

# Le Modèle Linéaire Gaussien Général

Application aux plans factoriels, aux modèles mixtes  
et aux modèles pour données répétées

(version de mars 2010)

Alain BACCINI

---

Institut de Mathématiques de Toulouse — UMR CNRS 5219  
Université Paul Sabatier — 31062 – Toulouse cedex 9.



# Table des matières

<b>1</b>	<b>Introduction à la modélisation statistique</b>	<b>9</b>
1.1	Notion de modélisation mathématique . . . . .	9
1.2	Principales méthodes de modélisation statistique . . . . .	10
1.3	Préliminaires à toute modélisation statistique . . . . .	11
1.3.1	“Nettoyage” des données . . . . .	12
1.3.2	Analyses univariées . . . . .	12
1.3.3	Analyses bivariées . . . . .	12
1.3.4	Analyses multivariées quantitatives . . . . .	13
1.3.5	Analyses multivariées qualitatives . . . . .	13
1.3.6	Bilan . . . . .	13
1.4	Formalisation de la notion de modèle statistique . . . . .	13
<b>2</b>	<b>Généralités sur le modèle linéaire</b>	<b>15</b>
2.1	Définitions et notations . . . . .	16
2.1.1	Le modèle linéaire . . . . .	16
2.1.2	Le modèle linéaire gaussien . . . . .	17
2.1.3	Notations . . . . .	17
2.1.4	Trois exemples basiques . . . . .	18
2.2	Estimation des paramètres . . . . .	19
2.2.1	Estimation de $\beta$ dans le cas général . . . . .	19
2.2.2	Moindres carrés ordinaires et moindres carrés généralisés . . . . .	19
2.2.3	Estimation de $\beta$ dans le cas gaussien . . . . .	19
2.2.4	Estimation d’une fonction linéaire de $\beta$ . . . . .	20
2.2.5	Valeurs prédites et résidus . . . . .	20
2.2.6	Estimation de $\sigma^2$ dans le cas général . . . . .	21
2.2.7	Estimation de $\sigma^2$ dans le cas gaussien . . . . .	22
2.2.8	Intervalle de confiance pour une fonction linéaire de $\beta$ . . . . .	22
2.2.9	Intervalle de confiance conjoints : méthode de Bonferroni . . . . .	22
2.3	Test d’une hypothèse linéaire en $\beta$ . . . . .	23
2.4	Contrôles d’un modèle linéaire . . . . .	23
2.4.1	Contrôles de la qualité d’un modèle . . . . .	24
2.4.2	Contrôles de la validité d’un modèle . . . . .	24
2.5	Panorama sur le modèle linéaire . . . . .	24
2.5.1	Le modèle linéaire gaussien de base . . . . .	24
2.5.2	Le modèle linéaire gaussien général . . . . .	25
2.5.3	Le modèle linéaire généralisé . . . . .	25
<b>3</b>	<b>L’analyse de variance univariée</b>	<b>27</b>
3.1	Cas d’un seul facteur . . . . .	28
3.1.1	Écriture initiale du modèle . . . . .	28
3.1.2	Paramétrage centré . . . . .	29
3.1.3	Paramétrage SAS . . . . .	29
3.1.4	Estimation des paramètres . . . . .	30
3.1.5	Test de l’effet du facteur $F$ . . . . .	31
3.1.6	Autres tests . . . . .	32

3.1.7	Illustration . . . . .	32
3.2	Cas de deux facteurs croisés . . . . .	35
3.2.1	Notations . . . . .	35
3.2.2	Écriture initiale du modèle . . . . .	37
3.2.3	Paramétrage centré . . . . .	37
3.2.4	Paramétrage SAS . . . . .	38
3.2.5	Estimation des paramètres . . . . .	39
3.2.6	Tests d'hypothèses . . . . .	40
3.2.7	Cas particulier d'un plan équilibré . . . . .	41
3.2.8	Illustration . . . . .	42
3.3	Cas de trois facteurs croisés . . . . .	46
3.3.1	Notations . . . . .	46
3.3.2	Modèle . . . . .	48
3.3.3	Estimations . . . . .	49
3.3.4	Tests . . . . .	49
3.4	Généralisation . . . . .	50
<b>4</b>	<b>Étude de quelques plans d'expériences incomplets</b>	<b>51</b>
4.1	La méthode des blocs . . . . .	52
4.1.1	Principes . . . . .	52
4.1.2	Plans en blocs complets équilibrés . . . . .	52
4.1.3	Plans en blocs incomplets équilibrés . . . . .	53
4.2	Les plans en carrés latins et gréco-latins . . . . .	55
4.2.1	Les plans en carrés latins . . . . .	56
4.2.2	Les plans en carrés gréco-latins . . . . .	58
4.3	Les plans à plusieurs facteurs à deux niveaux . . . . .	60
4.3.1	Introduction . . . . .	60
4.3.2	Cas $p = 2$ . . . . .	60
4.3.3	Cas $p = 3$ . . . . .	60
4.3.4	Cas $4 \leq p \leq 6$ . . . . .	62
4.3.5	Cas $p > 6$ . . . . .	65
4.4	Compléments . . . . .	66
<b>5</b>	<b>L'analyse de variance multivariée</b>	<b>67</b>
5.1	Écriture du modèle à un seul facteur . . . . .	68
5.1.1	Les données . . . . .	68
5.1.2	Le modèle . . . . .	68
5.2	Estimation des paramètres du modèle à un facteur . . . . .	69
5.2.1	Vraisemblance et log-vraisemblance . . . . .	69
5.2.2	Estimation maximum de vraisemblance . . . . .	70
5.2.3	Propriétés des estimateurs maximum de vraisemblance . . . . .	71
5.2.4	Indications sur la loi de Wishart . . . . .	71
5.3	Tests dans le modèle à un facteur . . . . .	71
5.3.1	Les matrices <b>H</b> et <b>E</b> . . . . .	71
5.3.2	Le test de Wilks . . . . .	72
5.3.3	Autres tests . . . . .	74
5.3.4	Cas particulier : $J = 2$ . . . . .	75
5.4	Illustration . . . . .	75
5.5	Modèle à deux facteurs croisés . . . . .	78
5.5.1	Données, modèle et paramétrages . . . . .	78
5.5.2	Tests et estimations . . . . .	78
5.5.3	Généralisation . . . . .	79
5.5.4	Illustration . . . . .	79

<b>6</b>	<b>Modèles à effets aléatoires et modèles mixtes</b>	<b>83</b>
6.1	Modèle à un facteur à effets aléatoires . . . . .	84
6.1.1	Écriture du modèle pour une observation . . . . .	84
6.1.2	Écriture matricielle du modèle . . . . .	85
6.1.3	Estimation de la moyenne . . . . .	86
6.1.4	Estimation des composantes de la variance . . . . .	86
6.1.5	Intervalles de confiance . . . . .	91
6.1.6	Test de l'effet du facteur . . . . .	92
6.1.7	Prévision d'un effet aléatoire . . . . .	92
6.1.8	Illustration . . . . .	92
6.2	Modèle à deux facteurs croisés à effets aléatoires . . . . .	97
6.2.1	Écritures du modèle . . . . .	97
6.2.2	Estimation des composantes de la variance dans le cas équilibré . . . . .	98
6.2.3	Tests des effets aléatoires dans le cas équilibré . . . . .	99
6.3	Modèles mixtes . . . . .	100
6.3.1	Écriture générale d'un modèle linéaire gaussien mixte . . . . .	100
6.3.2	Estimation des paramètres dans le cas équilibré . . . . .	102
6.3.3	Estimation des paramètres dans le cas déséquilibré . . . . .	102
6.3.4	Intervalles de confiance . . . . .	105
6.3.5	Tests de significativité des facteurs . . . . .	105
6.3.6	Prévisions dans les modèles mixtes . . . . .	107
6.3.7	Illustration . . . . .	107
<b>7</b>	<b>Modèles pour données répétées</b>	<b>113</b>
7.1	Introduction . . . . .	114
7.2	Analyses préliminaires . . . . .	114
7.2.1	ANOVA réalisée à chaque instant $t$ . . . . .	114
7.2.2	ANOVA réalisée sur la moyenne temporelle des observations . . . . .	115
7.3	Modèle à un facteur à effets fixes pour données répétées . . . . .	115
7.3.1	Principe . . . . .	115
7.3.2	Terminologie . . . . .	116
7.3.3	Mise en œuvre . . . . .	116
7.4	Les structures usuelles de covariance pour $\mathbf{R}$ . . . . .	117
7.5	Cas particulier : la structure "compound symmetry" . . . . .	119
7.5.1	Propriété préliminaire . . . . .	119
7.5.2	Conséquences . . . . .	119
7.5.3	Le test de sphéricité de Mauchly . . . . .	119
7.6	Modèles mixtes pour données répétées . . . . .	120
7.6.1	Principe . . . . .	120
7.6.2	Usage de la procédure MIXED . . . . .	121
7.6.3	Inférence . . . . .	121
7.7	Illustration . . . . .	122
<b>A</b>	<b>À propos de la méthode de Bonferroni</b>	<b>137</b>
A.1	Rappels sur la méthode de Bonferroni . . . . .	137
A.2	Les commandes means et lsmeans de la procédure GLM de SAS . . . . .	138
A.2.1	Principe général . . . . .	138
A.2.2	Tests des différences et méthode de Bonferroni . . . . .	139
A.2.3	Cas particulier du modèle additif : premières bizarreries . . . . .	142
A.2.4	Cas particulier d'un plan incomplet : nouvelles bizarreries . . . . .	145
A.3	Usage de lsmeans pour les graphiques d'interactions . . . . .	147
A.4	Les données "traitements" . . . . .	149

<b>B</b>	<b>Note sur les différents types de sommes de carrés</b>	<b>151</b>
B.1	Introduction . . . . .	151
B.2	Décomposition de la somme totale des carrés . . . . .	151
B.3	Exemple . . . . .	153
B.4	Traitement des données avec SAS . . . . .	153
B.4.1	Traitement initial . . . . .	153
B.4.2	Somme des carrés relative aux interactions . . . . .	155
B.4.3	Somme des carrés relative au facteur $F_2$ . . . . .	155
B.4.4	Somme des carrés relative au facteur $F_1$ . . . . .	157
B.4.5	Retour sur les sommes de type III . . . . .	159
B.4.6	Cas particulier du modèle additif . . . . .	164
B.5	Quelle philosophie suivre? . . . . .	165
<b>C</b>	<b>Un exercice sur les carrés latins</b>	<b>167</b>
<b>D</b>	<b>Indications sur les critères de choix de modèle</b>	<b>169</b>
D.1	Le $C_p$ de Mallows . . . . .	169
D.2	La déviance relative . . . . .	170
D.3	Le critère A.I.C. . . . .	170
D.4	Le critère B.I.C. . . . .	170
<b>E</b>	<b>Tests multidimensionnels pour données répétées</b>	<b>173</b>
E.1	Les données . . . . .	173
E.2	Traitement avec la commande <code>repeated</code> de la procédure GLM . . . . .	174
E.3	Traitement multivarié des variables d'évolution . . . . .	176
E.3.1	Introduction . . . . .	176
E.3.2	Tests des interactions . . . . .	177
E.4	Tests relatifs au temps . . . . .	178
E.4.1	Expression de la matrice $\mathbf{H}_T$ . . . . .	178
E.4.2	Application . . . . .	179
E.5	Bilan . . . . .	180
<b>F</b>	<b>Spécificité de la structure "compound symmetry"</b>	<b>183</b>
F.1	Étude des éléments propres d'une matrice particulière . . . . .	183
F.2	Application à la structure "compound symmetry" . . . . .	183
<b>G</b>	<b>Bibliographie</b>	<b>185</b>
G.1	Ouvrages généraux . . . . .	185
G.2	Articles spécialisés . . . . .	186

## Avant-propos

### Origine de ce document

Le présent document a été rédigé dans le cadre de l'enseignement "**Modèle Linéaire Gaussien Général**" du Master Professionnel (deuxième année) **Statistique et Économétrie** de Toulouse, commun aux deux universités Toulouse I Capitole et Paul Sabatier (Toulouse III).

Cet enseignement se déroule en 30 heures, 12 heures de cours et 18 heures de T.P. (devant des ordinateurs équipés du logiciel statistique SAS). Tous les modèles présentés sont ainsi illustrés au moyen du logiciel SAS.

Les trois premiers chapitres constituent essentiellement des révisions de notions en principe vues au niveau d'une première année de Master orienté vers la statistique. Le cœur de ce cours est constitué des chapitres 4 à 7.

### Remerciements

Ce cours doit beaucoup à J.R. Mathieu qui a mis en place le module "Modèle Linéaire Gaussien Général" lors de la création du D.E.S.S. (ancienne appellation de la deuxième année du Master Professionnel) "Statistique et Économétrie" et que nous tenons ici à remercier chaleureusement.



# Chapitre 1

## Introduction à la modélisation statistique

*Avant d'entrer dans le cœur de notre sujet, le modèle linéaire gaussien général, nous situons tout d'abord, dans ce chapitre d'introduction, la modélisation statistique au sein de la modélisation mathématique. Nous indiquons ensuite quelles sont les principales méthodes de modélisation statistique et nous précisons, parmi ces dernières, les méthodes traitées dans ce cours. Nous rappelons également les pré-traitements des données qui sont indispensables avant toute modélisation statistique. Enfin, nous donnons une formalisation plus mathématique de ce qu'est la modélisation statistique.*

### 1.1 Notion de modélisation mathématique

Une grande partie des mathématiques appliquées consiste, d'une certaine façon, à faire de la modélisation, c'est-à-dire à définir un (ou plusieurs) modèle(s), de nature mathématique, permettant de rendre compte, d'une manière suffisamment générale, d'un phénomène donné, qu'il soit physique, biologique, économique ou autre.

De façon un peu schématique, on peut distinguer la modélisation déterministe (au sein d'un modèle déterministe, on ne prend pas en compte de variations aléatoires) et la modélisation stochastique (qui prend en compte ces variations aléatoires en essayant de leur associer une loi de probabilité).

Les outils classiques de la modélisation déterministe sont les équations différentielles ordinaires (EDO) et les équations aux dérivées partielles (EDP), qui prennent en compte les variations d'un phénomène en fonction de facteurs tels que le temps, la température... Ces équations ont rarement des solutions explicites et leur résolution nécessite, le plus souvent, la mise en œuvre d'algorithmes numériques plus ou moins sophistiqués, permettant d'obtenir une solution, éventuellement approchée. C'est le champ d'application de ce que l'on appelle aujourd'hui le calcul scientifique.

La modélisation stochastique a pour but essentiel de préciser des lois de probabilité rendant compte des variations aléatoires de certains phénomènes, variations dues à des causes soit inconnues, soit impossible à mesurer (par exemple, parce qu'elles sont à venir).

Au sein de la modélisation stochastique, la modélisation probabiliste a surtout pour but de donner un cadre formel permettant, d'une part de décrire les variations aléatoires dont il est question ci-dessus, d'autre part d'étudier les propriétés générales des phénomènes qui les régissent. Plus appliquée, la modélisation statistique consiste essentiellement à définir des outils appropriés pour modéliser des données observées, en tenant compte de leur nature aléatoire.

Il faut noter que le terme de modélisation statistique est très général et que, à la limite, toute démarche statistique en relève. Toutefois, ce qui est traité dans ce cours est relativement précis et constitue une partie spécifique de la modélisation statistique.

## 1.2 Principales méthodes de modélisation statistique

Les méthodes de modélisation statistique sont, en fait, très nombreuses. Nous citons ci-dessous les principales, sachant que la croissance considérable des masses de données enregistrées dans différents secteurs (internet, biologie à haut débit, marketing...), le besoin d'exploiter ces données sur le plan statistique, ainsi que les outils modernes de calcul ont donné naissance ces dernières années (disons depuis le début du XXI<sup>e</sup> siècle) à de nombreuses méthodes, de plus en plus sophistiquées et, dans le même temps, de plus en plus "gourmandes" en temps calcul.

Dans les méthodes décrites ci-dessous, il y a presque toujours une variable privilégiée, en général appelée variable à expliquer, ou variable réponse, et notée  $Y$  (il s'agit d'une variable aléatoire). Le but est alors de construire un modèle permettant d'expliquer "au mieux" cette variable  $Y$  en fonction de variables explicatives observées sur le même échantillon.

### Le modèle linéaire (gaussien) de base

À la fois le plus simple, le plus ancien et le plus connu des modèles statistiques, il englobe essentiellement la régression linéaire, l'analyse de variance et l'analyse de covariance. Dans ce modèle, les variables explicatives (régresseurs ou facteurs) ne sont pas aléatoires (elles sont à effets fixes). Pour pouvoir être exploité pleinement, ce modèle nécessite l'hypothèse de normalité des erreurs, donc de la variable à expliquer (hypothèse gaussienne). Ce modèle est présenté en détail dans le chapitre 2.

### Le modèle linéaire généralisé

Il généralise le précédent à deux niveaux : d'une part, la loi des erreurs, donc de la variable réponse, n'est plus nécessairement gaussienne, mais doit appartenir à l'une des lois de la famille exponentielle ; d'autre part, la liaison linéaire entre l'espérance de la variable réponse et les variables explicatives se fait à travers une fonction particulière appelée fonction lien (spécifiée a priori). Ce modèle englobe différentes méthodes telles que la régression logistique, la régression Poisson, le modèle log-linéaire ou certains modèles de durée de vie.

### Les modèles non linéaires

De façon très générale, il s'agit de modèles permettant d'expliquer la variable réponse (aléatoire) au moyen des variables explicatives (non aléatoires dans les modèles usuels), à travers une fonction quelconque, inconnue (on est donc en dehors du cadre du modèle linéaire généralisé). Cette classe de modèles est très vaste et relève, en général, de la statistique non paramétrique. Citons, à titre d'exemple, la régression non paramétrique, les *GAM* (*Generalized Additive Models*) et les réseaux de neurones.

### Les modèles mixtes

On désigne sous ce terme des modèles permettant d'expliquer la variable aléatoire réponse au moyen de diverses variables explicatives, certaines étant aléatoires (on parle en général de facteurs à effets aléatoires) et intervenant dans la modélisation de la variance du modèle, d'autres ne l'étant pas (on parle de facteurs à effets fixes) et intervenant dans la modélisation de la moyenne. On trouve ainsi des modèles linéaires gaussiens mixtes, des modèles linéaires généralisés mixtes et des modèles non linéaires mixtes. Les premiers d'entre eux (les modèles linéaires gaussiens mixtes) seront introduits au chapitre 6 et utilisés encore au chapitre 7 de ce cours.

### Les modèles pour données répétées

On appelle données répétées, ou données longitudinales, des données observées au cours du temps sur les mêmes individus (en général, il s'agit de personnes ou d'animaux suivis dans le cadre d'une expérimentation médicale ou biologique). De façon claire, il est nécessaire de prendre en compte dans ces modèles une certaine dépendance entre les observations faites sur un même individu à différents instants. Les modèles linéaires ou linéaires généralisés, qu'ils soient standards ou mixtes, sont utilisés dans ce contexte ; nous aborderons les modèles linéaires mixtes pour données répétées au chapitre 7.

### Les modèles pour séries chronologiques

Les séries chronologiques sont les observations, au cours du temps, d'une certaine grandeur représentant un phénomène économique, social ou autre. Si données répétées et séries chronologiques ont en commun de rendre compte de l'évolution au cours du temps d'un phénomène donné, on notera que ces deux types de données ne sont pas réellement de même nature (dans une série chronologique, ce sont rarement des personnes ou des animaux que l'on observe). Pour les séries chronologiques, on utilise des modèles spécifiques : modèles AR (*Auto-Regressive*, ou auto-régressifs), MA (*Moving Average*, ou moyennes mobiles), ARMA, ARIMA (I pour *Integrated*)...

### L'analyse discriminante et la classification

S'il est plus courant d'utiliser ces méthodes dans un contexte d'exploration des données plutôt que dans un contexte de modélisation, l'analyse discriminante et la classification peuvent tout de même être utilisées dans la phase de recherche d'un modèle permettant d'ajuster au mieux les données considérées. C'est en particulier le cas lorsque la variable réponse du modèle envisagé est de nature qualitative.

### Les modèles par arbre binaire de régression et de classification

Ces méthodes (plus connues sous le nom de *CART*, pour *Classification And Regression Trees*) consistent à découper une population en deux parties, en fonction de celle des variables explicatives et du découpage en deux de l'ensemble de ses valeurs ou modalités qui expliquent au mieux la variable réponse. On recommence ensuite sur chaque sous-population ainsi obtenue, ce qui permet de définir, de proche en proche, un arbre binaire et de classer les variables explicatives selon l'importance de leur liaison avec la variable réponse (on parle d'arbre de régression en présence d'une variable réponse quantitative et d'arbre de classification en présence d'une variable réponse qualitative). De telles méthodes peuvent constituer un complément intéressant au modèle linéaire ou au modèle linéaire généralisé.

### Quelques autres modèles

Concernant les méthodes de modélisation statistique, on ne saurait être exhaustif dans cette introduction. Parmi les méthodes récentes, faisant un usage intensif de l'ordinateur, citons, pour mémoire, la régression *PLS* (*Partial Least Squares*), les méthodes d'agrégation, ou de combinaison, de modèles (*bagging*, *boosting*, *random forests*), les méthodes de régularisation et les SVM (*Support Vector Machines*).

Dans ce cours, nous n'aborderons qu'un petit nombre de modèles parmi ceux évoqués ci-dessus. En fait, tous les modèles qui seront abordés relèvent du modèle linéaire gaussien : le modèle de base dans les chapitres 2 et 3 ; le cas particulier des plans d'expériences au chapitre 4 et celui de l'analyse de variance multidimensionnelle au chapitre 5 ; les modèles mixtes au chapitre 6 et les modèles pour données répétées au chapitre 7.

On trouvera d'intéressants développements sur d'autres modèles statistiques dans Saporta (2006) ainsi que dans le document intitulé "Modélisation statistique et apprentissage", rédigé par Ph. Besse et disponible à l'adresse électronique suivante

<http://www.math.univ-toulouse.fr/~besse/>

rubrique "Enseignement".

## 1.3 Préliminaires à toute modélisation statistique

Quel que soit le modèle, ou le type de modèles, envisagé face à un jeu de données, quel que soit le problème qu'il s'agit de traiter, une modélisation statistique ne peut sérieusement s'envisager que sur des données "propres", c'est à dire pré-traitées, afin de les débarasser, autant que faire se peut, de tout ce qui peut nuire à la modélisation : codes erronés, données manquantes, données

aberrantes, variables inutiles, variables redondantes... C'est cet ensemble de pré-traitements que nous décrivons dans ce paragraphe.

On notera que cette phase est parfois appelée *datamanagement*, autrement dit "gestion des données".

### 1.3.1 "Nettoyage" des données

Avant toute chose, il faut disposer d'un fichier informatique contenant les données dans un format exploitable (texte ou excel, par exemple), les individus étant disposés en lignes et les variables en colonnes. Avec ce fichier, il faut essayer de repérer d'éventuels codes interdits ou aberrants : chaîne de caractères pour une variable numérique ; code "3" pour la variable sexe ; valeur 153 pour l'âge d'un groupe d'individus, etc. Une fois repérés, ces codes doivent être corrigés si possible, supprimés sinon.

Dans cette phase, il faut également essayer de repérer des données manquantes en grande quantité, soit sur une colonne (une variable), soit sur une ligne (un individu). Si quelques données manquantes ne sont pas vraiment gênantes dans la plupart des traitements statistiques, il n'en va pas de même lorsque cela concerne un fort pourcentage des observations d'une variable ou d'un individu. Dans ce cas, il est préférable de supprimer la variable ou l'individu (dont la colonne, ou la ligne, serait, de toutes façons, inexploitable).

### 1.3.2 Analyses univariées

Cette phase, souvent fastidieuse, consiste à étudier chaque variable l'une après l'autre, afin d'en connaître les principales caractéristiques et d'en repérer, le cas échéant, certaines anomalies.

Pour les variables quantitatives, on pourra faire un histogramme ou un diagramme en boîte et déterminer des caractéristiques telles que le minimum, le maximum, la moyenne, l'écart-type, la médiane et les quartiles. Cela peut conduire à supprimer une variable (si elle présente très peu de variabilité), à la transformer (par exemple, en prenant son logarithme si elle est à valeurs positives et très dissymétrique), ou encore à repérer des valeurs très particulières (que l'on devra, éventuellement, corriger ou éliminer).

Pour les variables qualitatives, on pourra faire un diagramme en colonnes des modalités et déterminer les effectifs et les fréquences de ces dernières. Cela pourra encore conduire à supprimer une variable (si tous les individus, ou presque, présentent la même modalité), ou à en regrouper des modalités "proches" (si certains effectifs sont trop faibles).

Ces analyses univariées permettent également de prendre connaissance des données et de fournir certaines indications pour la phase ultérieure de modélisation. Toutefois, il faut noter que ces analyses peuvent être inenvisageables avec des données "fortement multidimensionnelles", c'est-à-dire comportant des centaines, voire des milliers, de variables ; on rencontre aujourd'hui de telles données dans certains contextes particuliers.

### 1.3.3 Analyses bivariées

Ces analyses ont pour but d'étudier d'éventuelles liaisons existant entre couples de variables. Il peut s'agir de deux variables explicatives, dont on soupçonne qu'elles sont fortement corrélées, dans le but d'éliminer l'une des deux. Il peut aussi s'agir d'étudier les liens entre la variable à expliquer et chaque variable explicative (de façon systématique), pour avoir une première idée des variables explicatives susceptibles de jouer un rôle important lors de la modélisation. Enfin, ces analyses peuvent aussi permettre de repérer des points aberrants (ou extrêmes) qui n'ont pas pu l'être avec les analyses univariées.

Rappelons que, pour étudier la liaison entre deux variables quantitatives, on dispose, comme graphique, du nuage de points (ou diagramme de dispersion) et, comme indicateur de liaison, du coefficient de corrélation linéaire. Dans le cas d'une variable quantitative et d'une variable qualitative, on dispose du diagramme en boîtes parallèles et du rapport de corrélation. Enfin, dans le cas de deux variables qualitatives, on utilise en général un diagramme en colonnes de profils (profils-lignes ou profils-colonnes selon ce que l'on souhaite mettre en évidence) et des indicateurs de liaison liés au khi-deux (coefficients de Tschuprow ou de Cramér).

### 1.3.4 Analyses multivariées quantitatives

Elles consistent à déterminer la matrice des corrélations entre toutes les variables quantitatives considérées, notamment la variable à expliquer, lorsque celle-ci est quantitative. Cela peut permettre encore de supprimer des variables très corrélées, par exemple afin d'éviter de faire une régression sur de telles variables, dont on sait que les résultats seraient très instables, voire sans aucune signification. Cela permet aussi de prendre connaissance de la structure de corrélation entre les variables considérées, ce qui est toujours utile dans le cadre d'une modélisation.

On peut également envisager, à ce niveau, de réaliser une analyse en composantes principales (A.C.P.) de toutes ces variables, afin de préciser davantage, de façon globale, leurs relations linéaires.

### 1.3.5 Analyses multivariées qualitatives

C'est le pendant des analyses ci-dessus, cette fois pour les variables qualitatives. On peut, tout d'abord, déterminer la matrice des coefficients de Tschuprow (ou celle des coefficients de Cramér) et l'analyser comme une matrice de corrélations. Toutefois, il est bien connu que, dans la pratique, ces coefficients sont systématiquement petits : pratiquement toujours inférieurs à 0.5 et le plus souvent compris entre 0.1 et 0.3. Leur interprétation est donc, en général, assez délicate. Ils permettent néanmoins de repérer les liaisons les plus importantes, même si elles sont de l'ordre de 0.3, 0.4 ou 0.5.

Il est d'autant plus important d'envisager, dans ces analyses préliminaires, de réaliser une analyse des correspondances multiples (A.C.M.) entre variables qualitatives. Celle-ci permettra, le cas échéant, de confirmer une liaison forte entre certains couples de variables et, si nécessaire, d'en éliminer quelques-unes. L'A.C.M. permet également de regrouper certaines modalités d'une même variable lorsque celles-ci apparaissent proches dans l'ensemble des résultats et, par suite, de simplifier les données. Enfin, le tableau de Burt, fourni avec les résultats de l'A.C.M., permet de repérer des occurrences très faibles pour certains croisements de modalités et d'envisager encore d'autres regroupements.

### 1.3.6 Bilan

Une fois réalisées toutes les étapes préliminaires décrites ci-dessus, on dispose de données "mises au propre", simplifiées, et dont on commence à connaître certaines caractéristiques. On peut, à partir de ce moment là, envisager leur modélisation.

Les modèles susceptibles d'être adaptés aux données considérées, parmi tous ceux décrits dans le paragraphe précédent, sont nécessairement limités à ce stade là. Ils sont fonction de la nature des données ainsi que des questions posées par l'utilisateur, autrement dit de ses objectifs.

Insistons ici sur le fait que des données sont toujours recueillies (produites) par un utilisateur (biologiste, informaticien, gestionnaire...) dans un but bien précis. La modélisation statistique doit avoir pour objectif premier de répondre aux questions que s'est posé cet utilisateur lorsqu'il a décidé de recueillir les données. Une collaboration entre utilisateur et statisticien est donc, à ce niveau là, absolument indispensable.

## 1.4 Formalisation de la notion de modèle statistique

Même si nous ne l'utilisons que fort peu dans la suite de ce cours, nous donnons, dans ce dernier paragraphe, une formalisation de ce qu'est un modèle statistique, afin de relier cette notion au formalisme habituellement utilisé en calcul des probabilités.

La notion de modèle statistique correspond à la modélisation d'une succession d'expériences aléatoires, chacune associée à une observation de l'échantillon considéré. Ainsi, considérons  $n$  variables aléatoires réelles (v.a.r.)  $Y_i$ , chacune associée à une expérience aléatoire dont le résultat est la valeur observée de  $Y_i$  (en fait, on suppose ici que l'expérience considérée est quantitative, par exemple le résultat d'une certaine mesure ; cela étant, ce qui suit se généralise sans difficulté au cas qualitatif).

On suppose donc, au départ, que les v.a.r.  $Y_i$  sont définies sur un certain espace probabilisé  $(\Omega, \mathcal{A}, \Pi)$  et sont à valeurs dans  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ . Si l'on appelle  $Q$  la loi de probabilité conjointe des v.a.r.  $(Y_1, \dots, Y_n)$ , soit encore la loi induite sur  $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$  par  $Y = (Y_1, \dots, Y_n)$ , alors le modèle statistique associé à l'expérience considérée est, par définition :

$$(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, Q).$$

C'est donc l'espace probabilisé qui va rendre compte des expériences aléatoires réalisées. Ainsi, préciser le modèle (faire des hypothèses...) reviendra à préciser la loi de probabilité  $Q$ .

La première hypothèse que l'on fait généralement dans la pratique est celle de l'**indépendance** des différentes expériences, autrement dit l'indépendance mutuelle des v.a.r.  $Y_i$ ,  $i = 1, \dots, n$ . Si l'on appelle  $P_i$  la loi de probabilité induite par  $Y_i$  sur  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ , le modèle statistique peut alors se mettre sous la forme suivante :

$$(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \prod_{i=1}^n P_i).$$

On retiendra que c'est ce cadre général qui est celui du modèle linéaire et du modèle linéaire généralisé, l'hypothèse de linéarité concernant, dans les deux cas, la relation entre  $\mathbb{E}(Y_i)$  et les variables explicatives.

Une autre hypothèse, souvent faite dans la pratique, est que les  $Y_i$  ont toutes la même loi de probabilité (elles sont **identiquement distribuées**). Dans ce cas, on a  $P_i = P, \forall i = 1, \dots, n$ , et le modèle devient :

$$(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, P^n).$$

On a coutume de le noter  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, P)^{\otimes n}$  ou, plus simplement,  $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, P)^n$ . C'est ce qu'on appelle le **modèle d'échantillonnage** qui suppose les v.a.r. **indépendantes et identiquement distribuées** (i.i.d.). On notera que ce modèle ne peut servir de cadre au modèle linéaire que pour la loi des erreurs (les v.a.r.  $Y_i$  n'ont pas toutes, dans le modèle linéaire, la même espérance).

Dans la pratique, un modèle statistique n'est réellement opérationnel que si l'on précise la loi de probabilité  $P$  (cas i.i.d.) ou les lois  $P_i$  (cas seulement indépendant ; dans ce dernier cas, les  $P_i$  sont en général choisies dans une même famille de lois : normale, binomiale...). Après avoir ainsi précisé la loi de probabilité (ou la famille de lois de probabilité) du modèle, il reste d'abord à faire des tests, d'une part pour essayer de simplifier le modèle retenu, d'autre part pour tester la significativité de ce dernier, ensuite à en estimer les paramètres. C'est tout ce travail – choix de la loi de probabilité ou de la famille de lois, tests, choix du modèle, estimation des paramètres du modèle retenu, validation du modèle – qui constitue la modélisation statistique.

## Chapitre 2

# Généralités sur le modèle linéaire

L'objectif du chapitre 2 est uniquement de mettre en place les principaux éléments du modèle linéaire (essentiellement gaussien), à savoir l'estimation ponctuelle, l'estimation par intervalle de confiance et les tests.

Pour des compléments bibliographiques, nous renvoyons essentiellement à six ouvrages : trois en français et trois autres en langue anglaise. Azaïs & Bardet (2005) est un ouvrage consacré spécifiquement au modèle linéaire et constitue un excellent complément de ce cours ; Monfort (1997) propose une approche très mathématique, de la statistique en général et du modèle linéaire en particulier ; Saporta (2006) est d'un abord plus simple, le modèle linéaire ne constituant qu'une petite partie de cet ouvrage très complet et très intéressant ; Jorgensen (1993) couvre bien les chapitres 2 et 3 de ce cours ; Milliken & Johnson (1984) en couvre la presque totalité ; enfin, Rencher & Schaalje (2008) est notre ouvrage de référence sur le modèle linéaire. Cela étant, signalons que le nombre d'ouvrages consacrés, au moins partiellement, au modèle linéaire est considérable.

### Résumé

Précisons l'écriture du modèle linéaire pour tout individu  $i$  ( $i = 1, \dots, n$ ) d'un échantillon de taille  $n$  :

$$Y_i = \sum_{j=1}^p \beta_j X_i^j + U_i .$$

$Y_i$  est la variable aléatoire réelle réponse et  $U_i$  est la variable aléatoire réelle erreur, supposée  $\mathcal{N}(0, \sigma^2)$ , les  $U_i$  étant indépendantes (et donc i.i.d.). Les  $\beta_j$  sont des coefficients, des paramètres inconnus, à estimer. Les  $X_i^j$  sont les valeurs des variables explicatives qui ne sont en général pas considérées comme aléatoires : on suppose qu'il s'agit de valeurs choisies, contrôlées.

Matriciellement, on peut réécrire

$$Y = \mathbf{X}\beta + U ,$$

où  $Y$  et  $U$  sont des vecteurs aléatoires de  $\mathbb{R}^n$ ,  $\mathbf{X}$  est une matrice  $n \times p$  et  $\beta$  est le vecteur de  $\mathbb{R}^p$  des paramètres.

Si l'estimation ponctuelle est possible sans aucune hypothèse de distribution sur les erreurs du modèle, grâce à la méthode des moindres carrés, il n'en va pas de même pour l'estimation par intervalle de confiance et pour les tests : dans ce cas, l'hypothèse de normalité des erreurs (l'hypothèse gaussienne) est indispensable. De manière souvent implicite, l'hypothèse gaussienne sera faite dans tout ce cours car elle est quasiment partout indispensable.

**L'estimation ponctuelle** du vecteur des paramètres  $\beta$ , que ce soit par moindres carrés ou par maximum de vraisemblance dans le cas gaussien, conduit au résultat suivant :

$$\hat{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y.$$

On appelle valeurs prédites les  $\hat{Y}_i$ , coordonnées du vecteur aléatoire

$$\hat{Y} = \mathbf{X}\hat{B} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = \mathbf{H}Y,$$

où  $\mathbf{H}$  est la matrice de projection orthogonale sur le sous-espace vectoriel de  $\mathbb{R}^n$  engendré par les colonnes de  $\mathbf{X}$ .

On appelle résidus les  $\hat{U}_i$ , coordonnées du vecteur aléatoire

$$\hat{U} = Y - \hat{Y} = \mathbf{H}^\perp Y,$$

où  $\mathbf{H}^\perp = \mathbf{I}_n - \mathbf{H}$  est la matrice de projection orthogonale sur le sous-espace vectoriel de  $\mathbb{R}^n$  supplémentaire orthogonal au précédent.

L'estimateur de la variance du modèle ( $\sigma^2$ ), après correction de biais, est donnée par :

$$\hat{\Sigma}^2 = \frac{\sum_{i=1}^n \hat{U}_i^2}{n-p} = \frac{\|\hat{U}\|^2}{n-p}.$$

**L'estimation par intervalle de confiance** d'une fonction linéaire des paramètres,  $c'\beta = \sum_{j=1}^p c_j \beta_j$ , conduit à l'intervalle

$$c'\hat{\beta} \pm t [\hat{\sigma}^2 c'(\mathbf{X}'\mathbf{X})^{-1}c]^{1/2},$$

où  $t = t_{n-p}(1 - \frac{\alpha}{2})$  est le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi de Student à  $n - p$  degrés de liberté. Le coefficient de sécurité de cet intervalle est  $1 - \alpha$ , autrement dit son risque est  $\alpha$ .

**Le test** d'une hypothèse nulle  $\{H_0 : \mathbf{C}'\beta = 0\}$ , linéaire en  $\beta$ , contre l'alternative opposée, se fait au moyen de la statistique de Fisher (ou Fisher-Snedecor) qui s'écrit :

$$F = \frac{NUM}{q\hat{\Sigma}^2},$$

où  $q$  est le nombre de contraintes définies par  $H_0$  (autrement dit, le rang de  $\mathbf{C}$ , matrice de dimension  $p \times q$ , avec  $1 \leq q < p$ ) et où le numérateur  $NUM$  peut s'écrire sous l'une des formes suivantes

$$NUM = \|\hat{U}_0\|^2 - \|\hat{U}\|^2 = \|\hat{U}_0 - \hat{U}\|^2 = \|\hat{Y}_0 - \hat{Y}\|^2 = \|\hat{B}_0 - \hat{B}\|_{X,X}^2 = \hat{B}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\hat{B},$$

$\hat{B}_0$ ,  $\hat{Y}_0$  et  $\hat{U}_0$  désignant respectivement le vecteur des estimateurs, celui des valeurs prédites et celui des résidus dans le modèle sous  $H_0$ .



## 2.1 Définitions et notations

### 2.1.1 Le modèle linéaire

**Définition 1** On appelle modèle linéaire un modèle statistique qui peut s'écrire sous la forme

$$Y = \sum_{j=1}^p \beta_j X^j + U.$$

Dans la définition ci-dessus, les éléments intervenant ont les caractéristiques suivantes :

- $Y$  est une variable aléatoire réelle (v.a.r.) que l'on observe et que l'on souhaite expliquer, ou prédire (ou les deux à la fois); on l'appelle variable à expliquer, ou **variable réponse** (parfois aussi variable dépendante, ou variable endogène).
- Chaque variable  $X^j$  est une variable réelle (éventuellement ne prenant que les valeurs 0 et 1), non aléatoire dans le modèle de base, également observée; l'ensemble des  $X^j$  est censé expliquer  $Y$ , en être la cause (au moins partiellement); les variables  $X^j$  sont appelées variables explicatives, ou **prédicteurs** (parfois variables indépendantes, ou variables exogènes).  
Pour chaque variable  $X^j$ , l'expérimentateur est supposé choisir diverses valeurs caractéristiques (au moins deux) pour lesquelles il réalise une ou plusieurs expériences en notant les valeurs correspondantes de  $Y$ : il contrôle donc les variables  $X^j$ , pour cette raison appelées aussi **variables contrôlées**; en réalité, dans la pratique, ce n'est pas toujours exactement le cas.

- Les  $\beta_j$  ( $j = 1, \dots, p$ ) sont des coefficients, des **paramètres**, non observés; on devra donc les estimer au moyen de techniques statistiques appropriées.
- $U$  est le terme d'erreur du modèle; c'est une v.a.r. non observée pour laquelle on fait systématiquement les hypothèses suivantes :

$$\mathbb{E}(U) = 0 ; \text{Var}(U) = \sigma^2 > 0$$

( $\sigma^2$  est un paramètre inconnu, également à estimer). Lorsqu'on répète les observations de  $Y$  et des  $X^j$ , on suppose que la variance de  $U$  est constante ( $\sigma^2$ ); c'est ce que l'on appelle l'hypothèse d'**homoscédasticité**.

- Les hypothèses faites sur  $U$  entraînent les conséquences suivantes sur  $Y$  :

$$\mathbb{E}(Y) = \sum_{j=1}^p \beta_j X^j ; \text{Var}(Y) = \sigma^2.$$

- L'espérance mathématique de  $Y$  s'écrit donc comme une combinaison linéaire des  $X^j$  : la liaison entre les  $X^j$  et  $Y$  est de nature linéaire (linéaire en moyenne). C'est la raison pour laquelle ce modèle est appelé le *modèle linéaire*.

### 2.1.2 Le modèle linéaire gaussien

C'est un modèle linéaire dans lequel on fait l'hypothèse supplémentaire que la v.a.r.  $U$  est gaussienne, c'est-à-dire normale. On pose donc :

$$U \sim \mathcal{N}(0, \sigma^2),$$

cette hypothèse entraînant la normalité de  $Y$ .

Si l'on veut, dans un modèle linéaire, pouvoir construire des intervalles de confiance ou faire des tests concernant les paramètres (les  $\beta_j$  et  $\sigma^2$ ), cette hypothèse gaussienne est indispensable. Sauf indication contraire, elle sera faite dans toute la suite de ce cours.

### 2.1.3 Notations

Pour pouvoir faire, au minimum, l'estimation ponctuelle des paramètres  $\beta_j$  et  $\sigma^2$ , il est indispensable de répliquer, de manières indépendantes, les observations simultanées des variables  $X^j$  et  $Y$ .

Nous supposons donc par la suite que  $n$  observations indépendantes sont réalisées et nous écrirons le modèle, pour la  $i$ -ième observation ( $i = 1, \dots, n$ ), sous la forme :

$$Y_i = \sum_{j=1}^p \beta_j X_i^j + U_i \quad (\text{égalité entre v.a.r.}).$$

Les valeurs observées des variables seront notées par des minuscules, de sorte qu'on écrira :

$$y_i = \sum_{j=1}^p \beta_j x_i^j + u_i \quad (\text{égalité entre nombres réels}).$$

Par ailleurs, on notera  $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$  le vecteur aléatoire de  $\mathbb{R}^n$  correspondant à l'ensemble

de l'échantillon des v.a.r. réponses (la notation  $Y$  est identique à celle introduite en 2.1.1 pour une seule v.a.r. réponse, mais cela ne devrait pas entraîner de confusion puisqu'on travaillera dorénavant avec un échantillon),  $\mathbf{X} = (x_i^j)$  la matrice réelle,  $n \times p$ , des valeurs contrôlées des

prédicteurs,  $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$  le vecteur des paramètres dans  $\mathbb{R}^p$  et  $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$  le vecteur aléatoire

de  $\mathbb{R}^n$  contenant les erreurs du modèle (même remarque que ci-dessus).

Matriciellement, le modèle linéaire s'écrit donc

$$Y = \mathbf{X}\beta + U,$$

avec, dans le cas gaussien,

$$U \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n) \text{ et } Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n),$$

$\mathbf{I}_n$  désignant la matrice identité d'ordre  $n$ .

Par la suite, on supposera  $n > p$  (le nombre d'observations est au moins égal au nombre de paramètres à estimer),  $p \geq 1$  (il y a au moins une variable explicative dans le modèle) et  $\mathbf{X}$  de rang  $p$  (les variables  $X^j$  sont linéairement indépendantes).

**Remarque 1** *On notera que les v.a.r.  $U_i$  sont i.i.d. (indépendantes et identiquement distribuées) par hypothèse, alors que les v.a.r.  $Y_i$  sont indépendantes, de même variance, normales dans le cas gaussien, mais n'ont pas toutes la même moyenne (elles ne sont donc pas i.i.d.).*

**Remarque 2** *Dans le modèle linéaire, et plus particulièrement dans l'analyse de variance, la matrice  $\mathbf{X}$  est souvent appelée **matrice d'incidence**.*

### 2.1.4 Trois exemples basiques

#### Le modèle constant, ou modèle "blanc"

Il s'écrit :

$$Y_i = \beta + U_i \quad (Y = \beta \mathbf{1}_n + U).$$

Autrement dit,  $p = 1$  et  $\mathbf{X} = \mathbf{1}_n$  : l'unique prédicteur est la variable constante et égale à 1. Ce modèle n'a pas d'intérêt pratique, mais il est utilisé comme modèle de référence, celui par rapport auquel on comparera d'autres modèles.

#### Le modèle de régression linéaire simple

C'est le modèle suivant :

$$Y_i = \beta_1 + \beta_2 X_i^2 + U_i.$$

Ici,  $p = 2$  et  $\mathbf{X} = (\mathbf{1}_n \ X^2)$  : on a rajouté un "vrai" prédicteur quantitatif ( $X^2$ ) à la constante.

#### Le modèle d'analyse de variance à un facteur à deux niveaux

Ce modèle s'écrit :

$$Y_i = \beta_j + U_i,$$

lorsque la  $i$ -ième observation de  $Y$  est réalisée au niveau  $j$  ( $j = 1, 2$ ) du facteur (la variable explicative est ici qualitative à deux modalités; dans le contexte du modèle linéaire, on parle plutôt de *facteur* à deux *niveaux*). En fait, chaque niveau du facteur est remplacé par une variable indicatrice, de sorte que  $p = 2$ .

Matriciellement, ce modèle peut s'écrire

$$Y = \mathbf{X}\beta + U,$$

avec

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ et } \mathbf{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

Dans la matrice  $\mathbf{X}$  ci-dessus, les  $n_1$  premières lignes sont (1 0) s'il y a  $n_1$  observations réalisées au niveau 1 du facteur, les  $n_2$  suivantes étant (0 1) s'il y a  $n_2$  observations réalisées au niveau 2 du facteur ( $n_1 + n_2 = n$ ).

## 2.2 Estimation des paramètres

### 2.2.1 Estimation de $\beta$ dans le cas général

En l'absence d'hypothèse sur la distribution de  $U$ , on estime  $\beta$  par la méthode des moindres carrés. Elle consiste à poser :

$$\hat{\beta} = \text{Arg min } \|y - \mathbf{X}\beta\|^2, \beta \in \mathbb{R}^p. \quad (2.1)$$

(Cette écriture suppose que  $\mathbb{R}^n$  est muni de la norme euclidienne classique, autrement dit que l'on utilise le critère dit des *moindres carrés ordinaires*.)

On montre alors que ce problème admet la solution unique

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \text{ (estimation),}$$

valeur observée du vecteur aléatoire

$$\hat{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \text{ (estimateur).}$$

#### Propriétés de $\hat{B}$

- $\mathbb{E}(\hat{B}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(Y) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$  :  $\hat{B}$  est un estimateur sans biais de  $\beta$ .
- $\text{Var}(\hat{B}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\sigma^2}{n}\mathbf{S}_n^{-1}$ , avec  $\mathbf{S}_n = \frac{1}{n}\mathbf{X}'\mathbf{X}$  (matrice des variances-covariances empiriques lorsque les variables  $X^j$  sont centrées). On obtient un estimateur convergent, sous réserve que :

$$\lim_{n \rightarrow \infty} \det(\mathbf{S}_n) = d > 0.$$

### 2.2.2 Moindres carrés ordinaires et moindres carrés généralisés

Dans le point 2.1.3, on a posé  $\text{Var}(U) = \sigma^2\mathbf{I}_n$ . Supposons maintenant, de façon plus générale, que  $\text{Var}(U) = \sigma^2\mathbf{V}$ , où  $\mathbf{V}$  est une matrice connue, carrée d'ordre  $n$ , symétrique et strictement définie-positive. On peut alors se ramener au cas précédent en faisant intervenir la matrice  $\mathbf{V}^{-1}$  dans le critère des moindres carrés. Pour cela, on cherche le vecteur  $\hat{\beta}$  de  $\mathbb{R}^p$  solution de :

$$\hat{\beta} = \text{Arg min } \|y - \mathbf{X}\beta\|_{\mathbf{V}^{-1}}^2. \quad (2.2)$$

La solution est donnée par :

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}Y).$$

Le critère (2.1) est appelé critère des moindres carrés ordinaires (MCO), alors que le critère (2.2) est appelé critère des moindres carrés généralisés (MCG) (voir, par exemple, Monfort, 1997, chapitre 26). Le critère des moindres carrés généralisés sera utilisé au chapitre 6.

### 2.2.3 Estimation de $\beta$ dans le cas gaussien

#### Densité d'une loi multinormale

Soit  $Z$  un vecteur aléatoire à valeurs dans  $\mathbb{R}^n$ , de densité gaussienne, admettant  $\mu$  comme vecteur des moyennes ( $\mu \in \mathbb{R}^n$ ) et  $\Sigma$  comme matrice des variances-covariances ( $\Sigma$  est carrée d'ordre  $n$ , symétrique, strictement définie-positive). On rappelle la densité de  $Z$  :

$$f(z) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\det \Sigma)^{1/2}} \exp\left[-\frac{1}{2}(z - \mu)' \Sigma^{-1}(z - \mu)\right].$$

#### Vraisemblance d'un échantillon gaussien de taille $n$

Dans le cadre du modèle linéaire gaussien, le vecteur aléatoire  $Y$  admet pour espérance le vecteur  $\mathbf{X}\beta$  et pour matrice des variances-covariances  $\Sigma = \sigma^2\mathbf{I}_n$ . Sa vraisemblance s'écrit donc :

$$L(y, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)'(y - \mathbf{X}\beta)\right].$$

### Log-vraisemblance

Le logarithme (népérien) de la fonction ci-dessus s'écrit :

$$\begin{aligned} l(y, \beta, \sigma^2) &= \log[L(y, \beta, \sigma^2)] \\ &= -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} (y - \mathbf{X}\beta)' (y - \mathbf{X}\beta) \\ &= \text{constante} - n \log(\sigma) - \frac{1}{2\sigma^2} \|y - \mathbf{X}\beta\|^2. \end{aligned}$$

### Conséquences

Maximiser  $l(y, \beta, \sigma^2)$  selon  $\beta$ , pour trouver l'estimateur maximum de vraisemblance, revient donc à minimiser  $\|y - \mathbf{X}\beta\|^2$  selon  $\beta$ , et redonne l'estimateur  $\hat{B}$  introduit en 2.2.1. Ainsi, estimateurs moindres carrés ordinaires et maximum de vraisemblance sont identiques dans le modèle linéaire gaussien.

### Propriétés

L'estimateur  $\hat{B}$  de  $\beta$  demeure d'une part sans biais, d'autre part convergent, sous la même condition que précédemment. De plus, on peut, dans le cadre gaussien, préciser sa distribution : comme transformée linéaire d'un vecteur gaussien, elle est gaussienne, donc  $\mathcal{N}_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ . Enfin, on peut vérifier que  $\hat{B}$  est un estimateur efficace de  $\beta$  (sa variance est égale à la borne inférieure de l'inégalité de Cramér-Rao).

**Remarque 3** Si les prédicteurs  $X^j$  sont deux à deux orthogonaux, alors  $\mathbf{X}'\mathbf{X} = \text{diag}(\alpha_1 \cdots \alpha_p)$ , avec  $\alpha_j = \sum_{i=1}^n (x_i^j)^2 > 0$  (sinon, la  $j$ -ième colonne de  $\mathbf{X}$  serait nulle et  $\mathbf{X}$  ne serait pas de rang  $p$ ). Il vient donc  $(\mathbf{X}'\mathbf{X})^{-1} = \text{diag}(\frac{1}{\alpha_1} \cdots \frac{1}{\alpha_p})$  et l'on en déduit  $\hat{B}_j \sim \mathcal{N}(\beta_j, \frac{\sigma^2}{\alpha_j})$ , les  $\hat{B}_j$  étant donc mutuellement indépendants. Cette situation se rencontre, dans certains cas particuliers, en analyse de variance (voir chapitre 3).

### 2.2.4 Estimation d'une fonction linéaire de $\beta$

On considère maintenant un vecteur non nul  $c$  de  $\mathbb{R}^p$  et la forme linéaire  $c'\beta$ . On vérifie simplement, dans le modèle gaussien, que l'estimateur maximum de vraisemblance de  $c'\beta$  est  $c'\hat{B}$  et que  $c'\hat{B} \sim \mathcal{N}(c'\beta, \sigma^2 c'(\mathbf{X}'\mathbf{X})^{-1}c)$ . Il s'agit d'un estimateur sans biais, convergent (toujours sous la même condition) et efficace.

On utilise ce résultat pour estimer l'un des paramètres  $\beta_j$ , une différence entre deux paramètres  $\beta_j - \beta_k$ , etc.

### 2.2.5 Valeurs prédites et résidus

#### Valeurs prédites

On appelle vecteur des valeurs prédites le vecteur  $\hat{y}$  de  $\mathbb{R}^n$  défini par :

$$\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y.$$

Il s'agit du vecteur des prédictions (ou approximations)  $\hat{y}_i$  des  $y_i$  réalisées avec le modèle linéaire considéré; on parle aussi de *valeurs ajustées*.

En fait, en posant  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ , on remarque que  $\mathbf{H}$  est la matrice de la projection orthogonale (au sens de la métrique usuelle) sur le sous-espace vectoriel  $F_X$  de  $\mathbb{R}^n$  engendré par les colonnes de  $\mathbf{X}$ . Par suite,  $\hat{y} = \mathbf{H}y$  est la projection orthogonale de  $y$  sur  $F_X$ .

Dans le modèle gaussien, on obtient

$$\hat{Y} = \mathbf{H}Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{H});$$

en effet,  $\hat{Y}$  est gaussien comme transformé linéaire de  $Y$  gaussien,  $\mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta$  (le vecteur  $\mathbf{X}\beta$  étant, par définition, dans le sous-espace  $F_X$ ) et  $\sigma^2\mathbf{H}\mathbf{H}' = \sigma^2\mathbf{H}^2$  ( $\mathbf{H}$  est symétrique) =  $\sigma^2\mathbf{H}$  ( $\mathbf{H}$  est idempotente).

### Erreur-type (*standard error*) d'une valeur prédite

De façon usuelle, on note  $h_i$  le  $i$ -ième terme de la diagonale de  $\mathbf{H}$  ( $i = 1, \dots, n$ ). On obtient ainsi  $\hat{Y}_i \sim \mathcal{N}((\mathbf{X}\beta)_i, \sigma^2 h_i)$ . L'écart-type (*standard deviation*) de  $\hat{Y}_i$  est donc  $\sigma\sqrt{h_i}$  et on l'estime par  $\hat{\sigma}\sqrt{h_i}$  (voir le point suivant pour l'expression de  $\hat{\sigma}^2$ , donc de  $\hat{\sigma}$ ). La quantité  $\hat{\sigma}\sqrt{h_i}$  est appelée erreur-type de  $\hat{Y}_i$  et sera utilisée par la suite.

### Résidus

On appelle résidu le vecteur  $\hat{u}$  de  $\mathbb{R}^n$  défini par  $\hat{u} = y - \hat{y}$ . C'est l'écart entre l'observation du vecteur aléatoire  $Y$  et sa prédiction (son approximation) par le modèle considéré. Autrement dit, c'est une approximation du vecteur des erreurs  $U$ .

On obtient ainsi

$$\hat{U} = Y - \hat{Y} = (\mathbf{I}_n - \mathbf{H})Y = \mathbf{H}^\perp Y,$$

où  $\mathbf{H}^\perp$  est le projecteur orthogonal sur le sous-espace vectoriel  $F_X^\perp$  de  $\mathbb{R}^n$  supplémentaire orthogonal à  $F_X$ .

Dans le modèle gaussien, on obtient :

$$\hat{U} = \mathbf{H}^\perp Y \sim \mathcal{N}_n(0, \sigma^2 \mathbf{H}^\perp).$$

### Indépendance de $\hat{U}$ avec $\hat{Y}$ et avec $\hat{B}$

On a :

$$\text{Cov}(\hat{U}, \hat{Y}) = \text{Cov}(\mathbf{H}^\perp Y, \mathbf{H}Y) = \sigma^2 \mathbf{H}^\perp \mathbf{H} = 0.$$

Par suite,  $\hat{Y}$  et  $\hat{U}$  sont non corrélés, donc indépendants dans le cas gaussien. Il en est de même pour  $\hat{U}$  et  $\hat{B}$ .

### Résidus studentisés

Dans le cas gaussien, pour tout  $i$  ( $i = 1, \dots, n$ ), on a  $\hat{U}_i \sim \mathcal{N}(0, \sigma^2(1 - h_i))$ . L'écart-type de  $\hat{U}_i$  est donc  $\sigma\sqrt{1 - h_i}$  et son estimation, appelée erreur-type de  $\hat{U}_i$ , est  $\hat{\sigma}\sqrt{1 - h_i}$ .

On appelle alors  $i$ -ième résidu studentisé la quantité  $\hat{s}_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_i}}$ . Il s'agit de l'approximation de l'observation d'une loi  $\mathcal{N}(0, 1)$ , utilisée dans la validation du modèle.

**Remarque 4** On notera que si la construction de  $\hat{s}_i$  rappelle celle d'une observation de loi de Student, ce n'est pas ici le cas puisqu'il n'y a pas indépendance entre  $\hat{U}_i$  et  $\hat{\Sigma}^2 = \frac{\sum_{i=1}^n \hat{U}_i^2}{n - p}$  (voir l'expression de  $\hat{\Sigma}^2$  ci-dessous). Pour cette raison, on trouve dans la littérature statistique d'autres expressions pour les résidus studentisés; nous ne les introduisons pas ici car elles nous semblent peu utiles.

### 2.2.6 Estimation de $\sigma^2$ dans le cas général

Sans hypothèse gaussienne, on ne peut envisager d'utiliser le maximum de vraisemblance. Par ailleurs, les moindres carrés ne permettent pas d'estimer  $\sigma^2$ , dans la mesure où ce paramètre n'est pas lié à l'espérance de  $Y$ . On doit donc avoir recours à une estimation empirique (souvent appelée *plug-in*) : le paramètre  $\sigma^2$  représentant la variance de la variable erreur  $U$ , on l'estime par la variance empirique des résidus  $\hat{U}_i$ , soit  $\Sigma^{*2} = \frac{1}{n} \sum_{i=1}^n \hat{U}_i^2$  (la moyenne empirique des  $\hat{U}_i$  est nulle).

On peut alors vérifier que cet estimateur est biaisé et le corriger en posant  $\hat{\Sigma}^2 = \frac{1}{n - p} \sum_{i=1}^n \hat{U}_i^2$ , estimateur sans biais de  $\sigma^2$ . On ne peut toutefois rien dire ni sur sa variance ni sur sa convergence.

### 2.2.7 Estimation de $\sigma^2$ dans le cas gaussien

Dans ce cas, on applique la méthode du maximum de vraisemblance qui consiste à maximiser, selon  $\sigma^2$ , l'expression de  $l(y, \beta, \sigma^2)$  donnée en 2.2.3. On peut vérifier que cela conduit à la même expression  $\hat{\Sigma}^2$  que celle fournie par la méthode empirique. On utilise donc encore l'estimateur corrigé  $\hat{\Sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n \hat{U}_i^2$ , de façon à disposer d'un estimateur sans biais.

De plus, l'hypothèse gaussienne permet maintenant de montrer :

$$\frac{(n-p)\hat{\Sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n \hat{U}_i^2}{\sigma^2} = \frac{\|\hat{U}\|^2}{\sigma^2} \sim \chi_{n-p}^2.$$

On déduit de ce résultat :

- $\mathbb{E}(\hat{\Sigma}^2) = \sigma^2$  (résultat déjà connu) ;
- $\text{Var}(\hat{\Sigma}^2) = \frac{2\sigma^4}{n-p}$  :  $\hat{\Sigma}^2$  est donc un estimateur convergent ;
- par ailleurs, on peut vérifier que  $\hat{\Sigma}^2$  n'est pas efficace, mais est asymptotiquement efficace ; de plus, il s'agit d'un estimateur optimal pour  $\sigma^2$ , c'est-à-dire de variance minimum parmi les estimateurs sans biais (propriété générale de la famille exponentielle) ;
- enfin, dans le cas gaussien, on peut vérifier que les estimateurs  $\hat{B}$  et  $\hat{\Sigma}^2$  sont indépendants.

### 2.2.8 Intervalle de confiance pour une fonction linéaire de $\beta$

On ne peut envisager un tel intervalle que dans le cadre du modèle gaussien. Soit donc  $c$  un vecteur non nul de  $\mathbb{R}^p$  et  $c'\beta$  la forme linéaire associée. On a vu en 2.2.4 :

$$c'\hat{B} \sim \mathcal{N}(c'\beta, \sigma^2 c'(\mathbf{X}'\mathbf{X})^{-1}c).$$

La variance ci-dessus faisant intervenir le paramètre inconnu  $\sigma^2$ , on utilise  $\hat{\Sigma}^2$  et l'indépendance de  $c'\hat{B}$  et de  $\hat{\Sigma}^2$  pour obtenir une loi de Student, dont on déduit l'intervalle de confiance suivant, de coefficient de sécurité  $1 - \alpha$  :

$$c'\hat{\beta} \pm \hat{\sigma}[c'(\mathbf{X}'\mathbf{X})^{-1}c]^{1/2} t_{n-p}(1 - \frac{\alpha}{2}).$$

Dans l'expression ci-dessus, on notera que :

- $c'\hat{\beta}$  est l'estimation ponctuelle de  $c'\beta$  ;
- $\hat{\sigma}[c'(\mathbf{X}'\mathbf{X})^{-1}c]^{1/2}$  est l'erreur-type de  $c'\hat{\beta}$  ;
- $t_{n-p}(1 - \frac{\alpha}{2})$  est le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi de Student à  $n-p$  degrés de liberté (d.d.l.).

**Remarque 5** On peut tester l'hypothèse nulle  $\{H_0 : c'\beta = 0\}$  à partir de l'intervalle de confiance défini ci-dessus. Il suffit de regarder si l'intervalle contient, ou non, la valeur 0. En fait, cette démarche est équivalente au test de Student de cette hypothèse nulle (voir la remarque 8).

### 2.2.9 Intervalles de confiance conjoints : méthode de Bonferroni

En considérant  $c' = (0, \dots, 0, 1, 0, \dots, 0)$ , où le 1 est situé en  $j$ -ième position ( $j = 1, \dots, p$ ), on obtient, par la méthode ci-dessus, un intervalle de confiance de risque  $\alpha$  (c'est-à-dire de coefficient de sécurité  $1 - \alpha$ ) pour le paramètre  $\beta_j$ .

Pour construire simultanément des intervalles de confiance pour les  $p$  paramètres  $\beta_j$ , de risque inconnu mais majoré par  $\alpha$  ( $\alpha \in ]0, 1[$ ), on peut utiliser la méthode de Bonferroni. Elle consiste à construire un intervalle, pour chacun des paramètres  $\beta_j$ , selon la formule indiquée ci-dessus, en utilisant pour risque non pas  $\alpha$  mais  $\frac{\alpha}{p}$ . Toutefois, il faut noter que, dès que  $p$  vaut 5 ou plus, cette méthode est trop conservatrice : elle a tendance à ne pas rejeter l'hypothèse nulle d'égalité des paramètres  $\beta_j$ , autrement dit à regrouper la plupart des niveaux du facteur.

Nous donnons quelques développements de cette méthode dans l'Annexe A.

## 2.3 Test d'une hypothèse linéaire en $\beta$

Dans le modèle linéaire, on est souvent amené à tester une hypothèse nulle, linéaire en  $\beta$ , du type  $\{H_0 : \mathbf{C}'\beta = 0\}$ , où  $\mathbf{C}$  est une matrice  $p \times q$  de rang  $q$ , ( $1 \leq q < p$ ), ce qui revient à tester la réalité de  $q$  contraintes linéaires sur le paramètre  $\beta$  (par exemple,  $\beta_1 = 0$ ,  $\beta_2 = \beta_3$ , etc.). Le but est, en fait, de simplifier le modèle. On notera que cela revient à tester  $\{H_0 : \beta \in E_0\}$ , où  $E_0$  est un sous-espace vectoriel de  $\mathbb{R}^p$  de dimension  $p - q$ , ou encore  $\mathbb{E}(Y) = \mathbf{X}\beta \in F_0$ , où  $F_0$  est un sous-espace vectoriel de  $\mathbb{R}^n$  de dimension  $p - q$ .

On a vu :

$$\frac{(n-p)\hat{\Sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n \hat{U}_i^2}{\sigma^2} = \frac{\|\hat{U}\|^2}{\sigma^2} \sim \chi_{n-p}^2.$$

De la même manière, si  $H_0$  est vraie, on peut vérifier que

$$\frac{\|\hat{U}_0\|^2 - \|\hat{U}\|^2}{\sigma^2} \sim \chi_q^2,$$

avec  $\|\hat{U}_0\|^2 = \sum_{i=1}^n \hat{U}_{i0}^2$ ,  $\hat{U}_{i0} = Y_i - \hat{Y}_{i0}$  et  $\hat{Y}_{i0} = \mathbf{X}\hat{B}_0$ ,  $\hat{B}_0$  étant l'estimateur maximum de vraisemblance de  $\beta$  sous la contrainte  $\mathbf{C}'\beta = 0$ . De plus, sous  $H_0$ , les deux statistiques de khi-deux définies ci-dessus sont indépendantes.

On en déduit le test de  $H_0$  : rejet de  $H_0$  ssi (si, et seulement si)

$$F = \frac{\|\hat{U}_0\|^2 - \|\hat{U}\|^2}{\|\hat{U}\|^2} \times \frac{n-p}{q} > f_{q; n-p}(1-\alpha),$$

où  $f_{q; n-p}(1-\alpha)$  est le quantile d'ordre  $1-\alpha$  d'une loi de Fisher à  $q$  et  $n-p$  d.d.l. Ce test est de niveau  $\alpha$ .

### Autres expressions de $F$

On peut écrire la statistique  $F$  sous la forme  $\frac{NUM}{q \hat{\Sigma}^2}$ , puisque  $\hat{\Sigma}^2 = \frac{\|\hat{U}\|^2}{n-p}$ ; le numérateur peut alors prendre les expressions suivantes :

$$NUM = \|\hat{U}_0\|^2 - \|\hat{U}\|^2 = \|\hat{U}_0 - \hat{U}\|^2 = \|\hat{Y}_0 - \hat{Y}\|^2 = \|\hat{B}_0 - \hat{B}\|_{\mathbf{X}'\mathbf{X}}^2 = \hat{B}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\hat{B}.$$

La quantité  $\|\hat{U}\|^2$  correspond à ce qui est souvent appelé, dans les logiciels, *error sum of squares* (dans le modèle complet).

**Remarque 6** *Ce test est en général appelé test de Fisher, parfois test de Fisher-Snedecor, voire test de Snedecor.*

**Remarque 7** *Dans la pratique, les logiciels calculent la valeur observée  $f$  de la statistique  $F$  (sur les données considérées), puis la probabilité  $P[F_{q; n-p} > f]$  ( $F_{q; n-p}$  désigne une loi de Fisher à  $q$  et  $n-p$  d.d.l.), en général appelée  $p$ -value. On rejette alors  $H_0$  ssi la  $p$ -value est inférieure à  $\alpha$ .*

**Remarque 8** *Si  $q = 1$ , le test de Fisher ci-dessus peut se ramener à un test de Student, lui-même équivalent à l'intervalle de confiance construit en 2.2.8.*

### Critère de choix de modèle : le $C_p$ de Mallows

Lorsqu'on hésite à prendre en compte un effet faiblement significatif (dont la  $p$ -value est proche de  $\alpha$ ), on peut utiliser le critère  $C_p$  (voir l'Annexe D) pour décider : on calcule ce critère pour chacun des deux modèles (avec et sans cet effet) et on retient celui des deux qui minimise le  $C_p$ .

## 2.4 Contrôles d'un modèle linéaire

À l'issue de différents traitements statistiques (études exploratoires élémentaires, puis multidimensionnelles, modélisations avec tests et estimations des paramètres...), lorsqu'un modèle linéaire semble convenir à un jeu de données, un certain nombre de contrôles sont nécessaires avant de le retenir effectivement. Ces contrôles ont pour but d'apprécier la *qualité* et la *validité* du modèle envisagé. Ils peuvent, bien sûr, conduire à en changer.

### 2.4.1 Contrôles de la qualité d'un modèle

- *Significativité.* Le test de significativité du modèle est le test de l'hypothèse nulle correspondant au modèle constant (ou modèle blanc) au sein du modèle retenu (autrement dit, à la nullité de tous les paramètres  $\beta_j$ , à l'exception de celui correspondant au vecteur constant). Ce test doit être très significatif (c'est la condition minimale).
- *Valeur du  $R^2$ .* Le coefficient  $R^2 = \frac{\|\hat{Y}\|^2}{\|Y\|^2}$ , compris entre 0 et 1, mesure la qualité globale du modèle et doit être suffisamment proche de 1.
- *Graphique des valeurs prédites contre les valeurs observées.* En axes orthonormés, on représente le nuage des points ayant pour abscisses les valeurs observées ( $y_i$ ) et pour ordonnées les valeurs prédites par le modèle ( $\hat{y}_i$ ). Plus le nuage obtenu est proche de la première bissectrice, plus le modèle est globalement bon. On peut également faire figurer la première bissectrice sur ce graphique pour préciser les choses. Ce graphique fournit, d'une autre manière, une information analogue à celle fournie par le coefficient  $R^2$ . Mais, il permet aussi de contrôler que la forme générale du nuage (donc l'ensemble des observations de  $Y$ ) n'a rien de particulier. On en trouvera des exemples au chapitre 3 (Figures 3.1 et 3.3).

### 2.4.2 Contrôles de la validité d'un modèle

Ces contrôles se font à partir de ce qu'il est convenu d'appeler le **graphique des résidus**. C'est le graphique donnant le nuage des points ayant pour abscisses les valeurs prédites ( $\hat{y}_i$ ) et pour ordonnées les résidus studentisés ( $\hat{s}_i$ ), et dont on trouvera aussi des exemples au chapitre 3 (Figures 3.2 et 3.4).

Trois éléments sont contrôlés à travers ce graphique.

- *Le caractère linéaire des données.* Les données ayant été ajustées par un modèle linéaire, si leur structure est réellement linéaire, on ne doit retrouver aucune structure dans les résidus. Si on retrouve une forme en "U", on pourra essayer de remplacer  $Y$  par  $\log(Y)$  ou par  $\sqrt{Y}$  (à condition que  $Y$  soit à valeurs positives); pour une forme en "U renversé", on pourra essayer de remplacer  $Y$  par  $\exp(Y)$  ou par  $Y^2$ ; etc.
- *L'homoscédasticité.* La variance de la variable erreur  $U$  étant supposée constante d'une observation à l'autre, la variabilité des résidus studentisés doit être de même amplitude quelles que soient les valeurs  $\hat{y}_i$ , ce que l'on peut contrôler sur le graphique des résidus. Là encore, en cas de croissance des résidus en fonction des valeurs  $\hat{y}_i$ , on peut envisager la transformation de  $Y$  en  $\log(Y)$  ou en  $\sqrt{Y}$  (toujours sous la même condition).
- *La normalité.* Enfin, si les données sont réellement gaussiennes, les résidus studentisés sont approximativement distribués selon une loi normale réduite, et pas plus de 5% d'entre eux ne doivent sortir de l'intervalle  $[-2, +2]$ , ce qui est très facile à contrôler sur le graphique.

Il est donc conseillé de n'utiliser un modèle linéaire que s'il a passé avec succès l'ensemble des contrôles de qualité et de validité indiqués ci-dessus.

## 2.5 Panorama sur le modèle linéaire

### 2.5.1 Le modèle linéaire gaussien de base

Il s'agit du modèle développé dans les paragraphes précédents.

Précisons que si tous les prédictors  $X^j$  sont quantitatifs, on obtient ce que l'on appelle la *régression linéaire*. Celle-ci ne sera pas développée dans ce cours et nous renvoyons pour cela aux enseignements de première année de Master ou à la bibliographie mentionnée en début de chapitre.

Lorsque tous les prédictors sont qualitatifs, on parle alors de facteurs et le modèle linéaire recouvre ce que l'on appelle l'*analyse de variance*, ou ANOVA (acronyme anglais de *ANalysis Of VAriance*), ou encore les *plans factoriels*. Les cas les plus simples seront traités au chapitre 3, tandis que des cas plus particuliers seront abordés au chapitre 4.

Enfin, lorsqu'il y a mélange de prédictors quantitatifs et qualitatifs, on parle d'*analyse de covariance*, pour laquelle nous renvoyons encore aux enseignements de première année de Master ou à la bibliographie.

### 2.5.2 Le modèle linéaire gaussien général

C'est l'objet principal de ce cours. Il s'agit de diverses généralisations du modèle linéaire gaussien de base.

- Lorsque la variable réponse  $Y$  est multidimensionnelle, on obtient le modèle linéaire multivarié. Dans le chapitre 5, on s'intéressera au cas de prédicteurs  $X^j$  qualitatifs, ce qui nous donnera l'analyse de variance multivariée, ou MANOVA.
- Avec une variable réponse  $Y$  unidimensionnelle, on peut introduire, parmi les prédicteurs  $X^j$ , des variables aléatoires (et plus seulement des prédicteurs contrôlés). On définit ainsi les modèles à effets aléatoires et les modèles mixtes que nous traiterons au chapitre 6.
- On peut enfin considérer, pour chaque individu  $i$  pris en compte, des observations de  $Y_i$  répétées dans le temps. Ces observations sont naturellement corrélées, ce qui nécessite l'introduction de modèles spécifiques : les modèles pour données répétées, étudiés au chapitre 7.

### 2.5.3 Le modèle linéaire généralisé

Il s'agit d'une extension du modèle linéaire qui ne sera pas abordée dans ce cours. Pour mémoire, indiquons qu'il s'agit toujours d'expliquer une variable  $Y$  au moyen de prédicteurs  $X^j$ , en utilisant un échantillon de taille  $n$ , mais qu'il y a généralisation à deux niveaux :

- chaque v.a.r.  $Y_i$  de l'échantillon est distribuée selon une même loi de la *famille exponentielle* (normale, binomiale, Poisson, gamma...);
- la relation linéaire entre  $\mathbb{E}(Y_i)$  et les prédicteurs  $X^j$  se fait au moyen d'une fonction particulière  $g$ , monotone et dérivable, appelée *fonction lien*, de la façon suivante :

$$g[\mathbb{E}(Y_i)] = \sum_{j=1}^p \beta_j X^j.$$

#### Exemples

- Si l'on prend la loi normale comme loi de la famille exponentielle et la fonction identité comme fonction lien, on retrouve le modèle linéaire gaussien de base : le modèle linéaire généralisé en constitue donc bien une généralisation.
- Si l'on suppose maintenant  $Y_i \sim \mathcal{B}(n_i, p_i)$ , qu'on modélise  $\frac{Y_i}{n_i}$  et qu'on choisit la fonction *logit* comme fonction lien ( $g(x) = \log\left(\frac{x}{1-x}\right)$ ,  $x \in ]0, 1[$ ), on obtient la régression logistique :

$$\mathbb{E}\left(\frac{Y_i}{n_i}\right) = p_i ; g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^p \beta_j x_i^j.$$



## Chapitre 3

# L'analyse de variance univariée

*Le chapitre 3 est consacré aux plans factoriels. Il s'agit de l'appellation appropriée, bien qu'assez peu employée, de l'analyse de variance, appelée par les anglo-saxons "ANalysis Of VAriance" et, pour cette raison, bien connue sous l'acronyme d'ANOVA.*

*L'ANOVA correspond à un modèle linéaire gaussien dans lequel toutes les variables explicatives (les  $X^j$ ) sont qualitatives. Dans ce contexte, elles sont appelées **facteurs** (d'où le terme de plans factoriels) et leurs modalités sont appelées **niveaux**. Ces niveaux sont supposés choisis, fixés, par l'utilisateur, de sorte que l'on parle souvent de **facteurs contrôlés**. De son côté, la variable aléatoire réponse  $Y$  est toujours quantitative et supposée gaussienne.*

*Seuls seront traités dans ce chapitre les cas de l'analyse de variance à un facteur, à deux facteurs croisés et à trois facteurs croisés. Dans un dernier paragraphe, nous donnerons quelques indications sur les cas plus généraux dont certains seront étudiés au chapitre 4.*

*Les références bibliographiques du chapitre 3 sont les mêmes que celles du chapitre 2.*

### Résumé

Les problèmes abordés dans chacun des paragraphes de ce chapitre seront, à chaque fois, les trois problèmes clés du modèle linéaire gaussien : estimation ponctuelle, estimation par intervalle de confiance et tests. Ils seront traités dans cet ordre, en particulier parce qu'on a besoin de certaines estimations ponctuelles pour construire un intervalle de confiance et pour faire un test. Mais, dans la pratique, on commence en général par faire différents tests pour choisir le modèle le plus adapté aux données considérées, puis on détermine les estimations des paramètres dans le modèle ainsi choisi.

Les paramètres que l'on va utiliser en ANOVA vont représenter des effets particuliers du modèle pris en compte : effet général et effets principaux des niveaux du facteur dans un plan à un seul facteur ; effet général, effets principaux des niveaux de chaque facteur et effets d'interactions dans un plan à deux facteurs... Ces différents effets ne peuvent être pris en compte si on conserve le paramétrage standard du modèle linéaire (par exemple, dans un modèle à deux facteurs,  $Y_{ijk} = \beta_{jk} + U_{ijk}$ ). D'où la nécessité d'utiliser d'autres paramétrages. Il en existe plusieurs et nous en présentons deux dans ce chapitre : le paramétrage dit centré, car il fait intervenir des paramètres centrés, et le paramétrage SAS, utilisé systématiquement dans le logiciel SAS.

Ainsi, pour un plan à deux facteurs croisés, le paramétrage centré consiste à poser :  $\beta_{jk} = \mu + \alpha_j^1 + \alpha_k^2 + \gamma_{jk}$ . Le paramètre  $\mu$  représente l'effet général, les paramètres  $\alpha_j^1$  et  $\alpha_k^2$  les effets principaux des deux facteurs et les paramètres  $\gamma_{jk}$  les effets d'interactions. Les  $\alpha_j^1$  sont centrés selon  $j$ , les  $\alpha_k^2$  selon  $k$  et les  $\gamma_{jk}$  selon  $j$  et selon  $k$ .

Le paramétrage SAS, tel qu'on le trouve en particulier dans la procédure GLM, consiste, de son côté, à réécrire :  $\beta_{jk} = m + a_j^1 + a_k^2 + c_{jk}$ . Les paramètres  $m$ ,  $a_j^1$ ,  $a_k^2$  et  $c_{jk}$  représentent les mêmes notions que celles précisées ci-dessus, mais ils sont définis en se "callant" sur la dernière cellule, d'indice  $(J, K)$ .

### 3.1 Cas d'un seul facteur

Lorsque nécessaire, le facteur considéré sera noté  $F$ ; cette notation est certes la même que celle de la statistique du test de Fisher, mais, dans le contexte, il ne devrait pas y avoir de confusion; de plus, la notation du facteur sera peu utilisée. Par ailleurs, le nombre des niveaux de  $F$  sera noté  $J$  ( $J \geq 2$ ) et l'indice du niveau courant noté  $j$  ( $j = 1, \dots, J$ ).

Pour chaque niveau  $j$ , on réalise  $n_j$  observations indépendantes de la v.a.r. (quantitative) à expliquer  $Y$  ( $n_j \geq 1$ ), notées  $y_{ij}$ ,  $i = 1, \dots, n_j$ ; on pose enfin  $n = \sum_{j=1}^J n_j$ :  $n$  est le nombre total d'observations réalisées dans l'expérience.

Si  $n_j = n_0, \forall j, j = 1, \dots, J$ , on dit que le plan est **équilibré**; sinon, on parle de plan **déséquilibré**. Dans un plan équilibré,  $n_0$  s'appelle le nombre de **répétitions**.

**Remarque 9** *On a utilisé ci-dessus le terme de plan. C'est le terme utilisé dans tout le contexte de l'ANOVA, où l'on parle de plan d'expériences<sup>1</sup> ou de plan factoriel, voire, tout simplement, de plan. En fait, ce terme est d'origine industrielle et, dans un tel environnement, on parle également d'expérience planifiée, ce qui sous-entend, d'ailleurs, que les niveaux du (ou des) facteurs pris en compte sont totalement contrôlés (d'où le terme de facteur contrôlé).*

#### 3.1.1 Écriture initiale du modèle

On commence par écrire le modèle sous la forme :

$$Y_{ij} = \beta_j + U_{ij}.$$

- $\beta_j$  est le paramètre associé au niveau  $j$  du facteur  $F$ ; il est inconnu, à estimer; ce paramètre représente un effet non aléatoire, encore appelé **effet fixe**.
- $U_{ij}$  est la v.a.r. erreur associée à l'observation numéro  $i$  du niveau  $j$  de  $F$ ; on suppose  $U_{ij} \sim \mathcal{N}(0, \sigma^2)$ ,  $\sigma^2$  étant aussi un paramètre à estimer (il ne dépend pas de  $j$ , autrement dit le modèle est homoscédastique); par ailleurs, les v.a.r.  $U_{ij}$  sont supposées indépendantes (elles sont donc i.i.d.).
- $Y_{ij}$  est la v.a.r. réponse associée à l'observation numéro  $i$  du niveau  $j$  de  $F$ ; on obtient donc  $Y_{ij} \sim \mathcal{N}(\beta_j, \sigma^2)$ , les  $Y_{ij}$  étant indépendantes.

On peut réécrire le modèle sous la forme matricielle

$$Y = \mathbf{X}\beta + U,$$

où  $Y$  et  $U$  sont des vecteurs de  $\mathbb{R}^n$ ,  $\beta$  est un vecteur de  $\mathbb{R}^J$  (ici,  $p = J$ ) et  $\mathbf{X}$ , appelée **matrice d'incidence**, est une matrice  $n \times J$  ne comportant que des 0 et des 1; en fait, chaque colonne de  $\mathbf{X}$  est l'indicatrice du niveau correspondant de  $F$  et nous noterons  $Z^j$  l'indicatrice courante. On peut ainsi réécrire :

$$Y = \sum_{j=1}^J \beta_j Z^j + U.$$

**Exemple 1** *Considérons le cas  $J = 3$ ,  $n_1 = 2$ ,  $n_2 = 3$ ,  $n_3 = 1$  ( $n = 6$ ). Il vient :*

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

**Remarque 10** *Sous la dernière forme donnée ci-dessus, on voit que le modèle est équivalent à un modèle de régression multiple, sans coefficient constant, dont les régresseurs sont les  $J$  variables indicatrices  $Z^j$ .*

**Remarque 11** *On vérifie que les colonnes de  $\mathbf{X}$  sont deux à deux orthogonales. On en déduit que  $\mathbf{X}'\mathbf{X} = \text{diag}(n_1 \dots n_J)$  : il s'agit d'une matrice régulière.*

<sup>1</sup>Dans l'expression plan d'expériences, on trouve le terme d'expérience tantôt au singulier et tantôt au pluriel; nous préférons utiliser le pluriel, d'une part parce que le même plan peut servir à plusieurs expériences, d'autre part parce que le *petit Robert* cite l'expression "Laboratoire d'expériences".

### 3.1.2 Paramétrage centré

Le paramétrage initial ne permet pas de dissocier d'une part les effets des différents niveaux du facteur  $F$ , d'autre part l'effet général (et les choses seront encore plus problématiques en présence de deux facteurs ou plus). D'où la nécessité de réécrire le modèle, le problème étant qu'il existe plusieurs réécritures distinctes (mais, bien sûr, équivalentes).

Dans le paramétrage centré, on pose :

$$\mu = \frac{1}{J} \sum_{j=1}^J \beta_j \text{ (moyenne "non pondérée" des } \beta_j \text{)} ; \alpha_j = \beta_j - \mu.$$

On obtient ainsi  $\beta_j = \mu + \alpha_j$  et on réécrit le modèle sous la forme :

$$Y_{ij} = \mu + \alpha_j + U_{ij}.$$

On notera la relation  $\sum_{j=1}^J \alpha_j = 0$ .

- Le paramètre  $\mu$  est appelé l'effet général, ou encore l'effet moyen général.
- Les paramètres  $\alpha_j$  ( $j = 1, \dots, J$ ) sont appelés les effets principaux du facteur  $F$ , ou encore les effets différentiels. La littérature statistique anglo-saxonne parle fréquemment de *contrastes*, dans la mesure où il s'agit de paramètres de somme nulle.
- Dans  $\mathbb{R}^n$ , on peut réécrire le modèle sous la forme suivante :

$$\begin{aligned} Y &= \sum_{j=1}^J \beta_j Z^j + U = \mu \mathbb{1}_n + \sum_{j=1}^J \alpha_j Z^j + U = \mu \mathbb{1}_n + \sum_{j=1}^{J-1} \alpha_j Z^j - Z^J \sum_{j=1}^{J-1} \alpha_j + U \\ &= \mu \mathbb{1}_n + \sum_{j=1}^{J-1} \alpha_j (Z^j - Z^J) + U. \end{aligned}$$

On obtient maintenant un modèle de régression linéaire sur les  $J - 1$  variables  $Z^j - Z^J$ , avec coefficient constant.

#### Notation

On notera  $\beta_c$  le vecteur des  $J$  paramètres dans ce paramétrage ( $\mu$  et les  $\alpha_j$ ,  $j = 1, \dots, J - 1$ ) et  $\mathbf{X}_c$  la matrice d'incidence correspondante, de sorte qu'on pourra réécrire  $Y = \mathbf{X}_c \beta_c + U$ .

**Exemple 2** Dans l'exemple introduit plus haut,  $\mathbf{X}_c$  et  $\beta_c$  ont pour expression :

$$\mathbf{X}_c = (\mathbb{1}_n \quad (Z^1 - Z^3) \quad (Z^2 - Z^3)) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix} ; \beta_c = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

La matrice  $\mathbf{X}_c$  est toujours de rang 3, mais ses colonnes ne sont plus orthogonales. Toutefois, elles le seraient dans un plan équilibré.

### 3.1.3 Paramétrage SAS

Le principe de ce paramétrage est de se "caller" sur le dernier niveau  $J$  du facteur  $F$ . On pose ainsi

$$Y_{ij} = m + a_j + U_{ij},$$

avec  $m = \beta_J$  et  $a_j = \beta_j - \beta_J$ ,  $\forall j = 1, \dots, J$  (de sorte que  $a_J = 0$ ). On peut alors réécrire :

$$Y = \sum_{j=1}^J \beta_j Z^j + U = \beta_J \mathbb{1}_n + \sum_{j=1}^J a_j Z^j + U = m \mathbb{1}_n + \sum_{j=1}^{J-1} a_j Z^j + U.$$

On voit qu'il s'agit d'un modèle de régression sur les  $J - 1$  indicatrices  $Z^j$  ( $j = 1, \dots, J - 1$ ), avec coefficient constant. Pour cette raison, le paramètre  $m$  est appelé *intercept* dans SAS, comme le coefficient constant d'une régression.

**Notation**

On notera maintenant  $\beta_s$  le vecteur des  $J$  paramètres de ce paramétrage ( $m$  et les  $a_j$ ,  $j = 1, \dots, J-1$ ) et  $\mathbf{X}_s$  la matrice d'incidence correspondante, de sorte qu'on pourra réécrire  $Y = \mathbf{X}_s \beta_s + U$ .

**Exemple 3** En considérant toujours le même exemple,  $\mathbf{X}_s$  et  $\beta_s$  ont pour expression :

$$\mathbf{X}_s = (\mathbb{I}_n \ Z^1 \ Z^2) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}; \quad \beta_s = \begin{pmatrix} m \\ a_1 \\ a_2 \end{pmatrix}.$$

La matrice  $\mathbf{X}_s$  est encore de rang 3, ses colonnes n'étant pas non plus orthogonales. On notera qu'elles ne le seraient pas davantage dans le cas d'un plan équilibré.

**3.1.4 Estimation des paramètres**

En appliquant les résultats généraux relatifs à l'estimation dans le modèle linéaire gaussien, on obtient les résultats indiqués ci-dessous.

**Vecteur des paramètres dans le paramétrage initial.**

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y, \text{ avec } \mathbf{X}'\mathbf{X} = \text{diag}(n_1 \cdots n_J) \text{ et } \mathbf{X}'y = \begin{pmatrix} n_1 \bar{y}_{\bullet 1} \\ \vdots \\ n_J \bar{y}_{\bullet J} \end{pmatrix}, \text{ où } \bar{y}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}. \text{ On}$$

obtient ainsi  $\hat{\beta}_j = \bar{y}_{\bullet j}$ . De plus  $\hat{B} \sim \mathcal{N}_J(\beta, \sigma^2 \text{diag}(\frac{1}{n_1} \cdots \frac{1}{n_J}))$ , de sorte que les composantes  $\hat{B}_j$  sont indépendantes.

**Paramétrage centré.**

Il vient :  $\hat{\mu} = \frac{1}{J} \sum_{j=1}^J \bar{y}_{\bullet j}$  (*attention* : si les effectifs  $n_j$  ne sont pas tous égaux, autrement dit si le plan est déséquilibré,  $\hat{\mu}$  n'est pas la moyenne générale des observations de  $Y$ , notée  $\bar{y}_{\bullet\bullet}$ ). D'autre part,  $\hat{\alpha}_j = \bar{y}_{\bullet j} - \hat{\mu}$ , de sorte qu'on retrouve  $\sum_{j=1}^J \hat{\alpha}_j = 0$ .

**Paramétrage SAS.**

On obtient maintenant  $\hat{m} = \bar{y}_{\bullet J}$  et  $\hat{a}_j = \bar{y}_{\bullet j} - \bar{y}_{\bullet J}$ , de sorte qu'on vérifie bien  $\hat{a}_J = 0$ .

**Valeurs prédites.**

Elles sont définies par  $\hat{y}_{ij} = \hat{\beta}_j = \bar{y}_{\bullet j}$ . Elles ne dépendent pas du paramétrage considéré.

**Résidus.**

On obtient  $\hat{u}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{\bullet j}$  (même remarque que ci-dessus).

**Variance.**

Comme pour les valeurs prédites et les résidus, l'estimation de la variance ne dépend pas du paramétrage choisi. Il vient :

$$\hat{\sigma}^2 = \frac{1}{n-J} \sum_{j=1}^J \sum_{i=1}^{n_j} (\hat{u}_{ij})^2 = \frac{1}{n-J} \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2.$$

**Intervalle de confiance de  $\beta_j$ , de coefficient de sécurité  $1 - \alpha$ .**

On obtient l'intervalle de type Student suivant :

$$\bar{y}_{\bullet j} \pm \frac{\hat{\sigma}}{\sqrt{n_j}} t_{n-J}(1 - \frac{\alpha}{2}),$$

où  $t_{n-J}(1 - \frac{\alpha}{2})$  désigne le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi de Student à  $n - J$  d.d.l. On notera que  $\frac{\hat{\sigma}}{\sqrt{n_j}}$  est l'erreur-type de  $\hat{B}_j$ .

**Erreurs-types**

Un calcul simple permet de vérifier que l'erreur-type de  $\hat{Y}_{ij}$  est encore  $\frac{\hat{\sigma}}{\sqrt{n_j}}$ , tandis que celle de  $\hat{U}_{ij}$  est  $\hat{\sigma} \sqrt{\frac{n_j - 1}{n_j}}$ . On notera que ces erreurs-types sont constantes dans le cas équilibré.

**3.1.5 Test de l'effet du facteur  $F$** 

Tester la significativité du facteur  $F$  revient à tester la significativité du modèle envisagé. Dans le paramétrage initial du modèle, l'hypothèse nulle se met sous la forme  $\{H_0 : \beta_1 = \dots = \beta_J\}$ , l'alternative étant le contraire de  $H_0$ .

Dans les autres paramétrages,  $H_0$  est équivalente aux contraintes suivantes :

- dans le paramétrage centré,  $\alpha_1 = \dots = \alpha_J = 0$  ;
- dans le paramétrage SAS,  $a_1 = \dots = a_J = 0$ .

Dans tous les cas, le nombre de contraintes indépendantes,  $q$ , imposées par  $H_0$  est égal à  $J - 1$ .

Sous  $H_0$ , le modèle s'écrit  $Y_{ij} = \mu + U_{ij}$  (il s'agit du modèle constant, ou modèle blanc), et l'on obtient :

$$\hat{\mu}^0 = \bar{y}_{\bullet\bullet} = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} y_{ij}$$

(attention : si le plan est déséquilibré,  $\hat{\mu}^0 \neq \hat{\mu}$ ).

On prend alors pour expression de la statistique du test de Fisher la quantité

$$f = \frac{\|\hat{y} - \hat{y}^0\|^2}{q\hat{\sigma}^2},$$

$\hat{y}$  et  $\hat{y}^0$  désignant, dans  $\mathbb{R}^n$ , les vecteurs des valeurs prédites respectivement dans le modèle considéré (modèle avec le seul facteur  $F$ ) et dans le modèle sous  $H_0$  (modèle constant). Il vient (voir le 3.1.4)

$$f = \frac{1}{(J-1)\hat{\sigma}^2} \sum_{j=1}^J n_j (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2, \text{ avec } \hat{\sigma}^2 = \frac{1}{n-J} \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2.$$

On compare enfin cette statistique avec  $f_{J-1; n-J}(1 - \alpha)$ , quantile d'ordre  $1 - \alpha$  d'une loi de Fisher à  $J - 1$  et  $n - J$  d.d.l.

**Synthèse des résultats précédents : le tableau d'analyse de la variance**

Il est fréquent de résumer la construction de la statistique du test de Fisher au sein d'un tableau, appelé tableau d'analyse de la variance, qui se présente sous la forme ci-dessous (on notera que la plupart des logiciels fournissent ce tableau).

sources de variation	sommes des carrés	d.d.l.	carrés moyens	valeur de la statistique de Fisher
Facteur $F$	$SSF$	$J - 1$	$MSF = \frac{SSF}{J - 1}$	$\frac{MSF}{MSE}$
Erreur	$SSE$	$n - J$	$MSE = \frac{SSE}{n - J} = \hat{\sigma}^2$	—
Total	$SST$	$n - 1$	—	—

Les sommes de carrés apparaissant ci-dessus sont définies de la façon suivante :

$$SSF = \sum_{j=1}^J n_j (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2;$$

on notera que dans le cas d'un plan équilibré (tous les  $n_j$  sont égaux à  $n_0$  et, par suite,  $\hat{\mu} = \bar{y}_{\bullet\bullet}$ ), on peut encore écrire :  $SSF = n_0 \sum_{j=1}^J (\hat{\alpha}_j)^2$ ;

$$SSE = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2;$$

$$SST = SSF + SSE = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet\bullet})^2.$$

### 3.1.6 Autres tests

Si l'hypothèse nulle considérée ci-dessus a été rejetée, autrement dit si le modèle à un facteur est significatif, on peut être amené à tester d'autres hypothèses nulles, dans le but de simplifier le modèle. Par exemple, on peut considérer des hypothèses nulles du type :  $\beta_j = \beta_{j'}$  ;  $\alpha_j = 0$  ;  $a_j = 0$ .

Pour cela, on utilise en général un test de Student (rappelons que, lorsque  $q = 1$ , le test de Fisher est équivalent à un test de Student). En particulier, le logiciel SAS propose, pour chaque valeur de  $j$  ( $j = 1, \dots, J - 1$ ), le test de Student de l'hypothèse nulle  $\{H_0 : \beta_j = \beta_J\}$  ; cela se fait avec la procédure GLM, au sein de la commande `model`, lorsqu'on rajoute l'option `solution`.

Avec des options spécifiques de la commande `model`, on peut aussi tester la significativité de l'écart entre deux niveaux quelconques du facteur  $F$ . Enfin, on peut utiliser la technique de Bonferroni (présentée en 2.2.9) pour construire des intervalles de confiance conjoints pour les écarts  $\beta_j - \beta_{j'}$ .

### 3.1.7 Illustration

#### Les données

Il s'agit d'un exemple fictif d'analyse de variance à un seul facteur. La variable réponse, en première colonne, prend des valeurs entières comprises entre 10 et 25. Le facteur est à trois niveaux, notés 1,2,3 et figure en seconde colonne. Il y a 9 individus observés, donc 9 lignes dans le fichier des données reproduit ci-après.

```
11 1
13 1
15 2
18 2
21 2
19 3
20 3
22 3
23 3
```

### Le programme SAS

Le programme SAS ci-dessous permet de faire les principaux traitements (graphiques compris) dans le cadre d'une ANOVA à un seul facteur. Les commentaires permettent de discerner les différentes phases du programme.

```

* ----- ;
* options facultatives pour la mise en page des sorties ;
* ----- ;
options pagesize=64 linesize=76 nodate;
title;
footnote 'ANOVA 1 facteur - Exemple fictif';
* ----- ;
*           lecture des donnees ;
* (le fichier "fic.don" contient les donnees ;
* et se trouve dans le repertoire de travail) ;
* ----- ;
data fic;
infile 'fic.don';
input y f;
run;
* ----- ;
*           procedure GLM pour l'ANOVA ;
* ----- ;
proc glm data=fic;
class f;
model y = f / ss3;
run;
quit;
* ----- ;
*           on relance avec l'option "solution" ;
* ----- ;
proc glm data=fic;
class f;
model y = f / ss3 solution;
run;
quit;
* ----- ;
*           on relance avec la commande "output" ;
*           pour archiver diverses quantites ;
* ----- ;
proc glm data=fic noprint;
class f;
model y = f;
output out=sortie p=yy r=uu stdr=erty student=rest;
proc print data=sortie;
run;
quit;
* ----- ;
*           graphique valeurs predites vs valeurs observees ;
* ----- ;
proc gplot data=sortie;
axis1 label=('valeurs observees')
      order=(10 to 25 by 5) length=7cm;
axis2 label=('valeurs' justify=right 'predites')
      order=(10 to 25 by 5) length=7cm;
symbol1 i=none v=dot;
symbol2 i=r1 v=none;
plot yy*y y*y / haxis=axis1 vaxis=axis2
      hminor=4 vminor=4
      overlay;
run;

```

```

goptions reset=all;
quit;
* ----- ;
*           graphique des residus           ;
* ----- ;
proc gplot data=sortie;
axis1 label=('valeurs predites')
      order=(10 to 25 by 5) length=7cm;
axis2 label=('resisus' justify=right 'studentises')
      order=(-3 to 3 by 1) length=7cm;
symbol v=dot;
plot rest*yy / haxis=axis1 vaxis=axis2
              hminor=4 vminor=0
              vref=-2 vref=0 vref=2;
run;
goptions reset=all;
quit;

```

### Les sorties de la procédure GLM

```

PAGE 1                               The GLM Procedure
-----
                                Class Level Information

Class          Levels    Values
f                3      1 2 3

                                Number of observations      9

```

```

PAGE 2                               The GLM Procedure
-----

Dependent Variable: y

Source          DF          Sum of Squares    Mean Square    F Value    Pr > F
Model           2          108.0000000      54.0000000     10.80     0.0103
Error           6           30.0000000       5.0000000
Corrected Total 8          138.0000000

                                R-Square    Coeff Var    Root MSE     y Mean
                                0.782609    12.42260    2.236068     18.00000

Source          DF          Type III SS    Mean Square    F Value    Pr > F
f                2          108.0000000     54.0000000     10.80     0.0103

```

PAGE 3

The GLM Procedure

-----

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		21.00000000 B	1.11803399	18.78	<.0001
f	1	-9.00000000 B	1.93649167	-4.65	0.0035
f	2	-3.00000000 B	1.70782513	-1.76	0.1295
f	3	0.00000000 B	.	.	.

PAGE 4

-----

Obs	y	f	yy	uu	erty	rest
1	11	1	12	-1	1.58114	-0.63246
2	13	1	12	1	1.58114	0.63246
3	15	2	18	-3	1.82574	-1.64317
4	18	2	18	0	1.82574	0.00000
5	21	2	18	3	1.82574	1.64317
6	19	3	21	-2	1.93649	-1.03280
7	20	3	21	-1	1.93649	-0.51640
8	22	3	21	1	1.93649	0.51640
9	23	3	21	2	1.93649	1.03280

### Estimation des paramètres

L'option `solution` de la commande `model` de la procédure GLM permet d'obtenir l'estimation des paramètres du modèle correspondant. Ici, SAS nous fournit les valeurs de  $\hat{m}$  (21), de  $\hat{a}_1$  (-9) et de  $\hat{a}_2$  (-3). On en déduit les estimations des paramètres  $\beta_j$  ( $\hat{\beta}_1 = -9 + 21 = 12$  ;  $\hat{\beta}_2 = -3 + 21 = 18$  ;  $\hat{\beta}_3 = 21$ ), donc des valeurs prédites (`yy`) et des résidus (`uu`).

### La commande output

Dans le fichier obtenu avec l'option `out=` de la commande `output` de la procédure GLM (ici, il s'appelle "sortie"), on récupère les valeurs prédites (`p=`), les résidus (`r=`), les erreurs-types des résidus (`stdr=`) et les résidus studentisés (`student=`) (on laisse le soin au lecteur de retrouver toutes ces valeurs).

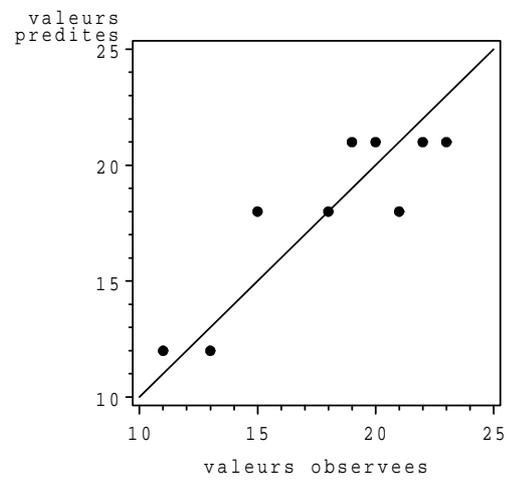
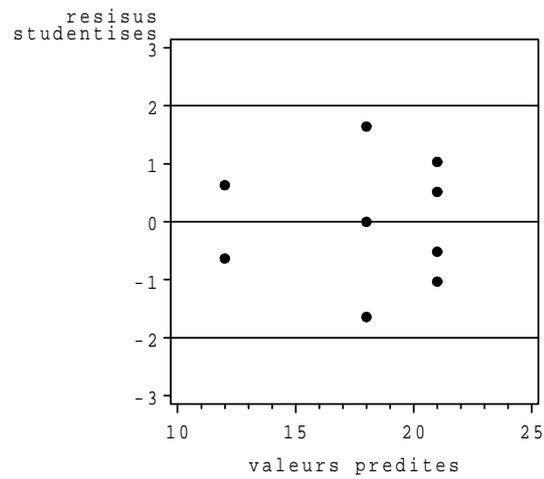
### Les graphiques

Les figures 3.1 et 3.2 donnent les graphiques tels qu'on les obtient en sortie de SAS.

## 3.2 Cas de deux facteurs croisés

### 3.2.1 Notations

- Le premier facteur est noté  $F_1$  et son nombre de niveaux est noté  $J$  ( $J \geq 2$ ) ; ces niveaux seront indicés par  $j$ .
- Le second facteur est noté  $F_2$  et son nombre de niveaux est noté  $K$  ( $K \geq 2$ ) ; les niveaux de  $F_2$  seront indicés par  $k$ .
- Les deux facteurs sont croisés, c'est-à-dire d'une part qu'ils sont symétriques (on peut permuter leurs rôles), d'autre part qu'on réalise des observations pour chacun des croisements  $(j, k)$  ; on notera  $n_{jk}$  le nombre d'observations réalisées pour le croisement  $(j, k)$ , le nombre total d'observations étant noté  $n$ , de sorte que l'on aura :  $n = \sum_{j=1}^J \sum_{k=1}^K n_{jk}$ . Lorsque  $n_{jk} = n_0 \forall (j, k)$ , on dit que le plan est **équilibré** ; sinon, on dit qu'il est **déséquilibré**.
- Chaque croisement  $(j, k)$  est en général appelé une *cellule* du plan factoriel.

FIG. 3.1 – *Graphique valeurs prédites vs valeurs observées.*FIG. 3.2 – *Graphique des résidus.*

- Les observations de la variable réponse  $Y$ , réalisées au sein de chaque cellule, seront supposées indépendantes et seront notées  $y_{ijk}$ ,  $i = 1, \dots, n_{jk}$ ; elles seront également supposées indépendantes sur l'ensemble des cellules.

**Remarque 12** On supposera, dans tout ce chapitre,  $n_{jk} \geq 1$ , autrement dit toute cellule est observée au moins une fois; on dit, dans ce cas, que le plan est **complet**. Lorsqu'au moins une cellule est telle que  $n_{jk} = 0$ , on dit que le plan est **incomplet**. Le cas des plans incomplets sera abordé au chapitre 4.

### 3.2.2 Écriture initiale du modèle

On écrit chaque v.a.r. sous la forme :

$$Y_{ijk} = \beta_{jk} + U_{ijk}.$$

Matriciellement, cela peut s'écrire :

$$Y = \mathbf{X}\beta + U.$$

$Y$  et  $U$  sont deux vecteurs de  $\mathbb{R}^n$ ,  $\beta$  est un vecteur de  $\mathbb{R}^{JK}$  (ici,  $p = JK$ ) et  $\mathbf{X}$ , matrice d'incidence, est de dimension  $n \times JK$ .

La matrice  $\mathbf{X}$  ne contient que des 0 et des 1, ses colonnes étant constituées des indicatrices  $Z^{jk}$  des cellules. Elle est de rang  $JK$  et ses colonnes sont deux à deux orthogonales. Les éléments  $\beta_{jk}$  du vecteur  $\beta$  sont les paramètres inconnus du modèle, à estimer. Enfin, on suppose toujours  $U_{ijk} \sim \mathcal{N}(0, \sigma^2)$ , les  $U_{ijk}$  étant i.i.d., le paramètre  $\sigma^2$  étant lui aussi inconnu et à estimer. On a donc  $Y_{ijk} \sim \mathcal{N}(\beta_{jk}, \sigma^2)$ , les  $Y_{ijk}$  étant indépendantes.

**Exemple 4** Considérons le cas très simple  $J = 2$ ,  $K = 3$  et  $n_0 = 1$  (une seule observation dans chacune des 6 cellules). Dans ce cas, la matrice d'incidence  $\mathbf{X}$  est tout simplement égale à la matrice identité  $\mathbf{I}_6$ .

### 3.2.3 Paramétrage centré

On introduit tout d'abord les quantités suivantes :

- $\beta_{j\bullet} = \frac{1}{K} \sum_{k=1}^K \beta_{jk}$ ; c'est la valeur moyenne des paramètres au niveau  $j$  de  $F_1$ ;
- $\beta_{\bullet k} = \frac{1}{J} \sum_{j=1}^J \beta_{jk}$ ; valeur moyenne des paramètres au niveau  $k$  de  $F_2$ ;
- $\beta_{\bullet\bullet} = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K \beta_{jk}$ ; valeur moyenne générale.

On notera que les différentes moyennes définies ci-dessus sont toujours des moyennes "non pondérées".

On définit ensuite les paramètres centrés de la façon suivante :

- $\mu = \beta_{\bullet\bullet}$  : c'est l'effet général, ou effet moyen général;
- $\alpha_j^1 = \beta_{j\bullet} - \beta_{\bullet\bullet}$  : effet principal, ou différentiel, du niveau  $j$  de  $F_1$ ;
- $\alpha_k^2 = \beta_{\bullet k} - \beta_{\bullet\bullet}$  : effet principal, ou différentiel, du niveau  $k$  de  $F_2$ ;
- $\gamma_{jk} = \beta_{jk} - \mu - \alpha_j^1 - \alpha_k^2 = \beta_{jk} - \beta_{j\bullet} - \beta_{\bullet k} + \beta_{\bullet\bullet}$  : effet d'interaction des niveaux  $j$  de  $F_1$  et  $k$  de  $F_2$ .

On vérifie, de façon immédiate, les relations de centrage suivantes :

$$\sum_{j=1}^J \alpha_j^1 = \sum_{k=1}^K \alpha_k^2 = 0; \quad \sum_{j=1}^J \gamma_{jk} = 0, \quad \forall k = 1, \dots, K; \quad \sum_{k=1}^K \gamma_{jk} = 0, \quad \forall j = 1, \dots, J.$$

Finalement, on peut réécrire le modèle sous la forme suivante :

$$Y_{ijk} = \beta_{jk} + U_{ijk} = \mu + \alpha_j^1 + \alpha_k^2 + \gamma_{jk} + U_{ijk}.$$

D'autre part, un raisonnement analogue à celui fait en 3.1.2, mais un peu plus lourd (il nécessite l'usage des indicatrices  $Z_1^j$  pour les niveaux de  $F_1$  et  $Z_2^k$  pour les niveaux de  $F_2$ ), nous permet maintenant d'écrire :

$$Y = \mu \mathbf{1}_n + \sum_{j=1}^{J-1} \alpha_j^1 (Z_1^j - Z_1^J) + \sum_{k=1}^{K-1} \alpha_k^2 (Z_2^k - Z_2^K) + \sum_{j=1}^{J-1} \sum_{k=1}^{K-1} \gamma_{jk} (Z_1^j - Z_1^J) (Z_2^k - Z_2^K) + U.$$

On retrouve ainsi un modèle de régression linéaire avec coefficient constant.

**Exemple 5** Reprenons le cas de l'exemple 4. Sous forme matricielle, on écrira le modèle  $Y = \mathbf{X}_c \beta_c + U$ , avec :

$$\mathbf{X}_c = \left( \begin{array}{c|c|c|c|c|c} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array} \right) ; \quad \beta_c = \begin{pmatrix} \mu \\ \alpha_1^1 \\ \alpha_1^2 \\ \alpha_2^2 \\ \gamma_{11} \\ \gamma_{12} \end{pmatrix}.$$

Le plan étant équilibré, deux colonnes de  $\mathbf{X}_c$  extraites de deux blocs différents sont toujours orthogonales. Par ailleurs, toujours parce que le plan est équilibré, toutes les colonnes, à l'exception de la première, sont centrées.

**Remarque 13** Le paramétrage centré fait intervenir un paramètre  $\mu$ ,  $(J-1)$  paramètres  $\alpha_j^1$  indépendants,  $(K-1)$  paramètres  $\alpha_k^2$  indépendants et  $(J-1)(K-1)$  paramètres  $\gamma_{jk}$  indépendants. Au total, il y a bien  $JK$  paramètres indépendants, comme dans le paramétrage initial en  $\beta_{jk}$ .

**Remarque 14** Considérons maintenant deux indices distincts  $j$  et  $j'$ , quelconques mais fixés. On peut écrire :

$$\beta_{jk} - \beta_{j'k} = (\alpha_j^1 - \alpha_{j'}^1) + (\gamma_{jk} - \gamma_{j'k}), \quad \forall k = 1, \dots, K.$$

On remarque que le premier terme est indépendant de  $k$  et que le second disparaît dans un modèle sans interaction. D'où l'idée de réaliser un graphique avec en abscisses les différents indices  $k$ , en ordonnées les valeurs moyennes  $\bar{y}_{\bullet jk}$  (estimations des  $\beta_{jk}$ , voir plus loin) et une "courbe" pour chaque indice  $j$  (courbes superposées). Si ces courbes sont sensiblement parallèles, on peut négliger les effets d'interactions dans le modèle considéré (voir les figures 3.5 et 3.6).

### 3.2.4 Paramétrage SAS

On réécrit maintenant :

$$\begin{aligned} \beta_{jk} &= \beta_{JK} + (\beta_{jK} - \beta_{JK}) + (\beta_{Jk} - \beta_{JK}) + (\beta_{jk} - \beta_{jK} - \beta_{Jk} + \beta_{JK}) \\ &= m + a_j^1 + a_k^2 + c_{jk}. \end{aligned}$$

Les paramètres définis ci-dessus vérifient ainsi les relations :

$$a_j^1 = a_K^2 = 0 ; \quad c_{jK} = 0, \quad \forall j = 1, \dots, J ; \quad c_{Jk} = 0, \quad \forall k = 1, \dots, K.$$

Bien sûr, il y a toujours  $JK$  paramètres indépendants dans ce nouveau paramétrage.

On peut encore vérifier que l'on obtient maintenant

$$Y = m\mathbb{I}_n + \sum_{j=1}^{J-1} a_j^1 Z_1^j + \sum_{k=1}^{K-1} a_k^2 Z_2^k + \sum_{j=1}^{J-1} \sum_{k=1}^{K-1} c_{jk} Z_1^j Z_2^k + U,$$

$m$  étant toujours appelé *intercept* dans SAS.

**Exemple 6** Reprenons une dernière fois l'exemple 4. Sous forme matricielle, le modèle s'écrit maintenant  $Y = \mathbf{X}_s \beta_s + U$ , avec :

$$\mathbf{X}_s = \left( \begin{array}{c|c|c|c|c|c} 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right) ; \quad \beta_s = \begin{pmatrix} m \\ a_1^1 \\ a_1^2 \\ a_2^2 \\ c_{11} \\ c_{12} \end{pmatrix}.$$

On notera que les colonnes de  $\mathbf{X}_s$  ne sont ni centrées ni orthogonales.

### 3.2.5 Estimation des paramètres

#### Paramétrage initial

À partir des résultats généraux relatifs au modèle linéaire, on déduit :

$$\hat{\beta}_{jk} = \bar{y}_{\bullet jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk}.$$

On sait que l'on a

$$\hat{B}_{jk} \sim \mathcal{N}\left(\beta_{jk}, \frac{\sigma^2}{n_{jk}}\right),$$

les différents estimateurs étant indépendants. Enfin, l'erreur-type de  $\hat{B}_{jk}$  est  $\frac{\hat{\sigma}}{\sqrt{n_{jk}}}$ , ce qui permet de construire un intervalle de confiance pour  $\beta_{jk}$ .

#### Paramétrage centré

Les estimations des paramètres se déduisent dans ce cas de celles ci-dessus. Il vient :

- $\hat{\mu} = \hat{\beta}_{\bullet\bullet} = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K \bar{y}_{\bullet jk}$ . Comme dans le cas d'un seul facteur, on notera que  $\hat{\mu}$  n'est égal à la moyenne générale des observations de  $Y$  que si le plan est équilibré.
- $\hat{\alpha}_j^1 = \hat{\beta}_{j\bullet} - \hat{\beta}_{\bullet\bullet} = \frac{1}{K} \sum_{k=1}^K \bar{y}_{\bullet jk} - \hat{\mu}$  (les  $\hat{\alpha}_j^1$  sont centrés selon  $j$ ).
- $\hat{\alpha}_k^2 = \hat{\beta}_{\bullet k} - \hat{\beta}_{\bullet\bullet} = \frac{1}{J} \sum_{j=1}^J \bar{y}_{\bullet jk} - \hat{\mu}$  (les  $\hat{\alpha}_k^2$  sont centrés selon  $k$ ).
- $\hat{\gamma}_{jk} = \hat{\beta}_{jk} - \hat{\alpha}_j^1 - \hat{\alpha}_k^2 - \hat{\mu} = \hat{\beta}_{jk} - \hat{\beta}_{j\bullet} - \hat{\beta}_{\bullet k} + \hat{\beta}_{\bullet\bullet}$   
 $= \bar{y}_{\bullet jk} - \frac{1}{K} \sum_{k=1}^K \bar{y}_{\bullet jk} - \frac{1}{J} \sum_{j=1}^J \bar{y}_{\bullet jk} + \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K \bar{y}_{\bullet jk}$   
 (les  $\hat{\gamma}_{jk}$  sont centrés selon  $j$  et selon  $k$ ).

#### Paramétrage SAS

De la même manière, on obtient maintenant :

- $\hat{m} = \bar{y}_{\bullet JK}$  ;
- $\hat{a}_j^1 = \bar{y}_{\bullet jk} - \bar{y}_{\bullet JK}$  ( $\hat{a}_j^1 = 0$ ) ;
- $\hat{a}_k^2 = \bar{y}_{\bullet jk} - \bar{y}_{\bullet JK}$  ( $\hat{a}_k^2 = 0$ ) ;
- $\hat{c}_{jk} = \bar{y}_{\bullet jk} - \bar{y}_{\bullet jk} - \bar{y}_{\bullet jk} + \bar{y}_{\bullet JK}$   
 ( $\hat{c}_{jk} = 0$ , dès que l'un au moins des deux indices est maximum).

#### Valeurs prédites et résidus

De façon standard, les valeurs prédites  $\hat{y}_{ijk}$  valent  $\hat{\beta}_{jk}$  ( $= \bar{y}_{\bullet jk}$ ) et les résidus  $\hat{u}_{ijk}$  valent  $y_{ijk} - \bar{y}_{\bullet jk}$ . L'erreur-type de  $\hat{y}_{ijk}$  vaut  $\frac{\hat{\sigma}}{\sqrt{n_{jk}}}$  et celle de  $\hat{u}_{ijk}$  vaut  $\hat{\sigma} \sqrt{\frac{n_{jk}-1}{n_{jk}}}$ . On notera que ces expressions (indépendantes du paramétrage choisi) ne sont plus valables avec un modèle autre que le modèle complet.

#### Variance

L'estimation de la variance  $\sigma^2$  est la suivante :

$$\hat{\sigma}^2 = \frac{1}{n - JK} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (\hat{u}_{ijk})^2 = \frac{1}{n - JK} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\bullet jk})^2.$$

### 3.2.6 Tests d'hypothèses

Dans un modèle à deux facteurs croisés, trois hypothèses nulles peuvent être envisagées afin de simplifier le modèle considéré.

#### Hypothèse $H_0$ : absence d'effet des interactions

Cette hypothèse peut prendre les différentes formes suivantes, équivalentes :

$$H_0 \iff \gamma_{jk} = 0, \forall(j, k) \iff c_{jk} = 0, \forall(j, k)$$

$$\iff \beta_{jk} - \beta_{j'k} \text{ est indépendant de } k, \forall(j, j') \iff \beta_{jk} - \beta_{jk'} \text{ est indépendant de } j, \forall(k, k').$$

Ainsi, avec le paramétrage centré,  $H_0$  conduit à réécrire le modèle sous la forme :

$$Y_{ijk} = \mu + \alpha_j^1 + \alpha_k^2 + U_{ijk}^0,$$

que l'on appelle le **modèle additif**.

La valeur de la statistique du test de  $H_0$  est

$$f = \frac{1}{(J-1)(K-1)\hat{\sigma}^2} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (\hat{y}_{ijk} - \hat{y}_{ijk}^0)^2,$$

où  $\hat{y}_{ijk}^0$  est la valeur prédite de  $y_{ijk}$  dans le modèle additif. Cette statistique est à comparer avec le quantile  $f_{(J-1)(K-1); n-JK} (1-\alpha)$ .

#### Hypothèse $H'_0$ : absence d'effet du facteur $F_1$

Cette autre hypothèse peut prendre les différentes formes suivantes :

$$H'_0 \iff \beta_{1\bullet} = \dots = \beta_{J\bullet} \iff \alpha_j^1 = 0, \forall j = 1, \dots, J \iff a_j^1 = 0, \forall j = 1, \dots, J.$$

#### Convention

Dans tout ce cours, nous ferons les tests de façon hiérarchique, c'est-à-dire que, dans le cas présent, nous ne testerons l'hypothèse  $H'_0$  que si, au préalable, l'hypothèse  $H_0$  n'a pas été rejetée ; ainsi, le modèle de référence pour tester  $H'_0$  sera le modèle additif. Cette façon de procéder n'est pas universelle, mais elle a le mérite d'être cohérente et simple à interpréter.

Sous  $H'_0$ , le modèle s'écrit donc :

$$Y_{ijk} = \mu + \alpha_k^2 + U_{ijk}^{0'}.$$

La valeur de la statistique du test est maintenant

$$f' = \frac{1}{(J-1)(\hat{\sigma}^0)^2} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (\hat{y}_{ijk} - \hat{y}_{ijk}^{0'})^2,$$

où  $(\hat{\sigma}^0)^2$  désigne l'estimation de  $\sigma^2$  dans le modèle additif,

$$(\hat{\sigma}^0)^2 = \frac{1}{n - (J + K - 1)} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ijk} - \hat{y}_{ijk}^0)^2,$$

et où  $\hat{y}_{ijk}^{0'}$  est la valeur prédite de  $y_{ijk}$  dans le modèle avec  $F_2$  comme seul facteur. La valeur  $f'$  est à comparer au quantile  $f_{J-1; n-(J+K-1)} (1-\alpha)$ .

**Hypothèse  $H_0''$  : absence d'effet du facteur  $F_2$** 

Cette dernière hypothèse peut prendre les différentes formes suivantes :

$$H_0'' \iff \beta_{\bullet 1} = \dots = \beta_{\bullet K} \iff \alpha_k^2 = 0, \forall k = 1, \dots, K \iff a_k^2 = 0, \forall k = 1, \dots, K.$$

On raisonne alors de façon symétrique par rapport à ce qui a été fait au point précédent.

**Remarque 15** *Le test de Fisher est également utilisé pour tester la significativité du modèle finalement retenu, autrement dit pour tester le modèle constant contre ce modèle.*

**Remarque 16** *Si l'on choisit le modèle additif pour un jeu de données, les estimations des paramètres  $\beta_{jk}$  avec le paramétrage centré sont obtenues directement en posant  $\hat{\beta}_{jk} = \hat{\mu} + \hat{\alpha}_j^1 + \hat{\alpha}_k^2$ , où les estimations  $\hat{\mu}$ ,  $\hat{\alpha}_j^1$  et  $\hat{\alpha}_k^2$  sont celles obtenues dans le modèle complet. Il suffit donc d'annuler les estimations des paramètres d'interactions pour trouver, à partir du modèle complet, les nouvelles estimations des  $\beta_{jk}$ . Par contre, il en va différemment avec le paramétrage SAS : la matrice d'incidence  $\mathbf{X}_s$  doit être modifiée (par suppression des colonnes relatives aux interactions), on doit ensuite calculer  $\hat{\beta}_s = (\mathbf{X}'_s \mathbf{X}_s)^{-1} \mathbf{X}'_s y$  et en déduire les nouvelles estimations des paramètres  $m$ ,  $a_j^1$  et  $a_k^2$ . Ainsi, le paramétrage centré apparaît comme plus naturel que le paramétrage SAS (ou que tout autre paramétrage).*

**3.2.7 Cas particulier d'un plan équilibré**

On dit qu'un plan à deux facteurs croisés est équilibré s'il vérifie  $n_{jk} = n_0, \forall (j, k)$ . Dans ce cas,  $n = n_0 JK$  et diverses écritures se simplifient. On obtient ainsi :

$$\hat{\beta}_{jk} = \bar{y}_{\bullet jk} = \frac{1}{n_0} \sum_{i=1}^{n_0} y_{ijk};$$

$$\hat{\beta}_{j\bullet} = \frac{1}{K} \sum_{k=1}^K \bar{y}_{\bullet jk} = \frac{1}{n_0 K} \sum_{k=1}^K \sum_{i=1}^{n_0} y_{ijk} = \bar{y}_{\bullet j\bullet}$$

(moyenne de toutes les observations de  $Y$  au niveau  $j$  du facteur  $F_1$ ) ;

$$\hat{\beta}_{\bullet K} = \bar{y}_{\bullet \bullet K} \text{ (même chose) ;}$$

$$\hat{\beta}_{\bullet \bullet} = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K \bar{y}_{\bullet jk} = \frac{1}{n} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_0} y_{ijk} = \bar{y}_{\bullet \bullet \bullet}$$

(moyenne générale de toutes les observations de  $Y$ ).

Les calculs des statistiques des tests vus précédemment peuvent encore se synthétiser, dans ce cas, sous la forme d'un tableau d'analyse de la variance (voir plus loin).

Les sommes de carrés apparaissant dans ce tableau sont définies de la façon suivante :

$$SSF_1 = n_0 K \sum_{j=1}^J (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet \bullet \bullet})^2 = n_0 K \sum_{j=1}^J (\hat{\alpha}_j^1)^2 ;$$

$$SSF_2 = n_0 J \sum_{k=1}^K (\bar{y}_{\bullet \bullet k} - \bar{y}_{\bullet \bullet \bullet})^2 = n_0 J \sum_{k=1}^K (\hat{\alpha}_k^2)^2 ;$$

$$SSF_{12} = n_0 \sum_{j=1}^J \sum_{k=1}^K (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet \bullet k} + \bar{y}_{\bullet \bullet \bullet})^2 = n_0 \sum_{j=1}^J \sum_{k=1}^K (\hat{\gamma}_{jk})^2 ;$$

$$SSE = \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_0} (y_{ijk} - \bar{y}_{\bullet jk})^2 ;$$

$$SST = \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_0} (y_{ijk} - \bar{y}_{\bullet \bullet \bullet})^2.$$

Tableau d'analyse de la variance (cas équilibré)

sources de variation	sommes des carrés	d.d.l.	carrés moyens	valeurs des statistiques de Fisher
$F_1$	$SSF_1$	$J - 1$	$MSF_1 = \frac{SSF_1}{J - 1}$	$\frac{MSF_1}{MSE}$
$F_2$	$SSF_2$	$K - 1$	$MSF_2 = \frac{SSF_2}{K - 1}$	$\frac{MSF_2}{MSE}$
$F_1 * F_2$	$SSF_{12}$	$(J - 1)(K - 1)$	$MSF_{12} = \frac{SSF_{12}}{(J - 1)(K - 1)}$	$\frac{MSF_{12}}{MSE}$
Erreur	$SSE$	$n - JK$	$MSE = \frac{SSE}{n - JK} = \hat{\sigma}^2$	—
Total	$SST$	$n - 1$	—	—

**Remarque 17** Dans le tableau ci-dessus, les tests de significativité des facteurs  $F_1$  et  $F_2$  sont faits avec, pour modèle de référence, le modèle complet (c'est-à-dire, le modèle comportant tous les effets initialement introduits ; on l'appelle encore le modèle plein). Si l'hypothèse de nullité des interactions n'est pas rejetée et qu'on souhaite faire ces tests (significativité de chaque facteur) avec le modèle additif comme référence, au dénominateur de la statistique de Fisher, on doit remplacer  $MSE$  par  $\frac{SSE + SSF_{12}}{n - J - K + 1} = (\hat{\sigma}^0)^2$ , estimation de  $\sigma^2$  dans le modèle additif.

**Remarque 18** Dans le cas déséquilibré, on ne peut pas construire de tableau analogue. En effet, le développement de  $SST$  (la somme des carrés totale) en fonction des autres sommes de carrés fait aussi intervenir dans ce cas les doubles produits (qui sont nuls dans le cas équilibré). Ces doubles produits ne pouvant pas être affectés à un effet spécifique, le tableau d'analyse de la variance n'a plus de raison d'être dans ce cas.

**Remarque 19** La définition même des sommes de carrés n'est plus très claire dans le cas déséquilibré. Cela conduit à introduire diverses sommes de carrés (trois), appelées de type I, de type II et de type III (toutes égales dans le cas équilibré). De plus, des sommes de carrés spécifiques au cas des plans incomplets existent également et sont appelées sommes de type IV. On trouvera en Annexe B quelques précisions sur les trois premiers types de sommes de carrés. Sauf indication contraire, il est recommandé d'utiliser les sommes de type III.

### 3.2.8 Illustration

#### Les données

Il s'agit d'un célèbre exemple (fictif) d'analyse de variance à deux facteurs croisés. La variable réponse, en dernière colonne, est le rendement laitier mesuré sur un échantillon de 40 vaches laitières de la même espèce. Il y a deux facteurs contrôlés, tous deux liés à l'alimentation des vaches : la dose, en première colonne, à 2 niveaux (1 = dose faible, 2 = dose forte) ; le régime alimentaire, en deuxième colonne, à 4 niveaux (1 = paille, 2 = foin, 3 = herbe, 4 = aliments ensilés). Pour chaque dose et chaque régime (8 cellules), on a observé le rendement de 5 vaches. On a donc affaire à un plan complet, équilibré, avec 5 répétitions. Les données sont reproduites ci-dessous.

1	1	8	2	1	8
1	1	11	2	1	9
1	1	11	2	1	8
1	1	10	2	1	10
1	1	7	2	1	9

1	2	12	2	2	10
1	2	13	2	2	7
1	2	14	2	2	10
1	2	11	2	2	12
1	2	10	2	2	11
1	3	10	2	3	11
1	3	12	2	3	9
1	3	12	2	3	11
1	3	13	2	3	11
1	3	14	2	3	12
1	4	17	2	4	17
1	4	13	2	4	19
1	4	17	2	4	17
1	4	14	2	4	16
1	4	13	2	4	21

### Le programme SAS

Le programme ci-dessous réalise la procédure GLM sur les données des vaches laitières, trace les deux graphiques de contrôle du modèle choisi (le modèle complet), puis les deux graphiques des interactions. On trouvera des compléments sur cet exemple en Annexe A.

```

options pagesize=64 linesize=76 nodate;
title;
footnote 'ANOVA 2 facteurs - vaches laitieres';
* ----- ;
data vach;
infile 'vach.don';
input f1 f2 y;
run;
* ----- ;
proc glm;
class f1 f2;
model y = f1 f2 f1*f2 / ss3 solution;
output out=sortie p=yy r=uu stdr=erty student=rest;
lsmeans f1 f2 f1*f2 / out=graph;
run;
quit;
* ----- ;
*          graphiques de controle du modele          ;
* ----- ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach1.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=sortie;
axis1 label=('valeurs observees') order=(6 to 22 by 2)
      minor=none length=7cm;
axis2 label=('valeurs' justify=right 'predites')
      order=(6 to 22 by 2) minor=none length=7cm;
symbol1 v=dot i=none;
symbol2 v=none i=rl;
plot yy*y y*y / haxis=axis1 vaxis=axis2 overlay;
run;
goptions reset=all;
quit;
* ----- ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach2.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=sortie;
axis1 label=('valeurs predites')

```

```

order=(6 to 20 by 2) minor=none length=7cm;
axis2 label=('resisus' justify=right 'studentises')
order=(-3 to 3 by 1) minor=none length=7cm;
symbol v=dot;
plot rest*yy / haxis=axis1 vaxis=axis2
vref=-2 vref=0 vref=2;

run;
goptions reset=all;
quit;
* ----- ;
*           graphiques des interactions           ;
* ----- ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach3.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=graph;
axis1 label=('premier facteur') order=(1 to 2 by 1)
minor=none length=6cm;
axis2 label=('moyenne' justify=right 'des effets')
order=(8 to 20 by 2) minor=none length=6cm;
symbol1 i=join v=dot;
symbol2 i=join v=triangle;
symbol3 i=join v=circle;
symbol4 i=join v=#;
symbol5 i=join v=%;
plot lsmean*f1=f2 / haxis=axis1 vaxis=axis2;
run;
goptions reset=all;
quit;
* ----- ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach4.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=graph;
axis1 label=('second facteur') order=(1 to 4 by 1)
minor=none length=6cm;
axis2 label=('moyenne' justify=right 'des effets')
order=(8 to 20 by 2) minor=none length=6cm;
symbol1 i=join v=dot;
symbol2 i=join v=triangle;
symbol3 i=join v=circle;
plot lsmean*f2=f1 / haxis=axis1 vaxis=axis2;
run;
goptions reset=all;
quit;

```

### Les sorties de la procédure GLM

PAGE 1

The GLM Procedure

-----

#### Class Level Information

Class	Levels	Values
f1	2	1 2
f2	4	1 2 3 4
Number of observations		40

PAGE 2

-----

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	331.6000000	47.3714286	17.54	<.0001
Error	32	86.4000000	2.7000000		
Corrected Total	39	418.0000000			

R-Square	Coeff Var	Root MSE	y Mean
0.793301	13.69306	1.643168	12.00000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	0.4000000	0.4000000	0.15	0.7029
f2	3	290.2000000	96.7333333	35.83	<.0001
f1*f2	3	41.0000000	13.6666667	5.06	0.0056

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	18.00000000 B	0.73484692	24.49	<.0001
f1 1	-3.20000000 B	1.03923048	-3.08	0.0042
f1 2	0.00000000 B	.	.	.
f2 1	-9.20000000 B	1.03923048	-8.85	<.0001
f2 2	-8.00000000 B	1.03923048	-7.70	<.0001
f2 3	-7.20000000 B	1.03923048	-6.93	<.0001
f2 4	0.00000000 B	.	.	.
f1*f2 1 1	3.80000000 B	1.46969385	2.59	0.0145
f1*f2 1 2	5.20000000 B	1.46969385	3.54	0.0013
f1*f2 1 3	4.60000000 B	1.46969385	3.13	0.0037
f1*f2 1 4	0.00000000 B	.	.	.
f1*f2 2 1	0.00000000 B	.	.	.
f1*f2 2 2	0.00000000 B	.	.	.
f1*f2 2 3	0.00000000 B	.	.	.
f1*f2 2 4	0.00000000 B	.	.	.

PAGE 3

-----

## Least Squares Means

f1 y LSMEAN

1	12.1000000
2	11.9000000

f2 y LSMEAN

1	9.1000000
2	11.0000000
3	11.5000000
4	16.4000000

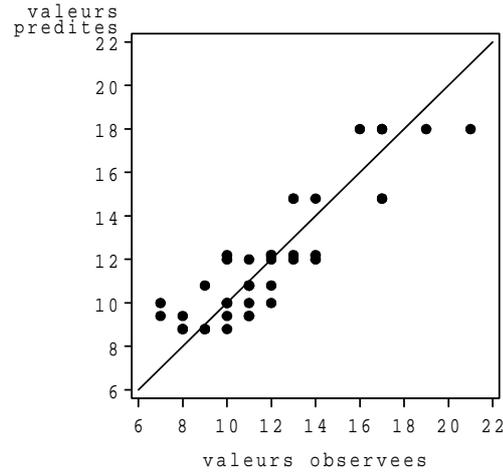


FIG. 3.3 – Graphique valeurs prédites vs valeurs observées.

f1	f2	y LSMEAN
1	1	9.4000000
1	2	12.0000000
1	3	12.2000000
1	4	14.8000000
2	1	8.8000000
2	2	10.0000000
2	3	10.8000000
2	4	18.0000000

### Commentaires

La commande `lsmeans` de la procédure GLM permet, d'une part, d'obtenir en sortie certaines moyennes des  $y_{ijk}$  (ici, selon les niveaux de chaque facteur, puis selon les cellules), d'autre part, de récupérer la table SAS, ici appelée `graph`, qui permet de réaliser les graphiques d'interactions.

### Les graphiques

Le programme ci-dessus produit les quatre graphiques 3.3 à 3.6.

## 3.3 Cas de trois facteurs croisés

### 3.3.1 Notations

- Les trois facteurs considérés sont notés  $F_1$ ,  $F_2$  et  $F_3$ .
- Le nombre de niveaux de  $F_1$  est noté  $J$ , celui de  $F_2$  est noté  $K$  et celui de  $F_3$  est noté  $L$  ( $J \geq 2$  ;  $K \geq 2$  ;  $L \geq 2$ ).
- Les niveaux de  $F_1$  sont indicés par  $j$ , ceux de  $F_2$  par  $k$  et ceux de  $F_3$  par  $\ell$ .
- Les trois facteurs étant croisés, on considère les  $JKL$  cellules (ou triplets)  $(j, k, \ell)$ .
- Dans chaque cellule, on réalise  $n_{jkl}$  observations de la variable à expliquer  $Y$  et on pose  $n = \sum_{j=1}^J \sum_{k=1}^K \sum_{\ell=1}^L n_{jkl}$ . On suppose toujours que le plan est complet :  $\forall (j, k, \ell), n_{jkl} \geq 1$  ; de plus, si  $\forall (j, k, \ell), n_{jkl} = n_0$ , alors le plan est équilibré.
- On notera  $Y_{ijkl}$  ( $i = 1, \dots, n_{jkl}$ ) les v.a.r. associées aux observations de  $Y$  dans la cellule  $(j, k, \ell)$ .

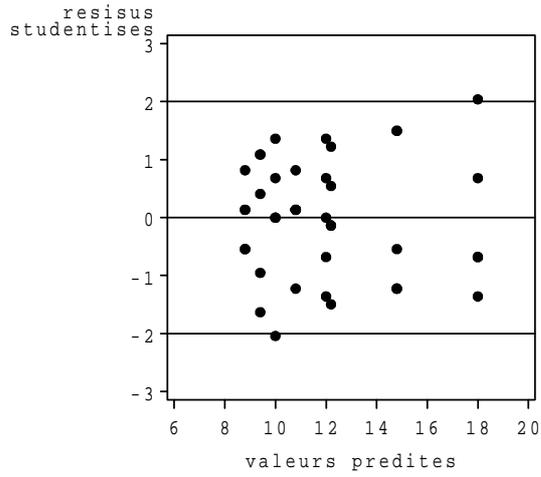


FIG. 3.4 – Graphique des résidus.

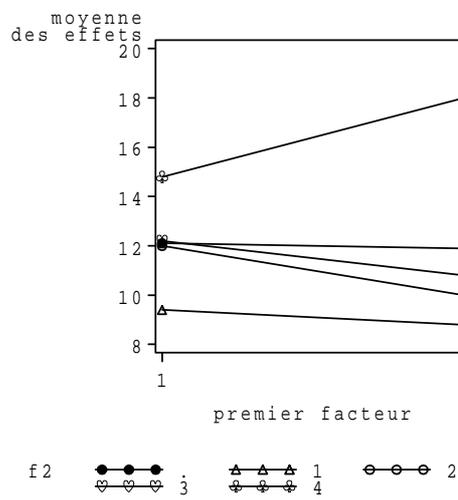


FIG. 3.5 – Premier graphique des interactions.

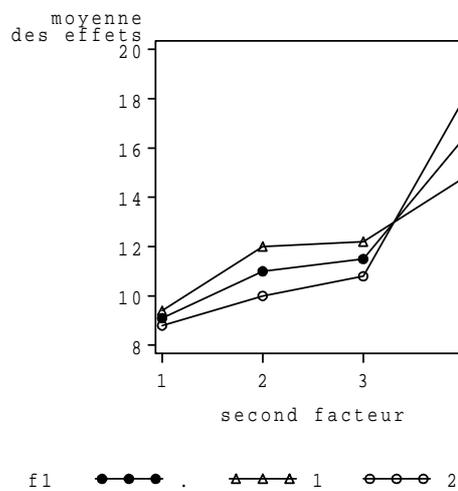


FIG. 3.6 – Second graphique des interactions.

### 3.3.2 Modèle

#### Paramétrage initial

Dans un premier temps, on écrit le modèle sous la forme

$$Y_{ijkl} = \beta_{jkl} + U_{ijkl},$$

avec toujours les mêmes hypothèses sur les v.a.r.  $U_{ijkl}$  (elles sont i.i.d.,  $\mathcal{N}(0, \sigma^2)$ ). Il y a donc  $JKL$  paramètres  $\beta_{jkl}$  indépendants à estimer, en plus de  $\sigma^2$  (ici,  $p = JKL$ ).

#### Paramétrage centré

Il fait intervenir toutes les moyennes partielles (non pondérées) des paramètres  $\beta_{jkl}$ . Définissons tout d'abord ces moyennes :

$$\begin{aligned} \beta_{\bullet k \ell} &= \frac{1}{J} \sum_{j=1}^J \beta_{jkl} ; \quad \beta_{j \bullet \ell} = \frac{1}{K} \sum_{k=1}^K \beta_{jkl} ; \quad \beta_{jk \bullet} = \frac{1}{L} \sum_{\ell=1}^L \beta_{jkl} ; \\ \beta_{\bullet \bullet \ell} &= \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K \beta_{jkl} ; \quad \beta_{\bullet k \bullet} = \frac{1}{JL} \sum_{j=1}^J \sum_{\ell=1}^L \beta_{jkl} ; \quad \beta_{j \bullet \bullet} = \frac{1}{KL} \sum_{k=1}^K \sum_{\ell=1}^L \beta_{jkl} ; \\ \beta_{\bullet \bullet \bullet} &= \frac{1}{JKL} \sum_{j=1}^J \sum_{k=1}^K \sum_{\ell=1}^L \beta_{jkl}. \end{aligned}$$

On réécrit alors le modèle sous la forme :

$$Y_{ijkl} = \mu + \alpha_j^1 + \alpha_k^2 + \alpha_\ell^3 + \gamma_{jk}^{12} + \gamma_{j\ell}^{13} + \gamma_{k\ell}^{23} + \delta_{jkl} + U_{ijkl} = \beta_{jkl} + U_{ijkl}.$$

- Le paramètre  $\mu$  est l'effet général (1 paramètre). Il est défini par :  $\mu = \beta_{\bullet \bullet \bullet}$ .
- Les paramètres  $\alpha_j^1$ ,  $\alpha_k^2$  et  $\alpha_\ell^3$  sont les effets principaux associés aux différents niveaux de chacun des trois facteurs. Ils sont définis par les relations suivantes :

$$\alpha_j^1 = \beta_{j \bullet \bullet} - \beta_{\bullet \bullet \bullet} ; \quad \alpha_k^2 = \beta_{\bullet k \bullet} - \beta_{\bullet \bullet \bullet} ; \quad \alpha_\ell^3 = \beta_{\bullet \bullet \ell} - \beta_{\bullet \bullet \bullet}.$$

Ils vérifient :

$$\sum_{j=1}^J \alpha_j^1 = \sum_{k=1}^K \alpha_k^2 = \sum_{\ell=1}^L \alpha_\ell^3 = 0.$$

Il y a donc  $(J-1) + (K-1) + (L-1)$  paramètres  $\alpha$  indépendants.

- Les paramètres  $\gamma_{jk}^{12}$ ,  $\gamma_{j\ell}^{13}$  et  $\gamma_{k\ell}^{23}$  sont les effets d'interactions d'ordre 2 (entre 2 facteurs). Ils sont définis par les relations suivantes :

$$\begin{aligned} \gamma_{jk}^{12} &= \beta_{jk \bullet} - \beta_{j \bullet \bullet} - \beta_{\bullet k \bullet} + \beta_{\bullet \bullet \bullet} ; \\ \gamma_{j\ell}^{13} &= \beta_{j \bullet \ell} - \beta_{j \bullet \bullet} - \beta_{\bullet \bullet \ell} + \beta_{\bullet \bullet \bullet} ; \\ \gamma_{k\ell}^{23} &= \beta_{\bullet k \ell} - \beta_{\bullet k \bullet} - \beta_{\bullet \bullet \ell} + \beta_{\bullet \bullet \bullet}. \end{aligned}$$

Ils vérifient :

$$\sum_{j=1}^J \gamma_{jk}^{12} = \sum_{k=1}^K \gamma_{jk}^{12} = \sum_{j=1}^J \gamma_{j\ell}^{13} = \sum_{\ell=1}^L \gamma_{j\ell}^{13} = \sum_{k=1}^K \gamma_{k\ell}^{23} = \sum_{\ell=1}^L \gamma_{k\ell}^{23} = 0.$$

Il y a donc  $(J-1)(K-1) + (J-1)(L-1) + (K-1)(L-1)$  paramètres  $\gamma$  indépendants.

- Les paramètres  $\delta_{jkl}$  sont les effets d'interactions d'ordre 3 (entre 3 facteurs). Ils sont définis par :

$$\delta_{jkl} = \beta_{jkl} - \beta_{\bullet k \ell} - \beta_{j \bullet \ell} - \beta_{jk \bullet} + \beta_{\bullet \bullet \ell} + \beta_{\bullet k \bullet} + \beta_{j \bullet \bullet} - \beta_{\bullet \bullet \bullet}.$$

Ils vérifient :

$$\sum_{j=1}^J \delta_{jkl} = \sum_{k=1}^K \delta_{jkl} = \sum_{\ell=1}^L \delta_{jkl} = 0.$$

Il y a donc  $(J-1)(K-1)(L-1)$  paramètres  $\delta$  indépendants.

– Au total, ce nouveau paramétrage fait intervenir

$$\begin{aligned}
1 &+ (J-1) + (K-1) + (L-1) \\
&+ (J-1)(K-1) + (J-1)(L-1) + (K-1)(L-1) \\
&+ (J-1)(K-1)(L-1) \\
&= JKL
\end{aligned}$$

paramètres indépendants, ce qui est cohérent.

### Paramétrage SAS

Le paramétrage SAS utilise toujours, dans ce cas, le principe consistant à prendre le dernier niveau de chaque facteur comme niveau de référence. Dans SAS, on réécrit le modèle sous la forme :

$$Y_{ijkl} = m + a_j^1 + a_k^2 + a_\ell^3 + c_{jk}^{12} + c_{j\ell}^{13} + c_{k\ell}^{23} + d_{jkl} + U_{ijkl} = \beta_{jkl} + U_{ijkl}.$$

De façon “logique” (compte tenu de ce qui a déjà été fait avec deux facteurs croisés), les paramètres sont définis de la manière suivante :

$$\begin{aligned}
m &= \beta_{JKL} ; \\
a_j^1 &= \beta_{jKL} - \beta_{JKL} ; \quad a_k^2 = \beta_{jK\ell} - \beta_{JKL} ; \quad a_\ell^3 = \beta_{JK\ell} - \beta_{JKL} ; \\
(a_j^1 &= a_k^2 = a_\ell^3 = 0) ; \\
c_{jk}^{12} &= \beta_{jkL} - \beta_{jKL} - \beta_{jK\ell} + \beta_{JKL} ; \quad c_{j\ell}^{13} = \beta_{jK\ell} - \beta_{jKL} - \beta_{jK\ell} + \beta_{JKL} ; \\
c_{k\ell}^{23} &= \beta_{JK\ell} - \beta_{jK\ell} - \beta_{JK\ell} + \beta_{JKL} ; \\
(c_{jK}^{12} &= 0, \forall j ; \quad c_{Jk}^{12} = 0, \forall k ; \quad c_{jL}^{13} = 0, \forall j ; \quad c_{J\ell}^{13} = 0, \forall \ell ; \quad c_{kL}^{23} = 0, \forall k ; \quad c_{K\ell}^{23} = 0, \forall \ell) ; \\
d_{jkl} &= \beta_{jkl} - \beta_{jK\ell} - \beta_{jK\ell} - \beta_{JK\ell} + \beta_{JKL} + \beta_{JKL} + \beta_{JK\ell} - \beta_{JKL} ; \\
(d_{jkl} &= 0, \text{ dès qu'au moins un des indices } j, k \text{ ou } \ell \text{ est maximum}).
\end{aligned}$$

Il y a toujours un total de  $JKL$  paramètres indépendants dans le paramétrage SAS.

### 3.3.3 Estimations

Selon le même principe que celui vu dans les cas de un ou de deux facteurs, on obtient, comme estimation de chaque paramètre  $\beta_{jkl}$ , la quantité :

$$\hat{\beta}_{jkl} = \bar{y}_{\bullet jkl} = \frac{1}{n_{jkl}} \sum_{i=1}^{n_{jkl}} y_{ijkl}.$$

Pour le paramétrage centré, on calcule ensuite toutes les moyennes partielles (toujours non pondérées) de ces estimations, respectivement notées :

$$\hat{\beta}_{\bullet k\ell} \quad \hat{\beta}_{j\bullet\ell} \quad \hat{\beta}_{jk\bullet} \quad \hat{\beta}_{\bullet\bullet\ell} \quad \hat{\beta}_{\bullet k\bullet} \quad \hat{\beta}_{j\bullet\bullet} \quad \hat{\beta}_{\bullet\bullet\bullet}.$$

Les estimations des paramètres  $\mu$ ,  $\alpha_j^1$ ,  $\alpha_k^2$ ,  $\alpha_\ell^3$ ,  $\gamma_{jk}^{12}$ ,  $\gamma_{j\ell}^{13}$ ,  $\gamma_{k\ell}^{23}$  et  $\delta_{jkl}$  s'obtiennent, à partir des estimations précédentes, en utilisant les mêmes formules que celles ayant permis de définir ces paramètres à partir des moyennes partielles des  $\beta_{jkl}$ .

Dans le paramétrage SAS, selon le même principe, on utilise les mêmes formules que celles données plus haut en remplaçant les  $\beta$  par leurs estimations. On obtient ainsi les estimations des paramètres définis par SAS.

### 3.3.4 Tests

Là encore, nous préconisons de procéder de façon hiérarchique pour faire les tests (de Fisher) de nullité des différentes catégories de paramètres.

- On commence donc par tester la nullité des effets d'interactions d'ordre 3 :

$$\{H_0 : \delta_{jkl} = 0, \forall(j, k, \ell)\}.$$

Si cette hypothèse est rejetée, on garde le modèle complet et les tests sont terminés.

- Si, au contraire, l'hypothèse ci-dessus n'est pas rejetée, on prend comme modèle de référence le modèle sans interactions d'ordre 3. Dans ce modèle, on teste successivement la nullité des différents effets d'interactions d'ordre 2 :

$$\{H_0^{12} : \gamma_{jk}^{12} = 0, \forall(j, k)\} ; \{H_0^{13} : \gamma_{j\ell}^{13} = 0, \forall(j, \ell)\} ; \{H_0^{23} : \gamma_{k\ell}^{23} = 0, \forall(k, \ell)\}.$$

- Si les trois hypothèses sont rejetées, on garde le modèle avec les trois séries d'effets d'interactions d'ordre 2 et les tests sont terminés.
- Si deux hypothèses sont rejetées, on enlève seulement du modèle les effets d'interactions supposés nuls et les tests sont terminés (chacun des trois facteurs doit être conservé car il intervient dans au moins une des séries d'interactions d'ordre deux).
- Si une seule hypothèse est rejetée, on enlève les effets d'interactions supposés nuls et on teste la nullité des effets du facteur n'intervenant pas dans les interactions (il y en a un et un seul). Si cette dernière hypothèse est rejetée, les tests sont terminés. Sinon, on enlève le facteur correspondant et on se retrouve dans un modèle d'ANOVA à deux facteurs (les tests sont également terminés).
- Si aucune hypothèse n'est rejetée, on prend alors le modèle additif (sans aucune interaction) pour référence et on teste séparément la nullité des effets de chaque facteur.

**Remarque 20** *On utilisera encore le test de Fisher pour tester la significativité du modèle retenu.*

**Remarque 21** *Dans le cadre d'un plan à trois facteurs, on appelle toujours modèle additif le modèle sans aucune interaction.*

### 3.4 Généralisation

Conceptuellement, il n'est pas difficile de définir des plans factoriels à quatre facteurs croisés ou plus. Toutefois, leur écriture devient très vite inextricable.

Il faut noter que, dans la pratique, notamment industrielle, il n'est pas rare de trouver de tels plans à au moins quatre facteurs. Toutefois, dans ce genre de situations, on a le plus souvent affaire à des plans incomplets, les plans complets étant trop coûteux à mettre en œuvre. Il convient alors de choisir de façon spécifique les cellules dans lesquelles on réalise les observations, de manière à obtenir un maximum de propriétés statistiques avec un minimum d'observations. L'étude de certains de ces plans factoriels incomplets est abordée dans le chapitre 4.

Enfin, signalons qu'on rencontre également, dans la pratique, des plans à deux ou plusieurs facteurs hiérarchisés (les niveaux d'un facteur sont conditionnés par les niveaux d'un autre facteur). Nous n'aborderons pas ce type de plans dans ce cours, mais signalons néanmoins qu'il est très simple de les mettre en œuvre avec la procédure GLM de SAS et que c'est dans ce cas que les sommes de carrés de type I s'avèrent utiles (voir l'Annexe B).

## Chapitre 4

# Étude de quelques plans d'expériences incomplets

*L'objet de ce chapitre est d'étudier certains plans factoriels particuliers, assez courants dans la pratique statistique, notamment industrielle. Ces plans d'expériences particuliers étant nombreux, nous avons du faire des choix qui, comme toujours, comportent un certain arbitraire. Nous détaillerons successivement la méthode des blocs, les plans en carrés latins et gréco-latins et les plans à plusieurs facteurs à deux niveaux (de type Plackett et Burman). Dans un dernier paragraphe, nous donnerons quelques indications sur d'autres dispositifs expérimentaux également courants.*

*La principale référence bibliographique de ce chapitre est l'ouvrage de John (1998). Un autre ouvrage de référence, très complet sur les plans d'expériences et rédigé en français, est celui de Dreesbeke et al. (1997). Enfin, on trouvera divers compléments intéressants dans les ouvrages de Azaïs & Bardet (2005), de Bergonzini & Duby (1995) de Goupy & Creighton (2006) et de Saporta (2006).*

### Résumé

La méthode des blocs consiste à introduire un facteur exogène (autre que le(s) facteur(s) d'intérêt), appelé le facteur bloc. Chaque bloc pris en compte (autrement dit chaque niveau du facteur bloc) est une "unité expérimentale" supposée homogène relativement à l'expérimentation considérée (des exemples sont donnés dans le premier paragraphe). Ce type de dispositif a pour avantage de permettre de contrôler une partie de la variabilité résiduelle du modèle étudié (donc de son erreur).

Les plans en carrés latins et en carrés gréco-latins sont des plans factoriels fractionnaires (plans factoriels incomplets dans lesquels on n'observe qu'une fraction de l'ensemble des cellules du plan). Ils sont définis par le croisement de trois facteurs (cas des carrés latins) ou de quatre facteurs (cas des carrés gréco-latins), ces facteurs ayant nécessairement tous le même nombre de niveaux. Cette particularité permet d'obtenir des propriétés intéressantes avec un nombre restreint d'observations.

Dans les plans à plusieurs facteurs à deux niveaux, on peut étudier un grand nombre de facteurs (jusqu'à dix et plus) avec peu d'observations (8, 12, 16...), grâce à des dispositifs expérimentaux particuliers, à savoir des plans (très) incomplets, équilibrés et orthogonaux : il s'agit des plans de Plackett et Burman qui permettent de faire des tests (essentiellement sur les effets principaux, mais pas uniquement) et d'estimer les paramètres des modèles retenus.

La mise en œuvre de ces plans peut se faire, de façon standard, au moyen de la procédure GLM du logiciel SAS.

## 4.1 La méthode des blocs

### 4.1.1 Principes

#### Objectif

L'idée essentielle à l'origine de la méthode des blocs est de *réduire la variabilité résiduelle* d'un modèle pour un plan factoriel. En effet, même dans le cas très simple d'une analyse de variance à un seul facteur avec un plan complet et équilibré, si le nombre de répétitions au sein de chaque niveau du facteur est assez élevé (disons supérieur à 5), il peut se faire que les valeurs de la variable réponse soit relativement dispersées, même pour un niveau fixé du facteur, ce qui va entraîner une estimation de la variance de l'erreur du modèle ( $\hat{\sigma}^2$ ) assez importante. Cela aura alors les conséquences suivantes :

- la statistique du test de l'effet du facteur (statistique de Fisher) aura un grand dénominateur et sera donc relativement petite : il sera ainsi assez difficile de mettre en évidence cet effet ;
- les erreurs-types des estimateurs des paramètres du modèle (proportionnelles à  $\hat{\sigma}$ ) seront grandes, de sorte que les estimations correspondantes seront peu précises.

L'introduction d'un facteur bloc, facteur exogène, a pour but de diminuer la variabilité résiduelle du modèle considéré (donc  $\hat{\sigma}^2$ ) et d'atténuer ainsi les problèmes évoqués ci-dessus.

#### Définition

On appelle *bloc* une condition expérimentale homogène relativement au phénomène étudié (le phénomène mesuré par la variable réponse  $Y$ ). L'idée est donc de constituer un certain nombre, noté  $B$  ( $B \geq 2$ ), de blocs distincts ; dans chacun d'eux, on observe la variable réponse pour différents niveaux du (des) facteur(s). En fait, l'ensemble des blocs considérés constitue un facteur supplémentaire.

Cette définition étant un peu vague, nous la précisons en donnant quelques exemples.

#### Exemples

- En agronomie, si l'on souhaite comparer les rendements (variable réponse) de différentes variétés d'une culture donnée (le facteur est la variété) dans une certaine zone géographique, un bloc sera une parcelle de terrain, homogène relativement aux facteurs non contrôlés susceptibles d'influencer le rendement : altitude, ensoleillement, humidité... Ainsi, un bloc-parcelle devra être sensiblement situé à la même altitude, avoir partout la même exposition, la même humidité...
- En médecine, si l'on souhaite étudier l'effet de différents médicaments (le facteur est le type de médicament) sur la guérison d'une maladie donnée (la variable réponse est le taux de guérison parmi les patients atteints de cette maladie), un bloc sera un ensemble de malades homogènes selon, par exemple, les facteurs âge et sexe (autrement dit un ensemble de malades de même sexe et appartenant à la même tranche d'âge).
- Dans une expérience de laboratoire se traduisant par une mesure très précise (la variable réponse) faite par un opérateur utilisant une certaine machine, un bloc sera constitué par un couple opérateur-machine.

#### Notion de randomisation

En statistique, la notion de *randomisation* (ce terme anglais est passé dans la langue française et signifie *répartition au hasard*) est très importante dans les plans d'expériences. Elle n'est pas propre à la méthode des blocs, même si elle est fondamentale dans ce cas. Son principe est de répartir au hasard les traitements qui seront administrés aux différents individus d'un même bloc. Nous précisons la notion de randomisation dans les différents plans d'expériences étudiés par la suite.

### 4.1.2 Plans en blocs complets équilibrés

Nous nous intéressons ici au cas d'un seul facteur  $F$ , comportant  $J$  niveaux ( $J \geq 2$ ) indicés par  $j$ . Toutefois, ce qui est décrit dans ce cas s'applique de la même manière dans le cas de plusieurs

facteurs croisés.

### Principe

On constitue  $B$  blocs au sein desquels chacun des niveaux du facteur est observé une fois et une seule. Chaque niveau est ainsi observé  $B$  fois, de sorte que, pour le seul facteur  $F$ , le plan est un plan complet, équilibré, avec  $B$  répétitions. De son côté, le plan croisant le facteur  $F$  et le facteur bloc est un plan complet, équilibré, sans répétition.

**Exemple 7** *Considérons un facteur  $F$  à 3 niveaux ( $J = 3$ ) pour lequel on a constitué 5 blocs ( $B = 5$ ). Nous donnons, dans le tableau ci-dessous, les 15 observations réalisées.*

niveaux du facteur $F \rightarrow$		1	2	3
<i>Blocs</i>	1	6	5	8
	2	10	9	12
	3	5	4	8
	4	9	7	10
	5	10	7	11

**Remarque 22** *De même qu'on peut envisager de considérer plusieurs facteurs croisés et d'observer une fois et une seule chaque cellule dans chaque bloc, il est aussi possible d'observer plusieurs fois chaque niveau de  $F$  dans chaque bloc (le nombre d'observations de chaque niveau dans chaque bloc restant toujours le même pour conserver un plan équilibré). Toutefois, cela augmente d'autant le nombre total d'observations, donc le coût de l'expérimentation.*

### Modèle

Indiquons par  $b$  les  $B$  blocs considérés et notons  $Y_{bj}$  la v.a.r. réponse dans le bloc  $b$ , au niveau  $j$  de  $F$ . Nous utilisons le paramétrage centré et nous écrivons le modèle sous la forme

$$Y_{bj} = \mu + \alpha_b^1 + \alpha_j^2 + U_{bj}$$

avec  $U_{bj} \sim \mathcal{N}(0, \sigma^2)$ , les  $U_{bj}$  étant indépendantes. Comme d'habitude, les paramètres vérifient les contraintes :  $\sum_{b=1}^B \alpha_b^1 = \sum_{j=1}^J \alpha_j^2 = 0$ .

Il s'agit en fait d'un modèle à 2 facteurs croisés (le facteur bloc et le facteur  $F$ ), sans interaction. En général, on ne considère pas les effets d'interactions entre le facteur et les blocs, d'une part car ils n'ont pas de signification concrète, d'autre part car il ne serait pas possible d'estimer la variance du modèle (et donc de faire des tests) dans un modèle avec interactions (qui serait un modèle saturé, c'est-à-dire comportant autant de paramètres que d'observations).

**Remarque 23** *Dans le modèle ci-dessus, on notera qu'il n'y a pas d'indice  $i$ , indice des individus dans la notation utilisée dans tout ce cours. Cela tient au fait qu'il n'y a, dans ce type de plan, qu'une seule observation par bloc et par niveau du facteur.*

### Randomisation

Dans le cas d'un plan en blocs à un seul facteur, complet et équilibré, la randomisation consiste, dans chaque bloc, à tirer au hasard les niveaux du facteur auquel seront observés les différents individus.

#### 4.1.3 Plans en blocs incomplets équilibrés

Toujours dans le cas d'un seul facteur et d'un dispositif en blocs, si le nombre de niveaux  $J$  du facteur est assez élevé, ainsi que le nombre  $B$  de blocs, un plan complet équilibré peut être trop coûteux pour l'expérimentateur. D'où l'idée, dans ce cas, de considérer un plan incomplet, mais toujours équilibré. Par construction, ces plans vont nécessairement vérifier diverses contraintes.

### Principe

- Le nombre  $J$  de niveaux du facteur  $F$  doit, dans ce cas, être au moins égal à 3 :  $J \geq 3$  (nous en précisons plus loin la raison, mais, de toutes façons, ce type de plan est conçu pour un facteur avec de nombreux niveaux).
- Le nombre de blocs considéré doit, de son côté, être au moins égal à  $J$  :  $B \geq J$ .
- Dans chaque bloc, on n'observe pas tous les niveaux de  $F$  (sinon, on aurait affaire à un plan complet), mais seulement un nombre restreint  $K$ , toujours le même pour équilibrer l'expérience; comme il est nécessaire d'observer au moins deux niveaux distincts de  $F$  dans chaque bloc (sinon, il y aurait confusion des effets de  $F$  et des effets des blocs), il vient :  $2 \leq K < J$  (on voit ainsi pourquoi il est nécessaire d'avoir  $J \geq 3$ ). Le nombre total d'observations est donc  $n = BK$ .
- Chaque niveau  $j$  du facteur  $F$  est observé dans le même nombre  $R$  de blocs ( $R$  est donc le nombre d'observations réalisées pour chaque niveau du facteur, autrement dit le nombre de répétitions : le plan est ainsi équilibré). Il vient  $2 \leq R < B$  et le nombre total d'observations vérifie  $n = JR = BK$  (on obtient ainsi  $B \geq 3$ ; la condition plus restrictive  $B \geq J$  provient de considérations sur le rang de la matrice d'incidence; voir, par exemple, John, 1998).

Pour obtenir un maximum d'homogénéité du plan considéré, on lui impose, en plus, la condition suivante :

- Chaque couple  $(j, j')$  de niveaux de  $F$  ( $j \neq j'$ ) est observé dans le même nombre  $L$  de blocs. Cette condition entraîne l'égalité suivante :  $BK(K-1) = LJ(J-1)$ . En effet :
  - . le nombre total de couples de niveaux de  $F$  est  $\frac{J(J-1)}{2}$ ;
  - . le nombre de "places" nécessaires dans l'ensemble des blocs est donc :  $L \frac{J(J-1)}{2}$ ;
  - . le nombre de couples expérimentables dans un bloc est :  $C_K^2 = \frac{K(K-1)}{2}$ ;
  - . le nombre total de couples expérimentables est donc :  $B \frac{K(K-1)}{2}$ ;
  - . on en déduit :  $BK(K-1) = LJ(J-1)$ .

### Conséquences

Comme on a vu  $n = JR = BK$ , on déduit de l'égalité ci-dessus :  $L = R \frac{K-1}{J-1}$ . Par conséquent, pour un nombre  $J$  de niveaux de  $F$  donné, les entiers  $B$ ,  $K$  et  $R$  doivent être choisis de telle sorte que  $B \geq J$ , que  $JR = BK$  et que  $R \frac{K-1}{J-1}$  soit entier.

Ainsi, les dispositifs en blocs incomplets équilibrés, caractérisés par les 5 entiers  $J$ ,  $B$ ,  $K$ ,  $R$  et  $L$ , répondent à des règles précises de construction et sont, pour cette raison, en nombre limité.

**Exemple 8** L'exemple ci-dessous correspond au cas  $J = 5$ ,  $B = 10$ ,  $K = 3$ ,  $R = 6$ ,  $L = 3$  et  $n = 30$ .

niveaux de $F \rightarrow$		1	2	3	4	5
<i>blocs</i>	1	52	51	60	-	-
	2	56	61	-	61	-
	3	49	54	-	-	65
	4	46	-	56	55	-
	5	48	-	53	-	52
	6	44	-	-	52	57
	7	-	45	51	51	-
	8	-	46	52	-	52
	9	-	47	-	50	50
	10	-	-	29	29	30

Pour des valeurs "raisonnables" du nombre  $J$  de niveaux du facteur  $F$  considéré, on peut lister les caractéristiques de tous les plans en blocs incomplets et équilibrés qu'il est possible de construire. Ainsi, pour les cinq premières valeurs possibles de  $J$ , nous donnons, dans le tableau ci-après, les caractéristiques de tous ces plans.

$J$	$B$	$K$	$R$	$L$	$n$
3	3	2	2	1	6
4	4	3	3	2	12
4	6	2	3	1	12
5	5	4	4	3	20
5	10	2	4	1	20
5	10	3	6	3	30
6	6	5	5	4	30
6	10	3	5	2	30
6	15	2	5	1	30
6	15	4	10	6	60
6	20	3	10	4	60
7	7	3	3	1	21
7	7	4	4	2	28
7	7	6	6	5	42
7	21	2	6	1	42

Pour les structures de plans avec  $J \geq 8$ , on se reportera aux tables de Fisher & Yates (1963).

### Modèle

Indiquons encore par  $j$  chaque niveau du facteur et par  $b$  chaque bloc. La v.a.r. réponse  $Y_{bj}$  n'est observée que si le couple  $(b, j)$  est tel que le niveau  $j$  est expérimenté dans le bloc  $b$ . On écrit alors, toujours avec le paramétrage centré,

$$Y_{bj} = \mu + \alpha_b^1 + \alpha_j^2 + U_{bj},$$

avec  $U_{bj} \sim \mathcal{N}(0, \sigma^2)$ , les  $U_{bj}$  étant indépendantes. Autrement dit, le modèle est le même que dans le cas complet.

### Randomisation

Encore une fois, la randomisation est assez systématique dans ce genre de situation expérimentale. Elle se fait, ici, à trois niveaux :

- les libellés des traitements (autrement dit les colonnes du tableau de l'exemple 8) sont tirés au hasard ;
- les configurations des lignes sont affectées aléatoirement aux blocs ;
- dans chaque bloc, on tire au hasard le traitement subi par chaque individu.

**Remarque 24** *Pour simplifier l'exposé, on n'a considéré, dans ce paragraphe, qu'un seul facteur  $F$ . Dans la pratique des plans d'expériences, on rencontre fréquemment des dispositifs en blocs incomplets et équilibrés associés à plusieurs facteurs. Cela complique le dispositif sans en changer le principe.*

**Remarque 25** *Nous verrons, au chapitre 6, les modèles à effets aléatoires et les modèles mixtes. On peut, d'une part utiliser la méthode des blocs avec un modèle mixte, d'autre part considérer, dans certains cas, que le facteur bloc est lui-même un facteur à effets aléatoires.*

## 4.2 Les plans en carrés latins et gréco-latins

Il s'agit de plans factoriels à trois facteurs croisés (cas des carrés latins) ou à quatre facteurs croisés (cas des carrés gréco-latins), incomplets, équilibrés et correspondant au cas particulier où tous les facteurs ont le même nombre de niveaux.

### 4.2.1 Les plans en carrés latins

#### Contexte

On considère un modèle d'ANOVA à trois facteurs croisés, dans lequel chacun des facteurs possède le même nombre de niveaux, noté  $J$ . On suppose qu'il n'y a pas d'effets d'interactions dans le modèle ou, du moins, que ces effets ne sont pas d'un intérêt majeur pour l'expérimentateur (en effet, il ne sera pas possible de les prendre en compte dans le modèle). On ne va donc considérer que des modèles additifs et ne s'intéresser qu'aux effets principaux.

#### Principe

Dans chaque cellule correspondant au croisement de deux niveaux des deux premiers facteurs, on réalise une observation et une seule pour un niveau déterminé du troisième facteur. On a ainsi  $J^2$  cellules observées parmi les  $J^3$  possibles, et une seule observation dans chaque cellule observée : c'est un plan incomplet fractionnaire (de fraction  $1/J$ , voir la remarque 27).

Dans les plans en carrés latins, le choix du niveau du troisième facteur couplé avec chacune des cellules obtenues par croisement des deux premiers facteurs se fait au moyen d'une table appelée **carré latin**. Le principe d'une telle table est illustré dans l'exemple ci-dessous ; il s'agit d'avoir une fois et une seule chacun des niveaux de  $F_3$  dans chaque ligne et dans chaque colonne de la table croisant  $F_1$  et  $F_2$ , de telle sorte que le plan obtenu soit équilibré.

**Exemple 9** *Considérons le cas  $J = 4$ . La table ci-dessous est un carré latin.*

niveaux de $F_2 \rightarrow$	1	2	3	4	
1	a	b	c	d	
niveaux de $F_1$	2	b	c	d	a
3	c	d	a	b	
4	d	a	b	c	

Les lettres contenues dans la table représentent les niveaux du facteur  $F_3$  ; on notera que les trois facteurs jouent des rôles symétriques.

Par permutation des lignes (sauf la première, qui représente un codage arbitraire des niveaux de  $F_3$ ) et permutation des colonnes, on génère, à partir du carré ci-dessus, un sous-ensemble (de cardinal  $3! \times 4! = 144$ ) de tous les carrés latins d'ordre quatre. Il existe quatre "modèles" de tables non équivalentes par permutations des lignes et des colonnes, donc 576 carrés latins d'ordre quatre au total (pour plus de détails, on pourra se reporter aux tables de Fisher & Yates, déjà mentionnées).

**Remarque 26** *La dénomination de carré latin provient du fait que, pour symboliser un tel plan d'expériences, on utilise un carré dans lequel figure des lettres latines.*

**Remarque 27** *En fait, chacun des trois facteurs considérés ayant  $J$  niveaux, le plan complet comprend  $J^3$  cellules dont on n'observe seulement que la fraction  $\frac{1}{J}$ , soit  $J^2$  observations. Ce type de plans, croisant plusieurs facteurs ayant le même nombre de niveaux (souvent 2 ou 3), dans lesquels on n'observe qu'une fraction donnée des cellules observables, s'appelle un plan fractionnaire.*

**Remarque 28** *Jusqu'à présent, nous n'avons fait que décrire les carrés latins. De façon plus précise, un carré latin est un plan d'expériences qui vérifie les caractéristiques suivantes :*

1. le plan croise trois facteurs, à  $J$  niveaux chacun ( $J \geq 2$ ), ces facteurs jouant des rôles symétriques ;
2. le plan factoriel est incomplet, fractionnaire, de fraction  $\frac{1}{J}$  ;
3. chaque niveau de chaque facteur est observé exactement  $J$  fois (de sorte qu'il s'agit d'un plan équilibré) ;

4. chaque niveau de chaque facteur est observé simultanément une fois et une seule avec chaque niveau de chacun des deux autres facteurs : c'est cette dernière propriété qui nécessite certaines précautions dans la construction d'un plan en carré latin et qui explique l'usage de tables comme celle donnée dans l'exemple ci-dessus.

Ainsi, la table ci-dessous ne constitue pas un carré latin, dans la mesure où elle permet de vérifier les propriétés 1, 2 et 3, mais pas la propriété 4 :

niveaux de $F_2 \rightarrow$		1	2	3	4
	1	a	a	a	a
niveaux de $F_1$	2	b	b	b	b
	3	c	c	c	c
	4	d	d	d	d

**Remarque 29** Enfin, signalons qu'un Sudoku est un carré latin d'ordre 9, dans lequel on a remplacé les lettres latines par les chiffres arabes de 1 à 9 et dans lequel on a rajouté une contrainte (n'ayant rien à voir avec les plans d'expériences) sur les 9 sous-carrés d'ordre 3 disjoints qui le composent. Le but du jeu est de retrouver l'intégralité d'un tel carré à partir d'une partie restreinte qui en est donnée.

**Remarque 30** En guise d'exercice sur les carrés latins, on trouvera en Annexe C un jeu mathématique paru dans le quotidien "Le Monde".

### Modèle

Comme indiqué plus haut, seul le modèle additif est pris en compte dans ce type de plan. En utilisant le paramétrage centré, ce modèle s'écrit sous la forme

$$Y_{jkl} = \mu + \alpha_j^1 + \alpha_k^2 + \alpha_l^3 + U_{jkl},$$

avec  $\sum_{j=1}^J \alpha_j^1 = \sum_{k=1}^J \alpha_k^2 = \sum_{l=1}^J \alpha_l^3 = 0$ ,  $U_{jkl} \sim \mathcal{N}(0, \sigma^2)$ , les  $U_{jkl}$  étant des v.a.r. indépendantes. Comme dans les modèles considérés dans le paragraphe 4.1, il n'y a pas d'indice  $i$  ici car on ne fait qu'une seule observation par cellule observée.

### Tests des effets

Compte tenu de la structure particulière du plan dans ce cas, la décomposition de la somme des carrés totale est elle-même particulière. Cette décomposition est explicitée dans le tableau d'analyse de la variance ci-dessous (elle est analogue à celle obtenue, au chapitre précédent, pour les plans complets et équilibrés, dans le cas d'un modèle additif).

sources de variation	sommes des carrés	d.d.l.	carrés moyens	valeurs des statistiques de Fisher
$F_1$	$SSF_1 = J \sum_{j=1}^J (\bar{y}_{j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	$J - 1$	$MSF_1 = \frac{SSF_1}{J - 1}$	$\frac{MSF_1}{MSE}$
$F_2$	$SSF_2 = J \sum_{k=1}^J (\bar{y}_{\bullet k \bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	$J - 1$	$MSF_2 = \frac{SSF_2}{J - 1}$	$\frac{MSF_2}{MSE}$
$F_3$	$SSF_3 = J \sum_{l=1}^J (\bar{y}_{\bullet\bullet l} - \bar{y}_{\bullet\bullet\bullet})^2$	$J - 1$	$MSF_3 = \frac{SSF_3}{J - 1}$	$\frac{MSF_3}{MSE}$
Erreur	$SSE$	$(J - 1)(J - 2)$	$MSE = \frac{SSE}{(J - 1)(J - 2)} = \hat{\sigma}^2$	—
Total	$SST = \sum_A (y_{jkl} - \bar{y}_{\bullet\bullet\bullet})^2$	$J^2 - 1$	—	—

Dans le tableau ci-dessus, la somme des carrés relative aux erreurs s'écrit :

$$SSE = \sum_A [y_{jkl} - (\bar{y}_{j\bullet\bullet} + \bar{y}_{\bullet k\bullet} + \bar{y}_{\bullet\bullet\ell} - 2\bar{y}_{\bullet\bullet\bullet})]^2$$

( $A$  désigne le sous-ensemble de  $J^3$  correspondant aux indices  $(j, k, \ell)$  observés).

Les estimations des paramètres (lorsque les effets correspondants sont significatifs) sont données ci-dessous :

$$\hat{\mu} = \bar{y}_{\bullet\bullet\bullet} = \frac{1}{J^2} \sum_A y_{jkl} ; \quad \hat{\alpha}_j^1 = \bar{y}_{j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}, \text{ avec } \bar{y}_{j\bullet\bullet} = \frac{1}{J} \sum_{k=1}^J y_{jkl}$$

(pour un  $j$  et un  $k$  fixés,  $\ell$  est aussi fixé) ;

$$\hat{\alpha}_j^2 = \bar{y}_{\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet} ; \quad \hat{\alpha}_j^3 = \bar{y}_{\bullet\bullet\ell} - \bar{y}_{\bullet\bullet\bullet}.$$

### Randomisation

Encore une fois, elle est quasiment systématique dans ce type de modèles. Elle se pratique à plusieurs niveaux :

- parmi les trois facteurs dont on dispose, on peut tirer au sort celui qui sera en lignes, celui qui sera en colonnes et celui qui sera le troisième ;
- on peut aussi tirer au sort les étiquettes des niveaux de chaque facteur ;
- partant d'un carré latin donné, on peut tirer au sort une permutation sur les lignes et une autre sur les colonnes ;
- enfin, on peut aussi tirer au sort l'affectation de chaque individu à une cellule.

#### 4.2.2 Les plans en carrés gréco-latins

On les appelle également plans en *carrés latins orthogonaux*.

#### Contexte

Il s'agit maintenant d'une ANOVA à quatre facteurs croisés, chaque facteur possédant toujours le même nombre de niveaux noté  $J$ . Là encore, on ne fera intervenir aucune interaction dans les modèles envisagés qui seront donc tous additifs.

#### Principe

Sur les  $J^4$  cellules envisageables, seules  $J^2$  sont observées et elles le sont une seule fois (on considère donc la fraction  $\frac{1}{J^2}$  du plan complet : c'est encore un plan fractionnaire). En fait, si l'on désigne par  $F_1, F_2, F_3$  et  $F_4$  les quatre facteurs considérés,  $(F_1, F_2, F_3)$  et  $(F_1, F_2, F_4)$  constituent chacun un carré latin, les différents niveaux de  $F_3$  et de  $F_4$  étant couplés une fois et une seule, de telle sorte que ces deux carrés latins soient "orthogonaux".

**Exemple 10** *Les carrés gréco-latins sont plus délicats à construire que les carrés latins. Nous donnons ci-dessous un exemple de carré gréco-latin d'ordre quatre ( $J = 4$ ).*

niveaux de $F_2 \rightarrow$	1	2	3	4	
1	$a \alpha$	$b \beta$	$c \gamma$	$d \delta$	
niveaux de $F_1$	2	$c \beta$	$d \alpha$	$a \delta$	$b \gamma$
3	$d \gamma$	$c \delta$	$b \alpha$	$a \beta$	
4	$b \delta$	$a \gamma$	$d \beta$	$c \alpha$	

Les lettres latines contenues dans la table représentent les niveaux du facteur  $F_3$ , tandis que les lettres grecques représentent ceux du facteur  $F_4$ .

On notera encore que les quatre facteurs jouent des rôles symétriques et que, par permutation des lignes et des colonnes, on génère une partie (un tiers) de l'ensemble de tous les carrés gréco-latins d'ordre quatre (sur lesquels on trouvera encore d'autres précisions dans les tables de Fisher & Yates).

**Modèle**

Le modèle (additif) pour un plan en carré gréco-latin s'écrit, avec le paramétrage centré, sous la forme suivante :

$$Y_{jklm} = \mu + \alpha_j^1 + \alpha_k^2 + \alpha_l^3 + \alpha_m^4 + U_{jklm},$$

avec  $\sum_{j=1}^J \alpha_j^1 = \sum_{k=1}^J \alpha_k^2 = \sum_{l=1}^J \alpha_l^3 = \sum_{m=1}^J \alpha_m^4 = 0$ ,  $U_{jklm} \sim \mathcal{N}(0, \sigma^2)$ , les  $U_{jklm}$  étant des v.a.r. indépendantes.

Test de significativité du modèle ci-dessus, tests relatifs à chacun des effets principaux, estimation des paramètres dans le modèle retenu et randomisation généralisent de façon naturelle ce qui a été explicité dans le cas des carrés latins.

**Remarque 31** *De façon claire, la dénomination de carré gréco-latin provient du fait que, pour symboliser un tel plan d'expériences, on utilise un carré dans lequel figure des lettres latines et des lettres grecques.*

**Remarque 32** *Ici encore, on peut préciser la définition des carrés gréco-latins : un carré gréco-latin est un plan d'expériences qui vérifie les caractéristiques suivantes :*

1. *le plan croise quatre facteurs, à  $J$  niveaux chacun ( $J \geq 3$ ), ces facteurs jouant des rôles symétriques (on vérifie, de façon immédiate, qu'il n'existe pas de carré gréco-latin d'ordre 2);*
2. *le plan factoriel est incomplet, fractionnaire, de fraction  $\frac{1}{J^2}$ ;*
3. *chaque niveau de chaque facteur est observé exactement  $J$  fois (il s'agit encore d'un plan équilibré);*
4. *chaque niveau de chaque facteur est observé simultanément une fois et une seule avec chaque niveau de l'un quelconque des trois autres facteurs.*

**Remarque 33** *En partant d'un carré gréco-latin, si on supprime l'un quelconque des quatre facteurs, le plan d'expériences ainsi obtenu est un carré latin. Si on en supprime deux (ou si on supprime un facteur dans un carré latin), le plan obtenu est un plan à deux facteurs croisés à  $J$  niveaux chacun, complet, équilibré, sans répétition.*

**Remarque 34** *Dans un carré latin ou dans un carré gréco-latin, un (voire plusieurs) des facteurs peut être un bloc.*

**Remarque 35** *Nous introduirons au chapitre 6 les modèles mixtes faisant intervenir à la fois des effets fixes (comme dans tous les modèles envisagés jusqu'ici) et des effets aléatoires. Dans les plans en carrés latins ou en carrés gréco-latins, on peut très bien considérer que certains des trois ou quatre facteurs intervenant sont à effets aléatoires, ce qui conduit à généraliser le modèle ci-dessus à un modèle mixte.*

**Remarque 36** *Donnons quelques précisions sur l'existence des carrés latins et gréco-latins. On peut construire des carrés latins d'ordre  $J$  ( $J$  niveaux pour chaque facteur), pour toute valeur de  $J$  à partir de 2. Les tables de Fisher & Yates en donnent toutes les configurations pour  $J$  variant de 4 à 12. Concernant les carrés gréco-latins, on ne peut en construire qu'à partir de  $J = 3$ . De plus, il faut noter la particularité suivante : il n'existe pas de carré gréco-latin d'ordre six<sup>1</sup>. Les tables de Fisher & Yates en donnent toutes les configurations pour  $J$  allant 3, 4 et 5, puis pour  $J$  allant 7, 8 et 9.*

<sup>1</sup>Cette particularité remonte au mathématicien suisse Leonhard EULER (1707–1783) qui chercha vainement à disposer 36 officiers de 6 grades différents (facteur  $F_3$ ), appartenant à 6 régiments différents (facteur  $F_4$ ), dans une grille à 6 lignes (facteur  $F_1$ ) et 6 colonnes (facteur  $F_2$ ), de telle sorte que chaque grade et chaque régiment apparaisse une fois et une seule dans chaque ligne et dans chaque colonne. En 1782, Euler conjectura le résultat suivant : il n'existe pas de carré gréco-latin d'ordre  $J = 4k + 2, k \in \mathbb{N}$ . Le résultat est trivial pour  $k = 0$ , soit  $J = 2$ . Pour  $k = 1$ , soit  $J = 6$ , il n'a été démontré qu'en 1900, par Gaston Tarry, mathématicien français né à Villefranche de Rouergue. Par contre, cette conjecture est fautive pour  $k \geq 2$  et ce n'est qu'en 1960 que Bose *et al.* ont montré la possibilité de construire des carrés gréco-latins pour tout  $J > 2$ , sauf pour  $J = 6$ .

## 4.3 Les plans à plusieurs facteurs à deux niveaux

### 4.3.1 Introduction

Dans la pratique des plans d'expériences, il est fréquent d'avoir un nombre important de facteurs à prendre en compte (par exemple plus de cinq) tout en étant limité dans le nombre d'observations réalisables (problèmes de coût, de temps...).

En fait, il existe des plans d'expériences spécifiques pour répondre à ce type de situations. Pour ce faire, on se limite en général à des facteurs à deux niveaux (*bas* et *haut*) ou, éventuellement, à trois niveaux (*bas*, *intermédiaire* et *haut*). Dans ce paragraphe, nous ne traiterons que des facteurs à deux niveaux et nous noterons  $p$  le nombre total de facteurs pris en compte.

### 4.3.2 Cas $p = 2$

Il s'agit d'un cas très simple d'analyse de variance à deux facteurs croisés, chacun à deux niveaux. Nous renvoyons donc au chapitre 3 pour plus de détails. Notons seulement les deux particularités suivantes :

- les écritures des tests et des estimations se simplifient, dans la mesure où l'on a  $J = K = 2$ ; en particulier, les effets de chaque facteur étant supposés centrés, il y a un seul paramètre à estimer par facteur ;
- dans le cas d'un plan complet, équilibré, sans répétition (une seule observation par cellule, soit quatre au total), on ne peut pas faire de test si l'on estime les interactions (il y a, dans ce cas, quatre paramètres à estimer et le modèle est saturé); on doit donc supprimer les interactions si l'on veut pouvoir faire des tests.

### 4.3.3 Cas $p = 3$

On a donc ici  $J = K = L = 2$ ; pour le cas général, nous renvoyons encore au chapitre 3. Dans ce paragraphe, nous allons nous intéresser, tout d'abord, au cas d'un *plan complet, équilibré, sans répétition* (huit observations au total, soit une par cellule), puis au cas d'un plan incomplet à 4 observations.

#### Plan d'expériences complet sans répétition

Les huit observations se présentent sous la forme suivante :

$\ell \rightarrow$	1		2	
$j \quad k \rightarrow$	1	2	1	2
1	$y_{111}$	$y_{121}$	$y_{112}$	$y_{122}$
2	$y_{211}$	$y_{221}$	$y_{212}$	$y_{222}$

#### Modèle

Si l'on prend en compte l'effet général (un paramètre), les effets principaux de chacun des trois facteurs (trois paramètres en tout) et les effets d'interactions d'ordre deux (3 paramètres), avec le paramètre de variance, on arrive à huit paramètres, soit le nombre d'observations. Le modèle prenant en compte ces trois types d'effets est donc un modèle saturé et il n'est pas possible, dans ce cas, d'estimer ou de tester les interactions d'ordre trois. Ce modèle s'écrit, selon le paramétrage centré :

$$Y_{jkl} = \mu + \alpha_j^1 + \alpha_k^2 + \alpha_\ell^3 + \gamma_{jk}^{12} + \gamma_{j\ell}^{13} + \gamma_{k\ell}^{23} + U_{jkl}.$$

Il est clair qu'un modèle additif, sans aucune interaction, est plus approprié à ce dispositif expérimental.

#### Tableau d'analyse de la variance pour le modèle additif

Compte tenu que chacun des trois facteurs pris en compte n'a que deux niveaux, les moyennes partielles, servant à estimer et à tester les différents effets, ont des expressions particulières. Par

exemple, on a, par définition :

$$\bar{y}_{jk\bullet} = \frac{1}{2}(y_{jk1} + y_{jk2}) ; \quad \bar{y}_{j\bullet\bullet} = \frac{1}{4} \sum_{k=1}^2 \sum_{\ell=1}^2 y_{jk\ell} ; \quad \bar{y}_{\bullet\bullet\bullet} = \frac{1}{8} \sum_{j=1}^2 \sum_{k=1}^2 \sum_{\ell=1}^2 y_{jk\ell} = \frac{1}{2}(\bar{y}_{1\bullet\bullet} + \bar{y}_{2\bullet\bullet}).$$

On en déduit des égalités du type :

$$\bar{y}_{1\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} = \frac{1}{2}(\bar{y}_{1\bullet\bullet} - \bar{y}_{2\bullet\bullet}) = -(\bar{y}_{2\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}).$$

Pour le tableau d'analyse de la variance, on part de l'égalité suivante :

$$y_{jk\ell} - \bar{y}_{\bullet\bullet\bullet} = y_{jk\ell} - (\bar{y}_{j\bullet\bullet} + \bar{y}_{\bullet k\bullet} + \bar{y}_{\bullet\bullet\ell} - 2\bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet\bullet\ell} - \bar{y}_{\bullet\bullet\bullet}).$$

On en déduit :

$$\begin{aligned} \sum_{j=1}^2 \sum_{k=1}^2 \sum_{\ell=1}^2 (y_{jk\ell} - \bar{y}_{\bullet\bullet\bullet})^2 &= \sum_{j=1}^2 \sum_{k=1}^2 \sum_{\ell=1}^2 [y_{jk\ell} - (\bar{y}_{j\bullet\bullet} + \bar{y}_{\bullet k\bullet} + \bar{y}_{\bullet\bullet\ell} - 2\bar{y}_{\bullet\bullet\bullet})]^2 \\ &+ 4 \sum_{j=1}^2 (\bar{y}_{j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + 4 \sum_{k=1}^2 (\bar{y}_{\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + 4 \sum_{\ell=1}^2 (\bar{y}_{\bullet\bullet\ell} - \bar{y}_{\bullet\bullet\bullet})^2 \end{aligned}$$

(le plan étant équilibré, on vérifie sans difficulté la nullité des doubles produits). Cette décomposition peut se réécrire sous la forme

$$SST = SSE + SSF_1 + SSF_2 + SSF_3,$$

où  $SST$  et  $SSE$  désignent respectivement la somme des carrés totale et la somme des carrés relative aux erreurs dans la modèle additif, tandis que  $SSF_1$ ,  $SSF_2$  et  $SSF_3$  désignent respectivement les sommes des carrés relatives à chacun des trois facteurs. On notera que les égalités du type

$$SSF_1 = 4 \sum_{j=1}^2 (\bar{y}_{j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 = 2 (\bar{y}_{1\bullet\bullet} - \bar{y}_{2\bullet\bullet})^2$$

découlent des égalités signalées plus haut, ce qui permet de simplifier le calcul de ces sommes de carrés.

On obtient encore le tableau d'analyse de la variance sous la forme suivante :

sources de variation	sommes des carrés	d.d.l.	carrés moyens	valeurs des statistiques de Fisher
$F_1$	$SSF_1$	1	$SSF_1$	$\frac{SSF_1}{MSE}$
$F_2$	$SSF_2$	1	$SSF_2$	$\frac{SSF_2}{MSE}$
$F_3$	$SSF_3$	1	$SSF_3$	$\frac{SSF_3}{MSE}$
Erreur	$SSE$	4	$MSE = \frac{SSE}{4} = \hat{\sigma}^2$	—
Total	$SST$	7	—	—

**Remarque 37** Dans les généralisations à plus de trois facteurs, on garde le caractère équilibré des plans d'expériences étudiés et donc ce type de tableau de décomposition.

**Plan d'expériences incomplet, fractionnaire, équilibré**

Si l'on souhaite juste estimer les effets  $\mu$ ,  $\alpha_j^1$ ,  $\alpha_k^2$  et  $\alpha_\ell^3$ , sans faire de test ni calculer d'erreur-type, on peut, à la limite, se contenter de quatre observations (deux par niveau de chaque facteur) selon le dispositif incomplet, fractionnaire de fraction  $\frac{1}{2}$ , suivant :

observations	$F_1$	$F_2$	$F_3$
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

(les chiffres à l'intérieur de la table désignent les niveaux des facteurs à prendre en compte).

Dans l'écriture vectorielle du modèle additif (selon le paramétrage centré), sous la forme  $Y = \mathbf{X}_c \beta_c + U$ , le vecteur  $\beta_c$  a pour transposé  $\beta'_c = (\mu \ \alpha_1^1 \ \alpha_1^2 \ \alpha_1^3)$  et la matrice  $\mathbf{X}_c$  s'écrit :

$$\mathbf{X}_c = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

On notera, tout d'abord, que les trois dernières colonnes de la matrice  $\mathbf{X}_c$  correspondent au dispositif présenté au-dessus, en remplaçant la valeur 2 par la valeur  $-1$  (ces trois colonnes sont ainsi centrées). Ensuite, les colonnes de cette matrice sont deux à deux orthogonales, ceci n'étant possible que parce que le plan est équilibré (il s'agit d'une condition nécessaire, mais pas suffisante). Enfin, on notera que ce dispositif est associé à ce qu'on appelle la table  $L_4$  que nous présentons ci-dessous (1 est remplacé par + et  $-1$  par  $-$ ) et dont la version  $L_8$  sera détaillée au point suivant.

La table  $L_4$  :

	$F_1$	$F_2$	$F_3$
1	+	+	+
2	+	-	-
3	-	+	-
4	-	-	+

**4.3.4 Cas  $4 \leq p \leq 6$** 

Dans la pratique, ce cas est, bien sûr, plus intéressant que les précédents dans la mesure où l'on peut raisonnablement envisager, tant que l'on a  $p \leq 3$ , un plan complet, équilibré avec au moins deux répétitions. Au delà, c'est d'autant plus difficile que  $p$  augmente.

**Dispositif expérimental**

L'idée est ici de généraliser ce qui a été fait plus haut pour trois facteurs et seulement quatre observations. Ainsi, avec huit observations, il est possible d'étudier des modèles additifs comportant jusqu'à six facteurs explicatifs, chacun ayant toujours seulement deux niveaux. En effet, il y a un paramètre à estimer par facteur, un paramètre pour l'effet général et un paramètre pour la variance. Toutefois, le dispositif expérimental est assez spécifique (même s'il n'est pas unique, voir plus bas) et les observations se présentent, par exemple, sous la forme décrite dans le tableau ci-dessous.

observations	$F_1$	$F_2$	$F_3$	$F_4$	$F_5$	$F_6$
1	1	1	1	1	1	1
2	1	1	2	1	2	2
3	1	2	1	2	2	1
4	1	2	2	2	1	2
5	2	1	1	2	1	2
6	2	1	2	2	2	1
7	2	2	1	1	2	2
8	2	2	2	1	1	1

**Modèle**

En considérant toujours le paramétrage centré, le modèle avec six facteurs s'écrit sous la forme :

$$Y_{hijklm} = \mu + \alpha_h^1 + \alpha_i^2 + \alpha_j^3 + \alpha_k^4 + \alpha_\ell^5 + \alpha_m^6 + U_{hijklm}.$$

Sous forme matricielle, il s'écrit encore  $Y = \mathbf{X}_c \beta_c + U$ , avec  $\beta_c' = (\mu \alpha_1^1 \alpha_1^2 \alpha_1^3 \alpha_1^4 \alpha_1^5 \alpha_1^6)$  et :

$$\mathbf{X}_c = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 \end{pmatrix}.$$

On notera les particularités suivantes de la matrice  $\mathbf{X}_c$  :

- à l'exception de la première colonne, qui ne contient que des 1, les autres colonnes correspondent à celles du dispositif présenté plus haut, en remplaçant chaque valeur 2 par la valeur -1 ;
- toujours à l'exception de la première colonne, les autres colonnes de  $\mathbf{X}_c$  sont centrées (parce que le plan est équilibré) ;
- les colonnes de  $\mathbf{X}_c$  sont deux à deux orthogonales (pour obtenir ce résultat, il est nécessaire, mais pas suffisant, que le plan soit équilibré).

**Estimations et tests**

Dans le modèle avec six facteurs, l'estimation de l'effet général  $\mu$  est  $\hat{\mu} = \bar{y}$  (moyenne des huit observations).

L'estimation de  $\alpha_1^s$  ( $s = 1, \dots, 6$ ) est :

$$\hat{\alpha}_1^s = \bar{y}_1^s - \bar{y} = \frac{\bar{y}_1^s - \bar{y}_2^s}{2},$$

où  $\bar{y}_1^s$  (respectivement  $\bar{y}_2^s$ ) est la moyenne des quatre observations réalisées au niveau 1 (respectivement au niveau 2) de  $F_s$  et où  $\bar{y}$  vérifie  $\bar{y} = \frac{\bar{y}_1^s + \bar{y}_2^s}{2}$ .

Enfin, l'estimation de  $\sigma^2$  vérifie

$$\hat{\sigma}^2 = \|\hat{U}\|^2 = \|Y - \mathbf{X}_c \hat{\beta}_c\|^2, \quad \text{avec } \hat{\beta}_c' = (\hat{\mu} \hat{\alpha}_1^1 \hat{\alpha}_1^2 \hat{\alpha}_1^3 \hat{\alpha}_1^4 \hat{\alpha}_1^5 \hat{\alpha}_1^6)$$

(il n'y a pas de dénominateur car le degré de liberté de  $\hat{U}$  est 1).

Pour le test de significativité du facteur  $F_s$  ( $s = 1, \dots, 6$ ), équivalent au test de l'hypothèse nulle  $\{H_0 : \alpha_1^s = 0\}$ , la valeur de la statistique du test est :

$$f = \frac{2 (\bar{y}_1^s - \bar{y}_2^s)^2}{\hat{\sigma}^2}.$$

**Base orthogonale de  $\mathbb{R}^8$** 

Considérons maintenant l'espace vectoriel réel  $\mathbb{R}^8$  muni de la base canonique  $\mathcal{C}$  et de la métrique identité (associée à la matrice identité d'ordre 8 relativement à  $\mathcal{C}$ ). Considérons d'autre part la matrice  $\mathbf{P}$  ci-dessous, carrée d'ordre 8 :

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix}.$$

On peut vérifier que les colonnes de  $\mathbf{P}$  sont deux à deux orthogonales. Il s'ensuit que  $\mathbf{P}$  est régulière, est donc une matrice de passage (de la base canonique à une nouvelle base, notée  $\mathcal{B}$ ), que les coordonnées de cette nouvelle base sur  $\mathcal{C}$  sont fournies par les colonnes de  $\mathbf{P}$  et que  $\mathcal{B}$  est une base orthogonale de  $\mathbb{R}^8$ . On notera encore que  $\mathbf{P}$  est obtenue en rajoutant une colonne orthogonale à la matrice  $\mathbf{X}_c$  définie plus haut et que ses colonnes sont centrées, à l'exception de la première. On obtient ainsi la propriété suivante.

**Propriété 1** *Tout plan qui expérimente six facteurs avec huit observations, de telle sorte que la matrice d'incidence associée soit constituée du premier vecteur-colonne de  $\mathbf{P}$  et de six autres de ses vecteurs-colonnes (n'importe lesquels, d'où la non unicité du plan) permet d'étudier les effets de ces six facteurs dans les mêmes conditions qu'avec le modèle défini plus haut.*

### Notion de confusion des effets et triangle des interactions

Supposons que, dans le modèle présenté plus haut, on souhaite prendre en considération un effet d'interaction, par exemple entre les deux facteurs  $F_1$  et  $F_2$ . Cet effet sera mesuré par un paramètre noté  $\gamma^{12}$  et l'on réécrira le modèle sous la forme matricielle suivante :

$$Y = \mathbf{X}^* \beta^* + U^*, \text{ avec } \mathbf{X}^* = (\mathbf{X}_c | X^{12}) \text{ et } \beta^* = \begin{pmatrix} \beta_c \\ \gamma^{12} \end{pmatrix}.$$

De façon standard, la colonne  $X^{12}$  est obtenue par produit des colonnes  $X^1$  et  $X^2$  et vaut :

$$X^{12} = X^1 \times X^2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} = X^4,$$

colonne associée au facteur  $F_4$ . Ainsi, on ne pourra estimer dans ce modèle ni  $\alpha_1^4$  (effet de  $F_4$ ) ni  $\gamma^{12}$  (effet d'interaction entre  $F_1$  et  $F_2$ ), mais seulement la somme  $\alpha_1^4 + \gamma^{12}$ . On dit qu'il y a confusion de ces deux effets. Avec le plan à seulement 8 observations considéré ici, il faut donc choisir, entre ces deux effets, celui qui sera introduit dans le modèle.

Pour savoir à quel facteur correspond toute interaction d'ordre deux dans ce modèle, on dresse un tableau, appelé *triangle des interactions*, qui fournit le résultat. Dans le tableau ci-dessous, on a considéré 7 facteurs (voir la remarque 39). De plus, pour une meilleure lisibilité, on a remplacé les facteurs  $F_1$  à  $F_7$  par les lettres de  $a$  à  $g$ . Pour la même raison, on a fait figurer le tableau entier (un carré) plutôt que sa partie supérieure ou inférieure (un triangle).

	$a$	$b$	$c$	$d$	$e$	$f$	$g$
$a$	—	$d$	$f$	$b$	$g$	$c$	$e$
$b$	$d$	—	$e$	$a$	$c$	$g$	$f$
$c$	$f$	$e$	—	$g$	$b$	$a$	$d$
$d$	$b$	$a$	$g$	—	$f$	$e$	$c$
$e$	$g$	$c$	$b$	$f$	—	$d$	$a$
$f$	$c$	$g$	$a$	$e$	$d$	—	$b$
$g$	$e$	$f$	$d$	$c$	$a$	$b$	—

La lecture du tableau est très simple : il y a confusion de l'effet d'interaction entre  $a$  et  $b$  et de l'effet du facteur  $d$ ; de même pour l'interaction entre  $c$  et  $e$  et le facteur  $b$ ...

**Remarque 38 : Où l'algèbre rejoint la statistique...** *Considérons l'ensemble*

$$H = \{1, a, b, c, d, e, f, g\}$$

*muni d'une opération interne, notée  $*$ , dont le tableau ci-dessus est la table, et rajoutons lui les propriétés suivantes :*

- 1 est l'élément neutre de  $*$  :  $\forall x \in H, 1 * x = x * 1 = x$  ;
- le produit de tout élément avec lui-même est égal à 1 :  $\forall x \in H, x * x = 1$ .

Alors,  $H$  est un groupe commutatif pour l'opération  $*$  (immédiat).

Nous ne développerons pas davantage les liens entre plans d'expériences et algèbre, mais il est clair que l'étude des propriétés mathématiques des plans d'expériences nécessite un large usage de l'algèbre (voir, par exemple, Collombier, 1996).

#### La table $L_8$

On appelle ainsi la table à 8 lignes et 7 colonnes obtenue à partir de la matrice  $\mathbf{P}$  définie plus haut en supprimant la première colonne (qui ne correspond pas à un facteur) et en remplaçant  $+1$  par  $+$  (niveau *haut* du facteur) et  $-1$  par  $-$  (niveau *bas* du facteur). Elle indique comment expérimenter de 4 à 6 facteurs à deux niveaux (en supprimant trois, deux ou une colonne) avec seulement 8 observations. Nous donnons ci-dessous un modèle de table  $L_8$ .

	$a$	$b$	$c$	$d$	$e$	$f$	$g$
1	+	+	+	+	+	+	+
2	+	+	-	+	-	-	-
3	+	-	+	-	-	+	-
4	+	-	-	-	+	-	+
5	-	+	+	-	+	-	-
6	-	+	-	-	-	+	+
7	-	-	+	+	-	-	+
8	-	-	-	+	+	+	-

Il existe différentes tables  $L_8$ , équivalentes du point de vue de l'expérimentation. L'une d'entre elles est obtenue, à partir d'une ligne fixe appelée *générateur*, par un procédé systématique de permutations circulaires (ce procédé se généralisant aux tables plus importantes décrites dans le point suivant). En fait, le générateur comportant quatre signes  $-$  pour trois signes  $+$ , et les permutations circulaires d'un ensemble de sept éléments ne permettant de générer que sept lignes, on doit rajouter une ligne (la première) ne comportant que des  $+$ . Nous donnons ci-dessous cette table.

	$a$	$b$	$c$	$d$	$e$	$f$	$g$
1	+	+	+	+	+	+	+
2	-	-	-	+	-	+	+
3	+	-	-	-	+	-	+
4	+	+	-	-	-	+	-
5	-	+	+	-	-	-	+
6	+	-	+	+	-	-	-
7	-	+	-	+	+	-	-
8	-	-	+	-	+	+	-

**Remarque 39** Si l'on se contente d'estimations ponctuelles, sans faire de test ni construire d'intervalle de confiance, la table  $L_8$  indique le plan à considérer pour étudier jusqu'à 7 facteurs avec un modèle sans interaction (ou 6 facteurs avec une interaction...).

#### 4.3.5 Cas $p > 6$

Comme on a construit une table  $L_4$  et une table  $L_8$  (et même plusieurs, mais équivalentes), il est possible de construire une table  $L_{12}$ , une table  $L_{16}$ ... L'indice de ces tables est le nombre total d'observations réalisées ( $n = 4$ ;  $n = 8$ ;  $n = 12$ ;  $n = 16$ ...) et est un multiple de 4 :  $n = 4 \times m$ ,  $m \in \mathbb{N}^*$ . La valeur 4 correspond au carré du nombre de niveaux considérés (2) et permet de conserver des plans équilibrés lorsqu'on augmente le nombre d'observations, ainsi que le caractère orthogonal des colonnes de  $\mathbf{X}_c$ .

La table  $L_{12}$  permet de construire un plan avec lequel on peut étudier jusqu'à dix facteurs sans interaction (onze si l'on ne fait que des estimations). Le triangle des interactions se généralise et

permet d'introduire certaines interactions, à condition de supprimer les facteurs correspondant à la confusion des effets. De même, la table  $L_{16}$  permet d'étudier jusqu'à 14 facteurs...

Tous ces plans sont appelés des *plans de Plackett et Burman*, car ils ont été introduits pour la première fois par ces deux auteurs (Plackett & Burman, 1946). Ils présentent tous la particularité d'être des *plans incomplets, équilibrés et orthogonaux* :

- ils sont incomplets, puisque  $p$  facteurs à 2 niveaux nécessitent un minimum de  $2^p$  observations pour obtenir un plan complet et qu'on en est loin dans les exemples présentés ici ;
- ils sont équilibrés, car chaque niveau (*bas* ou *haut*) de chaque facteur est observé  $\frac{n}{2}$  fois, si  $n$  est le nombre total d'observations du plan considéré ;
- ils sont orthogonaux, car la matrice d'incidence  $\mathbf{X}_c$  associée à chacun de ces plans est orthogonale (ses colonnes sont deux à deux orthogonales).

**Remarque 40** *Les plans de Plackett et Burman étudiés dans ce paragraphe sont, pour certains d'entre eux, des plans fractionnaires. En effet, avec  $p$  facteurs comportant tous deux niveaux ( $p \geq 3$ ), un plan complet sans répétition doit comporter  $2^p$  observations. Un plan fractionnaire n'expérimentera qu'une fraction déterminée,  $\frac{1}{2^k}$ , de ces observations, de sorte que l'on fera seulement  $2^{p-k}$  observations ( $k = 1, \dots, p-2$ ). Il s'agit des plans obtenus avec les tables  $L_4, L_8, L_{16}$ ... Les plans de Plackett et Burman offrent donc plus de possibilités en proposant également des tables telles que  $L_{12}$ .*

## 4.4 Compléments

Dès l'introduction de ce chapitre, nous avons indiqué que les dispositifs expérimentaux permettant de répondre à des situations concrètes particulières sont extrêmement nombreux et que nous nous sommes volontairement contentés ici de présenter les plus courants. Nous signalons ci-dessous quelques-uns des autres dispositifs également courants.

Tout d'abord, de façon analogue à ce qui a été développé pour des facteurs à deux niveaux, on peut considérer des plans incomplets, équilibrés et orthogonaux pour plusieurs facteurs à trois niveaux. Il existe encore des tables permettant de construire les dispositifs expérimentaux correspondant. Ainsi, la table  $L_9$  permet d'étudier jusqu'à 3 facteurs à trois niveaux (et même 4, si l'on n'estime pas  $\sigma^2$  et qu'on ne fait pas de tests) avec seulement 9 observations. La table  $L_{18}$  permet d'étudier jusqu'à 8 facteurs à trois niveaux avec 18 observations. La table  $L_{27}$  permet d'étudier jusqu'à 12 facteurs à trois niveaux (éventuellement 13), avec 27 observations... De la même manière, il existe des dispositifs permettant d'étudier simultanément plusieurs facteurs à deux niveaux et plusieurs facteurs à trois niveaux.

Un autre dispositif courant dans la pratique est le plan dit en *cross-over*. Dans sa version la plus simple, ce dispositif consiste à étudier un facteur à deux niveaux (par exemple, traitement et placebo) en deux périodes. Dans la première période, une partie de l'échantillon considéré reçoit le traitement, l'autre partie recevant le placebo. Dans la deuxième période, les rôles sont inversés (autrement dit croisés, d'où le terme de *cross-over*). Comme dans le cas des blocs, on réduit ainsi la variabilité résiduelle, ce qui permet d'améliorer l'efficacité du dispositif.

Pour mémoire, signalons encore d'autres dispositifs : les plans *split plot*, ou en parcelles divisées, les plans Taguchi, les surfaces de réponses... Pour plus de détails sur ces dispositifs, on se reportera, par exemple, à Azaïs & Bardet (2005), à Droesbeke *et al.* (1997), à John (1998) ou à Saporita (2006).

## Chapitre 5

# L'analyse de variance multivariée

Le modèle linéaire gaussien standard a pour objectif de modéliser la dépendance (supposée linéaire) d'une variable aléatoire réelle ( $Y$ ) par rapport soit à d'autres variables quantitatives contrôlées (cas de la régression), soit à des facteurs (cas de l'analyse de variance), soit à un mélange des deux (cas de l'analyse de covariance). Ce modèle s'étend sans difficulté majeure au cas où la réponse n'est pas unidimensionnelle, mais multidimensionnelle : la variable aléatoire réelle  $Y$  est alors remplacée par un vecteur aléatoire.

Dans ce cours, nous avons déjà repris et développé les méthodes d'analyse de la variance ; par contre, nous ne sommes revenus ni sur les méthodes de régression ni sur les méthodes d'analyse de la covariance. De la même manière, dans le cadre du modèle linéaire gaussien multivarié, nous détaillerons uniquement la généralisation de l'analyse de variance. Pour la régression et l'analyse de covariance, le passage au cas multidimensionnel se fait de la même façon. En particulier, on trouve les mêmes tests que ceux présentés ici.

Concernant la bibliographie, l'ouvrage de référence pour ce chapitre est celui de Seber (1984). On peut également indiquer l'ouvrage de Anderson (2003) et celui de Rencher (1995).

### Résumé

Le chapitre 5 est donc consacré aux plans factoriels avec réponse multidimensionnelle, autrement dit à l'analyse de variance multivariée (ou multidimensionnelle), encore appelée MANOVA (acronyme pour *Multivariate ANalysis Of VAriance*). Seuls seront traités dans ce chapitre les cas de un facteur et de deux facteurs croisés. Dans ce contexte, nous n'aborderons pas la généralisation des intervalles de confiance et nous nous consacrerons seulement à l'estimation ponctuelle des paramètres et aux tests d'hypothèses.

Pour ce qui est de l'estimation ponctuelle des paramètres, les principes et les résultats sont de même nature que ceux vus dans le cas unidimensionnel. Toutefois, la méthode du maximum de vraisemblance se trouve compliquée par le fait que chaque observation est maintenant la réalisation d'une loi gaussienne multidimensionnelle, ce qui alourdit l'écriture de la vraisemblance et nécessite des dérivations matricielles. Par ailleurs, l'expression des paramètres est maintenant matricielle et non plus vectorielle. Ainsi, si nous notons  $D$  la dimension du vecteur aléatoire réponse  $Y$  ( $D \geq 2$ ), on retrouve l'expression habituelle des estimateurs des paramètres

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

dans laquelle  $\mathbf{Y}$  est maintenant une matrice  $n \times D$ , de sorte que  $\hat{\beta}$  est une matrice  $p \times D$  ( $n$  est le nombre total d'observations et  $p$  est le nombre de colonnes de  $X$  :  $J$  dans le cas d'un seul facteur,  $JK$  dans le cas de deux facteurs croisés, etc.).

La loi normale centrée unidimensionnelle prise jusqu'à présent comme modèle pour les erreurs avait pour variance  $\sigma^2$ . Cette variance est ici remplacée par une matrice de variances-covariances  $\Sigma$ ,  $D \times D$ , pour la loi normale centrée, multidimensionnelle d'ordre  $D$ , des erreurs. Si on pose  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$  (matrice  $n \times D$  des valeurs prédites) et  $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}}$  (matrice  $n \times D$  des résidus), la

matrice  $\Sigma$  est estimée par  $\frac{1}{n-p} \hat{U}'\hat{U}$ , où  $\hat{U}'\hat{U}$  est distribuée selon une loi de Wishart, généralisation multidimensionnelle de la loi de khi-deux.

Pour ce qui est des tests, le test de Fisher, permettant de tester différentes hypothèses nulles en ANOVA unidimensionnelle, est maintenant remplacé par plusieurs tests (quatre dans SAS) dont les statistiques sont calculées à partir des valeurs propres des deux matrices remplaçant numérateur et dénominateur de la statistique de Fisher. Les tests fournis par SAS sont les tests de Wilks, de Lawley-Hotelling, de Pillai et de Roy. Dans les cas simples, ils sont tous les quatre équivalents. Dans les autres cas, les trois premiers sont voisins et très rarement contradictoires. Par contre, le quatrième est moins précis et est déconseillé. S'il faut en privilégier un, nous recommandons plus particulièrement le test de Wilks.

C'est encore la procédure GLM de SAS qui est utilisée pour mettre en œuvre la MANOVA.



Dans tout ce chapitre, l'objectif est de modéliser un vecteur aléatoire  $Y$  de  $\mathbb{R}^D$  ( $D \in \mathbb{N}$ ,  $D \geq 2$ ) au moyen d'une loi gaussienne sur  $\mathbb{R}^D$ .

## 5.1 Écriture du modèle à un seul facteur

### 5.1.1 Les données

- On considère ici un unique facteur, encore noté  $F$ , possédant  $J$  niveaux ( $J \geq 2$ ), indicés par  $j$  ( $j = 1, \dots, J$ ).
- Pour chaque niveau  $j$  de  $F$ , on réalise  $n_j$  observations du vecteur aléatoire  $Y$  de  $\mathbb{R}^D$  ( $n_j \geq 1$ ); on pose  $n = \sum_{j=1}^J n_j$ .
- On note  $Y_{ij}$  le vecteur aléatoire associé à la  $i$ -ième observation réalisée au niveau  $j$  de  $F$  :  $Y_{ij} \in \mathbb{R}^D$ .

L'objectif de la MANOVA est d'étudier l'influence des niveaux du facteur  $F$  sur les valeurs du vecteur réponse  $Y$ . Cette influence va être étudiée globalement, dans  $\mathbb{R}^D$ , d'où la nécessité d'avoir recours à des techniques multidimensionnelles, différentes de celles vues en ANOVA.

**Remarque 41** *Parallèlement à la MANOVA, il est habituel de faire une ANOVA pour chacune des  $D$  composantes du vecteur  $Y$  (le logiciel SAS le fait automatiquement). C'est un complément intéressant pour la MANOVA, mais cela ne la remplace pas. En particulier, les tests à regarder pour le choix d'un modèle adapté à un jeu de données sont les tests multidimensionnels.*

### 5.1.2 Le modèle

#### Écriture initiale

Pour chaque expérience  $(i, j)$  ( $i$ -ième observation réalisée au niveau  $j$  de  $F$ ), on écrit le vecteur aléatoire réponse  $Y_{ij}$  de  $\mathbb{R}^D$  sous la forme :

$$Y_{ij} = \beta_j + U_{ij}.$$

Attention, les trois éléments de cette écriture doivent être vus comme des **vecteurs-lignes** de  $\mathbb{R}^D$ , comme précisé ci-dessous.

- Le vecteur  $\beta_j = (\beta_j^1 \cdots \beta_j^d \cdots \beta_j^D)$  est un paramètre à estimer; il modélise la valeur de la réponse  $Y$  au niveau  $j$  de  $F$ .
- Le terme  $U_{ij} = (U_{ij}^1 \cdots U_{ij}^D)$  est le vecteur aléatoire des erreurs. On suppose que les  $U_{ij}$  sont i.i.d., de loi  $\mathcal{N}_D(0_D, \Sigma)$ , où  $\Sigma$  est une matrice symétrique et strictement définie-positive; on doit également estimer  $\Sigma$ . On notera que  $\Sigma$  ne dépend pas de  $j$ , autrement dit on est toujours dans le cadre d'un modèle homoscédastique.
- Les vecteurs aléatoires  $Y_{ij}$  sont donc indépendants, de loi  $\mathcal{N}_D(\beta_j', \Sigma)$ .

Finalement, il y a  $J \times D$  paramètres de moyenne  $\beta_j^d$  à estimer, ainsi que  $\frac{D(D+1)}{2}$  paramètres de variance  $(\Sigma)_d^{d'}$  ( $1 \leq d \leq D$  ;  $1 \leq d' \leq D$ ). Comme on dispose de  $nD$  observations, on doit veiller à ce que la taille  $n$  de l'échantillon utilisé vérifie :  $n \geq J + \frac{D+1}{2}$ .

### Écriture matricielle

L'ensemble des  $nD$  observations réalisées peut se mettre sous la forme matricielle suivante :

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}.$$

- Dans l'écriture ci-dessus,  $\mathbf{X}$  et  $\beta$  sont des matrices réelles (non aléatoires) de dimensions respectives  $n \times J$  et  $J \times D$ .
- Comme dans le cas unidimensionnel, les colonnes de la matrice d'incidence  $\mathbf{X}$  sont les indicatrices  $Z^j$  des niveaux du facteur  $F$ , de sorte que  $\mathbf{X}$  ne comporte que des 0 et des 1.
- Les termes  $\mathbf{Y}$  et  $\mathbf{U}$  sont des matrices aléatoires de dimension  $n \times D$ . Elles sont gaussiennes et vérifient :

$$\mathbb{E}(\mathbf{U}) = \mathbf{0}_{n \times D} \quad ; \quad \mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta \quad ; \quad \text{Var}(\mathbf{U}) = \text{Var}(\mathbf{Y}) = \mathbf{I}_n \otimes \Sigma.$$

Dans cette dernière écriture,  $\mathbf{I}_n$  désigne la matrice identité d'ordre  $n$  et  $\otimes$  le produit matriciel direct, ou produit de Kronecker. En fait, on a

$$\mathbf{I}_n \otimes \Sigma = \begin{pmatrix} \Sigma & \mathbf{0}_{D \times D} & \cdots & \mathbf{0}_{D \times D} \\ \mathbf{0}_{D \times D} & \Sigma & \cdots & \mathbf{0}_{D \times D} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0}_{D \times D} & \mathbf{0}_{D \times D} & \cdots & \Sigma \end{pmatrix},$$

où chacun des  $n^2$  termes de cette matrice est lui-même une matrice (un bloc matriciel), de dimension  $D \times D$ . La matrice  $\mathbf{I}_n \otimes \Sigma$  est donc carrée d'ordre  $nD$ .

### Paramétrage centré

Comme dans le cas unidimensionnel, ce paramétrage consiste à décomposer chaque vecteur-ligne  $\beta_j$  sous la forme :

$$\beta_j = \mu + \alpha_j, \text{ avec } \mu = \frac{1}{J} \sum_{j=1}^J \beta_j \text{ et } \alpha_j = \beta_j - \mu.$$

Le paramètre  $\mu$  est l'effet (moyen) général et le paramètre  $\alpha_j$  est l'effet principal (ou différentiel) du niveau  $j$  de  $F$ . Ces deux paramètres sont des vecteurs de  $\mathbb{R}^D$  et on notera que l'on a encore  $\sum_{j=1}^J \alpha_j = \mathbf{0}_D$ .

### Paramétrage SAS

Pour ce paramétrage, on pose  $m = \beta_J$  et  $a_j = \beta_j - \beta_J$  (de sorte que, encore une fois,  $a_J = \mathbf{0}_D$ ). Les paramètres  $m$  et  $a_j$  sont également des vecteurs de  $\mathbb{R}^D$ .

## 5.2 Estimation des paramètres du modèle à un facteur

### 5.2.1 Vraisemblance et log-vraisemblance

La vraisemblance de l'échantillon des  $y_{ij}$  s'écrit

$$\begin{aligned} L(y_{ij}, \beta, \Sigma) &= \prod_{j=1}^J \prod_{i=1}^{n_j} \frac{1}{(2\pi)^{D/2} (\det \Sigma)^{1/2}} \exp\left[-\frac{1}{2} (y_{ij} - \beta_j) \Sigma^{-1} (y_{ij} - \beta_j)'\right] \\ &= C_1 (\det \Sigma)^{-n/2} \exp\left[-\frac{1}{2} \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - X_j \beta) \Sigma^{-1} (y_{ij} - X_j \beta)'\right], \end{aligned}$$

où  $C_1$  est une constante et  $X_j$  un vecteur-ligne à  $J$  éléments, comportant un 1 en  $j$ -ième colonne et des 0 partout ailleurs (en fait,  $X_j$  est n'importe laquelle des lignes de la matrice  $\mathbf{X}$  correspondant aux observations du niveau  $j$  de  $F$ ).

La log-vraisemblance s'écrit

$$\begin{aligned} l(y_{ij}, \beta, \Sigma) &= \log[L(y_{ij}, \beta, \Sigma)] \\ &= C_2 - \frac{n}{2} \log(\det \Sigma) - \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_{\mathbf{I}_n, \Sigma^{-1}}^2, \end{aligned}$$

où  $\log$  désigne le logarithme népérien et où  $C_2 = \log(C_1)$ . Par ailleurs, on rappelle que si  $\mathbf{N}$  est une matrice  $n \times n$ , symétrique et strictement définie-positive, si  $\mathbf{P}$  est une matrice  $p \times p$ , également symétrique et strictement définie-positive, alors, pour toute matrice  $\mathbf{A}$  de dimension  $n \times p$ , on peut définir sa norme carrée par

$$\|\mathbf{A}\|_{\mathbf{N}, \mathbf{P}}^2 = \text{tr}(\mathbf{A}\mathbf{P}\mathbf{A}'\mathbf{N}),$$

où  $\text{tr}$  désigne la trace de la matrice correspondante (cette norme est appelée norme de Hilbert-Schmidt).

### 5.2.2 Estimation maximum de vraisemblance

Pour estimer les matrices  $\beta$  et  $\Sigma$ , on doit ici faire des dérivations matricielles. On admettra les résultats ci-dessous (qui sont également les résultats de l'estimation moindres carrés, en l'absence de l'hypothèse de normalité). Pour des précisions, on pourra se reporter à Seber (1984). Citons également le site "The Matrix Cookbook", très intéressant, à l'adresse suivante :

<http://matrixcookbook.com/>

#### Estimation des paramètres d'intérêt

On obtient  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\bar{y}_{\bullet 1} \dots \bar{y}_{\bullet j} \dots \bar{y}_{\bullet J})'$  (matrice  $J \times D$ ), où  $\bar{y}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$  est le vecteur-ligne de  $\mathbb{R}^D$ , moyenne des observations de  $Y$  au niveau  $j$  de  $F$ . On notera  $\hat{\mathbf{B}}$  l'estimateur correspondant défini par  $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ .

**Remarque 42** Signalons que  $\hat{\beta}$  peut s'obtenir colonne par colonne, au moyen des résultats, pour chaque colonne, d'une ANOVA unidimensionnelle à un facteur :  $\hat{\beta}^d = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^d$  est la solution de l'ANOVA de  $Y^d$  sur  $F$ . Ainsi, pour obtenir  $\hat{\beta}$ , on pourra utiliser les estimations unidimensionnelles de  $\hat{\beta}^d$  ( $d = 1, \dots, D$ ) fournies par SAS au début des sorties de la MANOVA.

#### Valeurs prédites

Pour un niveau  $j$  donné, et pour toute observation  $i$  faite à ce niveau, la valeur prédite correspondante est :  $\hat{y}_{ij} = \hat{\beta}_j = \bar{y}_{\bullet j}$  (vecteur de  $\mathbb{R}^D$ ). On notera  $\hat{\mathbf{Y}}$  la matrice aléatoire  $n \times D$  de l'ensemble des valeurs prédites.

#### Résidus

Il vient :  $\hat{u}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{\bullet j}$  (vecteur de  $\mathbb{R}^D$ ). On notera  $\hat{\mathbf{U}}$  la matrice aléatoire  $n \times D$  des résidus ainsi définis.

#### Estimation de la matrice des variances-covariances

Elle est obtenue, comme dans le cas unidimensionnel, à partir de la matrice des résidus :  $\hat{\Sigma} = \frac{1}{n-J} \hat{\mathbf{U}}' \hat{\mathbf{U}}$  (matrice  $D \times D$ ).

### 5.2.3 Propriétés des estimateurs maximum de vraisemblance

- Les matrices  $\hat{\mathbf{B}}$ , de dimension  $J \times D$ ,  $\hat{\mathbf{Y}}$ , de dimension  $n \times D$ , et  $\hat{\mathbf{U}}$ , de dimension  $n \times D$ , sont des matrices aléatoires gaussiennes, d'espérances respectives  $\beta$ ,  $\mathbf{X}\beta$  et  $\mathbf{0}_{n \times D}$  (leurs matrices de variances-covariances ne seront pas explicitées).
- La matrice aléatoire  $\hat{\mathbf{U}}$  est indépendante des matrices aléatoires  $\hat{\mathbf{B}}$  et  $\hat{\mathbf{Y}}$ .
- Enfin,  $(n - J)\hat{\Sigma} = \hat{\mathbf{U}}'\hat{\mathbf{U}}$  est une matrice aléatoire distribuée selon une loi de Wishart de dimension  $D$ , à  $n - J$  degrés de liberté et de matrice associée  $\Sigma$ ; cette loi de probabilité est notée  $W_D(n - J, \Sigma)$ .

### 5.2.4 Indications sur la loi de Wishart

Il s'agit, en quelques sortes, d'une généralisation multidimensionnelle de la loi de khi-deux. Dans  $\mathbb{R}^D$  ( $D \geq 2$ ), considérons  $m$  ( $m \geq D$ ) vecteurs (colonnes) aléatoires notés  $T_i$  ( $i = 1, \dots, m$ ), supposés i.i.d. selon une loi normale centrée, de matrice de variances-covariances  $\Sigma$  ( $D \times D$ , symétrique et strictement définie-positive). Alors, la matrice aléatoire  $\mathbf{W} = \sum_{i=1}^m T_i T_i'$  (de dimension  $D \times D$ ) définit une loi de Wishart de dimension  $D$ , à  $m$  d.d.l., de matrice associée  $\Sigma$ . Elle est notée  $W_D(m, \Sigma)$ .

**Remarque 43** *On notera que  $\Sigma$  n'est pas la matrice des variances-covariances de  $\mathbf{W}$ . En effet, la matrice aléatoire  $\mathbf{W}$  est constituée de  $D \times D = D^2$  éléments aléatoires et admet donc une matrice de variances-covariances de dimension  $D^2 \times D^2$  qui ne sera pas explicitée (mais qui ne peut être  $\Sigma$ ).*

#### Quelques propriétés de la loi de Wishart

- Telle qu'elle est définie ci-dessus, la loi de Wishart  $W_D(m, \Sigma)$  apparaît comme la loi de  $m$  fois la matrice des variances-covariances empiriques d'une loi normale centrée de  $\mathbb{R}^D$ , de matrice de variances-covariances  $\Sigma$ .
  - $\mathbb{E}(\mathbf{W}) = m\Sigma$  (immédiat).
  - Supposons :  $\mathbf{W}_1 \sim W_D(m_1, \Sigma)$ ;  $\mathbf{W}_2 \sim W_D(m_2, \Sigma)$ ;  $\mathbf{W}_1$  et  $\mathbf{W}_2$  indépendantes; alors :  $\mathbf{W}_1 + \mathbf{W}_2 \sim W_D(m_1 + m_2, \Sigma)$ . (Cette propriété est admise.)
- Pour plus de détails sur la loi de Wishart, on pourra encore se reporter à Seber (1984).

## 5.3 Tests dans le modèle à un facteur

La seule hypothèse considérée ici est la significativité du facteur  $F$ , autrement dit la significativité du modèle lui-même (puisque  $F$  est le seul facteur pris en compte, pour l'instant, dans le modèle). L'hypothèse nulle s'écrit sous l'une des formes suivantes :

$$\{H_0 : F \text{ n'a pas d'effet sur } Y\} \iff \{H_0 : \beta_1 = \dots = \beta_J\} \iff \{H_0 : \alpha_1 = \dots = \alpha_J = 0\} \iff \{H_0 : a_1 = \dots = a_J = 0\} \quad (\text{il y a chaque fois } J - 1 \text{ contraintes vectorielles dans } \mathbb{R}^D).$$

La mise en œuvre d'un test permettant de tester  $H_0$  contre son contraire  $H_1$ , avec un niveau  $\alpha$  fixé, nécessite de généraliser le test de Fisher qui ne peut plus être utilisé ici. Cette généralisation peut se faire de différentes manières et conduit à différents tests. Avant de les introduire, nous devons définir les deux matrices à partir desquelles ils sont construits.

### 5.3.1 Les matrices $\mathbf{H}$ et $\mathbf{E}$

#### Retour sur le cas où $Y$ est unidimensionnelle

Revenons sur le cas de l'ANOVA à un seul facteur  $F$  avec une variable réponse  $Y$  unidimensionnelle. Pour tester la significativité du modèle, donc du facteur, on a utilisé le test de Fisher, qui fait intervenir les quantités ci-dessous.

- $SSF = \sum_{j=1}^J n_j (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2$  : c'est la somme des carrés expliquée par le facteur  $F$ , ou somme des carrés inter-groupes, ou *between sum of squares*. Sous l'hypothèse nulle  $H_0$ , cette somme est nécessairement "petite" ; nous la noterons  $H$ , car elle est liée à l'hypothèse testée :  $H = SSF$ . Son d.d.l. est  $J - 1$ .

- $SSE = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2$  : c'est la somme des carrés résiduelle (non expliquée par  $F$ , donc par le modèle), ou somme des carrés intra-groupes, ou *pooled within sum of squares*. Elle représente la somme des carrés liée à l'erreur du modèle, pour cette raison notée  $E$  :  $E = SSE$ . Son d.d.l. est  $n - J$ .

La statistique du test de Fisher peut s'écrire sous la forme :

$$F = \frac{SSF}{SSE} \frac{n - J}{J - 1} = \frac{H}{E} \frac{n - J}{J - 1}.$$

(Ne pas confondre les deux notations  $F$ , tantôt pour le facteur, tantôt pour la statistique de Fisher.)

Enfin, rappelons la relation  $SST = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet\bullet})^2 = SSF + SSE = H + E$ , où  $SST$  est la somme des carrés totale, de d.d.l.  $n - 1$ .

### Généralisation au cas où $Y$ est multidimensionnelle

Dans le cas de la MANOVA en dimension  $D$  et à un seul facteur, on généralise les quantités  $H$  et  $E$  comme indiqué ci-dessous.

- La somme  $H$  est remplacée par la matrice  $\mathbf{H}$ ,  $D \times D$ , définie par :

$$\mathbf{H} = \sum_{j=1}^J n_j (\bar{\mathbf{y}}_{\bullet j} - \bar{\mathbf{y}}_{\bullet\bullet})' (\bar{\mathbf{y}}_{\bullet j} - \bar{\mathbf{y}}_{\bullet\bullet}).$$

Le d.d.l. associé, qui vaut toujours  $J - 1$ , sera noté  $\nu_H$ .

- La somme  $E$  est remplacée par la matrice  $\mathbf{E}$ ,  $D \times D$ , définie par :

$$\mathbf{E} = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})' (y_{ij} - \bar{y}_{\bullet j}).$$

Le d.d.l. associé, qui vaut toujours  $n - J$ , sera noté  $\nu_E$ . On notera que  $\mathbf{E} = \hat{\mathbf{U}}' \hat{\mathbf{U}}$ .

- La somme des carrés totale,  $SST$ , est remplacée par la somme des deux matrices définies ci-dessus :  $\mathbf{H} + \mathbf{E}$ ; son d.d.l. est  $n - 1$ . Nous n'utiliserons pas de notation particulière pour cette matrice.

En ANOVA à un seul facteur, la statistique de Fisher est proportionnelle au rapport  $\frac{H}{E}$ . En MANOVA à un seul facteur, pour tester la significativité de ce facteur, les tests utilisés s'appuient sur l'un des produits matriciels  $\mathbf{H}\mathbf{E}^{-1}$  ou  $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ .

**Remarque 44** Nous avons déjà donné l'expression de l'estimateur de la matrice des variances-covariances :  $\hat{\Sigma} = \frac{1}{n - J} \hat{\mathbf{U}}' \hat{\mathbf{U}}$ . On voit qu'on peut la réécrire sous la forme :  $\hat{\Sigma} = \frac{1}{n - J} \mathbf{E} = \frac{\mathbf{E}}{\nu_E}$ .

**Remarque 45** La matrice  $\hat{\Sigma}$  contient des variances et des covariances empiriques résiduelles, c'est-à-dire conditionnées par le facteur  $F$ . Si, à partir des éléments de  $\hat{\Sigma}$ , on calcule des coefficients de corrélations linéaires empiriques entre composantes de  $Y$ , il s'agira de corrélations résiduelles, plus couramment appelées corrélations partielles (conditionnelles à  $F$ ). On trouve ces corrélations partielles en sortie du logiciel SAS (en pratique, elles ont peu d'intérêt).

### 5.3.2 Le test de Wilks

#### Principe

Il s'agit du test le plus courant dans le contexte de la MANOVA qui est, en fait, une adaptation du test du rapport des vraisemblances.

Notons  $\theta$  le vecteur de tous les paramètres du modèle, de dimension  $(J \times D) + \frac{D(D+1)}{2}$ ,  $\hat{\theta}$  son estimation maximum de vraisemblance et  $\hat{\theta}_0$  son estimation maximum de vraisemblance sous la contrainte définie par  $H_0$  (autrement dit sous la contrainte linéaire :  $\beta_1 = \dots = \beta_J$ ). La statistique du test du rapport des vraisemblances est  $\frac{L(y, \hat{\theta}_0)}{L(y, \hat{\theta})}$ , dont on peut vérifier qu'elle vaut :

$\left[\frac{\det(\mathbf{E})}{\det(\mathbf{H} + \mathbf{E})}\right]^{n/2}$  (voir Seber, 1984). Le test de Wilks consiste à considérer la puissance  $2/n$  de cette quantité, autrement dit sa statistique est définie par

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{H} + \mathbf{E})} = \prod_{k=1}^s \frac{1}{1 + \lambda_k},$$

où les  $\lambda_k$  sont les valeurs propres de la matrice  $\mathbf{HE}^{-1}$  (ou  $\mathbf{E}^{-1}\mathbf{H}$ ) et où  $s = \inf(D, J - 1)$  est le nombre de valeurs propres non nulles de cette matrice.

La mise en œuvre de ce test va dépendre du cas de figure.

### Cas où on se ramène à un test de Fisher exact

Cela se produit dans trois cas particuliers.

- Cas d'un facteur à 2 niveaux :  $J = 2 \iff \nu_H = J - 1 = 1$  ( $D$  quelconque). On peut alors montrer :

$$\frac{1 - \Lambda}{\Lambda} \frac{\nu_E - D + 1}{D} = \frac{1 - \Lambda}{\Lambda} \frac{n - (D + 1)}{D} \sim F_{D ; n - (D + 1)}.$$

Les tables de la distribution de Fisher permettent donc de faire un test exact.

- Cas d'un facteur à 3 niveaux :  $J = 3 \iff \nu_H = 2$  ( $D$  quelconque). Il vient dans ce cas :

$$\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{\nu_E - D + 1}{D} = \frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n - (D + 2)}{D} \sim F_{2D ; 2(n - (D + 2))}.$$

Même chose, on peut faire un test de Fisher exact.

- Cas où  $Y$  est à 2 dimensions :  $D = 2$  ( $J$  quelconque). Il vient maintenant :

$$\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{\nu_E - 1}{\nu_H} = \frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n - (J + 1)}{J - 1} \sim F_{2\nu_H ; 2(\nu_E - 1)} (F_{2(J-1) ; 2(n - (J+1))}).$$

Toujours la même chose.

### Cas où on dispose de tables

Des tables du test de Wilks on été établies et permettent de faire encore un test exact dans de nombreux autres cas (on les trouve dans les ouvrages de statistique multidimensionnelle). Pour les niveaux 10%, 5%, 2, 5%, 1% et 0, 5%, on dispose de tables pour  $D$  variant de 3 à 10, pour  $\nu_H$  variant de 3 à 13 (et souvent plus) et pour  $\nu_E$  variant de 1 à 20, ainsi que pour les valeurs 30, 40, 60 et 120. On trouvera ces tables, par exemple, dans Seber (1984).

### Approximation de Fisher

Dans les cas où on ne dispose pas de tables permettant de faire un test exact, on pourra faire un test de Fisher approché (d'autant meilleur que  $n$  est grand) en utilisant le résultat suivant :

$$\phi = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{ft - g}{D\nu_H} \sim F_{D\nu_H ; ft - g} \text{ (approximativement).}$$

Dans l'expression ci-dessus, on a les définitions suivantes :

$$f = \nu_H + \nu_E - \frac{\nu_H + D + 1}{2} = (n - 1) - \frac{J + D}{2} ; g = \frac{D\nu_H}{2} - 1 ; t = \left[ \frac{D^2\nu_H^2 - 4}{D^2 + \nu_H^2 - 5} \right]^{1/2}.$$

**Remarque 46** Dans chacun des trois cas particuliers  $J = 2$ ,  $J = 3$  et  $D = 2$ , on pourra vérifier que l'expression ci-dessus redonne celle fournie plus haut. Dans ces trois cas, la distribution de Fisher n'est donc pas une approximation de la loi de la statistique de test, mais sa distribution exacte.

**Remarque 47** Concernant le test de Wilks (ainsi, d'ailleurs, que les suivants), le logiciel SAS fournit, en sortie de la procédure GLM, la valeur de  $\Lambda$ , la valeur  $\phi = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{ft - g}{D\nu_H}$ , les d.d.l.  $D\nu_H$  et  $ft - g$ , et une p-value représentant la probabilité qu'une loi de Fisher à  $D\nu_H$  et  $ft - g$  d.d.l. dépasse  $\phi$  ; le test réalisé à partir de cette p-value est donc, selon le cas, soit un test exact, soit l'approximation de Fisher indiquée ci-dessus.

### 5.3.3 Autres tests

Dans la littérature statistique, on trouve d'autres tests permettant d'éprouver la même hypothèse nulle. Ces tests sont automatiquement fournis par le logiciel SAS, en même temps que le test de Wilks. Nous donnons leur principe ci-dessous.

#### Le test de la trace de Lawley-Hotelling

La statistique de ce test est :

$$T^2 = \nu_E \text{ trace}(\mathbf{HE}^{-1}) = (n - J) \sum_{k=1}^s \lambda_k.$$

Pour un niveau de test  $\alpha = 5\%$ , pour des valeurs de  $D$  variant de 2 à 6, pour  $\nu_H = J - 1$  variant de  $D$  à 6, puis prenant les valeurs 8, 10, 12, 15, 20, 25, 40 et 60, enfin pour  $\nu_E = n - J$  variant de  $D$  à 8, puis prenant les valeurs 10, 20, 30  $\dots$  100 et 200, on dispose de tables pour la statistique  $\frac{T^2}{\nu_E} = \sum_{k=1}^s \lambda_k$ , permettant de faire un test exact.

Dans les autres cas, on utilise l'approximation suivante

$$\frac{T^2}{c \nu_E} = \frac{1}{c} \text{ trace}(\mathbf{HE}^{-1}) \sim F_{a; b} \text{ (approximativement)}$$

avec :

$$a = D \nu_H ; b = 4 + \frac{a + 2}{B - 1} ; B = \frac{(\nu_E + \nu_H - D - 1)(\nu_E - 1)}{(\nu_E - D - 3)(\nu_E - D)} ; c = \frac{a(b - 2)}{b(\nu_E - D - 1)}.$$

**Remarque 48** *Compte-tenu de l'expression de la statistique de ce test, celui-ci est la généralisation multidimensionnelle la plus naturelle du test de Fisher.*

#### Le test de la trace de Pillai

La statistique de ce test est

$$V = \text{trace} [\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}] = \sum_{k=1}^s \mu_k = \sum_{k=1}^s \frac{\lambda_k}{1 + \lambda_k},$$

où  $s = \inf(D, J - 1)$ , où les  $\mu_k$  sont les valeurs propres de la matrice  $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$  et les  $\lambda_k$  celles de la matrice  $\mathbf{HE}^{-1}$ .

Si l'on pose  $k_1 = \frac{1}{2}(|D - \nu_H| - 1)$  et  $k_2 = \frac{1}{2}(\nu_E - D - 1)$ , des tables permettent de réaliser un test exact de Pillai pour  $\alpha = 5\%$ ,  $s$  variant de 2 à 6,  $k_1$  et  $k_2$  variant de 0 à 10 ou bien prenant les valeurs 15, 20 ou 25.

Dans les autres cas, on utilisera l'approximation suivante :

$$\frac{V}{s - V} \frac{2k_2 + s + 1}{2k_1 + s + 1} = \frac{V}{s - V} \frac{n + \inf(D, J - 1) - (D + J)}{\sup(D, J - 1)} \sim F_{s(2k_1 + s + 1) ; s(2k_2 + s + 1)}$$

(approximativement).

#### Le test de la plus grande racine de Roy

La statistique de ce dernier test est  $\lambda_{\max}$ , la plus grande des valeurs propres de  $\mathbf{HE}^{-1}$ . On trouve diverses approximations qui permettent de mettre en œuvre ce test. Celle utilisée par SAS est la suivante :

$$S = \frac{\lambda_{\max} (\nu_H + \nu_E - r)}{r} \sim F_r ; \nu_H + \nu_E - r \text{ (approximativement) ,}$$

où  $r = \max(D, \nu_H)$ .

On notera que, dans ce cas, la loi de Fisher est un minorant pour la loi de  $S$ , ce qui signifie que la  $p$ -value calculée par SAS pour ce test est toujours plus petite que celle des autres tests. Pour cette raison, **nous déconseillons ce test.**

### 5.3.4 Cas particulier : $J = 2$

Il n'y a qu'une seule valeur propre non nulle dans ce cas particulier, et les différentes approximations par une loi de Fisher données ci-dessus sont toutes identiques (le résultat est simple, bien qu'un peu fastidieux, à vérifier). De plus, elles correspondent toutes les quatre au test exact de Fisher donné en 5.3.2. En fait, dans ce cas, la statistique utilisée vaut

$$\lambda_1 \frac{n - D - 1}{D}$$

et est distribuée selon une loi de Fisher à  $D$  et  $n - D - 1$  degrés de liberté.

## 5.4 Illustration

### Les données

Il s'agit d'un exemple fictif d'analyse de variance multidimensionnelle, de dimension 3, à un seul facteur. Le facteur est à deux niveaux, notés 1 et 2, et figure en première colonne. Les variables réponses figurent dans les trois colonnes suivantes et prennent des valeurs entières comprises entre 8 et 24. Il y a 8 individus observés, donc 8 lignes dans le fichier des données reproduit ci-dessous.

```
1 10 12 14
1 11 13 15
1 8 9 8
1 9 10 8
2 15 17 16
2 19 18 17
2 21 20 19
2 23 22 24
```

### Le programme SAS

Le programme SAS ci-dessous permet de faire la MANOVA de ces données de façon standard.

```
* ----- ;
* options facultatives pour la mise en page des sorties ;
* ----- ;
options pagesize=64 linesize=76 nodate;
title;
footnote 'MANOVA - donnees fictives - 1 facteur';
* ----- ;
*          lecture des donnees          ;
*      (le fichier "fic1.don" contient les donnees ;
*      et se trouve dans le repertoire de travail) ;
* ----- ;
data fic1;
infile 'fic1.don';
input f $ y1 y2 y3;
run;
* ----- ;
*          procedure GLM pour la MANOVA          ;
* ----- ;
proc glm data=fic1;
class f;
model y1-y3 = f / ss3 solution;
manova H = f;
run;
quit;
```

## Les sorties de la procédure GLM

PAGE 1

The GLM Procedure

-----

## Class Level Information

Class	Levels	Values
f	2	1 2

Number of observations	8
------------------------	---

PAGE 2

The GLM Procedure

-----

Dependent Variable: y1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	200.0000000	200.0000000	30.00	0.0015
Error	6	40.0000000	6.6666667		
Corrected Total	7	240.0000000			

R-Square	Coeff Var	Root MSE	y1 Mean
0.833333	17.80682	2.581989	14.50000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f	1	200.0000000	200.0000000	30.00	0.0015

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	19.50000000 B	1.29099445	15.10	<.0001
f 1	-10.00000000 B	1.82574186	-5.48	0.0015
f 2	0.00000000 B	.	.	.

PAGE 3

The GLM Procedure

-----

Dependent Variable: y2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	136.1250000	136.1250000	33.00	0.0012
Error	6	24.7500000	4.1250000		
Corrected Total	7	160.8750000			

R-Square	Coeff Var	Root MSE	y2 Mean
0.846154	13.42816	2.031010	15.12500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f	1	136.1250000	136.1250000	33.00	0.0012

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	19.25000000 B	1.01550480	18.96	<.0001
f 1	-8.25000000 B	1.43614066	-5.74	0.0012
f 2	0.00000000 B	.	.	.

PAGE 4 The GLM Procedure

-----

Dependent Variable: y3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	120.1250000	120.1250000	8.93	0.0244
Error	6	80.7500000	13.4583333		
Corrected Total	7	200.8750000			

R-Square	Coeff Var	Root MSE	y3 Mean
0.598009	24.25494	3.668560	15.12500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f	1	120.1250000	120.1250000	8.93	0.0244

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	19.00000000 B	1.83428006	10.36	<.0001
f 1	-7.75000000 B	2.59406374	-2.99	0.0244
f 2	0.00000000 B	.	.	.

PAGE 5

-----

Characteristic Roots and Vectors of: E Inverse \* H, where  
H = Type III SSCP Matrix for f  
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector y1	V'EV=1 y2	y3
26.4943529	100.00	-0.23580920	1.15536589	-0.45600123
0.0000000	0.00	-0.36273387	0.42959731	0.01073044
0.0000000	0.00	0.18534079	-0.44542771	0.23501557

MANOVA Test Criteria and Exact F Statistics  
 for the Hypothesis of No Overall f Effect  
 H = Type III SSCP Matrix for f  
 E = Error SSCP Matrix

S=1 M=0.5 N=1

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.03637111	35.33	3	4	0.0025
Pillai's Trace	0.96362889	35.33	3	4	0.0025
Hotelling-Lawley Trace	26.49435290	35.33	3	4	0.0025
Roy's Greatest Root	26.49435290	35.33	3	4	0.0025

## 5.5 Modèle à deux facteurs croisés

### 5.5.1 Données, modèle et paramétrages

On considère maintenant deux facteurs explicatifs notés  $F_1$  et  $F_2$ , à  $J$  et  $K$  niveaux respectivement. Les niveaux de  $F_1$  sont indicés par  $j$  ( $j = 1, \dots, J$ ) et ceux de  $F_2$  par  $k$  ( $k = 1, \dots, K$ ). Pour chaque couple  $(j, k)$  obtenu par croisement des deux facteurs, on réalise  $n_{jk}$  observations ( $n_{jk} \geq 1$ ) d'un vecteur aléatoire réponse à valeurs dans  $\mathbb{R}^D$  ( $D \geq 2$ ), ces vecteurs étant notés  $Y_{ijk}$  ( $i = 1, \dots, n_{jk}$ ). On pose  $n = \sum_{j=1}^J \sum_{k=1}^K n_{jk}$ .

Le modèle se met sous la forme

$$Y_{ijk} = \beta_{jk} + U_{ijk},$$

où chaque  $\beta_{jk}$  est un paramètre de  $\mathbb{R}^D$  et où  $U_{ijk} \sim \mathcal{N}_D(0, \Sigma)$ , les  $U_{ijk}$  étant indépendants (et donc i.i.d.). On a ainsi :  $Y_{ijk} \sim \mathcal{N}_D(\beta_{jk}, \Sigma)$ .

On peut réécrire le modèle sous la forme matricielle

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U},$$

où  $\mathbf{Y}$  est une matrice aléatoire  $n \times D$ ,  $\mathbf{X}$  est la matrice d'incidence, de dimension  $n \times JK$ , dont les colonnes sont les indicatrices des cellules  $(j, k)$ ,  $\beta$  est la matrice  $JK \times D$  des paramètres et  $\mathbf{U}$  la matrice aléatoire  $n \times D$  des erreurs.

Encore une fois, les paramètres  $\beta_{jk}$  peuvent être décomposés selon le paramétrage centré (en écrivant  $\beta_{jk} = \mu + \alpha_j^1 + \alpha_k^2 + \gamma_{jk}$ , avec les contraintes usuelles de centrage) ou encore selon le paramétrage SAS (en écrivant maintenant  $\beta_{jk} = m + a_j^1 + a_k^2 + c_{jk}$ , avec tous les paramètres dont au moins un des indice  $j$  ou  $k$  est maximum égaux à 0), tous les paramètres intervenant étant des vecteurs-lignes de  $\mathbb{R}^D$ .

### 5.5.2 Tests et estimations

Tout ce qui a été exposé dans le cas d'un seul facteur se généralise ici sans autre difficulté que celle due aux écritures. En particulier, on retrouve les mêmes tests.

- Pour les tests de nullité des effets d'interactions, on notera que l'on a maintenant  $\nu_H = (J-1)(K-1)$ ,  $\nu_E = n - JK$  et que le nombre de valeurs propres non nulles à prendre en compte est  $s = \inf(D, \nu_H)$ .
- Par ailleurs, on notera que lorsque le test de significativité de chaque facteur  $F_1$  et  $F_2$  est fait dans le cadre du modèle complet,  $\nu_E$  demeure égal à  $n - JK$  et les simplifications indiquées dans la remarque 46 ne s'appliquent plus.
- Dans un modèle additif à deux facteurs croisés,  $\nu_E$  vaut  $n - (J + K - 1)$ .
- Lorsqu'un facteur ne possède que  $J = 2$  niveaux, les 4 tests multidimensionnels sont encore tous identiques, mais l'expression de la statistique de test est plus compliquée que celle indiquée en 5.3.4, car le nombre de niveaux du second facteur intervient.
- Enfin, concernant les estimations de  $\hat{\beta}$ , on les obtient colonne par colonne, comme indiqué dans la remarque 42.

### 5.5.3 Généralisation

On peut encore envisager, avec une variable réponse  $Y$  multidimensionnelle d'ordre  $D$ , des modèles à trois facteurs croisés ou plus. Il n'y a aucune difficulté théorique, mais seulement des difficultés formelles (complexité des écritures) et pratiques : nombre très important de paramètres à estimer, donc d'observations à réaliser. Nous ne détaillons pas davantage ces extensions.

### 5.5.4 Illustration

#### Les données

Les données, toujours fictives, sont de même nature que les précédentes. La variable réponse est à 3 dimensions et figure dans les 3 dernières colonnes du fichier. Il y a maintenant 2 facteurs, le premier à 2 niveaux (notés 1 et 2), le second à 4 niveaux (notés 1, 2, 3 et 4). Les facteurs figurent dans les 2 premières colonnes du fichier. Pour chaque cellule (il y en a 8), on a réalisé 4 observations, de sorte que l'on dispose de 32 observations. Les données sont reproduites ci-dessous.

```

1 1  8  7 10
1 1  9 13 11
1 1  8  9  8
1 1  9 10  8
1 2 10 12 14
1 2 11 13 15
1 2 11 10 12
1 2 13 12 12
1 3 13 16 19
1 3 15 17 20
1 3 12 16 16
1 3 14 18 18
1 4 12 18 19
1 4 18 19 23
1 4 13 11 19
1 4 16 21 20
2 1 15 17 16
2 1 17 18 17
2 1 15 16 18
2 1 18 17 18
2 2 21 20 24
2 2 23 22 24
2 2 20 23 22
2 2 22 26 23
2 3 25 25 30
2 3 28 25 29
2 3 23 28 25
2 3 25 30 27
2 4 28 27 32
2 4 29 26 29
2 4 26 29 27
2 4 27 34 29

```

#### Le programme SAS

L'option `nouni` du programme ci-dessous permet d'éviter les traitements unidimensionnels.

```

options pagesize=64 linesize=76 nodate;
title;
footnote 'MANOVA - donnees fictives - 2 facteurs';
* ----- ;
*           lecture des donnees                ;
*   (le fichier "fic2.don" contient les donnees ;
*   et se trouve dans le repertoire de travail) ;
* ----- ;

```

```

data fic2;
infile 'fic2.don';
input f1 $ f2 $ y1 y2 y3;
run;
* ----- ;
*           procedure GLM pour la MANOVA           ;
* ----- ;
proc glm data=fic2;
class f1 f2;
model y1-y3 = f1 | f2 / nouni;
manova H = f1 | f2 / printh printe;
run;
quit;

```

### Les sorties de la procédure GLM

```

PAGE 1                               The GLM Procedure
-----
                                Class Level Information

Class          Levels   Values
-----
f1              2       1 2
f2              4       1 2 3 4

Number of observations      32

```

```

PAGE 2                               The GLM Procedure
-----
Multivariate Analysis of Variance

E = Error SSCP Matrix

          y1          y2          y3
y1        63          13          35
y2        13        159.75       -2.25
y3        35        -2.25          66

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 24          y1          y2          y3
y1          1.000000      0.129584      0.542782
              0.5370              0.0051
y2          0.129584      1.000000      -0.021912
              0.5370              0.9172
y3          0.542782      -0.021912      1.000000
              0.0051              0.9172

```

PAGE 3

The GLM Procedure  
Multivariate Analysis of Variance

tests de f1

-----

H = Type III SSCP Matrix for f1

	y1	y2	y3
y1	903.125	855.3125	775.625
y2	855.3125	810.03125	734.5625
y3	775.625	734.5625	666.125

Characteristic Roots and Vectors of: E Inverse \* H, where  
H = Type III SSCP Matrix for f1  
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1		
		y1	y2	y3
19.8676273	100.00	0.07148178	0.03487179	0.05101439
0.0000000	0.00	-0.11562870	-0.00337882	0.13836212
0.0000000	0.00	-0.06841360	0.07223796	0.00000000

MANOVA Test Criteria and Exact F Statistics  
for the Hypothesis of No Overall f1 Effect

H = Type III SSCP Matrix for f1  
E = Error SSCP Matrix

S=1 M=0.5 N=10

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.04792112	145.70	3	22	<.0001
Pillai's Trace	0.95207888	145.70	3	22	<.0001
Hotelling-Lawley Trace	19.86762728	145.70	3	22	<.0001
Roy's Greatest Root	19.86762728	145.70	3	22	<.0001

tests de f2

-----

H = Type III SSCP Matrix for f2

	y1	y2	y3
y1	352.375	408.5625	474.125
y2	408.5625	479.59375	553.4375
y3	474.125	553.4375	640.375

Characteristic Roots and Vectors of: E Inverse \* H, where  
H = Type III SSCP Matrix for f2  
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1		
		y1	y2	y3
13.1912003	99.67	0.02098208	0.03734450	0.09571189
0.0429976	0.32	-0.14092901	0.04193749	0.06806198
0.0007461	0.01	-0.05346802	-0.05737958	0.08918153

MANOVA Test Criteria and F Approximations  
for the Hypothesis of No Overall f2 Effect  
H = Type III SSCP Matrix for f2  
E = Error SSCP Matrix

S=3 M=-0.5 N=10

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.06751086	12.09	9	53.693	<.0001
Pillai's Trace	0.97150442	3.83	9	72	0.0005
Hotelling-Lawley Trace	13.23494405	31.40	9	31.536	<.0001
Roy's Greatest Root	13.19120030	105.53	3	24	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

tests des interactions

H = Type III SSCP Matrix for f1\*f2

	y1	y2	y3
y1	28.375	23.0625	6.125
y2	23.0625	23.34375	7.6875
y3	6.125	7.6875	4.375

Characteristic Roots and Vectors of: E Inverse \* H, where

H = Type III SSCP Matrix for f1\*f2  
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1		
		y1	y2	y3
0.57694177	85.26	0.13350497	0.02373892	-0.04939697
0.09184462	13.57	-0.06443491	0.05012982	0.12343212
0.00788528	1.17	0.03441843	-0.05804521	0.06380435

MANOVA Test Criteria and F Approximations for  
the Hypothesis of No Overall f1\*f2 Effect  
H = Type III SSCP Matrix for f1\*f2  
E = Error SSCP Matrix

S=3 M=-0.5 N=10

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.57625194	1.52	9	53.693	0.1660
Pillai's Trace	0.45780353	1.44	9	72	0.1872
Hotelling-Lawley Trace	0.67667167	1.61	9	31.536	0.1564
Roy's Greatest Root	0.57694177	4.62	3	24	0.0110

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

### Application

À titre d'application, on pourra, à partir des valeurs propres données ci-dessus dans le cadre du test de significativité des interactions, retrouver les valeurs de  $\Lambda$  (Wilks' Lambda), de  $F$  (F Value pour le test de Wilks), ainsi que les degrés de liberté (on remarquera les d.d.l. décimaux).

## Chapitre 6

# Modèles à effets aléatoires et modèles mixtes

*Un modèle mixte est un modèle comportant à la fois des facteurs à effets fixes, tels qu'ils ont été introduits au chapitre 3, et des facteurs à effets aléatoires, notion nouvelle, un peu particulière, introduite au début de ce chapitre. Les méthodes usuelles dans le modèle linéaire standard, estimations, tests et prévisions, deviennent assez délicates dans le cadre d'un modèle mixte. Certaines sont détaillées dans ce chapitre, d'autres seront simplement évoquées.*

*Les références bibliographiques les plus importantes sur les modèles linéaires mixtes sont les ouvrages de Miller (1997), Searle et al. (1992) et Verbeke & Molenberghs (2000).*

### Résumé

Une nouvelle notion est introduite dans ce chapitre : celle de facteur à effets aléatoires. Jusqu'à présent, les facteurs considérés dans les chapitres 3, 4 et 5 étaient des facteurs à effets fixes : les différents niveaux en étaient fixés une fois pour toutes et les effets associés étaient des paramètres à estimer, ces paramètres intervenant dans la moyenne du modèle. Les facteurs à effets aléatoires vont avoir, a priori, une grande quantité de niveaux, les observations réalisées correspondant à un nombre restreint de ces niveaux, pris aléatoirement. On va ainsi modéliser ces niveaux en tant qu'observations d'une variable aléatoire normale, de moyenne nulle (la moyenne du modèle sera définie par les effets fixes) et de variance inconnue, à estimer. Chaque facteur à effets aléatoires sera donc caractérisé par un paramètre de variance qu'il faudra estimer en plus de la variance des erreurs du modèle. D'où le nom de **composantes de la variance** qu'on rencontre également pour de tels modèles.

On appelle modèles mixtes des modèles comportant à la fois des facteurs à effets fixes (ces effets entrant dans la définition de la moyenne du modèle) et des facteurs à effets aléatoires (ces effets entrant, quant à eux, dans la définition de la variance du modèle). La nécessité d'estimer simultanément plusieurs paramètres de moyenne et plusieurs paramètres de variances dans les modèles mixtes va compliquer la procédure d'estimation. Ainsi, la méthode du maximum de vraisemblance, qui entraîne un biais systématique dans l'estimation de la variance, n'est pas la plus appropriée dans ce cas : on lui préfère, en général, la méthode dite du **maximum de vraisemblance restreint**.

Les tests de significativité des effets aléatoires sont encore des tests de Fisher dans le cas de plans d'expériences équilibrés. Mais, les statistiques de khi-deux intervenant au dénominateur de la statistique de Fisher diffèrent parfois de ce qu'on a vu au chapitre 3, afin de tenir compte de la particularité des modèles mixtes. D'autre part, dans le cas déséquilibré, les tests de Fisher sont en fait des tests approchés et sont d'un usage plus délicat.

## 6.1 Modèle à un facteur à effets aléatoires

Les deux premiers paragraphes de ce chapitre sont consacrés à des modèles comportant uniquement des facteurs à effets aléatoires. Ces modèles ont peu d'intérêt dans la pratique ; nous les traitons essentiellement dans un but pédagogique, afin d'introduire de manière progressive les divers constituants des modèles mixtes. Notons que le seul effet fixe d'un modèle à effets aléatoires est l'effet (moyen) général.

### 6.1.1 Écriture du modèle pour une observation

Le modèle standard d'analyse de variance (ANOVA) à un seul facteur (à effets fixes) a été défini, au chapitre 3, sous l'une des formes suivantes :

$$Y_{ij} = \beta_j + U_{ij} = \mu + \alpha_j + U_{ij} = m + a_j + U_{ij}.$$

L'écriture de gauche est l'écriture initiale, la suivante correspond au paramétrage centré, la dernière correspond au paramétrage SAS. Dans ces écritures, les différents paramètres intervenant ( $\beta_j$ ,  $\mu$ ,  $\alpha_j$ ,  $m$ ,  $a_j$ ) sont des réels non aléatoires qu'il convient d'estimer par une méthode appropriée, en général le maximum de vraisemblance. On les appelle parfois des **effets fixes** (c'est-à-dire non aléatoires) et il arrive que l'on parle de **modèle à effets fixes** pour désigner l'ANOVA. Dans un tel modèle, les seules variations aléatoires que l'on envisage sont celles liées à la variable aléatoire réelle (v.a.r.)  $U_{ij}$  qui représente l'erreur du modèle, autrement dit les variations aléatoires non identifiées, non expliquées.

Dans la pratique, il peut se faire que l'on souhaite intégrer dans le modèle des variations aléatoires liées à un phénomène connu ; par exemple, les réactions aléatoires de différents individus à l'absorption d'un médicament donné. Pour ce faire, on intègre une v.a.r. autre que  $U_{ij}$  dans le modèle qui devient ainsi un modèle à effets aléatoires. On ne mélangera pas les effets de variance, que l'on va chercher à prendre en compte à travers la v.a.r. en question, avec les effets de moyenne, que l'on continuera à attribuer aux facteurs à effets fixes mis dans le modèle. Ainsi, on modélisera ces effets aléatoires au moyen d'une v.a.r. de moyenne nulle et de variance inconnue, à estimer. Pour des raisons de cohérence, dans le cadre d'un modèle gaussien, on ne considèrera que des v.a.r. normales (gaussiennes), de sorte qu'un modèle à un seul facteur à effets aléatoires (et sans effet fixe autre que l'effet général) s'écrira sous la forme suivante :

$$Y_{ij} = \mu + A_j + U_{ij}$$

(on notera que, compte tenu de la spécificité d'un tel modèle, cette écriture constituera le seul paramétrage considéré ici).

Dans l'écriture ci dessus :

- $Y_{ij}$  est la v.a.r. réponse ;
- $\mu$  est un effet fixe, non aléatoire, à estimer (c'est l'effet général, unique effet fixe entrant dans ce modèle) ;
- $A_j$  ( $j = 1, \dots, J$ ) est une v.a.r.  $\mathcal{N}(0, \sigma_a^2)$  ; les différentes v.a.r.  $A_j$  sont supposées indépendantes et de même loi ; elles sont donc i.i.d., gaussiennes, centrées, de même variance inconnue ; ainsi, les  $J$  niveaux du facteur à effets aléatoires considéré sont  $J$  observations indépendantes de cette loi  $\mathcal{N}(0, \sigma_a^2)$  ;
- $U_{ij} \sim \mathcal{N}(0, \sigma^2)$ , les  $U_{ij}$  étant également i.i.d. ; pour un niveau  $j$  fixé du facteur aléatoire, on réalise  $n_j$  observations indicées par  $i$ , de sorte que l'indice  $i$  varie de 1 à  $n_j$  et que le nombre total d'observations est  $n = \sum_{j=1}^J n_j$  ;
- on suppose de plus que chaque v.a.r.  $A_j$  est indépendante de chaque v.a.r.  $U_{ij'}$ ,  $\forall (i, j, j')$ .

On déduit ainsi de ce modèle

$$\mathbb{E}(Y_{ij}) = \mu, \quad \text{Var}(Y_{ij}) = \sigma_a^2 + \sigma^2,$$

les deux termes de variance  $\sigma_a^2$  et  $\sigma^2$  étant inconnus et devant être estimés, de même que  $\mu$ .

Les problèmes que nous allons aborder dans ce paragraphe sont l'estimation des paramètres  $\mu$ ,  $\sigma_a^2$  et  $\sigma^2$ , le test de significativité du modèle, autrement dit du facteur à effets aléatoires, ainsi que la prédiction des effets  $\mu + A_j$ .

**Remarque 49** On notera que les deux termes de variance définis ci-dessus ne dépendent pas des indices  $i$  et  $j$  : le modèle reste donc homoscédastique.

**Remarque 50** Pour un même indice  $j$  et deux indices  $i$  et  $i'$  différents, on a :

$$\text{Cov}(Y_{ij}, Y_{i'j}) = \text{Cov}(A_j + U_{ij}, A_j + U_{i'j}) = \sigma_a^2 ;$$

il n'y a donc pas indépendance entre les deux v.a.r.  $Y_{ij}$  et  $Y_{i'j}$  : ceci est un élément nouveau et très important dans les modèles à effets aléatoires, comme dans les modèles mixtes.

**Remarque 51** De façon concrète, on décide de considérer un facteur comme étant à effets aléatoires lorsque les  $J$  niveaux de ce facteur ne sont pas les seuls qui intéressent l'expérimentateur, mais sont  $J$  niveaux pris au hasard dans une population de niveaux quasiment infinie ou, en tout cas, très importante. Le choix de ces niveaux est donc lui-même aléatoire et doit se traduire par l'introduction d'un facteur à effets aléatoires dans le modèle, afin de pouvoir appliquer ce dernier à tout niveau du facteur.

**Remarque 52** En pratique, comme facteurs à effets aléatoires, on trouve des individus (les "sujets" d'une étude médicale, biologique, génétique...), des animaux (même chose, avec une étude pharmacologique, vétérinaire...), des variétés végétales (étude agronomique ou autre), voire des blocs dans une étude agronomique.

### 6.1.2 Écriture matricielle du modèle

Comme indiqué plus haut, on suppose que l'on fait  $n_j$  observations (non indépendantes) de la v.a.r. réponse  $Y$  au niveau  $j$  du facteur aléatoire ; on supposera  $n_j \geq 1$ , de sorte que l'on ne considèrera ici que des plans complets ; le nombre total d'observations réalisées est noté  $n$  ( $n = \sum_{j=1}^J n_j$ ).

Sous forme matricielle, le modèle s'écrit :

$$Y = \mu \mathbb{1}_n + \mathbf{Z}A + U.$$

Dans l'écriture ci-dessus :

- $Y$  et  $U$  sont des vecteurs aléatoires de  $\mathbb{R}^n$  (que l'on supposera muni de la base canonique) dont les composantes sont, respectivement, les v.a.r.  $Y_{ij}$  et  $U_{ij}$  ;
- le vecteur  $\mathbb{1}_n$  est le vecteur de  $\mathbb{R}^n$  dont toutes les composantes sur la base canonique valent 1 ;
- la matrice  $\mathbf{Z}$ , de dimension  $n \times J$ , comporte dans ses colonnes les indicatrices des niveaux du facteur considéré : elle ne contient donc que des 0 et des 1 ;
- enfin, le vecteur

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_J \end{pmatrix}$$

est un vecteur aléatoire gaussien de  $\mathbb{R}^J$  (ici,  $p = J$ , comme en ANOVA à un seul facteur) :  $A \simeq \mathcal{N}_J(0, \sigma_a^2 \mathbf{I}_J)$ , où  $\mathbf{I}_J$  est la matrice identité d'ordre  $J$ .

On a :  $\text{Var}(\mathbf{Z}A) = \mathbf{Z}\text{Var}(A)\mathbf{Z}' = \sigma_a^2 \mathbf{Z}\mathbf{Z}'$  ; on en déduit :  $\text{Var}(Y) = \sigma_a^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 \mathbf{I}_n = \mathbf{V}$ . On obtient ainsi :  $Y \sim \mathcal{N}_n(\mu \mathbb{1}_n, \mathbf{V})$ . On dit que  $\sigma_a^2$  et  $\sigma^2$  sont les composantes de la variance  $\mathbf{V}$  de  $Y$ .

**Exemple 11** Considérons le cas très simple dans lequel  $J = 2$ ,  $n_1 = 2$  et  $n_2 = 3$  (on a donc  $n = 5$ ). Il vient :

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} ; \quad A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} ; \quad \mathbf{V} = \left( \begin{array}{cc|ccc} \sigma_a^2 + \sigma^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 + \sigma^2 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_a^2 + \sigma^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 + \sigma^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma^2 \end{array} \right).$$

**Remarque 53** Comme dans le modèle à un seul facteur à effets fixes, on notera que, au sein d'un même niveau du facteur à effets aléatoires, les observations de ce facteur sont constantes. Seules les observations des v.a.r.  $U_{ij}$  changent :  $y_{ij} = \mu + a_j + u_{ij}$  ;  $y_{i'j} = \mu + a_j + u_{i'j}$  ( $a_j$  est l'observation de la v.a.r.  $A_j$ ).

### 6.1.3 Estimation de la moyenne

#### Cas général

Le cas général correspond au cas où le plan considéré peut être déséquilibré, autrement dit au cas où les effectifs  $n_j$  ne sont pas nécessairement tous égaux. Alors, dans le modèle écrit ci-dessus, la matrice  $\mathbf{V}$  des variances-covariances du vecteur aléatoire associé aux observations a une structure irrégulière (ses blocs diagonaux n'ont pas tous la même dimension; ainsi, dans l'exemple ci-dessus, il y a un bloc carré d'ordre 2 et un autre carré d'ordre 3). Cela entraîne une modification de l'expression usuelle de  $\hat{\mu}$ , estimation de  $\mu$ . On notera, d'ailleurs, que cette expression (que nous donnons plus bas) est la même dans le cas gaussien, avec estimation par maximum de vraisemblance, et dans le cas sans hypothèse gaussienne, avec estimation par moindres carrés (au demeurant, dans ce dernier cas, la technique appropriée est la méthode des moindres carrés généralisés, associée à la métrique de Mahalanobis, définie par la matrice  $\mathbf{V}^{-1}$ ). On se reportera au point 6.3.3 pour plus de détails.

Posons  $\bar{Y}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ , moyenne empirique des composantes du vecteur aléatoire  $Y$  correspondant au niveau  $j$  du facteur. On peut vérifier que l'on a :

$$\mathbb{E}(\bar{Y}_{\bullet j}) = \mu; \quad \text{Var}(\bar{Y}_{\bullet j}) = \frac{1}{n_j^2} \text{Var}\left(\sum_{i=1}^{n_j} Y_{ij}\right) = \frac{\sigma_a^2 + \sigma^2}{n_j} + \frac{n_j(n_j - 1)\sigma_a^2}{n_j^2} = \sigma_a^2 + \frac{\sigma^2}{n_j} = \tau_j^2.$$

L'estimation (maximum de vraisemblance ou moindres carrés généralisés) du paramètre de moyenne  $\mu$  est alors :

$$\hat{\mu} = \sum_{j=1}^J w_j \bar{y}_{\bullet j}, \quad \text{avec } w_j = \frac{\frac{1}{\tau_j^2}}{\sum_{j=1}^J \frac{1}{\tau_j^2}}.$$

**Remarque 54** On notera que le calcul de  $\hat{\mu}$  donné ci-dessus nécessite la connaissance des poids  $w_j$ , donc des éléments  $\tau_j^2$ , autrement dit des composantes de la variance  $\sigma_a^2$  et  $\sigma^2$ . Comme ces composantes sont inconnues, elles doivent être estimées avant de pouvoir calculer l'estimation de la moyenne.

**Remarque 55** La justification de l'expression de  $\hat{\mu}$  est donnée en 6.3.3, dans la remarque 71.

#### Cas équilibré

Dans ce cas, on a  $n_j = n_0, \forall j$ , et  $n = n_0 J$ . Les éléments  $\tau_j$  ne dépendent plus de  $j$  et les poids  $w_j$  sont tous égaux à  $\frac{1}{J}$ . On en déduit immédiatement :

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_0} y_{ij} = \bar{y}_{\bullet\bullet} \quad (\text{moyenne générale des observations } y_{ij}).$$

On voit que, dans ce cas, le calcul effectif de  $\hat{\mu}$  peut se faire avant d'estimer les composantes de la variance.

### 6.1.4 Estimation des composantes de la variance

Dans les modèles à effets aléatoires, il existe différentes méthodes d'estimation des composantes de la variance (les paramètres  $\sigma_a^2$  et  $\sigma^2$  dans le cas d'un seul facteur). Dans les cas simples (en particulier celui d'un plan équilibré), les différentes méthodes peuvent être équivalentes, mais ce n'est plus vrai dans le cas général. De plus, il n'y a pas de méthode qui soit uniformément meilleure que les autres, d'où les difficultés.

**Estimation par ANOVA**

Indiquons tout de suite que cette appellation n'est pas vraiment appropriée, même si c'est la plus couramment utilisée dans la littérature statistique (raison pour laquelle nous l'utilisons). On trouve également l'appellation de "méthode de type I" dans le logiciel SAS, ce qui est tout aussi peu approprié. Il s'agit en fait d'une méthode d'estimation de type moments : on écrit un système d'équations en égalant moments empiriques et moments théoriques (d'ordre 1) de certaines sommes de carrés ; la solution de ce système fournit des valeurs pour les composantes de la variance et ces valeurs sont prises comme estimations de ces composantes. On pourrait donc encore l'appeler "méthode par sommes de carrés", ou "méthode par formes quadratiques", mais nous utiliserons plutôt l'appellation usuelle.

Considérons donc les deux sommes de carrés définies ci-dessous.

- $SSA$  est la somme des carrés des écarts pris en compte par le modèle, c'est-à-dire par le facteur aléatoire  $A$ . On a :

$$SSA = \sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2.$$

Le calcul de l'espérance mathématique de  $SSA$ , un peu lourd, conduit au résultat suivant :

$$\mathbb{E}(SSA) = (J - 1)\sigma^2 + \frac{1}{n}(n^2 - \sum_{j=1}^J n_j^2)\sigma_a^2.$$

On notera que, dans le cas équilibré ( $n_j = n_0, \forall j$ ), cette espérance se simplifie et devient :

$$\mathbb{E}(SSA) = (J - 1)(\sigma^2 + n_0\sigma_a^2).$$

Par ailleurs, les solutions du système que l'on va écrire feront intervenir le carré moyen relatif au facteur  $A$ , qui sera noté  $MSA$  :  $MSA = \frac{1}{J-1}SSA$  ( $SSA$  est une quantité à  $J-1$  degrés de liberté). On obtient :

$$\mathbb{E}(MSA) = \sigma^2 + \frac{1}{n(J-1)}(n^2 - \sum_{j=1}^J n_j^2)\sigma_a^2,$$

qui devient, dans le cas équilibré :

$$\mathbb{E}(MSA) = \sigma^2 + n_0\sigma_a^2.$$

- $SSE$  est la somme des carrés des écarts résiduels (ceux associés aux erreurs du modèle, c'est-à-dire non pris en compte par le modèle) :

$$SSE = \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2.$$

L'espérance de  $SSE$  s'écrit :

$$\mathbb{E}(SSE) = (n - J)\sigma^2.$$

On utilisera encore le carré moyen relatif aux erreurs, noté  $MSE$  :  $MSE = \frac{SSE}{n - J}$  ( $SSE$  est à  $n - J$  degrés de liberté). Son espérance vaut  $\mathbb{E}(MSE) = \sigma^2$ .

On définit ensuite un système de 2 équations à 2 inconnues en égalant les espérances mathématiques de  $MSA$  et  $MSE$  avec les valeurs observées de ces 2 carrés moyens qui seront respectivement notées  $MSA(y)$  et  $MSE(y)$ . Ce système est le suivant :

$$\begin{aligned} MSA(y) &= \sigma^2 + \frac{1}{n(J-1)}(n^2 - \sum_{j=1}^J n_j^2)\sigma_a^2 ; \\ MSE(y) &= \sigma^2. \end{aligned}$$

La résolution (immédiate) fournit les estimations ANOVA des composantes de la variance. L'estimation de la composante relative au facteur aléatoire  $A$  sera indiquée par 1, afin de ne pas la confondre avec les autres estimations obtenues plus loin.

$$\begin{aligned}\hat{\sigma}^2 &= MSE(y) ; \\ \hat{\sigma}_{a_1}^2 &= \frac{n(J-1)[MSA(y) - MSE(y)]}{n^2 - \sum_{j=1}^J n_j^2}.\end{aligned}$$

Par la suite, nous noterons  $\hat{\Sigma}^2$  la statistique (la v.a.r.) dont  $\hat{\sigma}^2$  est l'observation ( $\hat{\Sigma}^2 = MSE$ ) et  $\hat{\Sigma}_{a_1}^2$  la statistique dont  $\hat{\sigma}_{a_1}^2$  est l'observation :

$$\hat{\Sigma}_{a_1}^2 = \frac{n(J-1)[MSA - MSE]}{n^2 - \sum_{j=1}^J n_j^2}$$

(dans les expressions ci-dessus,  $\hat{\Sigma}^2$  et  $\hat{\Sigma}_{a_1}^2$  sont les estimateurs, tandis que  $\hat{\sigma}^2$  et  $\hat{\sigma}_{a_1}^2$  sont les estimations).

#### Propriétés des estimateurs ANOVA

- On a nécessairement :  $\hat{\sigma}^2 \geq 0$ . Par contre, le signe de  $\hat{\sigma}_{a_1}^2$  est celui de  $MSA(y) - MSE(y)$  et peut être négatif. Dans ce cas, par convention, on posera  $\hat{\sigma}_{a_1}^2 = 0$ , ce qui revient à considérer que chaque niveau  $A_j$  est presque sûrement nul (comme observation d'une loi normale de moyenne et de variance nulles). Le facteur aléatoire est donc sans effet et on doit, dans ce cas, le retirer du modèle.
- On peut montrer la relation suivante :

$$\frac{(n-J)\hat{\Sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi_{n-J}^2.$$

- On en déduit immédiatement que  $\hat{\Sigma}^2$  est un estimateur sans biais et convergent de  $\sigma^2$ .
- Par ailleurs, on vérifie sans difficulté que  $\hat{\Sigma}_{a_1}^2$  est un estimateur sans biais de  $\sigma_a^2$ .
- Lorsque le plan considéré est équilibré, l'expression de  $\hat{\Sigma}_{a_1}^2$  se simplifie et devient :

$$\hat{\Sigma}_{a_1}^2 = \frac{MSA - MSE}{n_0}.$$

- Toujours dans le cas d'un plan équilibré, on peut aussi montrer la relation suivante :

$$\frac{SSA}{n_0\sigma_a^2 + \sigma^2} \sim \chi_{J-1}^2$$

(voir, par exemple, Searle *et al.*, 1992). Par contre, il n'y a pas de résultat de ce type dans le cas déséquilibré.

- Les statistiques  $SSA$  et  $SSE$  sont indépendantes (que le plan soit équilibré ou non ; on pourra encore se reporter à Searle *et al.*, 1992).
- Par contre,  $\hat{\Sigma}_{a_1}^2$  et  $\hat{\Sigma}^2$  ne sont pas indépendantes (c'est immédiat, compte tenu de l'expression de  $\hat{\Sigma}_{a_1}^2$ ).
- On n'a pas de résultat de convergence concernant  $\hat{\Sigma}_{a_1}^2$ , qu'on soit dans le cas équilibré ou déséquilibré.
- Enfin, on sait écrire  $\text{Var}(\hat{\Sigma}_{a_1}^2)$  et  $\text{Cov}(\hat{\Sigma}_{a_1}^2, \hat{\Sigma}^2)$  dans tous les cas, mais les expressions sont compliquées, surtout dans le cas déséquilibré.

#### Estimation par maximum de vraisemblance

La vraisemblance de l'échantillon s'écrit :

$$L(y, \mu, \sigma_a^2, \sigma^2) = \frac{1}{(2\pi)^{n/2} (\det \mathbf{V})^{1/2}} \exp\left[-\frac{1}{2}(y - \mu \mathbf{1}_n)' \mathbf{V}^{-1} (y - \mu \mathbf{1}_n)\right],$$

où  $\mathbf{V}$  a été définie en 6.1.2 et vaut :  $\mathbf{V} = \sigma_a^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 \mathbf{I}_n$ .

La log-vraisemblance s'écrit donc :

$$l(y, \mu, \sigma_a^2, \sigma^2) = \log[L(y, \mu, \sigma_a^2, \sigma^2)] = \text{constante} - \frac{1}{2} \log(\det \mathbf{V}) - \frac{1}{2} (y - \mu \mathbb{I}_n)' \mathbf{V}^{-1} (y - \mu \mathbb{I}_n).$$

Il est possible d'expliciter  $\det \mathbf{V}$ , ainsi que  $\mathbf{V}^{-1}$ , en fonction de  $\sigma_a^2$  et  $\sigma^2$ . On peut alors dériver  $l(y, \mu, \sigma_a^2, \sigma^2)$  selon ces 2 variables et obtenir les équations de vraisemblance. Toutefois, les solutions de ces équations ne sont explicites que dans le cas équilibré.

*Cas équilibré*

La résolution des équations de vraisemblance conduit aux expressions suivantes :

$$\hat{\Sigma}^2 = MSE; \quad \hat{\Sigma}_{a2}^2 = \frac{1}{n} \left[ \left(1 - \frac{1}{J}\right) MSA - MSE \right].$$

L'expression de  $\hat{\Sigma}^2$  est la même que celle obtenue par ANOVA ; il en va donc de même pour ses propriétés. Par contre, ce n'est pas le cas pour l'expression ci-dessus de  $\hat{\Sigma}_{a2}^2$  qui est différente de celle obtenue par ANOVA dans le cas équilibré ( $\hat{\Sigma}_{a1}^2$ ). Il est encore possible que sa valeur calculée  $\hat{\sigma}_{a2}^2$  soit négative ; comme précédemment, elle est alors ramenée à 0. Le problème majeur de cet estimateur  $\hat{\Sigma}_{a2}^2$  est qu'il est biaisé pour  $\sigma_a^2$ , sans qu'on puisse en corriger le biais. En effet, on peut vérifier sans difficulté :  $\mathbb{E}(\hat{\Sigma}_{a2}^2) = \frac{1}{n_0 J^2} [n_0 (J-1) \sigma_a^2 - \sigma^2]$ . Enfin, signalons qu'on sait encore expliciter dans ce cas la matrice des variances-covariances du couple  $(\hat{\Sigma}^2, \hat{\Sigma}_{a2}^2)$ .

*Cas déséquilibré*

Dans ce cas, il n'y a pas de solution analytique des équations de vraisemblance et l'on a recours à une méthode numérique de résolution d'un système non linéaire (méthode itérative, de type Fisher scoring). La solution  $\hat{\Sigma}_{a2}^2$  obtenue dans ce cas est, en général, encore biaisée, sans que l'on puisse écrire explicitement le biais, et encore moins le corriger.

**Remarque 56** *Le caractère biaisé de l'estimateur maximum de vraisemblance  $\hat{\Sigma}_{a2}^2$  vient de ce que la log-vraisemblance donnée plus haut ne sépare pas le paramètre de moyenne ( $\mu$ ) des paramètres de variance ( $\sigma_a^2$  et  $\sigma^2$ ). Toutefois, il est possible de faire une décomposition appropriée de la vraisemblance : c'est le principe de la méthode du maximum de vraisemblance restreint.*

### Estimation par maximum de vraisemblance restreint

Il est donc possible de factoriser la vraisemblance  $L(y, \mu, \sigma_a^2, \sigma^2)$  selon deux termes : le premier contient  $\mu$ , et sa maximisation conduit à l'expression de  $\hat{\mu}$  donnée en 6.1.3 ; le second ne dépend que de  $\sigma_a^2$  et de  $\sigma^2$ . La méthode du maximum de vraisemblance restreint consiste à maximiser le logarithme de ce seul second terme pour obtenir une estimation des composantes de la variance. Cette méthode, introduite par Patterson & Thomson (1971), est maintenant connue sous l'acronyme de REML (pour *REstricted -ou REsidual- Maximum Likelihood*, appellation introduite par Harville (1977) ; on trouve aussi les termes de vraisemblance résiduelle et de vraisemblance marginale). Elle fournit, en général, des estimateurs des composantes de la variance ayant de meilleures propriétés que les estimateurs maximum de vraisemblance. Toutefois, comme pour ces derniers, il n'y a de solution explicite que dans le cas équilibré. On trouvera plus de détails sur cette méthode en 6.3.3.

*Cas équilibré*

La résolution des équations, obtenues par dérivation de la partie de la log-vraisemblance ne dépendant que des termes de variance, conduit aux expressions suivantes :

$$\hat{\Sigma}^2 = MSE; \quad \hat{\Sigma}_{a3}^2 = \frac{MSA - MSE}{n_0}.$$

Notons tout d'abord que l'expression de  $\hat{\Sigma}^2$  est la même que dans les deux cas précédents. Par ailleurs, dans le cas équilibré, on notera que :  $\hat{\Sigma}_{a3}^2 = \hat{\Sigma}_{a1}^2 \neq \hat{\Sigma}_{a2}^2$ . Les propriétés de  $\hat{\Sigma}_{a3}^2$  sont donc les mêmes que celles de  $\hat{\Sigma}_{a1}^2$  ; en particulier, il s'agit d'un estimateur sans biais. Bien sûr, sa valeur calculée  $\hat{\sigma}_{a3}^2$  peut être négative et est alors ramenée à 0.

*Cas déséquilibré*

Qu'on utilise la vraisemblance complète ou la vraisemblance restreinte, les équations obtenues n'ont pas de solution analytique dans le cas déséquilibré. On est donc conduit, comme précédemment, à utiliser un algorithme de résolution numérique. Bien sûr, on n'a pas de résultat sur le biais des estimateurs, mais des simulations ont mis en évidence que les estimateurs obtenus par maximum de vraisemblance restreint sont, en général, "meilleurs" que ceux obtenus par maximum de vraisemblance.

**Estimations MINQUE et MIVQUE**

Avant que les ordinateurs ne fournissent des solutions commodes au problème de la résolution numérique d'un système d'équations non linéaires sans solution analytique, un certain nombre de statisticiens ont cherché à obtenir, de façon relativement simple, des solutions "raisonnables", à défaut d'être optimales, pour les estimateurs des composantes de la variance dans les modèles à effets aléatoires et les modèles mixtes. Parmi eux, C.R. Rao a publié, dans les années 1970-72, quatre articles sur le sujet (voir la bibliographie) proposant deux classes d'estimateurs : les estimateurs MINQUE et MIVQUE. Nous donnons, dans ce paragraphe, quelques éléments succincts sur ces estimateurs aujourd'hui peu utilisés. Notons toutefois que le logiciel SAS utilise des estimations MIVQUE pour initialiser l'algorithme REML.

L'idée générale est d'estimer toute combinaison linéaire des composantes de la variance, du type  $c_1\sigma_a^2 + c_2\sigma^2$ , par une forme quadratique des observations  $y_{ij}$  (en remarquant que tout estimateur d'une variance fait intervenir, d'une façon ou d'une autre, des sommes de carrés des observations, ce qui est aussi l'idée de l'estimation par ANOVA). Si  $\mathbf{Q}$  désigne une matrice réelle, carrée d'ordre  $n$  et symétrique, on prend donc pour estimation de  $c_1\sigma_a^2 + c_2\sigma^2$  la forme quadratique  $y'\mathbf{Q}y$ . On impose alors à la matrice  $\mathbf{Q}$  diverses propriétés. Tout d'abord, elle doit être telle que  $y'\mathbf{Q}y$  soit invariant par translation sur  $\mu$  (autrement dit, l'estimation de la variance ne doit pas dépendre de la moyenne); ensuite, l'estimation fournie doit être sans biais (pour  $c_1\sigma_a^2 + c_2\sigma^2$ ). Ceci ne suffisant pas pour obtenir un estimateur unique, on doit imposer une autre condition d'optimalité. Selon la condition choisie, on obtient l'une des deux classes précisées ci-dessous.

Si l'on impose, en plus, à  $y'\mathbf{Q}y$  de minimiser une certaine norme, on obtient un estimateur appelé *Minimum Norm Quadratic Unbiased Estimator*, d'où l'acronyme MINQUE (ce fut la première méthode proposée; en général, on lui préfère la suivante).

Si l'on impose maintenant à  $y'\mathbf{Q}y$  d'être de variance minimum, on obtient alors un estimateur appelé *Minimum Variance Quadratic Unbiased Estimator*, autrement dit MIVQUE (cette seconde propriété est plus naturelle pour un estimateur).

Concrètement, ces deux procédures fonctionnent en une seule itération : on choisit tout d'abord une valeur initiale pour chaque composante de la variance; on met ensuite à jour ces composantes par un calcul unique faisant intervenir  $\mathbf{Q}$  et  $y$ . Par construction, les estimateurs ainsi définis sont sans biais, mais ils peuvent encore conduire à des solutions négatives. D'autre part, il est clair qu'ils dépendent du choix de la solution initiale.

**Remarque 57** Dans le cas équilibré, on notera tout d'abord que toutes les méthodes exposées ci-dessus conduisent à la même estimation de la composante résiduelle  $\sigma^2$  de la variance :

$$\hat{\Sigma}^2 = MSE.$$

D'autre part, en ce qui concerne l'autre composante  $\sigma_a^2$ , celle liée au facteur aléatoire  $A$ , les estimations par ANOVA, par maximum de vraisemblance restreint, ainsi que les estimations MINQUE et MIVQUE, conduisent au même résultat (toujours dans le cas équilibré) :

$$\hat{\Sigma}_{a1}^2 = \frac{MSA - MSE}{n_0}.$$

Par contre, la solution fournie par le maximum de vraisemblance est différente :

$$\hat{\Sigma}_{a2}^2 = \frac{1}{n_0J}[(1 - \frac{1}{J})MSA - MSE].$$

Cela met bien en évidence la faiblesse du maximum de vraisemblance dans ce cas : c'est la seule méthode conduisant à une estimation biaisée.

**Remarque 58** Dans le cas déséquilibré, les méthodes du maximum de vraisemblance et du maximum de vraisemblance restreint nécessitent l'usage d'un algorithme itératif, donc le choix de solutions initiales dont dépendent, bien sûr, les résultats obtenus. Il est souvent conseillé, et c'est le choix que nous préconisons, de prendre dans ce cas l'estimation par ANOVA comme solution initiale.

Il faut toutefois noter que ce n'est pas le choix du logiciel statistique SAS qui, dans les procédures VARCOMP et MIXED, utilise comme solution initiale une solution particulière de la méthode MIVQUE, appelée MIVQUE(0). La méthode MIVQUE nécessitant elle-même le choix de valeurs initiales (et opérant ensuite en une seule étape), on appelle MIVQUE(0) la méthode obtenue en prenant 1 pour valeur initiale de  $\sigma^2$  et 0 pour valeur initiale de  $\sigma_a^2$ .

### 6.1.5 Intervalles de confiance

#### Intervalle de confiance pour $\mu$

Dans le cas équilibré, on peut construire un intervalle de confiance pour  $\mu$  de la même façon que dans une ANOVA à un facteur (à partir de la loi de Student). En effet, on a dans ce cas

$$\hat{\mu} = \bar{Y}_{\bullet\bullet} \sim \mathcal{N}\left(\mu, \frac{n_0\sigma_a^2 + \sigma^2}{n}\right)$$

et, en utilisant le fait que  $\frac{SSA}{n_0\sigma_a^2 + \sigma^2} \sim \chi_{J-1}^2$  (indépendant de  $\hat{\mu}$ ), on obtient un intervalle de confiance, de type Student, de la forme :

$$\bar{Y}_{\bullet\bullet} \pm t \sqrt{\frac{MSA}{n}}.$$

Dans l'expression ci-dessus,  $t$  est le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi de Student à  $J - 1$  degrés de

liberté,  $MSA = \frac{1}{J-1} \sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$  et  $n = n_0 J$ .

Dans le cas déséquilibré, les choses sont plus délicates. En effet, dans ce cas, il vient :

$$\hat{\mu} = \sum_{j=1}^J w_j \bar{Y}_{\bullet j} \sim \mathcal{N}(\mu, v^2), \text{ avec : } v^2 = \frac{1}{\sum_{j=1}^J \frac{n_j}{n_j \sigma_a^2 + \sigma^2}}.$$

En utilisant les estimations des paramètres de variances  $\sigma_a^2$  et  $\sigma^2$ , on peut construire un intervalle approximatif à partir de la loi normale. Toutefois, on n'est même pas sûr que le risque asymptotique de cet intervalle soit  $\alpha$ , compte tenu qu'on n'est pas sûr de la convergence de l'estimateur  $\hat{\Sigma}_a^2$ .

#### Intervalle de confiance pour $\sigma^2$

Qu'on soit dans le cas équilibré ou déséquilibré, on a vu plus haut que :  $\frac{SSE}{\sigma^2} \sim \chi_{n-J}^2$ . En utilisant les quantiles appropriés de la loi de khi-deux, on en déduit donc un intervalle de confiance exact (non asymptotique) pour le paramètre  $\sigma^2$ .

#### Intervalle de confiance pour une fonction des deux paramètres de variance

Dans le cas équilibré, on peut construire des intervalles de confiance exacts pour les rapports  $\frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}$ ,  $\frac{\sigma^2}{\sigma_a^2 + \sigma^2}$  et  $\frac{\sigma_a^2}{\sigma^2}$ , à partir des lois de Fisher appropriées. Toutefois, en pratique, ces intervalles ne sont pas d'un grand intérêt. Par ailleurs, ils ne sont valables que dans le cas équilibré.

#### Intervalle de confiance pour $\sigma_a^2$

Il n'existe pas d'intervalle de confiance exact pour le paramètre  $\sigma_a^2$ . On peut juste obtenir un intervalle approché dans le cas équilibré, mais nous ne le détaillons pas ici. Nous renvoyons encore une fois à Searle *et al.* (1992) pour plus de détails.

### 6.1.6 Test de l'effet du facteur

Tester la significativité du facteur aléatoire  $A$ , supposé de loi  $\mathcal{N}(0, \sigma_a^2)$ , revient à tester  $\{H_0 : \sigma_a^2 = 0\}$  contre  $\{H_1 : \sigma_a^2 > 0\}$ , avec un niveau  $\alpha$  fixé (cela revient encore, dans le cas d'un seul facteur, à tester la significativité du modèle).

Sous l'hypothèse nulle  $H_0$ , qu'on soit dans le cas équilibré ou déséquilibré, c'est le modèle constant qui est considéré et la statistique  $F = \frac{MSA}{MSE}$  est distribuée selon une loi de Fisher à  $(J - 1)$  et  $(n - J)$  degrés de liberté, comme dans le cas d'un facteur à effets fixes. On peut donc utiliser le test de Fisher standard pour tester  $H_0$ .

Par contre, sous  $H_1$ , la distribution de  $F$  est différente selon que l'on a affaire à un facteur à effets fixes, à un facteur à effets aléatoires avec un plan équilibré, ou encore à un facteur à effets aléatoires avec un plan déséquilibré. C'est donc la détermination de la puissance de ce test qui est délicate (nous ne l'aborderons pas dans ce cours).

**Remarque 59** *Il est important de bien voir que le test de significativité du facteur, dans le cas d'un seul facteur, est le même que ce facteur soit à effets fixes ou à effets aléatoires. On ne peut donc pas espérer s'en servir pour choisir entre les deux types d'effets. Redisons ici que ce choix dépend des conditions expérimentales et absolument pas d'une quelconque technique statistique.*

### 6.1.7 Prédiction d'un effet aléatoire

Il convient de faire attention au fait que la prédiction de la v.a.r.  $Y$ , lorsqu'elle n'est pas observée mais qu'on connaît le niveau du facteur  $A$  auquel est réalisée son observation, est assez délicate. Soit  $j$  le niveau en question et  $Y_{\ell j}$  la v.a.r. correspondante à prédire. On dispose de  $\mathbb{E}(Y_{\ell j}) = \mu$ , mais une prédiction par  $\hat{\mu}$  ne tient pas compte du niveau auquel on se trouve et ne convient donc pas. Par ailleurs,  $\bar{y}_{\bullet j}$  correspond à la prédiction qu'on ferait dans le cas d'un facteur à effets fixes et ne convient pas davantage, pas plus que  $\mathbb{E}(A_j) = 0$ .

La solution consiste en fait à prévoir  $Y_{\ell j}$  par  $\hat{\mu} + \mathbb{E}(A_j / \bar{Y}_{\bullet j} = \bar{y}_{\bullet j})$ . Autrement dit, on fait intervenir l'espérance conditionnelle de la v.a.r.  $A_j$ , sachant que la v.a.r.  $\bar{Y}_{\bullet j}$  prend la valeur  $\bar{y}_{\bullet j}$ , moyenne observée, au sein de l'échantillon, à ce niveau du facteur.

On sait que  $A_j \sim \mathcal{N}(0, \sigma_a^2)$  et que  $Y_{ij} \sim \mathcal{N}(\mu, \sigma_a^2 + \sigma^2)$ . On a vu en 6.1.3 que  $\bar{Y}_{\bullet j} \sim \mathcal{N}(\mu, \sigma_a^2 + \frac{\sigma^2}{n_j})$  et on vérifie sans difficulté que  $\text{Cov}(A_j, \bar{Y}_{\bullet j}) = \sigma_a^2$  (dans les calculs, on doit faire attention au fait que  $Y_{ij}$  et  $Y_{i'j}$  ne sont pas indépendantes). Cela permet d'écrire la loi conjointe des v.a.r.  $A_j$  et  $\bar{Y}_{\bullet j}$  :

$$\mathcal{N}_2 \left( \begin{pmatrix} 0 \\ \mu \end{pmatrix}; \begin{pmatrix} \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \frac{\sigma^2}{n_j} \end{pmatrix} \right).$$

Les règles usuelles du calcul de l'espérance conditionnelle d'une composante de loi normale bidimensionnelle sachant la seconde composante permettent d'écrire

$$\mathbb{E}(A_j / \bar{Y}_{\bullet j} = \bar{y}_{\bullet j}) = 0 + (\bar{y}_{\bullet j} - \mu) \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma^2}{n_j}} = (\bar{y}_{\bullet j} - \mu) \frac{\sigma_a^2}{\tau_j^2},$$

qui sera "estimée" par  $(\bar{y}_{\bullet j} - \hat{\mu}) \frac{\hat{\sigma}_a^2}{\hat{\tau}_j^2}$ . Finalement, la prédiction de  $Y_{\ell j}$  sera donnée par :

$$\hat{Y}_{\ell j} = \hat{\mu} + (\bar{y}_{\bullet j} - \hat{\mu}) \frac{\hat{\sigma}_a^2}{\hat{\tau}_j^2}.$$

### 6.1.8 Illustration

Il s'agit d'un exemple fictif à un seul facteur aléatoire, ce facteur comportant 4 niveaux. Les niveaux, notés 1, 2, 3 et 4, figurent en première colonne du fichier des données. La variable réponse figure dans la colonne suivante. Le plan est complet, avec répétitions, déséquilibré. Il y a 13 individus observés, donc 13 lignes dans le fichier des données reproduit ci-dessous.

**Les données**

```

1 9
1 10
1 11
1 12
2 15
2 16
2 17
3 13
3 13
3 14
3 15
4 25
4 28

```

**Le programme SAS**

Le programme SAS ci-dessous permet de traiter ces données en tenant compte du caractère aléatoire du facteur. Outre la procédure GLM, il est ici nécessaire d'utiliser la procédure VARCOMP pour estimer les composantes de la variance ( $\sigma_a^2$  et  $\sigma^2$ ). On notera que la méthode d'estimation doit être spécifiée, la méthode par défaut étant MIVQUE(0).

```

* ----- ;
* options facultatives pour la mise en page des sorties ;
* ----- ;
options linesize=76 pagesize=64 nodate;
title;
footnote 'Effets aleatoires - Exemple fictif';
* ----- ;
*          lecture des donnees          ;
*      (le fichier "fic.don" contient les donnees ;
*      et se trouve dans le repertoire de travail) ;
* ----- ;
data fic;
infile 'fic.don';
input effet $ reponse;
run;
* ----- ;
*          procedure GLM          ;
* ----- ;
proc glm data=fic;
class effet;
model reponse = effet / ss3;
random effet;
run;
quit;
* ----- ;
*          procedure VARCOMP          ;
*      (chacune des 4 options est utilisee successivement) ;
* ----- ;
proc varcomp data=fic method=type1;
class effet;
model reponse = effet;
run;
* ----- ;
proc varcomp data=fic method=mivque0;
class effet;
model reponse = effet;
run;
* ----- ;

```

```

proc varcomp data=fic method=ml;
class effet;
model reponse = effet;
run;
* ----- ;
proc varcomp data=fic method=reml;
class effet;
model reponse = effet;
run;
quit;

```

### Les sorties de la procédure GLM

```

PAGE 1                               The GLM Procedure
-----
                                Class Level Information

Class           Levels   Values
effet            4       1 2 3 4

                                Number of observations   13

```

```

PAGE 2                               The GLM Procedure
-----

Dependent Variable: reponse

Source           DF          Sum of Squares    Mean Square    F Value    Pr > F
Model            3          354.0576923      118.0192308    74.54     <.0001
Error            9          14.2500000        1.5833333
Corrected Total  12          368.3076923

                                R-Square    Coeff Var    Root MSE    reponse Mean
                                0.961310    8.261603    1.258306    15.23077

```

```

PAGE 3                               The GLM Procedure
-----

Source           Type III Expected Mean Square
effet            Var(Error) + 3.1795 Var(effet)

```

### Les sorties de la procédure VARCOMP

```

PAGE 1                               Variance Components Estimation Procedure
-----

                                Class Level Information

Class           Levels   Values
effet            4       1 2 3 4

```

Number of observations 13

Dependent Variable: reponse

Type 1 Analysis of Variance

Source	DF	Sum of Squares	Mean Square
effet	3	354.057692	118.019231
Error	9	14.250000	1.583333
Corrected Total	12	368.307692	.

Type 1 Analysis of Variance

Source	Expected Mean Square
effet	Var(Error) + 3.1795 Var(effet)
Error	Var(Error)
Corrected Total	.

Type 1 Estimates

Variance Component	Estimate
Var(effet)	36.62097
Var(Error)	1.58333

PAGE 2

Variance Components Estimation Procedure

-----

Class Level Information

Class	Levels	Values
effet	4	1 2 3 4

Number of observations 13

MIVQUE(0) SSQ Matrix

Source	effet	Error	reponse
effet	31.90533	9.53846	906.47337
Error	9.53846	12.00000	368.30769

MIVQUE(0) Estimates

Variance Component	reponse
Var(effet)	25.23143
Var(Error)	10.63656

PAGE 3

## Variance Components Estimation Procedure

-----

## Class Level Information

Class	Levels	Values
effet	4	1 2 3 4
Number of observations		13
Dependent Variable:		reponse

## Maximum Likelihood Iterations

Iteration	Objective	Var(effet)	Var(Error)
0	29.6188217655	12.0015563076	5.0593729821
1	23.0175866155	33.3583321167	1.6232181526
2	23.0101144107	35.1435488604	1.5855418629
3	23.0101132719	35.1211357555	1.5859817377
4	23.0101132718	35.1208834116	1.5859866946

Convergence criteria met.

Maximum Likelihood  
Estimates

Variance Component	Estimate
Var(effet)	35.12088
Var(Error)	1.58599

## Asymptotic Covariance Matrix of Estimates

	Var(effet)	Var(Error)
Var(effet)	640.34038	-0.28152
Var(Error)	-0.28152	0.56084

PAGE 4

## Variance Components Estimation Procedure

-----

## Class Level Information

Class	Levels	Values
effet	4	1 2 3 4
Number of observations		13
Dependent Variable:		reponse

## REML Iterations

Iteration	Objective	Var(effet)	Var(Error)
0	26.7855408891	13.0016859999	5.4809873972
1	19.3392040165	34.5534023567	1.7968792220
2	19.1250010581	49.6309606128	1.5586243312
3	19.1198401813	47.1506894694	1.5841973236
4	19.1198374354	47.0942665176	1.5848199498
5	19.1198374354	47.0942665176	1.5848199498

Convergence criteria met.

## REML Estimates

Variance Component	Estimate
Var(effet)	47.09427
Var(Error)	1.58482

## Asymptotic Covariance Matrix of Estimates

	Var(effet)	Var(Error)
Var(effet)	1520.6	-0.28093
Var(Error)	-0.28093	0.55920

## 6.2 Modèle à deux facteurs croisés à effets aléatoires

On suppose maintenant que la v.a.r. réponse  $Y$  dépend de deux facteurs à effets aléatoires notés  $A$  et  $B$  et, éventuellement, de leur interaction qui sera notée  $C$ . On note encore  $J$  le nombre de niveaux observés de  $A$  ( $J \geq 2$ ), ces niveaux étant indicés par  $j$ , et  $K$  le nombre de niveaux observés de  $B$  ( $K \geq 2$ ), ces niveaux étant maintenant indicés par  $k$ . Les deux facteurs  $A$  et  $B$  sont croisés et on note  $n_{jk}$  ( $n_{jk} \geq 1$ ) le nombre d'observations réalisées dans la cellule  $(j, k)$  du plan obtenu par ce croisement. Enfin, on pose  $n = \sum_{j=1}^J \sum_{k=1}^K n_{jk}$  :  $n$  est le nombre total d'observations réalisées.

### 6.2.1 Écritures du modèle

Pour une observation  $Y_{ijk}$  de la v.a.r. réponse  $Y$ , le modèle s'écrit :

$$Y_{ijk} = \mu + A_j + B_k + C_{jk} + U_{ijk}.$$

- Comme précédemment,  $\mu$  est l'effet général; c'est l'unique effet fixe de ce modèle.
  - $A_j$  est l'effet aléatoire du niveau  $j$  du facteur  $A$  ( $j = 1, \dots, J$ ); on suppose :  $A_j \sim \mathcal{N}(0, \sigma_a^2)$ .
  - $B_k$  est l'effet aléatoire du niveau  $k$  du facteur  $B$  ( $k = 1, \dots, K$ ); on suppose de même :  $B_k \sim \mathcal{N}(0, \sigma_b^2)$ .
  - $C_{jk}$  est l'effet aléatoire de l'interaction de  $A$  et de  $B$  dans la cellule  $(j, k)$  (l'interaction entre deux facteurs à effets aléatoires est nécessairement elle-même à effets aléatoires); on suppose maintenant :  $C_{jk} \sim \mathcal{N}(0, \sigma_c^2)$ .
  - $U_{ijk}$  est la v.a.r. erreur du modèle et l'on suppose :  $U_{ijk} \sim \mathcal{N}(0, \sigma^2)$ .
- Enfin, les v.a.r.  $A_j$ ,  $B_k$ ,  $C_{jk}$  et  $U_{ijk}$  sont supposées mutuellement indépendantes.

On peut réécrire le modèle sous la forme matricielle suivante :

$$\begin{aligned} Y &= \mu \mathbb{1}_n + \mathbf{Z}_1 \begin{pmatrix} A_1 \\ \vdots \\ A_J \end{pmatrix} + \mathbf{Z}_2 \begin{pmatrix} B_1 \\ \vdots \\ B_K \end{pmatrix} + \mathbf{Z}_3 \begin{pmatrix} C_1 \\ \vdots \\ C_{JK} \end{pmatrix} + U \\ &= \mu \mathbb{1}_n + \mathbf{Z}_1 A + \mathbf{Z}_2 B + \mathbf{Z}_3 C + U. \end{aligned}$$

Dans l'écriture ci-dessus,  $Y$  et  $U$  sont des vecteurs aléatoires de  $\mathbb{R}^n$  et  $\mathbb{1}_n$  est le vecteur dont toutes les coordonnées (sur la base canonique de  $\mathbb{R}^n$ ) sont égales à 1. D'autre part,  $\mathbf{Z}_1$  (respectivement  $\mathbf{Z}_2$ , resp.  $\mathbf{Z}_3$ ) est la matrice des indicatrices des niveaux de  $A$  (resp. de  $B$ , resp. des cellules), de dimension  $n \times J$  (resp.  $n \times K$ , resp.  $n \times JK$ ).

La loi de probabilité du vecteur aléatoire  $Y$  a maintenant pour expression

$$Y \sim \mathcal{N}_n(\mu \mathbb{1}_n ; \sigma_a^2 \mathbf{Z}_1 \mathbf{Z}_1' + \sigma_b^2 \mathbf{Z}_2 \mathbf{Z}_2' + \sigma_c^2 \mathbf{Z}_3 \mathbf{Z}_3' + \sigma^2 \mathbf{I}_n), \text{ soit } \mathcal{N}_n(\mu \mathbb{1}_n ; \mathbf{V}),$$

en posant ici :

$$\mathbf{V} = \sigma_a^2 \mathbf{Z}_1 \mathbf{Z}_1' + \sigma_b^2 \mathbf{Z}_2 \mathbf{Z}_2' + \sigma_c^2 \mathbf{Z}_3 \mathbf{Z}_3' + \sigma^2 \mathbf{I}_n.$$

**Remarque 60** Ici encore, comme dans le cas d'un seul facteur, on notera qu'il y a des corrélations entre observations, en raison de la nature aléatoire des effets considérés. Ainsi, deux observations  $Y_{ijk}$  et  $Y_{i'jk}$  de la même cellule  $(j, k)$  ont en commun les v.a.r.  $A_j$ ,  $B_k$  et  $C_{jk}$ , de sorte que leur covariance vaut :  $\sigma_a^2 + \sigma_b^2 + \sigma_c^2$ . De même, deux observations  $Y_{ijk}$  et  $Y_{i'jk'}$  de la même ligne  $j$  ont en commun la v.a.r.  $A_j$ , de sorte que leur covariance vaut  $\sigma_a^2$ . Enfin, on a encore un résultat de même nature pour deux observations d'une même colonne.

### 6.2.2 Estimation des composantes de la variance dans le cas équilibré

Dans toute la suite de cette section 6.2, on supposera que l'on a affaire à un plan équilibré, autrement dit on posera :  $n_{jk} = n_0 \geq 1$ ,  $\forall (j, k)$ . Il s'ensuit  $n = n_0 JK$ , et l'estimateur du paramètre de moyenne est  $\hat{\mu} = \bar{Y}_{\dots} = \frac{1}{n_0 JK} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_0} Y_{ijk}$ . On se reportera au paragraphe 6.3.3 pour le cas déséquilibré (à partir de deux facteurs à effets aléatoires, en présence d'un plan déséquilibré, on a recours aux procédures générales d'estimation relatives aux modèles mixtes, sur lesquelles nous reviendrons en 6.3.3).

#### Estimation par ANOVA

Le principe de cette méthode est toujours le même : on définit un système d'équations en égalant les espérances mathématiques de certains carrés moyens avec leur moyenne empirique. Chaque carré moyen est relatif à un terme de variance ( $A$ ,  $B$ ,  $C$  ou  $U$ ).

On définit ainsi :

$$\begin{aligned} SSA &= n_0 K \sum_{j=1}^J (\bar{Y}_{\bullet j \bullet} - \bar{Y}_{\dots})^2 ; \quad MSA = \frac{SSA}{J-1}. \\ SSB &= n_0 J \sum_{k=1}^K (\bar{Y}_{\bullet \bullet k} - \bar{Y}_{\dots})^2 ; \quad MSB = \frac{SSB}{K-1}. \\ SSC &= n_0 \sum_{j=1}^J \sum_{k=1}^K (\bar{Y}_{\bullet j k} - \bar{Y}_{\bullet j \bullet} - \bar{Y}_{\bullet \bullet k} + \bar{Y}_{\dots})^2 ; \quad MSC = \frac{SSC}{(J-1)(K-1)}. \\ SSE &= \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_0} (Y_{ijk} - \bar{Y}_{\bullet j k})^2 ; \quad MSE = \frac{SSE}{(n_0-1)JK}. \end{aligned}$$

On peut ensuite calculer (c'est assez fastidieux, mais pas vraiment difficile) les espérances mathématiques de ces carrés moyens.

$$\mathbb{E}(MSA) = \sigma^2 + n_0 \sigma_c^2 + n_0 K \sigma_a^2 ; \quad \mathbb{E}(MSB) = \sigma^2 + n_0 \sigma_c^2 + n_0 J \sigma_b^2 ;$$

$$\mathbb{E}(MSC) = \sigma^2 + n_0\sigma_c^2 ; \quad \mathbb{E}(MSE) = \sigma^2.$$

Enfin, en égalant ces espérances avec les observations correspondantes des quantités  $MSA$ ,  $MSB$ ,  $MSC$  et  $MSE$  sur l'échantillon considéré (ces observations seront respectivement notées  $MSA(y)$ ,  $MSB(y)$ ,  $MSC(y)$  et  $MSE(y)$ ), on obtient un système linéaire de quatre équations à quatre inconnues, triangulaire, dont on déduit immédiatement les estimations ANOVA des composantes de la variance :

$$\hat{\sigma}^2 = MSE(y) ; \quad \hat{\sigma}_c^2 = \frac{MSC(y) - MSE(y)}{n_0} ;$$

$$\hat{\sigma}_b^2 = \frac{MSB(y) - MSC(y)}{n_0J} ; \quad \hat{\sigma}_a^2 = \frac{MSA(y) - MSC(y)}{n_0K}.$$

En remplaçant ensuite les observations des moyennes de carrés par les statistiques correspondantes, on obtient les expressions analogues pour les estimateurs (il suffit d'enlever les  $(y)$ ) ; ces derniers seront respectivement notés  $\hat{\Sigma}^2$ ,  $\hat{\Sigma}_c^2$ ,  $\hat{\Sigma}_b^2$  et  $\hat{\Sigma}_a^2$ .

#### Propriétés des estimateurs ANOVA

- Il peut arriver que les valeurs calculées des estimateurs  $\hat{\Sigma}_a^2$ ,  $\hat{\Sigma}_b^2$  ou  $\hat{\Sigma}_c^2$  soient négatives ; dans ce cas, elle sont mises à zéro et le facteur correspondant est supprimé du modèle.
- Les quatre estimateurs  $\hat{\Sigma}_a^2$ ,  $\hat{\Sigma}_b^2$ ,  $\hat{\Sigma}_c^2$  et  $\hat{\Sigma}^2$  sont sans biais (c'est immédiat d'après les formules ci-dessus).
- Parmi les estimateurs sans biais, ils sont de variance minimum (admis).
- On peut encore vérifier :  $\frac{SSE}{\sigma^2} \sim \chi_{(n_0-1)JK}^2$ .
- On ne sait pas expliciter la loi de probabilité des trois autres estimateurs.

#### Autres méthodes d'estimation

Comme on ne considère que le cas équilibré dans ce paragraphe, les estimateurs ANOVA, REML, MINQUE et MIVQUE sont identiques. Seuls les estimateurs maximum de vraisemblance sont différents. On obtient encore  $MSE$  comme estimateur de  $\sigma^2$  par maximum de vraisemblance, mais les estimateurs des trois autres composantes de la variance sont en général différents de ceux explicités ci-dessus. De plus, ils sont biaisés.

**Remarque 61** Dans le cas déséquilibré, les différentes méthodes d'estimation fournissent, en général, des résultats différents (voir le point 6.3.3).

### 6.2.3 Tests des effets aléatoires dans le cas équilibré

#### Propriétés préliminaires

On a déjà signalé :

$$\frac{(n_0 - 1)JK}{\sigma^2} MSE \sim \chi_{(n_0-1)JK}^2 .$$

On peut également vérifier :

$$\frac{(J - 1) MSA}{n_0K\sigma_a^2 + n_0\sigma_c^2 + \sigma^2} \sim \chi_{J-1}^2 ; \quad \frac{(K - 1) MSB}{n_0J\sigma_b^2 + n_0\sigma_c^2 + \sigma^2} \sim \chi_{K-1}^2 ;$$

$$\frac{(J - 1)(K - 1) MSC}{n_0\sigma_c^2 + \sigma^2} \sim \chi_{(J-1)(K-1)}^2 .$$

De plus, les quatre statistiques ci-dessus sont indépendantes. Cela permet de définir des statistiques de tests, distribuées selon des lois de Fisher, pour tester les différentes hypothèses nulles relatives au modèle à deux facteurs croisés. Comme déjà vu pour des facteurs à effets fixes, on procède de façon hiérarchique, en commençant par tester les interactions.

**Test de  $\{H_0^c : \sigma_c^2 = 0\}$  contre  $\{H_1^c : \sigma_c^2 > 0\}$**

Sous  $H_0^c$ , il est clair que  $F_c = \frac{MSC}{MSE} \sim F_{(J-1)(K-1); (n_0-1)JK}$ .  $F_c$  est donc la statistique du test de  $H_0^c$  contre  $H_1^c$  et c'est encore la même statistique que pour tester l'interaction entre deux facteurs à effets fixes. Si  $H_0^c$  est rejetée, on conserve le modèle complet, avec les deux facteurs aléatoires et les interactions. Sinon, on enlève les interactions du modèle et on conserve le modèle additif.

**Test de  $\{H_0^a : \sigma_a^2 = 0\}$  contre  $\{H_1^a : \sigma_a^2 > 0\}$  dans le modèle complet**

Sous  $H_0^a$ ,  $F_a = \frac{MSA}{MSC} \sim F_{J-1; (J-1)(K-1)}$ .  $F_a$  est donc la statistique du test de  $H_0^a$  contre  $H_1^a$ .

**Test de  $\{H_0^b : \sigma_b^2 = 0\}$  contre  $\{H_1^b : \sigma_b^2 > 0\}$  dans le modèle complet**

De façon symétrique, on a maintenant, sous  $H_0^b$  :  $F_b = \frac{MSB}{MSC} \sim F_{K-1; (J-1)(K-1)}$ .  $F_b$  est la statistique du test de  $H_0^b$  contre  $H_1^b$ .

**Remarque 62** *On notera que les dénominateurs des deux dernières statistiques de tests sont MSC et non MSE. Autrement dit, ces tests ne sont pas les mêmes que ceux qu'on ferait dans le cadre d'un modèle à effets fixes.*

**Remarque 63** *Toujours en ce qui concerne ces deux derniers tests, on voit qu'ils sont les mêmes, que  $H_0^c$  soit vraie ou non. Si on le souhaite, on peut donc les utiliser pour tester les effets principaux au sein du modèle additif. Toutefois, ceci n'est pas l'optique du logiciel SAS.*

**Remarque 64** *On ne dispose plus de tels tests dans le cas déséquilibré (voir le point 6.3.5).*

**Remarque 65** *Concernant les intervalles de confiance, on dispose toujours du même intervalle que dans le cas d'un seul facteur pour les paramètres  $\mu$  et  $\sigma^2$ . Par contre, on n'a pas de résultat précis pour les autres composantes de la variance.*

**Remarque 66** *Pour déterminer les valeurs prédites une fois un modèle choisi, nous renvoyons au point 6.3.6.*

## 6.3 Modèles mixtes

On appelle modèle mixte un modèle statistique dans lequel on considère à la fois des facteurs à effets fixes (qui vont intervenir au niveau de la moyenne du modèle) et des facteurs à effets aléatoires (qui vont intervenir au niveau de la variance du modèle). Un modèle est dit mixte lorsqu'il y a au moins un facteur de chaque nature. Dans le cadre de ce cours, nous ne considérons que des modèles linéaires gaussiens mixtes, mais la notion de modèle mixte se rencontre également dans d'autres contextes, notamment dans le modèle linéaire généralisé. Dans la suite de ce paragraphe, nous ne spécifierons pas le nombre de facteurs à effets fixes, ni celui de facteurs à effets aléatoires : ils seront quelconques.

### 6.3.1 Écriture générale d'un modèle linéaire gaussien mixte

Un modèle linéaire gaussien mixte, relatif à  $n$  observations, s'écrit sous la forme matricielle suivante :

$$Y = \mathbf{X}\beta + \sum_{k=1}^K \mathbf{Z}_k A_k + U = \mathbf{X}\beta + \mathbf{Z}\mathbf{A} + U.$$

Nous précisons ci-dessous les éléments de cette écriture.

- $Y$  est le vecteur aléatoire réponse de  $\mathbb{R}^n$ .
- $\mathbf{X}$  est la matrice  $n \times p$  relative aux effets fixes du modèle (figurant en colonnes);  $p$  est donc le nombre total d'effets fixes pris en compte dans le modèle; la matrice  $\mathbf{X}$  est analogue à la matrice d'incidence dans une ANOVA.

- $\beta$  est le vecteur des  $p$  effets fixes  $\beta_j$  ( $j = 1, \dots, p$ ); il s'agit de paramètres à estimer.
- $\mathbf{Z}_k$  est la matrice des indicatrices (disposées en colonnes) des niveaux du  $k$ -ième facteur à effets aléatoires ( $k = 1, \dots, K$ ); on notera  $q_k$  le nombre de niveaux de ce facteur;  $\mathbf{Z}_k$  est donc de dimension  $n \times q_k$ .
- Nous allons noter  $A_{k\ell}$  la v.a.r. associée au  $\ell$ -ième niveau du  $k$ -ième facteur à effets aléatoires ( $\ell = 1, \dots, q_k$ ); pour tout  $\ell$ , on a  $A_{k\ell} \sim \mathcal{N}(0, \sigma_k^2)$ ; pour un indice  $k$  donné (autrement dit, pour un facteur déterminé), les v.a.r.  $A_{k\ell}$  sont supposées indépendantes (donc i.i.d.); bien entendu, deux observations de la v.a.r. réponse  $Y$  faites au même niveau  $\ell$  du  $k$ -ième facteur sont corrélées, leur covariance comportant le terme  $\sigma_k^2$  et, éventuellement, d'autres composantes de la variance; par ailleurs, pour deux indices  $k$  et  $k'$  distincts,  $A_{k\ell}$  et  $A_{k'\ell'}$  sont indépendantes, pour tous les niveaux  $\ell$  et  $\ell'$ .
- Dans l'écriture matricielle ci-dessus, on a posé  $A_k = \begin{pmatrix} A_{k1} \\ \vdots \\ A_{kq_k} \end{pmatrix}$ , de sorte que l'on a  $A_k \sim \mathcal{N}_{q_k}(0, \sigma_k^2 \mathbf{I}_{q_k})$ , les vecteurs aléatoires  $A_k$  étant mutuellement indépendants.
- $U$  est le vecteur aléatoire des erreurs du modèle; il vérifie  $U \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ ; de plus, il est supposé indépendant des  $A_k$ .
- Pour obtenir la seconde écriture (simplifiée) du modèle, qui est la forme la plus générale d'un modèle linéaire mixte, on a posé :

$$\mathbf{Z} = (\mathbf{Z}_1 | \dots | \mathbf{Z}_K) \quad ; \quad A = \begin{pmatrix} A_1 \\ \vdots \\ A_K \end{pmatrix}.$$

$\mathbf{Z}$  est une matrice (connue) de dimension  $n \times q$ , avec  $q = \sum_{k=1}^K q_k$  ( $q$  est le nombre total d'effets aléatoires considérés);  $A$  est un vecteur aléatoire gaussien de  $\mathbb{R}^q$ .

**Remarque 67** *Il n'est pas très courant (et, dans la mesure du possible, il vaut mieux l'éviter) de considérer des effets d'interactions entre un facteur à effets fixes et un autre facteur à effets aléatoires. Toutefois, dans certains cas pratiques, on peut être amené à le faire. Dans ce cas, les interactions en question doivent nécessairement être considérées comme des effets aléatoires. Dans l'écriture générale des modèles mixtes, elles sont donc intégrées dans la partie  $A$ .*

### Moments de $Y$

- De façon évidente, il vient :  $\mathbb{E}(Y) = \mathbf{X}\beta$ .
- D'autre part :  $\text{Var}(Y) = \mathbf{V} = \text{Var}(\mathbf{Z}A) + \text{Var}(U) = \sum_{k=1}^K \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k' + \sigma^2 \mathbf{I}_n$ .

Finalement, on obtient  $Y \sim \mathcal{N}_n(\mathbf{X}\beta, \mathbf{V})$ , les composantes de  $Y$  n'étant pas indépendantes au sein d'un même niveau d'un facteur aléatoire donné.

**Remarque 68** *Il est courant de poser :  $\mathbf{G} = \text{diag}(\sigma_1^2 \mathbf{I}_{q_1} \dots \sigma_K^2 \mathbf{I}_{q_K})$ . Cela permet d'écrire :  $\sum_{k=1}^K \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k' = \mathbf{Z} \mathbf{G} \mathbf{Z}'$ ; d'où :  $\mathbf{V} = \mathbf{Z} \mathbf{G} \mathbf{Z}' + \sigma^2 \mathbf{I}_n$  (la partie  $\sigma^2 \mathbf{I}_n$  de la variance est parfois notée  $\mathbf{R}$ ).*

**Remarque 69** *Il arrive que l'on intègre le vecteur aléatoire des erreurs  $U$  dans la partie aléatoire du modèle mixte ci-dessus, en remarquant que  $U$  et chaque  $A_k$  sont de même nature. On peut en*

*effet écrire :  $U = \mathbf{I}_n U = \mathbf{I}_n \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ , où  $U_i \sim \mathcal{N}(0, \sigma^2)$ . On peut ainsi considérer que  $U$  représente*

*un facteur à  $n$  effets aléatoires. Dans ce cas, on le note  $A_0$ , on pose  $\mathbf{Z}_0 = \mathbf{I}_n$ ,  $\mathbf{Z}^* = (\mathbf{Z}_0 | \mathbf{Z})$  et  $A^* = (A_0 | A)'$ , ce qui permet de réécrire :  $Y = \mathbf{X}\beta + \mathbf{Z}^* A^*$ . Nous utiliserons peu cette écriture simplifiée, pour éviter de mélanger une partie du modèle avec son erreur. Toutefois, elle pourra être implicite, comme par exemple dans l'estimation des composantes de la variance par ANOVA développée en 6.3.3 (cette réécriture du modèle mixte permet surtout d'alléger les notations par la suite).*

### 6.3.2 Estimation des paramètres dans le cas équilibré

#### Estimation de $\beta$

Le vecteur  $\beta$  des  $p$  paramètres  $\beta_j$  correspondant aux effets fixes du modèle a, dans ce cas, toujours la même expression, quelle que soit la méthode utilisée pour estimer les composantes de la variance. Il s'agit de l'expression fournie par la méthode des moindres carrés ordinaires, que nous noterons  $\text{OLSE}(\beta)$  (pour *Ordinary Least Squares Estimator*) :  $\hat{B} = \text{OLSE}(\beta) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$  (dans ce cas, il n'est pas nécessaire d'avoir estimé  $\mathbf{V}$  pour obtenir l'estimation de  $\beta$ ).

#### Estimation des composantes de la variance

Il s'agit ici d'estimer les  $K + 1$  paramètres de variance  $\sigma_k^2$  ( $k = 1, \dots, K$ ) et  $\sigma^2$ . Cela revient à estimer la matrice  $\mathbf{V}$  des variances-covariances de la variable réponse  $Y$ .

Le plan étant équilibré, les méthodes ANOVA, REML, MINQUE et MIVQUE, qui généralisent ce qui a été exposé en 6.1.4, conduisent toutes au même résultat explicite, solution d'un système de  $K + 1$  équations linéaires. Les estimateurs obtenus sont sans biais et de variance minimum parmi les estimateurs sans biais et quadratiques en les observations. On peut néanmoins obtenir des valeurs négatives pour certaines composantes de la variance : elles sont alors mises à 0 et le facteur correspondant doit être retiré du modèle.

De son côté, la méthode du maximum de vraisemblance fournit des solutions en général différentes des précédentes et biaisées.

En 6.3.3, nous revenons plus en détails sur les méthodes d'estimation de  $\mathbf{V}$  dans le cas général.

### 6.3.3 Estimation des paramètres dans le cas déséquilibré

#### Estimation de $\beta$

L'expression que l'on obtient dans ce cas pour  $\hat{B}$  fait intervenir l'estimation de la matrice des variances-covariances  $\mathbf{V}$  de  $Y$ . Si l'expression est unique, la valeur correspondante dépend de la méthode d'estimation de  $\mathbf{V}$ . En fait, l'expression obtenue est aussi celle fournie par la méthode des moindres carrés généralisés, pour cette raison notée  $\text{GLSE}(\beta)$  (pour *Generalized Least Squares Estimator*) :

$$\hat{B} = \text{GLSE}(\beta) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}Y.$$

Dans l'expression ci-dessus,  $\hat{\mathbf{V}} = \sum_{k=1}^K \hat{\sigma}_k^2 \mathbf{Z}_k \mathbf{Z}_k' + \hat{\sigma}^2 \mathbf{I}_n$  ; on voit donc que l'on doit estimer les composantes de la variance avant de pouvoir estimer le vecteur  $\beta$  (la justification de l'expression ci-dessus de  $\hat{B}$  est donnée dans le point traitant de l'estimation de  $\mathbf{V}$  par maximum de vraisemblance).

Quelle que soit la méthode utilisée pour estimer les composantes de la variance,  $\text{GLSE}(\beta)$  est un estimateur sans biais de  $\beta$ .

**Remarque 70** *On peut vérifier que, dans le cas d'un plan équilibré, on obtient  $\text{GLSE}(\beta) = \text{OLSE}(\beta)$ .*

#### Estimation de $\mathbf{V}$ par ANOVA

Cette méthode consiste à généraliser ici ce qui a été vu en 6.1.4, les sommes de carrés étant maintenant remplacées par des matrices de formes quadratiques sur  $\mathbb{R}^n$ . Pour chacun des facteurs à effets aléatoires, ainsi que pour la v.a.r. erreur  $U$ , on définit donc une matrice réelle  $\mathbf{C}_h$ , carrée d'ordre  $n$ , symétrique et strictement définie positive. En affectant l'indice 0 à  $U$ , l'indice  $h$  va ainsi varier de 0 à  $K$ . Ces matrices  $\mathbf{C}_h$ , autrement dit ces formes quadratiques, seront choisies de telle sorte que les équations obtenues soient commodes à résoudre (voir plus loin).

Si l'on considère un vecteur aléatoire  $Y$  de  $\mathbb{R}^n$ , de loi  $\mathcal{N}_n(\mu, \mathbf{V})$ , on sait que  $\mathbb{E}(Y'\mathbf{C}_h Y) = \mu'\mathbf{C}_h\mu + \text{tr}(\mathbf{C}_h\mathbf{V})$ . En appliquant cette formule au modèle mixte écrit plus haut, il vient :

$$Y \sim \mathcal{N}_n(\mathbf{X}\beta, \mathbf{V}) \implies \mathbb{E}(Y'\mathbf{C}_h Y) = \beta'\mathbf{X}'\mathbf{C}_h\mathbf{X}\beta + \text{tr}(\mathbf{C}_h\mathbf{V}).$$

Pour que l'estimation de  $\mathbf{V}$ , que l'on va faire au moyen de ces formes quadratiques, soit déconnectée de celle de  $\beta$ , on choisit les matrices  $\mathbf{C}_h$  telles que  $\mathbf{X}'\mathbf{C}_h\mathbf{X} = 0$  (ce qui est équivalent à les choisir

telles que  $\mathbf{C}_h \mathbf{X} = 0$ , dès que  $\mathbf{C}_h$  est strictement définie positive). On obtient ainsi :

$$\mathbb{E}(Y' \mathbf{C}_h Y) = \text{tr}(\mathbf{C}_h \mathbf{V}) = \text{tr}(\mathbf{C}_h \sum_{k=0}^K \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k') = \sum_{k=0}^K \sigma_k^2 \text{tr}(\mathbf{Z}_k' \mathbf{C}_h \mathbf{Z}_k).$$

La méthode d'ANOVA consiste ainsi à suivre les étapes suivantes :

- on choisit  $K + 1$  matrices  $\mathbf{C}_h$  ( $h = 0, 1, \dots, K$ ) vérifiant les propriétés requises et linéairement indépendantes ;
- on appelle  $\mathbf{T}$  la matrice carrée d'ordre  $K + 1$  de terme général :  $\mathbf{T}_h^k = \text{tr}(\mathbf{Z}_k' \mathbf{C}_h \mathbf{Z}_k)$  ;
- on pose :

$$\gamma^2 = \begin{pmatrix} \sigma^2 \\ \sigma_1^2 \\ \vdots \\ \sigma_K^2 \end{pmatrix} ; \quad Q(y) = \begin{pmatrix} Q_0(y) \\ Q_1(y) \\ \vdots \\ Q_K(y) \end{pmatrix}, \quad \text{avec } Q_h(y) = y' \mathbf{C}_h y ;$$

- on résout le système  $\mathbb{E}(Y' \mathbf{C}_h Y) = y' \mathbf{C}_h y$  ( $h = 0, 1, \dots, K$ ), soit encore :  $\mathbf{T} \gamma^2 = Q(y)$  ;
- ce qui conduit à la solution :  $\hat{\gamma}^2 = \mathbf{T}^{-1} Q(y)$ .

On voit que toute la méthode repose sur le choix approprié des matrices  $\mathbf{C}_h$ . Il est clair que différents choix au niveau de ces matrices conduiront à différentes estimations des composantes de la variance. Il est maintenant courant d'utiliser l'une des méthodes proposées par C.R. Henderson (1953) et améliorées par la suite.

- La méthode dite Henderson I ne s'applique qu'aux modèles à effets aléatoires, autrement dit tels que  $\mathbf{X} = 0_n$ , ce qui évite d'imposer la propriété  $\mathbf{X}' \mathbf{C}_h \mathbf{X} = 0$  aux matrices  $\mathbf{C}_h$ . Elle est calquée sur la méthode décrite en 6.2.2, en adaptant les sommes de carrés au cas déséquilibré.
- La méthode Henderson II est une adaptation de la précédente qui définit des formes quadratiques spécifiques permettant d'annuler les effets fixes, donc associées à des matrices  $\mathbf{C}_h$  vérifiant  $\mathbf{X}' \mathbf{C}_h \mathbf{X} = 0$ . Toutefois, cette méthode ne peut s'appliquer s'il y a dans le modèle des interactions entre effets fixes et effets aléatoires.
- La **méthode Henderson III** est la plus générale (elle s'applique dans tous les cas), mais aussi la plus compliquée. C'est la plus utilisée dans la pratique et c'est, en particulier, celle qui est mise en œuvre dans le logiciel SAS.

Toutes ces méthodes fournissent des estimateurs sans biais des composantes de la variance. On trouvera une présentation très détaillée des méthodes de Henderson dans Searle *et al.* (1992).

### Estimation de $\mathbf{V}$ par maximum de vraisemblance

La log-vraisemblance du modèle mixte gaussien s'écrit :

$$l(y, \beta, \mathbf{V}) = c - \frac{1}{2} \log[\det(\mathbf{V})] - \frac{1}{2} (y - \mathbf{X}\beta)' \mathbf{V}^{-1} (y - \mathbf{X}\beta)$$

(généralisation de ce qui a été fait en 6.1.4). On en déduit :

$$\frac{\partial l}{\partial \beta} = \mathbf{X}' \mathbf{V}^{-1} y - \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} \beta \quad (\text{système de } p \text{ équations dont découlent les équations normales}).$$

On remarque ensuite :  $\frac{\partial \mathbf{V}}{\partial \sigma_k^2} = \mathbf{Z}_k \mathbf{Z}_k'$  ; on en déduit :

$$\frac{\partial l}{\partial \sigma_k^2} = -\frac{1}{2} \text{tr}(\mathbf{V}^{-1} \mathbf{Z}_k \mathbf{Z}_k') + \frac{1}{2} (y - \mathbf{X}\beta)' \mathbf{V}^{-1} \mathbf{Z}_k \mathbf{Z}_k' \mathbf{V}^{-1} (y - \mathbf{X}\beta)$$

(une équation pour chaque  $\sigma_k^2$ ,  $k = 0, 1, \dots, K$ ). Les premières équations de vraisemblance fournissent :

$$\hat{\beta} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} y = \text{GLSE}(\beta).$$

Les suivantes s'écrivent :

$$\text{tr}(\mathbf{V}^{-1} \mathbf{Z}_k \mathbf{Z}_k') = (y - \mathbf{X}\beta)' \mathbf{V}^{-1} \mathbf{Z}_k \mathbf{Z}_k' \mathbf{V}^{-1} (y - \mathbf{X}\beta), \quad k = 0, 1, \dots, K.$$

On obtient ainsi un système de  $K + 1 + p$  équations non linéaires à  $K + 1 + p$  inconnues que l'on résoud par une méthode numérique itérative (de type Fisher scoring). Ces procédures numériques fournissent en plus, à la convergence, la matrice des variances-covariances asymptotiques des estimateurs.

Les estimateurs obtenus par maximum de vraisemblance sont, en général, biaisés : la méthode produit un biais systématique. Ils peuvent être négatifs et sont alors ramenés à 0.

**Remarque 71** *On a vu que les  $p$  premières équations ci-dessus fournissent l'estimation maximum de vraisemblance de  $\beta$ ,  $\hat{\beta} = GLSE(\beta)$ , dans tous les cas de figure. Pour retrouver l'expression donnée pour  $\hat{\mu}$  en 6.1.3 dans le cas déséquilibré, il faut remarquer que, dans le cas particulier où  $\mu$  est le seul effet fixe,  $\mathbf{X} = \mathbb{1}_n$ , de sorte que  $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$  est la somme de tous les termes de  $\mathbf{V}^{-1}$ . Comme on a, dans ce cas,  $\mathbf{V} = \sigma_a^2\mathbf{Z}\mathbf{Z}' + \sigma^2\mathbf{I}_n$  (matrice bloc-diagonale), elle s'inverse par bloc, chaque bloc étant carré d'ordre  $n_j$ , de la forme*

$$\begin{pmatrix} \sigma_a^2 + \sigma^2 & \sigma_a^2 & \cdots \\ \sigma_a^2 & \sigma_a^2 + \sigma^2 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix},$$

dont l'inverse s'écrit

$$\frac{1}{(n_j\sigma_a^2 + \sigma^2)\sigma^2} \begin{pmatrix} (n_j - 1)\sigma_a^2 + \sigma^2 & -\sigma_a^2 & \cdots \\ -\sigma_a^2 & (n_j - 1)\sigma_a^2 + \sigma^2 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}.$$

Ainsi, la somme de toute ligne (ou de toute colonne) de  $\mathbf{V}^{-1}$  vaut  $\frac{1}{(n_j\sigma_a^2 + \sigma^2)}$  et le total de tous les termes de cette matrice vaut

$$\sum_{j=1}^J \frac{n_j}{n_j\sigma_a^2 + \sigma^2} = \sum_{j=1}^J \frac{1}{\tau_j^2}.$$

D'autre part,  $\mathbb{1}_n\mathbf{V}^{-1}\mathbf{y}$  vaut  $\sum_{j=1}^J \frac{\bar{y}_{\bullet j}}{\tau_j^2}$ , ce qui permet d'écrire :

$$\mu = \sum_{j=1}^J w_j \bar{y}_{\bullet j} \text{ et } \hat{\mu} = \sum_{j=1}^J \hat{w}_j \bar{y}_{\bullet j}.$$

### Estimation de $\mathbf{V}$ par maximum de vraisemblance restreint

On a pu constater, dans le point précédent, que les différentes équations obtenues lorsqu'on réalise l'estimation des paramètres par maximum de vraisemblance contiennent à la fois le vecteur  $\beta$  (des paramètres liés à l'espérance de  $Y$ ) et la matrice  $\mathbf{V}$  (des paramètres liés à la variance de  $Y$ ). Dans un modèle mixte, c'est ce mélange de paramètres de natures différentes dans les mêmes équations qui engendre un biais systématique dans l'estimation par maximum de vraisemblance des composantes de la variance. L'objet de la méthode du maximum de vraisemblance restreint est précisément de séparer les deux types de paramètres.

L'idée est la suivante : aux colonnes de la matrice  $\mathbf{X}$  sont associés des vecteurs de l'espace vectoriel  $F = \mathbb{R}^n$  ( $n$  est le nombre d'observations réalisées) ; on munit ce dernier de la métrique identité (associée à la matrice  $\mathbf{I}_n$  sur la base canonique), ce qui en fait un espace euclidien. En notant  $F_X$  le sous-espace vectoriel de  $F$  engendré par les colonnes de  $X$  ( $F_X$  est supposé de dimension  $p$ ), on sait que le projecteur orthogonal de  $F$  dans  $F_X$  a pour matrice associée  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Par ailleurs, le projecteur sur le s.e.v.  $F_X^\perp$ , supplémentaire orthogonal à  $F_X$  dans  $F$ , est  $\mathbf{H}^\perp = \mathbf{I}_n - \mathbf{H}$ . Ainsi, en projetant  $Y$  sur  $F_X^\perp$  et en travaillant avec cette projection (qui est, par définition, orthogonale à toute combinaison linéaire des colonnes de  $\mathbf{X}$ ), on s'affranchit de  $\beta$  dans l'estimation des composantes de la variance. Toutefois, le s.e.v.  $F_X^\perp$  étant de dimension  $m = n - p$ , le vecteur aléatoire projeté de  $Y$  sur  $F_X^\perp$  est multinormal d'ordre  $m$ . Écrit dans  $\mathbb{R}^n$ , c'est un vecteur aléatoire dégénéré (sa matrice des variances-covariances est singulière) qu'on doit transformer pour obtenir un vecteur aléatoire directement écrit dans  $\mathbb{R}^m$ . Soit donc  $\mathbf{M}_0$  une matrice  $m \times n$ , de rang  $m$ ,

réalisant cette transformation et soit  $\mathbf{M} = \mathbf{M}_0 \mathbf{H}^\perp$ . On considère finalement  $Y^* = \mathbf{M}Y$  ; on a ainsi  $Y^* \sim \mathcal{N}_m(\mu, \mathbf{V}^*)$ , avec :

$$\mu = \mathbf{M}\mathbf{X}\beta = \mathbf{M}_0(\mathbf{H}^\perp \mathbf{X}\beta) = 0 \text{ (par définition de } \mathbf{H}^\perp \text{)} ; \quad \mathbf{V}^* = \mathbf{M}\mathbf{V}\mathbf{M}' = \mathbf{M}_0 \mathbf{H}^\perp \mathbf{V} \mathbf{H}^\perp \mathbf{M}_0'.$$

En fait, on peut encore écrire :

$$\mathbf{V}^* = \sum_{k=0}^K \sigma_k^2 \mathbf{M}\mathbf{Z}_k \mathbf{Z}_k' \mathbf{M}' = \sum_{k=0}^K \sigma_k^2 \mathbf{Z}_k^* \mathbf{Z}_k^{*'} \quad (\text{en posant } \mathbf{Z}_k^* = \mathbf{M}\mathbf{Z}_k).$$

En réécrivant les  $K + 1$  dernières équations de vraisemblance relatives au vecteur aléatoire  $Y^*$ , il vient maintenant :

$$\text{tr}(\mathbf{V}^{*-1} \mathbf{Z}_k^* \mathbf{Z}_k^{*'}) = y^{*'} \mathbf{V}^{*-1} \mathbf{Z}_k^* \mathbf{Z}_k^{*'} \mathbf{V}^{*-1} y^*, \quad k = 0, 1, \dots, K \text{ (avec : } y^* = \mathbf{M}y \text{)}.$$

Il s'agit d'un système de  $K + 1$  équations non linéaires à  $K + 1$  inconnues (les composantes  $\sigma_k^2$  de la variance ;  $k = 0, 1, \dots, K$ ) dans lequel ne figure plus le vecteur  $\beta$ . En général, il n'admet pas de solution analytique et nécessite une procédure numérique itérative pour sa résolution. Les estimateurs  $\hat{\Sigma}_k^2$  ainsi obtenus ne sont pas nécessairement sans biais, mais ne comportent pas de biais systématique. Là encore, la procédure itérative fournit, à la convergence, la matrice des variances-covariances asymptotiques des estimateurs.

**Remarque 72** *Les équations écrites ci-dessus proviennent de l'annulation des dérivées partielles, selon les  $\sigma_k^2$ , de la log-vraisemblance du vecteur aléatoire  $Y^*$ , autrement dit de la log-vraisemblance de la projection de  $Y$  sur le s.e.v.  $F_{\bar{X}}^\perp$ , ou encore de la restriction de la vraisemblance de  $Y$  à ce sous-espace. Les estimateurs obtenus par maximisation de cette restriction de la vraisemblance sont, pour cette raison, appelés estimateurs du maximum de vraisemblance restreint (on devrait dire, de façon plus rigoureuse, du maximum de la vraisemblance restreinte).*

### Estimation de $\mathbf{V}$ par MINQUE et MIVQUE

Dans un modèle mixte, il est encore possible d'estimer les composantes de la variance en utilisant soit la méthode MINQUE, soit la méthode MIVQUE. Le principe général reste le même que celui exposé en 6.1.4. En fait, ces méthodes sont peu utilisées dans la pratique. Il faut néanmoins rappeler que SAS utilise la méthode dite MIVQUE(0) pour initialiser la procédure itérative utilisée avec la méthode REML (procédures VARCOMP et MIXED).

#### 6.3.4 Intervalles de confiance

Dans le cas d'un plan équilibré, on peut construire un intervalle de confiance exact, de type Student, pour toute fonction linéaire  $c'\beta$  du paramètre  $\beta$  relatif aux effets fixes ( $c \in \mathbb{R}^p$ ). Le principe est le même que celui indiqué au paragraphe 6.1.5. Dans le cas d'un plan déséquilibré, cela n'est plus possible.

Pour la variance des erreurs  $\sigma^2$ , un intervalle de confiance exact, de type khi-deux, peut être construit dans tous les cas, que ce soit pour un plan équilibré ou pour un plan déséquilibré.

Mais, pour les autres paramètres de variance, on ne dispose pas d'intervalle de confiance précis (pas même asymptotique).

#### 6.3.5 Tests de significativité des facteurs

Ces tests sont standards dans le cas équilibré, mais deviennent assez problématiques dans le cas déséquilibré.

##### Cas équilibré

Pour tester la significativité d'un facteur à effets fixes dans un modèle mixte, on utilise le test habituel de Fisher, tel qu'il a été défini en ANOVA : il reste valable dans ce cas (rappelons qu'il s'agit d'un test exact).

Pour tester la significativité d'un facteur à effets aléatoires, c'est-à-dire la nullité d'une composante de la variance, on utilise encore un test de Fisher analogue à ceux définis en 6.2.3.

Tous ces tests sont mis en œuvre par la procédure GLM de SAS, mais seuls ceux relatifs aux effets fixes le sont par la procédure MIXED.

### Cas déséquilibré

Il n'y a malheureusement pas de test exact, ni même de test asymptotique, qui permette de tester les effets, que ce soient les effets fixes ou les effets aléatoires, dans un modèle mixte avec un plan déséquilibré. Il existe seulement des tests approchés (dont on ne contrôle pas réellement le niveau, et encore moins la puissance). Nous donnons néanmoins ci-dessous quelques pistes qui permettront d'aider à choisir le modèle le plus approprié relativement à un jeu de données.

- *Le test de Fisher, avec degré de liberté calculé selon l'approximation de Satterthwaite.* Nous présentons ce test dans un cadre simple, le principe restant le même dans tout modèle mixte. On se place donc dans le cas d'un modèle à deux facteurs croisés à effets aléatoires. Notons  $A$  et  $B$  les deux facteurs considérés,  $J$  et  $K$  leurs nombres de niveaux,  $C$  leurs interactions et  $E$  les erreurs du modèle considéré. Enfin, on note encore  $MS$  les carrés moyens associés à chacun de ces effets. Il est possible d'écrire :

$$\begin{aligned}\mathbb{E}(MSA) &= \sigma^2 + \alpha_1\sigma_c^2 + \alpha_2\sigma_a^2 ; \\ \mathbb{E}(MSB) &= \sigma^2 + \alpha_3\sigma_c^2 + \alpha_4\sigma_b^2 ; \\ \mathbb{E}(MSC) &= \sigma^2 + \alpha_5\sigma_c^2 ; \\ \mathbb{E}(MSE) &= \sigma^2.\end{aligned}$$

On ne dispose pas d'expression explicite pour les coefficients  $\alpha_i$ , mais on sait les déterminer numériquement, en général en utilisant la méthode dite de Henderson III (c'est ce qui est fait dans le logiciel SAS).

Si l'on souhaite tester, par exemple, l'hypothèse nulle  $\{H_0 : \sigma_a^2 = 0\}$ , on peut réécrire, sous  $H_0$  :

$$\begin{aligned}\mathbb{E}(MSA) &= \sigma^2 + \alpha_1\sigma_c^2 \\ &= \sigma^2 + \alpha_1 \left[ \frac{\mathbb{E}(MSC) - \sigma^2}{\alpha_5} \right] \\ &= \sigma^2 \left[ 1 - \frac{\alpha_1}{\alpha_5} \right] + \frac{\alpha_1}{\alpha_5} \mathbb{E}(MSC) \\ &= \frac{\alpha_1}{\alpha_5} \mathbb{E}(MSC) + \left[ 1 - \frac{\alpha_1}{\alpha_5} \right] \mathbb{E}(MSE).\end{aligned}$$

En fait, on remplace les espérances (inconnues) des carrés moyens intervenant dans le terme de droite de cette expression par les valeurs empiriques correspondantes (par exemple,  $MSC(y)$  remplace  $\mathbb{E}(MSC)$ ). On pose alors :

$$MSA0 = \frac{\alpha_1}{\alpha_5} MSC(y) + \left[ 1 - \frac{\alpha_1}{\alpha_5} \right] MSE(y).$$

Il a été montré par Satterthwaite (1946) que toute quantité du type  $MSA0$  est approximativement distribuée selon une loi de khi-deux dont le degré de liberté  $q$  peut être calculé. En fait, si  $d_1$  est le degré de liberté de  $MSC$  et  $d_2$  celui de l'erreur du modèle considéré (donc de  $MSE$ ), la formule donnant  $q$  est la suivante :

$$q = \frac{\left[ \frac{\alpha_1}{\alpha_5} MSC(y) + \left[ 1 - \frac{\alpha_1}{\alpha_5} \right] MSE(y) \right]^2}{\frac{\left[ \frac{\alpha_1}{\alpha_5} MSC(y) \right]^2}{d_1} + \frac{\left[ 1 - \frac{\alpha_1}{\alpha_5} \right]^2 MSE(y)^2}{d_2}}.$$

Pour tester  $H_0$ , on fait donc le rapport des deux expressions  $MSA(y)$  et  $MSA0$  et on le compare à une loi de Fisher à  $J-1$  et  $q$  degrés de liberté. Si l'on ne peut pas affirmer que ces deux carrés moyens sont indépendants, le fait qu'ils soient calculés avec des sommes de carrés distinctes permet de penser que le résultat obtenu est une approximation correcte d'une loi de Fisher. En fait, des simulations (ainsi que l'expérience) ont montré que ce test approximatif de

Fisher fonctionne plutôt bien. C'est lui qui est mis en œuvre dans la procédure GLM du logiciel SAS, mais pas dans la procédure MIXED. Nous recommandons d'utiliser prioritairement ce test dans les applications.

- *Le test de Wald.* Il permet de tester la significativité des composantes de la variance dans les modèles mixtes. En particulier, c'est le test que l'on trouve dans la procédure MIXED de SAS. Clairement, nous déconseillons ce test. Non seulement il s'agit d'un test asymptotique (qui nécessite donc certaines précautions d'utilisation) mais, surtout, la loi de khi-deux obtenue comme loi limite sous  $H_0$  nécessite certaines conditions techniques qui ne sont manifestement pas vérifiées ici (l'hypothèse nulle d'un tel test est nécessairement du type  $\{H_0 : \sigma_a^2 = 0\}$ , autrement dit la valeur testée est située sur la frontière de l'espace paramétrique  $\mathbb{R}_+$ , ce qui empêche d'établir le résultat asymptotique qui est donc faux).
- *Une autre solution,* peu courante, consiste à calculer un effectif moyen  $n^*$ , en faisant la moyenne harmonique de l'ensemble des effectifs des différentes cellules (les cellules définies par le plan d'expériences considéré). On remplace ensuite chaque effectif par  $n^*$  et on opère comme en présence d'un plan équilibré avec  $n^*$  répétitions. Cette méthode est celle préconisée dans Miller (1997) lorsque les effectifs sont assez proches les uns des autres ; elle nous semble moins intéressante que la première méthode indiquée ci-dessus.
- Certains praticiens regardent ce que donnent les tests standards de Fisher en considérant tous les effets du modèle comme des effets fixes. Cela peut permettre de se faire une idée de l'importance des différents effets, mais doit seulement être considéré comme un complément aux tests de Fisher abordés plus haut.
- Une comparaison des estimations des différentes composantes de la variance dans le modèle complet (dans lequel on a pris en compte tous les effets considérés au départ) permet aussi de préciser l'importance relative des différents effets. En particulier, la comparaison de chacune de ces composantes avec l'estimation de la variance de l'erreur du modèle est un élément intéressant à prendre en compte dans le choix d'un modèle.
- Lorsque le choix entre plusieurs modèles n'est pas clair (autrement dit, lorsqu'on hésite à prendre en compte certains effets dans le modèle retenu), on peut aussi regarder les critères de choix de modèle tels que *AIC* ou *BIC* (en faisant attention à leur définition dans la procédure MIXED de SAS ; se reporter à l'Annexe D). En particulier, si l'un des modèles envisagés minimise chacun de ces deux critères, c'est le modèle qu'il faudra retenir.
- Enfin, signalons l'existence de tests exacts pour tester certaines hypothèses relatives aux effets d'un modèle mixte avec plan déséquilibré. Ces tests sont assez complexes, difficiles à mettre en œuvre, et leur principe dépasse le cadre de ce cours. De plus, ils ne figurent dans aucun des logiciels statistiques courants. On en trouvera une présentation très détaillée dans l'ouvrage de Khuri *et al.* (1998) auquel nous renvoyons le lecteur intéressé.

### 6.3.6 Prévisions dans les modèles mixtes

Sur la base du principe exposé dans le point 6.1.7, la prévision d'une v.a.r.  $Y$  non observée, au moyen d'un modèle mixte, fait appel à la notion d'espérance conditionnelle, afin d'obtenir un résultat optimal en un certain sens. C'est ce qu'on appelle le BLUP (*Best Linear Unbiased Predictor*), que nous ne développerons pas ici, mais que l'on trouvera détaillé dans Searle *et al.* (1992). D'autre part, il est possible d'obtenir les prévisions de type BLUP avec la procédure MIXED de SAS.

### 6.3.7 Illustration

Il s'agit d'un exemple fictif à deux facteurs : le premier, f1, supposé à effets fixes, comporte 3 niveaux (notés 1, 2, 3) ; le second, f2, supposé à effets aléatoires, en comporte 4 (notés 1, 2, 3, 4). La réponse,  $y$ , est constituée d'entiers compris entre 1 et 30. Le plan est complet, déséquilibré. Il y a 35 individus observés, les données étant reproduites ci-après.

**Les données**

```

1 1 10
1 1 12
1 1 14
1 2 17
1 2 18
1 3 15
1 3 14
1 3 14
1 4 12
1 4 10
1 4 13
1 4 11
2 1 7
2 1 8
2 2 13
2 2 11
2 2 12
2 3 9
2 3 8
2 3 10
2 3 9
2 4 7
2 4 6
3 1 19
3 1 17
3 1 14
3 2 23
3 2 25
3 2 24
3 2 22
3 3 19
3 3 20
3 4 16
3 4 15
3 4 15

```

**Le programme SAS**

Le programme SAS ci-dessous permet de traiter ces données en tenant compte du caractère aléatoire du second facteur, sans considérer d'interactions. Outre les procédures GLM et VARCOMP, on a utilisé ici la procédure MIXED.

```

* options facultatives pour la mise en page des sorties ;
* ----- ;
options linesize=76 pagesize=64 nodate;
title;
footnote 'Modele mixte - Exemple fictif';
* ----- ;
*          lecture des donnees                ;
*    (le fichier "fic.don" contient les donnees ;
*    et se trouve dans le repertoire de travail) ;
* ----- ;
data fic;
infile 'fic.don';
input f1 f2 y;
run;
* ----- ;
*          procedure GLM :                    ;
*          f1 en effets fixes et              ;
*          f2 en effets aleatoires            ;
* ----- ;

```





## REML Estimates

Variance Component	Estimate
Var(f2)	8.74710
Var(Error)	1.57409

## Asymptotic Covariance Matrix of Estimates

	Var(f2)	Var(Error)
Var(f2)	53.16422	-0.01969
Var(Error)	-0.01969	0.17087

## Les sorties de la procédure MIXED

## The Mixed Procedure

## Model Information

Data Set	WORK.FIC
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

## Class Level Information

Class	Levels	Values
f1	3	1 2 3
f2	4	1 2 3 4

## Dimensions

Covariance Parameters	2
Columns in X	4
Columns in Z	4
Subjects	1
Max Obs Per Subject	35
Observations Used	35
Observations Not Used	0
Total Observations	35

## Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	167.18609531	
1	2	124.34322114	0.00000121
2	1	124.34318122	0.00000000

Convergence criteria met.

Covariance Parameter  
Estimates

Cov Parm	Estimate
f2	8.7466
Residual	1.5741

## Fit Statistics

-2 Res Log Likelihood	124.3
AIC (smaller is better)	128.3
AICC (smaller is better)	128.8
BIC (smaller is better)	127.1

## Type 3 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
f1	2	29	181.42	<.0001

## Chapitre 7

# Modèles pour données répétées

*Les données répétées, ou données longitudinales, constituent un domaine à la fois important et assez particulier de la statistique. On entend par données répétées des données telles que, pour chaque individu considéré, on dispose d'observations à différents instants, autrement dit répétées dans le temps. Les principaux domaines d'application de ce type de données sont la médecine et la biologie, lors d'expérimentations humaines ou animales. La difficulté majeure dans le traitement statistique de ces données provient de ce qu'il n'est en général pas réaliste de supposer que les observations réalisées sur un même individu, au cours du temps, sont indépendantes. Il est donc nécessaire d'introduire une structure de covariance pour les variables aléatoires associées à chaque individu, afin de tenir compte de cette situation particulière. On va ainsi retrouver des modèles proches de ceux vus au chapitre 5 avec l'analyse de variance multidimensionnelle. Par ailleurs, il est fréquent, dans les modèles pour données répétées, de considérer, en plus des facteurs à effets fixes que l'on souhaite étudier dans le modèle, des effets aléatoires associés aux individus. On aura pour cela recours à des modèles mixtes, tels que nous les avons introduits au chapitre 6. Enfin, on notera que, dans le cadre de la modélisation des données répétées, le terme statistique d'individu est souvent remplacé par celui de **sujet**. Nous utiliserons indifféremment l'un ou l'autre par la suite.*

*Pour la bibliographie, nous conseillons les ouvrages de Brown & Prescott (1999), de Davis (2002) et de Verbeke & Molenberghs (2000).*

### Résumé

Ce chapitre est donc consacré à la modélisation de ce que l'on appelle les *données répétées*, ou encore les *données longitudinales*. Sauf cas particulier, il s'agit de données répétées au cours du temps, sur les mêmes individus.

De manière générale, on peut penser que les observations faites à différents instants, sur un individu donné, sont corrélées. D'où la nécessité d'introduire une "structure de covariance" pour ces observations dans les modèles pour données répétées (autrement dit, de considérer une matrice de variances-covariances  $\mathbf{R}$ , de dimension  $T \times T$ , si  $T$  est le nombre d'instantants d'observation). En ce sens, ces modèles ressemblent aux modèles de MANOVA vus au chapitre 5, puisque la réponse de chaque individu est multidimensionnelle. Mais, au lieu de correspondre à différentes composantes observées à un instant donné, cette réponse correspond à la même variable observée à différents instants. De plus, on va maintenant considérer le cadre général des modèles linéaires gaussiens mixtes, introduits au chapitre 6, pour pouvoir expliquer l'évolution au cours du temps de la variable réponse à la fois par des facteurs à effets fixes et par des facteurs à effets aléatoires, notamment des facteurs individuels. De façon naturelle, le temps lui-même interviendra comme facteur à effets fixes.

Les modèles pour données répétées sont donc assez complexes et d'un maniement plutôt délicat. Dans leur forme la plus générale (qui ne sera abordée qu'au paragraphe 7.6), ils nécessitent en effet l'estimation des effets fixes, celle des composantes de la variance et celle des éléments de la matrice  $\mathbf{R}$ . Néanmoins, certaines structures particulières pour la matrice  $\mathbf{R}$  pourront être envisagées,

réduisant ainsi le nombre de paramètres à estimer. Parallèlement, il faudra tester la significativité des différents effets initialement considérés, dans le but de simplifier au maximum le modèle finalement retenu. Enfin, si plusieurs modèles sont envisageables, il faudra procéder à un choix de modèle en utilisant des indicateurs du type AIC ou BIC.

On notera, pour terminer cette présentation générale, que la procédure de SAS la plus usuelle pour traiter les modèles pour données répétées est la procédure MIXED, mais que la procédure GLM peut aussi être utilisée de manière complémentaire.



## 7.1 Introduction

Nous noterons toujours  $Y$  la variable aléatoire réelle réponse que l'on souhaite modéliser. Elle est observée sur différents individus et à différents instants. L'objet de ce chapitre est de définir des modèles prenant en compte d'une part un ou plusieurs facteurs susceptibles d'avoir un effet sur les valeurs de  $Y$ , d'autre part le temps. Jusqu'au paragraphe 7.6, nous ne considérerons qu'un seul facteur à effets fixes autre que le temps (la généralisation à plusieurs facteurs de même nature ne pose pas de problème particulier). Ce facteur sera noté  $F$  (il pourra s'agir, par exemple, d'un traitement médical), le nombre de ses niveaux sera noté  $J$ , ces derniers étant indicés par  $j$ .

Au chapitre 3, pour une analyse de variance (ANOVA) à un seul facteur  $F$ , nous avons écrit le modèle sous la forme suivante :

$$Y_{ij} = \mu + \alpha_j + U_{ij} ; \quad j = 1, \dots, J ; \quad i = 1, \dots, n_j .$$

Maintenant, chaque mesure  $Y_{ij}$  est répétée à différents instants notés  $t$  ( $t = 1, \dots, T$ ) et les v.a.r. correspondantes sont notées  $Y_{ijt}$ , de sorte que le modèle est réécrit sous la forme suivante :

$$Y_{ijt} = \mu + \alpha_j^1 + \alpha_t^2 + \gamma_{jt} + U_{ijt} .$$

On pourrait envisager de traiter ce modèle comme un modèle d'ANOVA à deux facteurs croisés ( $F$  et le temps), mais cela reviendrait à supposer que les observations réalisées sur un même individu au cours du temps sont indépendantes, ce qui n'est pas réaliste (sauf cas très particulier). Il est donc nécessaire d'introduire ici des modèles spécifiques pour ce type de données, ces modèles devant prendre en compte une structure de covariance entre observations réalisées sur un même individu à différents instants.

**Remarque 73** *Dans tout ce chapitre, nous supposons d'une part que les instants d'observation sont les mêmes pour tous les individus, d'autre part que tous les individus sont observés à chaque instant considéré (autrement dit, il n'y a pas de données manquantes). Malheureusement, cette situation, relativement commode en ce qui concerne la modélisation, n'est pas la plus courante dans la pratique, ce qui complique d'autant plus les choses.*

## 7.2 Analyses préliminaires

Deux types d'analyses préliminaires sont possibles avec les données répétées. Elles n'abordent pas réellement le problème de la modélisation de ces dernières, mais peuvent apporter des compléments intéressants à cette modélisation. Notons que ces deux types d'analyses sont réalisées de façon systématique avec la procédure GLM de SAS.

### 7.2.1 ANOVA réalisée à chaque instant $t$

Dans une première approche, on peut réaliser l'ANOVA de  $Y$  en fonction de  $F$  à chaque instant d'observation  $t$ . Cela donne une première idée de l'influence de  $F$  sur  $Y$ , mais ne modélise pas l'évolution de  $Y$  au cours du temps. C'est donc largement insuffisant.

On notera que ces analyses sont les premières sorties fournies par la procédure GLM de SAS avec des données répétées, ce qui peut être trompeur (en laissant supposer que cela relève de la modélisation du temps, ce qui n'est pas le cas).

Toutefois, de même qu'il est conseillé de faire une étude univariée de chaque variable quantitative considérée dans une Analyse en Composantes Principales avant de réaliser cette dernière, de même les  $T$  ANOVA dont il est ici question sont utiles pour bien maîtriser les données considérées.

### 7.2.2 ANOVA réalisée sur la moyenne temporelle des observations

Il s'agit de l'ANOVA de  $Y$  en fonction de  $F$  réalisée sur les moyennes temporelles

$$\bar{y}_{ij\bullet} = \frac{1}{T} \sum_{t=1}^T y_{ijt} .$$

Bien entendu, cette analyse ne prend nullement en compte le temps, puisqu'on fait au préalable la moyenne des différentes valeurs obtenues aux différents instants d'observation. En fait, on teste ainsi l'influence marginale (indépendamment du temps) du facteur  $F$ . Comme précédemment, cela apporte un complément intéressant à la modélisation des données répétées, même si cela n'en fait pas partie.

On notera que, bizarrement, dans la procédure GLM de SAS, cette analyse marginale est faite en utilisant  $\sqrt{T}$  au lieu de  $T$  au dénominateur de l'expression ci-dessus. Toutefois, cela ne change strictement rien au résultat du test de significativité marginale du facteur  $F$  (test de Fisher usuel).

## 7.3 Modèle à un facteur à effets fixes pour données répétées

Contrairement aux analyses décrites en 7.2, le modèle présenté ici est spécifique aux données répétées. Il convient lorsqu'on ne souhaite pas prendre en compte des facteurs à effets aléatoires pour modéliser les données. La généralisation à plus d'un facteur à effets fixes autre que le temps ne pose pas de problème particulier.

### 7.3.1 Principe

Le principe général de ce modèle est le suivant : pour un niveau  $j$  du facteur  $F$  ( $j = 1, \dots, J$ ), et pour un individu  $i$  pris en compte à ce niveau là ( $i = 1, \dots, n_j$ ), on considère le vecteur aléatoire  $Y_{ij} = (Y_{ij1}, \dots, Y_{ijT})'$  de  $\mathbb{R}^T$  et l'on pose :

$$Y_{ij} \sim \mathcal{N}_T(\mu_j, \mathbf{R}) .$$

On notera que, contrairement à ce qui a été fait au chapitre 5, le vecteur aléatoire  $Y_{ij}$  est ici considéré comme un vecteur colonne. Dans l'écriture ci-dessus :

- le vecteur  $\mu_j$  désigne le vecteur moyenne de la loi gaussienne au niveau  $j$  de  $F$  ; les composantes de ce vecteur sont liées au temps :

$$\mu_j = \begin{pmatrix} \mu_{j1} \\ \vdots \\ \mu_{jT} \end{pmatrix} ;$$

par ailleurs, on pose :  $\mu_{jt} = \mu + \alpha_j^1 + \alpha_t^2 + \gamma_{jt}$  ( $\mu$  est l'effet général,  $\alpha_j^1$  est l'effet principal du niveau  $j$  du facteur  $F$ ,  $\alpha_t^2$  est l'effet principal de l'instant  $t$  et  $\gamma_{jt}$  est l'effet d'interaction entre le niveau  $j$  et l'instant  $t$  ; les paramètres  $\alpha_j^1$  et  $\alpha_t^2$  sont centrés et les paramètres  $\gamma_{jt}$  sont doublement centrés) ;

- la matrice  $\mathbf{R}$  est  $T \times T$ , symétrique et strictement définie positive ; elle représente la structure de covariance des observations répétées sur un même individu et doit être estimée ; nous verrons, au prochain paragraphe, les principales structures de covariance usuelles, la plupart permettant de diminuer sensiblement le nombre de paramètres à estimer. On notera ici que la matrice  $\mathbf{R}$  ne dépend pas de  $j$ , autrement dit que le modèle considéré est encore homoscedastique ; il est certes possible de considérer des modèles hétéroscedastiques

en présence de données répétées, mais cela entraîne encore une augmentation du nombre de paramètres à estimer et peut devenir très problématique.

Le modèle considéré s'écrit donc :

$$Y_{ijt} = \mu_{jt} + U_{ijt} ,$$

le paramètre de moyenne  $\mu_{jt}$  s'écrivant comme indiqué plus haut et la v.a.r. erreur  $U_{ijt}$  étant la  $t$ -ième composante du vecteur aléatoire gaussien centré  $U_{ij}$  de  $\mathbb{R}^T$ , de matrice de variances-covariances  $\mathbf{R}$  : les composantes de ce vecteur aléatoire ne sont donc pas indépendantes (sauf si  $\mathbf{R}$  est diagonale, ce qui n'est en général pas le cas).

Dans la pratique, on commence par tester les effets d'interactions temps  $\times$  facteur puis, s'ils ne sont pas significatifs, on teste successivement l'effet du temps et l'effet du facteur (ce dernier point a été précisé en 7.2.2). Des indications sur les tests liés au temps sont données dans le point suivant. Dans le modèle retenu, on estime les différents paramètres, dont les éléments de la matrice  $\mathbf{R}$ .

### 7.3.2 Terminologie

Nous précisons ici quelques points de terminologie utiles dans l'utilisation des logiciels statistiques tels que SAS. On notera que les individus (ou unités statistiques) sont souvent appelés "sujets" dans le contexte des données répétées.

- Les effets principaux du facteur  $F$  (les  $\alpha_j^1$ ), non liés au temps, sont appelés les effets entre les sujets, ou inter-sujets (*between subjects effects*). On a vu en 7.2.2 comment les tester.
- Les effets liés au temps, à savoir les effets principaux du temps (les  $\alpha_t^2$ ) et les effets d'interactions temps  $\times$  facteur (les  $\gamma_{jt}$ ), sont appelés les effets intra-sujets (*within subjects effects*). On les teste avec les tests multidimensionnels introduits en MANOVA (voir le chapitre 5), en privilégiant toujours le test de Wilks. Ces tests sont définis en dimension  $D = T - 1$  et portent sur des différences du type :  $Y_{ij2} - Y_{ij1}, Y_{ij3} - Y_{ij1}, \dots, Y_{ijT} - Y_{ij1}$ .
- Finalement, dans le modèle multidimensionnel complet, on distingue :
  - 1 effet général,  $\mu$  ;
  - $J$  effets principaux du facteur  $F$ , les  $\alpha_j^1$  (avec une contrainte de centrage) ;
  - $T$  effets principaux du temps, les  $\alpha_t^2$  (également avec une contrainte de centrage) ;
  - $J \times T$  effets d'interactions, les  $\gamma_{jt}$  (avec  $J + T - 1$  contraintes pour le double centrage).

Les deux premiers types d'effets sont des effets inter-sujets (indépendants du temps), tandis que les deux derniers types sont des effets intra-sujets (ils dépendent du temps).

En comptant les éléments de la matrice  $\mathbf{R}$ , cela fait au total  $(J \times T) + \frac{T(T+1)}{2}$  paramètres indépendants à estimer dans le modèle complet, avec  $n \times T$  observations ( $n = \sum_{j=1}^J n_j$ ). On veillera donc à disposer d'un nombre total de sujets  $n$  vérifiant :  $n > J + \frac{T+1}{2}$ .

### 7.3.3 Mise en œuvre

Le logiciel statistique SAS dispose de deux procédures permettant de mettre en œuvre les modèles pour données répétées : la procédure **GLM** et la procédure **MIXED**. Nous renvoyons aux paragraphes 7.6 et 7.7 pour des précisions et des illustrations.

On notera que la procédure **MIXED**, plus récente, est mieux adaptée à ces modèles. Toutefois, la procédure **GLM** présente plusieurs avantages qui rendent son utilisation intéressante, en complément de **MIXED** : tout d'abord, elle réalise les différents tests multidimensionnels relatifs aux effets liés au temps<sup>1</sup>, ce que ne fait pas **MIXED** ; ensuite, elle met en œuvre le test de sphéricité de Mauchly (voir le point 7.5.3), indiquant si la structure *compound symmetry* est acceptable ou non.

<sup>1</sup>Pour des précisions sur ces tests, on se reportera à l'Annexe E.

## 7.4 Les structures usuelles de covariance pour $\mathbf{R}$

La procédure MIXED de SAS propose différentes structures classiques pour la matrice  $\mathbf{R}$  des covariances entre v.a.r. associées aux observations répétées sur un même individu. La structure choisie se déclare dans la commande `repeated` de cette procédure, au sein de l'option `type=...`. Par défaut, MIXED utilise l'option `simple`, correspondant à la structure d'indépendance (voir plus loin).

Nous précisons ci-dessous les principales options possibles pour la structure de  $\mathbf{R}$ . Dans les cas nécessitant une illustration, nous la donnons en utilisant le cas particulier  $T = 4$  (on notera qu'on ne parle de données répétées que pour  $T \geq 2$  et que le choix de structure n'a vraiment d'intérêt que pour  $T \geq 3$ ). La liste donnée ci-dessous n'est pas exhaustive et l'on pourra trouver d'autres possibilités pour  $\mathbf{R}$  en consultant l'aide en ligne de SAS.

Pour un modèle donné, le choix de la structure de covariance peut se faire en utilisant un critère usuel de choix de modèle (AIC ou BIC, voir l'Annexe D).

### Absence de structure

C'est le cas le plus général : la matrice  $\mathbf{R}$  n'a aucune structure particulière dans ce cas (elle est simplement symétrique et définie positive) et tous ses éléments doivent être estimés, soit  $\frac{T(T+1)}{2}$ . C'est, bien entendu, le cas le plus coûteux en nombre de paramètres à estimer. L'option correspondant à cette structure est appelée `unstructured` dans MIXED et se déclare avec l'option `type=un`.

### Structure d'indépendance

À l'opposé de la précédente, cette structure est la plus simple qui soit. Elle consiste à poser  $\mathbf{R} = \sigma^2 \mathbf{I}_T$ , où  $\sigma^2$  est le seul paramètre (supposé strictement positif) à estimer et où  $\mathbf{I}_T$  est la matrice identité d'ordre  $T$ . Cette structure suppose donc que les observations réalisées sur tout individu aux différents instants sont indépendantes, ce qui n'est, en général, pas réaliste et n'a donc pas de réel intérêt dans la pratique (on notera que cela revient à faire une ANOVA dans laquelle le temps est un facteur ordinaire). Sauf cas très particulier, cette structure n'est donc pas utilisée dans les modèles pour données répétées. L'option correspondante s'appelle `simple` dans la procédure MIXED de SAS et se déclare avec l'option `type=simple`.

**Attention : c'est l'option par défaut.**

### Structure symétrie composée, ou “compound symmetry”

Même si elle n'est souvent pas très réaliste, on rencontre fréquemment cette structure car elle présente différentes propriétés intéressantes. Elle consiste à poser  $\mathbf{R} = \sigma_1^2 \mathbf{I}_T + \sigma_2^2 \mathbb{I}_{T \times T}$ , où  $\sigma_1^2$  et  $\sigma_2^2$  sont deux paramètres strictement positifs à estimer,  $\mathbb{I}_{T \times T}$  désignant la matrice carrée d'ordre  $T$  dont tous les éléments valent 1. Ainsi, la variance de la v.a.r. associée à toute observation vaut  $\sigma_1^2 + \sigma_2^2$ , la covariance des v.a.r. associées à deux observations réalisées sur le même individu à deux instants différents valant constamment  $\sigma_2^2$ . L'option correspondante s'appelle `compound symmetry` (symétrie composée) et se déclare avec l'option `type=cs`.

Outre la simplicité d'une telle structure de covariance, nous en verrons, au paragraphe 7.5, d'autres propriétés.

### Structure auto-régressive d'ordre 1

L'intérêt de cette structure est de ne nécessiter que l'estimation de deux paramètres (comme la précédente), tout en étant telle que la corrélation des v.a.r. associées à deux observations réalisées sur le même individu à deux instants différents soit inversement proportionnelle à l'écart entre ces deux instants. Pour  $T = 4$ , la matrice  $\mathbf{R}$  s'écrit

$$\mathbf{R} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix},$$

avec :  $\sigma^2 > 0$  ;  $0 < \rho < 1$ . Ainsi, la variance de la v.a.r. associée à toute observation vaut  $\sigma^2$ , la corrélation des v.a.r. associées à deux observations réalisées sur le même individu à deux instants différents  $i$  et  $i'$  valant  $\sigma^2 \rho^{|i-i'|}$ . L'option correspondante s'appelle auto-régressive d'ordre 1 et se déclare avec l'option `type=ar(1)`. Il s'agit d'une structure très courante en pratique.

### Structure de Toeplitz à deux bandes

Il n'y a encore que deux paramètres à estimer dans cette structure, un peu moins générale que la précédente. Pour  $T = 4$ , la matrice  $\mathbf{R}$  s'écrit

$$\mathbf{R} = \begin{pmatrix} \sigma_1^2 & \sigma_2 & 0 & 0 \\ \sigma_2 & \sigma_1^2 & \sigma_2 & 0 \\ 0 & \sigma_2 & \sigma_1^2 & \sigma_2 \\ 0 & 0 & \sigma_2 & \sigma_1^2 \end{pmatrix},$$

avec  $\sigma_1^2 > 0$ . Le paramètre  $\sigma_2$  peut être négatif, ce qui permet de traiter des cas un peu particuliers. La structure correspondante se déclare avec l'option `type=toep(2)`.

### Structure de Toeplitz générale

Moins simple que les précédentes, cette structure nécessite l'estimation de  $T$  paramètres. La matrice  $\mathbf{R}$  est dans ce cas une matrice dite de Toeplitz (quelconque). Elle s'écrit, toujours pour  $T = 4$ , sous la forme

$$\mathbf{R} = \begin{pmatrix} \sigma_1^2 & \sigma_2 & \sigma_3 & \sigma_4 \\ \sigma_2 & \sigma_1^2 & \sigma_2 & \sigma_3 \\ \sigma_3 & \sigma_2 & \sigma_1^2 & \sigma_2 \\ \sigma_4 & \sigma_3 & \sigma_2 & \sigma_1^2 \end{pmatrix},$$

avec  $\sigma_1^2 > 0$ . Les paramètres  $\sigma_2, \sigma_3$  et  $\sigma_4$  peuvent encore être négatifs. La structure correspondante se déclare avec l'option `type=toep`.

### Structure spatiale

La procédure MIXED de SAS propose différentes formes de structures spatiales (on pourra se reporter à la documentation de SAS pour plus de détails). La plus "naturelle" est une généralisation de la structure auto-régressive d'ordre 1. Elle consiste à poser (toujours pour  $T = 4$ ) :

$$\mathbf{R} = \sigma^2 \begin{pmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} & \rho^{t_4-t_1} \\ \rho^{t_2-t_1} & 1 & \rho^{t_3-t_2} & \rho^{t_4-t_2} \\ \rho^{t_3-t_1} & \rho^{t_3-t_2} & 1 & \rho^{t_4-t_3} \\ \rho^{t_4-t_1} & \rho^{t_4-t_2} & \rho^{t_4-t_3} & 1 \end{pmatrix},$$

avec :  $\sigma^2 > 0$  ;  $0 < \rho < 1$  ;  $t_1, t_2, t_3, t_4$  désignent les instants d'observation. Pour des instants régulièrement espacés ( $t = 1, \dots, T$ ), cette structure est identique à la structure auto-régressive d'ordre 1. Mais, dans le cas d'observations irrégulières (cas assez fréquent dans la pratique), ce n'est plus le cas et cette structure spatiale est alors commode. Elle se déclare avec l'option `type=sp(pow) (temps)`, si `temps` est le nom de la variable contenant les instants d'observation. On notera que `pow` signifie `power` et correspond à la fonction puissance utilisée ici. Les autres structures spatiales proposées par MIXED pour la matrice  $\mathbf{R}$  correspondent à d'autres fonctions que l'on doit déclarer à cet endroit là.

### Laquelle choisir ?

Supposons un modèle choisi (modèle complet, modèle additif...) au moyen des tests indiqués au paragraphe 7.3 (par exemple, en ne spécifiant pas au départ de structure sur la matrice  $\mathbf{R}$ ). Se pose alors le problème du choix de la structure de covariance "optimale" pour ce modèle. La façon usuelle de procéder consiste à regarder chacune des structures présentées ci-dessus et à choisir ensuite celle qui minimise un critère de choix de modèle : AIC ou BIC (lorsqu'il y a contradiction entre ces critères, les choses deviennent délicates...). On notera que les expressions des matrices

$\mathbf{H}$  et  $\mathbf{E}$  intervenant dans les tests multivariés introduits au chapitre 5 et utilisés ici ne font pas intervenir la matrice  $\mathbf{R}$ ; le choix de cette dernière n’a donc pas d’influence sur le résultat de ces tests, d’où l’intérêt de procéder comme indiqué (à condition, toutefois, que le nombre d’observations  $n \times T$  soit suffisant).

## 7.5 Cas particulier : la structure “compound symmetry”

### 7.5.1 Propriété préliminaire

On a vu, au paragraphe précédent, que la structure de covariance dite *compound symmetry* consiste à poser :  $\mathbf{R} = \sigma_1^2 \mathbf{I}_T + \sigma_2^2 \mathbf{I}_T$ . D’un point de vue algébrique, on peut montrer que, dans  $\mathbb{R}^T$ , la matrice  $\mathbf{R}$  définie ci-dessus n’admet que deux valeurs propres :  $\lambda_1 = \sigma_1^2$ ;  $\lambda_2 = \sigma_1^2 + T\sigma_2^2$ . La première de ces valeurs propres est d’ordre  $T - 1$ ; la seconde est d’ordre 1. Par conséquent, il existe une matrice  $\mathbf{C}$ , de dimension  $(T - 1) \times T$  et de rang  $T - 1$ , telle que, en posant  $Y_{ij}^* = \mathbf{C}Y_{ij}$ , il vient

$$Y_{ij}^* \sim \mathcal{N}_{T-1}(\mathbf{C}\mu_j, \mathbf{C}\mathbf{R}\mathbf{C}'),$$

avec :

$$\mathbf{C}\mathbf{R}\mathbf{C}' = \sigma_1^2 \mathbf{I}_{T-1}.$$

On se reportera à l’Annexe F pour les justifications de ce qui précède.

### 7.5.2 Conséquences

En supposant que la structure *compound symmetry* soit appropriée aux données considérées, si l’on projette chaque vecteur aléatoire  $Y_{ij}$  sur le sous-espace propre associé à la valeur propre  $\lambda_1$  définie ci-dessus, on obtient, pour le vecteur projeté, une structure de corrélation proportionnelle à la matrice identité, autrement dit on se ramène au cas standard où l’on peut considérer comme indépendantes les différentes observations d’un même individu réalisées aux différents instants. Les tests relatifs aux différents effets (facteur, temps et interactions) sont alors des tests de Fisher usuels, puisqu’on s’est ramené au cas standard unidimensionnel. La taille de l’échantillon correspondant est  $N = n(T - 1)$ , ce qu’on peut obtenir en travaillant sur des différences, par exemple entre chaque instant et l’instant initial.

Enfin, en posant  $U_{ijt} = A_{ij} + U_{ijt}^*$  dans le modèle défini au paragraphe 7.3, avec  $A_{ij} \sim \mathcal{N}(0, \sigma_2^2)$ , on montre, de façon immédiate, que le modèle se ramène à un modèle mixte dont les effets fixes sont ceux du temps, du facteur et des interactions, et les effets aléatoires (les  $A_{ij}$ ) sont ceux attachés aux individus observés. Cette remarque permet d’estimer les composantes de la variance de ce modèle mixte ( $\sigma_1^2$  et  $\sigma_2^2$ ) en utilisant, par exemple, la procédure VARCOMP de SAS. On a ainsi contourné toutes les difficultés provenant de la nature répétée des données...

### 7.5.3 Le test de sphéricité de Mauchly

Les commodités évoquées ci-dessus montrent que la structure de covariance *compound symmetry* présente des avantages certains lorsqu’on doit modéliser des données répétées (et cela était carrément vital lorsqu’on ne disposait pas de logiciel performant comme aujourd’hui). Encore faut-il s’assurer que cette structure soit valide. Un test permet de contrôler l’adéquation de cette structure aux données analysées : c’est le test dit de sphéricité de Mauchly.

Il consiste à estimer la matrice  $\mathbf{R}$  par  $\hat{\mathbf{R}}$ , sans supposer de structure (cas *unstructured*), puis à considérer une matrice  $\mathbf{C}$ , de dimension  $(T - 1) \times T$ , dont les lignes soient orthonormées (au sens de la métrique euclidienne classique de  $\mathbb{R}^T$ ). Le test consiste alors à contrôler la sphéricité de la matrice  $\mathbf{C}\hat{\mathbf{R}}\mathbf{C}'$ , autrement dit sa proportionnalité avec la matrice identité d’ordre  $T - 1$ . Il s’agit d’un test asymptotique de type khi-deux dont le degré de liberté est  $\frac{T(T - 1)}{2} - 1$ .

La *p-value* de ce test est fournie par la procédure GLM de SAS, lorsqu’on a mis l’option `printe` dans la commande `repeated`, à la rubrique *Sphericity Tests* et à la ligne *Orthogonal Components* (cette ligne correspond à la transformation réalisée au moyen de la matrice  $\mathbf{C}$ ).

**Remarque 74** On trouvera une présentation détaillée du test de sphéricité de Mauchly dans la section 7.2 de l’ouvrage de Rencher (1995).

**Remarque 75** *On notera que la procédure GLM de SAS propose, lorsque le test de sphéricité a rejeté la structure de symétrie composée, deux modifications possibles du test de Fisher unidimensionnel, pour pouvoir continuer à l'appliquer (Greenhouse-Geisser et Huynh-Feldt). Nous déconseillons l'usage de ces tests. Il faut signaler qu'on était réduit à les utiliser lorsqu'on ne disposait ni de la procédure MIXED de SAS, ni d'autres logiciels permettant de modéliser des données répétées avec des structures de covariance variées. Ce n'est aujourd'hui plus le cas, et cette sorte de pis-aller n'a plus de raison d'être.*

**Remarque 76** *Lorsqu'on modélise des données répétées avec la procédure GLM de SAS, on dispose donc d'un ensemble de techniques permettant de faire un certain nombre de choses. Toutefois, il faut voir que les seules structures de covariance envisageables avec GLM sont l'absence de structure, la structure d'indépendance (en général inappropriée) et la symétrie composée. Dans le cas où cette dernière est rejetée par le test de sphéricité, on n'a donc pas d'alternative à l'absence de structure (dont les estimations ne sont d'ailleurs pas fournies directement), alors que la procédure MIXED offre toutes les possibilités signalées au paragraphe précédent.*

**Remarque 77** *On notera enfin que, lorsque la structure symétrie composée est acceptable, autrement dit lorsque le test de Mauchly n'est pas significatif, cela ne signifie pas que c'est la structure la mieux adaptée aux données considérées : les principales structures de covariance doivent être envisagée avec la procédure MIXED, avant de choisir celle qui minimise soit le critère AIC, soit le critère BIC.*

En conclusion, on peut dire que la modélisation de données répétées avec SAS nécessite en général les deux procédures GLM et MIXED.

## 7.6 Modèles mixtes pour données répétées

### 7.6.1 Principe

Dans la pratique des modèles pour données répétées, il est courant d'avoir affaire à des situations bien plus complexes que celle envisagée jusqu'à présent (un seul facteur à effets fixes).

On peut tout d'abord envisager plusieurs facteurs croisés, ou encore un mélange de facteurs et de covariables. Rappelons qu'on appelle covariable une variable quantitative explicative (toujours supposée contrôlée, c'est-à-dire non aléatoire), telle qu'on en utilise en analyse de covariance. Nous ne développons pas davantage ce point qui ne présente aucune difficulté particulière : on choisit d'abord les effets significatifs que l'on met dans le modèle (en utilisant, par exemple, les tests multidimensionnels), puis la structure de covariance (dans le catalogue présenté au paragraphe 7.4, en utilisant les critères indiqués). On estime enfin les paramètres du modèle retenu. À ce niveau, précisons que le temps est, par nature, un facteur à effets fixes : les instants d'observation ne sont pas choisis au hasard, ils correspondent aux moments où l'on souhaite étudier le phénomène considéré.

Mais il est surtout fréquent de considérer des modèles mixtes, mêlant des facteurs (et/ou des covariables) à effets fixes, comme indiqué ci-dessus, et des facteurs à effets aléatoires, en général liés aux individus observés. De façon très générale, en renumérotant les individus selon un indice  $i$  variant de 1 à  $n$  (si  $n$  est le nombre total de sujets observés), sans tenir compte du niveau du (ou des) facteur(s) correspondant(s), on pourra écrire un tel modèle sous la forme suivante :

$$Y_i = \mathbf{X}_i\beta + \mathbf{Z}_iA + U_i.$$

Dans cette écriture :

- $Y_i$  est un vecteur aléatoire de  $\mathbb{R}^T$ , si  $T$  est le nombre d'instantants d'observation (de répétitions) ;
- $\mathbf{X}_i$  est la matrice d'incidence associée au  $i$ -ième sujet et relative aux effets fixes ; elle est de dimension  $T \times p$ , si  $p$  est le nombre total d'effets fixes pris en compte dans le modèle, y compris les effets liés au temps ; elle ne comporte que des 0 et des 1 (sauf s'il y a une ou plusieurs covariables) ; pour un indice  $i$  fixé (donc pour un individu fixé), les lignes de  $\mathbf{X}_i$  correspondent aux différents instants d'observation et ne changent que pour les effets liés au temps ;

- $\beta$  est le vecteur de  $\mathbb{R}^p$  contenant les paramètres correspondant aux effets fixes (y compris le temps);
- $\mathbf{Z}_i$  est la matrice d'incidence associée au  $i$ -ième sujet et relative aux effets aléatoires; elle est de dimension  $T \times q$ , si  $q$  est le nombre total d'effets aléatoires pris en compte dans le modèle; elle aussi ne contient que des 0 et des 1; de plus, ses lignes sont constantes pour un indice  $i$  fixé;
- $A$  est le vecteur aléatoire de  $\mathbb{R}^q$  représentant les effets aléatoires; on pose  $A \sim \mathcal{N}_q(0, \mathbf{G})$ , où  $\mathbf{G}$  définit la structure de covariance associée aux effets aléatoires; le plus souvent,  $\mathbf{G}$  est choisie bloc-diagonale; on notera que  $A$  a la même structure qu'en 6.3.1 et que  $\mathbf{G}$  correspond à la matrice introduite dans la remarque 68;
- $U_i$  est le vecteur aléatoire erreur de  $\mathbb{R}^T$ ; sa loi est  $\mathcal{N}_T(0, \mathbf{R})$ ; la structure de covariance associée aux données répétées est donc définie par la covariance de  $U_i$  (on notera que l'on n'a pas indicé  $\mathbf{R}$  par  $i$ , de sorte que l'on est encore dans le cadre d'un modèle homoscédastique); les vecteurs aléatoires  $U_i$  ( $i = 1, \dots, n$ ) sont supposés i.i.d. et indépendants de  $A$ ;
- Ainsi, chaque vecteur aléatoire  $Y_i$  de  $\mathbb{R}^T$  est distribué selon une loi  $\mathcal{N}_T(\mathbf{X}_i\beta, \mathbf{V})$ , avec  $\mathbf{V} = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}$ .

Ce type de modèle, très général, synthétise donc d'une part ce que nous avons vu au chapitre 6, avec les modèles mixtes, d'autre part les aspects spécifiques des modèles pour données répétées, étudiés précédemment dans ce chapitre. Nous ne le développons pas davantage sur le plan théorique, même s'il constitue le quotidien de la pratique des données répétées : c'est plutôt sa mise en œuvre qui nécessite des développements, ce qui relève de la pratique.

**Remarque 78** *Pour retrouver la formule donnée en 7.3.1 pour un modèle sans effets aléatoires, il faut, dans la formule ci-dessus, annuler le terme  $A$  et, dans l'autre formule, d'une part considérer les vecteurs de  $\mathbb{R}^T$  des  $Y_{ijt}$  ( $t = 1, \dots, T$ ), notés  $Y_{ij}$ , d'autre part renuméroter ces vecteurs selon un indice  $i$  variant de 1 à  $n$ ; le nouveau terme  $\mu_i$  de cette formule (noté  $\mu_j$  en 7.3.1) correspond alors à  $\mathbf{X}_i\beta$ .*

### 7.6.2 Usage de la procédure MIXED

Lorsqu'on utilise la procédure MIXED de SAS pour traiter des données répétées, les trois commandes fondamentales de cette procédure sont explicitées ci-dessous (la seconde sera omise dans le cas d'un modèle sans effet aléatoire).

#### La commande “model”

Sa syntaxe est `model y = f1 f2 ...`; elle sert à déclarer les effets fixes du modèle, y compris le temps; elle permet à SAS de construire la “super” matrice d'incidence  $\mathbf{X}$  constituée des matrices  $\mathbf{X}_i$  mises les unes en dessous des autres ( $\mathbf{X}$  est donc de dimension  $nT \times p$ ).

#### La commande “random”

Sa syntaxe est `random a1 a2 ...`; elle sert à déclarer les effets aléatoires, donc à construire la “super” matrice d'incidence  $\mathbf{Z}$  constituée des matrices  $\mathbf{Z}_i$  mises les unes en dessous des autres ( $\mathbf{Z}$  est donc de dimension  $nT \times q$ ); c'est dans cette commande que l'on doit déclarer la structure de covariance  $\mathbf{G}$ , avec l'option `type=...`

#### La commande “repeated”

Sa syntaxe est `repeated / <options>`, deux options étant ici indispensables : `sub=...`, pour déclarer la variable contenant les étiquettes des sujets (variable restant constante lorsque les observations sont répétées sur le même sujet); `type=...`, pour déclarer la structure de covariance intra-sujet (voir le paragraphe 7.4), ce qui permet de construire la matrice  $\mathbf{R}$  associée.

### 7.6.3 Inférence

On a déjà signalé, au chapitre 6, que les tests au sein des modèles mixtes sont assez délicats à mettre en œuvre. Pour un modèle mixte relatif à des données répétées, les choses sont, bien sûr,

encore plus délicates puisque trois ensembles d'éléments sont maintenant à choisir pour déterminer le modèle retenu : les effets fixes, les effets aléatoires et la structure de covariance  $\mathbf{R}$ . Disons tout de suite qu'il n'y a pas de méthode réellement satisfaisante. On peut néanmoins envisager de procéder par étape : on commence, comme indiqué à la fin du point 7.4, par tester les effets fixes puis, avec les effets retenus, on choisit la structure de covariance. Enfin, on estime les composantes de la variance correspondant aux effets aléatoires et on retient les effets les plus importants.

Concernant les estimations, c'est le plus souvent la méthode REML (maximum de vraisemblance restreint) qui est utilisée. On trouvera divers compléments sur l'inférence dans les modèles mixtes pour données répétées dans l'ouvrage de Verbeke & Molenberghs (2000).

## 7.7 Illustration

### Les données

Il s'agit d'un exemple fictif de données répétées. Il comporte 20 individus (leurs codes, de 01 à 20, sont en première colonne), un facteur fixe à 4 niveaux (codés de 1 à 4, en seconde colonne), et une variable réponse, observée à 3 instants : les réponses aux instants 1, 2 et 3 sont dans les 3 dernières colonnes.

Les données sont reproduites ci-dessous.

```
01 1 60 70 80
02 1 65 70 75
03 1 60 75 75
04 1 70 75 80
05 1 75 75 90
06 2 60 65 85
07 2 80 90 100
08 2 65 80 95
09 2 85 90 95
10 2 90 90 90
11 3 90 95 100
12 3 85 90 95
13 3 85 85 100
14 3 80 90 100
15 3 70 80 90
16 4 80 80 104
17 4 85 95 114
18 4 90 95 109
19 4 95 100 114
20 4 90 90 129
```

### Le programme SAS

Le programme SAS ci-dessous permet de réaliser différents traitements de ces données, en utilisant les procédures GLM (avec la commande REPEATED), puis MIXED (après transformation indispensable des données), enfin VARCOMP, pour retrouver certains résultats de MIXED et mieux comprendre le fonctionnement de ces traitements.

```
* ----- ;
*          IMPORTATION DES DONNEES          ;
* ----- ;
data fic1;
infile 'fic.don';
input indiv trait y1-y3;
run;
* ----- ;
*          GLM AVEC REPEATED                ;
* ----- ;
proc glm data=fic1;
class trait;
```

```

model y1-y3 = trait / ss3;
repeated temps / printh printe;
run;
* ----- ;
*           et avec CONTRAST           ;
* ----- ;
proc glm data=fic1;
class trait;
model y1-y3 = trait / ss3 nouni;
repeated temps contrast(1)/ printm printh printe;
run;
quit;
* ----- ;
*   TRANSFORMATION DES DONNEES POUR MIXED ;
* ----- ;
data fic2;
set fic1;
y=y1; temps=1; output;
y=y2; temps=2; output;
y=y3; temps=3; output;
* ----- ;
*           PROC MIXED - REML           ;
*           COMPOUND SYMMETRY           ;
* ----- ;
proc mixed data=fic2 method=reml;
class trait temps;
model y = trait | temps;
repeated / sub=indiv type=cs;
run;
* ----- ;
*           sans interactions           ;
* ----- ;
proc mixed data=fic2 method=reml;
class trait temps;
model y = trait temps / s;
repeated / sub=indiv type=cs r;
run;
quit;
* ----- ;
*           PROC VARCOMP               ;
* ----- ;
proc varcomp data=fic2 method=reml;
class indiv trait temps;
model y = temps trait indiv / fixed=2;
run;
quit;

```

### Les sorties de la procédure GLM

PAGE 1

-----

#### The GLM Procedure

##### Class Level Information

Class	Levels	Values
trait	4	1 2 3 4
Number of Observations Read		20
Number of Observations Used		20

PAGE 2

-----

Dependent Variable: y1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1320.000000	440.000000	5.87	0.0067
Error	16	1200.000000	75.000000		
Corrected Total	19	2520.000000			

R-Square	Coeff Var	Root MSE	y1 Mean
0.523810	11.10289	8.660254	78.00000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trait	3	1320.000000	440.000000	5.87	0.0067

PAGE 3

-----

Dependent Variable: y2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1010.000000	336.666667	6.19	0.0054
Error	16	870.000000	54.375000		
Corrected Total	19	1880.000000			

R-Square	Coeff Var	Root MSE	y2 Mean
0.537234	8.778501	7.373941	84.00000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trait	3	1010.000000	336.666667	6.19	0.0054

PAGE 4

-----

Dependent Variable: y3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2950.000000	983.333333	22.16	<.0001
Error	16	710.000000	44.375000		
Corrected Total	19	3660.000000			

	R-Square	Coeff Var	Root MSE	y3 Mean	
	0.806011	6.939017	6.661456	96.00000	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
trait	3	2950.000000	983.333333	22.16	<.0001

PAGE 5

-----

## Repeated Measures Analysis of Variance

## Repeated Measures Level Information

Dependent Variable	y1	y2	y3
Level of temps	1	2	3

## Partial Correlation Coefficients from the Error SSCP Matrix / Prob &gt; |r|

DF = 16	y1	y2	y3
y1	1.000000	0.817215 <.0001	0.476687 0.0530
y2	0.817215 <.0001	1.000000	0.445327 0.0732
y3	0.476687 0.0530	0.445327 0.0732	1.000000

E = Error SSCP Matrix

temps\_N represents the contrast between the nth level of temps and the last

	temps_1	temps_2
temps_1	1030	755
temps_2	755	880

## Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob &gt; |r|

DF = 16	temps_1	temps_2
temps_1	1.000000	0.793025 0.0001
temps_2	0.793025 0.0001	1.000000

## Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	2	0.3688221	14.961612	0.0006
Orthogonal Components	2	0.7564513	4.186756	0.1233

PAGE 6

-----

H = Type III SSCP Matrix for temps

temps\_N represents the contrast between the nth level of temps and the last

	temps_1	temps_2
temps_1	6480	4320
temps_2	4320	2880

MANOVA Test Criteria and Exact F Statistics  
for the Hypothesis of no temps Effect

H = Type III SSCP Matrix for temps

E = Error SSCP Matrix

S=1 M=0 N=6.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.13552715	47.84	2	15	<.0001
Pillai's Trace	0.86447285	47.84	2	15	<.0001
Hotelling-Lawley Trace	6.37859532	47.84	2	15	<.0001
Roy's Greatest Root	6.37859532	47.84	2	15	<.0001

H = Type III SSCP Matrix for temps\*trait

temps\_N represents the contrast between the nth level of temps and the last

	temps_1	temps_2
temps_1	450	555
temps_2	555	690

MANOVA Test Criteria and F Approximations for  
the Hypothesis of no temps\*trait Effect

H = Type III SSCP Matrix for temps\*trait

E = Error SSCP Matrix

S=2 M=0 N=6.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.55370370	1.72	6	30	0.1507
Pillai's Trace	0.45037037	1.55	6	32	0.1940
Hotelling-Lawley Trace	0.79866221	1.94	6	18.326	0.1275
Roy's Greatest Root	0.78934068	4.21	3	16	0.0225

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

PAGE 7

-----

## Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trait	3	4890.000000	1630.000000	12.98	0.0001
Error	16	2010.000000	125.625000		

PAGE 8

-----

## Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temps	2	3360.000000	1680.000000	69.82	<.0001
temps*trait	6	390.000000	65.000000	2.70	0.0309
Error(temps)	32	770.000000	24.062500		

Source	Adj Pr > F	
	G - G	H - F
temps	<.0001	<.0001
temps*trait	0.0446	0.0309
Error(temps)		
Greenhouse-Geisser Epsilon	0.8042	
Huynh-Feldt Epsilon	1.0480	

## La procédure GLM avec CONTRAST

PAGE 1

-----

The GLM Procedure  
Repeated Measures Analysis of Variance

## Repeated Measures Level Information

Dependent Variable	y1	y2	y3
Level of temps	1	2	3

## Partial Correlation Coefficients from the Error SSCP Matrix / Prob &gt; |r|

DF = 16	y1	y2	y3
y1	1.000000	0.817215 <.0001	0.476687 0.0530
y2	0.817215 <.0001	1.000000	0.445327 0.0732
y3	0.476687 0.0530	0.445327 0.0732	1.000000

temps\_N represents the contrast between the nth level of temps and the 1st

M Matrix Describing Transformed Variables

	y1	y2	y3
temps_2	-1.000000000	1.000000000	0.000000000
temps_3	-1.000000000	0.000000000	1.000000000

E = Error SSCP Matrix

temps\_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3
temps_2	400	275
temps_3	275	1030

Partial Correlation Coefficients from the Error SSCP Matrix of the  
Variables Defined by the Specified Transformation / Prob > |r|

DF = 16	temps_2	temps_3
temps_2	1.000000	0.428434 0.0862
temps_3	0.428434 0.0862	1.000000

PAGE 2

-----

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	2	0.6579784	6.278748	0.0433
Orthogonal Components	2	0.7564513	4.186756	0.1233

H = Type III SSCP Matrix for temps

temps\_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3
temps_2	720	2160
temps_3	2160	6480

MANOVA Test Criteria and Exact F Statistics

for the Hypothesis of no temps Effect

H = Type III SSCP Matrix for temps

E = Error SSCP Matrix

S=1 M=0 N=6.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.13552715	47.84	2	15	<.0001
Pillai's Trace	0.86447285	47.84	2	15	<.0001
Hotelling-Lawley Trace	6.37859532	47.84	2	15	<.0001
Roy's Greatest Root	6.37859532	47.84	2	15	<.0001

H = Type III SSCP Matrix for temps\*trait

temps\_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3
temps_2	30	-105
temps_3	-105	450

PAGE 3

-----

MANOVA Test Criteria and F Approximations for  
the Hypothesis of no temps\*trait Effect

H = Type III SSCP Matrix for temps\*trait

E = Error SSCP Matrix

S=2 M=0 N=6.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.55370370	1.72	6	30	0.1507
Pillai's Trace	0.45037037	1.55	6	32	0.1940
Hotelling-Lawley Trace	0.79866221	1.94	6	18.326	0.1275
Roy's Greatest Root	0.78934068	4.21	3	16	0.0225

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

PAGE 4

-----

Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trait	3	4890.000000	1630.000000	12.98	0.0001
Error	16	2010.000000	125.625000		

PAGE 5

-----

Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temps	2	3360.000000	1680.000000	69.82	<.0001
temps*trait	6	390.000000	65.000000	2.70	0.0309
Error(temps)	32	770.000000	24.062500		

Source	Adj Pr > F	
	G - G	H - F
temps	<.0001	<.0001
temps*trait	0.0446	0.0309
Error(temps)		
Greenhouse-Geisser Epsilon		0.8042
Huynh-Feldt Epsilon		1.0480

### Les sorties de la procédure MIXED

Il est tout d'abord nécessaire de procéder à une transformation des données, pour disposer les différents instants d'observation en lignes et non plus en colonnes. Ensuite, on a choisi la structure de covariance *compound symmetry* (puisque le test de Maucly n'est pas significatif) et utilisé d'abord un modèle avec interactions, puis un modèle additif.

PAGE 1

-----

#### The Mixed Procedure

##### Model Information

Data Set	WORK.FIC2
Dependent Variable	y
Covariance Structure	Compound Symmetry
Subject Effect	indiv
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

##### Class Level Information

Class	Levels	Values
trait	4	1 2 3 4
temps	3	1 2 3

##### Dimensions

Covariance Parameters	2
Columns in X	20
Columns in Z	0
Subjects	20
Max Obs Per Subject	3

##### Number of Observations

Number of Observations Read	60
Number of Observations Used	60
Number of Observations Not Used	0

## Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	350.36360353	
1	1	334.64512218	0.00000000

Convergence criteria met.

## Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	indiv	33.8542
Residual		24.0625

PAGE 2

-----

## Fit Statistics

-2 Res Log Likelihood	334.6
AIC (smaller is better)	338.6
AICC (smaller is better)	338.9
BIC (smaller is better)	340.6

## Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	15.72	<.0001

## Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trait	3	16	12.98	0.0001
temps	2	32	69.82	<.0001
trait*temps	6	32	2.70	0.0309

PAGE 3

-----

## Model Information

Data Set	WORK.FIC2
Dependent Variable	y
Covariance Structure	Compound Symmetry
Subject Effect	indiv
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

## Class Level Information

Class	Levels	Values
trait	4	1 2 3 4
temps	3	1 2 3

## Dimensions

Covariance Parameters	2
Columns in X	8
Columns in Z	0
Subjects	20
Max Obs Per Subject	3

## Number of Observations

Number of Observations Read	60
Number of Observations Used	60
Number of Observations Not Used	0

## Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	388.88556695	
1	1	376.20962184	0.00000000

Convergence criteria met.

## Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3
1	62.2259	31.6996	31.6996
2	31.6996	62.2259	31.6996
3	31.6996	31.6996	62.2259

PAGE 4

-----

## Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	indiv	31.6996
Residual		30.5263

## Fit Statistics

-2 Res Log Likelihood	376.2
AIC (smaller is better)	380.2
AICC (smaller is better)	380.4
BIC (smaller is better)	382.2

## Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	12.68	0.0004

## Solution for Fixed Effects

Effect	trait	temps	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept			108.00	3.0647	16	35.24	<.0001
trait	1		-25.0000	4.0927	16	-6.11	<.0001
trait	2		-14.0000	4.0927	16	-3.42	0.0035
trait	3		-9.0000	4.0927	16	-2.20	0.0429
trait	4		0	.	.	.	.
temps		1	-18.0000	1.7472	38	-10.30	<.0001
temps		2	-12.0000	1.7472	38	-6.87	<.0001
temps		3	0	.	.	.	.

## Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trait	3	16	12.98	0.0001
temps	2	38	55.03	<.0001

## Les sorties de la procédure VARCOMP

Elles n'apportent rien de nouveau. Elles permettent seulement de retrouver certains résultats de la procédure MIXED, moyennant une déclaration particulière des individus (voir le programme plus haut).

## Variance Components Estimation Procedure

## Class Level Information

Class	Levels	Values
indiv	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
trait	4	1 2 3 4
temps	3	1 2 3

Number of Observations Read 60  
 Number of Observations Used 60

Dependent Variable: y

## REML Iterations

Iteration	Objective	Var(indiv)	Var(Error)
0	207.2392071872	31.6995614035	30.5263157895
1	207.2392071872	31.6995614035	30.5263157895

Convergence criteria met.

## REML Estimates

Variance Component	Estimate
Var(indiv)	31.69956
Var(Error)	30.52632

## Asymptotic Covariance Matrix of Estimates

	Var(indiv)	Var(Error)
Var(indiv)	224.63890	-16.34835
Var(Error)	-16.34835	49.04505

# Annexes



## Annexe A

# À propos de la méthode de Bonferroni

*L'objet de cette annexe est de rappeler la méthode de Bonferroni et de préciser les commandes `means` et `lsmeans` de la procédure GLM de SAS, dans le cadre de la modélisation linéaire gaussienne. En particulier, bien que très proches, ces deux commandes ne fonctionnent pas de la même façon pour la mise en œuvre de la méthode de Bonferroni et présentent, l'une comme l'autre, des bizarreries que nous indiquons.*

### A.1 Rappels sur la méthode de Bonferroni

Dans le cadre du modèle linéaire gaussien, le test usuel, permettant de tester toute hypothèse nulle linéaire en les paramètres, est le test de Fisher ; ce test se ramène à un test de Student, plus simple, lorsque l'hypothèse nulle ne porte que sur un seul paramètre.

Dans la pratique, un problème qui se pose souvent avec ces tests est celui des tests multiples. L'idée est la suivante : si l'on travaille avec un niveau de test (une erreur de première espèce) de 5%, et si l'on fait un seul test, alors, lorsque la valeur de la statistique de test dépasse la valeur limite (fournie par les tables ou par les logiciels), ce qui conduit à rejeter l'hypothèse nulle  $H_0$ , il y a un risque (une probabilité) de 0,05, pour que l'on rejette à tort cette hypothèse ; ce risque est donc contrôlé et faible. Par contre, si l'on répète ce même test pour divers paramètres avec les mêmes données (dans le cadre de la même expérience), le risque de rejeter à tort augmente d'autant et devient grand si l'on fait de nombreux tests. De plus, ce risque n'est plus contrôlé (à la limite, si l'on fait 100 tests différents avec les mêmes données et si l'on rejette 5 fois  $H_0$ , il y a de fortes chances pour que ces rejets soient faits à tort).

En particulier, dans le cadre d'une ANOVA (analyse de variance, ou plan factoriel), une fois un modèle choisi, il est fréquent qu'on cherche à savoir si les différences des moyennes de la variable réponse entre les niveaux d'un facteur pris en compte dans le modèle sont significatives ou non. Plus ces niveaux sont nombreux, plus il faut faire de tests pour répondre à cette question, et plus le risque de conclure à tort à la significativité des différences est grand.

Pour se prémunir contre l'augmentation du risque de première espèce dans les tests multiples, il existe différentes méthodes, la plus célèbre étant celle dite de Bonferroni. Cette dernière consiste à majorer le risque réel pour l'ensemble des tests (impossible à déterminer) par le niveau de test choisi (en général 5%), en divisant pour cela le niveau de chacun des tests effectués par le nombre total de tests. Ainsi, si on dispose d'un facteur à 4 niveaux et si on veut tester toutes les différences deux à deux entre les niveaux, il y a 6 tests à réaliser, de sorte que chacun est fait avec un niveau de 5/6, soit 0,83%, ce qui est faible. Pour 5 niveaux, on doit réaliser chaque test au niveau de 0,1% et ainsi de suite. La méthode est trop conservatrice (elle a tendance à ne rejeter que très rarement  $H_0$ , à cause du niveau très faible de chaque test) et s'avère peu intéressante dès que le nombre de niveaux est supérieur ou égal à 5 : en gros, la correction est alors telle que le remède est pire que le mal...

On notera que des méthodes nouvelles pour diminuer le risque de rejeter à tort  $H_0$ , autrement dit pour se prémunir des faux positifs, ont été mises au point assez récemment. Elles sont nettement plus performantes que la méthode de Bonferroni et nous ne conseillons d'utiliser cette dernière, le cas échéant, que pour des facteurs à 3 ou 4 niveaux.

## A.2 Les commandes means et lsmeans de la procédure GLM de SAS

### A.2.1 Principe général

Lorsqu'on réalise une ANOVA avec la procédure GLM de SAS, on peut utiliser les deux commandes `means` et `lsmeans` pour calculer les moyennes partielles de la variable réponse  $Y$  au sein des niveaux des facteurs entrant dans le modèle considéré et pour tester la significativité des différences entre ces moyennes partielles.

Dans un premier temps, nous allons utiliser le célèbre exemple des rendements de vaches laitières pour illustrer l'usage de ces deux commandes (se reporter au chapitre 3 pour la présentation de ces données). Nous considérons que le modèle le mieux adapté à ces données est le modèle complet (effets de chacun des deux facteurs et des interactions) et nous exécutons le programme SAS suivant :

```
data vach;
infile 'vach.don';
input f1 f2 y;
run;
* ----- ;
proc glm data=vach;
class f1 f2;
model y = f1 f2 f1*f2 / ss3;
means f2;
lsmeans f2;
run;
```

En voici les principaux résultats :

```

                                The GLM Procedure

                                Class Level Information

                                Class          Levels    Values
                                -----
                                f1              2         1 2
                                f2              4         1 2 3 4

                                Number of Observations Read          40

Dependent Variable: y

                                Sum of
Source          DF          Squares    Mean Square    F Value    Pr > F
-----
Model           7          331.6000000    47.3714286    17.54    <.0001
Error          32           86.4000000     2.7000000
Corrected Total 39          418.0000000

                                R-Square    Coeff Var    Root MSE      y Mean
-----
                                0.793301    13.69306    1.643168      12.00000
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	0.4000000	0.4000000	0.15	0.7029
f2	3	290.2000000	96.7333333	35.83	<.0001
f1*f2	3	41.0000000	13.6666667	5.06	0.0056

\*\*\*\*\*

Level of f2	N	Mean	Std Dev
1	10	9.1000000	1.37032032
2	10	11.0000000	1.94365063
3	10	11.5000000	1.43372088
4	10	16.4000000	2.54732976

\*\*\*\*\*

Least Squares Means

f2	y LSMEAN
1	9.1000000
2	11.0000000
3	11.5000000
4	16.4000000

Les deux dernières parties des résultats ci-dessus sont celles correspondant respectivement aux commandes `means` et `lsmeans`. On pourra vérifier que les moyennes fournies sont tout simplement celles des 10 observations correspondant à chaque niveau du second facteur. On notera que `means` fournit en plus l'estimation usuelle de l'écart-type des 10 observations de chaque niveau de `f2` (avec un dénominateur en  $n - 1$ ).

### A.2.2 Tests des différences et méthode de Bonferroni

Les deux commandes considérées permettent aussi de mettre en œuvre les tests de comparaison, avec ou sans la correction de Bonferroni. Commençons par regarder ce que fait la commande `means` pour tester les différences entre niveaux du facteur `f2` et comment elle met en œuvre cette correction.

```
proc glm data=vach;
class f1 f2;
model y = f1 | f2 / ss3;
means f2 / t;
means f2 / bon;
run;
```

Voici les résultats.

t Tests (LSD) for y

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	32
Error Mean Square	2.7
Critical Value of t	2.03693
Least Significant Difference	1.4968

Means with the same letter are not significantly different.

t Grouping	Mean	N	f2
A	16.4000	10	4
B	11.5000	10	3
B	11.0000	10	2
C	9.1000	10	1

\*\*\*\*\*

Bonferroni (Dunn) t Tests for y

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	32
Error Mean Square	2.7
Critical Value of t	2.81234
Minimum Significant Difference	2.0666

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	f2
A	16.4000	10	4
B	11.5000	10	3
B	11.0000	10	2
C B	11.0000	10	2
C			
C	9.1000	10	1

L'option `t` de la commande `means` réalise les tests de Student (*t-tests*) de comparaison des moyennes de  $Y$  pour les différents niveaux de  $f_2$  (chaque test est fait avec un niveau de 5%) et présente les résultats sous forme de regroupement ou non de ces niveaux. Le principe est exactement le même avec l'option `bon`, mais chaque test est maintenant fait avec un niveau six fois plus petit et les résultats sont, bien sûr, différents.

Regardons maintenant ce fait la commande `lsmeans` dans les mêmes conditions.

```
proc glm data=vach;
class f1 f2;
model y = f1 | f2 / ss3;
lsmeans f2 / pdiff;
lsmeans f2 / tdiff;
lsmeans f2 / adj=bon tdiff;
run;
```

Voici les résultats.

Least Squares Means		
f2	y LSMEAN	LSMEAN Number
1	9.1000000	1
2	11.0000000	2
3	11.5000000	3
4	16.4000000	4

Least Squares Means for effect f2  
 Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: y

i/j	1	2	3	4
1		0.0145	0.0026	<.0001
2	0.0145		0.5011	<.0001
3	0.0026	0.5011		<.0001
4	<.0001	<.0001	<.0001	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

\*\*\*\*\*

Least Squares Means

f2	y LSMEAN	LSMEAN Number
1	9.1000000	1
2	11.0000000	2
3	11.5000000	3
4	16.4000000	4

Least Squares Means for Effect f2  
 t for H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: y

i/j	1	2	3	4
1		-2.58557	-3.26599	-9.93404
		0.0145	0.0026	<.0001
2	2.585573		-0.68041	-7.34847
	0.0145		0.5011	<.0001
3	3.265986	0.680414		-6.66806
	0.0026	0.5011		<.0001
4	9.934042	7.348469	6.668055	
	<.0001	<.0001	<.0001	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

\*\*\*\*\*

Least Squares Means

Adjustment for Multiple Comparisons: Bonferroni

f2	y LSMEAN	LSMEAN Number
1	9.1000000	1
2	11.0000000	2
3	11.5000000	3
4	16.4000000	4

Least Squares Means for Effect f2  
 t for H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: y

i/j	1	2	3	4
1		-2.58557 0.0869	-3.26599 0.0156	-9.93404 <.0001
2	2.585573 0.0869		-0.68041 1.0000	-7.34847 <.0001
3	3.265986 0.0156	0.680414 1.0000		-6.66806 <.0001
4	9.934042 <.0001	7.348469 <.0001	6.668055 <.0001	

Les deux premiers tableaux sont comparables, le premier ne donnant que les *p-values* des tests de Student de comparaison des moyennes deux à deux, le second donnant en plus les statistiques de tests (il s'agit ici de tests multiples, non corrigés). Le troisième et dernier tableau reprend les mêmes statistiques de tests que le second, mais il ne donne pas les mêmes *p-values*, car il fait la correction de Bonferroni et utilise un niveau de test de 0,83%. De façon plus précise, il multiplie les précédentes *p-values* par 6 (nombre de différences testées), ce qui est assez discutable (c'est approximatif).

Bien entendu les deux présentations faites par `means` et par `lsmeans` sont équivalentes, bien que différentes; ceci est vrai que ce soit sans la correction ou avec la correction de Bonferroni.

### A.2.3 Cas particulier du modèle additif : premières bizarreries

Bien que les effets d'interactions soient significatifs dans ce modèle et que les effets principaux du facteur f1 ne le soient pas, considérons néanmoins, pour illustrer notre propos, le modèle additif avec les deux facteurs et sans les interactions. Dans ce modèle, intéressons nous aux tests de comparaison des niveaux du second facteur, avec la correction de Bonferroni, et comparons encore les sorties des commandes `means` et `lsmeans`.

Voici le programme SAS :

```
proc glm data=vach;
class f1 f2;
model y = f1 f2 / ss3;
means f2;
lsmeans f2;
run;
* ----- ;
proc glm data=vach;
class f1 f2;
model y = f1 f2 / ss3 solution;
means f2 / bon;
lsmeans f2 / adj=bon tdiff;
run;
```

Et en voici les principales sorties;

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	290.6000000	72.6500000	19.96	<.0001
Error	35	127.4000000	3.6400000		
Corrected Total	39	418.0000000			



Means with the same letter are not significantly different.

Bon Grouping	Mean	N	f2
A	16.4000	10	4
B	11.5000	10	3
B			
C B	11.0000	10	2
C			
C	9.1000	10	1

\*\*\*\*\*

Least Squares Means  
Adjustment for Multiple Comparisons: Bonferroni

f2	y LSMEAN	LSMEAN Number
1	9.1000000	1
2	11.0000000	2
3	11.5000000	3
4	16.4000000	4

Least Squares Means for Effect f2  
t for H0: LSMEAN(i)=LSMEAN(j) / Pr > |t|

Dependent Variable: y

i/j	1	2	3	4
1		-2.22683 0.1949	-2.81284 0.0480	-8.55573 <.0001
2	2.226834 0.1949		-0.58601 1.0000	-6.3289 <.0001
3	2.812843 0.0480	0.586009 1.0000		-5.74289 <.0001
4	8.555732 <.0001	6.328898 <.0001	5.742889 <.0001	

Laissons de côté les résultats concernant le modèle, qui n'est manifestement pas adapté aux données, et regardons les résultats des deux commandes `means` et `lsmeans`.

En utilisation basique, sans test de comparaison, ces deux commandes redonnent exactement les mêmes résultats qu'avec le modèle complet, autrement dit on obtient, avec chacune des deux commandes, les moyennes de la variable réponse pour l'ensemble des observations réalisées dans chaque niveau du second facteur et, uniquement avec la première commande, les écarts-types des mêmes observations. On peut déjà s'étonner de ces résultats inchangés, compte tenu que les estimations des paramètres ont, elles, changé.

Regardons maintenant les résultats lorsqu'on demande, en plus, les tests de comparaison, avec la correction de Bonferroni. Avec `means`, les moyennes de la variable réponse dans les 4 niveaux de f2 sont toujours les mêmes, mais les valeurs critiques et la différence significative minimum ont changé par rapport au modèle complet, ce qui est normal puisque le modèle a changé. Avec `lsmeans`, les moyennes de la variable réponse dans les 4 niveaux de f2 sont encore les mêmes, mais différences et *p-values* ont également changé, pour tenir compte du changement de modèle. On pourra vérifier que ces différences sont celles obtenues en divisant l'écart entre les estimations des paramètres des effets principaux de f2 par leur erreur-type, ce qui est logique. On pourra encore vérifier que les résultats des deux commandes sont ici parfaitement cohérents (voir la différence significative minimum dans `means` et la *p-value* du test entre les niveaux 1 et 3 dans `lsmeans`, pour une différence très proche de la valeur en question).

En conclusion, les tests de comparaison des niveaux du facteur f2, avec la correction de Bonferroni, sont corrects avec les deux commandes, mais les valeurs moyennes, en particulier celles appelées valeurs ajustées dans `lsmeans`, sont bizarres, dans la mesure où elles ne correspondent pas aux estimations fournies par le modèle. On pourra en effet vérifier que, dans le modèle additif, les estimations des niveaux de f2 sont respectivement 9.0 10.9 11.4 16.3 (ajouter l'estimation de l'effet général à celle de chaque niveau du facteur). Pour terminer, on notera que les différences deux à deux entre ces deux séries d'estimations sont les mêmes, de sorte que cette bizarrerie n'a pas de conséquence sur les tests de comparaison.

#### A.2.4 Cas particulier d'un plan incomplet : nouvelles bizarreries

Pour illustrer les particularités des commandes `means` et `lsmeans` dans le cas d'un plan incomplet, nous allons considérer le cas d'un plan en blocs, incomplet et équilibré, avec un seul facteur. L'exemple traité est celui de 5 traitements (le facteur) testés sur 10 patients (les blocs). Les données et leur description sont fournies dans la section 4.

Les effets d'interactions ne pouvant pas être pris en compte dans ce cas (pas assez d'observations) et ne présentant, au demeurant, que peu d'intérêt, c'est le modèle additif qui est considéré ici, et il s'avère que ce modèle est très significatif, possède un très bon coefficient  $R^2$  (0,97) et que chacun des 2 facteurs (bloc et traitement) est très significatif. Ce modèle est donc bien adapté aux données (de plus, on peut vérifier que le graphique des résidus ne met en évidence aucune particularité).

Pour étudier les différences entre les moyennes de  $Y$  aux différents niveaux du facteur traitement, regardons ce que donne la méthode de Bonferroni avec les deux commandes `means` et `lsmeans`. Exécutons le programme SAS suivant :

```
data traitmt;
infile 'traitmt.don';
input bloc trait y;
run;
* ----- ;
proc glm data=traitmt;
class bloc trait;
model y = bloc trait / ss3 solution;
means trait / bon;
lsmeans trait / adj=bon tdiff;
run;
```

En voici les résultats :

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	2091.200000	160.861538	34.88	<.0001
Error	16	73.800000	4.612500		
Corrected Total	29	2165.000000			
	R-Square	Coeff Var	Root MSE	y Mean	
	0.965912	4.295346	2.147673	50.00000	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
bloc	9	2076.200000	230.688889	50.01	<.0001
trait	4	345.200000	86.300000	18.71	<.0001

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		55.06666667 B	1.46714008	37.53	<.0001
bloc	1	4.80000000 B	1.86681547	2.57	0.0205
bloc	2	-2.46666667 B	1.86681547	-1.32	0.2050
bloc	3	1.33333333 B	1.81107703	0.74	0.4723
bloc	4	-3.00000000 B	1.81107703	-1.66	0.1171
bloc	5	-0.46666667 B	1.81107703	-0.26	0.7999
bloc	6	6.46666667 B	1.81107703	3.57	0.0026
bloc	7	-2.66666667 B	1.86681547	-1.43	0.1724
bloc	8	10.26666667 B	1.81107703	5.67	<.0001
bloc	9	-24.93333333 B	1.81107703	-13.77	<.0001
bloc	10	0.00000000 B	.	.	.
trait	1	-9.80000000 B	1.35830777	-7.21	<.0001
trait	2	-6.80000000 B	1.35830777	-5.01	0.0001
trait	3	-1.00000000 B	1.35830777	-0.74	0.4723
trait	4	-2.40000000 B	1.35830777	-1.77	0.0963
trait	5	0.00000000 B	.	.	.

\*\*\*\*\*

Bonferroni (Dunn) t Tests for y

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	16
Error Mean Square	4.6125
Critical Value of t	3.25199
Minimum Significant Difference	4.0323

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	trait
A	51.000	6	5
A			
A	50.500	6	2
A			
A	50.000	6	3
A			
A	49.500	6	4
A			
A	49.000	6	1

\*\*\*\*\*

Least Squares Means

Adjustment for Multiple Comparisons: Bonferroni

trait	y LSMEAN	LSMEAN Number
1	44.2000000	1
2	47.2000000	2
3	53.0000000	3
4	51.6000000	4
5	54.0000000	5

Least Squares Means for Effect trait  
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: y

i/j	1	2	3	4	5
1		-2.20863	-6.47865	-5.44796	-7.21486
		0.4213	<.0001	0.0005	<.0001
2	2.208631		-4.27002	-3.23932	-5.00623
	0.4213		0.0059	0.0514	0.0013
3	6.47865	4.270019		1.030694	-0.73621
	<.0001	0.0059		1.0000	1.0000
4	5.447955	3.239325	-1.03069		-1.7669
	0.0005	0.0514	1.0000		0.9631
5	7.21486	5.006229	0.73621	1.766904	
	<.0001	0.0013	1.0000	0.9631	

Cette fois, les résultats de la commande `means` sont incorrects, les moyennes calculées pour chaque niveau du facteur traitement étant les moyennes de toutes les observations faites avec ce traitement, sans tenir compte de l'effet bloc. Néanmoins, la valeur critique du `t` et la différence significative minimum sont, pour leur part, correctes.

Si l'on regarde maintenant les résultats fournis par la commande `lsmeans`, ils sont corrects en ce qui concerne les tests (réalisés avec la correction de Bonferroni). Mais, les moyennes fournies dans la colonne "y LSMEAN" sont bizarres, pour ne pas dire incompréhensibles... En effet, on s'attend à trouver ici les estimations des valeurs moyennes de  $Y$ , pour chaque traitement, fournies par le modèle considéré; autrement dit, la somme de l'effet général, ou *intercept*, et des effets principaux des niveaux du traitement. Or ce n'est pas ce que donne SAS, l'*intercept* de 55.07 étant remplacé par 54, valeur a priori inexplicable. En fait, les "pseudo-moyennes" données ici par SAS sont déterminées à partir des estimations des effets principaux du facteur traitement dans le modèle considéré, en leur ajoutant une quantité  $y_0$  telle que la moyenne des valeurs ainsi obtenues soit égale à la moyenne générale de  $Y$  : si l'on note  $\hat{a}_j$  les estimations SAS des 5 effets principaux et  $\bar{y}$  la moyenne générale de  $Y$  sur l'ensemble des 30 observations réalisées, il vient :

$$y_0 = \bar{y} - \frac{\sum_{j=1}^5 \hat{a}_j}{5}.$$

En conclusion, déjà que l'usage de la correction de Bonferroni n'est pas vraiment judicieuse dans bien des cas, la façon dont procède SAS, avec les commandes `means` et `lsmeans` de la procédure GLM, est assez déconcertante : il convient donc de manipuler ces commandes avec beaucoup de précautions.

### A.3 Usage de `lsmeans` pour les graphiques d'interactions

Nous venons de voir que la commande `lsmeans` présente certaines bizarreries. Il serait dommage de ne pas signaler ici qu'elle présente par ailleurs l'avantage de permettre la réalisation des graphiques d'interactions dans les modèles d'ANOVA à deux facteurs croisés, ou plus.

Pour illustrer cette possibilité, nous reprenons l'exemple du rendement des vaches laitières pour lequel nous allons réaliser le premier graphique d'interactions (les deux se font de la même manière).

Voici le programme SAS qui permet de faire ce graphique :

```
proc glm data=vach noprint;
class f1 f2;
model y = f1 | f2 / ss3;
lsmeans f1 | f2 / out=graph;
run;
* ----- ;
proc print data=graph;
run;
```

```

* ----- ;
proc gplot data=graph;
axis1 label=('premier facteur') order=(1 to 2 by 1)
      minor=none length=6cm;
axis2 label=('moyenne' justify=right 'des y')
      order=(6 to 22 by 2) minor=none length=6cm;
symbol1 i=join v=dot cv=black;
symbol2 i=join v=triangle cv=black;
symbol3 i=join v=circle cv=black;
symbol4 i=join v=# cv=black;
symbol5 i=join v=% cv=black;
plot lsmean*f1=f2 / haxis=axis1 vaxis=axis2;
run;
goptions reset=all;
quit;

```

Afin de comprendre comment se passent les choses, voici le contenu du fichier “graph” :

Obs	_NAME_	f1	f2	LSMEAN	STDERR
1	y	1	.	12.1	0.36742
2	y	2	.	11.9	0.36742
3	y	.	1	9.1	0.51962
4	y	.	2	11.0	0.51962
5	y	.	3	11.5	0.51962
6	y	.	4	16.4	0.51962
7	y	1	1	9.4	0.73485
8	y	1	2	12.0	0.73485
9	y	1	3	12.2	0.73485
10	y	1	4	14.8	0.73485
11	y	2	1	8.8	0.73485
12	y	2	2	10.0	0.73485
13	y	2	3	10.8	0.73485
14	y	2	4	18.0	0.73485

On notera que les valeurs nécessairement manquantes des colonnes f1 et f2 de ce fichier expliquent le message d’erreur (non fatale) obtenu systématiquement dans la fenêtre “log” de SAS au moment de la réalisation du graphique. Il ne faut donc pas tenir compte de ce message.

Le graphique obtenu est donné par la Figure A.1.

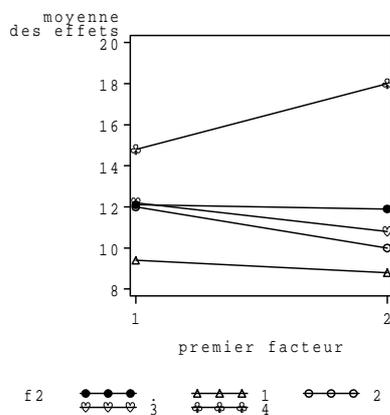


FIG. A.1 – Graphique des interactions

## A.4 Les données “traitements”

On a souhaité expérimenter 5 traitements sur 10 patients, mais il n’est pas possible d’administrer plus de 3 traitements distincts à un patient donné, quel qu’il soit. On a donc réalisé un plan en blocs, incomplet, équilibré, en administrant 3 traitements à chaque patient, les 10 patients constituant ainsi 10 blocs. Le fichier “traitmt.don” contient :

- en première colonne, le numéro du patient, donc du bloc (codé de 1 à 10);
- en deuxième colonne, le numéro du traitement (codé de 1 à 5);
- en troisième colonne, la mesure biologique réalisée à l’issue de l’expérience.

Voici le fichier “traitmt.don” :

```
1 1 52
1 2 50.5
1 3 59.5
2 2 46
2 3 52
2 5 52
3 1 45.5
3 3 56
3 4 54.5
4 2 47
4 4 50
4 5 50
5 1 48
5 3 53
5 5 52
6 1 49
6 2 54
6 5 65
7 2 45
7 3 51
7 4 51
8 1 55.5
8 2 60.5
8 4 61
9 3 28.5
9 4 28.5
9 5 30
10 1 44
10 4 52
10 5 57
```



## Annexe B

# Note sur les différents types de sommes de carrés

Le but de cette annexe est de préciser la signification des différents types de sommes de carrés (ainsi que les “philosophies” sous-jacentes) que l’on trouve dans la plupart des logiciels de statistique, en particulier SAS, dans le contexte de l’analyse de variance (ANOVA). Pour illustrer notre propos, nous nous placerons en ANOVA à deux facteurs croisés.

### B.1 Introduction

Considérons un modèle d’analyse de variance à deux facteurs croisés dans lequel :

- le premier facteur, noté  $F_1$ , possède  $J$  niveaux ( $J \geq 2$ ) qui seront indicés par  $j$  ;
- le second facteur, noté  $F_2$ , possède  $K$  niveaux ( $K \geq 2$ ) qui seront indicés par  $k$  ;
- au croisement du niveau  $j$  de  $F_1$  et du niveau  $k$  de  $F_2$ , on réalise  $n_{jk}$  observations ( $n_{jk} \geq 1$ ) d’une v.a.r.  $Y$  (le plan d’expérience est donc complet, pas nécessairement équilibré) ;
- chaque observation est notée  $y_{ijk}$  ( $i = 1, \dots, n_{jk}$  ;  $j = 1, \dots, J$  ;  $k = 1, \dots, K$ ) ;
- on pose :  $n_{j+} = \sum_{k=1}^K n_{jk}$  (effectif marginal du niveau  $j$  de  $F_1$ ) ;  $n_{+k} = \sum_{j=1}^J n_{jk}$  (effectif marginal du niveau  $k$  de  $F_2$ ) ;  $n = \sum_{j=1}^J \sum_{k=1}^K n_{jk}$  (nombre total d’observations).

Introduisons les différentes moyennes partielles empiriques des observations  $y_{ijk}$  :

$$\begin{aligned}\bar{y}_{\bullet jk} &= \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk} ; \\ \bar{y}_{\bullet j\bullet} &= \frac{1}{n_{j+}} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} y_{ijk} ; \\ \bar{y}_{\bullet\bullet k} &= \frac{1}{n_{+k}} \sum_{j=1}^J \sum_{i=1}^{n_{jk}} y_{ijk} ; \\ \bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} y_{ijk} .\end{aligned}$$

Considérons maintenant la somme des carrés totale du modèle, quantité à  $n - 1$  degrés de liberté :

$$SST = \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 .$$

Nous allons tout d’abord expliciter la décomposition de la quantité  $SST$ .

### B.2 Décomposition de la somme totale des carrés

Remarquons tout d’abord l’égalité suivante :

$$y_{ijk} - \bar{y}_{\bullet\bullet\bullet} = (y_{ijk} - \bar{y}_{\bullet jk}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet}) .$$

En élevant au carré, il vient :

$$\begin{aligned} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 &= (y_{ijk} - \bar{y}_{\bullet jk})^2 + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet})^2 \\ &+ (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{\ell=1}^6 DP_{\ell}(ijk), \end{aligned}$$

où les  $DP_{\ell}(ijk)$  ( $\ell = 1, \dots, 6$ ) représentent les doubles produits du développement de ce carré.

Par triple sommation, on obtient :

$$\begin{aligned} SST &= \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\bullet jk})^2 + \sum_{j=1}^J n_{j+} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{k=1}^K n_{+k} (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet})^2 \\ &+ \sum_{j=1}^J \sum_{k=1}^K n_{jk} (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{\ell=1}^6 \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} DP_{\ell}(ijk). \end{aligned}$$

Pour détailler les doubles produits, remarquons tout d'abord que si les quantités  $x_{jk}$  sont des réels indépendants de  $i$ , on peut écrire :

$$\sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} x_{jk} (y_{ijk} - \bar{y}_{\bullet jk}) = \sum_{j=1}^J \sum_{k=1}^K x_{jk} \left[ \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\bullet jk}) \right] = 0.$$

Il s'ensuit que les sommes des trois premiers doubles produits (ceux dans lesquels la quantité  $y_{ijk} - \bar{y}_{\bullet jk}$  est en facteur) sont nulles. Par contre, les trois autres ne sont, en général, pas nulles. La quatrième s'écrit :

$$\sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} DP_4(ijk) = 2 \sum_{j=1}^J \sum_{k=1}^K n_{jk} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet}) = SDP_1.$$

Les sommes des deux derniers doubles produits peuvent être regroupées dans l'expression suivante :

$$\begin{aligned} &\sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} [DP_5(ijk) + DP_6(ijk)] = SDP_2 \\ &= 2 \sum_{j=1}^J \sum_{k=1}^K n_{jk} (\bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet k} - 2\bar{y}_{\bullet\bullet\bullet}) (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet}). \end{aligned}$$

Enfin, un développement des parenthèses figurant dans les expressions  $SDP_1$  et  $SDP_2$  ci-dessus définies permet d'obtenir (les calculs sont simples, mais assez fastidieux) :

$$\begin{aligned} SDP_1 &= 2 \left( \sum_{j=1}^J \sum_{k=1}^K n_{jk} \bar{y}_{\bullet j\bullet} \bar{y}_{\bullet\bullet k} - n \bar{y}_{\bullet\bullet\bullet}^2 \right); \\ SDP_2 &= 4(n \bar{y}_{\bullet\bullet\bullet}^2 - \sum_{j=1}^J \sum_{k=1}^K n_{jk} \bar{y}_{\bullet j\bullet} \bar{y}_{\bullet\bullet k}). \end{aligned}$$

Finalement, la somme de tous les doubles produits figurant dans  $SST$  s'écrit :

$$SDP = 2(n \bar{y}_{\bullet\bullet\bullet}^2 - \sum_{j=1}^J \sum_{k=1}^K n_{jk} \bar{y}_{\bullet j\bullet} \bar{y}_{\bullet\bullet k}).$$

Explicitons maintenant les sommes de carrés en introduisant les quantités suivantes :

– somme des carrés due au facteur  $F_1$  (quantité à  $J - 1$  degrés de liberté) :

$$SSF_1 = \sum_{j=1}^J n_{j+} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 ;$$

– somme des carrés due au facteur  $F_2$  (quantité à  $K - 1$  degrés de liberté) :

$$SSF_2 = \sum_{k=1}^K n_{+k} (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet})^2 ;$$

- somme des carrés due aux interactions (quantité à  $(J - 1)(K - 1)$  degrés de liberté) :

$$SSF_{1*2} = \sum_{j=1}^J \sum_{k=1}^K n_{jk} (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet})^2 ;$$

- somme des carrés due aux erreurs (ou résiduelle ; quantité à  $n - JK$  degrés de liberté) :

$$SSE = \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\bullet jk})^2 .$$

On peut finalement réécrire la somme des carrés totale sous la forme :

$$SST = SSF_1 + SSF_2 + SSF_{1*2} + SSE + SDP.$$

**Remarque 79** Dans le cas particulier d'un plan équilibré ( $n_{jk} = n_0, \forall(j, k)$ ), on vérifie sans difficulté que  $SDP = 0$ . La décomposition de  $SST$  est alors d'interprétation évidente. Par contre, ce n'est pas le cas avec les plans déséquilibrés pour lesquels la quantité  $SDP$  est en général non nulle.

Lorsque la quantité  $SDP$  est non nulle, il n'est pas possible de l'affecter à une unique source de variation ( $F_1, F_2$  ou  $F_1 * F_2$ ). Ceci explique les difficultés rencontrées pour spécifier les sources de variation dans un modèle relatif à un plan déséquilibré. Pour cette raison, on a recours à d'autres raisonnements pour spécifier ces sources, ce qui explique l'existence de plusieurs types de sommes de carrés, selon la philosophie choisie.

### B.3 Exemple

On considère le jeu de données ci-dessous, dans lequel la variable réponse quantitative  $Y$  est expliquée par deux facteurs croisés,  $F_1$  à deux niveaux et  $F_2$  à trois niveaux. Il y a au total 18 observations dans un plan complet déséquilibré (il s'agit d'un exemple fictif, dans lequel les observations de  $Y$  ont été choisies pour faciliter les calculs "à la main", de façon à permettre un certain contrôle des résultats fournis par le logiciel SAS).

Outre les valeurs initiales des  $y_{ijk}$ , le tableau ci-dessous donne toutes les sommes et moyennes partielles (dans chaque cellule, chaque ligne, chaque colonne), ainsi que la somme et la moyenne globales.

	niveau 1 de $F_2$	niveau 2 de $F_2$	niveau 3 de $F_2$	sommes	moyennes
niveau 1 de $F_1$	10 14 18	36 40 44 48	82 86		
sommes	<b>42</b>	<b>168</b>	<b>168</b>	<b>378</b>	
moyennes	14	42	84		42
niveau 2 de $F_1$	22 26	24 28 32	60 64 68 72		
sommes	<b>48</b>	<b>84</b>	<b>264</b>	<b>396</b>	
moyennes	24	28	66		44
sommes	<b>90</b>	<b>252</b>	<b>432</b>	<b>774</b>	
moyennes	18	36	72		43

À partir du tableau ci-dessus, on peut calculer facilement les expressions suivantes :

$$SSF_1 = 18 ; \quad SSF_2 = 8514 ; \quad SSF_{1*2} = 1050 ; \quad SSE = 240 ; \quad SDP_1 = 180 ; \quad SDP_2 = -360.$$

On en déduit :  $SST = 9642$ . Dans le modèle "complet" (également appelé modèle "plein" et comportant un effet général, les effets de  $F_1$ , les effets de  $F_2$  et les effets d'interactions), la somme des carrés relative au modèle vaudra donc :  $9642 - 240 = 9402$ .

## B.4 Traitement des données avec SAS

### B.4.1 Traitement initial

Nous avons utilisé la procédure `GLM` du logiciel `SAS` pour traiter ces données selon un modèle d'analyse de variance à deux facteurs croisés, avec interactions. À la suite de la commande `model`, nous avons rajouté les options `ss1`, `ss2`, `ss3` et `ss4` pour obtenir les sommes de carrés de type I, de type II, de type III et de type IV (ces dernières uniquement dans le premier traitement, puisqu'elles ne se distinguent des précédentes que dans les plans incomplets).

En supposant les données contenues dans le fichier `deseq.don`, le programme SAS est le suivant :

```

data deseq;
infile 'deseq.don';
input f1 f2 y;
run;
proc glm data=deseq;
class f1 f2;
model y = f1 f2 f1*f2 / ss1 ss2 ss3 ss4;
run;
quit;

```

En voici les résultats.

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
f1	1	18.000000	18.000000	0.90	0.3615
f2	2	8801.112108	4400.556054	220.03	<.0001
f1*f2	2	582.887892	291.443946	14.57	0.0006

Source	DF	Type II SS	Mean Square	F Value	Pr > F
f1	1	305.112108	305.112108	15.26	0.0021
f2	2	8801.112108	4400.556054	220.03	<.0001
f1*f2	2	582.887892	291.443946	14.57	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	223.384615	223.384615	11.17	0.0059
f2	2	8664.968610	4332.484305	216.62	<.0001
f1*f2	2	582.887892	291.443946	14.57	0.0006

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
f1	1	223.384615	223.384615	11.17	0.0059
f2	2	8664.968610	4332.484305	216.62	<.0001
f1*f2	2	582.887892	291.443946	14.57	0.0006

On constate tout d'abord que le modèle est très significatif et que le coefficient  $R^2$  est très proche de 1 : on a donc un très bon ajustement du modèle aux données.

Ensuite, on voit que les sommes de carrés de type IV sont identiques à celles de type III : c'est normal, puisque les sommes de type IV ne sont différentes des sommes de type III que dans le cas

de certains plans incomplets. Nous ne les ferons donc plus figurer dans les traitements qui vont suivre, mais on pourra trouver des précisions sur les sommes de type IV dans Milliken & Johnson (1984) ou dans Azais (1994).

Nous allons maintenant détailler la façon d'obtenir les autres sommes de carrés figurant ci-dessus.

**Remarque 80** *Insistons encore ici sur le fait que, dans un plan complet équilibré, les quatre types de sommes sont toujours identiques.*

### B.4.2 Somme des carrés relative aux interactions

Le principe de calcul est très simple et assez naturel : on fait la différence entre la somme des carrés relative aux erreurs dans le modèle additif (sans interactions) et celle, également relative aux erreurs, dans le modèle complet (avec interactions). Le résultat obtenu représente donc la somme des carrés relative aux interactions. On constate que c'est la même, quel que soit le type de somme (ici : 582.89).

Contrôlons ce résultat en mettant en œuvre le modèle additif.

```
proc glm data=deseq;
class f1 f2;
model y = f1 f2 / ss1 ss2 ss3;
run;
quit;
```

On obtient les résultats suivants.

Dependent Variable: y

Model: y = f1 f2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8819.112108	2939.704036	50.01	<.0001
Error	14	822.887892	58.777707		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.914656	17.82945	7.666662	43.00000

On voit qu'en faisant la différence entre 822.89 et 240, on retrouve bien la quantité 582.89.

### B.4.3 Somme des carrés relative au facteur $F_2$

Le principe général reste le même : on fait la différence entre la somme des carrés relative aux erreurs dans un modèle où on a enlevé  $F_2$  et un modèle de référence, dans lequel il figure. Le problème est que le modèle de référence varie : c'est le modèle complet (effets de  $F_1$ , de  $F_2$  et des interactions) dans le cas des sommes de type III et c'est le modèle additif (effets de  $F_1$  et de  $F_2$  seulement) dans le cas des sommes de type I et de type II.

On a donc besoin des résultats de deux modèles : le modèle avec  $F_1$  et les interactions, et le modèle avec seulement  $F_1$ . Mettons ces deux modèles en œuvre.

```
proc glm data=deseq;
class f1 f2;
model y = f1 f1*f2 / ss1 ss2 ss3;
run;
proc glm data=deseq;
class f1 f2;
```

```

model y = f1 / ss1 ss2 ss3;
run;
quit;

```

Étudions en les résultats.

The GLM Procedure

Dependent Variable: y

Model: y = f1 f1\*f2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
f1	1	18.000000	18.000000	0.90	0.3615
f1*f2	4	9384.000000	2346.000000	117.30	<.0001

Source	DF	Type II SS	Mean Square	F Value	Pr > F
f1	1	18.000000	18.000000	0.90	0.3615
f1*f2	4	9384.000000	2346.000000	117.30	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	223.384615	223.384615	11.17	0.0059
f1*f2	4	9384.000000	2346.000000	117.30	<.0001

Dependent Variable: y

Model: y = f1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18.000000	18.000000	0.03	0.8648
Error	16	9624.000000	601.500000		
Corrected Total	17	9642.000000			

Pour les sommes de type I et de type II, on obtient :  $9624 - 822.89 = 8801.11$  : on retrouve bien le résultat de la première sortie.

Par contre, on ne peut pas retrouver ici la somme des carrés de type III pour  $F_2$ , puisque la somme des carrés relative aux erreurs dans le modèle avec  $F_1$  et les interactions reste la même que dans le modèle complet : 240. C'est **un des paradoxes de SAS**, qui permet de déclarer un modèle avec un seul facteur et les interactions, mais ne le traite pas comme tel, puisqu'il rajoute l'effet du

facteur enlevé,  $F_2$ , dans les effets d'interactions : ceux-ci se trouvent ainsi avoir 4 degrés de liberté (2 pour  $F_2$  et 2 pour les interactions) et une somme de carrés égale à 9384, que l'on obtient en additionnant 8801.11 et 582.89, sommes relatives respectivement à  $F_2$  et aux interactions dans le modèle complet, si l'on considère les sommes de type I ou II. Il faudra donc avoir recours à un artifice pour retrouver la somme de type III relative à  $F_2$  (voir le point B.4.5).

#### B.4.4 Somme des carrés relative au facteur $F_1$

Tout d'abord, remarquons que ces sommes sont toutes différentes, selon le type I, II ou III considéré. Cela provient de ce que les philosophies sont ici toutes trois différentes.

##### Type III

Le type III conserve la même philosophie : le modèle de référence étant le modèle complet, on calcule la différence entre la somme des carrés relatives aux erreurs dans le modèle avec les effets de  $F_2$  et ceux des interactions et la même somme dans le modèle complet. On se heurte encore à la même difficulté : il n'est pas possible d'obtenir directement la première somme des carrés. Nous aurons donc recours, là encore, au même artifice que précédemment (voir encore le point B.4.5).

##### Type II

Pour les sommes de type II, on doit faire la différence entre la somme des carrés relative aux erreurs dans le modèle avec les seuls effets de  $F_2$  et la même somme dans le modèle additif. Mettons en œuvre le modèle avec le seul facteur  $F_2$ .

```
proc glm data=deseq;
class f1 f2;
model y = f2 / ss1 ss2 ss3;
run;
```

En voici les résultats.

##### The GLM Procedure

Dependent Variable: y

Model: y = f2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8514.000000	4257.000000	56.61	<.0001
Error	15	1128.000000	75.200000		
Corrected Total	17	9642.000000			

On obtient  $1128 - 822.89 = 305.11$ , qui est, dans le modèle complet, la somme des carrés de type II relative à  $F_1$ .

##### Type I

Pour le type I, il convient de préciser la philosophie globale de cette approche. Elle suppose en effet que les effets déclarés dans la commande `model y = f1 f2 f1*f2` sont ordonnés. Autrement dit, les premiers effets à prendre en compte sont ceux de  $F_1$ , puis ceux de  $F_2$ , enfin ceux des interactions. Par conséquent, pour calculer la somme des carrés relative à l'un de ces trois effets, on considère la différence des sommes des carrés relatives aux erreurs dans deux modèles : "le plus grand modèle" ne contenant pas cet effet (ici, le modèle constant) et "le plus petit modèle" le contenant (ici, le modèle avec seulement  $F_1$ ). Dans le modèle constant, la somme des carrés relative aux erreurs est la somme des carrés totale à savoir 9642. Le modèle ne comportant que  $F_1$  a déjà

été étudié en B.4.3 (somme des carrés relative aux erreurs dans ce modèle : 9624). D'où la somme des carrés relative à  $F_1$  pour le type I :  $9642 - 9624 = 18$ . On peut remarquer qu'il s'agit en fait de la somme des carrés relative à  $F_1$  telle que nous l'avons définie au paragraphe B.2 ( $SSF_1$ ) et dont la valeur a été donnée au paragraphe B.3.

### Autre illustration de la philosophie de type I

De façon à bien comprendre la philosophie des sommes de type I, déclarons maintenant le modèle complet en commençant par  $F_2$  :

```
proc glm data=deseq;
class f1 f2;
model y = f2 f1 f1*f2 / ss1 ss2 ss3;
run;
```

En voici les résultats.

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

	R-Square	Coeff Var	Root MSE	y Mean
	0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
f2	2	8514.000000	4257.000000	212.85	<.0001
f1	1	305.112108	305.112108	15.26	0.0021
f1*f2	2	582.887892	291.443946	14.57	0.0006

Source	DF	Type II SS	Mean Square	F Value	Pr > F
f2	2	8801.112108	4400.556054	220.03	<.0001
f1	1	305.112108	305.112108	15.26	0.0021
f1*f2	2	582.887892	291.443946	14.57	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f2	2	8664.968610	4332.484305	216.62	<.0001
f1	1	223.384615	223.384615	11.17	0.0059
f1*f2	2	582.887892	291.443946	14.57	0.0006

Mise à part l'inversion de l'ordre des facteurs  $F_1$  et  $F_2$ , les sommes de type II et de type III sont identiques à ce qu'elles étaient dans le modèle initial. Par contre, il n'en va pas de même pour les sommes de type I : seule celle relative aux interactions est inchangée ; celle relative à  $F_1$  a augmenté (elle est passée de 18 à 305.11), tandis que celle relative à  $F_2$  a baissé (elle est passée de 8801.11 à 8514). Remarquons en passant que la somme relative à  $F_2$  (8514) est maintenant égale à celle définie au paragraphe B.2 ( $SSF_2$ ) et donnée au paragraphe B.3.

**Remarque 81** En additionnant les sommes des carrés relatives aux différents effets dans le type I, on retrouve, quel que soit l'ordre de déclaration des facteurs, la somme des carrés relative au modèle complet, à savoir 9402. Ceci est une propriété générale, et seules les sommes de type I possèdent cette propriété, comme on peut le constater en faisant les additions : pour le type II, on trouve 9689.11 ; pour le type III, 9471.24.

### B.4.5 Retour sur les sommes de type III

Revenons maintenant aux sommes de type III et essayons de retrouver les sommes de carrés relatives à chacun des deux facteurs  $F_1$  et  $F_2$ . Pour cela, il faut passer par l'intermédiaire d'un modèle de régression sur indicatrices, mais pas n'importe quelles indicatrices !

#### Introduction d'indicatrices

Nous allons introduire les indicatrices des niveaux de  $F_1$ , celles des niveaux de  $F_2$ , et celles des cellules obtenues par croisement de  $F_1$  et de  $F_2$ , comme cela se fait dans le cadre du *paramétrage* dit *centré* de l'ANOVA (paramétrage associé à un effet moyen général, des effets principaux pour chaque niveau de chaque facteur, leur somme étant nulle, et des effets d'interactions doublement centrés).

Pour  $F_1$ , c'est en fait la différence entre l'indicatrice du niveau 1 et celle du niveau 2 qui doit intervenir (une seule variable car un seul degré de liberté) ; nous la noterons  $L_1$ . Pour  $F_2$ , on doit utiliser deux variables (deux degrés de liberté) : la différence entre l'indicatrice du niveau 1 et celle du niveau 3 et la différence entre l'indicatrice du niveau 2 et celle du niveau 3 ; nous les noterons respectivement  $C_1$  et  $C_2$ . Enfin, pour les cellules, c'est le produit des précédentes qui seront utilisés (il n'y en a que deux) :  $LC_1 = L_1 \times C_1$  ;  $LC_2 = L_1 \times C_2$ .

Voici un programme SAS permettant de créer ces différences d'indicatrices :

```
data indil;
set deseq;
L1 = 0;
if f1 = 1 then L1 = 1;
if f1 = 2 then L1 = -1;
C1 = 0;
if f2 = 1 then C1 = 1;
if f2 = 3 then C1 = -1;
C2 = 0;
if f2 = 2 then C2 = 1;
if f2 = 3 then C2 = -1;
LC1 = L1 * C1;
LC2 = L1 * C2;
run;
```

En voici les résultats :

Obs	f1	f2	y	L1	C1	C2	LC1	LC2
1	1	1	10	1	1	0	1	0
2	1	1	14	1	1	0	1	0
3	1	1	18	1	1	0	1	0
4	1	2	36	1	0	1	0	1
5	1	2	40	1	0	1	0	1
6	1	2	44	1	0	1	0	1
7	1	2	48	1	0	1	0	1
8	1	3	82	1	-1	-1	-1	-1
9	1	3	86	1	-1	-1	-1	-1
10	2	1	22	-1	1	0	-1	0
11	2	1	26	-1	1	0	-1	0
12	2	2	24	-1	0	1	0	-1
13	2	2	28	-1	0	1	0	-1
14	2	2	32	-1	0	1	0	-1

15	2	3	60	-1	-1	-1	1	1
16	2	3	64	-1	-1	-1	1	1
17	2	3	68	-1	-1	-1	1	1
18	2	3	72	-1	-1	-1	1	1

### Régression sur les indicatrices

Faisons maintenant la régression de  $Y$  sur l'ensemble de ces indicatrices :

```
proc glm data=indii;
model y = L1 C1 C2 LC1 LC2 / ss1 ss2 ss3;
run;
```

Les résultats sont les suivants :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
L1	1	18.000000	18.000000	0.90	0.3615
C1	1	8388.765957	8388.765957	419.44	<.0001
C2	1	412.346150	412.346150	20.62	0.0007
LC1	1	480.760233	480.760233	24.04	0.0004
LC2	1	102.127660	102.127660	5.11	0.0432

Source	DF	Type II SS	Mean Square	F Value	Pr > F
L1	1	223.384615	223.384615	11.17	0.0059
C1	1	4443.428571	4443.428571	222.17	<.0001
C2	1	588.255319	588.255319	29.41	0.0002
LC1	1	579.428571	579.428571	28.97	0.0002
LC2	1	102.127660	102.127660	5.11	0.0432

Source	DF	Type III SS	Mean Square	F Value	Pr > F
L1	1	223.384615	223.384615	11.17	0.0059
C1	1	4443.428571	4443.428571	222.17	<.0001
C2	1	588.255319	588.255319	29.41	0.0002
LC1	1	579.428571	579.428571	28.97	0.0002
LC2	1	102.127660	102.127660	5.11	0.0432

On remarquera tout d'abord que l'on obtient exactement les mêmes résultats généraux qu'avec le modèle complet d'ANOVA (sommés des carrés relatives au modèle et aux erreurs, degrés de liberté, coefficient  $R^2$ ...), ce qui est logique.

On retrouve également des résultats identiques pour les sommes de carrés de type I : 18 pour  $L_1$  (donc pour  $F_1$ ) ;  $8388.77 + 412.35 = 8801.12$  pour  $C_1 + C_2$  (donc pour  $F_2$ ) ;  $480.76 + 102.13 = 582.89$  pour  $LC_1 + LC_2$  (donc pour les interactions).

Par contre, il n'en va pas de même pour les sommes de type II et de type III. Tout d'abord, on remarque qu'elles sont maintenant identiques, ce qui s'explique par le fait que le seul modèle par rapport auquel on peut, dans ce cadre, se référer est le modèle complet (la notion de modèle additif n'a plus de sens dans le cadre d'une régression). Ensuite, ces sommes s'obtiennent toujours en faisant la différence des sommes de carrés relatives aux erreurs au sein de deux modèles : le modèle dans lequel on enlève seulement l'effet considéré ( $L_1, C_1 \dots$ ) et le modèle complet considéré ci-dessus. Nous laissons le soin au lecteur de vérifier ce résultat.

### Somme des carrés de type III relative au facteur $F_2$

Refaisons maintenant la régression de  $Y$  sur les seules indicatrices  $L_1, LC_1$  et  $LC_2$ , autrement dit sur le facteur  $F_1$  et sur les interactions.

```
proc glm data=indi1;
model y = L1 LC1 LC2 / ss1 ss2 ss3;
run;
```

En voici les résultats :

The GLM Procedure					
Dependent Variable: y					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	737.031390	245.677130	0.39	0.7646
Error	14	8904.968610	636.069186		
Corrected Total	17	9642.000000			
	R-Square	Coeff Var	Root MSE	y Mean	
	0.076440	58.65212	25.22041	43.00000	

En faisant la différence  $8904.97 - 240 = 8664.97$ , on retrouve maintenant la somme des carrés de type III relative à  $F_2$  dans le modèle complet d'ANOVA.

### Somme des carrés de type III relative au facteur $F_1$

Dans la même optique, faisons maintenant la régression de  $Y$  sur les indicatrices  $C_1, C_2, LC_1$  et  $LC_2$  (autrement dit, sur  $F_2$  et sur les interactions).

```
proc glm data=indi1;
model y = C1 C2 LC1 LC2 / ss1 ss2 ss3;
run;
```

On obtient :

## The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9178.615385	2294.653846	64.38	<.0001
Error	13	463.384615	35.644970		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.951941	13.88451	5.970341	43.00000

La différence  $463.38 - 240 = 223.38$  redonne la somme des carrés de type III relative à  $F_1$ .

**Encore un paradoxe de SAS !**

Les indicatrices utilisées ci-dessus, ou plutôt les différences d'indicatrices (indicatrice d'un niveau moins indicatrice du dernier niveau), apparaissent naturellement dans le paramétrage centré du modèle d'ANOVA.

On peut se demander ce qu'il se passe si on remplace ces indicatrices par celles qui apparaissent naturellement dans le *paramétrage* du modèle d'ANOVA *réalisé par SAS*. Il s'agit simplement des indicatrice des niveaux des facteurs, à l'exception du dernier, et des produits de ces indicatrices pour les interactions.

Voici encore un programme permettant d'obtenir ces indicatrices :

```
data indi2;
set deseq;
LS1 = 0;
if f1 = 1 then LS1 = 1;
CS1 = 0;
if f2 = 1 then CS1 = 1;
CS2 = 0;
if f2 = 2 then CS2 = 1;
LCS1 = LS1 * CS1;
LCS2 = LS1 * CS2;
run;
```

Et voici la table SAS obtenue :

Obs	f1	f2	y	LS1	CS1	CS2	LCS1	LCS2
1	1	1	10	1	1	0	1	0
2	1	1	14	1	1	0	1	0
3	1	1	18	1	1	0	1	0
4	1	2	36	1	0	1	0	1
5	1	2	40	1	0	1	0	1
6	1	2	44	1	0	1	0	1
7	1	2	48	1	0	1	0	1
8	1	3	82	1	0	0	0	0
9	1	3	86	1	0	0	0	0
10	2	1	22	0	1	0	0	0
11	2	1	26	0	1	0	0	0
12	2	2	24	0	0	1	0	0
13	2	2	28	0	0	1	0	0
14	2	2	32	0	0	1	0	0
15	2	3	60	0	0	0	0	0

16	2	3	64	0	0	0	0	0
17	2	3	68	0	0	0	0	0
18	2	3	72	0	0	0	0	0

Faisons maintenant la régression de  $Y$  sur les cinq indicatrices ci-dessus.

```
proc glm data=indi2;
model y = LS1 CS1 CS2 LCS1 LCS2 / ss1 ss2 ss3;
run;
```

En voici les résultats :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LS1	1	18.000000	18.000000	0.90	0.3615
CS1	1	4324.500000	4324.500000	216.22	<.0001
CS2	1	4476.612108	4476.612108	223.83	<.0001
LCS1	1	570.887892	570.887892	28.54	0.0002
LCS2	1	12.000000	12.000000	0.60	0.4536

Source	DF	Type II SS	Mean Square	F Value	Pr > F
LS1	1	432.000000	432.000000	21.60	0.0006
CS1	1	2352.000000	2352.000000	117.60	<.0001
CS2	1	2475.428571	2475.428571	123.77	<.0001
LCS1	1	495.157895	495.157895	24.76	0.0003
LCS2	1	12.000000	12.000000	0.60	0.4536

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LS1	1	432.000000	432.000000	21.60	0.0006
CS1	1	2352.000000	2352.000000	117.60	<.0001
CS2	1	2475.428571	2475.428571	123.77	<.0001
LCS1	1	495.157895	495.157895	24.76	0.0003
LCS2	1	12.000000	12.000000	0.60	0.4536

Encore une fois, les résultats généraux n'ont pas changé. De même, les sommes de carrés de type I permettent de retrouver celles du modèle complet en ANOVA : 18 pour  $LS_1$ , c'est-à-dire pour  $F_1$  ;  $4324.50 + 4476.61 = 8801.11$  pour  $CS_1 + CS_2$ , c'est-à-dire pour  $F_2$  ;  $570.89 + 12.00 = 582.89$  pour  $LCS_1 + LCS_2$ , c'est-à-dire pour les interactions.

Par contre, les sommes de carrés de type II et de type III, encore une fois égales, ne redonnent pas les résultats de l'ANOVA avec le modèle complet et sont différentes de ce qu'elles étaient avec la première série d'indicateurs.

Intéressons nous maintenant au modèle de régression de  $Y$  sur les indicateurs  $LS_1$ ,  $LCS_1$  et  $LCS_2$ .

```
proc glm data=indi2;
model y = LS1 LCS1 LCS2 / ss1 ss2 ss3;
run;
```

Voici les résultats :

The GLM Procedure					
Dependent Variable: y					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5898.000000	1966.000000	7.35	0.0034
Error	14	3744.000000	267.428571		
Corrected Total	17	9642.000000			
	R-Square	Coeff Var	Root MSE	y Mean	
	0.611699	38.03080	16.35324	43.00000	

On pourra vérifier que la somme des carrés relative aux erreurs dans ce modèle (3744) ne permet pas de retrouver la somme des carrés de type III relative au facteur  $F_2$  dans le modèle complet d'ANOVA. Ainsi, l'usage d'indicateurs permet de retrouver ces sommes, mais uniquement les indicateurs utilisés dans le paramétrage centré, celles utilisées dans le paramétrage SAS ne le permettant pas.

#### B.4.6 Cas particulier du modèle additif

Revenons sur le modèle additif, déjà traité en B.4.2. Voici les sommes de carrés que l'on obtient dans ce modèle :

Source	DF	Type I SS	Mean Square	F Value	Pr > F
f1	1	18.000000	18.000000	0.31	0.5887
f2	2	8801.112108	4400.556054	74.87	<.0001
Source	DF	Type II SS	Mean Square	F Value	Pr > F
f1	1	305.112108	305.112108	5.19	0.0389
f2	2	8801.112108	4400.556054	74.87	<.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	305.112108	305.112108	5.19	0.0389
f2	2	8801.112108	4400.556054	74.87	<.0001

On remarque que les sommes de carrés de type II et de type III sont égales dans ce cas. Ce résultat est général et s'explique par le fait que le modèle de référence est ici le modèle additif, que ce soit pour le type II ou pour le type III.

## B.5 Quelle philosophie suivre ?

À l'issue de cette étude, on peut légitimement se poser la question : quelles sommes de carrés utiliser ? On se doute que la réponse n'est pas univoque et qu'elle est liée à la fois au type de données dont on dispose et à la philosophie que l'on souhaite suivre.

Tout d'abord, nous laisserons de côté les sommes de type IV qui ne concernent que les plans incomplets et déséquilibrés que nous n'avons pas envisagés ici (encore faut-il signaler que le type IV est parfois critiqué dans le contexte des plans incomplets déséquilibrés).

Ensuite, les sommes de type I sont spécifiques des modèles dans lesquels il existe un ordre naturel entre les facteurs. Pour des données de ce type, ce sont bien sûr ces sommes qu'il faut considérer. Dans les autres cas, il n'est pas courant de les utiliser (même si elles ont de bonnes propriétés, comme on l'a signalé).

**Remarque 82** *Il convient de ne pas confondre ce qu'on a appelé ici "ordre" entre les facteurs (on considère que l'un est plus important que l'autre, les interactions ayant nécessairement un moindre niveau d'importance) et ce qu'on appelle habituellement facteur hiérarchisé (la définition des niveaux du facteur hiérarchisé dépend du niveau de l'autre facteur dans lequel on se trouve ; de tels facteurs sont aussi ordonnés, mais de façon plus "structurelle"). Dans la procédure GLM de SAS, il est possible de faire un traitement spécifique pour des facteurs dont l'un est hiérarchisé à l'autre. C'est d'ailleurs dans ce contexte que les sommes de type I prennent tout leur sens.*

Reste donc le choix entre les sommes de type II et de type III pour les cas standards, mais déséquilibrés. Il est à noter que ce choix ne se pose que dans le cadre des modèles avec interactions, les deux types étant équivalents pour les modèles additifs. D'une façon générale, il est préconisé d'utiliser les sommes de type III de préférence à celles de type II. En particulier, on remarquera que SAS ne fournit par défaut que les sommes de type I et de type III.

Terminons cette discussion par la remarque ci-dessous dans laquelle on va préciser un peu plus les choses.

**Remarque 83** *La discussion sur le choix des sommes de carrés à utiliser dans la pratique est l'occasion de revenir sur la pratique des tests relatifs aux différents effets dans un modèle complexe comme une ANOVA à au moins deux facteurs. Considérons encore, pour simplifier, une ANOVA à deux facteurs croisés.*

*On peut préconiser la démarche consistant à tester en premier lieu les interactions, puis à passer au modèle additif si elles ne sont pas significatives. Cette démarche, assez naturelle, n'est pas la seule utilisée dans la pratique statistique. De plus, elle a le défaut suivant : elle conduit, lors des tests des effets principaux de chacun des deux facteurs, à prendre en compte dans le numérateur de l'estimateur de la variance (donc dans le dénominateur de la statistique de Fisher) les sommes de carrés, certes faibles mais non nulles, relatives aux interactions. Cela peut conduire à un biais dans la statistique de Fisher, donc dans la décision relative aux effets principaux.*

*D'où une autre démarche, tout aussi courante, qui consiste à tester chaque facteur au sein du modèle complet (avec interactions) et qu'on appelle souvent "non pooling", autrement dit non regroupement (des sommes de carrés des différents effets dans le numérateur de l'estimateur de la variance). Dans ce contexte, les sommes de type II ne sont pas justifiées (en fait, il n'y a pas de contexte dans lequel elles soient réellement justifiées).*

*Dans la pratique, on peut envisager de mener en parallèle les deux démarches ci-dessus. Lorsqu'elles conduisent à la même décision, il n'y a pas de problème. En cas de décisions contradictoires, il convient d'être très prudent et d'étudier en détails les deux modèles en présence, en particulier en utilisant les critères de choix de modèle (voir l'Annexe D).*

En guise de conclusion générale, **nous préconisons d'utiliser systématiquement les sommes de type III**. S'il existe un ordre entre les facteurs, notamment dans le cas de facteurs hiérarchisés, on devra aussi considérer les sommes de type I, voire les privilégier en cas de contradiction dans les décisions issues des tests. Si on est en présence d'un plan incomplet et déséquilibré, on devra considérer, en plus des sommes de type III, les sommes de type IV. Enfin, les sommes de type II sont déconseillées dans tous les cas.



## Annexe C

# Un exercice sur les carrés latins

*On propose, dans cette annexe, un exercice sur les carrés latins sous forme de jeu mathématique.*

Le quotidien “Le Monde” publie chaque semaine différents jeux dont un jeu mathématique intitulé *affaire de logique*. Nous reproduisons, à la page suivante, le problème numéro 533, paru dans l’édition datée du 22 mai 2007, ainsi que sa solution parue une semaine plus tard. Ce jeu est, en fait, un exercice intéressant sur les carrés latins.

**AFFAIRE DE LOGIQUE N° 533**

**CARRÉS LATINS SYMÉTRIQUES**

Dans un carré latin de 9 cases sur 9, chacun des chiffres de 1 à 9 figure une fois et une seule sur chaque ligne et sur chaque colonne.

Mais on peut encore y ajouter des conditions supplémentaires.

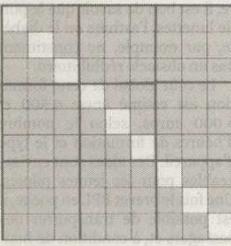
- Ainsi, dans un Sudo-ku, ces mêmes chiffres figurent une fois et une seule dans chacun des 9 carrés de trois cases sur trois de la grille.
- Dans un carré latin diagonal, les mêmes chiffres figurent une fois et une seule sur la diagonale principale (en blanc).
- Dans un carré latin symétrique, les chiffres occupent des positions symétriques par rapport à la diagonale principale.

Un carré latin symétrique  $9 \times 9$  est-il diagonal parfois, toujours, jamais ?

Un carré latin symétrique  $8 \times 8$  est-il diagonal parfois, toujours, jamais ?

ELISABETH BUSSE  
ET GILLES COHEN  
© POLE 2007

**Solution dans *Le Monde* du 29 mai.**



**Solution du jeu n° 533 paru dans *Le Monde* du 22 mai.**

Un carré latin symétrique  $9 \times 9$  est toujours diagonal, un  $8 \times 8$  ne l'est jamais.

Appelons  $d$  le nombre de lignes (et donc de colonnes) de la grille.

On constate aisément, grâce à la symétrie, que, dans un carré latin symétrique, un même chiffre apparaît sur les cases grises autant de fois au-dessus de la diagonale qu'en dessous, donc un nombre pair de fois.

Or chaque chiffre est présent une fois par ligne, donc au total  $d$  fois.

- Si  $d$  est impair, chacun des  $d$  chiffres, présent un nombre pair de fois dans les cases grises, apparaît donc un nombre impair de fois sur la diagonale. Et ce ne peut être qu'une fois, puisqu'il y a  $d$  chiffres à distribuer entre  $d$  cases.
- Si  $d$  est pair, chaque chiffre, présent un nombre pair de fois dans les cases grises, apparaît donc un nombre pair de fois sur la diagonale. Un carré latin symétrique pair ne peut donc être diagonal.

Tout carré latin symétrique impair est donc diagonal.

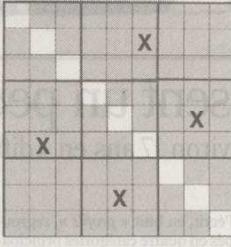


FIG. C.1 – Extraits du Monde des 22/05 et 29/05 2007.

## Annexe D

# Indications sur les critères de choix de modèle

*Les critères de choix de modèle sont assez utilisés dans la pratique statistique et permettent de choisir entre plusieurs modèles concurrents lors de la modélisation d'un jeu de données. Nous présentons ici les quatre principaux.*

La littérature statistique fournit de nombreux critères de choix de modèle. La plupart d'entre eux ont été définis dans le cadre du modèle linéaire gaussien et, plus particulièrement, dans celui de la régression linéaire. Toutefois, leur champ d'application est souvent plus large. Leur principe général est de minimiser une somme de deux termes :

- le premier terme est d'autant plus petit que le modèle s'ajuste bien aux données ; à la limite, le modèle saturé serait le meilleur au sens de ce seul premier terme (rappelons qu'un modèle saturé comporte autant de paramètres que d'observations et s'ajuste ainsi parfaitement aux données) ; toutefois, sauf cas particulier, un tel modèle n'a aucun intérêt pratique ;
- le second terme est fonction du nombre total de paramètres du modèle et pénalise ainsi les modèles surajustés aux données (comme le modèle saturé) ; on l'appelle le terme de pénalité ; à la limite, c'est au contraire le modèle constant qui serait le meilleur au sens de ce seul second critère.

Ces critères proposent donc un équilibre entre le surajustement (entraînant un faible biais du modèle, par rapport aux données étudiées, mais une forte variance, par rapport à son application à d'autres données), et la simplicité (entraînant une faible variance, mais un fort biais). Parmi tous les modèles possibles, celui qui minimise le critère choisi est celui qui réalise le meilleur équilibre entre les deux objectifs ci-dessus.

Le cheminement théorique conduisant à ces critères est en général assez complexe et leur formule finale est obtenue après diverses approximations et simplifications (pour plus de détails, nous renvoyons à l'ouvrage de J.M. Azaïs & J.M. Bardet, 2005). Nous donnons ci-dessous les quatre principaux critères sous une forme simplifiée, telle qu'on les trouve, par exemple, dans le logiciel SAS.

### D.1 Le $C_p$ de Mallows

Il a été défini dans Mallows (1973), dans le contexte de la régression. Il s'applique à tout modèle linéaire gaussien.

Pour un jeu de données de taille  $n$  et comportant au total  $q$  variables explicatives (ou régresseurs, ou variables indépendantes), le coefficient  $C_p$  associé à un modèle comportant seulement  $p$  régresseurs parmi les  $q$  ( $1 \leq p \leq q$ ) est défini par :

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2p - n = (n - q) \frac{SSE_p}{SSE} + 2p - n.$$

Dans ces expressions,  $SSE$  désigne la somme des carrés relative aux erreurs dans le modèle complet

(avec les  $q$  régresseurs),  $SSE_p$  la somme analogue dans la modèle avec seulement  $p$  régresseurs et  $\hat{\sigma}^2$  l'estimation de la variance de la loi gaussienne dans le modèle complet. On a donc :  $\hat{\sigma}^2 = \frac{SSE}{n - q}$ . Ainsi, plus on met de régresseurs dans le modèle, plus son erreur ( $SSE_p$ ) diminue (donc plus diminue le terme d'ajustement aux données  $\frac{SSE_p}{\hat{\sigma}^2}$ ), mais plus le terme de pénalité ( $2p$ ) augmente. On notera que le terme  $-n$  n'a aucune influence sur le modèle choisi et pourrait être supprimé. On notera encore que, dans le modèle complet,  $C_p = C_q = q$ .

Dans une analyse de variance,  $p$  doit être remplacé par le nombre de paramètres indépendants du modèle considéré (en dehors de la variance  $\sigma^2$ ).

## D.2 La déviance relative

Ce critère est spécifique au modèle linéaire généralisé. Pour un jeu de données fixé, soit  $M$  le modèle considéré et  $D(M)$  sa déviance :

$$D(M) = 2[\log L(M_s) - \log L(M)].$$

Dans l'expression ci-dessus,  $\log$  désigne le logarithme népérien,  $L$  la vraisemblance et  $M_s$  le modèle saturé. Par ailleurs, notons  $d_M$  le degré de liberté de l'erreur du modèle  $M$  (autrement dit, de sa déviance). On appelle alors déviance relative du modèle  $M$  le rapport  $\frac{D(M)}{d_M}$ .

Ici, le critère n'est pas une somme (entre un terme d'ajustement et un terme de pénalité) mais un rapport (entre un terme d'erreur et son degré de liberté). Dans le cadre du modèle linéaire généralisé, la minimisation de la déviance relative est un critère très pertinent de choix de modèle.

## D.3 Le critère A.I.C.

A.I.C. signifie *Akaike Information Criterion*. Ce critère a été défini dans Akaike (1974), et peut s'écrire sous la forme suivante :

$$AIC = -2\log(L) + 2k.$$

Dans cette expression,  $\log$  désigne toujours le logarithme népérien et  $L$  la vraisemblance du modèle considéré, tandis que  $k$  désigne le nombre de paramètres indépendants dans ce modèle.

Plus le modèle est complexe, pour bien s'ajuster aux données, plus  $L$  et  $\log(L)$  sont grands, et donc plus le premier terme est petit ; mais, dans le même temps, plus le second est grand.

Introduit dans le cadre du modèle linéaire gaussien, ce critère peut s'appliquer dans un cadre plus général (et il en est de même pour le critère B.I.C.).

## D.4 Le critère B.I.C.

B.I.C. signifie *Bayesian Information Criterion*, ce critère ayant été défini dans Schwarz (1978), à partir d'une approche bayésienne. Il peut s'écrire sous la forme suivante :

$$BIC = -2\log(L) + k\log(n).$$

Ici,  $n$  désigne le nombre d'observations considérées. On rencontre aussi ce critère sous le nom de *SC* (Schwarz Criterion).

**Remarque 84** On notera qu'on rencontre parfois, dans la littérature statistique, d'autres expressions pour ces critères. Elles sont, bien sûr, équivalentes lorsqu'on compare différents modèles sur le même jeu de données (elles conduisent au choix du même modèle). Les expressions données ci-dessus ont l'avantage d'être simples à retenir et d'être celles figurant dans le logiciel SAS.

**Remarque 85** Lorsqu'il s'agit de choisir un modèle dans un cadre très général (hors du modèle linéaire gaussien ou du modèle linéaire généralisé), si les deux critères A.I.C. et B.I.C. sélectionnent le même modèle, c'est clairement celui qu'il faut choisir. En cas de contradiction, les choses sont assez délicates.

Nous recommandons alors d'utiliser la minimisation du  $C_p$  de Mallows si l'on est dans le modèle linéaire gaussien et celle de la déviance relative si l'on est dans le modèle linéaire généralisé.

Dans un cadre plus général, il est très difficile de trancher. Notons toutefois que, dès que  $\log(n) > 2$  (autrement dit, dès que  $n \geq 8$ ), la pénalité est plus importante dans le critère B.I.C., ce qui conduit à choisir, selon ce critère, un modèle plus parcimonieux qu'avec le critère A.I.C. Pour cette raison, on peut préférer le modèle sélectionné par le critère A.I.C. dans une optique descriptive (pour la description des données étudiées, on privilégie la minimisation du biais) et celui sélectionné par le critère B.I.C. dans une optique prévisionnelle (pour la prédiction de  $Y$  sur des données sur lesquelles elle n'est pas observée, on privilégie la minimisation de la variance). Mais, il ne s'agit là que d'indications générales.

**Remarque 86** Il convient d'être prudent dans l'utilisation pratique des critères A.I.C. et B.I.C. En particulier, dans la procédure MIXED de SAS,  $L$  désigne la vraisemblance si l'on utilise le maximum de vraisemblance pour estimer les paramètres et la vraisemblance restreinte si l'on utilise le maximum de cette dernière (méthode REML). Par ailleurs, toujours dans la procédure MIXED de SAS avec la méthode REML, les deux critères ci-dessus sont, en fait, définis de la façon suivante :

$$AIC = -2\log(REML) + 2k',$$

où  $k'$  désigne le nombre de paramètres indépendants uniquement dans la structure de covariance du modèle (cela peut se comprendre, mais peut aussi conduire à des confusions) ;

$$BIC = -2\log(REML) + k' \log(m),$$

où  $m$  désigne le nombre de niveaux du facteur à effets aléatoires (ce qui est plus difficile à comprendre !).

Enfin, dans les modèles pour données répétées sans facteur à effets aléatoire, SAS prend pour  $n$  le nombre de sujets (qui n'est pas le nombre total d'observations), ce qui est normal.

**Remarque 87** Notons encore que la procédure MIXED de SAS fournit systématiquement un autre critère de choix de modèle noté A.I.C.C. (pour A.I.C. corrected). Nous déconseillons d'utiliser ce critère dont l'expression (y compris dans la documentation en ligne de SAS) n'est pas très claire.

**Remarque 88** En guise de conclusion, signalons que, dans la pratique, les critères présentés ici sont souvent utilisés non pas pour choisir un modèle parmi tous les modèles possibles (ce qui, d'un point de vue numérique, devient inapplicable dès que le nombre de variables explicatives est élevé), mais pour choisir entre deux, trois ou quatre modèles concurrents, après sélection au moyen des tests (par exemple, dans une démarche de type backward, forward ou stepwise). C'est cet usage, sélectif, des critères de choix de modèle que nous préconisons.



## Annexe E

# Tests multidimensionnels pour données répétées

*L'objet de cette annexe est de détailler les tests multidimensionnels réalisés par la procédure GLM de SAS dans le traitement de données répétées. Nous allons illustrer tout ce qui suit au moyen d'un exemple (fictif) de données répétées.*

### E.1 Les données

Voici ces données :

```
1 10 15 18 24
1 12 14 15 18
1 14 18 20 24
1 13 15 19 21
1 11 13 16 19
2 21 30 42 50
2 24 36 45 56
2 23 27 30 35
2 26 35 38 45
2 29 38 49 57
2 28 38 45 54
3 50 53 57 59
3 51 54 58 60
3 54 58 62 68
3 50 51 54 57
3 53 54 57 63
3 51 54 55 56
3 52 53 56 58
```

En première colonne figure un unique facteur, noté  $F$ , à trois niveaux, notés 1, 2 et 3. Le facteur  $F$  est supposé à effets fixes. Les tests que nous allons détailler concernant les effets fixes, nous avons, pour simplifier, considéré un modèle à effets fixes avec un unique facteur. Par contre, certains résultats étant un peu particuliers si le facteur ne comporte que deux niveaux, nous avons considéré un facteur à trois niveaux. De plus, nous avons volontairement considéré un plan déséquilibré, afin d'avoir les résultats les plus généraux possible. Ainsi, les données comportent respectivement 5, 6 et 7 observations dans les trois niveaux de  $F$ , soit un échantillon de 18 observations (18 lignes dans le fichier ci-dessus).

Il y a ensuite, dans les quatre colonnes suivantes du fichier, une variable réponse  $Y$  observée à quatre instants différents (ces variables seront par la suite notées  $Y_1$ ,  $Y_2$ ,  $Y_3$  et  $Y_4$ ).

## E.2 Traitement avec la commande `repeated` de la procédure GLM

Nous faisons ici un traitement de ces données avec la procédure GLM de SAS, au sein de laquelle nous devons mettre la commande `repeated`. Les données ci-dessus sont dans un fichier appelé `repet.don` et contenu dans le répertoire dans lequel nous avons ouvert SAS.

```
data repet;
infile 'repet.don';
input F $ Y1-Y4;
run;
* ----- ;
*      modelisation avec GLM      ;
* ----- ;
proc glm data=repet;
class F;
model Y1-Y4 = F / ss3;
repeated temps contrast(1) / printh printe;
run;
```

L'option `ss3` de la commande `model` permet de n'avoir en sortie que les sommes de type 3 (les seules qui nous intéressent ici). L'élément `contrast(1)` de la commande `repeated` permet de calculer les évolutions par rapport au temps 1, au lieu de le faire par rapport au temps 4, comme cela est fait par défaut. Enfin, les options `printh` et `printe` permettent d'obtenir les matrices **H** et **E** qui interviennent dans la définition des statistiques des tests multidimensionnels.

Nous donnons ci-dessous une partie des résultats obtenus. Nous ne faisons pas figurer les premiers d'entre eux qui sont les ANOVA unidimensionnelles réalisées, à chaque instant, selon le facteur *F*. Notons simplement que ces quatre ANOVA sont très significatives (toutes les *p-values* sont inférieures à  $10^{-4}$ ) et sont associées à des coefficients  $R^2$  tous supérieurs à 0.9. On peut donc penser qu'il y aura un effet marginal de *F* (indépendamment du temps) très significatif.

The GLM Procedure  
Repeated Measures Analysis of Variance

Repeated Measures Level Information

Dependent Variable	Y1	Y2	Y3	Y4
Level of temps	1	2	3	4

E = Error SSCP Matrix

temps\_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3	temps_4
temps_2	52.262	80.476	113.190
temps_3	80.476	196.248	255.019
temps_4	113.190	255.019	367.848

H = Type III SSCP Matrix for temps

temps\_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3	temps_4
temps_2	391.24276814	758.20605251	1166.7347575
temps_3	758.20605251	1469.3598576	2261.0650645
temps_4	1166.7347575	2261.0650645	3479.3486426

MANOVA Test Criteria and Exact F Statistics  
for the Hypothesis of no temps Effect  
H = Type III SSCP Matrix for temps  
E = Error SSCP Matrix

S=1 M=0.5 N=5.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.09087595	43.35	3	13	<.0001
Pillai's Trace	0.90912405	43.35	3	13	<.0001
Hotelling-Lawley Trace	10.00401131	43.35	3	13	<.0001
Roy's Greatest Root	10.00401131	43.35	3	13	<.0001

H = Type III SSCP Matrix for temps\*F

temps\_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3	temps_4
temps_2	157.73809524	271.19047619	388.80952381
temps_3	271.19047619	469.53015873	671.98095238
temps_4	388.80952381	671.98095238	962.15238095

MANOVA Test Criteria and F Approximations  
for the Hypothesis of no temps\*F Effect  
H = Type III SSCP Matrix for temps\*F  
E = Error SSCP Matrix

S=2 M=0 N=5.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.23139352	4.68	6	26	0.0024
Pillai's Trace	0.80107053	3.12	6	28	0.0182
Hotelling-Lawley Trace	3.18134407	6.69	6	15.676	0.0012
Roy's Greatest Root	3.13661494	14.64	3	14	0.0001

The GLM Procedure  
Repeated Measures Analysis of Variance  
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
F	2	17971.10754	8985.55377	181.81	<.0001
Error	15	741.33690	49.42246		

Signalons que le dernier test réalisé (celui relatif à l'effet marginal du facteur  $F$ ) est, comme prévu, très significatif, avec encore une  $p$ -value inférieure à  $10^{-4}$ . Les test multidimensionnels sont, de leur côté, très significatifs pour le temps et assez significatifs pour les interactions entre le temps et le facteur. Par conséquent, tous les effets déclarés dans ce modèle sont significatifs.

Interrogeons nous maintenant sur la façon dont sont construits ces tests multidimensionnels, autrement dit sur la façon dont sont obtenues les matrices  $\mathbf{E}$  pour les erreurs du modèle,  $\mathbf{H}_T$  pour les tests relatifs au temps et  $\mathbf{H}_{T*F}$  pour ceux relatifs aux interactions. Pour cela, nous devons considérer ce que nous allons appeler les **évolutions**, c'est-à-dire les différences entre les observations de la variable réponse  $Y$  à chaque instant  $t$  variant de 2 à 4 (de façon générale, de 2 à  $T$ ) et les observations de cette même variable  $Y$  à l'instant initial 1 (instant initial souvent appelé *baseline* dans le "jargon" de la statistique médicale et parfois noté 0).

## E.3 Traitement multivarié des variables d'évolution

### E.3.1 Introduction

Définissons donc les trois variables d'évolution suivantes :

$$Z_2 = Y_2 - Y_1; Z_3 = Y_3 - Y_1; Z_4 = Y_4 - Y_1.$$

Nous allons maintenant réaliser, toujours avec la procédure GLM de SAS, mais cette fois avec la commande `manova`, une analyse multivariée des trois variables  $Z_2$ ,  $Z_3$  et  $Z_4$ , en testant la significativité du facteur  $F$ .

Voici le programme SAS pour réaliser cette analyse :

```
* ----- ;
*      calcul des evolutions :      ;
*      Z2=Y2-Y1 Z3=Y3-Y1 Z4=Y4-Y1  ;
* ----- ;
data evol;
set repet;
Z2 = Y2 - Y1;
Z3 = Y3 - Y1;
Z4 = Y4 - Y1;
run;
* ----- ;
*      MANOVA des evolutions      ;
* ----- ;
proc glm data=evol;
class F;
model Z2-Z4 = F / ss3;
manova H = F / printh printe;
run;
```

Et en voici les résultats.

```

                                The GLM Procedure

                                Class Level Information

                                Class          Levels    Values
                                F              3         1 2 3

                                Number of Observations Read          18

                                E = Error SSCP Matrix

                                Z2              Z3              Z4

                                Z2    52.261904762    80.476190476    113.19047619
                                Z3    80.476190476    196.24761905    255.01904762
                                Z4    113.19047619    255.01904762    367.84761905
```

H = Type III SSCP Matrix for F

	Z2	Z3	Z4
Z2	157.73809524	271.19047619	388.80952381
Z3	271.19047619	469.53015873	671.98095238
Z4	388.80952381	671.98095238	962.15238095

Characteristic Roots and Vectors of: E Inverse \* H, where

H = Type III SSCP Matrix for F

E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector		V'EV=1	
		Z2	Z3	Z3	Z4
3.13661494	98.59	0.09861045	-0.00613949		0.02147662
0.04472912	1.41	-0.19155986	0.13254247		-0.01485006
0.00000000	0.00	-0.10786212	-0.18554264		0.17317314

MANOVA Test Criteria and F Approximations  
for the Hypothesis of No Overall F Effect

H = Type III SSCP Matrix for F

E = Error SSCP Matrix

S=2 M=0 N=5.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.23139352	4.68	6	26	0.0024
Pillai's Trace	0.80107053	3.12	6	28	0.0182
Hotelling-Lawley Trace	3.18134407	6.69	6	15.676	0.0012
Roy's Greatest Root	3.13661494	14.64	3	14	0.0001

Notons tout d'abord que, comme toujours, les premiers résultats fournis par la procédure GLM dans un contexte multidimensionnel sont les ANOVA univariées de chacune des variables  $Z_2$ ,  $Z_3$  et  $Z_4$  par rapport au facteur  $F$ . Elles ne figurent pas ci-dessus, mais nous pouvons indiquer qu'elles sont toutes les trois très significatives ( $p$ -values inférieures ou égales à  $10^{-4}$ ) et possèdent un bon coefficient  $R^2$  (compris entre 0.70 et 0.75).

### E.3.2 Tests des interactions

Regardons maintenant les deux matrices  $\mathbf{E}$  et  $\mathbf{H}$  obtenues à l'issue de cette analyse. La matrice  $\mathbf{E}$  de ce modèle est la même que celle obtenue avec les données initiales (les quatre observations de la variable  $Y$  et la commande `repeated`), ce qui signifie que les deux modèles sont équivalents. En effet, en prenant en compte, dans le modèle ci-dessus, les variables d'évolution, on fait bien intervenir le facteur temps et, en déclarant le facteur  $F$ , ce dernier intervient également. Les résidus de ce nouveau modèle sont donc logiquement les mêmes que dans le modèle initial, dans lequel intervenaient le temps, le facteur et les interactions. Par ailleurs, on constate également que la matrice  $\mathbf{H}$  du modèle relatif aux évolutions (le modèle ci-dessus) est identique à la matrice  $\mathbf{H}_{T*F}$  associée aux interactions dans le modèle initial. Par conséquent, dans le modèle pour données répétées traité avec GLM, les tests multidimensionnels relatifs aux interactions `temps * facteur` correspondent aux tests relatifs au facteur dans le modèle prenant en compte les évolutions, ce qui est logique. Par ailleurs, on a ici  $J = 3$  (donc  $\nu_H = 2$ ),  $\nu_E = n - J = 15$  et  $D = 3$ , ce qui permet de retrouver les 4 statistiques de Fisher et leurs d.d.l. en fonction des formules données en 5.3.2 et 5.3.3.

Une autre façon, plus rigoureuse, de voir les choses est d'écrire le modèle initial, sur les v.a.r.

$Y_{ijt}$ , selon le paramétrage centré :

$$Y_{ijt} = \mu + \alpha_j^1 + \alpha_t^2 + \gamma_{jt} + U_{ijt} ,$$

où  $\mu$  est l'effet (moyen) général, les  $\alpha_j^1$  sont les effets principaux (centrés selon  $j$ ) du facteur, les  $\alpha_t^2$  les effets principaux (centrés selon  $t$ ) du temps, les  $\gamma_{jt}$  les effets (doublement centrés) d'interactions et les  $U_{ijt}$  des v.a.r. erreurs telles que les vecteurs  $U_{ij} = (U_{ij1} \cdots U_{ijT})'$  de  $\mathbb{R}^T$  sont indépendants, gaussiens, centrés, avec une structure de covariance  $\Sigma$  constante (indépendante de  $i$  et de  $j$ ). On obtient alors :

$$\begin{aligned} Z_{ijt} = Y_{ijt} - Y_{ij1} &= (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + (U_{ijt} - U_{ij1}) \\ &= (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + E_{ijt} , \end{aligned}$$

en posant  $E_{ijt} = U_{ijt} - U_{ij1}$ , les  $E_{ijt}$  étant toujours des v.a.r. erreurs telles que les vecteurs  $E_{ij} = (E_{ij2} \cdots E_{ijT})'$  de  $\mathbb{R}^{T-1}$  sont indépendants, gaussiens, centrés, seule leur structure de covariance ayant changé. Notons maintenant  $\bar{Z}_{\bullet jt}$  la moyenne des quantité  $Z_{ijt}$  sur l'indice  $i$  et  $\bar{Z}_{\bullet \bullet t}$  la moyenne des mêmes quantités sur les deux indices  $i$  et  $j$ . Il vient :

$$\bar{Z}_{\bullet jt} = (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + \bar{E}_{\bullet jt} \quad \text{et} \quad \bar{Z}_{\bullet \bullet t} = (\alpha_t^2 - \alpha_1^2) + \bar{E}_{\bullet \bullet t} ,$$

puisque les quantités  $\gamma_{jt}$  sont doublement centrées. Les tests multidimensionnels (dans  $\mathbb{R}^{T-1}$ ) de significativité du facteur  $F$  font intervenir les deux matrices  $\mathbf{H}$  et  $\mathbf{E}$  dont les termes généraux sont définis ci-après. Le terme général de  $\mathbf{H}$  est

$$\sum_{j=1}^J n_j (\bar{Z}_{\bullet jt} - \bar{Z}_{\bullet \bullet t}) (\bar{Z}_{\bullet jt'} - \bar{Z}_{\bullet \bullet t'}) = \sum_{j=1}^J n_j [(\gamma_{jt} - \gamma_{j1}) + (\bar{E}_{\bullet jt} - \bar{E}_{\bullet \bullet t})][(\gamma_{jt'} - \gamma_{j1}) + (\bar{E}_{\bullet jt'} - \bar{E}_{\bullet \bullet t'})];$$

celui de  $\mathbf{E}$  est

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (Z_{ijt} - \bar{Z}_{\bullet jt}) (Z_{ijt'} - \bar{Z}_{\bullet jt'}) = \sum_{j=1}^J \sum_{i=1}^{n_j} (E_{ijt} - \bar{E}_{\bullet jt}) (E_{ijt'} - \bar{E}_{\bullet jt'}) .$$

On voit ainsi pourquoi les tests multidimensionnels relatifs au facteur  $F$  sur les données d'évolution sont en fait des tests relatifs aux interactions sur les données initiales.

Intéressons nous maintenant à la matrice  $\mathbf{H}_T$  intervenant dans les tests relatifs au temps dans le modèle pour données répétées traité avec GLM, tel qu'il a été introduit en E.2. Elle est beaucoup plus délicate à obtenir et nous l'explicitons dans le paragraphe suivant.

## E.4 Tests relatifs au temps

Il s'agit des tests multidimensionnels obtenus dans le modèle initial. Ils font intervenir la matrice  $\mathbf{E}$ , définie dans le point précédent, et la matrice  $\mathbf{H}_T$  que nous précisons ci-dessous.

### E.4.1 Expression de la matrice $\mathbf{H}_T$

L'hypothèse nulle correspond à la non influence du temps sur les mesures  $Y_{ijt}$ , autrement dit à la constance de ces mesures au cours du temps, autrement dit encore à la nullité des variables d'évolution  $Z_{ijt}$ . C'est sur ces dernières que l'on va définir l'hypothèse nulle, puisque les matrices  $\mathbf{H}_{T*F}$  et  $\mathbf{E}$  ont déjà été définies à partir des variables d'évolution.

En utilisant le paramétrage centré  $Z_{ijt} = (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + E_{ijt}$ , la non influence du temps correspond, en fait, à la nullité des  $T-1$  paramètres  $\alpha_t^2 - \alpha_1^2$  (on notera, ici encore, qu'un tel test n'a de sens que dans la mesure où les interactions entre le temps et le facteur sont elles-même supposées nulles). L'hypothèse nulle peut donc s'énoncer sous la forme suivante :

$$\{H_0 : \alpha_2^2 - \alpha_1^2 = \cdots = \alpha_T^2 - \alpha_1^2 = 0\}.$$

Compte tenu de l'expression donnée plus haut pour les  $\bar{Z}_{\bullet \bullet t}$ , on peut vérifier sans difficulté que, pour tout  $t$ , l'estimateur maximum de vraisemblance de  $\alpha_t^2 - \alpha_1^2$  est  $\bar{Z}_{\bullet \bullet t} = \frac{1}{J} \sum_{j=1}^J \bar{Z}_{\bullet jt}$ , de sorte

que, en pratique,  $H_0$  s'écrit  $\{H_0 : \bar{Z}_{\bullet\bullet 2} = \dots = \bar{Z}_{\bullet\bullet T} = 0\}$ , soit encore  $\{H_0 : \sum_{j=1}^J \bar{Z}_{\bullet j 2} = \dots = \sum_{j=1}^J \bar{Z}_{\bullet j T} = 0\}$ , que l'on peut réécrire sous la forme  $\{H_0 : \mathbf{C}'\mathbf{Z} = 0\}$ , où  $\mathbf{Z}$  est la matrice  $J \times (T-1)$  de terme général  $\bar{Z}_{\bullet jt}$  et où  $\mathbf{C} = \mathbb{1}_J$ , vecteur de  $\mathbb{R}^J$  dont toutes les composantes sont égales à 1.

Pour définir la matrice  $\mathbf{H}_T$  correspondant à l'hypothèse nulle considérée, il est préférable ici de revenir à l'expression de la statistique du test de Fisher dans le modèle linéaire gaussien classique (voir le 2.3). Rappelons que la matrice  $\mathbf{H}$  intervenant dans les tests multidimensionnels généralise le numérateur  $N$  de la statistique de Fisher et que ce dernier peut s'écrire de différentes façons, l'une des plus commodes étant la suivante :

$$N = \hat{B}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\hat{B}.$$

Dans cette expression,  $\mathbf{X}$  est la matrice d'incidence du modèle : elle est  $n \times p$ , si  $n$  est la taille de l'échantillon considéré et si  $p$  désigne le nombre total d'effets fixes (indépendants) pris en compte dans le modèle (ici,  $p = J$ ) ;  $\hat{B}$  est l'estimateur maximum de vraisemblance du vecteur  $\beta$  de  $\mathbb{R}^p$  des paramètres du modèle ;  $\mathbf{C}$  est une matrice  $p \times q$  de rang  $q$  ( $1 \leq q < p$ ) définissant l'hypothèse nulle sous la forme  $\{H_0 : \mathbf{C}'\beta = 0\}$ .

Dans le cas considéré ici, nous avons écrit l'hypothèse nulle sous la forme  $\{H_0 : \mathbf{C}'\mathbf{Z} = 0\}$  et la transposition du terme  $N$  au cas multidimensionnel d'ordre  $T-1$  nous donne l'expression suivante pour la matrice  $\mathbf{H}_T$  :

$$\mathbf{H}_T = \mathbf{Z}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\mathbf{Z}.$$

Cette expression de  $\mathbf{H}_T$  se simplifie en remarquant que la matrice d'incidence  $\mathbf{X}$ , de dimension  $n \times J$ , comporte en colonnes les indicatrices des niveaux du facteur  $F$  (les indicatrices des cellules dans le cas général), de sorte que  $\mathbf{X}'\mathbf{X} = \text{diag}(n_1 \dots n_J)$ ,  $(\mathbf{X}'\mathbf{X})^{-1} = \text{diag}(\frac{1}{n_1} \dots \frac{1}{n_J})$ ,  $\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C} = \frac{1}{n_1} + \dots + \frac{1}{n_J}$  et  $[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} = \frac{1}{\frac{1}{n_1} + \dots + \frac{1}{n_J}} = n^*$ , où  $n^*$  est la moyenne harmonique des effectifs  $n_j$ , divisée par  $J$ . Finalement, il vient :

$$\mathbf{H}_T = n^*\mathbf{Z}'\mathbb{1}_{J \times J}\mathbf{Z},$$

où  $\mathbb{1}_{J \times J}$  est la matrice carrée d'ordre  $J$  ne comportant que des 1.

**Remarque 89** On notera que, dans un plan équilibré avec  $n_0$  observations par cellule ( $n = Jn_0$ ), il vient :  $n^* = \frac{n_0}{J}$ . De plus, dans ce cas, le terme général de la matrice  $\mathbf{H}_T$  s'écrit  $n\bar{Z}_{\bullet\bullet t}\bar{Z}_{\bullet\bullet t}$ .

**Remarque 90** Il est important de remarquer que l'hypothèse nulle  $\{H_0 : \mathbf{C}'\mathbf{Z} = 0\}$  est définie par une seule contrainte sur  $\mathbf{Z}$ , la matrice  $\mathbf{C}$  ne comportant qu'une seule colonne ( $q = 1$ ). En fait,  $H_0$  exprime le centrage, dans  $\mathbb{R}^J$ , des  $T-1$  vecteurs  $\bar{Z}_{\bullet t}$  ( $t = 2, \dots, T$ ), de coordonnées  $\bar{Z}_{\bullet jt}$ . L'hypothèse nulle signifie donc que les vecteurs  $\bar{Z}_{\bullet t}$  sont dans un hyperplan de  $\mathbb{R}^J$ , ce qui correspond bien à une contrainte unique. Par conséquent, le d.d.l. associé à  $H_0$  (noté  $\nu_H$  dans les tests multidimensionnels) vaut toujours 1 dans ce cas :  $\nu_H = 1$ . Par suite, la matrice  $\mathbf{H}_T\mathbf{E}^{-1}$  n'admet qu'une seule valeur propre.

## E.4.2 Application

Revenons maintenant aux données traitées depuis le début. Pour chaque variable d'évolution, nous calculons ses moyennes partielles relativement aux trois niveaux du facteur  $F$ , afin de déterminer la matrice  $\mathbf{Z}$  définie plus haut.

Voici le programme SAS réalisant ce calcul :

```
proc means data=evol;
var Z2-Z4;
by F;
run;
```

Et voici les résultats obtenus :

----- F=1 -----

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Z2	5	3.0000000	1.4142136	2.0000000	5.0000000
Z3	5	5.6000000	1.8165902	3.0000000	8.0000000
Z4	5	9.2000000	3.0331502	6.0000000	14.0000000

----- F=2 -----

Variable	N	Mean	Std Dev	Minimum	Maximum
Z2	6	8.8333333	2.6394444	4.0000000	12.0000000
Z3	6	16.3333333	5.7154761	7.0000000	21.0000000
Z4	6	24.3333333	7.4475947	12.0000000	32.0000000

----- F=3 -----

Variable	N	Mean	Std Dev	Minimum	Maximum
Z2	7	2.2857143	1.2535663	1.0000000	4.0000000
Z3	7	5.4285714	1.8126539	4.0000000	8.0000000
Z4	7	8.5714286	2.9920530	5.0000000	14.0000000

La matrice  $\mathbf{Z}$ , de dimension  $J \times (T-1)$ , soit ici  $3 \times 3$ , s'obtient à partir des moyennes partielles (Mean) obtenues ci-dessus. Il vient :

$$\mathbf{Z}' = \begin{pmatrix} 3.00 & 8.83 & 2.29 \\ 5.60 & 16.33 & 5.43 \\ 9.20 & 24.33 & 8.57 \end{pmatrix}.$$

On en déduit :

$$\mathbf{Z}'\mathbb{I}_{3 \times 3}\mathbf{Z} = \begin{pmatrix} 199.35 & 386.32 & 594.48 \\ 386.32 & 748.67 & 1152.07 \\ 594.48 & 1152.07 & 1772.81 \end{pmatrix}.$$

On peut par ailleurs vérifier que  $n^* = \frac{210}{107}$ , ce qui permet de calculer (aux erreurs d'arrondi près) :

$$\mathbf{H}_T = n^* \mathbf{Z}'\mathbb{I}_{3 \times 3}\mathbf{Z} = \begin{pmatrix} 391.24 & 758.21 & 1166.73 \\ 758.21 & 1469.36 & 2261.07 \\ 1166.73 & 2261.07 & 3479.35 \end{pmatrix}.$$

On retrouve bien ainsi la matrice  $\mathbf{H}_T$  fournie en E.2 par la procédure GLM de SAS avec la commande `repeated`.

**Remarque 91** Pour le calcul des 4 statistiques de Fisher associées aux tests multidimensionnels et de leurs d.d.l., on utilisera ici  $\nu_H = 1$  comme indiqué plus haut,  $\nu_E = n - J$  (15 ici, puisque la matrice  $\mathbf{E}$  est la même, quelle que soit l'hypothèse testée dans le modèle) et  $D = T - 1$  (3 ici).

## E.5 Bilan

Dans les points E.3 et E.4 ci-dessus, on a explicité le calcul des matrices  $\mathbf{H}_T$ ,  $\mathbf{H}_{T*F}$  et  $\mathbf{E}$  permettant de déterminer les statistiques des tests multidimensionnels (et, principalement, le test de Wilks), ces matrices intervenant dans la procédure GLM du logiciel statistique SAS lorsqu'on déclare la commande `repeated` pour traiter des données répétées. La logique de ces tests apparaît

ainsi clairement, et il semble tout indiqué de les utiliser pour tester la significativité des effets fixes dans des modèles linéaires mixtes pour données répétées.

On notera que ces tests multidimensionnels peuvent être des tests approchés mais, en général, les approximations obtenues sont bonnes. On notera également que ces tests ne dépendent pas de la structure de covariance choisie pour les données répétées (matrice  $\mathbf{R}$ ) et qu'on peut donc les mettre en œuvre avant de choisir cette dernière. Pour le choix de la matrice  $\mathbf{R}$ , c'est la procédure MIXED de SAS qui est la plus appropriée, de même que pour l'estimation des composantes de la variance correspondant aux effets aléatoires.



## Annexe F

# Spécificité de la structure “compound symmetry”

*Cette structure de covariance, utilisée dans la modélisation des données répétées, comporte des particularités que nous détaillons ici (et qui lui ont valu son succès dans la pratique, du moins lorsque la procédure MIXED n'existait pas).*

### F.1 Étude des éléments propres d'une matrice particulière

Deux nombres réels quelconques  $a$  et  $b$  étant donnés, considérons la matrice carrée d'ordre  $n$  ( $n \geq 2$ ) suivante :

$$\mathbf{M} = \begin{pmatrix} a+b & a & \cdots & a \\ a & a+b & \cdots & a \\ \vdots & & \ddots & \vdots \\ a & \cdots & a & a+b \end{pmatrix} = a \mathbb{I}_{n \times n} + b \mathbf{I}_n,$$

où  $\mathbb{I}_{n \times n}$  désigne la matrice carrée d'ordre  $n$  dont tous les éléments valent 1 et où  $\mathbf{I}_n$  désigne la matrice identité d'ordre  $n$ .

Considérons par ailleurs l'espace vectoriel réel  $\mathbb{R}^n$  muni de la base canonique  $\mathcal{C} = (c_1, c_2, \dots, c_n)$ , notons  $\mathbb{I}_n$  le vecteur de  $\mathbb{R}^n$  dont toutes les coordonnées sur la base canonique valent 1 et désignons par  $\alpha$  et  $\beta$  deux réels quelconques non nuls. Il est alors immédiat de vérifier que le vecteur  $\alpha \mathbb{I}_n$  est vecteur propre de  $\mathbf{M}$  associé à la valeur propre  $na + b$  et encore facile de vérifier que tout vecteur de la forme  $\beta(\mathbb{I}_n - nc_i)$  est aussi vecteur propre de  $\mathbf{M}$ , associé quant à lui à la valeur propre  $b$ . Comme on a la relation  $\sum_{i=1}^n c_i = \mathbb{I}_n$ , il est clair que le sous-espace vectoriel de  $\mathbb{R}^n$  engendré par les vecteurs de la forme  $\beta(\mathbb{I}_n - nc_i)$  est de dimension  $n - 1$ . On en déduit que  $\mathbf{M}$  n'admet que deux valeurs propres distinctes :  $\lambda_1 = b$ , d'ordre  $n - 1$ , et  $\lambda_2 = na + b$ , d'ordre 1.

### F.2 Application à la structure “compound symmetry”

Dans la modélisation de données répétées au cours du temps ( $T$  instants d'observation), il est courant d'utiliser, comme structure de covariance entre les observations réalisées au cours du temps sur un même individu, la structure dite “compound symmetry”. Elle est associée à une matrice  $\mathbf{R}$ , carrée d'ordre  $T$ , de même structure que la matrice  $\mathbf{M}$  ci-dessus, avec  $a = \sigma_2^2$  et  $b = \sigma_1^2$ ,  $\sigma_1^2$  et  $\sigma_2^2$  désignant deux composantes de la variance (strictement positives). La matrice  $\mathbf{R}$  admet donc seulement deux valeurs propres,  $\lambda_1 = \sigma_1^2$ , d'ordre  $T - 1$ , et  $\lambda_2 = \sigma_1^2 + T\sigma_2^2$ , d'ordre 1.

Notons  $w_1, \dots, w_{T-1}$  une base orthonormée (au sens de la métrique euclidienne classique de  $\mathbb{R}^T$ ) du sous-espace propre associé à  $\lambda_1$ , et notons  $\mathbf{W}$  la matrice  $T \times (T - 1)$  contenant les vecteurs  $w_k$  ( $k = 1, \dots, T - 1$ ) disposés en colonnes. La matrice  $\mathbf{W}$  est de rang  $T - 1$  et vérifie les égalités :

$$\mathbf{R}\mathbf{W} = \lambda_1 \mathbf{W} ; \quad \mathbf{W}'\mathbf{W} = \mathbf{I}_{T-1}.$$

En posant maintenant  $\mathbf{C} = \mathbf{W}'$ , en considérant un vecteur aléatoire  $Y$  de  $\mathbb{R}^T$ , de loi  $\mathcal{N}_T(\mu, \mathbf{R})$ , et en définissant  $Y^* = \mathbf{C}Y$ , il vient :  $Y^* \sim \mathcal{N}_{T-1}(\mathbf{C}\mu, \mathbf{C}\mathbf{R}\mathbf{C}')$ , avec :

$$\mathbf{C}\mathbf{R}\mathbf{C}' = \mathbf{W}'\mathbf{R}\mathbf{W} = \lambda_1 \mathbf{W}'\mathbf{W} = \lambda_1 \mathbf{I}_{T-1} = \sigma_1^2 \mathbf{I}_{T-1}.$$

C'est cette propriété, correspondant à la structure d'indépendance en dimension  $T - 1$  pour les composantes du vecteur  $Y^*$ , qui est exploitée dans l'étude des données répétées avec la structure "compound symmetry" sur  $Y$ .

**Remarque 92** *C'est l'adéquation d'une matrice de covariance  $\mathbf{R}$ , de la forme  $\mathbf{R} = \sigma_1^2 \mathbf{I}_n + \sigma_2^2 \mathbf{I}_{n \times n}$ , aux données considérées qui est testée avec le test de sphéricité de Mauchly.*

# Annexe G

## Bibliographie

### G.1 Ouvrages généraux

- T.W. Anderson, “An introduction to multivariate statistical analysis”, Wiley, 2003.
- J.-M. Azaïs & J.-M. Bardet, “Le modèle linéaire par l'exemple”, Dunod, 2005.
- J.-C. Bergonzini & C. Duby, “Analyse et planification des expériences”, Masson, 1995.
- H. Brown & R. Prescott, “Applied mixed models in medicine”, Wiley, 1999.
- D. Collombier, “Plans d'expérience factoriels”, Springer, 1996.
- C.S. Davis, “Statistical methods for the analysis of repeated measurements”, Springer, 2002.
- J.-J. Dreesbeke, J. Fine & G. Saporta, “Plans d'expériences”, Technip, 1997.
- R.A. Fisher & F. Yates, “Statistical tables”, Oliver and Boyd, 1963.
- J. Goupy & L. Creighton, “Introduction aux plans d'expériences”, Dunod, 2006.
- P.W.M. John, “Statistical design and analysis of experiments”, SIAM, 1998.
- B. Jorgensen, “The theory of linear models”, Chapman & Hall, 1993.
- A.I. Khuri, T. Mathew & B.M. Sinha, “Statistical tests for mixed linear models”, Wiley, 1998.
- R.G. Miller Jr., “Beyond ANOVA”, Chapman & Hall, 1997.
- G.A. Milliken & D.E. Johnson, “Analysis of messy data”, Volume I : designed experiments, Van Nostrand Reinhold, 1984.
- A. Monfort, “Cours de statistique mathématique”, Économica, 1997.
- A.C. Rencher, “Methods of multivariate analysis”, Wiley, 1995.
- A.C. Rencher & G.B. Schaalje, “Linear models in statistics”, Wiley, 2008.
- G. Saporta, “Probabilités, analyse des données et statistique”, Technip, 2006.
- S.R. Searle, G. Casella & C.E. McCulloch, “Variance components”, Wiley, 1992.

- G.A.F. Seber, “Multivariate observations”, Wiley, 1984.
- G. Verbeke & G. Molenberghs, “Linear mixed models for longitudinal data”, Springer, 2000.

## G.2 Articles spécialisés

- H. Akaike, “A new look at the statistical model identification”, IEEE transactions on automatic control, 19 (6), 716-723, 1974.
- J.-M. Azais, “Analyse de variance non orthogonale; l'exemple de SAS/GLM”, Revue de Statistique Appliquée, 42 (2), 27-41, 1994.
- R.C. Boze, S.S. Shrikhande & E.T. Parker, “Further results on the construction of mutually orthogonal latin squares and the falsity of Euler's conjecture”, Canadian Journal of Mathematics, 12, 189-203, 1960.
- D.A. Harville, “Maximum likelihood approaches to variance component estimation and to related problems”, Journal of the American Statistical Association, 72 (358), 320-338, 1977.
- C.R. Henderson, “Estimation of variance and covariance components”, Biometrics, 9 (2), 226-252, 1953.
- C.L. Mallows, “Some comments on CP”, Technometrics, 15 (4), 661-675, 1973.
- H.D. Patterson & R. Thompson, “Recovery of inter-block information when block sizes are unequal”, Biometrika, 58 (3), 545-554, 1971.
- R.L. Plackett & J.P. Burman, “The design of optimum multifactorial experiments”, Biometrika, 33 (4), 305-325, 1946.
- C.R. Rao, “Estimation of heteroscedastic variances in linear models”, Journal of the American Statistical Association, 65 (329), 161-172, 1970.
- C.R. Rao, “Estimation of variance and covariance components : MINQUE theory”, Journal of Multivariate Analysis, 1, 257-275, 1971.
- C.R. Rao, “Minimum Variance Quadratic Unbiased Estimation of variance components”, Journal of Multivariate Analysis, 1, 445-456, 1971.
- C.R. Rao, “Estimation of variance and covariance components in linear models”, Journal of the American Statistical Association, 67 (337), 112-115, 1972.
- F.E. Satterthwaite, “An approximate distribution of estimates of variance components”, Biometrics Bulletin, 2 (6), 110-114, 1946.
- G. Schwarz, “Estimating the dimension of a model”, The Annals of Statistics, 6 (2), 461-464, 1978.