\left(\frac{Y_n}{\gamma_n} \leq -n^{1/4}\right) &= P\left(\exp\left(-\frac{Y_n}{\gamma_n}\right) \geq \exp\left(n^{1/4}\right)\right) \\ &\leq \mathbb{E}\left[\exp\left(-\frac{Y_n}{\gamma_n}\right)\right] e^{-n^{1/4}}. \end{aligned}$$

Mais $Y_n = \left(Q_{(m^\perp)}.\varepsilon^*\right)'.Q_{(m^\perp)}.\varepsilon^*$ donc

$$\begin{aligned} \mathbb{E}\left[\exp\left(-\frac{Y_n}{\gamma_n}\right)\right] &= \frac{1}{(2\pi.\sigma_*^2)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\,\frac{x'.x}{\sigma_*^2} - \gamma_n^{-1}.\left(Q_{(m^\perp)}.x\right)'.Q_{(m^\perp)}.x\right) dx \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\,x'.\left(I_n + 2\,\sigma_*^2.\gamma_n^{-1}.Q_{(m^\perp)}\right).x\right) dx \\ &= \left|I_n + 2\,\sigma_*^2.\gamma_n^{-1}.Q_{(m^\perp)}\right|^{-1/2} = \left|I_n + 2\,\sigma_*^2.\gamma_n^{-1}.J_{(m^\perp)}\right|^{-1/2}, \end{aligned}$$

où $J_{(m^\perp)}$ est la matrice identité à laquelle on a enlevé $|m|$ unités sur la diagonale (c'est la matrice $Q_{(m^\perp)}$ diagonalisée). Comme $\gamma_n \underset{n\to+\infty}{\longrightarrow} 0$, on en déduit que $\mathbb{E}\left[\exp\left(-\dfrac{Y_n}{\gamma_n}\right)\right] \underset{n\to+\infty}{\longrightarrow} 1$ et donc

$$P\left(\frac{Y_n}{\gamma_n} \leq -n^{1/4}\right) = \mathcal{O}(e^{-n^{1/4}}).$$

Enfin, $\widehat{\sigma}^2_{(m^*)} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} 1$ donc toujours en utilisant l'inégalité de Bienaymé-Tchébitchev,

$$P\left(\frac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)} - n^{1/4} \geq 0\right) \leq \mathbb{E}\left[e^{\widehat{\sigma}^2_{(m^*)}}\right].e^{-b_n.\gamma_n.n^{1/4}} = o(e^{-n^{1/4}}).$$

On vient bien de montrer que $p_n(m)$ décroît plus vite que toute fonction puissance quand $n \to \infty$ et quand $m \in \mathcal{M}_1$.

ii. Soit $m \in \mathcal{M}_2 \cup \{m^*\}$. Alors :

$$\begin{array}{rcl} p_n(m) & = & P(\widehat{m} = m) \\ & = & P\left(AIC(m) \leq AIC(m') \text{ pour tout } m' \in \mathcal{M}_2 \cup \{m^*\}\right) \\ & = & P\left(AIC(m^*) - AIC(m) \geq AIC(m^*) - AIC(m') \text{ pour tout } m' \in \mathcal{M}_2 \cup \{m^*\}\right) \\ & \underset{n\to+\infty}{\longrightarrow} & P\left(\xi^{(a)}(|m| - |m^*|) \geq \xi^{(a)}(|m'| - |m^*|) \text{ pour tout } m' \in \mathcal{M}_2 \cup \{m^*\}\right). \end{array}$$

iii. Enfin, on peut quantifier la probabilité de sur-ajuster pour le modèle sélectionné par le critère GIC. En reprenant le calcul fait précédemment, on obtient que pour $m \in \mathcal{M}_2 \cup \{m^*\}$,

$$\begin{array}{rcl} P\left[\text{BIC}(m) \leq \text{BIC}(m^*)\right] & \underset{n\to+\infty}{\sim} & P\left[C_k > 2(|m| - |m^*|).c_n\right] \text{ où } C_k \sim \chi^2(|m| - |m^*|) \\ & \underset{n\to+\infty}{\longrightarrow} & 0, \end{array}$$

ce qui achève la preuve. ■

Conséquence : Sous l'Hypothèse 1, le critère de sélection de modèle (AIC ou Cp) ne sous-ajuste pas asymptotiquement mais sur-ajuste. De plus, le risque quadratique obtenu à partir du modèle sélectionné est tel que

$$R_n \underset{n\to+\infty}{\longrightarrow} \sum_{m\in\mathcal{M}_2\cup\{m^*\}} R_n(m) \geq \mathbb{E}\left(||\mu^* - \widehat{Y}_{m^*}||^2\right);$$

on n'a pas pu obtenir asymptotiquement qu'un risque quadratique supérieur à celui que l'on aurait obtenu en ayant la connaissance a priori du vrai modèle. En revanche, pour le critère GIC, le modèle sélectionné converge vers le vrai modèle, et d'après la propriété montrée précédemment,

$$R_n \underset{n\to+\infty}{\longrightarrow} |m^*|.\sigma_*^2,$$

qui est le risque quadratique que l'on aurait obtenu si l'on avait eu connaissance a priori du vrai modèle.

2 Sélection de modèle en régression linéaire fonctionnelle gaussienne

2.1 Présentation

On suppose maintenant que le vrai modèle est toujours

$$Y = \mu^* + \varepsilon^*, \text{ où } \varepsilon^* \sim \mathcal{N}(0, \sigma_*^2 I_n), \tag{5.4}$$

mais cette fois-ci

$$\mu_i^* = s^*(x_i) \text{ où } s^* \text{ est une fonction de } I \text{ dans } \mathbb{R},$$

avec $(x_1, \cdots, x_n) \in I^n$ connus. En général, on supposera que I est un intervalle de $\mathbb{R}$, et en particulier $I = [0, 1]$. On suppose également que $s^* \in \mathcal{H}$, où $\mathcal{H}$ est un espace fonctionnel connu,

possédant une base orthonormée dénombrable $(\phi_j)_{j\in\mathbb{N}^*}$ et ainsi il existe une famille unique de réels $\beta = (\beta_j^*)_{j\in\mathbb{N}^*}$ telle que :

$$s^* = \sum_{j=1}^{\infty} \beta_j^*.\phi_j.$$

Remarque : Une grande partie du travail suivant dépendra du choix de l'espace fonctionnel $\mathcal{H}$. Un exemple souvent rencontré est celui d'un espace de Hilbert. En revanche, si $\mathcal{H}$ est de dimension finie, on est ramené au cas paramétrique, puisqu'alors

$$s^*(x_i) = \sum_{j=1}^{p} \beta_j^*.\phi_j(x_i) = \sum_{j=1}^{p} \beta_j^*.x_{ij}.$$

et on retrouve la matrice $X^{(m_p)}$ précédente.

On supposera donc que la famille de modèles utilisée est

$$Y = \mu + \varepsilon, \quad \text{où} \quad \mu = (s(x_i))_{1\leq i\leq n} \quad \text{et} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{5.5}$$

où $s \in \mathcal{H}$ est une fonction inconnue que l'on cherche à estimer. Pour ce faire, on utilisera sa décomposition dans la base $(\phi_j)_{j\in\mathbb{N}^*}$, ce qui revient à estimer la famille $\beta = (\beta_j^*)_{j\in\mathbb{N}^*}$. Cependant, comme l'on ne dispose que d'un nombre n d'observations, on ne peut au mieux estimer qu'un nombre inférieur à n de paramètres. On considérera donc à nouveau une famille finie de sous-ensembles de $\mathbb{N}^*$ dont le cardinal de chacun est inférieur à $\mathrm{rg}(X)$ avec $X = (\phi_j(x_i))_{1\leq i\leq n, j\in\mathbb{N}^*}$, soit

$$\mathcal{M}_n \subset \{m \subset \mathbb{N}^*,\ |m| \leq \mathrm{rg}(X)\} \quad \text{et} \quad |\mathcal{M}_n| < \infty.$$

On estimera donc β par $\widehat{\beta}(m)$ avec $m \in \mathcal{M}_n$, où comme dans le cas paramétrique

$$\widehat{\beta}(m) = ((X^{(m)})'.X^{(m)})^{-1}(X^{(m)})'.Y$$

La différence avec le cas paramétrique est qu'en général on ne peut estimer le bon ensemble m, puisque celui-ci en général ne sera pas de cardinal fini. On ne pourra donc au mieux qu'estimer une fonction approchant la fonction s^*, et qui sera :

$$\widehat{s} = \sum_{j=1}^{|m|} (\widehat{\beta}(m))_j.\phi_{m_j} \quad \text{où} \quad m = \{m_j,\, j = 1, \cdots, |m|\}.$$

2.2 Quelques résultats

Cas des décompositions ℓ^2

On s'intéresse ici aux résultats proposés par Shibata (1981). Par ailleurs, on passera sans repréciser de l'écriture de β sous la forme $(\beta(m))_{1\leq j\leq |m|}$ à celle sous la forme $(\beta_j)_{j\in\mathbb{N}^*}$ suivant le contexte. On commence par supposer que :

Hypothèse ℓ^2 : Les suites $(x_{ij})_{j\in\mathbb{N}^*} = (\phi_j(x_i))_{j\in\mathbb{N}^*}$ appartiennent à ℓ^2 pour tout $i = 1, \cdots, n$, soit $\sum_{j\in\mathbb{N}^*} \phi_j(x_i)^2 < \infty$.

On définit alors la norme $||.||_n$ telle que pour toute suite $\beta = (\beta)_{n\in\mathbb{N}^*} \in \ell^2$:

$$||\beta||_n = \sum_{i=1}^{n} \left(\sum_{j=1}^{\infty} x_{ij}.\beta_j \right)^2 .$$

Cette norme généralise en quelque sorte la norme matricielle précédemment définie. On définit alors le risque quadratique pour $m \in \mathcal{M}_n$:

$$R_n(m) = \mathbb{E}\left(||\widehat{\beta}(m) - \beta^*||_n^2 \right) = \mathbb{E}\left(||\beta^* - \beta_n^{(m)}||_n^2 \right) + |m|.\sigma^2,$$

avec $\beta_n^{(m)} = \text{Argmin}_{\beta\in<X^{(m)}>} ||\beta^* - \beta||_n^2$. On définit alors le "meilleur modèle possible" pour un n fixé par la formule :

$$m_n^* = \text{Argmin}_{m\in\mathcal{M}_n}\ R_n(m).$$

Ce modèle est là-encore inconnaissable, mais on essaye de l'estimer par un critère de type Cp de Mallows, soit pour $m \in \mathcal{M}_n$:

$$S_n(m) = \frac{1}{n}\,||Y - \widehat{\beta}(m)||^2\,(n + 2|m|)$$

et ainsi on estime le meilleur modèle par :

$$\widehat{m}_n = \text{Argmin}_{m\in\mathcal{M}_n}\ S_n(m).$$

On en arrive alors au théorème suivant :

Théorème : Sous les notations et hypothèses précédentes, et si :

i. $\text{rg}\left((X^{(m)})'.(X^{(m)})\right) = |m|$ pour tout $m \in \mathcal{M}_n$;

ii. $\max_{m\in\mathcal{M}_n} |m| = o(n)$;

iii. $\forall \delta \in [0,1[$, alors $\sum_{m\in\mathcal{M}_n} \delta^{R_n(m)} = 0$;

$$\text{alors}\quad \lim_{n\to\infty} \frac{||\widehat{\beta}(\widehat{m}_n) - \beta^*||_n^2}{R_n(m_n^*)} = \lim_{n\to\infty} \frac{||\widehat{\beta}(m) - \beta^*||_n^2}{\inf_{m\in\mathcal{M}_n} ||\widehat{\beta}(\widehat{m}_n) - \beta^*||_n^2} = 1.$$

Ce théorème nous indique donc qu'asymptotiquement on sélectionne un modèle minimisant le risque quadratique pour la famille de modèles choisie. La condition iii. du théorème reste cependant un peu obscure ; sur les exemples suivants, on voit qu'elle peut être facilement vérifiée :

- Si $I = [0,1[$, $x_i = \dfrac{i-1}{n}$ pour $i = 1,\cdots,n$ et $f(x) = \sum_{k=1}^{\infty} \beta_k.x^k$, avec la famille $\mathcal{M}_n = \{\{1,\cdots,k\}\}_{1\le k\le k_n}$ où $k_n = o(n)$ alors le Théorème est vérifié dès que β^* admet une infinité de coordonnées non nulles. Par exemple, pour $f(x) = \log(1+x)$, pour $f(x) = \exp x$,..., mais pas pour f polynôme !
- Si $I =]-1,1[$ et $\mathcal{H}$ est l'ensemble des fonctions de I dans $\mathbb{R}$ qui sont $\mathcal{C}^2$, en décomposant f dans la base des cos et sin

2.3 Estimation adaptative

Le problème des estimations précédentes tient dans le fait que l'on pose une base a priori pour des fonctions appartenant à un espace fonctionnel fixé ; ces estimations sont donc totalement liée au choix de cet espace. De plus le cadre paramétrique et non-paramétrique sont distincts et les résultats donnés ne sont qu'asymptotiques. Nous allons présenter une méthodologie permettant de dépasser ses limitations et d'intégrer l'ensemble de ces problèmes dans un cadre unique.
Cependant, avant d'évoquer en détail ces méthodes, nous allons voir quelques rappels et définitions qui nous seront utiles.

Rappels

Produit scalaire : Soit X un espace vectoriel réel ou complexe. Cet espace X est dit pré-hilbertien s'il existe une forme bilinéaire $(.,.)$ appelée produit scalaire et vérifiant les prorpiétés suivantes :

i. $\forall (x,y) \in X^2$, $\forall \alpha \in \mathbb{C}$, alors $(\alpha.x, y) = \alpha(x,y)$;
ii. $\forall (x,y) \in X^2$, $(x,y) = (y, \overline{x})$;
iii. $\forall (x,y,z) \in X^3$, $(x+y,z) = (x,z) + (y,z)$;
iv. et $\forall x \in X$ alors $(x,x) \geq 0$; $(x,x) = 0 \Longleftrightarrow x = 0$.

Espace de Hilbert : Un espace préhilbertien complet (c'est-à-dire tel que toute suite de Cauchy converge) est appelé un espace de Hilbert.

Exemples : • l'espace $\mathbb{L}^2(S, \mathcal{B}(S), m)$ où S est un espace topologique, avec le produit scalaire

$$(f,g) = \int_S f.\overline{g}.dm.$$

• l'espace $\ell^2(\mathbb{R})$ pour les suites de nombres réels muni du produit scalaire

$$((x_n)_n, (y_n)_n) = \sum_{n=1}^{\infty} x_n.y_n.$$

Définition : Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, H_1 un espace de Hilbert et $Y : \Omega \to H_1$ une fonction aléatoire telle que $Y = Y_s$ où $s \in \mathcal{H}$ un espace de Hilbert, de norme $||.||$. Alors, pour $\mathcal{H}_\alpha$ un sous-espace connu de $\mathcal{H}$, le risque quadratique minimax d'estimation de s sur $\mathcal{H}_\alpha$ est

$$R(\mathcal{H}_\alpha) = \inf_{\widehat{s}=f(Y_s)} \sup_{s \in \mathcal{H}_\alpha} \mathbb{E}\left(||\widehat{s} - s||^2\right), \quad \text{où la fonction } f \text{ est mesurable.}$$

Remarque : On peut écrire que si $\widehat{s} \in \mathcal{H}_\alpha$, $\mathbb{E}\left(||\widehat{s} - s||^2\right) = ||s_{\mathcal{H}_\alpha} - s||^2 + \mathbb{E}\left(||\widehat{s} - s_{\mathcal{H}_\alpha}||^2\right)$, d'où

$$R(\mathcal{H}_\alpha) \leq \sup_{s \in \mathcal{H}_\alpha} \left(\mathbb{E}\left(||\widehat{s} - s||^2\right) + ||s_{\mathcal{H}_\alpha} - s||^2\right).$$

Le risque minimax n'est pas exprimable en général, mais il peut être majoré ou minoré.

Exemple : On se place dans le cadre du problème de régression fonctionnelle,

$$y_i = s(x_i) + \varepsilon_i \quad \text{pour } i = 1, \cdots, n, \quad \text{avec } x_i = \frac{i}{n},$$

et avec les ε_i qui sont des variables gaussiennes centrées indépendantes de même variance σ^2. Pour une fonction s, on note $||.||_n$ la norme telle que :

$$||s||_n = \frac{1}{n}\sum_{i=1}^{n} s^2(x_i).$$

Enfin, on définit l'espace des fonctions holderiennes

$$\mathcal{H}_\alpha(L) = \left\{ s : [0,1] \leftarrow \mathbb{R} \mid |s(x) - s(y)| \leq L.|x-y|^\alpha,\ \forall (x,y) \in [0,1]^2 \right\},$$

avec $\alpha > 0$ et $L > 0$. On peut, après un travail certain..., montrer que le risque quadratique minimax pour cet espace fonctionnel s'écrit :

$$R_n(\mathcal{H}_\alpha(L)) = C.\left(\left(\frac{\sigma^2}{n}\right)^\alpha .L\right)^{\frac{2}{1+\alpha}} \quad \text{où} \quad C > 0.$$

Deux attitudes sont possibles pour estimer la fonction s^*.

i. On peut directement faire des estimations de s^* à partir du sous-espace $\mathcal{H}_\alpha(L)$. Ainsi, on note :

$$\tilde{s}_n = \text{Argmin}_{t \in \mathcal{H}_\alpha(L)} \sum_{i=1}^{n} (y_i - t(x_i))^2,$$

et ainsi $\tilde{s}_n$ dépend de α et de L. On montre alors (Van de Geer, 1990) que :

$$\mathbb{E}(||s^* - \tilde{s}_n||_n^2) \sim C'.R_n(\mathcal{H}_\alpha(L)) \quad \text{où} \quad C' > 0.$$

ii. On peut également estimer s sur un espace plus grand que $\mathcal{H}_\alpha(L)$. Ainsi, en travaillant sur $\mathbb{L}^2([0,1], \mu_n)$ avec $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$, on peut estimer s sur S_m sous-espace de $\mathbb{L}^2([0,1], \mu_n)$ défini pour $m \in \mathbb{N}^*$ par $S_m = \left\langle \left\{ \mathbb{I}_{\left[\frac{j-1}{m}, \frac{j}{m}\right]} \right\}_{1 \leq j \leq m} \right\rangle$. Ainsi, pour $m \in \mathbb{N}^*$, soit :

$$\widehat{s}_m = \text{Argmin}_{t \in S_m} \sum_{i=1}^{n} (y_i - t(x_i))^2 \ .$$

$$\begin{aligned} \text{Alors} \quad \mathbb{E}(||s^* - \widehat{s}_m||_n^2) &= ||s^* - s_m^*||_n^2 + \mathbb{E}(||\widehat{s}_m - s_m^*||_n^2) \\ &= ||s^* - s_m^*||_n^2 + \frac{m}{n}\,\sigma^2. \end{aligned}$$

Ainsi, lorsque s^* appartient à $\mathcal{H}_\alpha(L)$, alors $||s^* - s_m^*||_n^2 \leq \frac{L^2}{m^{2\alpha}}$. Donc, si on minimalise le risque quadratique précédent, soit maintenant $\frac{m}{n}\,\sigma^2 + \frac{L^2}{m^{2\alpha}}$, on obtient l'estimation suivante de m :

$$\widehat{m}_{\alpha,L} = \left(\left(\frac{L^2}{\sigma^2} n\right)^{\frac{1}{1+2\alpha}}\right).$$

En utilisant cette estimation, on obtient encore :

$$\mathbb{E}(||s^* - \widehat{s}_{\widehat{m}_{\alpha,L}}||_n^2) \sim C''.R_n(\mathcal{H}_\alpha(L)) \quad \text{où} \quad C'' > 0.$$

Définition : On dira ainsi qu'un estimateur $\widehat{s}$ est adaptatif dans le sens du risque quadratique minimax si $\widehat{s}$ ne dépend pas de θ, paramètre d'un espace fonctionnel $\mathcal{H}_\theta$, et si
$\mathbb{E}(||s^* - \widehat{s}||^2) \leq C.R(\mathcal{H}_\alpha)$ où $C > 0$.

Chapitre 6

Problèmes spécifiques à l'analyse de la variance

1 Cadre général

Comme nous l'avons vu précédemment, l'analyse de la variance consiste à expliquer une variable quantitative Y par un certain nombre de variables qualitatives ou facteurs. En voici deux exemples que nous écrivons déjà avec les notations (claires) que nous utiliserons ultérieurement :

- variable à expliquer : rendement ;
  - par 2 facteurs : variété × lieu ;
  - par 3 facteurs : variété × lieu × année ;
  - par 5 facteurs : famille génétique × individu × lieu × année × testeur.
- variable à expliquer : salaire annuel d'un cadre ;
  - par 7 facteurs : domaine d'activité × tranche d'âge × taille de l'entreprise × région d'activité × niveau de diplôme × fonction exercée × sexe.

Comme nous le voyons ci-dessus, un nombre de facteurs élevé peut permettre de s'adapter finement à des situations relativement complexes et il ne faut pas hésiter à les utiliser si on dispose d'un logiciel d'analyse non orthogonale (non équirépété) qui permet de faire les calculs.

Cependant, une nouvelle difficulté se dresse devant nous : il est possible que certains facteurs agissent sur la variable à expliquer de façon conjointe. On dira alors que l'analyse de la variance s'effectue avec des facteurs croisés, ce que nous allons étudier maintenant.

2 Deux facteurs croisés

2.1 Présentation

Commençons par l'exemple suivant : on considère des cultures quelconques et l'on affecte chaque plante d'un indice i pour désigner sa variété (l'indice i varie de 1 à I) et d'un indice j pour désigner le lieu de sa plantation (l'indice j varie de 1 à J). Pour chaque combinaison (i, j), on observe n_{ij} répétitions et on supposera que toutes les combinaisons sont représentées, c'est-à-dire que n_{ij} n'est jamais nul. On pose un modèle général dépendant de la valeur du couple (i, j)

$$Y_{ijk} = \theta_{ij} + \epsilon_{ijk}, \quad \text{où} \tag{6.1}$$

- i est l'indice de variété variant de 1 à I ;
- j est l'indice de lieu variant de 1 à J ;
- k est l'indice de répétition variant de 1 à n_{ij}.

Pour l'instant nous avons posé un modèle qui pourrait se traiter par une analyse de la variance à *un* facteur, ce facteur possédant $I \times J$ modalités. Les différentes valeurs θ_{ij} du facteur θ sont donc estimées par les moyenne empiriques correspondantes, soit :

$$\widehat{\theta}_{ij} = Y_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}.$$

Pour faire apparaître les deux facteurs originels, on définit différents effets par ce que nous appellerons la "décomposition marginale" :

- la moyenne générale $\mu = \theta_{..}$, estimée par $\widehat{\theta}_{..} = Y_{...}$;
- l'effet de la modalité i du premier facteur $\alpha_i = \theta_{i.} - \theta_{..}$, estimé par $\widehat{\theta}_{i.} - \widehat{\theta}_{..}$;
- l'effet de la modalité j du deuxième facteur $\beta_j = \theta_{.j} - \theta_{..}$, estimé par $\widehat{\theta}_{.j} - \widehat{\theta}_{..}$;
- la quantité manquante pour "arriver" à θ_{ij} est appelé l'interaction. En effet, il n'y a pas de raison que l'on ait $\theta_{ij} = \theta_{i.} + \theta_{.j} + \theta_{..}$. On rajoutera donc ce terme d'interaction γ_{ij} tel que :

$$\gamma_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..} = (\theta_{ij} - \theta_{..}) - (\theta_{i.} - \theta_{..}) - (\theta_{.j} - \theta_{..})$$

Le modèle initial : $Y_{ijk} = \theta_{ij} + \epsilon_{ijk}$ s'écrit donc plus précisément sous la forme :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{6.2}$$

avec $k \in \{1, \cdots, n_{ij}\}$ pour $(i, j) \in \{1, \cdots, I\} \times \{1, \cdots, J\}$,, avec les contraintes suivantes que, par définition (et pour rendre unique l'écriture du modèle), on greffe au modèle :

i. on pose $\sum_i^I \alpha_i = 0$ et $\sum_j^J \beta_j = 0$;

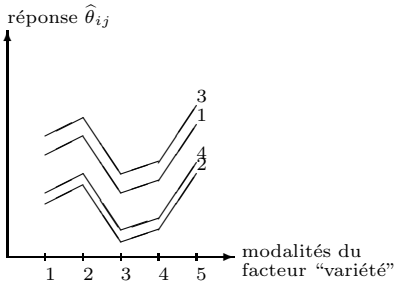
ii. pour tout $j = 1, \cdots, J$, on pose $\sum_i \gamma_{ij} = 0$;

iii. pour tout $i = 1, \cdots, I$, on pose $\sum_j \gamma_{ij} = 0$.

Définition 6.1 *Soit le modèle (6.2) avec les contraintes ci-dessus. Lorsque les paramètres d'interaction γ_{ij} sont nuls pour tout $(i,j) \in \{1,\cdots,I\} \times \{1,\cdots,J\}$, le modèle est dit additif. Sinon, on dit que le modèle est avec interaction, ou encore qu'il possède des facteurs croisés. Enfin, on appelle "effets principaux" les estimations et tests liés aux paramètres α_i et β_j, c'est-à-dire les effets pour chaque facteur pris isolément.*

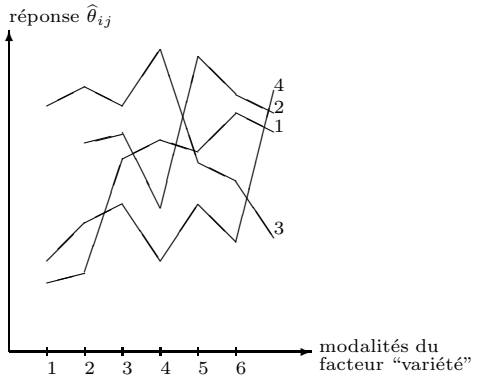
Il y a une grande différence entre les deux types de modèles. En premier lieu, la dimension paramétrique (le nombre de paramètres distincts à estimer) est beaucoup plus faible dans le cas sans interaction que dans le cas avec interaction. En effet, cette dimension vaut, dans le cas du modèle additif, $I + J - 1$ (par exemple si $I = J$=10, dim= 19). Dans le cas du modèle avec interaction, elle vaut $1 + (I-1) + (J-1) + (I-1)(J-1) = I \times J$ (exemple, dim= 100). Ce premier élément nous indique un premier avantage substanciel à supposer le modèle additif.
En second lieu, ces deux types de modèles correspondent à des comportement différents entre les facteurs. On peut ainsi représenter les différentes valeurs de $\widehat{\theta}_{ij}$ du modèle en fonction des modalités i et j des deux facteurs. Illustrons ceci par un exemple dans lequel un facteur (par exemple la variété de la plante) a 6 modalités et l'autre (par exemple son lieu de culture) en a 4 :

Modèle additif



Les courbes sont approximativement parallèles. Cela signifie que pour toutes les modalités de l'un des facteurs, la différence de réponse moyenne $\widehat{\theta}_{ij}$ entre deux modalités fixés de l'autre facteur reste approximativement constante. Par exemple, quelque soit la variété de la plante, la différence de réponse entre les lieux 3 et 1 est constante. Cela peut également s'écrire $\forall i = 1, \cdots, I, \forall j = 1, \cdots, J$, $\widehat{\theta}_{ij} - \widehat{\theta}_{i1} \simeq C_j$; la réponse $\widehat{\theta}_{ij}$ est bien l'addition des effets de i et des effets de j : le modèle est additif.

Modèle avec interaction



Le comportement des différentes courbes ne semble plus présenter de particularité. La réponse $\widehat{\theta}_{ij}$ semble dépendre de chaque composants i et j, les propriétés d'additivité précédentes ne sont plus applicables.

2.2 Modèle additif avec deux facteurs dans le cas équirépété

Nous allons maintenant traiter plus en détail le cas du modèle additif. Remarquons en premier lieu que lorsque $n_{ij} = 1$, alors le nombre de données est insuffisant pour poser le modèle complet avec interaction ; ainsi on a forcément $\gamma_{ij} = 0$ et si cela est vrai pour tout (i, j) on est dans le cadre du modèle additif. Rappelons l'expression de ce modèle, qui est, dans le cas équirépété :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \text{pour} \quad (i, j, k) \in \{1, \cdots, I\} \times \{1, \cdots, I\} \times \{1, \cdots, C\}, \tag{6.3}$$

avec les contraintes :$\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$ (on note $C = n_{ij}$ pour tout (i, j) et $n = I.J.C$ le nombre de données). Ce modèle n'est pas régulier (voir le chapitre 3) et pour donner l'expression explicite de ses estimateurs nous allons faire appel à la notion d'orthogonalité (ici tout ce qui concerne la notion d'orthogonalité est lié au produit scalaire euclidien usuel sur $\mathbb{R}^n$). On définit les espaces E_0, E_1 et E_2 les sous-espaces vectoriels de $\mathbb{R}^n$ suivants :

i. $E_0 = [\mathbb{I}] = \{(\lambda, \cdots, \lambda)' \text{ où } \lambda \in \mathbb{R}\}$.
C'est le sous-espace vectoriel de $\mathbb{R}^n$ formé des vecteurs dont toutes les coordonnées sont égales ;

ii. $E_1 = \{(a_1, \cdots, a_1, a_2, \cdots, a_I)' \text{ où } (a_1, \cdots, a_I) \in \mathbb{R}^I \text{ et } \sum_i a_i = 0\}$.
Notons que l'on répète $J.C$ fois chaque a_i dans l'écriture $(a_1, \cdots, a_1, a_2, \cdots, a_I)'$ ci-dessus. E_1 est le sous-espace vectoriel de $\mathbb{R}^n$ formé des vecteurs dont les coordonnées sont fonctions de i et dont la somme est nulle ;

iii. $E_2 = \{[B, B, \cdots, B]' \text{ où } B = (b_1, \cdots, b_1, b_2, \cdots, b_J) \text{ avec } (b_1, \cdots, b_J) \in \mathbb{R}^J \text{ et } \sum_j b_j = 0\}$.
Notons que l'on répète C fois de suite chaque b_j dans l'écriture de B et que l'on répète I fois B. E_2 est le sous-espace vectoriel de $\mathbb{R}^n$ formé des vecteurs dont les coordonnées sont fonctions de j et dont la somme est nulle.

On a alors les propriétés suivantes qui sont faciles à montrer :

- E_0, E_1 et E_2 sont des sous-espaces orthogonaux entre eux ;
- $E_0 + E_1$ (somme des sous-espaces) correspond au modèle restreint au premier facteur ;
- $E_0 + E_2$ correspond au modèle restreint au second facteur ;
- $E_0 + E_1 + E_2$ correspond au modèle à deux facteurs.

L'orthogonalité entre les trois espaces implique que dans le cas du modèle additif à deux facteurs, l'estimation de Y par moindres carrés est :

$$\widehat{Y} = P_{E_0+E_1+E_2}(Y) = P_{E_0}(Y) + P_{E_1}(Y) + P_{E_2}(Y).$$

Il est clair aussi en reprenant les notations habituelles que :

$$P_{E_0}(Y) + P_{E_1}(Y) = P_{E_0+E_1}(Y) = (Y_{1..}, \cdots, Y_{1..}, Y_{2..}, \cdots, Y_{I..})',$$

où chaque $Y_{i..}$ est répété $J.C$ fois (même raisonnement pour $P_{E_0+E_2}(Y)$). Enfin,

$$P_{E_0}(Y) = Y_{...},$$

moyenne sur toutes les données, qui, dans le cas équirépété, correspond également à la moyenne des différentes $Y_{i..}$ ou des différentes $Y_{.j.}$.

De tout cela, il est facile d'en déduire que

$$\widehat{Y_{ij}} = Y_{...} + (Y_{i..} - Y_{...}) + (Y_{.j.} - Y_{...}) = Y_{i..} + Y_{.j.} - Y_{...}.$$

Tout ceci permet d'en déduire la table d'analyse de la variance suivante :

Source	Somme de carrés	Degrés de liberté	$\widehat{F}$
Facteur 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2$	$I-1$	$\dfrac{(n-I-J+1)\sum_{i,j,k}(Y_{i..} - Y_{...})^2}{(I-1)\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2}$
Facteur 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2$	$J-1$	$\dfrac{(n-I-J+1)\sum_{i,j,k}(Y_{.j.} - Y_{...})^2}{(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2}$
Résiduelle	$\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2$	$n-I-J+1$	

2.3 Modèle avec interaction dans le cas équirépété

On va supposer ici que pour tout (i,j), $n_{ij} = C$ et que le modèle est, a priori, avec interactions (donc n'est pas additif) et donc s'écrit

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

avec les contraintes décrites plus haut. On veut tester les différentes hypothèses stipulant la présence ou non d'un des facteurs ou celle des facteurs croisés. Quantitativement, et en rapport avec ce modèle (6.2) :

- $(H_0^{(1)})$: "tous les coefficients α_i sont nuls" ;
- $(H_0^{(2)})$: "tous les coefficients β_i sont nuls" ;
- $(H_0^{(3)})$: "tous les coefficients γ_{ij} sont nuls".

Pour effectuer ces différents tests, et en reprenant les mêmes notations et la même démarche que dans le cas additif, on commence par noter le sous-espace vectoriel de $\mathbb{R}^n$ correspondant à la part des facteurs croisés :

$E_3 = \{(\gamma_{11}, \cdots, \gamma_{11}, \gamma_{12}, \cdots, \gamma_{IJ})' \text{ avec } (\gamma_{ij})_{ij} \in \mathbb{R}^{I.J} \text{ et } \forall i,\ \sum_j \gamma_{ij} = 0,\ \forall j,\ \sum_i \gamma_{ij} = 0\}$.
Notons que l'on répète C fois de suite chaque γ_{ij} dans l'écriture ci-dessus.

Comme dans le cas du modèle additif, on montre facilement les propriétés suivantes :

- E_0, E_1 et E_2 sont des sous-espaces orthogonaux entre eux ;
- E_0 et E_3 sont des sous-espaces orthogonaux ;
- $E_1 + E_2 \subset E_3$: la part de chacun des deux facteurs et contenu dans celle des facteurs croisés.

En général, E_3 est beaucoup plus "grand" que $E_1 + E_2$ (il faut penser les dimensions en termes de nombre de paramètres utilisés). On peut alors montrer également que :

$$\widehat{Y} = P_{E_0+E_1+E_2+E_3}(Y) = P_{E_0}(Y) + P_{E_3}(Y) = (Y_{11.}, .$$

On a toujours :

$$P_{E_0}(Y) = Y_{...},$$

moyenne sur toutes les données et les mêmes résultats pour $P_{E_0+E_1}(Y)$ et $P_{E_0+E_2}(Y)$. Enfin, on obtient :

$$\widehat{Y_{ijk}} = \widehat{\gamma}_{ij} = (Y_{ij.} - Y_{...}) - (Y_{i..} - Y_{...}) - (Y_{.j.} - Y_{...}) = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$$

On en déduit les statistiques présentées dans le tableau d'analyse de la variance ci-dessous :

Source	Somme de carrés	Degrés de liberté	$\widehat{F}$
Facteur 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2$	$I-1$	$\frac{(n-I.J)}{(I-1)} \frac{\sum_{i,j,k}(Y_{i..} - Y_{...})^2}{\sum_{i,j,k}(Y_{i,j,k} - Y_{ij.})^2}$
Facteur 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2$	$J-1$	$\frac{(n-I.J)}{(J-1)} \frac{\sum_{i,j,k}(Y_{.j.} - Y_{...})^2}{\sum_{i,j,k}(Y_{i,j,k} - Y_{ij.})^2}$
Facteur 1 × Facteur 2	$\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$	$(I-1)(J-1)$	$\frac{(n-I.J)\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}{(I-1)(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2}$
Résiduelle	$\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2$	$n - I.J$	

(pour ne pas alourdir ce tableau, nous n'avons pas complété les carrés moyens qui s'obtiennent naturellement en divisant la somme des carrés par les degrés de liberté). Il est clair que la statistique $\widehat{F}$ présentée pour le Facteur 1 est utilisée pour tester l'hypothèse $(H_0^{(1)})$, et de même avec Facteur 2 et Facteur 1 $\times$ Facteur 2 pour les hypothèses $(H_0^{(2)})$ et $(H_0^{(3)})$.

Remarque : Il est clair que le modèle avec interaction tel que nous l'avons écrit en (6.2) est équivalent au modèle :

$$Y_{ijk} = \mu + \gamma'_{ij} + \varepsilon_{ijk},$$

puisque la projection dans $E_0 + E_3$ est la même que celle dans $E_0 + E_1 + E_2 + E_3$ (on pourrait également montrer que l'on a équivalence également avec le modèle $Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}$, sans conditions sur les θ_{ij}). Cependant, pour mettre en avant les effets propres à chaque facteur (effets principaux), on préférera travailler en général avec le modèle (6.2), sauf si l'un des effets ne se conçoit pas sans l'autre (auquel cas on est dans le cadre d'un modèle dit "hiérarchique" : voir plus loin...).

2.4 Quel modèle choisir ?

Les tests à faire en analyse de la variance ne sont pas uniques et plusieurs écoles existent. En premier lieu il faut d'abord regarder si l'interaction est significative car il est beaucoup plus intéressant d'utiliser un modèle additif pour décrire les données. Notre attitude par rapport à l'interaction sera tout ou rien : ou elle est présente ou elle est totalement absente. Il existe des modèles intermédiaires dits de structuration de l'interaction qui permettent d'éviter un choix si violent, mais qui sortent du cadre de cet ouvrage.

On considère les deux situations suivantes :

Cas 1 L'interaction est significative, c'est-à-dire que l'hypothèse $(H_0^{(3)})$ est rejetée par le fait que le $\widehat{F}$ correspondant n'est pas dans l'intervalle de confiance à niveau choisi (par exemple 95%). Alors il est clair que les deux facteurs sont pertinents et qu'aucun des deux ne peut être enlevé du modèle. Cependant, votre logiciel préféré (quel qu'il soit) vous proposera toujours le test sur les effets principaux (c'est-à-dire sur les hypothèses $(H_0^{(2)})$ et $(H_0^{(3)})$). Ce test n'est en aucun cas le test de l'absence de l'effet puisqu'il est déjà présent dans l'interaction (voir la remarque de la section sur le modèle avec interaction), mais beaucoup d'utilisateurs s'accordent pour dire que ce test a un intérêt descriptif : il permet d'apprécier par l'intermédiaire des $\widehat{F}$, le rapport des ordres de grandeur entre les effets principaux et l'interaction. Cette démarche est à rapprocher de l'estimation de composantes de la variance dans des modèles mixtes.

Cas 2 L'interaction est non significative (donc $(H_0^{(3)})$ est acceptée). Dans ce cas on doit encore tester les effets principaux. La logique voudrait que l'on se place alors dans le modèle additif. Cependant cela est déconseillé car on désire se protéger contre un possible manque de puissance du test de l'interaction. On ne désire pas qu'une faible interaction non décelée vienne fausser l'estimateur $\widehat{\sigma}^2$ de σ^2 comme ce serait le cas dans le modèle additif. Pour ces raisons, on garde l'estimateur $\widehat{\sigma}^2$ du modèle complet et on définit l'absence d'effets principaux comme la nullité des effets α_i et β_j définis dans la décomposition marginale (donc en reprenant les statistiques pour chacun des deux facteurs dans le tableau d'analyse de la variance du modèle avec interaction.

2.5 Différences entre des expériences équirépétées et non-équirépétées

Jusqu'à présent, nous n'avons pas supposé l'équirépétition $n_{ij} = (\text{cte})$. Nous avons cherché délibérement à masquer les difficultés qui apparaissent dans le cas non équirépété. Voyons les

différences entre les deux cas :

– Quand les données sont équirépétées, il y a unicité de la définition des effets principaux et il y a unicité de la table d'analyse de la variance. De plus, les estimateurs sont des moyennes ordinaires, et on a en particulier :

$$\widehat{\theta_{i.}} = Y_{i..} \text{ et donc } \alpha_i = Y_{i..} - Y_{...}$$

Ce genre de calcul peut être mené par des programmes spécialisés très rapides : PROC ANOVA de SAS, ANOVA de GENSTAT. Ces programmes sont adaptés à l'équirépétition, mais il ne tournent pas dans le cas non équirépété.

– Quand les données ne sont plus équirépétées, il n'y a plus unicité de la définition des effets principaux. La définition que nous avons donnée précédemment correspond à un choix (c'est la décomposition dite de type III). Nous avons retenu la solution qui nous paraît être celle adoptée par la majorité des statisticiens. En revanche, la table d'analyse de la variance n'a en aucun cas l'expression simple du cas équirépété.
Plus précisément, dans le cas non équirépété, la somme des carrés associée à l'intéraction garde tout de même une définition unique : elle est définie comme la différence de sommes de carrés résiduelles entre modèle additif et interactif. Cependant les estimateurs du modèle additif n'ont pas d'expression simple sous forme de moyennes ou bien de moyennes de moyennes. Il faut alors résoudre un système d'équations linéaires. Cette somme a donc une expression compliquée. Le lecteur curieux peut consulter le livre de Searle (1987) p. 87-93.
L'expression de la somme des carrés associée au premier facteur dans la table d'analyse de la variance est (par exemple) :

$$SC = \sum_{i=1}^{I} w_i \left(\left[\frac{\sum w_i \widehat{\theta}_{i.}}{\sum w_i} \right] \widehat{\theta}_{i.} \right)^2$$

où, pour chaque $i = 1, \cdots, I$, le poids w_i est défini par le fait que :

$$\text{Var}\left(\widehat{\theta}_{i.}\right) = \frac{\sigma^2}{w_i} \text{ et donc } \frac{\sigma^2}{w_i} = \frac{\sigma^2}{J^2} \sum_{j=1}^{I} \frac{1}{n_{ij}}.$$

La somme des carrés associée à un effet découle du choix que nous avons fait de la définition de cet effet. Elle correspond par exemple pour le premier facteur au test de l'hypothèse

"$\theta_{i.}$ ne dépend pas de i"

Dans la syntaxe de SAS il s'agit des sommes de carrés de type III. L'utilisation de ces sommes de carrés de type III est conseillée dans les cas où les inégalités d'effectifs n'ont pas de sens particulier. Leur intérêt est que l'hypothèse testée ne dépend pas de ces effectif n_{ij}. Or ces effectifs sont souvent inégaux pour des raisons de contraintes expérimentales ou tout simplement à cause de données manquantes. Pour plus de détails sur les différentes décompositions, on consultera Azaïs (1994). Notons que les estimateurs dans le modèle interactif sont construits à partir de moyennes de moyennes. On construit les $\widehat{\theta}_{ij} = Y_{ij.}$ qui sont des moyennes et on fait ensuite des moyennes ($\widehat{\theta}_{i.}$ par ex) de ces moyennes.

3 Extensions

La partie précédente donne les grands traits prérequis à toute analyse de la variance avec deux facteurs. Nous allons commencer par une étude plus précise des résultats obtenus, qui permet d'améliorer la compréhension du modèle.

3.1 Comparaisons multiples

Si on teste la significativité d'un facteur (appelons le "traitement" par exemple), deux résultats peuvent se produire :

- Si l'effet (facteur) "traitement" est non significatif, on n'ira pas plus loin dans l'analyse et on considérera l'expérience comme négative pour ce qui concerne l'influence de ce facteur.

- Si l'effet traitement est significatif, on désire en général pousser l'analyse plus loin en classant les différents traitements ou en les comparant à un témoin.

Si on est en présence de I traitements différents (par exemple $I = 10$), il faut examiner les $I(I-1)/2$ (exemple $= 45$) comparaisons possibles de deux traitements deux à deux. Pour comparer deux traitements définis à priori, par exemple les modalités 1 et 2, on peut utiliser le test de student T général défini lors du chapitre 3. Rappelons d'abord que dans le cas du modèle avec interaction (6.2), on peut toujours écrire ce modèle sous la forme

$$Y = X.\theta + \varepsilon,$$

où $\theta = (\mu, \alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J, , \gamma_{11}, \cdots, \gamma_{IJ})'$ et X une matrice avec des 0 et des 1 d'ordre $(n, I * J)$ (plus les contraintes sur les paramètres). On testera donc la nullité de la combinaison linéaire $\alpha_1 - \alpha_2$ (avec les mêmes notations, la matrice C est telle que $C = (0, 1, -1, 0, \cdots, 0)'$). On utilise ainsi la statistique :

$$\widehat{T}_{12} = \frac{\widehat{\alpha_1} - \widehat{\alpha_2}}{\sqrt{\widehat{\mathrm{Var}}(\alpha_1 - \alpha_2)}} = \frac{Y_{1..} - Y_{2..}}{\sqrt{\widehat{\sigma}^2.C'.(X'.X)^{-1}.C}},$$

où $\widehat{\beta}_1$ est l'estimateur de la valeur du traitement 1, $\widehat{\beta}_2$ est celui du traitement 2, et bien -sûr $\widehat{\sigma}^2 = \dfrac{1}{n - I.J} ||Y - \widehat{Y}||^2$. On sait ainsi que

$$\widehat{T}_{12} \overset{\mathcal{L}}{\sim} T(n - I.J).$$

Ce test de Student est parfaitement valide pour comparer deux traitements choisis a priori. En revanche, il n'est plus du tout utilisable pour comparer le traitement qui donne en apparence les résultats les meilleurs, avec celui qui donne en apparence les résultats les plus mauvais. En effet, chaque test a une probabilité α (le niveau du test) de déclarer présente une différence qui n'existe pas. Au total, sur les $I(I-1)/2$ comparaisons, la probabilité d'en déclarer une significative par "hasard" devient importante. Pour contrôler un risque global sur les $I(I-1)/2$ comparaisons deux à deux ou sur les $(I-1)$ comparaisons à un témoin, il existe diverses méthodes. Commençons par les méthodes de comparaison deux à deux.

Méthodes de comparaison deux à deux :

i. Méthode de Tuckey : elle est adaptée au cas équirépété. Elle donne des intervalles de confiance simultanés pour les différences entre paramètres $\alpha_i - \alpha_j$ où $1 \leq i < j \leq I$ (le risque est global sur les $I(I-1)/2$ comparaisons). Dans le cas équirépété, c'est la méthode la plus précise.

ii. Méthode de Newman et Keuls : elle présente une légère modification de la méthode précédente, mais ell ne donne plus d'intervalles de confiance.

iii. Méthode de Bonferroni : cette méthode est la plus simple et peut être appliquée à tous les cas. Si on à $I(I-1)/2$ comparaisons à faire et que l'on veut un risque global de niveau α, on fait toutes les comparaisons par un test de Student classique, mais au niveau

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

Cette méthode est particulièrement adaptée au cas ou I est petit et le dispositif déséquilibré.

iv. Méthode de Scheffé : c'est une méthode très sûre qui consiste à construire un ellipsoïde de confiance pour le vecteur des paramètres θ ou pour une sous-partie de ce vecteur (par exemple les effets traitements). Cet ellipsoïde de confiance se projette en intervalles de confiance pour les différences. Là encore, on laissera à l'ordinateur le soin d'effectuer les calculs. La méthode de Scheffé a l'avantage (qui se paye par des intervalles légèrement plus grands) de donner des intervalles de confiance pour n'importe quelle combinaison des traitements.

Prenons l'exemple de trois traitements dans un dispositif équirépété que l'on veut tester avec un niveau α :

i. La méthode de Tuckey donne des intervalles de confiance pour : $\alpha_1 - \alpha_2, \alpha_3 - \alpha_2, \alpha_1 - \alpha_3$

ii. La méthode de Bonferroni demande d'otenir des comparaisons deux à deux, mais avec un niveau de confiance de $\alpha' = \alpha/3$.

iii. La méthode de Scheffé donne des intervalles de confiance pour toutes les combinaisons linéaires possibles entre les paramètres. Par exemple, en plus des trois ci-dessus, on peut obtenir un intervalle de confiance pour : $\alpha_1 - 2\alpha_2 + \alpha_3$.

Méthodes de comparaison à un témoin :

i. Méthode de Bonferroni. C'est le même principe que précédemment, mais il y a maintenant $I-1$ comparaisons à effectuer, et on prend donc pour niveau de confiance : $\alpha' = \alpha/(I-1)$.

ii. L'équivalent de la méthode de Tuckey dans le cas équilibré est la méthode de Dunnet, qui construit des intervalles de confiance pour les combinaisons

$$\alpha_i - \alpha_1 \text{ avec } 2 \leq i \leq t,$$

si on suppose que le facteur témoin est le traitement 1.

Tous les détails sur ces différentes méthodes se trouvent dans l'ouvrage de Miller (1966).

3.2 Plusieurs facteurs, facteurs croisés et hiérarchisés

Dans le cas où l'analyse porte sur plus de deux facteurs, on décompose la réponse en : effets principaux, interactions doubles, triples, etc ... et on utilise la même stratégie que précédemment. Un cas particulier doit toutesfois être précisé : il faut distinguer le cas de facteurs croisés et de facteurs hiérarchisés :

Définition 6.2 *Deux facteurs sont dit croisés si chacun d'eux a un sens indépendamment de l'autre.*
Le facteur B est hiérarchisé au facteur A si un indice du facteur B ne signifie rien de concret, tant que l'on ne connait pas l'indice associé du facteur A.

Pour bien distinguer entre ces deux types d'interaction, voici d'abord des exemples typiques de facteurs généralement croisés :

- variété × lieu
- variété × testeur
- concentration × température (réaction chimique).

Voici maintenant des exemples typiques de facteur hiérarchisés :

- famille génétique / numéro de descendant
- bloc / sous-bloc
- lapine / numéro de la portée / numéro du lapin dans la portée.

Le notion de hiérarchie se conçoit par le fait qu'il ne peut pas y avoir d'effet propre du numéro de lapin dans la portée, car il n'y a aucun rapport entre tous les lapins classés n^o4 dans les différents portées. De même, il n'y a aucun rapport entre le sous-bloc n^o2 du premier bloc et le sous-bloc n^o2 du second bloc. Quand on est en présence de facteurs hiérarchisés il faut prendre soin de le déclarer avec la syntaxe propre au logiciel que l'on utilise. En effet, il ne faut pas, dans ce cas, définir un effet propre du facteur hiérarchisé. La décomposition est différente de celle du modèle croisé dans le sens où ce qui serait l'effet principal du facteur hiérarchisé est incorporé à l'interaction. On élimine alors dans le modèle les termes correspondants aux effets principaux seconds dans l'ordre de la hiérarchie, pour obtenir le modèle suivant :

$$Y_{ijk} = \mu + \alpha_i + \gamma'_{ij} + \varepsilon_{ijk}, \tag{6.4}$$

avec les contraintes $\sum_i \alpha_i = 0$, $\sum_j \gamma'_{ij} = 0$ et $\sum_i \gamma'_{ij} = 0$.

3.3 Tester l'inhomogénéité des variances

Les graphes :

– résidus ($\widehat{\varepsilon}_{ij}$) contre le niveau (ici i) de l'unique facteur ;

– résidu ($\widehat{\varepsilon}_{ijk}$) contre réponse estimée ($\widehat{Y_{ijk}}$) (s'il y a plusieurs facteurs) ;

peuvent montrer une plus grande dispersion dans certaines régions de l'expérience ; on suspectera alors la non validité du second postulat, c'est-à-dire une inhomogénéité des variances. Le test classique utilisé dans ce cas-là est le test de Bartlet basé sur une méthode de maximum de vraisemblance. Cependant ce test est déconseillé car il n'est pas robuste à la non normalité. On utilisera donc plutôt le test de Levene (1966) ou sa modification basée sur les carrés. Dans le cas d'un unique facteur, le principe du test est le suivant :

Soit Y_{ij} la j-ème observation de la modalité i. On déduit de l'analyse de la variance les résidus $\widehat{\epsilon}_{ij}$. On effectue alors une analyse de la variance sur les $|\widehat{\epsilon}_{ij}|$. Si les variances sont homogènes, ces quantités doivent être d'espérance $(cte)\sigma$, c'est-à-dire constantes. Sinon, leur valeur "moyenne" varie avec la valeur de i. C'est donc la valeur du test de Fisher appliqué aux valeurs absolues des résidus qui donne le test de l'homoscédasticité. On peut également préférer travailler sur la variable $(\widehat{\epsilon}_{ij})^2$ et dans le cas de plusieurs facteurs, on testera les différents effets principaux.

Comme en régression, les deux seuls remèdes en cas d'inhomogénéité des variances sont :

– les transformations de la variables Y avec les mêmes règles ;

– le recours au modèle linéaire généralisé.

3.4 Plan à mesures répétées : le cas particulier du split-plot

Souvent, la mesure sur une unité statistique (donc pour des niveaux fixés des différents facteurs) n'est pas unique. Que faire alors ? Pour commencer, étudions l'exemple, très simple, qui suit :

Exemple 1 : Test bactériologique sur des fragments de dents (Calas 1993)

On soumet des fragments de dents à une contamination microbienne, puis à une désinfection à l'aide de différents produits (facteur "traitement"). Pour mesurer l'infection résiduelle, on observe au microscope électronique un certain nombre de régions ou "spots" de la dent dont on compte le nombre de germes. Analysons les sources de variabilité de cette expérience :

– les fragments de dents sont différents les uns des autres. C'est la première source de variabilité ;

– le choix du spot dans la dent est aléatoire : on peut "tomber par hasard" sur une zone infectée ou non.

Du fait qu'il peut y avoir plusieurs mesures sur chaque fragment de dent, nous sommes donc en présence de mesures répétées. On dira de manière équivalente que l'on est en présence d'un plan à deux sources d'erreur. Ici la solution sera simple puisque le facteur d'intérêt (le traitement) ne varient pas quand varie l'indice de répétition (les différents spots). On a donc la possibilité de se ramener a un modèle classique en calculant les moyennes par fragment : l'indice de répétition a alors disparu. Voyons maintenant un exemple avec plusieurs facteurs :

Exemple 2 : Moelleux de gâteaux (Cochran & Cox, 1950)

Nous reprenons cet exemple classique dû à Cochran & Cox (1950). On desire optimiser le moelleux de gâteaux au chocolat en fonction de 3 recettes et de la température de cuisson qui prend

6 valeurs de 10°C en 10°C de 175°C à 225°C. On réalise l'expérience suivante : chaque jour on prépare 3 pâtes correspondant à chacune des 3 recettes. La quantité de pâte est suffisante pour pouvoir confectionner 6 gâteaux pour chacune des 3 recettes. Chacun de ces gâteaux est cuit à une température différente. Il y a donc 18 gâteaux en tout. Cette expérience est répétée 15 jours durant. On mesure ensuite le moelleux du gâteau en mesurant l'angle de rupture d'une tranche.

Analyse de variabilité : La description de l'expérience nous indique les deux sources d'erreurs :

- l'erreur de composition de chaque pâte : erreur de pesée et de variabilité de composition des ingrédients : lait, oeufs, farine.
- l'erreur attachée à la mesure de l'angle : erreur de mesure, erreur quant à l'épaisseur de la tranche, erreur sur la température de cuisson.

On peut donc faire la liste des différents facteurs pouvant intervenir sur le moelleux du gateau :

i. les effets principaux, qui sont les facteurs : répétition (rep), recette (rec) et température ($temp$) ;

ii. les facteurs croisés deux à deux : $rec * temp$, $rep * rec$ et $temp * rep$;

iii. les trois facteurs croisés : $rep * rec * temp$;

Pour tenir compte de la répétition des mesures, il n'est pas possible ici de calculer une somme (ou une moyenne) comme dans l'exemple des dents. En effet, cela reviendrait à faire la somme des angles de rupture des 6 gâteaux issus de la même pâte mais cuits à 6 températures différentes, et donc à supprimer le facteur température qui est un facteur d'intérêt. On va donc utiliser une décomposition en strates. Soit P_b le projecteur sur le sous-espace engendré par l'effet bloc qui sera ici l'effet aléatoire le moins fin, $rec * temp$; et soit Q_b le projecteur sur le sous-espace orthogonal à $[rec * temp]$. On scinde le modèle en deux sous-modèles

- l'un portant sur P_bY appelé modèle interbloc ;
- l'autre portant sur Q_bY appelé modèle intrabloc.

On peut alors travailler sur chacun de ces sous-modèles.

i. **Modèle intrabloc :** pour Q_bY, par définition, l'effet $rec * rep$ disparait. On a donc un modèle à une seule source d'erreur ($temp$?). On peut montrer que tout se passe comme dans un modèle linéaire classique. Mais la façon la plus simple de l'analyser est de remarquer que ce modèle est équivalent à celui où l'on suppose le terme bloc : $rec * rep$ comme fixe.[1] Dans ce modèle l'effet rec n'est pas estimable car la projection par Q_b le fait disparaitre. Mais sont totalement estimables l'effet $temp$ et l'interaction $temp * rec$.

ii. **Modèle interbloc :** La projection P_bY revient dans le cas équirépété à calculer les moyennes par bloc, c'est-à-dire par pâte confectionnée. Les deux effets aléatoires $rec * rep$ et $rec * rep * temp$ s'additionnent dans un seul effet qui dépend du couple $rec * rep$. Ce modèle est encore

[1] fixe est à prendre au sens de non-aléatoire.

un modèle d'analyse de la variance classique dans lequel il ne reste qu'un seul effet, l'effet *rec*, dont on pourra estimer le paramètre.

Certains logiciels effectuent ces décompositions de manière automatique. Dans les autres cas, on effectura les deux décompositions "à la main". Dans le cas où le dispositif est déséquilibré, on n'a plus la belle propriété d'estimabilité totale ou de non-estimabilité totale des différents effets *rec*, $rec * temp$ et *temp* dans le modèle inter ou intrabloc. On devra avoir recours à des méthodes plus sophistiquées, notamment celle du maximum de vraisemblance...

4 Exercices

Exercice 1 : Pour introduire des effets différentiels dans un modèle d'analyse de la variance à **un** facteur,

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}$$

quel type de contrainte doit on utiliser ?

$$\sum_{i=1}^{I} \alpha_i = 0 \text{ ou bien } \sum_{i=1}^{I} n_i \alpha_i = 0 \text{ ?}$$

Solution
Par cohérence avec la décomposition marginale utilisée dans le cas où il y a deux facteurs, on serait tenté de poser la première contrainte. Ce n'est pas la bonne réponse. En effet :

- L'estimation et donc la définition précise du paramètre μ importe peu, puisque ce paramètre de moyenne générale est souvent sans intérêt. N'oublions pas que le but d'une expérience est de comparer ; toute l'attention est donc focalisée sur les effets différentiels α_i. Le premier type de contraintes a peu d'intérêt.

- Le second type n'a d'autre intérêt que calculatoire. Si on l'utilise, l'estimateur $\widehat{\mu}$ de μ n'est autre que la moyenne générale : $Y_{..}$, qui est l'estimateur sous l'hypothèse nulle et qui est nécessaire pour la construction du test de Fisher. En particulier

$$SCM = \sum_{ik} (Y_{i.} - Y_{..})^2 = \sum_{ik} (\widehat{\alpha}_i)^2$$

  ne serait pas vrai sinon.

Cette réponse est en contradiction avec celle que l'on fait pour deux facteurs. Nous voyons donc une fois de plus sur cet exemple, l'intérêt de travailler avec des modèles réguliers, comme l'est le modèle de notre première présentation.

Chapitre 7

Analyse de la covariance

Dans de nombreuses situations, l'ensemble des variables explicatives se compose à la fois des variables quantitatives et qualitatives. Supposons par exemple que l'on veuille expliquer la taille de fillettes de 6 à 10 ans en fonction de leur âge.

Pour une fillette donnée, dans la plage d'âge considérée, un modèle de régression linéaire est raisonnable. Par contre il est bien connu qu'il y a des individus plus grand que d'autre et des individus dont la taille va augmenter plus vite. Si on posait un modèle de régression unique, les postulats du modèle linéaire ne seraient pas respectés : par exemple tous les termes ϵ_i associés à un individu grand auraient tendance à être positifs (le premier postulat$E(\epsilon_i) = 0$ ne serait plus vrai). Il faut donc que les paramètres de la régression changent d'un indivudu à un autre. C'est le modèle avec "hétérogénéité des pentes" (qui est le modèle le plus complexe) que l'on va poser. SoitY_{ij} la taille de l'individu i à l'âge numéro j $(i = 1, ..., I\ ; j = 1, ..., J)$ On pose

$$Y_{ij} = \mu_i + \beta_i \text{âge}_j + \epsilon_{ij}.$$

Les questions statistiques qui se posent sont

- (i) est ce que les β_i sont différents ? : hétérogénéité des pentes,
- (ii) est ce que les μ_i sont différents ? : hétérogénéité des constantes.

On peut introduire les effets différentiels :

$$\beta_i = \beta + \gamma_i \quad ; \quad \mu_i = \mu + \alpha_i$$

avec

$$\sum_i \gamma_i = \sum_i \alpha_i = 0$$

pour obtenir le modèle

$$Y_{ij} = \mu + \alpha_i + \beta \text{âge}_j + \gamma_i \text{âge}_j + \epsilon_{ij} \tag{7.1}$$

qui est strictement équivalent au premier. Le dernier terme $\gamma_i \text{âge}_j$ qui dépend des deux variables peut être considéré comme une interaction entre le facteur individu et la variable quantitative âge. La notation informatique du modèle 7.1 est d'ailleurs

taille = constante + individu + âge + âge * individu.

On répond d'abord à la question (i) en testant l'hypothèse :

- pour tout $i = 1, .., I\ ;\ \gamma_i = 0$

Si cet effet est non significatif on peut alors répondre à la question (ii) en testant :

- pour tout $i = 1, .., I, \alpha_i = 0$

Dans ce dernier cas le facteur individu disparaît complètement du modèle. Remarquons que si on admet la nullité des γ_j le modèle (7.1) devient

$$Y_{ij} = \mu + \alpha_i + \beta \text{âge}_j + \epsilon_{ij}$$

qui revient à rajouter la simple covariable âge au modèle d'analyse de la variance à un facteur : individu. Dans les cas les plus généraux ont peut avoir plusieurs facteurs avec une structure croisée ou hiérarchique ainsi que plusieurs variables intervenant de manière linéaire, polynômiale ou plus complexe encore. Le principe est toujours de faire des régressions intragroupe et de faire apparaître des effets différentiels intergroupe.

Chapitre 8

Orthogonalité

Un modelèle linéaire comprend le plus souvent une décomposition naturelle des paramètres θ et conséquemment une décomposition de la matrice X du modèle. On va s'intéresser dans ce chapitre à l'othogonalité éventuelle des espaces associés. Le problème est plus ou moins délicat suivant que le modèles est régulier ou non.

Exemple 8.1 *Condidérons le modèle de régression linéaire multiple sur trois variables $Z^{(1)}, Z^{(2)}$ et $Z^{(3)}$:*

$$Y_i = \alpha + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \beta_3 Z_i^{(3)} + \epsilon_i \quad , \quad i = 1, n > 4.$$

le vecteur θ comprend 4 coordonnées : $\alpha, \beta_1, .., \beta_3$ et la matrice X quatres colonnes. On peut considérer la décomposition, plus précisément on parlera par la suite de partition, en quatres éléments. La partition de la matrice revient alors à l'écrire comme concaténation de 4 vecteurs colonnes. L'orthogonalité del la partition correspondra alors strictement à l'orthogonalité des 4 droites

$$[\mathbb{1}],\ [Z^{(1)}],\ [Z^{(2)}], [Z^{(3)}]$$

Exemple 8.2 *Soit le modèle de régression quadratique sur deux variables $Z^{(1)}$ et $Z^{(2)}$*

$$Y_i = \alpha + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \gamma_1 \big(Z_i^{(1)}\big)^2 + \gamma_2 \big(Z_i^{(2)}\big)^2 + \delta \big(Z_i^{(1)} Z_i^{(2)}\big) + \epsilon_i \quad , \quad i = 1, n > 6.$$

On peut définir la partition correspondant à

- *la constante α*
- *les effets linéaires β_1, β_2*
- *les effets carrés γ_1, γ_2*
- *les effets produits δ.*

L'orthogonalité de la partition est alors définie comme l'orthogonalité des espaces

$$[\mathbb{1}],\ [Z^{(1)} : Z^{(2)}], [\big(Z^{(1)}\big)^2 : \big(Z^{(2)}\big)^2],\ [\big(Z^{(1)} Z^{(2)}\big)].$$

Cette partition sera particulièrement étudiée au chapitre consacré au surfaces de réponses.

En conséquence on voit bien à partir de ces deux exemples **qu'il n'y a pas de modèles orthogonaux mais des partitions orthogonales**. On peut parler de modèle orthogonal seulement dans la cas où, comme dans l'exemple 8.1, on considère la partition la plus fine.

Formalisons ces exemples dans une définition.

Définition 8.1 *Soit le moèle linéaire régulier*

$$Y = X\theta + \epsilon$$

dans $\mathbb{R}^n$ *et considérons une partition en* m *termes*

$$Y = X_1\theta_1 + ... + X_m\theta_m + \epsilon.$$

On dit que cette partiton est orthogonale si les espaces

$$[X_1], ..., [X_m]$$

sont orthogonaux.

La conséquence de l'orthogonalité est la suivante
- Les différents estimateurs $\hat{\theta}_1, ..., \hat{\theta}_m$ sont non-corrélés (indépendants sous l'hypothèse gaussienne).
- l'expression de l'estimateur $\hat{\theta}_l$, $l = 1, ..., m$ ne dépend pas de la présence ou non des autres termes $\theta_{j'}$ dans le modèle.

Ces propriétés découlent directement du caractère bloc-diagonal de la matrice d'information du fait de l'orthogonalité.

Par ailleurs, sous l'hypothèse gaussienne, l'orthogonalité donne une indépendance approximative entre les tests sur les différents effets. Les tests portant sur deux effets orthogonaux ne sont liés que par l'estimation du σ^2.

Un exemple d'application de la seconde propriété est le suivant : soit un modèle de régression multiple et supposons que la partition la plus fine soit orthogonale, alors l'expression de l'estimateur du coefficient β_l de la variable $Z^{(l)}$ vaut

$$\hat{\beta}_i = \sum_{i=1}^{n} \frac{Z_i^{(l)} Y_i}{\left(Z_i^{(l)}\right)^2}.$$

Il suffit pour le vérifier de considérer le modèle où l'on n'a mis que la variable $Z^{(l)}$.

Cependant on veut pouvoir définir une notion d'orthogonalité pour des modèles d'analyse de la variance comme le modèle à deux facteurs croisés qui ne sont pas réguliers.

Exemple 8.3 *Soit le modèle d'analyse de la variance à deux facteurs croisés :*

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

associé à la partion naturelle, μ, α, β, γ. *La notion d'orthogonalité de cette partition ne peut pas se traiter indépendamment du choix du système de contraintes.*

Nous sommes donc amené à poser la définition suivante

Définition 8.2 *Soit le modèle linéaire non régulier :*

$$Y = X\theta + \epsilon \;\; ; \;\; rg(X) = k < p = dim(\theta).$$

on dit que les contraintes $C\theta = 0$, *où* C *est une matrice de dimension convenable, forment des* **contraintes d'identifiabilité** *si ,*
- $rg(C) = p - k$
- $Ker(X) \cap Ker(C) = 0$

Les conditions de la définition impliquent que sur $Ker(C)$, la matrice X est injective et a rang k.

Exemple 8.4 *Soit le modèle suivant d'analyse de la variance à un facteur*

$$Y_{i,j} = \mu_i + \epsilon_{i,j}; \quad i = 1, ..., 4 \quad j = 1, 2$$

Ce modèle est régulier de dimension 4 , mais si on le paramétrise avec des effets différentiels :

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}; \quad i = 1, ..., 4 \quad j = 1, 2$$

il n'est plus régulier : on a $p = 5$ et $k = 4$. par exemple, la contrainte $\sum_{i=1,...,4} \alpha_i = 0$ rend le modèle identifiable.

Nous sommes donc maintenant en mesure d'établir les relations entre les systèmes de contraintes et l'orthogonalité

Définition 8.3 *Considérons la partition suivante d'un modèle linéaire*

$$Y = X_1\theta_1 + ... + X_m\theta_m + \epsilon.$$

et soit un système de contraintes $C_1\theta_1 = 0, ..., C_m\theta_m = 0$ qui rendent le modèle identifiable. On dit que ces contraintes **rendent la partition orthogonale** *si les espaces*

$$V_i = \{X_i\theta_i : \theta_i \in Ker(C_i)\} \quad , \quad i = i, m$$

sont orthogonaux.

Cette définition prend tout son sens avec l'exemple incontournable suivant qui concerne le modèle d'analyse de la variance à deux facteurs croisés.

Proposition 8.1 *Soit le modèle d'analyse de la variance à deux facteurs croisés,*

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{i,j,k}; \quad i = 1, I, \ j = 1, J, \ k = 1, n_{ij},$$

avec $n_{ij} \geq 1$, $\sum_{ij} n_{ij} = n > IJ$. Il existe des contraintes qui rendent la partition $\mu, \alpha, \beta, \gamma$ orthogonale si et seulement si

$$n_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{8.1}$$

où, par exemple, $n_{i+} = \sum_i n_{i,j}$. Et dans ce cas les contraintes sont

$$\sum_i n_{i+}\alpha_i = 0 \quad (i); \quad \sum_j n_{+j}\beta_j = 0 \quad (ii); \quad \forall i \sum_j n_{ij}\gamma_{i,j} = 0 (iii) \quad ; \quad \forall j \sum_i n_{ij}\gamma_{i,j} = 0 (iv)$$

Un énoncé tres proche est possible pour le modèle additif. Notez bien que pour le système de containtes qui correspond à la décomposition de type III que nous avons présenté en section ? ? ? qui est nous le rappelons

$$\sum_i \alpha_i = 0 \quad ; \quad \sum_j \beta_j = 0 \quad ; \quad \forall i \sum_j \gamma_{i,j} = 0 \quad ; \quad \forall j \sum_i \gamma_{i,j} = 0$$

il n'y a orthogonalité d'apres la proposition 8.1 que si le modèles est **équirépété** c'est à dire $n_{ij} = cte$.

Démonstration :

a) Clairement les conditions d'orthogonalité de μ et α sont équivalentes à (i), celles de μ et β sont équivalentes à (ii) , celles de de l'espace engendré par μ plus α avec l'espace engendré par γ est équivalente à (iii) et enfin l'orthogonalité de l'espace engendré par μ plus β avec l'espace engendré par γ est équivalente à (iv).

b) il reste donc à examiner l'orthogonalité de l'espace engendré par α muni des containtes ci-dessus (on le notera $[\alpha]$) avec l'espace engendré par β ($[\beta]$ avec le meme type de contraintes.

Définissons d'abord les espaces suivants

$$A := \{(\alpha_1, ..., \alpha_I) \in \mathbb{R}^I : \sum_i n_{i+}\alpha_i = 0\},$$

et symétriquement

$$B := \{(\beta_1, ..., \beta_J) \in \mathbb{R}^J : \sum_i n_{+j}\beta_j = 0\}.$$

Soient maintenant YA et YB deux éléments quelconques de $[\alpha]$ et $[\beta]$, on a

$$YA_{ijk} = \alpha_i \text{ avec } \alpha \in A$$

$$YB_{ijk} = \beta_j \text{ avec } \beta \in B$$

Donc si (8.1) est vraie,

$$< YA, YB >= \sum_{ij} n_{ij}\alpha_i\beta_j = \sum_{ij} \frac{n_{i+}\alpha_i n_{+j}\beta_j}{n_{++}} = 0.$$

c) Réciproquement, si $[\alpha]$ et $[\beta]$ sont orthogonaux, pour tout α dans A et β dans B on a

$$\sum_{ij} n_{ij}\alpha_i\beta_j = 0 \tag{8.2}$$

Fixons α, comme la relation (8.2) est vraie pour tout β et que la seule relation vérifiée par les β est $\sum_j n_{+j}\beta_j = 0$, cela implique que

$$(\sum_i n_{ij}\alpha_i) \text{ est proportionnel à } n_{+j} \text{ quand j varie}$$

En sommant en j on voit que le coefficient de proportionalité est forcément nul. On donc montré que pour tout vecteur α dans A

$$\text{Pour tout } j = i, J \ \sum_{ij} n_{ij}\alpha_i = 0$$

À nouveau cela implique que le vecteur n_{ij} est proportionnel a n_{i+}. Notant C_j le coefficient de proportionnalité :

$$n_{ij} = C_j n_{i+}$$

et sommant en j on obtient

$$C_j n_{++} = n_{+j}$$

et il suffit de re-injecter cette formule dans la précédente pour obtenir le résultat ■.

Exemple 8.5 (régression polynômiale orthogonalisée) *On considère le modèle de régression quadratique de taille n à une variable*

$$Y = \alpha + \beta_1 Z + \beta_2 Z^2 + \epsilon \tag{8.3}$$

où la variable Z prend des valeurs régulièrement espacées de 1 à n. La moyenne de Z vaut $\bar{Z} = (n+1)/2$; On note $<>$ le produit scalaire de $\mathbb{R}^n$ et on définit les variables suivantes :

$$T^0 := \mathbb{1} \quad ; \quad T^1 := Z - \bar{Z} \quad ; \quad T^1 := (Z-\bar{Z})^2 - 1/n < (Z-\bar{Z})^2, \mathbb{1} > .$$

Le modèle 8.3 est équivalent au modèle

$$Y = \gamma_0 T^0 + \gamma_1 T^1 + \gamma_2 T^2 + \epsilon \tag{8.4}$$

La matrice d'information $(X'X)$ de ce nouveau modèle vaut

$$\begin{pmatrix} \|T^0\|^2 & < T^0, T^1 > & < T^0, T^2 > \\ < T^1, T^0 > & \|T^1\|^2 & < T^1, T^2 > \\ < T^2, T^0 > & < T^2, T^1 > & \|T^2\|^2 \end{pmatrix}$$

Cette matrice apparait comme la matrice de produits scalaires des vecteurs T^0 ; T^1 ; T^2. Par parité les produits scalaires $< T^0, T^1 >$ et $< T^1, T^2 >$ sont nuls. Par ailleurs T^2 a été construit de sorte que $< T^2, \mathbb{1} = 0$.

Le modèle est donc orthogonal puisque ses régresseurs sont orthogonaux entre eux et que la matrice d'information $(X'X)$ est diagonale.

Comme conséquence immédiate on obtient une formule explicite du coefficient de régression de Y sur T^i : il est le même que si T^i était seule variable. On a donc dans le modèle (8.4)

$$\widehat{\gamma_i} = \frac{< T^i, Y >}{\|T^i\|^2}.$$

Chapitre 9

Exemples informatiques

Le modèle linéaire est traité dans SAS essentiellement par les procédures ANOVA (analyse de la variance orthogonale), REG (régression), GLM (modèles d'analyse de la variance non-orthogonaux, analyse de la covariance, modèles lineaires généraux) et MIXED (modèles mixtes). Plutot que d'écrire un résumé de la volumineuse documentation SAS nous avons choisi de donner quelques exemples incontournables qui reprennent les exemples des chapitres précédents.

Hormis deux exemples basiques, nous supposerons que le lecteur est familier avec la contruction de tables data SAS et nous ne détaillerons pas cet aspect.

1 Modèles élémentaires

Le modèle de régression linéaire simple du chapitre ? se traite de la façon suivante :

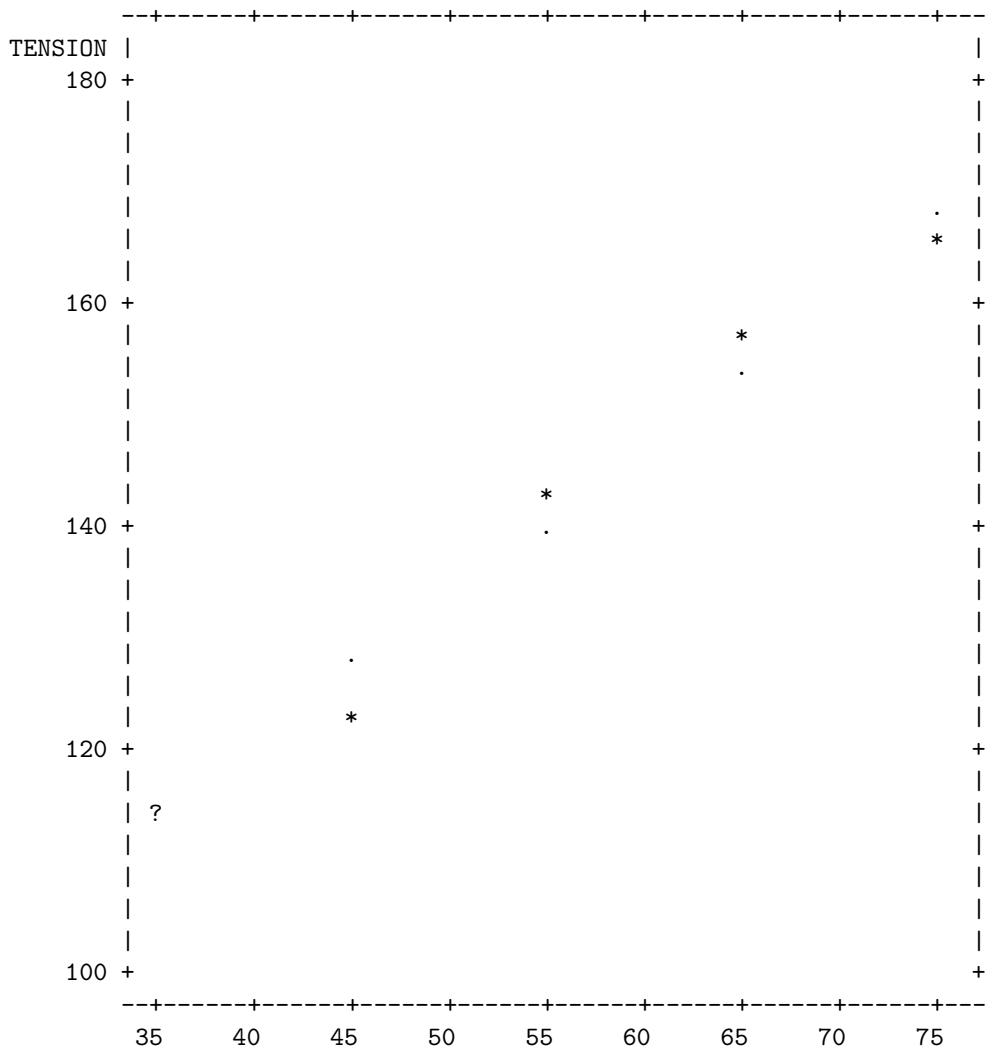
```
data sasuser.tension;
input  age tension;
cards;
35 114
45 124
55 143
65 158
75 166
;
run;

proc reg data=sasuser.tension;
model tension=age;
plot tension*age='*' p.*age/overlay symbol='.';
run;quit;
```

Interprétons la seconde partie du programme

- La première ligne déclare que la procédure va travailler sur la table "tension" contenue dans le répertoire "sasuser".
- La seconde ligne déclare la variable a expliquer (tension) et la variable explicative (age).
- La troisième ligne demande de manière très simple (mais malheureusement en basse résolution) un graphique qui superpose le nuage de points avec la droite de regression. Ce graphique n'est utilisable qu'en regression linéaire simple. Dans le cas général voir les exemples qui suivent.

Le résultat est (extrait)

```
                          The SAS System                                1
                                                 11:01 Friday, August 31, 2001


                        Analysis of Variance

                          Sum of          Mean
Source           DF      Squares        Square      F Value       Prob>F

Model             1   1904.40000   1904.40000      180.797       0.0009
Error             3     31.60000     10.53333
C Total           4   1936.00000


                        Parameter Estimates

                   Parameter      Standard    T for H0:
Variable  DF        Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1      65.100000    5.82837885        11.169        0.0015
AGE        1       1.380000    0.10263203        13.446
0.0009
```



Le graphique montre clairement que la dépendance entre les variables (indiquée par des *) est bien linéaire (le ? est dû à la superposition d'un point et d'une étoile). L'analyse de la variance montre que la pente est significativement non nulle : F=180,... ou T= 13.4... ce qui revient au même.

L'exemple de l'analyse de la variance à un facteur obéit à la même logique

```
 data ;
infile 'foret3';
input  hauteur foret;
run;

proc glm ;
class foret;
model hauteur= foret;
mean foret;
```

```
output out= sortie r= residu p=predite;
run;

proc gplot data = sortie;
plot residu*predite;
run;quit;
```

les petites différences sont

- l'étape data va chercher les données sur un fichier unix qui s'appelle "foret3". Pour avoir des graphiques de résidus plus réaliste nous n'avons pas utilisé les données du chapitre ? mais un jeu plus important extrait de Dacunha-Castelle & Duflo (1990).
- pour l'etape procédure, On a dû bien sûr utiliser un procédure adaptée a l'analyse de variance. par souci de cohérence avec ce qui suit nous avons préféré utiliser proc GLM plutôt que proc ANOVA.
- la ligne CLASS déclare que la variable (foret) (le numéro de la forêt) est qualitative.
- la ligne MEANS demande explicitement les moyennes qui ne sont pas données par défaut.
- la ligne ouput realise une sortie des résultats de l'analyse : les résidus et le valeur prédites sur un le data (sortie). Ces données sont reprise par GPLOT qui donne un graphique haute résolution.

On obtient (extrait)

```
                          The SAS System                              6
                                        11:01 Friday, August 31, 2001

                  General Linear Models Procedure

Dependent Variable: HAUTEUR

Source                   DF    Sum of Squares     F Value      Pr > F

Model                     2       50.44718376        7.13      0.0027

Error                    33      116.78253846

Corrected Total          35      167.22972222

                  R-Square             C.V.            HAUTEUR Mean

                  0.301664          7.523912              25.0027778


                  General Linear Models Procedure

           Level of          -----------HAUTEUR-----------
           FORET        N       Mean              SD

           1           13     25.9923077       1.39072016
           2           13     25.4538462       1.82647903
           3           10     23.1300000       2.43905719
```

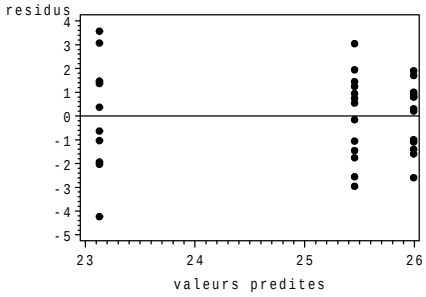


FIG. 9.1 – graphique des résidus

Le graphique des résidus ne met pas en évidence d'écart manifeste aux postulats. Le tableau d'analyse de la variance montre que les trois forêts ont des hauteurs significativement différentes (F = 7.13). Les hauteurs estimées sont 25.9..., 25.4... 23.1....

2 Régression multiple

Nous utilisons un jeu de donnée sur la chenille processionnaire du pin issu de Tomassone *et al* (1992). Il s'agit d'expliquer le nombre de nids de chenillles processionnaires (X11) ou son log (X12) en fonction de caractéristiques de la placette qui sont

- altitude en m : X1
- pente en degrés : X2
- nombre de pins dans la placette : X3
- hauteur de l'arbre échantillonné au centre de la placette : X4
- diamètre de cet arbre : X5
- note de densité de peuplement : X6
- orientation de la placette (1=sud 2=autre) : X7
- hauteur en m des arbres dominants : X8
- nombre de strates de végétation : X9
- mélange du peuplement (1=mélangé, 2=non mélangé) X10

On suppose les données rangées dans le data "sasuser.process" et on lance le programme suivant

```
proc reg data= sasuser.process;
model X12 = X1-X10/selection = rsquare best =1  cp  aic bic;
model X11 X12 = X1 X2 X4 X5/tol vif r;
plot r.*p.;
run;quit;
```

Le programme est volontairement chargé en options pour illustrer les possibilités de SAS. La première ligne n'appelle pas de commentaire, dans la seconde l'option (que l'on peut considérer comme monobloc " selection = rsquare best =1" revient à faire tourner l'algorithme de Furnival et Wilson. Cette option permet le tri automatique de régresseur est la meilleure qui soit et SAS supporte sans problèmes jusqu'à 15 régresseurs ce qui couvre la plupart des applications. Elle est clairement préférable aux options : "Backward" "Forward" "Stepwise" et "Maxsquare" qui ont le même propos. Il reste à choisir la taille optimale du modèle. Pour ce faire nous avons demandé l'impression du cp de Mallows (CP) ainsi que les critères d'Akaike (AIC) et de Schwarz (BIC). La troisième ligne utilise le modèle retenu par la ligne précédente avec le critère du CP. On y a illustré diverses possibilites : celle de mettre plusiieurs variables à expliquer et également la

demande des diagnostics de multicolinéarité et la demande des résidus. le graphique qui suit est en basse résolution c'est le graphique classique résidus (r.) contre valeur prédite(p.). La sortie est la suivante :

```
N = 32     Regression Models for Dependent Variable: X12


        R-square      C(p)       AIC       BIC Variables in Model
   In

     1  0.3722390   28.4409    2.9658    2.8586 X9
-------------------------------------------------
     2  0.5030220   18.6824   -2.5099   -2.3740 X1 X9
-----------------------------------------------------
     3  0.5917772   12.7026   -6.8053   -5.8632 X1 X2 X9
---------------------------------------------------------
     4  0.7118930    3.9032  -15.9567  -11.7138 X1 X2 X4 X5
-------------------------------------------------------------
     5  0.7247899    4.7437  -15.4222  -10.0751 X1 X2 X4 X5 X10
-----------------------------------------------------------------
     6  0.7434261    5.0681  -15.6660   -8.5451 X1 X2 X3 X4 X5 X9
---------------------------------------------------------------------
     7  0.7574359    5.8085  -15.4628   -6.4225 X1 X2 X3 X4 X5 X9 X10
-------------------------------------------------------------------------
     8  0.7605669    7.5270  -13.8786   -3.5949 X1 X2 X3 X4 X5 X8 X9 X10
-----------------------------------------------------------------------------
     9  0.7660609    9.0331  -12.6214   -0.7368 X1 X2 X3 X4 X5 X6 X8 X9 X10
---------------------------------------------------------------------------------
    10  0.7664287   11.0000  -10.6717    2.3033 X1 X2 X3 X4 X5 X6 X7 X8 X9
                                                X10
-------------------------------------------------------------------------------
```

Cette partie est la réponse à la premiere ligne et illustre le choix de modèles. L'option selection = rsquare best = 1 donne le meilleur modèle à 1 régresseur, puis celui a 2 régresseur etc. Dans notre exemple, on peut remarquer que les ensembles de variables choisies ne sont pas hiérarchiques : par exemple, on ne passe pas du modèle à 3 régresseurs à celui à 4 en rajoutant une variable.

Le vrai problème, ensuite est de choisir la taille du modèle. Dans notre exemple les critères de CP AIC ou BIC minimum sont remarquablement cohérents et choisissent un modèle à 4 régresseurs.

Nous ne donnons ci-dessous la sortie que pour la variable X11, celle pour X12 a exactement le même format

```
                          The SAS System                              11
                                                 15:11 Friday, August 31, 2001
Model: MODEL2
Dependent Variable: X11        nb  nids de procession. par arbre

                          Analysis of Variance

                          Sum of          Mean
   Source          DF      Squares       Square      F Value       Prob>F
```

Model	4	13.10487	3.27622	11.508	0.0001
Error	27	7.68670	0.28469		
C Total	31	20.79157			

Root MSE	0.53357	R-square	0.6303
Dep Mean	0.81406	Adj R-sq	0.5755
C.V.	65.54360		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	6.603087	1.02423313	6.447	0.0001
X1	1	-0.002813	0.00078216	-3.596	0.0013
X2	1	-0.045647	0.01345751	-3.392	0.0022
X4	1	-0.755095	0.21590679	-3.497	0.0016
X5	1	0.168475	0.05153843	3.269	0.0029

Variable	DF	Tolerance	Variance Inflation
INTERCEP	1	.	0.00000000
X1	1	0.89499376	1.11732623
X2	1	0.97975049	1.02066803
X4	1	0.17630697	5.67192545
X5	1	0.18093121	5.52696248

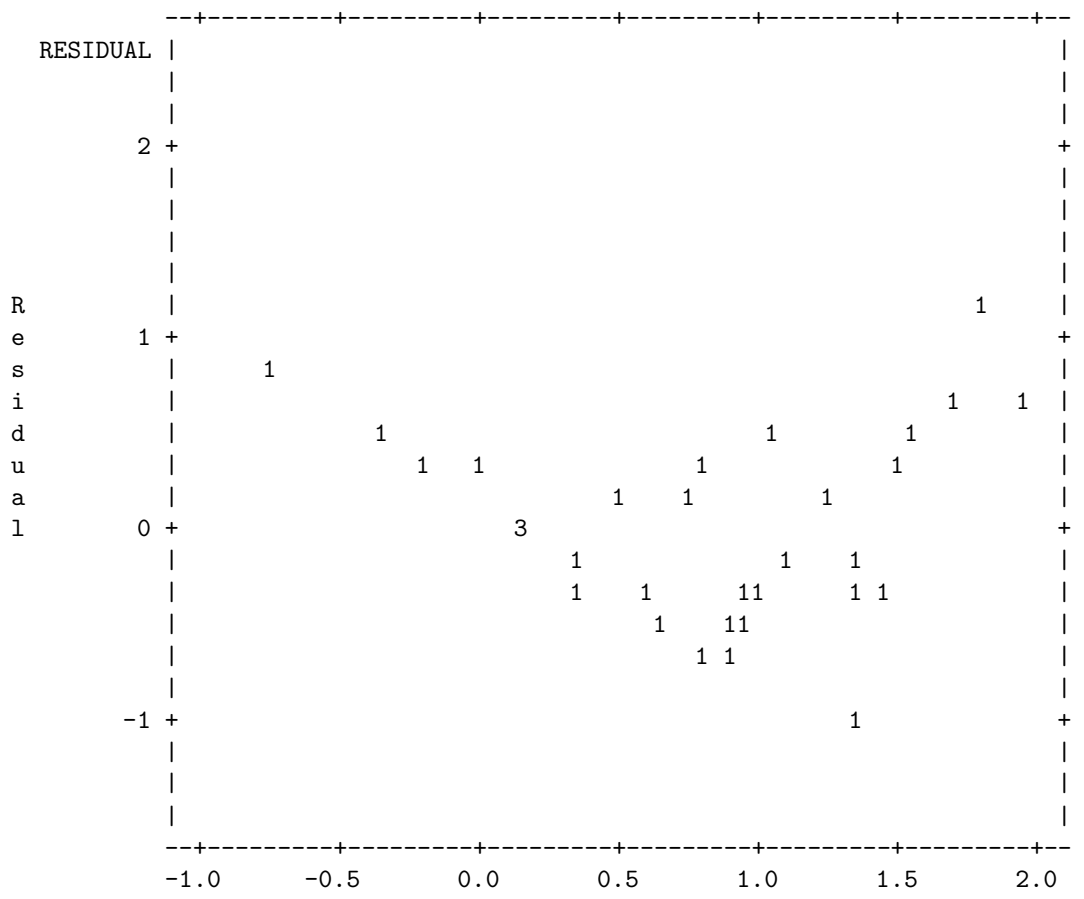
Variable	DF	Variable Label
INTERCEP	1	Intercept
X1	1	altitude
X2	1	pente
X4	1	hauteur
X5	1	diametre

Obs	Dep Var X11	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual
1	2.3700	1.6964	0.162	0.6736	0.508	1.325
2	1.4700	1.2602	0.180	0.2098	0.502	0.418
3	1.1300	1.3642	0.209	-0.2342	0.491	-0.477
4	0.8500	1.0823	0.165	-0.2323	0.507	-0.458
5	0.2400	0.3341	0.166	-0.0941	0.507	-0.186
6	1.4900	1.0255	0.108	0.4645	0.522	0.889
7	0.3000	0.0136	0.257	0.2864	0.468	0.612

```
         8    0.0700   -0.1807     0.268    0.2507     0.462     0.543
         9    3.0000    1.8174
**************************extrait de la sortie ********************************

 Obs    -2-1-0 1 2            D

         1  |      |**    |     0.036
         2  |      |      |     0.004
         3  |      |      |     0.008
         4  |      |      |     0.004
         5  |      |      |     0.001
         6  |      |*     |     0.007
         7  |      |*     |     0.023
         8  |      |*     |     0.020
         9  |      |****  |     0.150
        10  |      |*     |     0.048
        11  |    **|      |     0.063
        12  |     *|      |     0.027
        13  |      |**    |     0.094
        14  |      |**    |     0.030
```

**********extrait de la sortie **********

15:11 Friday, August 31, 2001



Remarquons d'abord que le graphique de résidus pour X11 est pathologique. Que l'on se rassure celui pour X12 ne l'est pas ! C'est pour cela que l'on a travaillé sur les log dans la première analyse.

Les valeurs des estimateurs montrent un effet positif (c.à d. une diminution) de l'altitude de la pente et de la hauteur des arbres. Cela peut éventuellement s'interpréter comme une difficulté d'accès à certaines placettes qui serait protectrice. On constate que toutes les variables retenues par la procédure précédente sont significatives. Les indicateurs de colinéarité (on vérifie que TOL est l'inverse de VIF) sont relativement raisonnables.

Les deux derniers tableaux sont dûs à l'analyse de résidus demandé par l'option "/r". Le premier ne pose pas de problemes d'interpretation. Le segond est conçu pour la recherche de corrélations entre résidus consécutifs. Il comprend une représentation des résidus par un diagrammene en bâtons ainsi que la valeur du D de Cook : mesure d'influence de la mesure sur le paramètre estimé.

3 Analyse de la variance a deux facteurs

Les données que nous allons utiliser sont extraites de Calas Rochd Druilhet et Azaïs (1998). Dans cette expérience on compare l'action de deux traitements (facteur trait) désinfectants sur des échantillons de dents contaminées au préalable par deux sources de germes (facteur germe). La réponse est le nombre moyen de germes restant. Elle est mesuré par microscopie électronique. Pour des raisons d'homogénéite de la variance, on travaillera plutôt sur le log de ce nombre (LNBAC). D'autres facteurs devraient en fait être incorporés dans le modèle d'analyse : l'âge de la dent, la vache dont est issu la dent etc..., mais par souci de simplicité nous les omettrons ici.

En supposant les données présentes dans la table "sasuer.dents" l'analyse se fait par

```
proc glm data=sasuser.dents;
class trait germe;
model log=trait germe trait*germe;
output out=sortie predicted=p student=r;
lsmeans trait germe trait*germe/ out=graph;
run; quit;
```

La seconde ligne déclare trait et germe en qualitatif. la troisième déclare le modèle standard d'analyse de la variance à deux facteurs avec interaction. La quatième ligne écrit, sur une table SAS, les résidus et les valeurs prédites en vue de préparer un graphique haute résolution. La cinquième ligne demande les moyennes ajustées : lsmeans et les écrit dans la table "graph".

le commandes suivantes sont typiquement UNIX Elles permettent de construire des graphiques haute résolutiuon et de les écrire en postcript encapsule dans les fichiers 'grafi.eps'. Nous laissons au lecteur à titre d'exercice le soin de découvrir leur logique.

```
/*   -------------------------------   */
 goptions device=psepsf gend='0a'x gaccess=gsasfile;
 filename gsasfile 'graf1.eps';
 proc gplot data=sortie;
 axis1 label=('valeurs predites') length=10cm;
 axis2 label=('residus') length=10cm;
 symbol v=dot;
 plot r*p / haxis=axis1 vaxis=axis2 vref=0;              ;
 run;goptions reset=all;quit;

 /*   ---------------------   */
 goptions device=psepsf gend='0a'x gaccess=gsasfile;
 filename gsasfile 'graf2.eps';
 proc gplot data=graph;
 axis1 label=('germe') order=(1 to 2 by 1) minor=none length=10cm;
 axis2 label=('moyenne' justify=right 'des effets') length=10cm;
 symbol1 i=join v=dot cv=black;
 symbol2 i=join v=triangle cv=black;
 symbol3 i=join v=circle cv=black;

 plot lsmean*germe =trait/ haxis=axis1 vaxis=axis2;
 run;goptions reset=all; quit;
```

```
/*   --------------------   */
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'graf3.eps';
proc gplot data=graph;
axis1 label=('trait') order=(1 to 2 by 1) minor=none length=10cm;
axis2 label=('moyenne' justify=right 'des effets') length=10cm;
symbol1 i=join v=dot cv=black;
symbol2 i=join v=triangle cv=black;
symbol3 i=join v=circle cv=black;

plot lsmean*trait=germe / haxis=axis1 vaxis=axis2;
run;goptions reset=all; quit;
```

Ces programmes donnent les résultats suivants

```
                                    10:00 Monday, September 3, 2001  17

                        General Linear Models Procedure
                            Class Level Information

                           Class    Levels    Values

                           TRAIT         2    1 2

                           GERME         2    1 2

                      Number of observations in data set = 64

                        General Linear Models Procedure

Dependent Variable: LNBAC
                                   Sum of          Mean
Source                     DF       Squares        Square   F Value     Pr > F

Model                       3    26.8258464     8.9419488     22.43     0.0001

Error                      60    23.9183125     0.3986385

Corrected Total            63    50.7441589

                   R-Square          C.V.      Root MSE           LNBAC Mean

                   0.528649       98.91739       0.63138             0.63829

Source                     DF     Type I SS   Mean Square   F Value     Pr > F

TRAIT                       1     0.1037806     0.1037806      0.26     0.6118
GERME                       1    16.7022612    16.7022612     41.90     0.0001
TRAIT*GERME                 1    10.0198046    10.0198046     25.14     0.0001
```

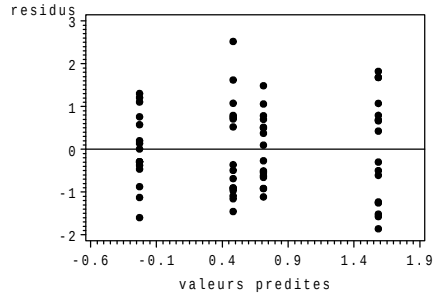


FIG. 9.2 – graphique des résidus

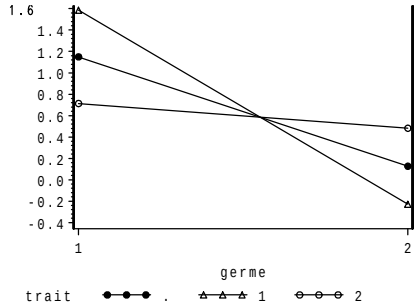


FIG. 9.3 – graphique des interactions

Least Squares Means

TRAIT	LNBAC LSMEAN
1	0.67855719
2	0.59801969

GERME	LNBAC LSMEAN
1	1.14914344
2	0.12743344

TRAIT	GERME	LNBAC LSMEAN
1	1	1.58508813
1	2	-0.22797375
2	1	0.71319875
2	2	0.48284063

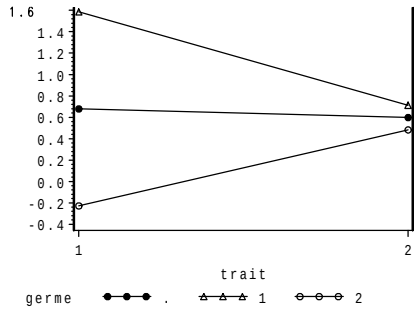


FIG. 9.4 – graphique des interactions dans l'autre sens

Ce jeu de données est parfaitement équilibré : pour un couple traitement-germe, il y a exactement 16 observations. Pour cette raison nous ne présentons pas le tableau d'analyse de la variance de type III car il est identique à celui de type I. On voit dans le tableau de type I que l'interaction est significative et que par conséquent les deux facteurs doivent être conservés, par contre l'effet traitement est non significatif ce qui est intriguant. On comprend mieux ce qui se passe en regardant les moyennes ajustées ou les graphiques correspondants : un traitement est efficace sur un germe et le second sur l'autre.

Ici nous avons donné un exemple de données équirepétées avec des facteurs à deux niveaux. Quand les données ne sont pas équirépétées, le tableau d'analyse de la variance n'est pas unique, il est alors conseillé d'utiliser le tableau de type III qui correspond à la décomposition du chapitre ?. Le recours à la directive "lsmeans" (et non pas "means") est nécessaire. Dans le cas où les facteurs ont plus de deux niveaux, et dans le cas où les effets sont significatifs, une comparaison de moyennes est nécessaire. Elle se fait, dans le cas équilibré, par les options

```
means trait germe trait*germe  /tuckey;
```

Dans le cas déséquilibré, le mieux est le plus souvent de faire un méthode de Bonferroni à la main après avoir demandé les comparaisons deux à deux par

```
lsmeans trait germe trait*germe  /tdiff;
```

Exercice 9.1 *Nous allons illustrer l'abominable complexité de l'option* `/solution` *en analyse de la variance à deux facteurs. Voici un exemple volontairement simple et dont les données on été inventées. L'utilisation de* `/solution` *donne la valeur -12 à la fin du tableau pour a*b 1 1. Comment s'interprète t'elle ?*

```
data;
input a b y;
cards;
    1    1     5
    1    1     3
    1    2    25
    1    2    27
    1    2    32
    2    1    12
    2    2    21
```

```
    2    2    27
;
proc glm ;
class a b;
model y= a b a*b/solution;
lsmeans a b a*b;
run; quit;
Dependent Variable: Y
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	792.000000	264.000000	22.96	0.0055
Error	4	46.000000	11.500000		
Corrected Total	7	838.000000			

R-Square	C.V.	Root MSE	Y Mean
0.945107	17.84824	3.39116	19.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	6.857143	6.857143	0.60	0.4831
B	1	555.428571	555.428571	48.30	0.0023
A*B	1	61.714286	61.714286	5.37	0.0814

10:00 Monday, September 3, 2001 21

General Linear Models Procedure

Dependent Variable: Y

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		24.00000000 B	10.01	0.0006	2.39791576
A	1	4.00000000 B	1.29	0.2659	3.09569594
	2	0.00000000 B	.	.	.
B	1	-12.00000000 B	-2.89	0.0446	4.15331193
	2	0.00000000 B	.	.	.
A*B	1 1	-12.00000000 B	-2.32	0.0814	5.18009009
	1 2	0.00000000 B	.	.	.
	2 1	0.00000000 B	.	.	.
	2 2	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized
inverse was used to solve the normal equations. Estimates followed

```
by the letter 'B' are biased, and are not unique estimators of the
parameters.
                      A   B            Y
                                   LSMEAN

                      1   1     4.0000000
                      1   2    28.0000000
                      2   1    12.0000000
                      2   2    24.0000000
```

4 Analyse de la covariance

Les données que nous considérons sont extraites de Tanner ,J. M. (1964), The Physique of the Olympic Athlete. George Allen and Unwin, London. On a mesuré année par année la taille de fillettes de 6 à 10 ans. De plus ces fillettes sont regroupées en trois classes, suivant la taille de leur mère (facteur tmere :p petite, m moyenne, g grande). On indique par ailleurs dans le fichier -le numéro de l'individu dans le groupe de taille de mère :ind -ensuite les 5 tailles de l'individu de 6 à 10 ans.

Extrait des données :

```
tmere      ind    taille de 6 a 10 ans

p          1      111   116.4  121.7  126.3  130.5
p          2      110   115.8  121.5  126.6  131.4
p          3      113.7 119.7  125.3  130.1  136.0
p          4      114.0 118.9  124.6  129.1  134.0
p          5      114.5 122.0  126.4  131.2  135.0
p          6      112.0 117.3  124.4  129.2  135.2
m          1      116   122    126.6  132.6  137.6
m          2      117.6 123.2  129.3  134.5  138.9
```

La variable à expliquer : la taille, est fonction de variables qualitatives (tmere, ind) et d'une variable quantitative l'age. C'est une situation d'analyse de la covariance. On remarque de plus que le facteur individu est hiérachisé au facteur tmere.

Le fichier de donnée n'est pas sous forme standard, voici la manière de le lire.

```
data sasuser.agemere;
infile '~/dess/agemere';
 input tmere$ ind@@;
    do i=6 to 10;
    input taille @@;
    age =i;
    output;
    end;
proc print;run;
```

Pour prendre en compte la variable quantitative (age) et le facteur (tmere) on pose d'abord un modèle standard d'hétérogénité des pentes.

```
proc glm data=sasuser.agemere;
  class tmere ind ;
model taille = tmere  age age*tmere /solution;
output out =a r=re p=es;
proc gplot data=a;
plot re*es;
run;
```

En fait ce modèle n'est pas le bon : indépendamment de la taille de la mère, il est clair qu'il y a des individus plus ou moins grands ou qui grandissent plus vite. On vérifie d'ailleurs cela sur le graphe de résidus que nous ne donnons pas.

Il est nécessaire de faire apparaître le facteur (ind) qui est hiérachisé au facteur (tmere).

```
proc glm data=sasuser.agemere;
class tmere ind ;
model taille = tmere ind(tmere) age age*tmere age*ind(tmere)/solution;
lsmeans tmere  ind(tmere);
output out =a r=re p=es;
proc gplot data=a;
plot re*es;
run;
```

l'option `/solution`, dont l'utilité est en général trés limitée, est nécessaire ici pour obtenir le coefficient de (age).

```
                          The SAS System                                 11
                                        09:37 Thursday, November 27, 1997

                      General Linear Models Procedure

Dependent Variable: TAILLE
                                  Sum of          Mean
Source                     DF        Squares        Square  F Value    Pr > F

Model                      39     8909.01740     228.43634   480.09    0.0001

Error                      60       28.54900       0.47582

Corrected Total            99     8937.56640

                   R-Square           C.V.      Root MSE        TAILLE Mean

                   0.996806       0.537826       0.68979            128.256


Source                     DF      Type I SS   Mean Square  F Value    Pr > F

TMERE                       2     1534.58354     767.29177  1612.58    0.0001
IND(TMERE)                 17      774.71086      45.57123    95.77    0.0001
AGE                         1     6535.67445    6535.67445 13735.70    0.0001
AGE*TMERE                   2       33.34026      16.67013    35.03    0.0001
AGE*IND(TMERE)             17       30.70829       1.80637     3.80    0.0001
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TMERE	2	1.95739	0.97869	2.06	0.1368
IND(TMERE)	17	27.67190	1.62776	3.42	0.0002
AGE	1	6468.29463	6468.29463	13594.09	0.0001
AGE*TMERE	2	33.34026	16.67013	35.03	0.0001
AGE*IND(TMERE)	17	30.70829	1.80637	3.80	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		76.98000000 B	43.44	0.0001	1.77211456
AGE		5.83000000 B	26.73	0.0001	0.21813222
AGE*TMERE	g	1.11000000 B	3.60	0.0006	0.30848555
	m	-0.12000000 B	-0.39	0.6987	0.30848555
	p	0.00000000 B	.	.	.
AGE*IND(TMERE)	1 g	-1.72000000 B	-5.58	0.0001	0.30848555
	2 g	-0.43000000 B	-1.39	0.1685	0.30848555
	3 g	-0.57000000 B	-1.85	0.0696	0.30848555
	4 g	-0.48000000 B	-1.56	0.1250	0.30848555
	5 g	-1.12000000 B	-3.63	0.0006	0.30848555
	6 g	-0.52000000 B	-1.69	0.0971	0.30848555
	7 g	0.00000000 B	.	.	.
	1 m	-0.33000000 B	-1.07	0.2890	0.30848555
	2 m	-0.32000000 B	-1.04	0.3037	0.30848555
	3 m	0.43000000 B	1.39	0.1685	0.30848555
	4 m	-0.49000000 B	-1.59	0.1175	0.30848555
	5 m	-0.06000000 B	-0.19	0.8464	0.30848555
	6 m	-0.23000000 B	-0.75	0.4588	0.30848555
	7 m	0.00000000 B	.	.	.
	1 p	-0.94000000 B	-3.05	0.0034	0.30848555
	2 p	-0.47000000 B	-1.52	0.1329	0.30848555
	3 p	-0.33000000 B	-1.07	0.2890	0.30848555
	4 p	-0.81000000 B	-2.63	0.0110	0.30848555
	5 p	-0.81000000 B	-2.63	0.0110	0.30848555
	6 p	0.00000000 B	.	.	.

Least Squares Means

TMERE	TAILLE LSMEAN
g	133.111429
m	127.511429
p	123.460000

IND	TMERE	TAILLE LSMEAN
1	g	130.940000
2	g	134.160000
3	g	131.940000
4	g	133.580000
5	g	133.540000
6	g	128.160000
7	g	139.460000
1	m	126.960000
2	m	128.700000
3	m	133.620000
4	m	124.520000
5	m	128.920000
6	m	124.940000
7	m	124.920000
1	p	121.180000
2	p	121.060000
3	p	124.960000
4	p	124.120000
5	p	125.820000
6	p	123.620000

le graphe de résidus, omis pour des raisons de concision, a un aspect satisfaisant Le résidu maximum est de 2cm ce qui est peu en pratique. Tous les effets sont significatifs sauf tmere en type I. Ce dernier point mérite une explication soignée. En effet le dispositif est équilibré, chaque individu est observé exactement une fois et une seule chaque année. Les composantes du modèles seraient orthogonales si on avait pris soin d'orthogonaliser la covariable (age) par rapport aux constantes. Cela peut se faire en centrant l'age : on introduit la variable (agec) : age-8 qui est centrée. Le modèle est alors orthogonal.

En fait la decomposition de type I va faire ceci automatiquement puisque age est othogonalisé par rapport à l'intercept ou tmere.

En consequence le test de type III de (tmere) est fait avec (age) et teste l'hypothese : "est-ce-que les 3 droites de régression moyennes des trois groupes de fillettes concourent à l'origine". L 'origine correspond a une taille théorique à la naissance qui a peu de sens car la croissance d'un être humain dans les toutres premières années n'est pas linéaire. Cette hypothèses est peu interprétable.

Ici le test de type I est préférable car il est fait implicitement avec 5agec) et teste l'hypothese : "est-ce-que les 3 droites de régression moyennes des trois groupes de fillettes concourent à 8 ans". Autrement dit "Y a t'il des groupes plus grands que d'autres". il s'agit là clairement de tester l'existence d'un partie héréditaire dans la taille.

Exercice 9.2 *Vérifiez les affirmation ci dessus en introduisant la variable (agec).*

- A l'aide de votre calcullette ou de matlab (ce qui est malheureusement un comble) calculez la croissance moyenne par an des différentes fillettes à partir de la sortie de /solution . Réflechissez au contraintes à utiliser en analyse de la variance à un facteur, voir exercice ? ?

5 Modèles mixtes

Nous reprenons l'exemple précédent mais, plus raisonnablemen,t nous supposons que les individus ont été échantillonnés parmi les trois populations : filletes dont la mère est petite, fillettes dont la mère est moyenne, fillettes dont la mère est grande. Nous voulons donc faire de l'inférence au nivceau de population et non plus des individus. On réalise l'analyse suivante

```
 data;
 set sasuser.agemere;
 agec= age-8;
  run;

  proc mixed ;
   class tmere ind ;
 model taille = tmere agec agec*tmere /solution;
 random ind(tmere) agec*ind(tmere);
 lsmeans tmere/diff;
 run;quit;
Ici la construction de (agec) est indispensable car la décmposition de
   type I n'existe pas dans proc MIXED. Comme dans proc GLM on déclare
   d'abord les variables qualitatives. La seule différence est que les
   effets aléatoires sont declarés dans un ligne à part : la ligne
   random. Ceci correspond a la forme la plus simple d'utilisation de
   proc MIXED. Il existe des formes beaucoup plus sophistiquées
   adaptées au situations de mesures répétées que nous ne détaillerons
   pas. La ligne LSMEANS demande les moyennes ajustées pour le facteur
   (tmere) avec comparaisons.

                              The SAS System                              11
                                          09:36 Friday, January 30, 1998


                        The MIXED Procedure

                      Class Level Information

                   Class     Levels  Values

                   TMERE          3  g m p
```

IND 7 1 2 3 4 5 6 7

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	322.29586816	
1	1	147.51999619	0.00000000

Convergence criteria met.

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
IND(TMERE)	18.95495193	9.01908204	3.12620865	2.88	0.0039
AGEC*IND(TMERE)	0.27963566	0.13305531	0.06256403	2.13	0.0334
Residual	1.00000000	0.47581667	0.08687184	5.48	0.0001

Model Fitting Information for TAILLE

Description	Value
Observations	100.0000
Variance Estimate	0.4758
Standard Deviation Estimate	0.6898
REML Log Likelihood	-160.140
Akaike's Information Criterion	-163.140
Schwarz's Bayesian Criterion	-166.955
-2 REML Log Likelihood	320.2804

Solution for Fixed Effects

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
INTERCEPT	123.46000000	1.23249377	17	100.17	0.0001
TMERE g	9.65142857	1.67960589	17	5.75	0.0001
TMERE m	4.05142857	1.67960589	17	2.41	0.0274
TMERE p	0.00000000	.	.	.	.
AGEC	5.27000000	0.17351127	17	30.37	0.0001
AGEC*TMERE g	0.97857143	0.23645601	17	4.14	0.0007
AGEC*TMERE m	0.29714286	0.23645601	17	1.26	0.2259
AGEC*TMERE p	0.00000000	.	.	.	.

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
TMERE	2	17	16.84	0.0001
AGEC	1	17	3572.36	0.0001
AGEC*TMERE	2	17	9.23	0.0019

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
TMERE g	133.11142857	1.14106751	17	116.66	0.0001
TMERE m	127.51142857	1.14106751	17	111.75	0.0001
TMERE p	123.46000000	1.23249377	17	100.17	0.0001

Differences of Least Squares Means

Level 1	Level 2	Difference	Std Error	DDF	T	Pr > |T|
TMERE g	TMERE m	5.60000000	1.61371314	17	3.47	0.0029
TMERE g	TMERE p	9.65142857	1.67960589	17	5.75	0.0001
TMERE m	TMERE p	4.05142857	1.67960589	17	2.41	0.0274

On voit dans l'ordre que - la convergence a été obtenu au bout de la première itération

- l'effet aléatoire (ind) est tres important. Le facteur d'hétérogénéité des pentes (agec*ind(tmere)) parait moins important mais il ne faut pas oublier qu'il est multiplié par (agec) qui peut prendre la valeur 2. Le test de Wald de la nullité (souvent très conservatif pour des petits échantillons) montre que les effets aléatoires sont significatifs.

-La solution des effets fixes est moins horrible que dans le cas d'effet fixes. Un petit calcul est quand meme necessair pour déduire que la croissance moyenne est de $5.27 + 1/3(0.98 + .3) = 5.70...$ cm par an.

- tous les effetes sont significatifs. On peut en déduire par ordre d'évidence que 1) les fillettes grandissent 2) la taille a une composante héréditaire : plus la mère est grande, plus la fille est grande et plus elle grandit rapidement.

- le tableau des LSMEANS permet de voir que chacun des groupes est significativement différent.

Chapitre 10

Bibliographie.

Arnold(1981). *The theory of linear models and multivariate analysis.* Willey, N.Y.

Azais J.M. (1994). Analyse de variance non orthogonale. L'exemple de SAS/GLM. *Rev. Stat. Appli.* XLII, 27-41.

Bonnet et Lansiaux (1992). Mémoire de maitrise, Université Paul Sabatier, Toulouse.

Calas P.(1994). *Adhérence de streptococcus sanguis et de Prevotella intermédia à la dentine radiculaire de bovin. Influence des traitements de surface.* Thèse de l'Université Paul Sabatier. Toulouse.

Calas P., Rochd T., Druilhet P. & Azaïs (1998). In vitro adhesion of two strains of prevotella nigrescens to the dentin of the root canal : the part played by different irrigations solutions.*Journal of Endodontics*, 24,2, 112, 115.

Cochran W.G., Cox G.M. (1950). *Experimental Designs.* Wiley, N.Y.

Coursol J. (1980). *Technique Statistique des modèles linéaires.* C.I.M.P.A. Nice

Dacunha-Castelle D. & Duflo M. (1990). *Probabilités et statistique, 1. Problèmes à temps fixe.* Masson 2ème édition.

Drapper N. et Smith H. (1966). *Applied Regression Analysis.* Wiley, NY.

Furnival G. et Wilson R. (1974). Regression by leaps and bounds. *Technometrics 16,4,499-511.*

Levene H. (1960). Robust tests for equality of variances. *In Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling, Olkin and al. ed.* Standford University Press, California, 278-292.

Miller R.G. (1966). *Simultaneous Statistical Inference.* Mc Graw-Hill, N.Y.

Searle S.R. (1987). *Linear models for unbalanced data.* Wiley, N.Y.

Tanner ,J. M. (1964), The Physique of the Olympic Athlete. George Allen and Unwin, London.

Tomassone, R., Audrain, S., Lesquoy, E. & Millier C. (1992). *La régression nouveaux regards sur une ancienne méthode statistique.*. Masson, Paris.

Annexe A

Rappels de Probabilités

Nous rappelons dans cette partie les propriétés minimales utilisées dans les chapitres précédents.

1 Règles opératoires du calcul de variance

L'espérance donne la postion moyenne d'un variable aléatoire. C'est un opérateur linéaire. Si α et β sont des scalaires,

$$\mathrm{E}(\alpha X + \beta Y) = \alpha \mathrm{E}(X) + \beta \mathrm{E}(Y),$$

quelles que soient les relations entre les variables aléatoires X et Y.

La variance mesure l'écart quadratique à l'espérance. Elle est quadratique et invariante par addition d'une constante :

$$\mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X).$$

La variance de la somme de variables aléatoires fait intervenir la covariance :

$$\mathrm{Var}(\alpha X + \beta Y) = \alpha^2 \mathrm{Var}(X) + \beta^2 \mathrm{Var}(Y) + 2\alpha\beta \mathrm{Cov}(X, Y).$$

Si les variables sont non corrélées, (en particulier si elles sont indépendantes) on obtient :

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Si $X_1, ..., X_n$ sont n copies indépendantes d'une même variable aléatoire X, on a

$$\mathrm{Var}(X_1 + ... + X_n) = n\mathrm{Var}(X).$$

$$\mathrm{Var}(\bar{X}) := \frac{1}{n}(X_1 + ... + X_n) = \frac{\mathrm{Var}(X)}{n}.$$

2 Lois de probabilités

Loi normale ou gaussienne centrée réduite : $N(0, 1)$.
C'est la loi sur $\mathbb{R}$ de densité

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Son espérance est nulle, sa variance vaut 1.

Loi normale ou gaussienne de moyenne m et de variance σ^2 :
Si X suit la loi $N(0,1)$, $Y = m + \sigma X$ suit par définition la loi $N(m, \sigma^2)$, loi normale d'espérance m et de variance σ^2. Sa densité est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

De la loi normale, on déduit un certain nombre de lois.

- **Loi du χ^2 à d degrés de libertés** : $\chi^2(d)$. Soient $X_1, ..., X_d$, d variables aléatoires indépendantes de loi $N(0,1)$, alors
$$S = X_1^2 + ... + X_d^2$$
suit une loi du χ^2 à d degrés de libertés. Cette loi est d'espérance d et de variance $2d$. C'est la loi Gamma de paramètres $(n/2, 2)$. Si S suit une loi $\chi^2(d)$, par définition $Y = \sigma^2 S$ suit une loi $\sigma^2\chi^2(d)$.
- **Loi de Student à d degrés de libertés** : $T(d)$. C'est la loi du quotient indépendant
$$T = \frac{N}{\sqrt{S/d}}$$
où N suit une loi $N(0,1)$ et S suit une loi $\chi^2(d)$, indépendante.
Interprétation : S est proche de son espérance qui vaut d et ce d'autant plus que d est grand par la loi des grands nombres. Le dénominateur est donc proche de 1. Il s'ensuit que la loi $T(d)$ est d'autant plus proche d'une loi normale que d est grand.
- **Loi de Fisher à n_1 et n_2 degrés de liberté** :F_{n_1,n_2}. Soient S_1 et S_2 deux variables aléatoires indépendantes de loi $\chi^2(n_1)$ et $\chi^2(n_2)$, alors par définition :
$$F = \frac{S_1/n_1}{S_2/n_2}$$
suit une loi F_{n_1,n_2}.
Interprétation : Par les mêmes considérations que précédemment, la loi F est d'autant plus proche de 1 que le nombre de degrés de liberté est élevé.

Si T suit une loi $T(d)$, alors T^2 suit une loi $F(1,d)$. Si F suit une loi $F(n_1, n_2)$, alors la loi de n_1F/n_2 est une loi beta de seconde espèce de paramètres $(n_1/2, n_2/2)$.

3 Vecteurs aléatoires :

Les vecteurs et matrices seront le plus souvent noté par des majuscules latines $X, Y, ...$ ou des minuscules grecques ϵ, θ. Quand il y aura risque d'ambiguïté nous identifierons les vecteurs par une flèche et les matrices par une écriture en **gras**. Par exemple :

"les quantités $\alpha, \vec{Z}, \mathbf{X}$"

implique que α est un scalaire, Z un vecteur, X une matrice.

Si $\mathbf{M}$ est une matrice, nous noterons $[\mathbf{M}]$ l'espace vectoriel engendré par les colonnes de $\mathbf{M}$.
Si E est un espace vectoriel nous noterons P_E le projecteur orthogonal sur E. Cette notation désigne aussi bien le projecteur que la matrice associée. $\mathbf{X}'$ est la matrice transposée de la matrice $\mathbf{X}$. Enfin $\mathbf{I}d$ est la matrice carrée identité : et $\mathbb{1}$ le vecteur dont toutes les coordonnées sont égales à 1.

Si $\vec{Z}$ est un vecteur aléatoire de taille n, on définit $\mathrm{E}(\vec{Z})$, le vecteur dont les coordonnées sont les espérances des coordonnées de $\vec{Z}$, par exemple

$$\mathrm{E}\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} \mathrm{E}(Z_1) \\ \mathrm{E}(Z_2) \\ \mathrm{E}(Z_3) \end{pmatrix}$$

$\mathrm{Var}\left(\vec{Z}\right)$ est la matrice n, n dont les éléments diagonaux sont les variances et les éléments non diagonaux sont les covariances des coordonnées de $\vec{Z}$. Cette matrice est parfois appellée matrice de variance-covariance. Par exemple :

$$\mathrm{Var}\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \mathrm{Cov}(Z_1, Z_3) \\ \mathrm{Cov}(Z_1, Z_2) & \mathrm{Var}(Z_2) & \mathrm{Cov}(Z_2, Z_3) \\ \mathrm{Cov}(Z_1, Z_3) & \mathrm{Cov}(Z_2, Z_3) & \mathrm{Var}(Z_3) \end{pmatrix}$$

On vérifie le résultat suivant :
Si $\mathbf{C}$ est une matrice p, n et $\vec{Z}$ un vecteur de taille n, alors $\mathbf{C}\vec{Z}$ est un vecteur de taille p de matrice de variance

$$\mathrm{Var}(\mathbf{C}\vec{Z}) = \mathbf{C}\mathrm{Var}(\vec{Z})\mathbf{C}'.$$

En particulier, si p vaut 1, alors $\mathbf{C} = \vec{h}'$ où $\vec{h}$ est un vecteur de taille n, alors

$$\mathrm{Var}(\vec{h}'\vec{Z}) = \vec{h}'\mathrm{Var}(\vec{Z})\vec{h}.$$

Notez que cette dernière quantité est scalaire.

4 Lois normales isotropes

Soient $Y_1 ... Y_n$ indépendantes de loi $N(0, \sigma^2)$, on considère le vecteur

$$\vec{Y} = (Y_1 ... Y_n).$$

À cause de l'indépendance, la densité de $\vec{Y}$ est le produit des densités des coordonnées (résultat admis)

$$\begin{aligned} f_{\vec{Y}}(y_1, ..., y_n) &= f_{Y_1}(y_1) ... f_{Y_n}(y_n) \\ &= \left(2\pi\sigma^2\right)^{-n/2} exp - \frac{1}{2\sigma^2}(y_1^2 + ... + y_n^2) \\ &= \left(2\pi\sigma^2\right)^{-n/2} exp - \frac{\|\vec{y}\|^2}{2\sigma^2}. \end{aligned}$$

On voit donc que la densité de $\vec{Y}$ ne dépend que de la norme $\|\vec{Y}\|$: elle est constante sur toutes les sphères centrées en zéro. Cela implique qu'elle est invariante par rotation ou symétrie : elle est invariante par toutes les isométries de $\mathbb{R}^n$. Rappelons que les isométries correspondent à des changements de bases ortho-normées (BON). Donc si nous exprimons les coordonnées de $\vec{Y}$ dans une autre BON, nous obtiendrons encore des lois $N(0, \sigma^2)$ indépendantes. C'est le résultat que nous utiliserons le plus souvent.

Proposition A.1 *Soient E_1 et E_2, deux sous espaces orthogonaux de $E = \mathbb{R}^n$ de dimensions k_1 et k_2 et soit $\vec{Y}$ un vecteur de $\mathbb{R}^n$ de loi normale isotrope de variance σ^2, alors $\|P_{E_1}(\vec{Y})\|^2$ suit une loi $\sigma^2\chi^2(k_1)$ indépendante de $P_{E_2}(\vec{Y})$.*

Démonstration : Soient $\vec{e}_1, ..., \vec{e}_{k_1}$ et $\vec{e}_{k_1+1}, ..., \vec{e}_{k_1+k_2}$ deux BON de E_1 et E_2, cette base peut être complétée en

$$\vec{e}_1, ..., \vec{e}_{k_1}, \vec{e}_{k_1+1}, ..., \vec{e}_{k_1+k_2}, \vec{e}_{k_1+k_2+1}, ..., \vec{e}_n$$

BON de $\mathbb{R}^n$.
Soient $Z_1, ...Z_n$, les coordonnées de $\vec{Y}$ dans cette base ; elles sont indépendantes de loi $N(0, \sigma^2)$. Comme

$$\|P_{E_1}(\vec{Y})\|^2 = Z_1^2 + ... + Z_{k_1}^2 = \sigma^2 \left((Z_1/\sigma)^2 + ... + ((Z_{k_1}/\sigma)^2 \right)$$

$$P_{E_2}(\vec{Y}) = Z_{k_1+1}\vec{e}_{k_1+1} + ... + Z_{k_1+k_2}\vec{e}_{k_1+k_2}$$

On voit directement l'indépendance et le fait que la loi de $\|P_{E_1}(\vec{Y})\|^2$ est bien une loi $\sigma^2\chi^2(k_1)$. ■

Proposition A.2 *Soit $\vec{Y}$ un vecteur normal isotrope de $\mathbb{R}^n$ de variance σ^2. Soit $\vec{C}'\vec{Y}$ une combinaison linéaire, alors*

$$\vec{C}'\vec{Y} \text{ suit la loi } N(0, \sigma^2\vec{C}'\vec{C}) = N(0, \sigma^2\|\vec{C}\|^2)$$

Démonstration : $\vec{C}$ est un vecteur de $\mathbb{R}^n$, $\vec{e}_1 = \vec{C}/\|\vec{C}\|$ peut être complété en une BON. La coordonnée de $\vec{Y}$ sur $\vec{e}_1$ suit une loi $N(0, \sigma^2)$ donc

$$\vec{C}'\vec{Y} = \|\vec{C}\| < \vec{e}_1, \vec{Y} >$$

Suit une loi $N(0, \sigma^2\|\vec{C}\|^2)$. ■

5 Vecteurs gaussiens généraux

Soit $\mathbf{C}$ une matrice (n, n), $\vec{Y}$ un vecteur gaussien isotrope de variance 1, et de dimension n, $\vec{\mu}$ un vecteur de $\mathbb{R}^n$ alors par définition

$$\vec{Z} = \vec{\mu} + \mathbf{C}\vec{Y}$$

suit une loi normale d'espérance $\vec{\mu}$ et de variance $\mathbf{CC}'$ notée $N(\vec{\mu}, \mathbf{CC}')$.

Proposition A.3 *Si $\vec{Z}$ est un vecteur gaussien d'espérance $\vec{\mu}$ et de matrice de variance Σ et si $\vec{h}'\vec{Z}$ est une combinaison linéaire :*

$$\vec{h}'\vec{Z} = \sum_i \vec{h}_i\vec{Z}_i$$

alors $\vec{h}'\vec{Z}$ suit une loi normale unidimensionnelle d'espérance $\vec{h}'\vec{\mu}$ et de variance $\vec{h}'\Sigma\vec{h}$.

Démonstration : Par définition $\vec{Z} = \vec{\mu} + \mathbf{T}\vec{Y}$ avec $\mathbf{TT}' = \Sigma$ et $\vec{Y}$ isotrope de variance 1. Dans ces conditions

$$\vec{h}'\vec{Z} = \vec{h}'\vec{\mu} + \vec{h}'\mathbf{T}\vec{Y}.$$

Comme $\vec{h}'\mathbf{T}$ est un vecteur, il suffit d'appliquer la proposition A.2. ■

De même on montre

Proposition A.4 *Si $\mathbf{M}$ est une matrice n, p, si $\vec{Z}$ est un vecteur normal de dimension n, et de loi*

$$N(\vec{\mu}, \Sigma)$$

alors $\mathbf{M}\vec{Z}$ suit la loi

$$N(\mathbf{M}\vec{\mu}, \mathbf{M}\Sigma\mathbf{M}').$$

Pour un présentation plus détaillée de notions sur les vecteurs gaussiens on peut consulter Toulouse P.(1999) *Thèmes de probabilités et statistique, chap.2.* Dunod.

Table des matières