Le Modèle Linéaire et ses Extensions

L. Bel, JJ Daudin, M. Etienne, E. Lebarbier, T. Mary-Huard, S. Robin, C. Vuillet

14 septembre 2016

Avant-propos

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

— il est capable de bien représenter (au moins en première approximation) un grand nombre d'expériences,

— sa simplicité permet de bien connaître ses propriétés.

De ce fait son emploi est très fréquent et il joue un rôle pivot dans toute la modélisation statistique. La façon dont on traite les modèles plus généraux en découle directement. Il est donc essentiel de bien comprendre ce modèle fondamental de la statistique qui fait partie de la culture de base de tout ingénieur ou chercheur dont les résultats expérimentaux s'expriment de façon chiffrée.

Lorsque les hypothèses de base sont trop contraignantes, des extensions de ce modèle permettent de prendre en compte certaines situations particulières. Nous présentons les deux extensions principales, le modèle linéaire généralisé et le modèle mixte. De plus bien avant l'analyse des résultats, la conception des expériences est essentielle pour espérer atteindre les objectifs de recherche attendus. C'est pourquoi nous présentons également les éléments de base des plans d'expériences.

La théorie est une chose, mais l'application pratique sur des exemples réels est essentielle pour bien comprendre et être capable d'appliquer ces méthodes. Nous avons séparé chaque méthode en deux chapitres. Le premier expose de façon synthétique la théorie et le deuxième traite plusieurs exemples et donne à l'occasion les compléments théoriques nécessaires. Ainsi le lecteur qui veut vérifier un élément théorique peut le retrouver facilement. Chaque exemple a été traité avec SAS ou R et les programmes de l'un ou l'autre de ces logiciels sont donnés.

Ce livre est issu de plusieurs polycopiés rédigés par les enseignants d'AgroParisTech pour les élèves de cet établissement. Il s'adresse donc à des étudiants ingénieurs, ou en licence ou master mais aussi à des utilisateurs de ces modèles statistiques qui veulent rafraîchir ou étendre leurs connaissances, vérifier un point particulier. Il nécessite des bases en algèbre linéaire et calcul des probabilités, dont certaines sont rappelées en annexe. Il suppose connues les notions de base de l'inférence statistique, estimateur, estimation, tests d'hypothèses (cf [18]). Il a été conçu comme un ouvrage de référence en français sur le sujet, mais il ne l'épuise pas. Des références sont indiquées pour aller plus loin sur des thèmes que nous n'avons pas pu développer faute de place.

Les deux premiers chapitres sont fortement inspirés du polycopié *Modèle linéaire* de C. Duby dont nous avons repris le plan et certaines parties.

Nous remercions Julie Aubert, Pierre Barbillon, Maud Delattre, Sophie Donnet, Jean-Benoist Leger, Sarah Ouadah, Laure Sansonnet et Jessica Tressou d'avoir relu cet

ouvrage et permis de l'améliorer. Il n'en reste pas moins que les auteurs sont les seuls responsables des erreurs qui auraient résisté à ces relectures vigilantes.

Table des matières

Avant-propos **3**

1 Le modèle linéaire **12**
1.1 Introduction . . . 12
1.2 Modélisation . . . 13
1.2.1 Modèle de régression . . . 13
1.2.2 Modèle d'analyse de la variance (anova) . . . 14
1.2.3 Modèle d'analyse de la covariance (ancova) . . . 15
1.2.4 Présentation unifiée du modèle linéaire . . . 16
1.3 Estimation des paramètres . . . 16
1.3.1 Estimation des paramètres de l'espérance θ . . . 17
1.3.2 Propriétés de l'estimateur $\widehat{\theta}$ . . . 23
1.3.3 Estimation de la variance σ^2 et propriétés des estimateurs . . . 26
1.3.4 Intervalles de confiance . . . 27
1.4 Tests d'hypothèses . . . 28
1.4.1 Tests sur les paramètres . . . 28
1.4.2 Tests de modèles emboîtés . . . 29
1.4.3 Tests à l'aide des réductions . . . 32
1.5 Notion d'orthogonalité . . . 37
1.5.1 Propriétés liées à l'orthogonalité . . . 38
1.5.2 Orthogonalité dans le modèle d'analyse de la variance . . . 38
1.5.3 Orthogonalité dans le cas de la régression multiple . . . 40
1.5.4 Orthogonalité dans le cas de l'analyse de la covariance . . . 40
1.6 Qualité d'ajustement et comparaison de modèles . . . 41
1.6.1 Le coefficient de détermination . . . 41
1.6.2 Critères de comparaison de modèles . . . 41
1.7 Diagnostic . . . 42
1.8 Limites et extensions du modèle linéaire . . . 44
1.8.1 Généralisations du modèle linéaire . . . 45
1.8.2 Les variables explicatives sont aléatoires . . . 45

2 Exemples de modèles linéaires **47**
2.1 Régression linéaire simple et polynomiale . . . 47
2.1.1 Présentation du problème . . . 47
2.1.2 Modèle de régression linéaire simple . . . 48
2.1.3 Modèle de régression polynomiale . . . 48
2.1.4 Influence de l'âge . . . 49

2.1.5 Programme. . . . 52
2.2 Régression linéaire multiple . . . 53
2.2.1 Problème et description des données . . . 53
2.2.2 Remarques. . . . 55
2.2.3 Modèle . . . 56
2.2.4 Influence des différentes variables sur le nombre de nids . . . 57
2.2.5 Sélection de variables . . . 58
2.2.6 Programme. . . . 60
2.3 Analyse de la variance à un facteur . . . 62
2.3.1 Description du problème et des données . . . 62
2.3.2 Effet du statut . . . 63
2.3.3 Comparaison des groupes de statuts. . . . 67
2.3.4 Programme R . . . 69
2.3.5 Programme SAS . . . 69
2.4 Analyse de la variance à deux facteurs : cas équilibré . . . 70
2.4.1 Objectif et dispositif expérimental . . . 70
2.4.2 Analyse descriptive . . . 71
2.4.3 Analyse de l'effet de la rotation et de la fertilisation . . . 72
2.4.4 Conclusion . . . 79
2.4.5 Programme R . . . 80
2.4.6 Programme SAS . . . 82
2.5 Analyse de la variance à deux facteurs : cas déséquilibré . . . 82
2.5.1 Objectif et dispositif . . . 82
2.5.2 Analyse descriptive . . . 84
2.5.3 Analyse de la variance à 2 facteurs . . . 85
2.5.4 Tests des effets et comparaison des copies . . . 87
2.5.5 Programme R . . . 92
2.6 Analyse de la covariance . . . 95
2.6.1 Problème et des données . . . 95
2.6.2 Station Châtelet . . . 95

3 Modèle linéaire généralisé 115
3.1 Introduction . . . 115
3.2 Modélisation . . . 116
3.2.1 Famille exponentielle naturelle . . . 116
3.2.2 Modèle . . . 117
3.2.3 Choix de la fonction de lien . . . 117
3.3 Estimation des paramètres . . . 118
3.3.1 Vraisemblance . . . 118
3.3.2 Propriétés de l'estimateur du maximum de vraisemblance . . . 119
3.4 Tests d'hypothèses . . . 120
3.4.1 Test de modèles emboîtés . . . 120
3.4.2 Tests de $\theta_j = \theta_{0j}$ . . . 121
3.4.3 Test de $C\theta = 0$ . . . 122
3.5 Qualité d'ajustement et choix de modèles . . . 122
3.5.1 Le pseudo R^2 . . . 122
3.5.2 Le χ^2 de Pearson généralisé . . . 123

3.5.3 Choix de modèle . . . 123
3.5.4 Diagnostic, résidus . . . 123
3.6 Régression logistique . . . 124
3.6.1 Régression logistique, cas où la réponse est binaire . . . 125
3.6.2 Régression multilogistique . . . 128
3.6.3 Surdispersion . . . 129
3.7 Preuve de la propriété 3.2.1, p. 116 . . . 130

4 Exemples de modèles linéaires généralisés 132
4.1 Loi de Bernoulli : pollution par l'ozone . . . 132
4.1.1 Contexte et données . . . 132
4.1.2 Résultats . . . 133
4.1.3 Programme R . . . 135
4.2 Loi binomiale : équité sociale . . . 137
4.2.1 Contexte et données . . . 137
4.2.2 Résultats . . . 138
4.2.3 Programme R . . . 139
4.3 Loi binomiale : courbe dose-réponse . . . 141
4.3.1 Contexte et données . . . 141
4.3.2 Résultats . . . 141
4.3.3 programme SAS . . . 143
4.4 Loi de Bernoulli, test de Hosmer-Lemeshow : asthme des enfants dans les écoles . . . 145
4.4.1 Contexte et données . . . 145
4.4.2 Résultats . . . 146
4.4.3 Programme SAS . . . 149
4.5 Loi de Poisson avec offset : biodiversité des fourmis en Guyane . . . 149
4.5.1 Contexte et données . . . 149
4.5.2 Résultats . . . 149
4.5.3 Richesses spécifiques . . . 151
4.5.4 Comparaison des richesses spécifiques . . . 152
4.5.5 Programme R . . . 152
4.6 Poisson tronquée en 0 : portées d'agneaux . . . 154
4.6.1 Contexte et données . . . 154
4.6.2 Résultats . . . 155
4.6.3 Comparaison des génotypes . . . 156
4.6.4 Programme R. . . . 157
4.7 Loi Gamma : roulements à billes, durées de vie . . . 159
4.7.1 Contexte et données . . . 159
4.7.2 Résultats . . . 159
4.7.3 Programme R . . . 161
4.8 Loi multinomiale : condamnations à mort en Floride . . . 161
4.8.1 Contexte et données . . . 161
4.8.2 Modèles pour les tables de contingence de dimension 3 . . . 161
4.8.3 Résultats . . . 162
4.8.4 Paradoxe de Simpson . . . 163
4.8.5 Programme SAS . . . 164

5 Modèle mixte, modélisation de la variance **166**
5.1 Modèle . . . 166
5.1.1 Composantes de la variance . . . 166
5.1.2 Mesures répétées . . . 168
5.1.3 Spatial . . . 169
5.2 Estimation . . . 170
5.2.1 Estimation de θ à ψ connu . . . 171
5.2.2 Estimation conjointe de θ et ψ par maximum de vraisemblance . 172
5.2.3 Autres méthodes d'estimation de Σ . . . 175
5.2.4 Prédiction des effets aléatoires . . . 181
5.3 Tests . . . 182
5.3.1 Test sur les paramètres de variance-covariance . . . 183
5.3.2 Test sur les paramètres de moyenne . . . 187
5.4 Analyse de la validité du modèle . . . 193
5.4.1 Analyse des résidus . . . 194
5.4.2 Effets aléatoires . . . 195
5.4.3 Adéquation du modèle par simulation . . . 195

6 Modèle mixte : Exemples **196**
6.1 Héritabilité . . . 196
6.1.1 Données et questions . . . 196
6.2 Aptitude à la dispersion des samares du frêne . . . 199
6.2.1 Présentation de l'expérience . . . 199
6.2.2 Analyse des composantes de la variance . . . 200
6.2.3 Répartition le long du cours . . . 204
6.3 Mesures répétées, nutrition humaine . . . 208
6.3.1 Problématique et données . . . 208
6.3.2 Analyse statistique . . . 209
6.4 Avalanches . . . 215
6.4.1 Problématique et données . . . 215
6.4.2 Analyse statistique . . . 216
6.4.3 Prise en compte d'effets aléatoires . . . 218
6.4.4 Programme R . . . 219

7 Plans d'expériences **222**
7.1 Pourquoi des plans d'expériences ? vocabulaire de base. . . . 222
7.2 Critères de qualité d'un plan . . . 225
7.2.1 Puissance d'un test . . . 225
7.2.2 Cas de plusieurs questions, plans optimaux . . . 228
7.3 Contrôle de l'hétérogénéité des unités expérimentales . . . 229
7.3.1 Bloc complet . . . 230
7.3.2 Carré Latin . . . 230
7.4 Plans d'expériences utilisant la confusion d'effets . . . 231
7.4.1 Définitions . . . 231
7.4.2 Confusion due au nombre trop petit de niveaux d'un ou de plusieurs facteurs de contrôle d'hétérogénéité . . . 235

7.4.3 Confusion d'effets pour traiter le cas d'un grand nombre d'unités de traitement, plans fractionnaires . . . 239
7.5 Plans d'expériences et modèle mixte . . . 247
7.5.1 Expériences à facteurs aléatoires . . . 247
7.5.2 Plans d'expériences avec effets bloc aléatoires : split-plot . . . 249
7.6 Plans pour les surfaces de réponse . . . 250
7.6.1 Introduction . . . 250
7.6.2 Plan composite centré . . . 252
7.6.3 Validation du modèle . . . 254
7.6.4 Solution optimale . . . 255
7.6.5 Plans Box-Behnken . . . 256
7.6.6 Pour aller plus loin . . . 257

8 Exemples de plans d'expériences 259
8.1 Blocs complets : 2 variétés de mais dans 8 lieux . . . 259
8.1.1 Plans d'expériences et données . . . 259
8.1.2 Analyse de l'expérience . . . 260
8.2 Bloc incomplet équilibré, Champagnes . . . 262
8.2.1 Présentation . . . 262
8.2.2 Analyse des résultats . . . 264
8.2.3 Résultats de l'analyse . . . 265
8.2.4 Programme SAS . . . 269
8.3 *Change-over design*, croissance de moutons . . . 270
8.3.1 Plan d'expériences et données . . . 270
8.3.2 Analyse des résultats . . . 271
8.4 Plan fractionnaire, fabrication de biscuit . . . 272
8.4.1 Plan d'expériences et données . . . 272
8.4.2 Analyse des résultats . . . 273
8.5 Blocs incomplets partiellement équilibrés, expression du génome de *Teleost* fish . . . 275
8.5.1 Plan d'expériences et données . . . 276
8.5.2 Analyse des résultats . . . 277
8.6 Plan fractionnaire Thermostat . . . 281
8.6.1 Plan d'expériences et données . . . 281
8.6.2 Analyse des résultats et conclusions . . . 282
8.7 Conception de produits robustes . . . 285
8.8 Plan fractionnaire puis surface de réponse, mesure de polluants . . . 286
8.8.1 Recherche de facteurs influents . . . 287
8.8.2 Optimisation des valeurs des facteurs . . . 288
8.9 Split-plot : effet secondaire d'un fongicide . . . 289
8.9.1 Plan d'expérience et données . . . 289
8.9.2 Analyse descriptive des données . . . 290
8.9.3 Analyse des résultats . . . 290
8.10 Plusieurs variables réponses . . . 294
8.11 Mélange, résistance de tissu . . . 298
8.11.1 Plan d'expériences et données . . . 298
8.11.2 Analyse des résultats . . . 299

8.12 programmes SAS et packages R (fin 2013) 300

9 Annexe1 : espaces euclidiens de dimension finie 301
9.1 Introduction 301
9.2 Sous espaces d'un espace vectoriel euclidien 301
9.2.1 Sous espaces orthogonaux d'un espace euclidien 301
9.2.2 Supplémentaire orthogonal d'un sous espace vectoriel F 301
9.3 Base orthonormée de $\mathbb{R}^n$ 302
9.3.1 Procédé d'orthonormalisation de Schmidt (ou Gram-Schmidt) . . 302
9.3.2 Matrice orthogonale 303
9.4 Projecteur orthogonal 303
9.4.1 Projecteur 303
9.4.2 Projecteur orthogonal 304

10 Annexe2 : lois normale multidimensionnelle, χ^2, Student, Fisher 306
10.1 Vecteurs aléatoires 306
10.2 Lois gaussiennes multidimensionnelles 307
10.2.1 Loi normale unidimensionnelle 307
10.2.2 Cas particulier d'un échantillon gaussien 307
10.2.3 Loi gaussienne multidimensionnelle, vecteur gaussien 307
10.2.4 Loi conditionnelle 308
10.2.5 Loi du χ^2 310
10.2.6 Loi de Student 310
10.2.7 Loi de Fisher 311
10.2.8 Loi du χ^2 décentrée 311

11 Annexe3 : loi des estimateurs 313
11.1 Théorème de Cochran 313
11.2 Loi des sommes de carrés du modèle linéaire 314
11.2.1 Décompositions de $\mathbb{R}^n$ associées au modèle linéaire 314
11.2.2 Loi de la somme des carrés résiduelle 315
11.2.3 Loi de la somme des carrés du modèle sous $H_0 = (\theta_1 = ... = \theta_s = 0)$ 315
11.2.4 Modèles emboîtés : loi de la différence entre les sommes de carrés des modèles. 315
11.3 Espérance et matrice de variance-covariance de $\widehat{\theta}$ 316
11.3.1 Espérance de $\widehat{\theta}$ 316
11.3.2 Matrice de variance-covariance de $\widehat{\theta}$ 317
11.4 Espérance de la somme des carrés résiduelle 318
11.5 Loi de $(\widehat{\theta}_1 - a)/\sqrt{\widehat{\mathbb{V}(\widehat{\theta}_1)}}$ 318
11.5.1 Loi sous H_0 318
11.5.2 Loi sous H_1 319
11.6 Loi de $\frac{(SCM_1 - SCM_0)/(p_1 - p_0)}{SCR_1/\nu_1}$ 319
11.6.1 Loi sous H_0 319
11.6.2 Loi sous H_1 319
11.7 Démonstration du théorème de Gauss-Markov 319
11.8 Démonstration du résultat (1.14, p. 24) 320

12 Annexe4 : algorithme de Newton Raphson **322**
12.1 Description de l'algorithme 322
12.2 Cas de la fonction de lien naturel 323

Index **323**

Bibliographie **327**

Chapitre 1

Le modèle linéaire

1.1 Introduction

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

$$Y_i = m_i + E_i$$

où

- i est le numéro de l'unité expérimentale,

  m_i est l'espérance de Y_i et inclut l'effet de variables explicatives,

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

Selon la nature des variables incluses dans la partie explicative m_i du modèle, on distingue trois grandes catégories de modèle linéaire :

- Lorsque les variables explicatives sont quantitatives, le modèle est appelé modèle de régression : simple s'il n'y a qu'une seule variable explicative, multiple sinon. Des exemples sont présentés dans le paragraphe 1.2.1, p. 13 et dans deux exemples détaillés dans les parties 2.1, p. 47 et 2.2, p. 53 du chapitre 2, p. 47.
- Lorsque les variables explicatives sont qualitatives, elles sont appelées **facteurs** et le modèle ainsi construit est un modèle d'analyse de la variance. Ce modèle est construit sur un exemple dans le paragraphe 1.2.2, p. 14 ci-dessous, puis étudié en détail dans les exemples 2.3, p. 62, 2.4, p. 70, 2.5, p. 82 du chapitre suivant.
- Lorsque les variables explicatives sont à la fois de nature quantitatives et qualitatives, le modèle ainsi construit est un modèle d'analyse de la covariance . Il est brièvement présenté dans le paragraphe 1.2.3, p. 15, puis étudié en détail au travers d'un exemple dans l'exemple présenté page 95.

1.2 Modélisation

1.2.1 Modèle de régression

Les brochets sont des prédateurs supérieurs qui cumulent l'ensemble des pesticides présents aux différents niveaux trophiques. Une étude cherche à comprendre si de plus, ils cumulent ces pesticides au cours de leur vie. Dans cet objectif, on souhaite quantifier le lien entre la concentration en DDT et l'âge, variable $x^{(1)}$. Un modèle de régression simple qui étudie le lien entre ces variables s'écrit :

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + E_i, \quad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

Ce modèle étant linéaire en ses paramètres, il peut se mettre sous une forme matricielle. Le vecteur $Y = (Y_1, \ldots, Y_n)'$ est le vecteur des variables à expliquer, le vecteur $E = (E_1, \ldots, E_n)'$ est le vecteur des erreurs résiduelles. Le vecteur de paramètres θ est défini par $\theta = (\beta_0, \beta_1)'$. Enfin la variable explicative et le terme constant sont stockés dans la matrice d'incidence, parfois appelée matrice de design, X, qui s'écrit donc

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_2^{(1)} \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n^{(1)} \end{bmatrix}.$$

Le **modèle de régression simple** s'écrit alors sous la forme

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

Si l'on souhaite inclure davantage de variables explicatives, on se trouve dans le cadre d'un modèle de régression linéaire multiple qui s'écrit

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)} + E_i, \quad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

En écrivant $\theta = (\beta_0, \beta_1, \ldots, \beta_p)'$ et

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(p)} \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(p)} \end{bmatrix},$$

on écrit **le modèle de régression multiple** sous sa forme matricielle

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

1.2.2 Modèle d'analyse de la variance (anova)

Pour comparer les rendements de cinq variétés de blé, 4 parcelles sont ensemencées pour chacune des cinq variétés étudiées, puis le rendement final y est mesuré. La variable explicative variété est qualitative, elle est souvent appelé facteur explicatif. Ce facteur possède cinq niveaux. Le modèle d'analyse de la variance à un facteur s'écrit alors, sous sa forme régulière :

$$Y_{ik} = \mu_i + E_{ik}, \quad i = 1, \dots I, k = 1, \dots n_i, \quad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \tag{1.1}$$

où i désigne le niveau i du facteur et k le numero de l'observation au sein de ce niveau i. I désigne le nombre total de niveaux de ce facteur, n_i le nombre d'observations pour le niveau i et $n = \sum_{i=1}^{I} n_i$ le nombre total d'observations. Dans le cas présent, on a $I = 5$, $n_i = 4,\ i = 1, \dots, 5$ et $n = 20$. Lorsqu'on veut dissocier un effet commun à toutes les variétés et un effet différentiel de chaque espèce par rapport à un comportement de référence, le modèle peut s'écrire sous sa forme singulière (forme singulière qui permettra une généralisation plus simple au cas à plus de deux facteurs) :

$$Y_{ik} = \mu + \alpha_i + E_{ik}, \quad E_{ik} \sim \mathcal{N}(0, \sigma^2). \tag{1.2}$$

Sous cette forme, le modèle possède un paramètre supplémentaire et n'est plus identifiable [1]. Ce problème est abordé dans le paragraphe 1.3.1, p. 21 de ce chapitre.

En posant

$$Y = (Y_{11} \dots, Y_{1n_1}, Y_{21} \dots, Y_{2n_2}, Y_{I1} \dots, Y_{In_I})',$$

$$E = (E_{11} \dots, E_{1n_1}, E_{21} \dots, E_{2n_2}, E_{I1} \dots, E_{In_I})',$$

puis $\theta = (\mu, \alpha_1, \dots, \alpha_I)'$,

et en notant $\mathbf{1}_{n_i}$ le vecteur de taille n_i ne contenant que des 1.

1. Un modèle est identifiable si pour deux jeux de paramètres différents θ_1 et θ_2 on a forcément que la loi des observations sous θ_1 est différente de la loi des observations sous θ_2 .

$$X = \begin{bmatrix}
1 & & 0 & 0 & 0 & 0 & 0 \\
\vdots & \mathbf{1}_{n_i} & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & & 0 & \vdots & \vdots & \vdots & \vdots \\
\vdots & 0 & & \vdots & \vdots & \vdots & \\
\vdots & \vdots & \mathbf{1}_{n_2} & \vdots & \vdots & \vdots & \\
\vdots & \vdots & & 0 & \vdots & \vdots & \\
\vdots & \vdots & 0 & & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \mathbf{1}_{n_3} & \vdots & \vdots & \\
\vdots & \vdots & \vdots & & 0 & \vdots & \\
\vdots & \vdots & \vdots & 0 & & \vdots & \\
\vdots & \vdots & \vdots & \vdots & \mathbf{1}_{n_4} & \vdots & \\
\vdots & \vdots & \vdots & \vdots & & 0 & \\
\vdots & \vdots & \vdots & \vdots & 0 & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \mathbf{1}_{n_5} & \\
1 & 0 & 0 & 0 & 0 & &
\end{bmatrix},$$

le **modèle d'analyse de la variance** se met sous la forme matricielle suivante :

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

1.2.3 Modèle d'analyse de la covariance (ancova)

Si dans l'expérience précédente, on veut prendre en compte une quantité d'azote x différente dans chacune des parcelles de l'expérience, il s'agit alors de proposer un modèle qui permet d'utiliser à la fois une variable quantitative et un facteur pour expliquer la variabilité du rendement et savoir si la réponse à l'azote est la même ou non pour toutes les variétés.
La forme régulière du modèle d'analyse de la covariance est donnée par

$$Y_{ik} = \mu_i + \beta_i x_{ik} + E_{ik}, \quad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

En écrivant $Y = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{In_I})'$,
$E = (E_{11}, \ldots, E_{1n_1}, E_{21}, \ldots, E_{I1}, \ldots, E_{In_I})'$, $\theta = (\mu_1, \ldots, \mu_I, \beta_1, \ldots, \beta_I)'$ et

$$X = \begin{bmatrix}
1 & 0 & \ldots & 0 & x_{11} & 0 & \ldots & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
1 & 0 & \ldots & 0 & x_{1n_1} & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 & 0 & x_{21} & \ldots & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{In_I}
\end{bmatrix},$$

le **modèle d'analyse de la covariance** se met alors sous la forme matricielle suivante :

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

Le modèle d'Ancova peut aussi s'écrire sous forme singulière :

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + E_{ik}, \quad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

1.2.4 Présentation unifiée du modèle linéaire

Ainsi quel que soit le modèle linéaire considéré et la nature des variables explicatives qui y sont incluses, l'écriture matricielle du modèle linéaire est :

$$Y = X\ \theta + E, \tag{1.3}$$

où

- Y, de dimension $(n, 1)$, contient les variables aléatoires représentant la variable à expliquer pour les n expériences. C'est un vecteur aléatoire.
- E, de dimension $(n, 1)$, contient les variables aléatoires résiduelles du modèle, rangées dans le même ordre que Y. Les E_i sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$ ou autrement dit le vecteur E, de dimension n suit une loi normale n-dimensionnelle centrée de matrice de variance $\sigma^2 I_n$.
- θ, de dimension $(p + 1, 1)$, contient $p + 1$ paramètres fixes et inconnus.
- X, de dimension $(n, p + 1)$, est une matrice (fixe et connue) contenant les valeurs des variables explicatives. La ligne i contient les variables explicatives concernant l'individu i, la colonne j contient lela variable explicative j pour les individus 1 à n. Dans le cas de facteurs qualitatifs, ces valeurs sont des 1 ou des 0. X s'appelle "matrice du plan d'expérience" ou "matrice de design".
- x_i est le vecteur ligne correspondant à la ième ligne de X.

Attention : l'aspect linéaire du modèle linéaire n'est pas aussi réducteur qu'on peut le penser, c'est la linéarité en chacun des paramètres qui est essentielle. Ainsi, le modèle $Y = \theta_0 + \theta_1 x + \theta_2 x^2 + E$ est encore un modèle linéaire. "Modèle linéaire" signifie que $\mathbb{E}[Y]$ est une combinaison linéaire des paramètres du modèle et les coefficients de ces combinaisons sont quelconques.

1.3 Estimation des paramètres

Une fois le modèle posé, la question qui se pose ensuite est l'estimation des paramètres inconnus du modèle. Les paramètres sont de deux sortes, ceux qui relèvent de l'espérance et sont contenus dans le vecteur θ, et le paramètre σ^2 qui mesure la variabilité qui subsiste lorsque l'on a enlevé à la variabilité totale des observations tout ce qui est expliqué par le modèle. Il met donc en jeu $p + 1$ paramètres pour l'espérance (p pour l'effet des p variables ou des p niveaux et 1 pour la constante) et 1 paramètre pour la variance (σ^2).
La méthode d'estimation des paramètres classiquement utilisée est la méthode du maximum de vraisemblance. Les variables aléatoires Y_i ayant été supposées indépendantes

et de loi gaussienne, la vraisemblance de l'échantillon $y = (y_1, \ldots, y_n)$ s'écrit :

$$\mathcal{L}(y;\theta,\sigma^2)\,^2 = \prod_{i=1}^{n} f(y_i;\theta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i - x_i\theta)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{\sum_{i=1}^{n}(y_i - x_i\theta)^2}{2\sigma^2}},$$

et la log-vraisemblance

$$\ell(y;\theta,\sigma^2) \quad = \quad -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\theta)^2\,.$$

Les valeurs de θ et σ^2 qui maximisent cette log-vraisemblance sont solutions du système d'équations aux dérivées partielles suivant :

$$\begin{cases} \dfrac{\partial \ell(y;\theta_j,\sigma^2)}{\partial \theta_j} = 0 \quad \text{pour } j = 0, \ldots, p \\ \dfrac{\partial \ell(y;\theta,\sigma^2)}{\partial \sigma^2} = 0 \end{cases} \tag{1.4}$$

Remarquons que la maximisation de la log-vraisemblance en θ peut se faire indépendamment de σ^2. En effet, maximiser la log-vraisemblance en θ revient à minimiser $\sum_{i=1}^{n}(y_i - x_i\theta)^2$ qui peut s'écrire sous la forme suivante :

$$\|y - X\theta\|^2,$$

où $\|.\|^2$ correspond à la norme euclidienne de $\mathbb{R}^n$: si u est un vecteur de $\mathbb{R}^n$, $u = (u_1, \ldots, u_n)$, la norme de u vaut $\|u\|^2 = \sum_{i=1}^{n} u_i^2$. D'un point de vue géométrique, cette norme s'interprète comme la distance séparant l'oginie d'un repère O et le point U le point de coordonnées $(u_1, \ldots, u_n)$.

Ainsi nous nous intéresserons dans un premier temps à l'estimation de θ (les paramètres de l'espérance) puis à celle de la variance σ^2.

Dans la suite, pour simplifier l'écriture, l'estimateur et l'estimation des paramètres θ seront notés de la même façon $\widehat{\theta}$.

1.3.1 Estimation des paramètres de l'espérance θ

Cette méthode d'estimation est connue sous le nom de méthode des moindres carrés ordinaires (MCO) et l'estimateur résultant porte alors le nom d'estimateur des moindres carrés. Remarquons que dès que la distribution est supposée gaussienne, la méthode du maximum de vraisemblance est équivalente à la méthode des moindres carrés pour l'estimation du paramètre de la moyenne. L'estimateur $\widehat{\theta}$ est tel qu'il rend

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

2. La vraisemblance dépend également de $x = (x_1, \ldots, x_n)$ mais dans toute la suite on travaille conditionnellement aux variables explicatives et cette dépendance sera donc omise dans toutes les notations.

Théorème 1.3.1. *Soit $< X >$ le sous-espace linéaire de $\mathbb{R}^n$ engendré par les vecteurs colonnes de la matrice X. L'estimateur des moindres carrés du paramètre θ, noté $\widehat{\theta}$, est tel que :*

$$\begin{aligned} \widehat{Y} &= PY \quad \textit{est le projeté orthogonal de } Y \textit{ sur } < X > \\ &= X\widehat{\theta}, \end{aligned}$$

où P est le projecteur orthogonal sur $< X >$. L'estimateur $\widehat{\theta}$ vérifie le système

$$X'X\widehat{\theta} = X'Y, \tag{1.5}$$

où X' est la transposée de la matrice X. Ce système s'appelle traditionnellement ***système des équations normales****.*

Ce système correspond exactement au système des dérivées partielles (1.4).

Preuve théorème 1.3.1. $< X >$ est le sous-espace linéaire de $\mathbb{R}^n$ engendré par les vecteurs colonnes de la matrice X, c'est-à-dire que tout élément de $< X >$ s'écrit comme une combinaison linéaire de ces vecteurs. D'après la proposition 9.4.3 de l'Annexe 9, p. 301, la projection orthogonale sur $< X >$ minimise l'écart entre n'importe quel élément de $< X >$ et Y c'est-à-dire $\|Y - X\widehat{\theta}\|^2 = min_{U \in < X >} \|Y - U\|^2$. Cette projection est illustrée dans la Figure 1.1. Pour obtenir une forme explicite de $\widehat{\theta}$, il suffit d'écrire les relations d'orthogonalité

$$\forall k = 1, .., p+1 \qquad \left\langle X^k, Y - X\widehat{\theta} \right\rangle = X^{k\prime}(Y - X\widehat{\theta}) = 0,$$

où X^k désigne la colonne k de X. Cette égalité s'écrit plus synthétiquement

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

□

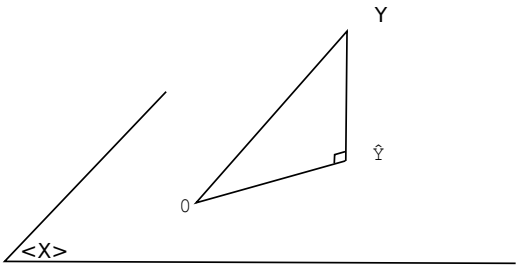


FIGURE 1.1 – Représentation de la projection de Y sur $< X >$.

On ne peut résoudre le système des équations normales (1.5, p. 18) que si la matrice carrée $X'X$ de dimension $(p+1) \times (p+1)$ admet une inverse $(X'X)^{-1}$. Or cette dernière n'existe que si la matrice $X'X$ est de plein rang, c'est-à-dire que son rang vaut exactement $p+1$. Noter que le rang de la matrice $X'X$ est le même que celui de X et

rappelons qu'il est égal au nombre de colonnes de la matrice X qui sont linéairement indépendantes (qui ne sont pas combinaisons linéaires des autres colonnes). Notons

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

Cela revient à dire que le système (1.5, p. 18) est un système linéaire de r équations indépendantes à $p+1$ inconnues. Si r est strictement inférieur à $p+1$, le système ne peut pas avoir une solution unique.

Dans les deux sous-sections suivantes, nous distinguons les cas où X est de plein rang ou non et donnons la forme de la matrice du projecteur P associée.

Cas où $r = p+1$

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

$$\widehat{\theta} = (X'X)^{-1}X'Y. \tag{1.6}$$

On a donc que

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

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

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

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

Remarquons que dans un modèle de régression multiple , les fortes corrélations entre variables peuvent rendre la matrice $X'X$ difficile à inverser, pour des raisons d'instabilité numérique (liée au mauvais conditionnement de la matrice $X'X$). Dans ce cas une approche possible consiste à choisir certaines variables comme représentantes du groupe des varaibles avec lesquelles elles sont fortements corrélées et à travailler uniquement avec ces représentantes.

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

Ce cas correspond aux situations où les vecteurs colonnes de la matrice X sont liées par $p+1-r$ relations linéaires indépendantes. C'est le cas des modèles d'analyse de la variance et d'analyse de la covariance écrits sous leurs formes singulières. Dans ce cas, la matrice $X'X$ n'est pas inversible et donc il existe une infinité de solutions $\widehat{\theta}$ vérifiant le système. Nous sommes dans un cas d'indétermination du système. La stratégie consiste alors à choisir, parmi cette infinité de solutions, une solution particulière en ajoutant des contraintes sur les paramètres de θ. Le nombre de contraintes indépendantes à ajouter est égale au nombre d'équations indépendantes manquantes dans le système. Cependant, les contraintes choisies donneront un sens particulier aux estimations des

paramètres qu'il faudra utiliser pour l'interprétation.
En général, on ne s'intéresse pas directement aux estimations de ces paramètres (notamment car ils n'ont pas de sens intrinsèque) mais plutôt à des combinaisons linéaires invariantes de ces paramètres (qui ne dépendent pas des contraintes choisies).

On introduit donc des contraintes linéaires sur les paramètres :

$$\underset{(h\times(p+1))}{H}\quad\underset{((p+1)\times 1)}{\theta}\quad=\quad 0$$

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

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

dont l'écriture condensée est :

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

La méthode des moindres carrés mène alors au système suivant :

$$G'G\widehat{\theta}=X'X\ \widehat{\theta}+H'H\widehat{\theta}=X'X\widehat{\theta}=X'Y. \tag{1.7}$$

Puisque $G'G$ est de plein rang (construite pour cela), elle est inversible et on obtient la solution suivante :

$$\widehat{\theta}=(G'G)^{-1}X'Y. \tag{1.8}$$

La matrice $(G'G)^{-1}$ est appelée une "inverse généralisée" de $X'X$ et se note $(X'X)^{-}$. Le projecteur orthogonal P est dans ce cas égal à :

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

Remarque 1. Si on change la matrice H, c'est-à-dire les contraintes, l'expression des estimateurs changent. Il est donc important de toujours revenir à l'expression de ces estimateurs pour interpréter correctement les estimations associées. Par contre, ce n'est pas le cas du projecteur P qui lui ne change pas, cela parce que le sous-espace $<X>$ sur lequel on projette ne dépend pas des contraintes. En particulier, puisque $\widehat{Y}=PY$, la prédiction par le modèle de Y ne dépend pas du système de contraintes choisies.

Remarque 2. Le choix des contraintes peut se faire selon deux objectifs : faciliter les calculs numériques ou donner un sens particulier aux paramètres.

Illustration dans le cadre d'un modèle d'Anova à 1 facteur

Le modèle d'analyse de la variance à 1 facteur est donné sous sa forme régulière par l'équation (1.1) et sous sa forme singulière par l'équation (1.2). Sous sa forme régulière, la version matricielle du modèle est donné par :

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & 0 & \dots & 0 \\ 1 & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_I \end{bmatrix} + \begin{bmatrix} E_{11} \\ \vdots \\ E_{1n_1} \\ \vdots \\ E_{I1} \\ \vdots \\ E_{In_I} \end{bmatrix} \tag{1.9}$$

La matrice X est de plein rang $r = I$, la matrice $X'X$ est inversible (si tant est que tous les niveaux sont bien représentés, c'est-à-dire que pour tout $n_i > 0$)et à partir de l'expression de $\widehat{\theta}$ donnée par l'équation (1.6), on obtient les estimateurs des moyennes suivantes :

$$\widehat{\mu}_i = Y_{i\bullet} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad i = 1, \dots, I,$$

et les estimations :

$$\widehat{\mu}_i = y_{i\bullet} := \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad i = 1, \dots, I.$$

Dans l'écriture singulière de ce modèle, la matrice X ainsi que le paramètre de moyenne θ changent :

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & 0 & \dots & 0 \\ 1 & 1 & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \\ 1 & 0 & \dots & 1 & \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{bmatrix} + \begin{bmatrix} E_{11} \\ \vdots \\ E_{1n_1} \\ \vdots \\ E_{I1} \\ \vdots \\ E_{In_I} \end{bmatrix} \tag{1.10}$$

Dans ce cas, la matrice X n'est plus de plein rang (la première colonne étant la somme des I suivantes, la matrice est de rang $r = I$, mais possède $I + 1$ colonnes). Il faut donc imposer une contrainte sur les paramètres afin de lever le problème d'identifiabilité [3]. Plusieurs contraintes sont possibles. Dans la plupart des logiciels, celle utilisée par défaut consiste en l'annulation d'une des modalités, appelée modalité de référence. Prenons l'exemple du logiciel SAS qui choisit de prendre la dernière modalité comme référence, soit à poser $\alpha_I = 0$. Cette contrainte est particulièrement intéressante dans

3. Voir Note 1

le cas où cette modalité sert de témoin dans l'expérience. Cette contrainte s'écrit sous la forme $H\theta = 0$ où la matrice H de dimension $1 \times (I+1)$ est :

$$H = [0\ 0\ \ldots\ 1]\,,$$

rendant la matrice G de rang $I+1$. Si on écrit le système aux équations normales associées (1.7, p. 20), on obtient le système :

$$\begin{cases} n\widehat{\mu} + \sum_i n_i \widehat{\alpha}_i = \sum_{i,j} y_{ij} \\ n_i\widehat{\mu} + n_i\widehat{\alpha}_i = \sum_j y_{ij} \quad \text{pour } i = 1, \ldots, I-1 \\ n_I\widehat{\mu} + (n_I+1)\widehat{\alpha}_I = \sum_j y_{Ij}, \end{cases}$$

ce qui revient à écrire le système aux dérivées partielles (1.4, p. 17) dédiées à θ auquel est ajouté la contrainte. On obtient alors :

$$\begin{cases} \widehat{\mu} = y_{I\bullet} = \dfrac{1}{n_I}\sum_j y_{Ij} \\ \widehat{\alpha}_i = y_{i\bullet} - y_{I\bullet} \quad \text{pour } i = 1, \ldots, I-1 \\ \widehat{\alpha}_I = 0 \end{cases}$$

Une autre façon (plus simple) d'obtenir ce résultat est de les déduire à partir des estimations du modèle régulier. En effet, pour passer de l'un à l'autre, on a posé que $\mu_i = \mu + \alpha_i$ et on a obtenu que $\widehat{\mu}_i = y_{i\bullet}$. On obtient alors

$$\begin{cases} \widehat{\mu}_i = \widehat{\mu} + \widehat{\alpha}_i \quad \text{pour } i = 1, \ldots, I-1 \\ \widehat{\mu_I} = \widehat{\mu} + \widehat{\alpha}_I = \widehat{\mu} \end{cases}$$

Ainsi si la contrainte $\alpha_I = 0$ est choisie, $\widehat{\mu}$ s'interprétera comme la réponse moyenne du groupe I et $\widehat{\alpha}_i$ comme des écarts de réponses moyennes entre le groupe i et le groupe I de référence.

Remarquons que si c'est la modalité 1 qui est prise comme référence (comme c'est le cas par défaut dans le logiciel R), les estimations seront différentes :

$$\begin{cases} \widehat{\mu} = y_{1\bullet} \\ \widehat{\alpha}_i = y_{i\bullet} - y_{1\bullet} \quad \text{pour } i = 2, \ldots, I \\ \widehat{\alpha}_1 = 0 \end{cases}$$

Leur interprétation en sera donc tout aussi différente.

Combinaisons linéaires estimables

Une combinaison linéaire des paramètres θ s'écrit

$$\underset{1\times 1}{\psi} \quad = \quad \sum_{j=0}^{p} c_j\,\theta_j \quad = \quad \underset{1\times(p+1)}{\mathrm{C}} \quad \underset{(p+1)\times 1}{\theta}$$

Lorsque la combinaison linéaire est telle que $\sum_{j=0}^{p} c_j = 0$, elle porte le nom spécifique de **contraste**. Ces contrastes apparaissent souvent dans le contexte des modèles d'analyse de la variance. En effet tester l'égalité entre deux niveaux i et i' d'un facteur dans une analyse de variance à un facteur comme présentée dans le modèle 1.2, p. 14, revient à tester, $\alpha_i = \alpha_{i\prime}$ ou encore $C\theta = 0$ avec $C_j = 0$ pour $j \notin \{i, i\prime\}$, $C_i = 1$ et $C_{i\prime} = -1$. Ainsi tester l'égalité de deux niveaux au sein d'un même facteur, revient à tester la nullité d'un contraste particulier.

On peut généraliser à tout vecteur ψ de dimension q de combinaisons linéaires des paramètres θ :

$$\begin{array}{ccccc} \psi & = & C & & \theta \\ q \times 1 & & q \times (p+1) & & (p+1) \times 1 \end{array}$$

Définition 1.3.1. Une combinaison linéaire des paramètres, $C\theta$, est estimable *s'il existe K tel que $C = KX$. On a alors $C\theta = \mathbb{E}(KY)$, c'est-à-dire qu'il existe un estimateur (sans biais) de $C\theta$.*

1.3.2 Propriétés de l'estimateur $\widehat{\theta}$

Espérance et matrice de variance

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

Théorème 1.3.2. *L'espérance et la matrice de variance-covariance de l'estimateur des moindres carrés des paramètres θ du modèle (1.3, p. 16) (dont l'expression est donnée par l'équation (1.6) dans le cas où X est de plein rang et par (1.8) sinon) valent respectivement*

$$\mathbb{E}[\widehat{\theta}] = \theta.$$

$\widehat{\theta}$ est donc un estimateur sans biais de θ, y compris lorsque $\widehat{\theta}$ est défini à partir d'un inverse généralisé et ce quel que soit cet inverse généralisé.

$$\mathbb{V}[\widehat{\theta}] = (X'X)^{-1}\sigma^2 \qquad si \qquad r = p+1, \tag{1.11}$$

ou

$$\mathbb{V}[\widehat{\theta}] = (X'X)^{-}X'X(X'X)^{-}\sigma^2 \qquad si \qquad r < p+1. \tag{1.12}$$

On peut détailler ce dernier résultat : notons $V = (X'X)^{-1}$ si la matrice X est de plein rang et $V = (X'X)^{-}X'X(X'X)^{-}$ si la matrice n'est pas de plein rang. v_{lj} désigne le terme général de cette matrice, $l, j = 1, \ldots, p+1$. On a alors

$$\mathbb{V}[\widehat{\theta}_j] = v_{jj}\sigma^2 \quad \text{et} \quad \mathbb{C}\text{ov}(\widehat{\theta}_j, \widehat{\theta}_l) = v_{jl}\sigma^2$$

On trouvera la démonstration de ce théorème en Annexe 11.3, p. 316.

Discussion sur la variance de $\widehat{\theta}$

La matrice de variance-covariance $\mathbb{V}[\widehat{\theta}]$ contient sur la diagonale les variances des estimateurs des paramètres θ_j et hors de la diagonale les covariances de ces estimateurs deux à deux. C'est donc une matrice symétrique. La formule de la variance des estimateurs permet d'apprécier la précision de ces estimateurs. Cette variance dépend de $(X'X)^{-1}$ (ou $(X'X)^{-}X'X(X'X)^{-}$) (et de σ^2), ainsi la précision des estimateurs dépendra de la qualité du plan d'expérience. Les plans d'expériences optimaux (voir partie 7.2.2, p. 228 du chapitre sur les plans d'expérience) cherchent à minimiser cette variance.

Regardons ce que l'on obtient dans le cas de deux modèles : un modèle d'anova à 1 facteur écrit sous sa forme régulière et un modèle de régression multiple .

Anova à 1 facteur. Si on suppose que le facteur n'a que deux modalités avec n_i observations pour le modalité i, alors la matrice $\mathbb{V}[\widehat{\theta}] = \mathbb{V}[(\widehat{\mu}_1\ \widehat{\mu}_2)']$ est une matrice de dimension 2×2 qui vaut

$$\begin{bmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{bmatrix}$$

La précision de l'estimateur de la moyenne μ_i est logiquement inversement proportionnelle au nombre d'observations pour cette modalité : plus il y a d'observations, plus l'estimation sera précise.

Modèle de régression multiple . La corrélation partielle permet de mesurer le lien qui peut exister entre deux variables, les autres étant contrôlées. Si on considère 3 variables U, V et W, le coefficient de corrélation partielle entre U et V conditionnellement à W, noté $cor(U, V|W)$, est la corrélation simple entre U et V, notée $cor(U, V)$, étant donné l'effet linéaire de W enlevé. En d'autres termes, si on régresse U et V par W, $U = a_1 + b_1 W + E$ et $V = a_2 + b_2 W + F$, alors on obtient que $cor(U, V|W) = cor(E, F)$. Son expression est :

$$cor(U, V|W) = \frac{cor(U, V) - cor(U, W)cor(V, W)}{\sqrt{(1 - cor(U, W)^2)(1 - cor(V, W)^2)}}. \tag{1.13}$$

On peut montrer que la matrice de variance-covariance des estimateurs des paramètres $\beta_1, \beta_2, \ldots, \beta_p$ est liée à la corrélation partielle emprique entre les variables associées. Plus précisément, on a que

$$cor(\widehat{\beta}_i, \widehat{\beta}_j) = -\widehat{cor}(X^{(i)}, X^{(j)}|X_{\setminus(i,j)}), \tag{1.14}$$

où $cor(X^{(i)}, X^{(j)}|X_{\setminus(i,j)})$ désigne la corrélation partielle empirique entre les variables $X^{(i)}$ et $X^{(j)}$ sachant toutes les autres. La démonstration de ce résultat est donnée en Annexe 11.8, p. 320. Ainsi, plus les variables sont corrélées partiellement, plus les estimateurs des paramètres associés le seront aussi (mais de signe opposé), ce qui rend leur interprétation plus que délicate.

Optimalité de $\widehat{\theta}$

Une fois un estimateur défini, on s'intéresse naturellement à savoir s'il est optimal au sens où c'est celui qui est le plus précis (i.e. qui a la plus petite variance parmi la classe d'estimateurs considérée). Le théorème de Gauss-Markov donné ci-dessous indique que

l'estimateur des moindres carrés $\widehat{\theta}$ est optimal parmi tous les estimateurs linéaires sans biais. Cette propriété est vraie même si la loi des erreurs n'est pas gaussienne.

Théorème 1.3.3. *(Gauss-Markov) $\widehat{\theta}$ est le meilleur estimateur linéaire sans biais de θ au sens suivant :*

$$\forall \tilde{\theta},\ \forall C,\ \mathbb{V}(C\widehat{\theta}) \leq \mathbb{V}(C\tilde{\theta}),$$

où C est un vecteur $(1, p+1)$ et $\tilde{\theta}$ un estimateur linéaire sans biais de θ.

2 cas particuliers s'en déduisent :
- $\forall j \in (1, p+1)$, $\mathbb{V}(\theta_j)$ est minimale parmi les estimateurs linéaires sans biais,
- pour toute combinaison linéaire estimable des paramètres $C\theta$, $C\widehat{\theta}$ est l'estimateur linéaire sans biais de variance minimale.

Le paramètre estimé (et son estimateur) peuvent dépendre des contraintes si $r \leq p+1$, alors que la combinaison linéaire estimable n'en dépend pas.

La démonstration de ce résultat est donnée en Annexe 11.7, p. 319.

Loi de l'estimateur $\widehat{\theta}$

Dans les deux cas ($r = p+1$ ou $r < p+1$), les estimateurs des paramètres sont des estimateurs linéaires, c'est-à-dire qu'ils sont des combinaisons linéaires de Y qui est un vecteur gaussien ($Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$). Il en est donc de même pour le vecteur des estimateurs $\widehat{\theta}$ (cf Annexe 10, p. 306).

D'après ce qui précède, on a donc que

$$\widehat{\theta} \sim \mathcal{N}_{p+1}(\theta, \mathbb{V}[\widehat{\theta}]), \tag{1.15}$$

où l'expression de $\mathbb{V}[\widehat{\theta}]$ est donnée par (1.11) ou (1.12).

On peut en extraire la loi d'un $\widehat{\theta}_j$ particulier :

$$\widehat{\theta}_j \sim \mathcal{N}(\theta_j, v_{jj}\sigma^2), \tag{1.16}$$

où v_{jj} est défini dans le paragraphe précédent.

Loi de l'estimateur d'une combinaison linéaires de θ

Si la quantité d'intérêt s'exprime comme une combinaison linéaire des paramètres ψ

$$\psi = C\theta \ \text{ d'estimateur } \ \hat{\psi} = C\widehat{\theta},$$

où C est un vecteur ligne de dimension $p+1$, alors

$$\begin{aligned} \mathbb{E}[\hat{\psi}] &= C\mathbb{E}[\widehat{\theta}] = C\theta, \\ \mathbb{V}[\hat{\psi}] &= C\mathbb{V}[\widehat{\theta}]C', \end{aligned}$$

et

$$\hat{\psi} \sim \mathcal{N}(\psi, C\mathbb{V}[\widehat{\theta}]C').$$

Prédiction

Connaissant $\widehat{\theta}$, la prédiction par le modèle de Y_i est obtenue simplement par l'estimation de son espérance :

$$\widehat{Y}_i = \hat{\mathbb{E}}[Y_i] = x_i\widehat{\theta},$$

de valeur $\hat{y}_i = x_i\widehat{\theta}$ (où $\widehat{\theta}$ est ici l'estimation de θ).

De cela, on peut en déduire un prédicteur de l'erreur pour l'observation i, E_i : en effet, une valeur naturelle est l'écart entre Y_i et son prédicteur :

$$\hat{E}_i = Y_i - \widehat{Y}_i.$$

$\hat{E}_i$ est une variable aléatoire dont une réalisation est $\hat{e}_i = y_i - \hat{y}_i$. Plus généralement,

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

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

$\widehat{E}$ est souvent appelé vecteur des résidus et par abus la réalisation de ce vecteur est également appelée vecteurs de résidus. Les résidus désignent donc à la fois les prédicteurs de E et leurs prédictions.

1.3.3 Estimation de la variance σ^2 et propriétés des estimateurs

Revenons sur le système aux dérivées partielles (1.4, p. 17) permettant d'obtenir les estimateurs de θ et de σ^2 du maximum de vraisemblance. Connaissant l'estimation de θ, la vraisemblance sera maximale si :

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\theta}\right)^2 = 0 \quad \Leftrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\theta}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Notons

$$SCR = (Y - \widehat{Y})'(Y - \widehat{Y}) = \widehat{E}'\widehat{E} = \sum_{t=1}^{n}\hat{E}_i^2.$$

Cette somme est appelée "Somme des Carrés Résiduelle " car c'est la somme des carrés des résidus . L'estimateur du maximum de vraisemblance de σ^2, noté S^2_{MV}, est donc donné par :

$$S^2_{\text{MV}} = \frac{SCR}{n}.$$

Si on utilise les hypothèses faites sur les variables aléatoires $\boldsymbol{E}$, on peut montrer que

$$\mathbb{E}[SCR] = (n - r)\sigma^2, \tag{1.17}$$

où r est le rang de la matrice X, i.e. la dimension de l'espace $< X >$. La démonstration est donnée en Annexe 11.4, p. 318. On peut comprendre géométriquement la raison du terme $n - r$: Y évolue dans un espace de dimension n alors que $\widehat{Y}$ appartient à l'espace $< X >$ de dimension r, le rang de la matrice X. $\widehat{E}$ appartient donc à un espace de

dimension $n-r$.
Ainsi, l'estimateur S^2_{MV} est un estimateur biaisé puisque $\mathbb{E}[S^2_{\text{MV}}] = \frac{n-r}{n}\sigma^2$. On en tire facilement un estimateur non biaisé S^2 :

$$S^2 = \frac{SCR}{n-r} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-r}. \tag{1.18}$$

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

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-r}.$$

Par le théorème de Cochran (donné en Annexe 11.1, p. 313), on peut montrer que :

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

De plus S^2 et $\widehat{\theta}$ sont des vecteurs aléatoires indépendants (cf démonstration du théorème 11.1, p. 313).

1.3.4 Intervalles de confiance

La connaissance de la loi de l'estimateur $\widehat{\theta}$ permet de construire des intervalles de confiance sur les paramètres θ ou sur une combinaison linéaire de ces paramètres $\psi = C\theta$.

Intervalle de confiance sur un paramètre θ_j. D'après la loi de $\widehat{\theta}_j$ donnée par l'équation (1.16, p. 25), on a que

$$\frac{\widehat{\theta}_j - \theta_j}{\sigma\sqrt{v_{jj}}} \sim \mathcal{N}(0,1).$$

La variance σ^2 étant inconnue, on la remplace par son estimateur. Comme $\frac{(n-r)S^2}{\sigma^2} \sim \chi^2(n-r)$ et est indépendant de $\widehat{\theta}$, on obtient :

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{S^2 v_{jj}}} \sim \mathcal{T}_{n-r},$$

où $\mathcal{T}_{n-r}$ est la loi de Student à $n-r$ degrés de liberté (degrés de liberté de la loi de l'estimateur S^2). Ainsi l'intervalle de confiance symétrique de niveau α de θ_j est donné par

$$IC_{1-\alpha}(\theta_j) = \left[\hat{\theta}_j - t_{1-\alpha/2,n-r}S\sqrt{v_{jj}}, \hat{\theta}_j + t_{1-\alpha/2,n-r}S\sqrt{v_{jj}}\right],$$

où $t_{1-\alpha/2,n-r}$ représente le quantile d'ordre $1-\alpha/2$ de la loi de Student à $(n-r)$ degrés de liberté. On confond souvent cet intervalle de confiance avec sa réalisation, qui n'est elle plus aléatoire :

$$IC_{1-\alpha}(\theta_j) = \left[\hat{\theta}_j - t_{1-\alpha/2,n-r}\hat{\sigma}\sqrt{v_{jj}}, \hat{\theta}_j + t_{1-\alpha/2,n-r}\hat{\sigma}\sqrt{v_{jj}}\right],$$

Intervalle de confiance d'une combinaison linéaire des paramètres. De la même façon, on peut construire l'intervalle de confiance symétrique de niveau α d'une combinaison linéaire des paramètres $\psi = C\theta$:

$$IC_{1-\alpha}(\psi) = \left[\widehat{\psi} - t_{1-\alpha/2,n-r}\sqrt{\widehat{\mathbb{V}}[\widehat{\psi}]}, \widehat{\psi} + t_{1-\alpha/2,n-r}\sqrt{\widehat{\mathbb{V}}[\widehat{\psi}]}\right],$$

où $\widehat{\mathbb{V}}[\widehat{\psi}] = C\widehat{\mathbb{V}}[\widehat{\theta}]C'$ avec $\widehat{\mathbb{V}}[\widehat{\theta}]$ la variance de $\widehat{\theta}$ pour laquelle σ^2 a été remplacée par son estimateur. A nouveau, on s'intéresse le plus souvent à la réalisation de cet intervalle de confiance.

Prenons l'exemple de $\psi = \alpha_1 - \alpha_2$ dans le cadre du modèle anova à 1 facteur. Son estimateur est $\widehat{\psi} = \hat{\alpha}_1 - \hat{\alpha}_2 = Y_{1\bullet} - Y_{2\bullet}$. De plus, $\mathbb{V}[\widehat{\psi}] = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$ et $\widehat{\mathbb{V}}[\widehat{\psi}] = S^2(\frac{1}{n_1} + \frac{1}{n_2})$. L'intervalle de confiance de ψ au niveau α $IC_{1-\alpha}(\psi)$, est :

$$\left[(Y_{1\bullet} - Y_{2\bullet}) - t_{1-\alpha/2,n-r}S\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}, (Y_{1\bullet} - Y_{2\bullet}) + t_{1-\alpha/2,n-r}S\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}\right].$$

1.4 Tests d'hypothèses

Dans toute la partie qui va suivre, M désignera un modèle $Y = X\theta + E$, r le rang de la matrice X (contenant le vecteur constant $\mathbf{1}$), et le vecteur des paramètres θ est $\theta = (\theta_0, \ldots, \theta_p)'$.

1.4.1 Tests sur les paramètres

La loi des estimateurs obtenue dans le paragraphe précédent permet de construire un test sur un paramètre mais également, puisqu'on a la loi conjointe des estimateurs, il est possible de construire un test sur une combinaison linéaire de paramètres. La signification d'un paramètre peut dépendre du système de contraintes choisies. De ce fait, un test portant sur ce paramètre s'interprétera en conséquence. Cet aspect est détaillé dans le chapitre d'exemples, en particulier dans le cas de l'analyse de la variance.

Test sur un seul paramètre

Pour tester l'égalité d'un paramètre à une valeur a définie a priori, le test présenté dans la proposition suivante peut être utilisé. Le plus souvent a vaut 0 comme par exemple dans une régression multiple pour savoir si la variable X_j est utile pour prédire Y.

Proposition 1.4.1. *Soit H_0, l'hypothèse nulle du test définie par $H_0 = \{\theta_j = a\}$. Sous l'hypothèse H_0*

$$\frac{\widehat{\theta}_j - a}{\sqrt{S^2 v_{jj}}} \sim \mathcal{T}_{n-r}. \tag{1.19}$$

Démonstration. D'après l'équation (1.16, p. 25), sous H_0, $\widehat{\theta}_j - a$ suit une loi normale centrée de variance $\sigma^2 v_{jj}$. Ainsi $\frac{\widehat{\theta}_j - a}{\sqrt{\sigma^2 v_{jj}}} \overset{H_0}{\sim} \mathcal{N}(0,1)$. Puisque SCR est indépendant de $\widehat{\theta}$

et SCR/σ^2 suit une loi du χ^2 à $n-r$ degrés de liberté,

$$\frac{\widehat{\theta}_j - a}{\sqrt{v_{jj}\frac{SCR}{n-r}}} \overset{H_0}{\sim} \mathcal{T}_{n-r}.$$

□

Test d'une combinaison linéaire des paramètres

En analyse de la variance, le test précédent est peu utile (voire dangereux car il peut être mal interprété, si on ne prend pas en compte les contraintes utilisées). Il est souvent plus intéressant de tester l'hypothèse nulle $H_0 = \{C\theta = a\}$ où $C\theta$ est une combinaison linéaire invariante et a est une valeur définie a priori.

Proposition 1.4.2. *Soit H_0, l'hypothèse nulle du test définie par $H_0 = \{C\theta = a\}$, où $C\theta$ est une combinaison linéaire des paramètres. Sous l'hypothèse H_0,*

$$\frac{C\widehat{\theta} - a}{\sqrt{C\widehat{\mathbb{V}}(\hat{\theta})C'}} \sim \mathcal{T}(n-r). \tag{1.20}$$

Démonstration. La preuve de cette proposition est similaire à celle de la proposition 1.4.1, p. 28. D'après l'équation (1.15, p. 25), $C\widehat{\theta}$ suit une loi normale centrée sur $C\theta$ et de variance $\sigma^2 CVC'$. Il s'en suit immédiatement le résultat, puisque σ^2 est remplacée par son estimateur $SCR/(n-r)$. □

Exemple : Dans le modèle $Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$ ($i = 1,2$ et $j = 1,2$, $k = 1,\ldots,n_{ij}$), on veut savoir s'il y a une différence entre les niveaux 1 et 2 du premier traitement dont les effets sont notés α_1 et α_2. On est donc amené à tester $H_0 = \{\alpha_1 - \alpha_2 = 0\}$. On a $\theta' = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2)'$, $C = [0,1,-1,0,0]$ et $a = 0$. $C\theta$ est une combinaison linéaire invariante, le test proposé ne dépend pas du système de contraintes choisi.

1.4.2 Tests de modèles emboîtés

Le pouvoir explicatif d'un modèle est examiné en testant ce modèle contre des versions simplifiées de lui-même de manière à identifier des variables réellement pertinentes. Cette procédure s'effectue au travers du test de modèles emboîtés décrit ici. Notons M_q et M_r deux modèles dont les matrices de design sont $X^{(q)}$ et $X^{(r)}$ respectivement de rangs q et r.

Définition 1.4.1. *M_q est emboîté dans M_r si et seulement si le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $X^{(q)}$ est contenu dans le sous-espace vectoriel engendré par les colonnes de $X^{(r)}$.*

Autrement dit M_q est emboîté dans M_r si et seulement si M_q est un cas particulier de M_r, obtenu en annulant certains paramètres ou certaines combinaisons linéaires de paramètres de M_r. Un test de modèles emboîtés consistera à tester l'hypothèse $H_0 = \{Y$ est issu du modèle $M_q\}$ contre $H_1 = \{Y$ est issu du modèle $M_r\}$

Exemples de modèles emboîtés

Analyse de variance 1 $Y_{ijk} = \mu + \alpha_i + E_{ijk}$ est emboîté dans le modèle $Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$. Le test de modèles emboîtés est un test de l'hypothèse $H_0 = \{\beta_1 = \beta_j,\ j = 1, \ldots, J\}$, c'est-à-dire l'absence d'effet du deuxième facteur.

Analyse de variance 2 $Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$ est emboîté dans le modèle $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$. Le test de modèles emboîtés est un test de l'hypothèse $H_0 = \{\gamma_{11} = \gamma_{ij} = 0,\ i = 1, \ldots, I, j = 1, \ldots, J\}$, c'est-à-dire l'absence d'interaction.

Regression multiple $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_q x_i^{(q)} + E_i$ est emboîté dans $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)} + E_i$ si $q < p$. Le test de modèles emboîtés est un test de l'hypothèse $H_0 = \{\beta_j = 0,\ j = q+1, \ldots, p\}$.

Regression polynomiale $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i$ est emboîté dans $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \beta_3 x^{(1)}{}_i^2 + \beta_4 x^{(2)}{}_i^2 + E_i$. Le test de modèles emboîtés est un test de l'hypothèse $H_0 = \{\beta_j = 0,\ j = 3, 4\}$, qui teste les termes quadratiques du polynôme.

À la différence du test pour un paramètre (paragraphe 1.4.1, p. 28), ce test permet de tester d'un seul coup l'égalité de plusieurs paramètres du modèle, comme par exemple tous les paramètres associés à un facteur. Il y a des règles hiérarchiques à respecter pour que ce test ait un sens. Par exemple on ne teste pas la nullité de l'effet principal d'un facteur (qui correspond à l'égalité de tous les paramètres liés à ce facteur) si une interaction impliquant ce facteur est incluse dans le modèle.

Les quantités pivot qui vont servir à comparer les modèles sont définies en terme de variabilité expliquée par ces modèles. Cette variabilité est issue de la décomposition de la variabilité de Y, le premier niveau de décomposition consistant à séparer la variabilité résiduelle (SCR) de la variabilité expliquée par le modèle considéré (SCM). La somme de ces deux quantités constitue la variabilité totale (SCT) comme énoncé au théorème 1.4.1. La somme des carrés expliquée par le modèle (SCM) pourra à son tour être décomposée pour faire apparaître la somme des carrés expliquée par le modèle M_q et le gain apporté par le passage de M_q à M_r. Les sommes des carrés associées à chacun de ces modèles seront indicées par q et r.

Théorème 1.4.1. *La variabilité de Y se décompose en la somme de deux termes :*

$$\begin{array}{ccccc} \|Y - \overline{Y}\|^2 & = & \|\widehat{Y} - \overline{Y}\|^2 & + & \|Y - \widehat{Y}\|^2 \\ SCT & = & SCM & + & SCR \end{array}$$

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

Pour tester la pertinence d'un modèle M_q emboîté dans M_r, on décompose SCM_r en 2 parties, SCM_q la somme des carrés expliquée par le modèle M_q et SCM_{r-q} l'écart entre SCM_r et SCM_q, qui mesure l'augmentation de la somme des carrés expliquée par le passage du petit modèle M_q au modèle emboîtant M_r. La proposition 1.4.3 présente le test du modèle M_q emboîté dans M_r.

Proposition 1.4.3. *Soit M_q un modèle emboîté dans M_r et $< X^{(q)} >$ et $< X^{(r)} >$ les espaces associés à ces deux modèles respectivement. On note q le rang de la matrice $X^{(q)}$ et r celui de $X^{(r)}$, $\hat{Y}^{(q)} = X^{(q)}\hat{\theta}^{(q)}$, la prédiction de Y dans le modèle M_q et on définit*

$$SCM_{r-q} = \left\|\hat{Y}^{(r)} - \hat{Y}^{(q)}\right\|^2 = SCM_r - SCM_q.$$

Sous l'hypothèse $H_0 = \{M_q$ et M_r sont équivalents$\}$, la statistique de test est

$$F = \frac{\frac{SCM_{r-q}}{r-q}}{\frac{SCR_r}{n-r}} \overset{H_0}{\sim} \mathcal{F}(r-q, n-r). \tag{1.21}$$

Démonstration. D'après les propositions 11.2.3, p. 315 et 11.2.1, p. 315, on a $\frac{SCM_{r-q}}{\sigma^2} \overset{H_0}{\sim} \chi^2(r-q)$, $\frac{SCR_r}{\sigma^2} \sim \chi^2(n-r)$ et ces deux variables aléatoires sont indépendantes. La définition 10.2.5, p. 311 permet de conclure. □

Interprétation de la proposition 1.4.3. Le modèle M_r est plus général que le modèle M_q, donc il s'ajuste mieux aux données que ce dernier, on a donc $SCM_r - SCM_q \geq 0$. La question qui se pose est la suivante : est-ce que l'augmentation du pouvoir explicatif (sur Y) apportée par le modèle M_r par rapport au modèle M_q est suffisamment forte pour justifier le choix de M_r de préférence à M_q ? Comme M_r correspond à un espace de dimension r plus grand que M_q (de dimension q), la quantité utile est $(SCM_r - SCM_q)/(r-q) = SCM_{r-q}/(r-q)$, l'amélioration du pouvoir explicatif par dimension supplémentaire. Cette statistique a une loi de probabilité qui dépend de la variance résiduelle σ^2 qu'il faut éliminer.

Vision géométrique du test. Notons $\mathbf{1}$ le vecteur de R^n pour lequel chaque composante vaut 1. On se place dans l'espace orthogonal à $\mathbf{1}$, $< \mathbf{1}^\perp >$. On peut alors schématiser le sous-espace $E_q =< \mathbf{1}^{\perp X^{(q)}} >$ engendré par les colonnes de $X^{(q)}$ et orthogonal à $\mathbf{1}$ par une droite passant par le point $\bar{Y}$. Cette droite appartient au sous-espace $< \mathbf{1}^{\perp X^{(r)}} >$ (schématisé par un plan sur la figure 1.2) engendré par les colonnes de $X^{(r)}$ et orthogonal à $\mathbf{1}$.
L'idée du test de l'hypothèse H_0 consiste à voir si le vecteur $\hat{Y}^{(r)}$ est suffisamment proche du sous-espace $< X^{(q)} >$, auquel cas on peut conclure que c'est le modèle M_q qui est préféré. C'est le cas si $\left\|P_{<X^{(r)}>}Y - P_{<X^{(q)}>}(Y)\right\|^2 = SCM_{r-q}$ est petite.

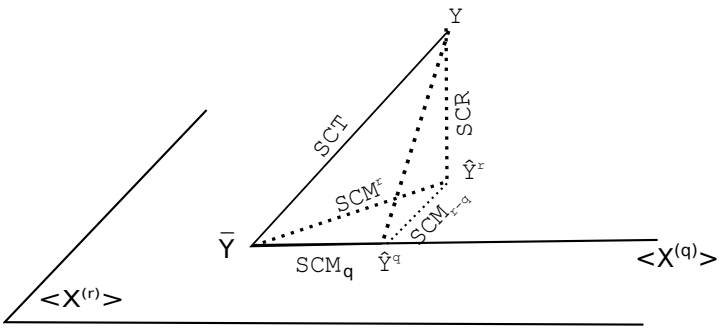


FIGURE 1.2 – Représentation graphique du test de modèles emboîtés .

Cas particulier : test du modèle complet. Le test du modèle complet consiste à tester le modèle naïf appelé M_0 (sans variables explicatives) contre le modèle proposé ou complet appelé M_{comp}. Le modèle M_0 est un modèle emboîté de M_{comp} et le test du modèle complet est donc simplement un test de modèles emboîtés ($M_q = M_0$, c'est-à-dire que $X^{(0)} = \mathbf{1}$, et $\theta^{(0)} = \mu$, et $M_r = M_{comp}$ pour lequel on note X la matrice de design). De plus, notons SCM, SCR et SCT les différentes sommes de carrés associées à ce modèle M_{comp}. Le test est présenté dans la proposition 1.4.4.

Proposition 1.4.4. *Soit*

$$\begin{aligned} H_0 &= \{ \ M_0 \text{ et } M_{comp} \text{ sont équivalents } \} \\ &= \left\{ Y = X^{(0)}\theta^{(0)} + E \right\} \\ &= \{ \ \theta_1 = \theta_2 = \ldots = \theta_p = 0 \ \} \end{aligned}$$

et

$$\begin{aligned} H_1 &= \{ \ M_{comp} \text{ est meilleur que } M_0 \} \\ &= \{ Y = X\theta + E \text{ avec un } i > 0 \text{ tel que } \theta_i \neq 0 \} \\ &= \{ \text{Il existe } i > 0 \text{ tel que } \theta_i \neq 0 \} . \end{aligned}$$

Alors, sous H_0,

$$\frac{SCM}{\sigma^2} \sim \chi^2(r-1), \tag{1.22}$$

et donc

$$\frac{SCM/(r-1)}{SCR/(n-r)} \overset{H_0}{\sim} \mathcal{F}(r-1, n-r). \tag{1.23}$$

Démonstration. La preuve est une application directe de la proposition 1.4.3, p. 31. Il suffit de remarquer que $Y^{(0)} = \bar{Y}$, donc $SCM_{r-1} = \left\| \hat{Y} - \bar{Y} \right\|^2 = SCM$. □

1.4.3 Tests à l'aide des réductions

Pour tester l'effet d'un facteur (resp. d'une variable), il est naturel de tester la nullité des paramètres associés à chaque niveau du facteur (resp le paramètre associé à la variable). Cette approche est insuffisante pour définir correctement l'hypothèse H_0 testée, il faut préciser les hypothèses portant sur les autres variables. La notion de réduction, définie ci-dessous, permet de bien spécifier le rôle des autres variables. L'approche générale est présentée dans la section qui suit et la déclinaison de cette approche dans le cas d'une analyse de la variance et d'une régression multiple se trouve au paragraphe 1.4.3, p. 34.

Cas général

Dans un souci de généralité, les variables (et/ou niveaux des facteurs) seront notés dans cette section $(Var_1, \ldots, Var_l)$ et leurs effets $(a_1, \ldots, a_l)$.

Définition 1.4.2. *Soit un modèle contenant les effets $(a_1, \ldots, a_l)$ des variables/facteurs $(Var_1, \ldots, Var_l)$. On appelle réduction associée à l'introduction de $a_{q_1}, \ldots, a_{q_d}$ dans*

un modèle contenant les effets $a_{i_1}, a_{i_2}, \ldots, a_{i_m}$*, notée* $R(a_{q_1}, \ldots, a_{q_d} | a_{i_1}, a_{i_2}, \ldots, a_{i_m})$*, la norme suivante*

$$\begin{aligned} R(a_{q_1}, \ldots, a_{q_d} | a_{i_1}, a_{i_2}, \ldots, a_{i_m}) &= \| P_{<i_1, i_2, \ldots, i_m, q_1, \ldots, q_d>} Y - P_{<i_1, i_2, \ldots, i_m>} Y \|^2 \\ &= SCM_{i_1, i_2, \ldots, i_m, q_1, \ldots, q_d} - SCM_{i_1, i_2, \ldots, i_m}, \end{aligned} \tag{1.24}$$

$P_{<i_1,i_2>}$ désignant la projection sur l'espace engendré par les variables Var_{i_1}, Var_{i_2} et SCM_{i_1,i_2} la somme des carrés du modèle associée aux variables Var_{i_1}, Var_{i_2}.

Un schéma de différentes réductions dans un modèle simple est donné dans la figure 1.3.

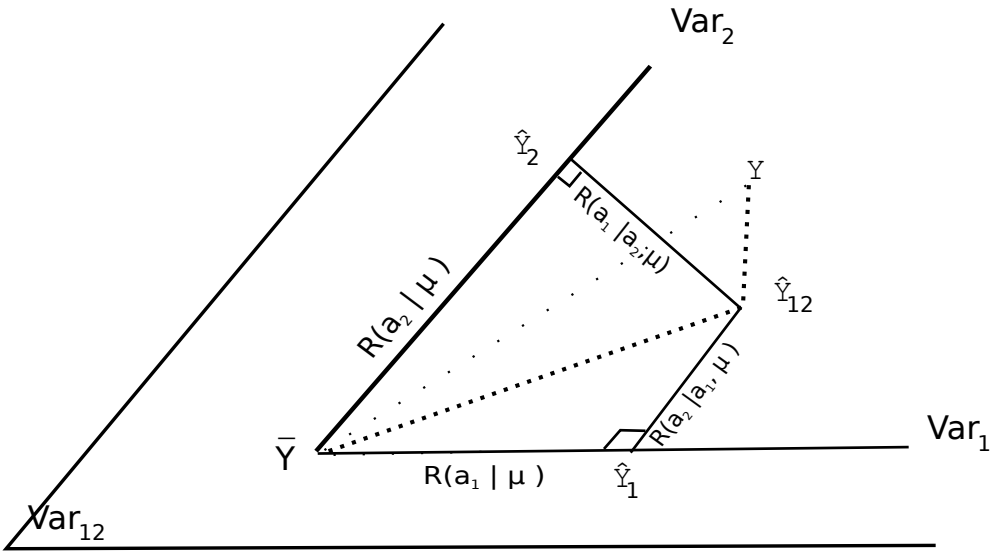


FIGURE 1.3 – Représentation des différentes réductions. Var_1 et Var_2 peuvent désigner soit une variable soit un facteur, μ représente la constante. $R(a_1|\mu)$ est la réduction de variabilité expliquée par l'introduction de a_1 dans le modèle nul M_0, $R(a_2|\mu, a_1)$ celle expliquée par l'introduction de a_2 dans le modèle contenant déjà a_1 et la constante. Notons que Y n'est pas dans le plan engendré par Var_1 et Var_2 et que $SCM = R(a_1, a_2|\mu)$.

Proposition 1.4.5. *Les réductions se décomposent successivement de la façon suivante :*

$$R(a_1, a_2, \ldots, a_p|\mu) = R(a_1|\mu) + R(a_2|a_1, \mu) + \ldots + R(a_p|a_1, \ldots, a_{p-1}, \mu).$$

La preuve cette proposition est une application directe du théorème de Pythagore (cf figure 1.3).

Un test de l'effet d'une variable/facteur est associé à chaque réduction. Considérons la réduction $R(a_{q_1}, \ldots, a_{q_d} | a_{i_1}, a_{i_2}, \ldots, a_{i_m})$, l'hypothèse H_0 associée s'écrit

$$H_0 = \left\{ P_{<i_1, i_2, \ldots, i_m, q_1, \ldots, q_d>} X\theta = P_{<i_1, i_2, \ldots, i_m>} X\theta \right\}.$$

La statistique de test est donnée par

$$F = \frac{\frac{R(a_{q_1},\ldots,a_{q_d}|a_{i_1},a_{i_2},\ldots,a_{i_m})}{q}}{\frac{SCR}{n-r}},$$

où

— q est la différence des dimensions des espaces $< a_{q_1},\ldots,a_{q_d}a_{i_1},a_{i_2},\ldots,a_{i_m} >$ et $< a_{i_1},a_{i_2},\ldots,a_{i_m} >$,

— SCR est la somme des carrés du modèle complet,

— r est la dimension de l'espace engendré par le modèle complet.

Le théorème de Cochran assure que

$$F \overset{H_0}{\sim} \mathcal{F}(q, n-r).$$

Remarquons que ce test est exactement un test de modèles emboîtés si le modèle complet est celui associé aux paramètres $a_{q_1},\ldots,a_{q_d}a_{i_1},a_{i_2},\ldots,a_{i_m}$.

Quelques réductions classiques

On se propose d'illustrer les tests associés aux réductions sur l'exemple d'une analyse de la variance à deux facteurs et d'une régression multiple . Les deux modèles considérés s'écrivent

$$\begin{aligned}(M1)\ &: Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}, \quad E_{ijk} \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)\\ &\quad i = 1,\ldots I, \quad j = 1,\ldots J, \quad k = 1,\ldots,n_{ij}\\ (M2)\ &: Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + E_i, \quad E_i \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)\\ &\quad i = 1,\ldots,n\end{aligned}$$

Certaines réductions sont classiquement utilisées et fournies en standard en sortie des logiciels. Parmi les plus classiques, on trouve les réductions de Type I et les réductions de Type II. Les tests associés sont souvent présentés au sein d'une table d'analyse de la variance. Les paragraphes suivants présentent cette table pour les deux modèles M_1 et M_2.

Somme des carrés de Type I. Les réductions de Type I proviennent de la décomposition de la somme des carrés du modèle en réductions successives. Dans le modèle M_1,

$$SCM = R(\alpha,\beta,\gamma|\mu) = R(\alpha|\mu) + R(\beta|\alpha,\mu) + R(\gamma|\alpha,\beta,\mu)$$

Les tests des différents effets associés à cette décomposition sont présentés dans la table 1.1. Dans le modèle M_2,

$$SCM = R(\beta_1,\ldots,\beta_p|\beta_0) = R(\beta_1|\beta_0) + R(\beta_2|\beta_1,\beta_0) + \ldots + R(\beta_p|\beta_{p-1},\ldots,\beta_1,\beta_0)$$

La table d'analyse de la variance correspondante à ces tests est donnée dans la table 1.2.

Remarquons que l'ordre d'introduction des variables/facteurs dans un modèle leur confère des rôles différents. Cette asymétrie dans le traitement de chacune des variables est parfois considérée comme un inconvénient de la décomposition de Type I. Les décompositions de Type II proposées ci-après symétrisent leurs rôles.

Effet	Réduction Type I	DDL	F	Question
α	$R(\alpha\|\mu)$	$I-1$	$\frac{\frac{R(\alpha\|\mu)}{I-1}}{\frac{SCR}{n-IJ}}$	Est-il pertinent d'ajouter l'effet du facteur 1 à un modèle ne contenant que la constante ?
β	$R(\beta\|\mu,\alpha)$	$J-1$	$\frac{\frac{R(\beta\|\mu,\alpha)}{J-1}}{\frac{SCR}{n-IJ}}$	Est-il pertinent d'ajouter l'effet du facteur 2 à un modèle contenant la constante et l'effet du facteur 1 ?
γ	$R(\gamma\|\mu,\alpha,\beta)$	$(I-1)\times$ $(J-I)$	$\frac{\frac{R(\gamma\|\mu,\alpha,\beta)}{(I-1)\times(J-1)}}{\frac{SCR}{n-IJ}}$	Est-il pertinent d'ajouter l'effet de l'interaction entre les deux facteurs à un modèle contenant la constante et les effets des deux facteurs ?

TABLE 1.1 – Table d'analyse de la variance des réductions de Type I du modèle M_1.

Somme des carrés de Type II. L'idée des Types II consiste essentiellement à considérer la réduction portée par une variable/facteur conditionnellement aux autres. Les tables 1.3 et 1.4 présentent les tables d'analyse de la variance pour les modèles M_1 et M_2 respectivement.

Effet	Réduction Type I	DDL	F	Question
β_1	$R(\beta_1\|\beta_0)$	1	$\frac{\frac{R(\beta_1\|\beta_0)}{1}}{\frac{SCR}{n-p-1}}$	Est-il pertinent d'ajouter l'effet de la variable $x^{(1)}$ à un modèle ne contenant que la constante ?
β_2	$R(\beta_2\|\beta_1,\beta_0)$	1	$\frac{\frac{R(\beta_2\|\beta_1,\beta_0)}{1}}{\frac{SCR}{n-p-1}}$	Est-il pertinent d'ajouter l'effet de la variable $x^{(2)}$ à un modèle contenant la constante et la variable $x^{(1)}$?
...				...
β_p	$R(\beta_p\|\beta_1,\ldots\beta_{p-1},\beta_0)$	1	$\frac{\frac{R(\beta_p\|\beta_1,..\beta_{p-1},\beta_0)}{1}}{\frac{SCR}{n-p-1}}$	Est-il pertinent d'ajouter l'effet de la variable $x^{(p)}$ à un modèle contenant la constante et les variables $x^{(1)}$, ..., $x^{(p-1)}$?

TABLE 1.2 – Table d'analyse de la variance des réductions de Type I du modèle M_2.

Effet	Réduction Type II	DDL	F	Question
α	$R(\alpha\|\beta,\mu)$	$I-1$	$\frac{\frac{R(\alpha\|\beta,\mu)}{I-1}}{\frac{SCR}{n-IJ}}$	Est-il pertinent d'ajouter l'effet du facteur 1 à un modèle contenant la constante et l'effet du facteur 2 ?
β	$R(\beta\|\mu,\alpha)$	$J-1$	$\frac{\frac{R(\beta\|\mu,\alpha)}{J-1}}{\frac{SCR}{n-IJ}}$	Est-il pertinent d'ajouter l'effet du facteur 2 à un modèle contenant la constante et l'effet du facteur 1 ?
γ	$R(\gamma\|\mu,\alpha,\beta)$	$(I-1)\times$ $(J-I)$	$\frac{\frac{R(\gamma\|\mu,\alpha,\beta)}{(I-1)\times(J-1)}}{\frac{SCR}{n-IJ}}$	Est-il pertinent d'ajouter l'effet de l'interaction entre les deux facteurs à un modèle contenant la constante et les effets des deux facteurs ?

TABLE 1.3 – Table d'analyse de la variance des réductions de Type II du modèle M_1.

Quelques remarques sur les décompositions de Type I et de Type II. Les conclusions portant sur l'effet d'une variable/facteur en Type I et en Type II peuvent être différentes puisque les tests proposés sont différents. Cette différence porte sur la définition de l'effet d'une variable/facteur.

Dans un modèle d'analyse de la variance (M_1), on regarde en premier lieu le test

Effet	Réduction Type II	DDL	F	Question
β_1	$R(\beta_1\|\beta_0,\beta_2,\ldots\beta_p)$	1	$\frac{\frac{R(\beta_1\|\beta_0,\beta_2,..\beta_p)}{1}}{\frac{SCR}{n-p-1}}$	Est-il pertinent d'ajouter l'effet de la variable $x^{(1)}$ à un modèle contenant la constante et toutes les autres variables ?
β_2	$R(\beta_2\|\beta_0,\beta_1,\beta_3,..\beta_p)$	1	$\frac{\frac{R(\beta_2\|\beta_0,\beta_1,\beta_3,..\beta_p)}{1}}{\frac{SCR}{n-p-1}}$	Est-il pertinent d'ajouter l'effet de la variable $x^{(2)}$ à un modèle contenant la constante et toutes les autres variables ?
...				...
β_p	$R(\beta_p\|\beta_0,..\beta_{p-1})$	1	$\frac{\frac{R(\beta_p\|\beta_0,\beta_1,..\beta_{p-1})}{1}}{\frac{SCR}{n-p-1}}$	Est-il pertinent d'ajouter l'effet de la variable $x^{(p)}$ à un modèle contenant la constante et toutes les autres variables ?

TABLE 1.4 – Table d'analyse de la variance des réductions de Type II du modèle M_2.

portant sur l'interaction . Si le test est significatif, on conclut à un effet de chacun des facteurs ne serait-ce qu'au travers de leur interaction . Pour cette raison, on ne considère pas $R(\alpha|\beta,\gamma,\mu)$ qui de toute façon est nulle. Pour différencier l'effet d'un facteur à travers son effet principal ou son interaction, il est possible de définir des réductions contraintes (Type III par exemple) qui ne sont pas présentées ici.

Dans un modèle de régression multiple (M_2), les tests de Type II pour chaque variable sont exactement équivalents aux tests de nullité des coefficients présenté dans la proposition 1.4.1, p. 28.

La somme des réductions en Type II n'a pas de raison d'être égale à la somme des carrés du modèle.

On peut remarquer en s'appuyant sur la figure 1.4 que si les sous-espaces engendrés par Var_1 et Var_2 et orthogonalisés par rapport à la constante, sont orthogonaux, alors $R(\alpha_1|\mu) = R(\alpha_1|\mu,\alpha_2)$ et $R(\alpha_2|\mu) = R(\alpha_2|\mu,\alpha_1)$. On parle dans ce cas de dispositif orthogonal et les tests impliquants les réductions citées seront équivalents. Dans le cas particulier du modèle M_1, les réductions de Type I et de Type II sont les mêmes.

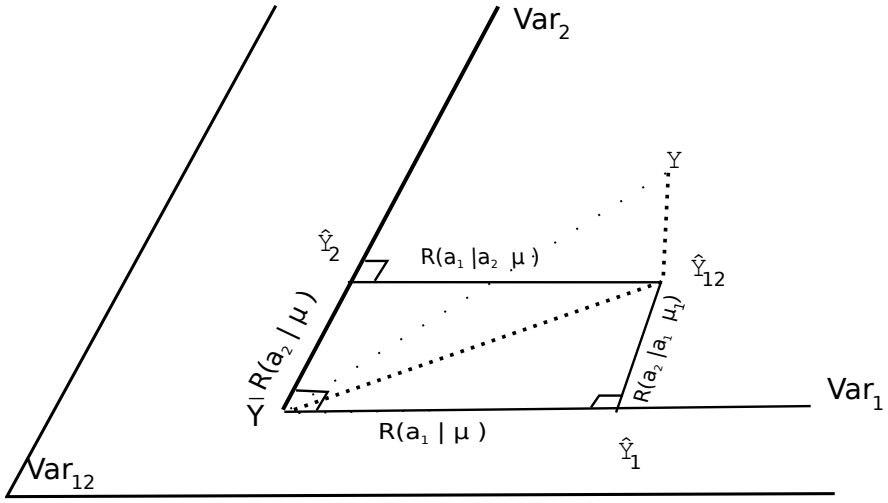


FIGURE 1.4 – Représentation des différentes réductions. Var_1 et Var_2 peuvent désigner soit une variable soit un facteur, μ représente la constante. Dans le cas orthogonal présenté ici, $R(a_1|\mu) = R(a_1|\mu,a_2)$ et $R(a_2|\mu) = R(a_2|\mu,a_1)$. Les réductions sont les normes au carré des vecteurs correspondants.

1.5 Notion d'orthogonalité

Rappelons que le modèle linéaire s'écrit sous forme générale $\mathbb{E}(Y) = X\theta$. La matrice X peut s'écrire par bloc $X = [\mathbf{1}\ X^1\ X^2 \dots X^p]$, chacun étant associé à un effet (un facteur en anova, une variable en régression, l'un ou l'autre en ancova, $\mathbf{1}$ étant associé à la constante). L'espace engendré par X peut toujours se décomposer sous la forme suivante :

$$< X >=< \mathbf{1} > + < \mathbf{1}^{\perp X^1} > + < \mathbf{1}^{\perp X^2} > + \dots + < \mathbf{1}^{\perp X^p} >,$$

où $< \mathbf{1}^{\perp X^j} >$ est le sous-espace de X^j orthogonal à $< \mathbf{1} >$.

La matrice X peut donc se réécrire sous la forme :

$$X = [\mathbf{1}\ \mathbf{1}^{\perp X^1}\ \mathbf{1}^{\perp X^2}\ \ldots\ \mathbf{1}^{\perp X^p}],$$

La notion d'orthogonalité est générale. Cependant pour simplifier l'exposé, on se restreint au cas d'un modèle sans interaction.

Définition 1.5.1. *Un plan d'expériences est orthogonal pour les effets principaux, si pour tous les couples j, j', correspondants à des effets principaux (une variable en régression ou un facteur seul en anova) $< \mathbf{1}^{\perp X^j} >$ et $< \mathbf{1}^{\perp X^{j'}} >$ sont orthogonaux.*

1.5.1 Propriétés liées à l'orthogonalité

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

Propriété 1.5.1. *Dans un plan orthogonal pour les effets principaux,*

- $\forall (j, j'), R(\alpha_{j'}|\mu, \alpha_j) = R(\alpha_{j'}|\mu)$, *les sommes de carrés de Type I et II sont égales. La somme des carrés due à un terme donné (c'est le numérateur du test 1.21, p. 31 de nullité de ce terme du modèle d'analyse de la variance) ne dépend pas du modèle "emboîtant".*
- *la matrice de variance des estimateurs est diagonale par blocs, les estimateurs appartenant à deux blocs différents sont non corrélés . En effet l'estimation de ces paramètres correspond à des projections dans des sous-espaces orthogonaux.*
- *les estimations des paramètres concernant un terme ne changent pas lorsque l'on considère un sous-modèle.*

1.5.2 Orthogonalité dans le modèle d'analyse de la variance

Considérons le modèle d'analyse de la variance à deux facteurs sans interaction écrit sous la forme suivante :

$$\mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j,$$

avec $i = 1, \ldots, I$, $j = 1, \ldots, J$ et $k = 1, \ldots, n_{ij}$ où n_{ij} est le nombre d'observations pour la modalité i du facteur associé à α et la modalité j du facteur associé à β. La matrice X s'écrit :

$$
X = \begin{array}{c}
\begin{array}{cccccccccccc}
\mathbf{1} & A^1 & A^2 & A^3 & \dots & A^I & B^1 & B^2 & \dots & \dots & B^J
\end{array} \\
\left(\begin{array}{cccccccccccc}
\mathbf{1}_{n_{11}} & \mathbf{1}_{n_{11}} & \mathbf{0}_{n_{11}} & \dots & \dots & \mathbf{0}_{n_{11}} & \mathbf{1}_{n_{11}} & \mathbf{0}_{n_{11}} & \dots & \dots & \mathbf{0}_{n_{11}} \\
\mathbf{1}_{n_{12}} & \mathbf{1}_{n_{12}} & \vdots & \dots & \dots & \vdots & \mathbf{0}_{n} & \mathbf{1}_{n_{12}} & \ddots & & \vdots \\
\vdots & \vdots & \vdots & \dots & \dots & \vdots & \vdots & \ddots & & \ddots & \vdots \\
\vdots & \vdots & \vdots & \dots & \dots & \vdots & \vdots & & \ddots & & \mathbf{0}_{n_{i,J-1}} \\
\mathbf{1}_{n_{1J}} & \mathbf{1}_{n_{1J}} & \mathbf{0}_{n_{1J}} & \dots & \dots & \mathbf{0}_{n_{1J}} & \mathbf{0}_{n_{1J}} & \dots & \dots & \mathbf{0}_{n_{1J}} & \mathbf{1}_{n_{1J}} \\
\mathbf{1}_{n_{21}} & \mathbf{0}_{n_{21}} & \mathbf{1}_{n_{21}} & \mathbf{0}_{n_{21}} & \dots & \mathbf{0}_{n_{21}} & \mathbf{1}_{n_{21}} & \mathbf{0}_{n_{21}} & \dots & \dots & \mathbf{0}_{n_{21}} \\
\mathbf{1}_{n_{22}} & \mathbf{0}_{n_{22}} & \mathbf{1}_{n_{22}} & \mathbf{0}_{n_{22}} & \dots & \mathbf{0}_{n_{22}} & \mathbf{0}_{n_{22}} & \mathbf{1}_{n_{22}} & \ddots & & \vdots \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & \ddots & & \mathbf{0}_{n_{1,J-1}} \\
\mathbf{1}_{n_{2J}} & \mathbf{0}_{n_{2J}} & \mathbf{1}_{n_{2J}} & \mathbf{0}_{n_{2J}} & \dots & \mathbf{0}_{n_{2J}} & \mathbf{0}_{n_{2J}} & \dots & \dots & \mathbf{0}_{n_{2J}} & \mathbf{1}_{n_{2J}} \\
 & & & & & \vdots & & & & & \\
 & & & & & \vdots & & & & & \\
 & & & & & \vdots & & & & & \\
 & & & & & \vdots & & & & & \\
\mathbf{1}_{n_{I1}} & \mathbf{0}_{n_{I1}} & \mathbf{0}_{n_{I1}} & \dots & \mathbf{0}_{n_{I1}} & \mathbf{1}_{n_{I1}} & \mathbf{1}_{n_{I1}} & \mathbf{0}_{n_{I1}} & \dots & \dots & \mathbf{0}_{n_{I1}} \\
\mathbf{1}_{n_{I2}} & \mathbf{0}_{n_{I2}} & \dots & \dots & \mathbf{0}_{n_{I2}} & \mathbf{1}_{n_{I2}} & \mathbf{0}_{n_{22}} & \mathbf{1}_{n_{22}} & \ddots & & \vdots \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & \ddots & & \mathbf{0}_{n_{I,J-1}} \\
\mathbf{1}_{n_{IJ}} & \mathbf{0}_{n_{IJ}} & \dots & \dots & \mathbf{0}_{n_{IJ}} & \mathbf{1}_{n_{IJ}} & \mathbf{0}_{n_{IJ}} & \dots & \dots & \mathbf{0}_{n_{IJ}} & \mathbf{1}_{n_{IJ}}
\end{array}\right)
\end{array}
$$

où $A = [A^1\ A^2\ \dots\ A^I]$ sont les colonnes de X associées à l'effet α et $B = [B^1\ B^2\ \dots\ B^J]$ celles associées à l'effet β. Comme proposé ci-dessus, la matrice X peut s'écrire :

$$X = \left[\mathbf{1}\ \mathbf{1}^{\perp A}\ \mathbf{1}^{\perp B}\right].$$

Le plan d'expérience est donc orthogonal si les espaces $< \mathbf{1}^{\perp A} >$ et $< \mathbf{1}^{\perp B} >$ sont orthogonaux. Cette propriété s'exprime plus simplement sous la forme donnée dans la proposition 1.5.1.

Proposition 1.5.1. *En analyse de la variance à deux facteurs , le plan d'expérience est orthogonal ssi*

$$n_{ij} = \frac{n_{i+} \times n_{+j}}{n},$$

où n_{ij} est le nombre d'observations pour la modalité i du facteur associé à α et la modalité j du facteur associé à β et $n_{i+} = \sum_{j=1}^{J} n_{ij}$ (respectivement $n_{+j} = \sum_{i=1}^{I} n_{ij}$) est le nombre total d'observations pour la modalité i du facteur associé à α (respectivement j du facteur associé à β).

Cette condition tient toujours dans le cas d'un modèle d'analyse de la variance à deux facteurs avec interaction car la condition d'orthogonalité ne porte que sur les facteurs principaux.

Un plan complet équilibré (i.e. où le nombre d'observations est le même pour tous les croisements des deux modalités, i.e. n_{ij} est égale à une constante) est donc un plan orthogonal.

Démonstration. On veut déterminer les conditions sur les effectifs n_{ij} pour que l'espace $< \mathbf{1}^{\perp A} >$ soit orthogonal à l'espace $< \mathbf{1}^{\perp B} >$.

Or

$$< \mathbf{1}^{\perp A} > \perp < \mathbf{1}^{\perp B} > \Leftrightarrow < \mathbf{1}^{\perp A} > \cap < B^{\perp} > = < \mathbf{1}^{\perp A} > .$$

Mais

$$u \in < \mathbf{1}^{\perp A} > \Leftrightarrow u = \sum_{i=1}^{I} a_i A^i \text{ et } \sum_{i=1}^{I} a_i n_{i+} = 0, \tag{1.25}$$

et

$$v \in < \mathbf{1}^{\perp A} > \cap < B^{\perp} > \Leftrightarrow u = \sum_{i=1}^{I} a_i A^i \text{ et } \forall j \sum_{i=1}^{I} n_{ij} a_i = 0. \tag{1.26}$$

Les espaces $< \mathbf{1}^{\perp A} > \cap < B^{\perp} >$ et $< \mathbf{1}^{\perp A} >$ sont égaux ssi

$$\forall a_i : \left\{ \forall j \sum_{i=1}^{I} n_{ij} a_i = 0 \Leftrightarrow \sum_{i=1}^{I} a_i n_{i+} = 0 \right\} \Leftrightarrow \{ \forall j, \quad n_{ij} = \lambda n_{i+} \} .$$

Finalement puisque $\sum_{i=1}^{I} n_{ij} = n_{+j} = \lambda n$, $\lambda = n_{+j}/n$ et par suite $n_{ij} = n_{i+} n_{+j}/n$. □

1.5.3 Orthogonalité dans le cas de la régression multiple

Dans ce modèle, la condition d'orthogonalité se traduit par le fait que les variables explicatives sont non corrélées entre elles. En effet, un élément qui appartient à l'espace $< \mathbf{1}^{\perp X^j} >$ est un vecteur $u = aX^j$ tel que

$$< u, \mathbf{1} > = 0 \quad \Leftrightarrow \quad \sum_i x_{ij} = 0 \quad \Leftrightarrow \quad x_{\bullet j} = 0,$$

où $x_{\bullet j}$ est la moyenne de la $j$ème variable. Et dire que $< \mathbf{1}^{\perp X^j} >$ et $< \mathbf{1}^{\perp X^{j'}} >$ sont orthogonaux signifient que

$$\sum_i (x_{ij} - x_{\bullet j})(x_{ij'} - x_{\bullet j'}) = 0.$$

On reconnaît la covariance empirique. Ce cas est peu fréquent car en général les variables explicatives ne sont pas maîtrisées (à moins que cela soit fixé par l'expérimentateur).

1.5.4 Orthogonalité dans le cas de l'analyse de la covariance

Pour simplifier la présentation, considérons un modèle d'analyse de la covariance avec un seul facteur et une seule covariable x sans interaction. Pour assurer l'orthogonalité, il faut que la covariable soit centrée au sein de chaque modalité i du facteur (ce

résultat est identique si on prend en compte une interaction entre la covariable et le facteur). En effet,

$$< X > \perp < \mathbf{1} > \Leftrightarrow x_{\bullet\bullet} = 0,$$

où $x_{\bullet\bullet}$ est la moyenne globale de la covariable, i.e. la covariable est centrée.

$$< A > \perp < \mathbf{1} > \Leftrightarrow \sum_i a_i n_i = 0,$$

et

$$< \mathbf{1}^{\perp X} > \perp < \mathbf{1}^{\perp A} > \Leftrightarrow \sum_i a_i n_i x_{i\bullet} = 0,$$

soit $x_{i\bullet} = 0$ quelque soit i.

1.6 Qualité d'ajustement et comparaison de modèles

1.6.1 Le coefficient de détermination

Une mesure naturelle pour quantifier la qualité de l'ajustement du modèle aux données est de regarder la part de variabilité expliquée par le modèle par rapport à la variabilité totale. On va donc s'intéresser au rapport de la somme des carrés du modèle sur la somme des carrés totale. Ce rapport, généralement noté R^2, s'appelle le coefficient de détermination :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}. \tag{1.27}$$

Il est toujours compris entre 0 et 1, et plus R^2 est proche de 1, meilleur est l'ajustement. Géométriquement ce rapport est égal au carré du cosinus de l'angle du vecteur $Y - \overline{Y}$ avec le sous-espace $< X >$.

On s'intéresse essentiellement à cette quantité lorsque l'on est dans un objectif de prédiction, i.e. pour juger de la précision des prédictions que l'on obtiendra par ce modèle.

Si l'on cherche à comparer le pouvoir prédictif d'un modèle sur Y avec le même modèle mais sur une transformation de Y (par exemple le log), on ne peut directement comparer les R^2 obtenus. Pour que cela soit rendu possible, il faut recalculer le R^2 du modèle sur $\log Y$ en utilisant la transformation inverse (et donc obtenir un R^2 sur Y).

1.6.2 Critères de comparaison de modèles

Le coefficient de détermination augmente avec le nombre de variables (ou de paramètres à estimer). Ainsi si on doit choisir entre un modèle avec p variables et le même modèle mais avec une variable de plus en se fiant au R^2, on choisira toujours le modèle avec le plus de variables. Ce critère ne prend en compte que l'ajustement du modèle et pas le prix de cet ajustement, i.e. le nombre de paramètres à estimer. Nous présentons ici deux critères qui mettent en jeu ces deux considérations.

coefficient de détermination ajusté. Le R^2 ajusté, noté R^2_{adj}, est donc une version du R^2 ajusté au nombre de paramètres à estimer. Pour un modèle à p variables (i.e. $p+1$ paramètres à estimer), il est défini de la façon suivante :

$$R^2_{adj} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = R^2 - \frac{p}{n-p-1}(1-R^2). \tag{1.28}$$

Ce nouveau R^2 n'augmentera que si la variable ajoutée améliore nettement l'ajustement, i.e. est donc explicative.

Critère AIC. AIC (Akaike Information Criterion, [2]) est un critère de sélection de modèles bien connu. Pour un modèle M, il est défini de la façon suivante :

$$AIC(M) = -2\mathcal{L}_M(y; \hat{\theta}, \hat{\sigma}^2) + 2D_M, \tag{1.29}$$

où $\mathcal{L}_M(y; \hat{\theta}, \hat{\sigma}^2)$ est la log-vraisemblance du modèle M calculée en son maximum, qui est aussi une mesure d'ajustement du modèle aux données, et D_M est le nombre de paramètres à estimer dans le modèle M, qui représente le prix de l'estimation. Le modèle qui a le plus petit AIC fait donc un bon compromis entre un bon ajustement du modèle aux données et un nombre raisonnable de paramètres à estimer.

1.7 Diagnostic

Plusieurs hypothèses ont été considérées lors de la construction du modèle. L'analyse de ce modèle (en particulier les tests effectués) repose sur celles-ci, il est donc important de les vérifier/valider avant toute analyse. Avant d'énumérer les différentes hypothèses, ainsi que les outils utilisés pour la validation de chacune, rappelons la définition du levier, des résidus bruts et des résidus standardisés.

Levier

Le prédicteur de Y est $\widehat{Y} = PY$. Notons h_{ij} le terme général de la matrice P, alors :

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j.$$

h_{ij} est une mesure de l'importance de l'observation y_j pour la prédiction de y_i. Par exemple en analyse de la variance à 1 facteur, chaque individu est caractérisé par le niveau du facteur k et l'indice de répétition l, soit $i = (k, l)$. On a $\hat{y}_i = \hat{y}_{kl} = \frac{1}{n_k} \sum_{t=1}^{n_k} y_{kt}$ et $h_{ij} = \frac{1}{n_k}$ si l'individu j a le niveau k du facteur et $h_{ij} = 0$ sinon. Ainsi plus le groupe auquel appartient l'observation contient d'observations, moins elle a d'importance.
En régression linéaire simple,

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2},$$

ainsi plus la valeur de la covariable x_j est éloignée du centre ($\bar{x}$), plus cette observation sera influente.

La valeur h_{ii} est ce que l'on appelle le levier qui mesure l'importance du rôle que joue l'observation y_i pour sa propre prédiction.

Résidus bruts et standardisés . Le résidu $\hat{e}_i$ de l'observation i se calcule en mesurant l'écart entre l'observation y_i et sa prédiction $\hat{y}_i$:

$$\hat{e}_i = y_i - \hat{y}_i. \tag{1.30}$$

Ce résidu est appelé résidu brut. Même sous l'hypothèse d'homoscédasticité, les résidus bruts n'ont pas tous la même variance :

$$\mathbb{V}[\hat{E}_i] = \sigma^2(1 - h_{ii}).$$

On en calcule des versions standardisées en normalisant les résidus bruts par leurs écart-types estimés afin de les rendre comparables :

$$\tilde{r}_i = \frac{\hat{e}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}. \tag{1.31}$$

Hypothèses du modèle et validation. Les hypothèses du modèle sont

1. $\mathbb{E}[Y] = X\theta$, la moyenne est une combinaison linéaire des variables explicatives. Cette hypothèse ne se vérifie que dans les cas particuliers de modèles qui font intervenir des variables quantitatives (régression ou ancova). L'hypothèse que la variable à expliquer y est supposée être linéaire en les variables explicatives $x^{(j)}$ n'est pas toujours satisfaite. Si ce n'est pas le cas et si cela n'a pas déjà été repéré lors de l'analyse descriptive (en visualisant le graphique des couples $(x_i^{(j)}, y_i)$), un diagnostic peut être fait à partir de ce que l'on appelle le graphe des résidus qui consiste à tracer les résidus standardisés $\tilde{r}_i$ en fonction des prédictions $\hat{y}_i$. En effet, des termes oubliés dans la modélisation se retrouvent systématiquement dans les résidus. Si une structure particulière apparaît, on peut mettre en question la supposition $\mathbb{E}[Y] = X\theta$. Quelquefois il suffit de transformer une variable explicative x par une fonction ($\log x$, $\sqrt{x}$, x^2...) ou d'ajouter cette variable transformée pour résoudre le problème.

2. Les erreurs sont indépendantes. En pratique, cette hypothèse est en général supposée. Elle se déduit du plan d'expérience mené : on dira que cette hypothèse est vérifiée si les observations sont issues d'expériences menées dans des conditions indépendantes. Par exemple, si un individu est utilisé deux fois dans l'expérience, on perd l'hypothèse d'indépendance. On peut vouloir tester cette hypothèse. Il existe plusieurs tests dont le test de Durbin-Watson, ([23, 24]).
3. $E_i \sim \mathcal{N}$, les erreurs suivent une loi gaussienne. Cette hypothèse peut être vérifiée en comparant la distribution empirique des résidus à la distribution théorique soit graphiquement (qqplot, histogramme par exemple), soit en effectuant un test comme par exemple
   — Shapiro-Wilks ([53])
   — Kolmogorov-Smirnov ([13]),
   — test du χ^2 ([11]),

avec la limite que ces tests supposent l'indépendances des variables observées, qui dans ce contexte sont les résidus, ceux-ci n'étant pas indépendants par construction.

4. $\mathbb{V}[E_i] = \sigma^2$, la variance des erreurs est constante (on parle d'homoscédasticité). La validation de cette hypothèse se fait via le graphe des résidus. Typiquement, si la variabilité des résidus augmente avec les prédictions, cette hypothèse n'est pas validée. Afin de stabiliser la variance, une procédure classique consiste à transformer les données. Les deux transformations les plus courantes sont :

$$\begin{cases} \log Y & \text{si } \sigma \text{ est proportionnel à } \mathbb{E}[Y] \\ \sqrt{Y} & \text{si } \sigma^2 \text{ est proportionnel à } \mathbb{E}[Y] \end{cases}$$

Détection de valeurs atypiques. Il se peut que certaines observations soient suspectes. On peut d'une part s'interroger sur la validité de ces mesures et d'autre part elles peuvent dans ce cas avoir une "trop" grande influence sur l'estimation des paramètres. L'un des critères les plus utilisés pour détecter ces points est la distance de Cook qui mesure l'influence d'une observation sur l'ensemble des prédictions en prenant en compte l'effet levier et l'importance des résidus :

$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_{(i)j} - \hat{y}_j)^2}{(p+1)\hat{\sigma}^2} = \frac{h_{ii}}{(p+1)(1-h_{ii})} \tilde{r}_i,$$

où $\hat{y}_{(i)j}$ est la prédiction de l'observation j obtenue sans l'observation i.

Cook [14] propose de considérer un individu comme atypique dès lors que sa distance de Cook $C_i > 1$. On peut cependant regarder avec plus de précision son résidu et son effet levier. On peut alors vouloir la supprimer.

Attention la suppression d'un point atypique doit toujours être justifiée par un argument non statistique (erreur de mesure probable, animal malade, condition exceptionnelle...).

1.8 Limites et extensions du modèle linéaire

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

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

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

— Un écart modéré de E à la loi normale a peu d'impact sur la qualité des résultats, d'autant moins que la "vraie" distribution des variables résiduelles est symétrique (ceci grâce au théorème de la limite centrale qui nous assure qu'une somme de variables aléatoires sous de bonnes conditions tend vers une loi normale).

— En analyse de la variance, l'inégalité de la variance résiduelle pour différents niveaux d'un facteur peut avoir des conséquences importantes si le dispositif n'est pas équilibré. Si le dispositif est équilibré, les résultats obtenus seront convenables.
— La supposition d'indépendance des erreurs est importante. Les résultats de l'estimation et des tests y sont sensibles, surtout si les corrélations sont élevées.

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

1.8.1 Généralisations du modèle linéaire

Modèles non linéaires

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

Il est plus adapté dans ce cas de choisir de modéliser l'espérance de la variable à expliquer par une fonction non linéaire. On a alors un modèle "non linéaire" ([34, 5, 58]).

Modèle linéaire général

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

Modèle linéaire généralisé

Le modèle est le suivant : $\mathbb{E}[g(Y)] = X\theta$ et $Y \sim \mathcal{L}(X\theta, \sigma^2)$, où g est une fonction monotone appelée fonction de lien et $\mathcal{L}$ est une loi de probabilité à choisir parmi les lois normales, de Poisson, Binomiale ou Gamma (plus généralement n'importe quelle loi de la famille exponentielle). Ce type de modèle est détaillé dans les chapitres 3, p. 115 et 4, p. 132.

1.8.2 Les variables explicatives sont aléatoires

Cas de la régression

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

Modèle mixte

Quand les variables E_i ne sont pas indépendantes, on utilise le modèle mixte. D'autre part, il arrive que les effets de certains facteurs soient modélisés comme des variables aléatoires. Le modèle mixte est traité dans les chapitres 5, p. 166 et 6, p. 196.

Chapitre 2

Exemples de modèles linéaires

Ce chapitre présente des exemples d'applications du modèle linéaire. Il vise à illustrer sur différents exemples les concepts présentés dans le chapitre précédent et à s'attarder sur les spécificités de chacun des modèles présentés.

2.1 Régression linéaire simple et polynomiale

2.1.1 Présentation du problème

Pour étudier la possible bioaccumulation de pesticides chez les brochets, on a mesuré le taux de DDT dans la chair de brochets de différents âges, capturés dans une même rivière. On cherche à décrire la relation potentielle entre le taux de DDT et l'âge des brochets. Le taux de DDT (variable `TxDDT`) pour le brochet i est noté y_i, son âge x_i (variable `Age`) et n désigne le nombre total de brochets, ici $n = 15$.

Un extrait des données est donné dans la table 2.1 et le graphe du taux de DDT des brochets en fonction de leur âge en figure 2.1. On remarque que plus le brochet est âgé, plus le taux de DDT est élevé. Ainsi il semble exister une relation entre ces quantités, de type plutôt quadratique que linéaire. On peut aussi déjà remarquer que la variabilité du taux de DDT par âge augmente elle aussi avec l'âge.

	Age	TxDDT
1	2	0.20
2	2	0.25
3	2	0.18
4	3	0.19
5	3	0.29
6	3	0.28
7	4	0.31
8	4	0.33

TABLE 2.1 – Extrait des données

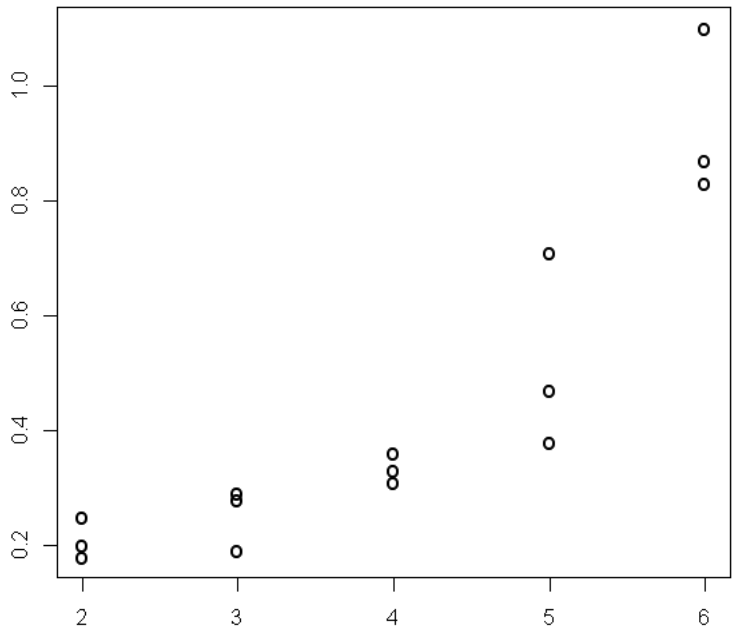


FIGURE 2.1 – Graphe des observations : taux de DDT en fonction de l'âge.

2.1.2 Modèle de régression linéaire simple

Si on suppose que le lien entre le taux de DDT et l'âge est une fonction affine le modèle approprié pour la modéliser est le modèle de régression linéaire simple : on suppose que les observations y_i sont des réalisations des n variables aléatoires indépendantes Y_i telles que

$$Y_i = a + bx_i + E_i, \qquad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n = 15.$$

L'analyse de ce modèle n'est possible que si les hypothèses du modèle linéaire sont recevables (se référer à la partie 1.7, p. 42 du chapitre Modèle Linéaire). La figure 2.2 représente le graphe des résidus standardisés en fonction des prédictions. Les résidus laissent apparaître une tendance quadratique qui indique que la modélisation du phénomène par un modèle de régression linéaire simple n'est pas adaptée, l'analyse descriptive avait déjà permis de soulever ce point. Afin de prendre en compte cette relation quadratique, on va considérer un modèle de régression polynomiale d'ordre 2.

2.1.3 Modèle de régression polynomiale

Le modèle s'écrit :

$$Y_i = a + bx_i + cx_i^2 + E_i, \qquad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n = 15.$$

Le graphe des résidus en fonction des prédictions est donné en haut à gauche de la figure 2.3. On observe que la tendance quadratique a bien disparu. Cependant, on peut

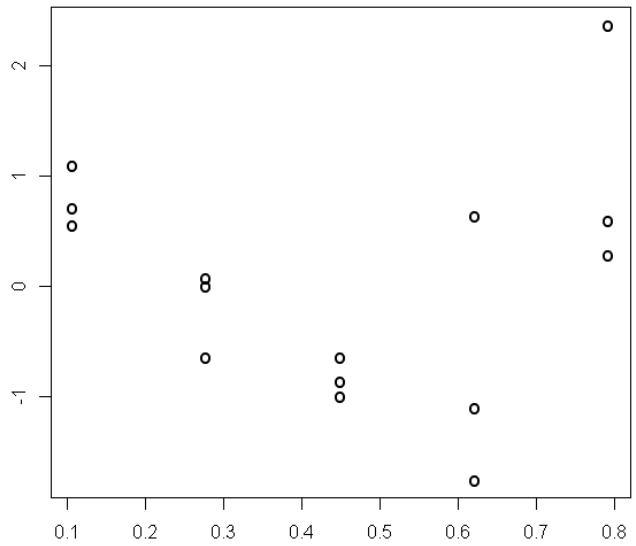


FIGURE 2.2 – Graphe des résidus (résidus standardisés en fonction des prédictions) dans le cadre de la régression linéaire simple.

remarquer, sur ce graphe comme sur le graphe de la racine carrée des résidus standardisés en fonction des prédictions (en bas à gauche), que l'hypothèse d'homoscédasticité n'est pas vérifiée (augmentation de la variabilité des résidus). En considérant une transformation logarithmique des données (figure 2.4), on observe que la transformation a bien joué son rôle de stabilisation de la variance. De plus, le graphique quantiles contre quantiles (en haut à droite), permet de vérifier que l'hypothèse de normalité des résidus est raisonnable ; et le graphique, en bas à droite, ne met en évidence aucun individu atypique (i.e. dont la distance de Cook serait plus grande que 1 par exemple).

Ainsi le modèle que l'on va considérer est le suivant :

$$\log Y_i = a + bx_i + cx_i^2 + E_i, \qquad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n = 15. \qquad \text{(M.2.1)}$$

Un modèle de régression polynomiale n'est autre qu'un cas particulier d'un modèle de régression multiple : il ne dépend en réalité que d'une seule variable x mais dont chaque puissance x^k est traitée comme une variable explicative dans un modèle de régression multiple.

Les estimations des différents paramètres sont données dans la table 2.2.

2.1.4 Influence de l'âge

La table 2.3 présente les résultats du test du modèle le plus simple contre le modèle complet, i.e. le test des hypothèses :

$$\mathrm{H}_0 = \{\log Y_i = a + E_i\} \quad \text{contre} \quad \mathrm{H}_1 = \{\log Y_i = a + bx_i + cx_i^2 + E_i\}.$$

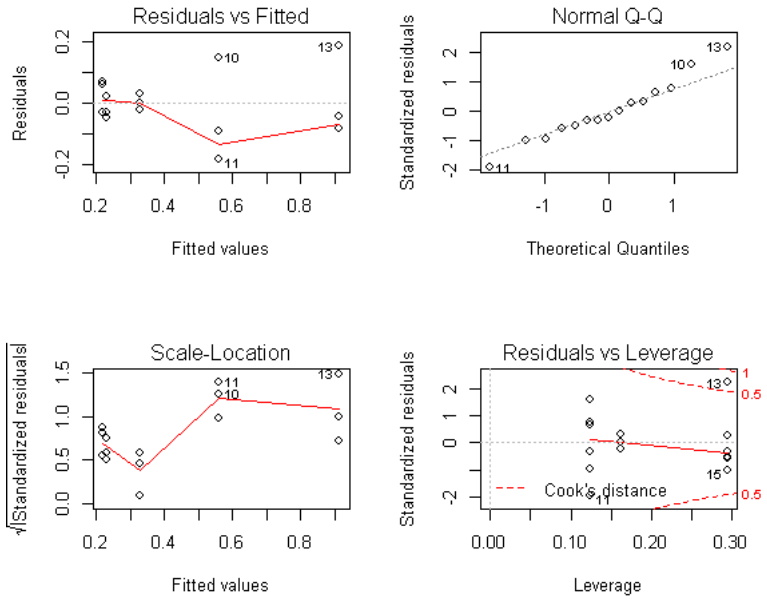


FIGURE 2.3 – Graphiques de diagnostic des résidus dans le cadre de la régression polynomiale sur Y.

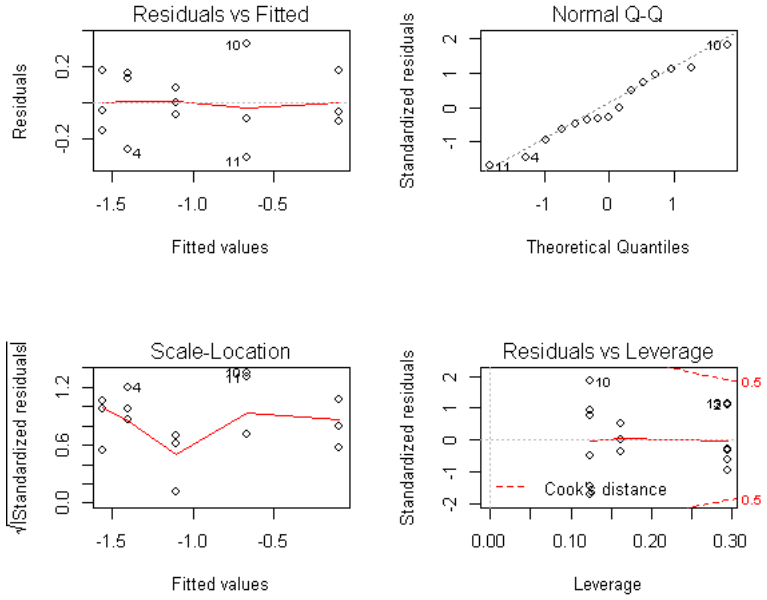


FIGURE 2.4 – Graphiques de diagnostic des résidus dans le cadre de la régression polynomiale sur $\log Y$.

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	-1.4548	0.4333	-3.36	0.0057
Age	-0.1945	0.2356	-0.83	0.4251
Age2	0.0704	0.0291	2.42	0.0325

TABLE 2.2 – Estimation des paramètres du modèle (M.2.1). La variable Age2 est l'âge au carré.

La probabilité critique est très proche de 0 et nous permet de conclure que l'âge a bien une influence sur le log du taux de DDT (au moins une des deux variables x ou x^2 contribue à expliquer le taux de DDT).

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	14	4.72				
2	12	0.43	2	4.29	60.18	0.0000

TABLE 2.3 – Table d'analyse de la variance du modèle (M.2.1).

Pour savoir si la relation est quadratique ou simplement linéaire, on cherche à étudier l'influence de la variable x^2 en plus de la variable x seule. Rappelons que les tests de l'effet des différentes variables peuvent être effectués à l'aide de différentes sommes des carrés (se référer à la partie 1.4.2, p. 29 du chapitre Modèle Linéaire). La table 2.4 donne les résultats des tests de type I et II. Pour tester si la relation qui lie le taux de DDT et l'âge est quadratique, il est nécessaire de tester l'apport de la variable x^2 dans un modèle qui contient déjà x. C'est le test qui est présenté sur la ligne `Age2` dans les tests de type I et II.

Type I					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	4.08	4.08	114.52	0.0000
Age2	1	0.21	0.21	5.84	0.0325
Residuals	12	0.43	0.04		

Type II				
	Sum Sq	Df	F value	Pr(>F)
Age	0.02	1	0.68	0.4251
Age2	0.21	1	5.84	0.0325
Residuals	0.43	12		

TABLE 2.4 – Tables d'analyse de la variance du modèle (M.2.1) : type I et type II.

Mais ce test d'intérêt peut aussi être traité comme un test sur le paramètre c dont les hypothèses sont :

$$\mathrm{H}_0 = \{c = 0\} \quad \text{contre} \quad \mathrm{H}_1 = \{c \neq 0\}$$

se lit dans la ligne Age2 de la table 2.2. On conclut qu'il existe bien une relation qua-

dratique entre le log du taux de DDT et l'âge des brochets.

Remarque 1. Les tests sur les paramètres donnés dans la table 2.2 et les tests de type II sont équivalents : chaque ligne correspond au test de l'influence de la variable correspondante en plus de toutes autres variables. On peut vérifier que les deux statistiques de test sont liées puisque $t^2 = F$.

Remarque 2. On voit qu'il existe une différence entre les deux types pour le test associé à la variable Age. En effet, en type I, le test porte sur l'influence de la variable Age par rapport à la constante, alors qu'en type II on cherche à savoir si la même variable âge a une influence en plus de la variable Age2 ; ce qui explique la différence des résultats.

2.1.5 Programme.

Programme R

```
########### #Lecture du jeu de données ###########
brochet=read.table('brochet.txt',header=TRUE,sep='')

#############Graphe du taux de DDT en fonction de l'âge ###########
plot(brochet$Age,brochet$DDT,col="blue",pch="+")

#############Modèle de régression linéaire simple ###########
#Modèle
reg<-lm(DDT~Age,data=brochet) #Diagnostic par(mfrow=c(2,2))
plot(reg)


#############Modèle de régression polynomiale sur le log ###########
# Création des variables Age2 et LogDDT
brochet.plus=data.frame(Age=brochet$Age,DDT=brochet$DDT, +
Age2=brochet$Age*brochet$Age,LogDDT=log(brochet$DDT))
#Modèle
poly<-lm(LogDDT~Age+Age2,data=brochet.plus)
#Diagnostic
par(mfrow=c(2,2)) plot(poly)
#Estimation des paramètres et tests sur les paramètres
summary(poly)
#Tests de type I et II
anova(poly)
Anova(poly)
```

Programme SAS

```
#############Lecture du jeu de données ###########
data BROCHET;
     infile 'brochet.txt' expandtabs firstobs=2;
     input Age TxDDT;
run;

#############Graphe du taux de DDT en fonction de l'âge ###########
proc gplot data=BROCHET;
     plot TxDDT * Age;
run; quit;
```

```
############Modèle de régression linéaire simple ###########
#Modèle
proc GLM data=BROCHET;
     model TxDDT = Age  / solution SS1 SS2;
     output out=REG p=Predite R=Residu STUDENT=ResidStandard;
run;
#Graphe des résidus
proc gplot data=REG;
    plot ResidStandard*Predite / vref=0;
run;quit;


############Modèle de régression polynomiale sur le log ###########
# Création des variables Age2 et LogDDT
data BROCHET; set BROCHET;
    Age2=Age*Age;
    LogDDT=log(TxDDT);
run;
# Modèle
proc GLM data=BROCHET;
     model LogDDT = Age Age2 / solution SS1 SS2;
     output out=POLY p=Predite R=Residu STUDENT=ResidStandard;
run;
#Graphe des résidus
proc gplot data=POLY;
    plot ResidStandard*Predite / vref=0;
run;quit;
```

2.2 Régression linéaire multiple

2.2.1 Problème et description des données

Objectif. La chenille processionnaire du pin est la chenille d'un papillon nocturne de la famille des Notodontidés. La chenille de ce papillon se développe de préférence sur des pins et peut causer des dégâts considérables. Pour tenter de comprendre les causes de l'infestation par cette chenille, on souhaite étudier l'influence de certaines caractéristiques de peuplements forestiers sur son développement.

Données. Ces données sont issues de [55]. On dispose d'un échantillon de $n = 33$ parcelles forestières d'une surface de 10 hectares. Chaque parcelle a été échantillonnée en placettes de 5 ares sur lesquelles plusieurs mesures ont été effectuées. La mesure retenue pour la parcelles est la moyenne de cette mesure sur les différentes placettes. Les différentes mesures effectuées sont les suivantes :

— l'altitude (en mètre) (variable `Altitude`),
— la pente (en degrés) (variable `Pente`),
— le nombre de pins (variable `NbPins`),
— la hauteur de l'arbre échantillonné au centre (variable `Hauteur`),
— le diamètre de cet arbre (variable `Diametre`),
— la note de densité de la végétation (variable `Densite`),
— l'orientation (variable `Orient`), allant de 1 (sud) à 2 (autre),

— la hauteur des arbres dominants (variable `HautMax`),
— le nombre de strates de végétation (variable `NbStrat`),
— le mélange du peuplement (variable `Melange`), allant de 1 (pas mélangé) à 2 (mélangé),
— le nombre de nids de processionnaires par arbre (variable `NbNids`).

Un extrait des données est présenté dans la table 2.5. Une remarque immédiate et surprenante face à ces données porte sur les parcelles 13 et 14 : il n'y a aucun pin alors que l'on observe des valeurs pour les autres variables, notamment le nombre de nids, qui est d'ailleurs assez élevé. Une explication possible est la suivante : pour chaque parcelle, le nombre de pin est la moyenne des mesures faites sur un échantillon de placettes de 5 ares, tandis que le nombre de nids est un nombre moyen par arbre. Si le nombre de pins moyen par parcelle était par exemple de 0.3, il se peut que cette valeur ait été arrondie à 0.

On peut alors s'interroger sur la pertinence de ces deux observations et sur la sensibilité des résultats à ces deux observations.

	Alti tude	Pente	Nb Pins	Hau- teur	Dia- metre	Den- site	Orient	Haut Max	Nb Strat	Me- lange	Nb Nids
1	1200	22	1	4.0	14.8	1.0	1.1	5.9	1.4	1.4	2.37
2	1342	28	8	4.4	18.0	1.5	1.5	6.4	1.7	1.7	1.47
3	1231	28	5	2.4	7.8	1.3	1.6	4.3	1.5	1.7	1.13
4	1254	28	18	3.0	9.2	2.3	1.7	6.9	2.3	1.6	0.85
5	1357	32	7	3.7	10.7	1.4	1.7	6.6	1.8	1.3	0.24
.../...											
12	1182	41	32	5.4	21.6	3.3	1.4	11.3	2.8	2.0	0.70
13	1179	15	0	3.2	10.5	1.0	1.7	4.0	1.1	1.6	2.64
14	1256	21	0	5.1	19.5	1.0	1.8	5.8	1.1	1.4	2.05
15	1251	26	2	4.2	16.4	1.1	1.7	6.2	1.3	1.8	1.75
.../...											
29	1208	23	2	3.5	11.5	1.1	1.7	5.4	1.3	2.0	1.09
30	1198	28	15	3.9	11.3	2.0	1.6	7.4	2.8	2.0	0.18
31	1228	31	6	5.4	21.8	1.3	1.7	7.0	1.5	1.9	0.35
32	1229	21	11	5.8	16.7	1.7	1.8	10.0	2.3	2.0	0.21
33	1310	36	17	5.2	17.8	2.3	1.9	10.3	2.6	2.0	0.03

TABLE 2.5 – Extrait des données.

Description des données. La table 2.6 présente les corrélations entre les différentes variables. Deux groupes au sein desquels les variables sont fortement corrélées se distinguent : {Hauteur, Diametre} et {Densite, NbStrat, NbPins, HautMax}. On remarque que les variables au sein de ces deux groupes sont corrélées négativement avec la variable d'intérêt, le NbNids. On pouvait s'attendre à retrouver de fortes liaisons entre certaines de ces variables, comme par exemple entre NbPins et Densité. D'autres moins évidentes apparaissent comme la corrélation positive entre NbPins et NbStrat.

L'utilisation directe du coefficient de corrélation pour interpréter les liens entre les variables n'est pas toujours pertinente. En effet, le lien entre deux variables peut venir du fait que ces deux variables sont liées à une troisième. On peut alors s'intéresser au coefficient de corrélation partielle (définie dans le chapitre Modèle Linéaire, équation 1.13, p. 24) qui permet, en effet, de mesurer la relation qui existe entre les deux va-

	Alti tude	Pente	Nb Pins	Hau- teur	Dia- metre	Den- site	Orient	Haut Max	Nb Strat	Me- lange	Nb Nids
Altitude	1.00	0.12	0.54	0.32	0.28	0.51	0.27	0.36	0.36	-0.13	-0.53
Pente	0.12	1.00	0.32	0.14	0.11	0.30	-0.15	0.26	0.33	0.13	-0.46
NbPins	0.54	0.32	1.00	0.41	0.29	**0.98**	0.13	**0.76**	**0.88**	0.19	-0.56
Hauteur	0.32	0.14	0.41	1.00	0.90	0.44	0.06	0.77	0.46	-0.12	-0.36
Diametre	0.28	0.11	0.29	0.90	1.00	0.31	-0.08	0.60	0.27	-0.09	-0.16
Densite	0.51	0.30	**0.98**	0.44	0.31	1.00	0.15	**0.81**	**0.91**	0.11	-0.57
Orient	0.27	-0.15	0.13	0.06	-0.08	0.15	1.00	0.06	0.06	0.13	-0.21
HautMax	0.36	0.26	**0.76**	0.77	0.60	**0.81**	0.06	1.00	**0.85**	0.00	-0.55
NbStrat	0.36	0.33	**0.88**	0.46	0.27	**0.91**	0.06	**0.85**	1.00	0.15	-0.64
Melange	-0.13	0.13	0.19	-0.12	-0.09	0.11	0.13	0.00	0.15	1.00	-0.11
NbNids	-0.53	-0.46	-0.56	-0.36	-0.16	-0.57	-0.21	-0.55	-0.64	-0.11	1.00

TABLE 2.6 – Matrice des corrélations entre variables. Deux groupes au sein desquels les variables sont fortement corrélées se distinguent (les corrélations au sein de ces groupes sont soulignées ou en gras).

riables corrigée de l'influence des autres variables. La table 2.7 présente les corrélations partielles entre les variables deux à deux, étant données toutes les autres. On remarque, par exemple, que la corrélation partielle entre NbPins et NbStrat est plus faible que la corrélation entre ces deux variables. Le lien fort entre ces deux variables était lié au lien qu'elles partagent avec la variable Densité.

	Alti tude	Pente	Nb Pins	Hau- teur	Dia- metre	Den- site	Orient	Haut Max	Nb Strat	Me- lange	Nb Nids
Altitude	1.00	-0.22	0.30	-0.16	0.26	-0.10	0.22	-0.06	-0.19	-0.38	-0.52
Pente	-0.22	1.00	0.13	-0.13	0.15	-0.06	-0.17	0.02	-0.06	-0.04	-0.46
NbPins	0.30	0.13	1.00	0.33	-0.23	**0.91**	-0.33	**-0.34**	**-0.04**	0.53	0.11
Hauteur	-0.16	-0.13	0.33	1.00	0.86	-0.42	0.41	0.59	0.07	-0.36	-0.27
Diametre	0.26	0.15	-0.23	0.86	1.00	0.30	-0.43	-0.20	-0.25	0.35	0.32
Densite	-0.10	-0.06	**0.91**	-0.42	0.30	1.00	0.40	**0.43**	**0.30**	-0.49	-0.01
Orient	0.22	-0.17	-0.33	0.41	-0.43	0.40	1.00	-0.18	-0.25	0.39	-0.06
HautMax	-0.06	0.02	**-0.34**	0.59	-0.20	**0.43**	-0.18	1.00	**0.38**	0.11	0.04
NbStrat	-0.19	-0.06	**-0.04**	0.07	-0.25	**0.30**	-0.25	**0.38**	1.00	0.17	-0.31
Melange	-0.38	-0.04	0.53	-0.36	0.35	-0.49	0.39	0.11	0.17	1.00	-0.18
NbNids	-0.52	-0.46	0.11	-0.27	0.32	-0.01	-0.06	0.04	-0.31	-0.18	1.00

TABLE 2.7 – Matrice des corrélations partielles. Les corrélations au sein des groupes précédemment identifiés comme présentant de forte corrélation sont souligées ou en gras.

2.2.2 Remarques.

Il est d'usage en régression multiple de travailler sur les données centrées réduites afin de rendre certains résultats directement interprétables. En effet, dans ce cas,

1. les coefficients de régression, qui sont en fait des coefficients de régression partielle, au sens où chacun mesure l'effet de la variable explicative associée sur la variable à expliquer lorsque les autres sont contrôlées, seront insensibles à l'unité ou l'échelle des variables et donc directement interprétables et comparables.
2. la matrice $X'X$ est égale à la matrice de corrélation empirique entre les variables à n près. Rappelons que pour obtenir les estimations des paramètres θ, il faut

inverser cette matrice. Ainsi plus la corrélation entre variables est importante, plus on aura des problèmes de conditionnement [1]

3. la matrice de variance des estimateurs des paramètres correspond exactement à l'opposé de la matrice des corrélations partielles (cf chapitre Modèle Linéaire, 1, p. 12). Ainsi les paramètres de variables fortement corrélées partiellement seront moins bien estimés (grande variance de leurs estimateurs) rendant leur interprétation plus que délicate.

Dans toute la suite, on travaillera sur les données centrées réduites.

2.2.3 Modèle

On cherche à expliquer le nombre de nids de processionnaires à partir de $p = 10$ variables explicatives toutes quantitatives décrites précédemment. On considère donc le modèle de régression linéaire multiple dont on rappelle l'écriture ci-dessous :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + E_i \qquad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, 33. \quad \text{(M.2.2)}$$

Le graphe des résidus standardisés en fonction des prédictions, donné figure 2.5 (à gauche) indique une tendance dans la moyenne et dans la variance, corrigées par le passage au logarithme du nombre de nids (cf figure 2.5 (à droite)). Ainsi le nouveau modèle considéré est le modèle (M.2.2) où Y_i représente maintenant le logarithme du nombre de nids pour la $i$ème observation.

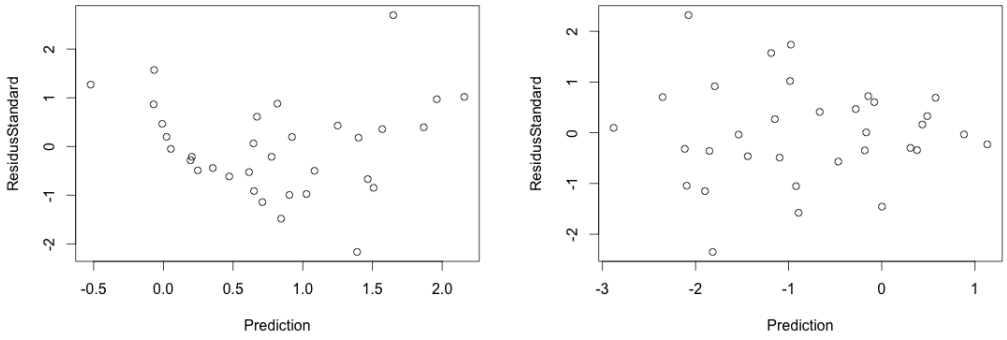


FIGURE 2.5 – Graphe des résidus dans le cadre de la régression multiple sur Y (à gauche) et sur $\log Y$ (à droite).

Les autres graphes de diagnostic donnés dans la figure 2.6 montrent que l'hypothèse de normalité est raisonnable, et qu'aucun individu n'est atypique.

1. Situation où une matrice théoriquement inversible pose des problèmes numériques lors de l'inversion car le rapport entre sa plus grande et sa plus petite valeur propre est très important. On parle de mauvais conditionnement de la matrice.

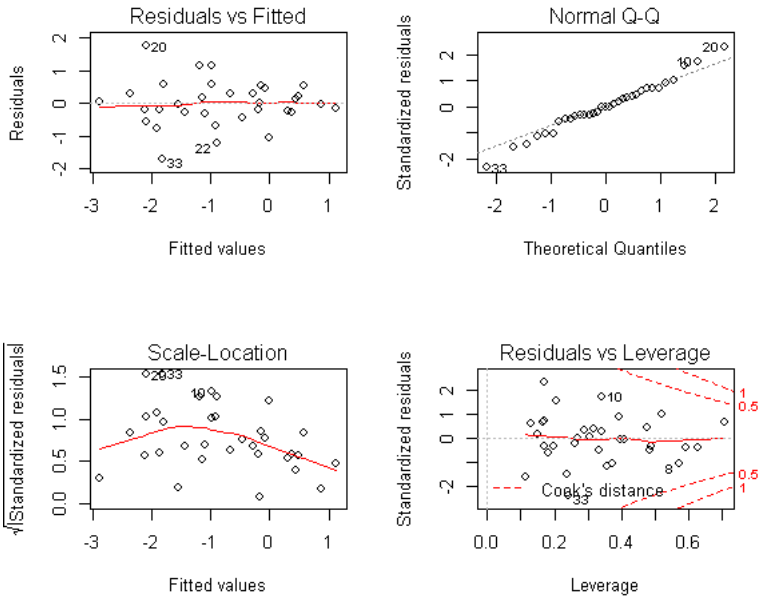


FIGURE 2.6 – Graphiques de diagnostic des résidus dans le cadre de la régression multiple sur $\log Y$.

2.2.4 Influence des différentes variables sur le nombre de nids

Le test du modèle complet, i.e. des hypothèses :

$$\mathrm{H}_0 = \{Y_i = a + E_i\} \quad \text{contre} \quad \mathrm{H}_1 = \{Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + E_i\},$$

ou encore en termes de paramètres :

$$\mathrm{H}_0 = \{\beta_1 = \beta_2 = \ldots = \beta_p = 0\} \quad \text{contre} \quad \mathrm{H}_1 = \{\exists j \mid \beta_j \neq 0\},$$

donne une probabilité critique égale à 0.0007441, l'hypothèse H_0 est rejetée indiquant qu'au moins une des variables contribue à expliquer le log du nombre de nids. La question qui se pose ensuite est de savoir si il y en a plusieurs et le cas échéant, lesquelles.
La table 2.2.4 donne les estimations des différents paramètres ainsi que les résultats des tests de nullité de ces paramètres. Il faut noter que le test portant sur une variable j consiste à tester les hypothèses suivantes :

$$\mathrm{H}_0 = \{\beta_j = 0\} \quad \text{contre} \quad \mathrm{H}_1 = \{\beta_j \neq 0\}$$

ou encore en termes de modèles :

$$\mathrm{H}_0 = \{Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{j-1} x_{(j-1)i} + \beta_{j+1} x_{(j+1)i} + \ldots + \beta_p x_{pi} + E_i\}$$

contre

$$\mathrm{H}_1 = \{Y_i = \beta_0 + \beta_1 x_{1i} + .. + \beta_{j-1} x_{(j-1)i} + \beta_j x_{ji} + \beta_{j+1} x_{(j+1)i} + .. + \beta_p x_{pi} + E_i\}.$$

Ainsi, ce test consiste à regarder si la variable x_j a un effet sur Y en présence de toutes les autres variables. On peut remarquer que le test portant sur la variable la plus corrélée au nombre de nids, à savoir NbStrat, n'est pas significatif. Ceci est dû à la redondance d'information. En effet, cette variable est fortement corrélée à d'autres variables (Densite, NbPins), l'information est donc déjà présente dans le modèle H_0 : ajouter le NbStrat n'apporte donc aucune information supplémentaire. Ainsi éliminer, à partir de ces tests, toutes les variables dont les coefficients sont significativement nuls est incorrect. Une procédure adaptée serait d'éliminer au fur et à mesure les variables non explicatives (i.e. dont la p.value est la plus grande) jusqu'à n'obtenir que des tests significatifs. Cependant quand le nombre de variables est assez élevé, comme c'est le cas ici, cette procédure "à la main" peut être assez longue et il existe des procédures de sélection de variables disponibles dans les logiciels à cet effet.
Notons que ces tests sont des tests de modèles emboîtés et qu'ils correspondent exactement (dans le cas de la régression multiple) aux tests de type II (cf chapitre Modèle Linéaire).

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	-0.81	0.14	-5.65	0.00
Altitude	-0.58	0.20	-2.88	0.01
Pente	-0.39	0.16	-2.45	0.02
NbPins	0.71	0.96	0.74	0.46
Hauteur	-1.38	0.59	-2.33	0.03
Diametre	1.02	0.45	2.26	0.03
Densite	-0.32	1.13	-0.29	0.78
Orient	-0.04	0.19	-0.19	0.85
HautMax	0.44	0.56	0.79	0.44
NbStrat	-0.72	0.49	-1.47	0.16
Melange	-0.13	0.19	-0.69	0.49

TABLE 2.8 – Estimation des paramètres du modèle.

2.2.5 Sélection de variables

La sélection de variables est une perspective naturelle à plus d'un titre.
Il y a deux situations :

— soit certaines ne contribuent pas à l'explication de la variable à expliquer et ces variables n'ont pas d'intérêt dans le modèle,
— soit des variables sont très corrélées entre elles et apportent donc une redondance d'information, dans ce cas il n'est pas pertinent de les mettre toutes.

Il faut bien noter que l'on cherche toujours à privilégier le modèle le plus simple possible permettant ainsi une interprétation facile et pouvant éviter à l'expérimentateur des coûts d'acquisition de certaines données s'il souhaite prédire Y. De plus, un trop grand nombre de variables conduit d'une part à une imprécision dans l'estimation des coefficients de régression et d'autre part peut mener à une augmentation de la variance résiduelle puisque le nombre de degrés de libertés diminue, celui-ci étant donné par $n-p-1$, p étant le nombre de variables présentes dans le modèle. L'objectif est donc de déterminer à partir de toutes les variables explicatives un sous-ensemble où chacune

apporte sa propre information (une information nouvelle). Il faut donc se donner un critère de sélection afin d'évaluer chaque modèle et de choisir le meilleur (au sens donc de ce critère). Une première possibilité consisterait alors à évaluer et donc comparer tous les modèles possibles et choisir celui qui optimise le critère. Malheureusement, cette solution est peut-être très longue, voire impossible dès lors que le nombre de variables est très grand (le nombre de régressions possibles étant de 2^p). Nous présentons ici les méthodes les plus utilisées dans des cas où cette recherche est impossible. Ces méthodes sont dites "pas à pas" dans lesquelles les variables sont introduites ou supprimées dans la régression l'une après l'autre au sens du critère choisi. Ces méthodes sont des heuristiques et aucune ne garantit d'atteindre la solution optimale.

Critères. Plusieurs critères sont possibles (et différents selon les logiciels), comme par exemple, le R^2 ajusté, le critère de sélection de modèles AIC... (cf partie 1.6, p. 41 du chapitre Modèle Linéaire pour la définition de ces critères, p.41).

Méthodes pas à pas. Il existe 3 stratégies :

- **Ascendante ou forward**. Cette stratégie consiste à partir du modèle le plus simple ($Y_i = \beta_0 + E_i$) et à ajouter au fur et à mesure la variable qui améliore le plus critère. L'algorithme s'arrête dès qu'il n'existe plus de variables dans ce cas.
- **Descendante ou backward**. C'est la stratégie symétrique de la précédente qui consiste à partir du modèle complet et à éliminer au fur et à mesure la variable la moins informative au vu du critère considéré et des autres variables. De la même façon que précédemment, l'algorithme s'arrête dès que la suppression de la variable n'améliore plus le critère.
- **Stepwise**. Cette stratégie combine les deux stratégies précédentes dans le sens où, à chaque étape on considère la possibilité d'ajouter une variable ou d'en retirer une. L'opération qui produit le meilleur critère est effectuée. Le modèle initial est au choix le modèle nul, le modèle complet ou n'importe quel autre modèle candidat.

Remarque pour les logiciels R et SAS. Dans cet exemple, on a choisi d'optimiser le critère AIC (comme le logiciel `R`). On peut également choisir d'optimiser le R^2_{adj}. Le logiciel `SAS` permet aussi d'obtenir la meilleure régression pour un nombre donné de variables ("best subset") pour les critères R^2, R^2_{adj}, ou CP (le critère du CP de Mallows, version moindres carrés du critère AIC). Pour les procédures "pas à pas", il procède en deux temps : il propose d'inclure ou de supprimer une variable selon le critère (par exemple la corrélation partielle), puis effectue ensuite un test de nullité des paramètres associés aux variables du modèle en cours (un test non significatif arrête les algorithmes).

Résultat pour le nombre de nids. La table 2.9 présente les différentes étapes de la stratégie stepwise obtenues par le logiciel R. Le modèle sélectionné est le suivant :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + E_i \qquad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \tag{2.1}$$

où x_1 est la variable NbStrat, x_2 l'Altitude, x_3 la Pente, x_4 le NbPins et x_5 l'Orientation. La table 2.10 présente les estimations des différents paramètres de ce modèle ainsi

que les tests de nullité de ces paramètres. On peut remarquer que la variable Orient n'apporte pas d'information supplémentaire par rapport aux autres variables (le test est non significatif) et pourrait donc être enlevée de ce modèle. Au regard des estimations des paramètres, on peut conclure que pour avoir le moins de parasites possible par pins, il faut qu'il y ait beaucoup de végétation, que le site soit en altitude, de pente forte et qu'il y ait peu de pins.

2.2.6 Programme.

Programme R

```
########## Lecture du jeu de données
Chenilles <- read.table("Chenilles.txt",header=TRUE, sep=" ")
head(Chenilles)
########## Création de la variable log(NbNids)
LogNbNids=log(Chenilles$NbNids)
########## Graphe de la variable à expliquer en fonction des autres
par(mfrow=c(1,1))
plot(Chenilles$NbStrat,Chenilles$NbNids)
plot(log(Chenilles$NbStrat),Chenilles$NbNids)
plot(Chenilles$Pente,Chenilles$NbNids)
```

```
Start:  AIC=15.44
LogNbNids ~ 1

           Df Sum of Sq    RSS     AIC
+ NbStrat   1   17.4987 32.097  3.0848
+ HautMax   1   14.5358 35.060  5.9985
+ Altitude  1   14.1222 35.474  6.3855
+ Densite   1   13.8412 35.755  6.6459
+ NbPins    1   13.2969 36.299  7.1444
+ Pente     1    9.1464 40.450 10.7171
+ Hauteur   1    8.9707 40.625 10.8602
<none>                 49.596 15.4443
+ Orient    1    2.6164 46.980 15.6558
+ Diametre  1    2.0025 47.594 16.0842
+ Melange   1    0.0659 49.530 17.4004

Step:  AIC=3.08
LogNbNids ~ NbStrat

           Df Sum of Sq    RSS     AIC
+ Altitude  1    5.7642 26.333 -1.4474
+ Pente     1    3.0899 29.007  1.7445
<none>                 32.097  3.0848
+ Orient    1    1.8378 30.260  3.1390
+ Hauteur   1    1.4585 30.639  3.5501
+ HautMax   1    0.2153 31.882  4.8627
+ Melange   1    0.1338 31.964  4.9470
+ Diametre  1    0.0945 32.003  4.9874
+ Densite   1    0.0368 32.061  5.0469
+ NbPins    1    0.0020 32.095  5.0828
- NbStrat   1   17.4987 49.596 15.4443

Step:  AIC=-1.45
LogNbNids ~ NbStrat + Altitude

           Df Sum of Sq    RSS     AIC
+ Pente     1    3.0701 23.263 -3.5382
+ Densite   1    2.2734 24.060 -2.4269
+ NbPins    1    1.9416 24.392 -1.9749
<none>                 26.333 -1.4474
+ Hauteur   1    0.6000 25.733 -0.2081
+ Orient    1    0.5599 25.773 -0.1566
+ HautMax   1    0.0482 26.285  0.4922
+ Diametre  1    0.0383 26.295  0.5046
+ Melange   1    0.0119 26.321  0.5376
- Altitude  1    5.7642 32.097  3.0848
- NbStrat   1    9.1407 35.474  6.3855

...... ...... ......

Step:  AIC=-5.44
LogNbNids~NbStrat+Altitude+Pente+NbPins+Orient

           Df Sum of Sq    RSS     AIC
<none>                 19.454 -5.4382
- Orient    1    1.3806 20.835 -5.1757
+ Hauteur   1    0.5057 18.949 -4.3074
+ Densite   1    0.2830 19.172 -3.9218
+ HautMax   1    0.0867 19.368 -3.5855
+ Melange   1    0.0138 19.441 -3.4616
+ Diametre  1    0.0000 19.454 -3.4382
- NbPins    1    2.5935 22.048 -3.3084
- Pente     1    4.3177 23.772 -0.8237
- Altitude  1    6.4076 25.862  1.9570
- NbStrat   1    6.6742 26.129  2.2954
```

TABLE 2.9 – Les différentes étapes de la méthode de sélection de variables stepwise obtenues avec le logicel R.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-0.81	0.15	-5.50	0.00
NbStrat	-0.99	0.32	-3.04	0.01
Altitude	-0.56	0.19	-2.98	0.01
Pente	-0.40	0.16	-2.45	0.02
NbPins	0.68	0.36	1.90	0.07
Orient	-0.22	0.16	-1.38	0.18

TABLE 2.10 – Estimation des paramètres et test de nullité des paramètres pour le modèle sélectionné.

```
plot(log(Chenilles$Pente),Chenilles$NbNids)
plot(Chenilles$Hauteur,Chenilles$NbNids)
plot(Chenilles$Diametre,Chenilles$NbNids)
######### Matrice de corrélations
cor(Chenilles)
######## Matrice de corrélations partielles
library(ppcor)  #Installation du package ppcor
pcor(Chenilles)

########## Données centrées réduites
Chenilles.cr=as.data.frame(scale(as.matrix(Chenilles)))
########## Modèle de régression multiple sur les données centrées réduites
Nids.lm=lm(NbNids~Altitude+Pente+NbPins+Hauteur+Diametre+Densite+Orient+...
... HautMax+NbStrat+Melange,data=Chenilles.cr)
par(mfrow=c(2,2))
plot(Nids.lm)

Nids.log.lm=lm(LogNbNids~Altitude+Pente+NbPins+Hauteur+Diametre+Densite+...
... Orient+HautMax+NbStrat+Melange,data=Chenilles.cr)
par(mfrow=c(2,2))
plot(Nids.log.lm)
summary(Nids.log.lm)

########## Sélection de variables
model.nul=lm(LogNbNids~1,data=Chenilles.cr)
select.variables=step(model.nul,scope=~Altitude+Pente+NbPins+Hauteur+Diametre+...
... Densite+Orient+HautMax+NbStrat+Melange,direction="both",data=Chenilles.cr)
summary(select.variables)
```

Programme SAS

```
########## Lecture du jeu de données
data Chenilles;
     infile 'Chenilles.txt' expandtabs firstobs=2;
     input Altitude Pente NbPins Hauteur Diametre Densite Orient HautMax NbStrat ...
     ... Melange NbNids LogNids;
run;
########## Graphe de la variable à expliquer en fonction des autres
proc gplot data=Chenilles;
plot NbNids*(Altitude Pente Diametre);
run;
######### Matrice de corrélations
proc Corr data=Chenilles;
run;
```

```
########## Données centrées réduites
proc standard data=Chenilles out=Chenilles_cr mean=0 std=1;
Var Altitude Pente NbPins Hauteur Diametre Densite Orient HautMax NbStrat Melange;
run;

########## Modèle de régression multiple sur les données centrées réduites
proc Reg data=Chenilles_cr;
     model NbNids = Altitude Pente NbPins Hauteur Diametre Densite Orient
                    HautMax NbStrat Melange;
     plot residual. * predicted. / vref=0;
run;

proc Reg data=Chenilles_cr;
     model LogNids = Altitude Pente NbPins Hauteur Diametre Densite Orient ...
                 ...     HautMax NbStrat Melange / SS1 SS2;
     plot residual. * predicted. / vref=0;
run;

########## Sélection de variables
proc Reg data=Chenilles_cr;
     model LogNids = Altitude Pente NbPins Hauteur Diametre Densite Orient ...
                  ...    HautMax NbStrat Melange
                      / selection=stepwise slentry=0.05 slstay=0.05;
run;

##########
```

2.3 Analyse de la variance à un facteur

2.3.1 Description du problème et des données

L'objectif est d'étudier l'influence du statut de domination d'un arbre sur son diamètre. Ici on s'intéresse, en particulier, aux alisiers pour lesquels 3 statuts sont définis : co-dominant (de la même hauteur qu'un ou plusieurs arbres voisins), dominant (plus haut que les arbres avoisinants) et dominé (plus bas que les arbres avoisinants). Pour cela, on dispose d'un échantillon de $n = 104$ arbres choisis aléatoirement en forêt de Rambouillet (Yvelines). Sur chacun d'entre eux, on a mesuré son diamètre en cm (variable `Diametre`) et on a noté son statut (variable `Statut`). Un extrait des données recueillies est présenté dans la table 2.11.

La table 2.12 donne la répartition des observations dans les différents statuts. On observe un déséquilibre entre ces effectifs. L'échantillon ayant été tiré aléatoirement, on peut supposer que ces effectifs reflètent les proportions réelles des différents statuts dans la forêt considérée. Sur la figure 2.7, on observe une légère différence des diamètres moyens surtout entre le statut dominant et les deux autres : $y_{1\bullet} = 28.7$, $y_{2\bullet} = 35.1$ et $y_{3\bullet} = 29.2$ où les indices $1, 2$ et 3 correspondent respectivement aux statuts co-dominant, dominant et dominé. La variabilité semble être à peu près la même dans les différents groupes de statuts (l'écart-type vaut 16.3 cm pour les co-dominants, 18.7 pour les dominants et 13.2 cm pour les dominés).

	Diametre	Statut
1	8.0	codomina
2	15.0	codomina
3	22.0	codomina
4	20.0	codomina
5	17.0	codomina
../..		
100	16.0	domine
101	10.0	domine
102	52.0	domine
103	9.0	domine
104	8.0	domine

TABLE 2.11 – Extrait des données.

	Effectif	
codomina	25	n_1
dominant	15	n_2
domine	64	n_3

TABLE 2.12 – Effectifs.

Le déséquilibre des effectifs aura plusieurs conséquences : (i) le test du modèle (et donc de l'effet du statut), qui est un test global, ne sera pas de puissance maximale, (ii) il en sera de même pour le test de toutes les comparaisons possibles des diamètres moyens, (iii) la précision de l'estimateur du diamètre moyen par statut étant de $1/n_i$ pour le statut i (cf chapitre modèle linéaire, discussion sur la variance de $\widehat{\theta}$), elle ne sera pas la même pour les différents statuts.

2.3.2 Effet du statut

Modèle.

Le `Statut` est un facteur qui possède 3 niveaux (variable qualitative). Ainsi pour étudier son influence sur le diamètre, on va considérer un modèle d'analyse de la variance dont on rappelle ci-dessous l'écriture singulière (utilisée par tous les logiciels) :

$$Y_{ik} = \mu + \alpha_i + E_{ik}, \qquad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{M.2.3}$$

où l'indice i représente le statut (1 = co-dominant, 2 = dominant, 3 = dominé), l'indice k est le numéro de l'arbre au sein du statut i ($k = 1, \ldots, n_i$) et la variable Y_{ik} désigne le diamètre (supposé aléatoire) du k-ème arbre de i-ème statut.

La figure 2.8 nous permet de valider l'hypothèse d'homoscédasticité.

Estimation des paramètres.

Dans l'écriture régulière du modèle d'analyse de la variance à 1 facteur :

$$Y_{ik} = \mu_i + E_{ik}, \qquad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{M.2.4}$$

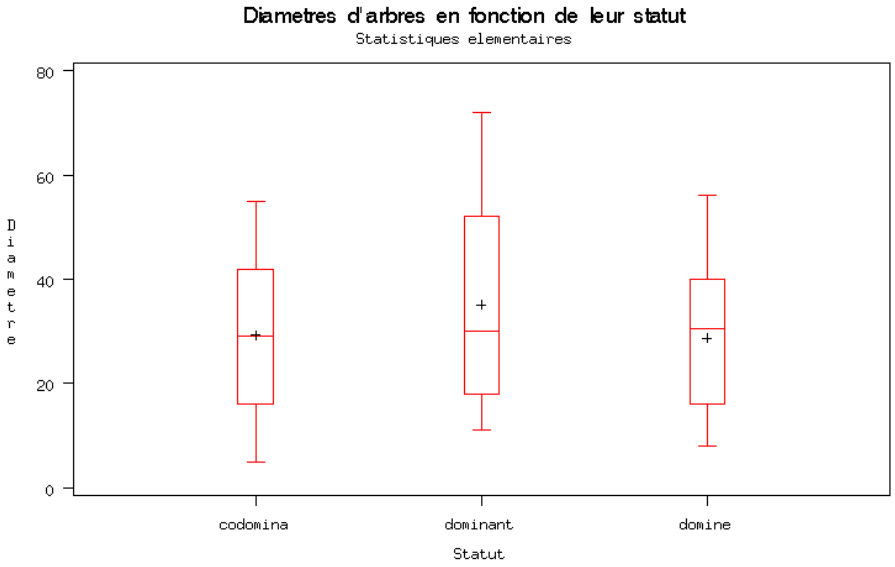


FIGURE 2.7 – Répartition du diamètre en fonction du statut de dominance des alisiers.

les paramètres μ_i ont des estimateurs naturels évidents (qui sont aussi ceux des moindres carrés et du maximum de vraisemblance) : on estime l'espérance du diamètre dans chaque statut par le diamètre moyen observé dans ce statut, soit

$$\hat{\mu}_i = Y_{i\bullet},$$

où $Y_{i\bullet} = \sum_{k=1}^{n_i} Y_{ik}/n_i$.
L'estimation des paramètres μ et α_i du modèle (M.2.3) est plus problématique à cause de la non-identifiabilité de ce modèle. On sait, d'après la partie 1.3.1 du chapitre Modèle linéaire qu'il faut appliquer une contrainte à ces paramètres pour pouvoir les estimer et que dans ce cas l'interprétation des valeurs estimées dépend de ces contraintes (ainsi que les tests sur ces paramètres). Différentes contraintes peuvent être considérées :

— $\boldsymbol{\alpha_I = 0}$ est la contrainte utilisée par le logiciel SAS pour des raisons de simplicité numérique, puisqu'elle revient à supprimer la dernière colonne de la matrice X qui devient, de ce fait, de plein rang. Elle aboutit aux estimateurs

$$\hat{\mu} = Y_{I\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{I\bullet}.$$

Par construction, on a donc $\hat{\alpha}_I = 0$. Les estimations de α_i s'interprètent comme des écarts à un groupe de référence qui est choisi ici comme étant le dernier, ici le statut 3, c'est à dire le satut `domine`.
La table 2.14 donne les estimations obtenues avec cette contrainte, ainsi que leurs écarts types. La note au bas de cette table nous rappelle que ces estimations ne sont pas les seules possibles. Cette table propose pour chaque paramètre un test de nullité de ce paramètre. Par exemple, dans la ligne 1 (intercept), l'hypothèse $H_0 = \{\mu = 0\}$ est rejetée (probabilité critique < 0.0001). Puisque $\alpha_I = 0$, alors μ est le diamètre moyen des arbres dominés et le rejet de H_0 signifie que le diamètre moyen des arbres dominés est significativement non nul. Dans cet

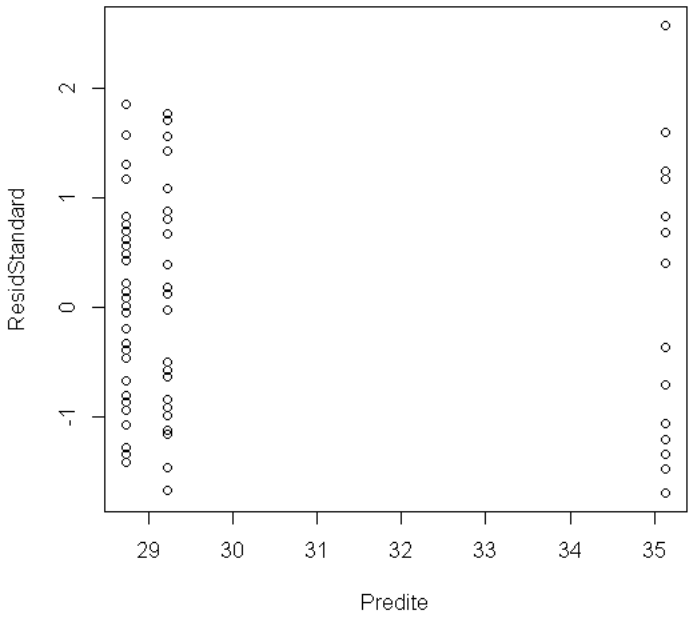


FIGURE 2.8 – Graphe des résidus standardisés.

exemple, comme dans beaucoup d'autres, ce test ne présente aucun intérêt. Dans la ligne 2 (Statut codomina), on teste l'hypothèse $H_0 = \{\alpha_1 = 0\}$ qui est acceptée (probabilité critique = 0.89), ce qui signifie que le diamètre moyen des arbres co-dominants n'est pas significativement différent de celui des arbres dominés.

— $\boldsymbol{\alpha_1 = 0}$ est la contrainte utilisée par le logiciel R. Le groupe de référence est ici le statut 1 et les expressions des estimateurs sont alors les suivantes :

$$\hat{\mu} = Y_{1\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{1\bullet}.$$

Les estimations de α_i s'interprètent comme des écarts au groupe 1.

— $\boldsymbol{\sum_i \alpha_i = 0}$ est sans doute la contrainte la plus naturelle, puisqu'elle suppose que les effets des différents niveaux se compensent globalement. Si le dispositif est *équilibré* ($n_1 = n_2 = \cdots = n_I$), on aboutit aux estimateurs, naturels eux aussi :

$$\hat{\mu} = Y_{\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{\bullet\bullet}.$$

μ est alors estimé par la moyenne générale et α_i par l'écart entre la moyenne du groupe i et la moyenne générale.
Cependant, dans le cas *déséquilibré* (comme dans l'exemple de ce chapitre), cette contrainte donne

$$\hat{\mu} = \frac{1}{I}\sum_i Y_{i\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - \hat{\mu}.$$

Dans ce cas, μ n'est plus estimé par la moyenne générale, mais par une moyenne qui donne un poids différent aux individus des différents groupes. L'interprétation

des $\hat{\alpha_i}$ est alors plus délicate : ce sont des écarts à une valeur moyenne, qui n'est pas la moyenne générale.

— $\sum_i \boldsymbol{n_i \alpha_i = 0}$ donne toujours les estimateurs naturels :

$$\hat{\mu} = Y_{\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{\bullet\bullet}.$$

Les $\hat{\alpha}_i$ sont alors les écarts à la moyenne générale. La définition de la contrainte montre cependant que ces estimateurs "naturels" peuvent donner des valeurs plus grandes aux $\hat{\alpha}_i$ des petits groupes, puisque ces derniers pèsent peu dans la moyenne générale.

La table 2.13 rappelle les effectifs et moyennes par statut et donne les estimations des paramètres obtenues avec ces différentes contraintes. On voit que, par exemple, $\hat{\alpha}_1$ change de signe selon les contraintes considérées ($\alpha_I = 0$ et $\sum_i \alpha_i = 0$), rappelant que son interprétation n'a de sens que via la contrainte utilisée.

		global	statut 1	statut 2	statut 3
effectif		104	25	15	64
moyenne		29.77	29.22	35.13	28.73
modèle	contrainte	–	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\mu_3}$
$(M.2.4)$	–	–	29.22	35.13	28.73
modèle	contrainte	$\hat{\mu}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
$(M.2.3)$	$\alpha_I = 0$	28.73	+0.49	+6.40	0
$(M.2.3)$	$\alpha_1 = 0$	29.22	0	+5.9133	−0.4856
$(M.2.3)$	$\sum_i \alpha_i = 0$	31.03	−1.81	+4.10	−2.29
$(M.2.3)$	$\sum_i n_i\alpha_i = 0$	29.77	−0.55	+5.36	−1.04

TABLE 2.13 – Estimations des paramètres μ et α_i avec différentes contraintes.

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		28.73437500 B	1.85268310	15.51	<.0001
Statut	codomina	0.48562500 B	3.49563548	0.14	0.8898
Statut	dominant	6.39895833 B	4.25176308	1.51	0.1354
Statut	domine	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

TABLE 2.14 – Estimation des paramètres du modèle dans le logiciel SAS.

Test de l'effet du statut.

Il est important de noter que le test de l'effet statut étudié ici est une notion *globale*. On étudie l'effet du statut *en général* sur le diamètre et non, par exemple, l'effet du statut "dominant".

On souhaite tester les hypothèses suivantes, écrites en termes de paramètres puis de modèles :

$$H_0 = \{\alpha_1 = \alpha_2 = \alpha_3 = 0\} = \{Y_{ik} = \mu + E_{ik}\}$$

$$\text{contre} \quad H_1 = \{\exists i : \alpha_i \neq 0\} = \{Y_{ik} = \mu + \alpha_i + E_{ik}\}$$

Dans le cadre d'une analyse de la variance à 1 seul facteur, on voit que le test de l'effet de ce facteur est exactement le test du modèle complet. La table 2.15 donne la table d'analyse de la variance avec le résultat de ce test. Les expressions de ces différentes quantités sont données dans la table 2.16.

	DF	Sum of Squares	Mean Square	F value	Pr > F
Model	2	507.68	253.84	1.16	0.3190
Error	101	22187.25	219.67		
Total	103	22694.93			

TABLE 2.15 – Table d'analyse de la variance décrivant l'effet du statut sur le diamètre des alisiers.

	DDL	Somme de carrés	Carré moyen	Stat. de test	Proba. critique
SCM	$I-1$	$\sum_i \sum_k (Y_{i\bullet} - Y_{\bullet\bullet})^2$	$SCM/(I-1)$	$F = \frac{SCM/(I-1)}{SCR/(n-I)}$	$\Pr\{\mathcal{F}_{I-1,n-I} > F\}$
SCR	$n-I$	$\sum_i \sum_k (Y_{ik} - Y_{i\bullet})^2$	$\hat{\sigma}^2 = SCR/(n-I)$		
SCT	$n-1$	$\sum_i \sum_k (Y_{ik} - Y_{\bullet\bullet})^2$			

TABLE 2.16 – Définition des sommes de carrés et carrés moyen dans le modèle d'analyse de la variance à un facteur.

La statistique de test F observée vaut 1.16, ce qui indique que la variabilité due au statut est du même ordre de grandeur que la variabilité résiduelle. La probabilité critique vaut 0.319, on ne rejette pas l'hypothèse H_0 et on conclut que le statut n'a pas d'effet significatif sur le diamètre.
Ce résultat ne prouve pas l'absence d'effet du statut sur le diamètre, il signifie que la variabilité individuelle du diamètre des arbres peut tout à fait produire "par hasard" les différences de diamètres moyens observés entre les statuts. Il peut exister un effet statut mais cette expérience ne permet pas de le détecter.

2.3.3 Comparaison des groupes de statuts.

L'analyse précédente révèle que l'effet du statut sur le diamètre n'est pas significatif, ce qui conclut l'étude. Dans un cas contraire où l'effet du facteur est significatif, il pourrait être intéressant de comparer les réponses moyennes entre les différents groupes afin de voir ce qui rend le facteur significatif.

Comparaison de deux groupes. On sait (cf [18], p. 70) que dans un modèle gaussien avec variance homogène, le test de l'hypothèse $H_0 = \{\mu_i = \mu_j\}$ contre $H_1 = \{\mu_i \neq \mu_j\}$

est fondé sur la statistique de test

$$T_{ij} = (Y_{i\bullet} - Y_{j\bullet}) \Big/ \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

où $\hat{\sigma}^2$ est la variance (supposées commune) estimée sur les mesures effectuées dans les deux groupes. Sous H_0, cette statistique de test suit une loi de student à $n_i + n_j - 2$ degrés de liberté :

$$T_{ij} \underset{H_0}{\sim} \mathcal{T}_{n_i+n_j-2}.$$

On sait enfin que la puissance de ce test croît avec le nombre de degrés de liberté qui rend compte de la précision avec laquelle sont estimées les espérances μ_i et μ_j.

Modèle d'analyse de la variance. Dans le cadre du modèle d'analyse de la variance, on peut améliorer la puissance de ces tests en utilisant une estimation de la variance fondée sur l'ensemble des groupes. Cette amélioration n'a de sens qu'à cause de l'hypothèse d'homoscédasticité. Pour comparer les groupes i et j, on utilise la même statistique de test T_{ij} mais en utilisant la variance estimée sur l'ensemble des données ($\hat{\sigma}^2$ donnée dans la table d'analyse de la variance (table 2.16)). Ainsi, les données issues des autres groupes contribuent à la définition de la statistique de test. Dans ce cas, T_{ij} suit une loi de student à $n - I$ degrés de libertés :

$$T_{ij} \underset{H_0}{\sim} \mathcal{T}_{n-I}.$$

Comparaisons multiples. On veut maintenant comparer tous les groupes deux à deux. En comparant I groupes, on effectue $I(I-1)/2 = 3$ comparaisons. Pour chaque comparaison, on prend une décision (acceptation ou rejet de H_0) aléatoire qui peut correspondre à une erreur. On se trouve confronté ici à un problème de tests multiples : en effectuant chacune des comparaisons à un niveau $\alpha = 5\%$, on encourt globalement un risque α^* de se tromper au moins une fois qui est supérieur à α.

On remarque que H_0 s'écrit sous la forme

$$H_0 = \bigcap_{i=1}^{I-1} \bigcap_{j=i+1}^{I} H_0^{i,j} = \bigcap_{i=1}^{I-1} \bigcap_{j=i+1}^{I} \{\mu_i = \mu_j\}$$

$H_0^{i,j}$ désigne ainsi un test élémentaire de comparaison entre deux populations. Afin de contrôler le niveau de test global, il faut faire chaque test élémentaire à un niveau plus faible. Une méthode bien connue est la méthode de Bonferroni qui est fondée sur l'inégalité suivante :

$$\alpha^* = P(\text{ Rejet } H_0 | H_0) = P\left(\bigcup_{i=1}^{I-1} \bigcup_{j=i+1}^{I} \text{ Rejet } H_0^{i,j} | H_0\right) \leq \alpha \times I(I-1)/2.$$

Cette inégalité dit simplement que le risque de rejeter $H_0^{i,j}$ pour au moins l'un des couples (i, j) tests est inférieur au risque de rejeter une hypothèse élémentaire $H_0^{i,j}$ lors d'une comparaison, (α), multiplié par le nombre de comparaisons $(I(I-1)/2)$. Ainsi, si

on veut limiter le risque de faire une erreur (de première espèce) à un niveau $\alpha^* = 5\%$, il faut effectuer chacun des tests $H_0^{i,j}$ au niveau

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

2.3.4 Programme R

```
###Lecture des données
Data=read.table("Arbres.don",header=FALSE,sep="")
colnames(Data)=c("Diametre","Statut")

###Graphiques et statistiques élémentaires.}
table(Data$Statut) #donne les effectifs par groupe de statut.
by(Data$Diametre,Data$Statut,mean) #donne les diamètres moyens
                                    par statut.
boxplot(Data$Diametre~Data$Statut) #affiche les boîtes à moustaches
                                    par groupe.

###Analyse de la variance.
arbres.lm=lm(Diametre~Statut,data=Data)
summary(arbres.lm)
anova(arbres.lm) #renvoie la table d'analyse de la variance
                 de ce modèle.
###Graphe des diagnostics de R.
par(mfrow=c(2,2))
plot(arbres.lm)
```

2.3.5 Programme SAS

```
###Lecture et affichage des données.}
data ARBRES; #permet de définir le tableau SAS
     infile 'Arbre.don' firstobs=2;
     input Diametre Statut$;
proc Print data=ARBRES; run;
    ARBRES} à partir des données lues dans le fichier {\tt
    Arbre.don'}.

### Graphiques et statistiques élémentaires
proc BoxPlot data=ARBRES; #affiche les boîtes à moustaches
                          par groupe(statut).
     plot Diametre * Statut;
proc Means data=ARBRES;  #effectif, moyenne, l'écart type, ...
     var Diametre;
proc Sort data=ARBRES; #trie par statut,necessaire pour proc Means
     by Statut;
proc Means data=ARBRES;#donne les statistique élémentaires par
  groupe de statut.
```

```
     var Diametre;
     by Statut;
run;

###Analyse de la variance.}
proc GLM data=ARBRES;
     class Statut;
     model Diametre = Statut / solution;
     means Statut / bon;
     output out=ANOVA p=Predite r=Residu;
run;
# class est l'instruction permettant de spécifier la nature
# qualitative de la variable explicative (si cette instruction est
# absente, la variable sera considérée comme quantitative) et il
# ne s'agira plus d'une analyse de la variance.
# means effectue la comparaison des moyennes des différents
# statuts. L'option bon permet de choisir la méthode de Bonferroni
# pour les tests   multiples.
# output permet de récupérer dans le tableau ANOVA
# les prédictions  et les résidus

### Analyse des résidus.
proc gplot data=ANOVA;
     plot Predite * Residu / vref=0;
run;
proc univariate data=ANOVA normal plot;
var Residu;
run;
```

2.4 Analyse de la variance à deux facteurs : cas équilibré

2.4.1 Objectif et dispositif expérimental

On souhaite étudier l'effet du niveau de fertilisation et de la rotation de culture sur le poids des grains de colza. On compare pour cela $I = 2$ niveaux de fertilisation notés 1 pour faible et 2 pour fort et $J = 3$ types de rotation maïs / blé / colza / blé :

A : sans enfouissement de paille (j=1),

B : avec enfouissement de paille (j=2),

C : avec 4 années de prairie temporaire entre chaque succession sans enfouissement de paille (j=3).

Dans cette partie, on désignera par *traitement* la combinaison Fertilisation*Rotation. Le traitement "1B" désigne la combinaison de la fertilisation "1" et de la rotation "B". On souhaite donc ici comparer $IJ = 2 \times 3 = 6$ traitements.

Chacune des 6 combinaisons Rotation*Fertilisation a été appliquée sur $K = 10$ parcelles, soit un total de $n = IJK = 60$ parcelles. Un tel dispositif est appelé "plan factoriel *complet*" car il permet de croiser tous les niveaux des deux facteurs. De plus

chaque traitement a été reçu par le même nombre de parcelles c'est donc un plan *équilibré*, avec

$$\forall (i,j), \quad n_{ij} \equiv K.$$

2.4.2 Analyse descriptive

La table 2.17 donne les moyennes du poids des grains en fonction de la rotation et de la fertilisation et la figure 2.9 représente leur distribution sous forme de boîtes à moustaches.

		rotation			
		$j=1$	$j=2$	$j=3$	total
fertilisation	$i=1$	$y_{11\bullet} = 24,11$	$y_{12\bullet} = 24,00$	$y_{13\bullet} = 28,64$	$y_{1\bullet\bullet} = 25,58$
	$i=2$	$y_{21\bullet} = 15,81$	$y_{22\bullet} = 19,84$	$y_{23\bullet} = 31,75$	$y_{2\bullet\bullet} = 22,47$
	total	$y_{\bullet 1\bullet} = 19,96$	$y_{\bullet 2\bullet} = 21,92$	$y_{\bullet 3\bullet} = 30,20$	$y_{\bullet\bullet\bullet} = 24,03$

TABLE 2.17 – Tableau des moyennes des poids de grains de colza en fonction de la fertilisation et de la rotation.

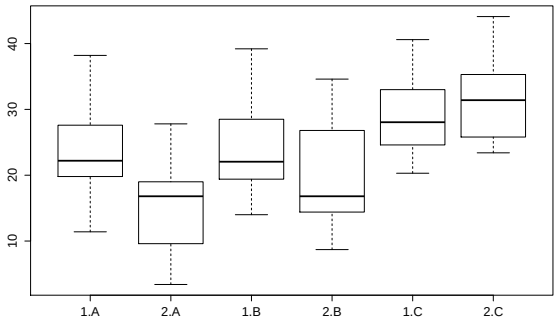


FIGURE 2.9 – Répartition du poids des grains en fonction de la fertilisation et de la rotation.

On observe des différences de moyennes assez fortes (de 15.8 pour le traitement 2A à 31.75 pour le traitement 2C). La figure 2.9 montre que la variabilité est sensiblement la même dans les différents groupes. Cette observation conforte l'hypothèse d'homoscédasticité qui est nécessaire pour les différents modèles présentés dans cette partie.

Graphe d'interaction

La combinaison de deux facteurs n'est pas forcément additive : l'effet spécifique du traitement (ij), peut-être autre chose que la simple somme de l'effet de la fertilisation i et de la rotation j. La différence entre l'effet du traitement (ij) et la somme des effets de chacun des facteurs est appelé terme d'interaction. En l'absence de ce terme, l'écart entre les deux niveaux de fertilisation doit être le même, quelle que soit la rotation :

$$\mu_{1j} - \mu_{2j} = \text{constante}.$$

La figure 2.10 présente "le graphe d'interaction" pour les poids des grains de colza, c'est-à-dire les moyennes observées pour chacun des traitements. L'écart entre les deux types de fertilisation pour les différents systèmes de rotation n'étant pas constant, il faudra inclure un terme d'interaction dans le modèle. Le graphe d'interaction ne permet cependant pas de se prononcer quant à la significativité de cet effet.

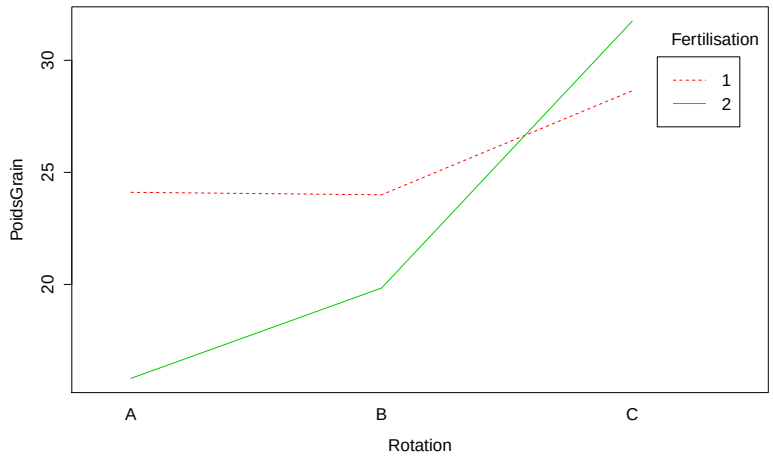


FIGURE 2.10 – Représentation de l'interaction observée entre les deux facteurs.

2.4.3 Analyse de l'effet de la rotation et de la fertilisation

Modélisation

Pour décrire l'effet conjoint des deux facteurs (fertilisation et rotation) sur le poids des grains, il est possible d'utiliser la version régulière du modèle d'analyse de la variance à 2 facteurs :

$$Y_{ijk} \overset{ind}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2),$$

ou de manière équivalente

$$Y_{ijk} = \mu_{ij} + E_{ijk}, E_{ijk} \quad \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \tag{M.2.5}$$

i indique le niveau de fertilisation et varie de 1 à 2, j indique le type de rotation utilisée et varie de 1 à 3, et k est le numéro de la culture au sein du niveau i et du niveau j, il varie de 1 à 10.

L'écriture sous forme singulière ne permet pas de dissocier les effets spécifiques de chacun des facteurs.

Pour faire apparaître les différents effets, on écrit classiquement le modèle d'analyse de la variance sous la forme suivante, classique dans tous les logiciels :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk} \qquad E_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, 2,\ j = 1, \ldots, 3, k = 1, \ldots 10, \tag{M.2.6}$$

où

- α_i représente l'*effet principal* de la fertilisation,
- β_j représente l'*effet principal* de la rotation,
- γ_{ij} est le terme d'*interaction*,
- σ^2 est la variance résiduelle.

Le modèle M.2.6 est plus explicite mais, comme le modèle d'analyse de la variance à un facteur M.2.3, p. 63, il n'est pas identifiable, c'est à dire que l'estimation de ses paramètres nécessite le recours à un système de contraintes.

Sous forme matricielle, le modèle s'écrit

$$Y = X\theta + E, \qquad E \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

avec $\theta = (\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_{11}, \ldots, \gamma_{IJ})'$ et

$$X = \begin{array}{c} \begin{array}{cccccccccccc} \mathbf{1} & A^1 & A^2 & B^1 & B^2 & B^3 & \Gamma^{11} & \Gamma^{12} & \Gamma^{13} & \Gamma^{21} & \Gamma^{22} & \Gamma^{23} \end{array} \\ \left(\begin{array}{cccccccccccc} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & 0 & \vdots & 1 & 0 & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 0 & 1 & \vdots & 0 & 1 & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 & 0 & \vdots & 1 & 0 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & 1 & \vdots & 0 & 1 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & 1 & 0 & 0 & \vdots & 1 & \vdots & \vdots & 1 & 0 & \vdots & \vdots \\ \vdots & 0 & 1 & 1 & \vdots & 0 & \vdots & \vdots & 0 & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & 0 & \vdots & \vdots & \vdots & \vdots & 1 & 0 & \vdots \\ \vdots & \vdots & \vdots & 0 & 1 & \vdots & \vdots & \vdots & \vdots & 0 & 1 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 & 0 & \vdots & \vdots & \vdots & \vdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & 1 & \vdots & \vdots & \vdots & \vdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

Le problème d'identifiabilité se traduit ici par le fait que la matrice X, de dimension $(n, (1+I)(1+J)) = (60, 12)$ n'est pas de rang plein, son rang vaut $IJ = 6$. Pour rendre le modèle identifiable, il faudra poser $12 - 6$ contraintes linéaires indépendantes.

Analyse des résidus

Avant toute chose, il faut vérifier que les hypothèses du modèle sont valides. La figure 2.11 présente les graphiques de diagnostic proposés par défaut par le logiciel `R`.

Les hypothèses portant sur la distribution des erreurs n'ont aucune raison d'être remises en cause ici. En effet le graphique en bas à gauche, qui représente en abscisse

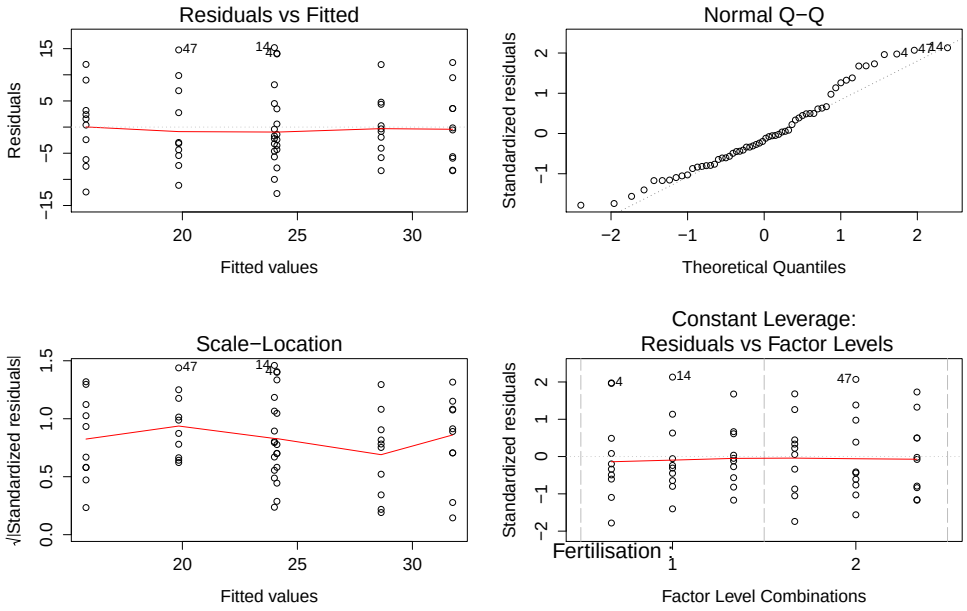


FIGURE 2.11 – Graphiques de diagnostic des résidus pour le modèle M.2.6 d'analyse de la variance à deux facteurs avec interaction pour le poids des grains de colza.

les valeurs prédites et en ordonnées la racine carrée des résidus standardisés permet de vérifier que la variance de ces résidus est similaire dans tous les groupes.

Le graphique quantiles contre quantiles, permet de vérifier que l'hypothèse de normalité des résidus est raisonnable.

Tandis que le quatrième graphique, en bas à droite, qui permet d'identifier des individus atypiques, ne met en évidence aucun individu inquiétant (i.e. dont la distance de Cook serait plus grande que 1 par exemple).

Remarque. D'ordinaire le quatrième graphique, en bas à droite, présente l'effet levier en abcsisse et les résidus standardisés en ordonnée. Mais l'effet levier associé à un individu dans une analyse de la variance est donné par l'inverse de l'effectif de son groupe (terme h_{ii} de la matrice de projection P comme présenté p.42). Puisque dans cet exemple tous les groupes ont les mêmes effectifs, les effets levier sont tous égaux. C'est ce qui est indiqué dans le titre de ce graphique et par conséquent ce sont les traitements qui sont représentés en abscisse.

Test du modèle complet

Le test du modèle complet présenté dans la table 2.18 constitue le premier outil pour évaluer l'effet des deux facteurs et de leur interaction sur le poids des grains.

Le rang de la matrice X est $r = 6$. D'après la proposition 1.4.4, p. 32, sous l'hypothèse $H_0 = \{Y_{ijk} = \mu + E_{ijk}\}$,

$$\frac{SCM/(IJ-1)}{SCR/(n-IJ)} \sim \mathcal{F}_{5,60-6}.$$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	59	4712.29				
2	54	3052.47	5	1659.82	5.87	0.0002

TABLE 2.18 – Table d'analyse de la variance du modèle à deux facteurs avec interaction.

On retrouve les 54 degrés de liberté associés à la somme des carrés résiduelle et les 5 degrés de liberté associés à SCM.

La statistique de Fisher vaut 5.87 (la variabilité expliquée par le traitement est presque 6 fois supérieure à la variabilité résiduelle), ce qui correspond à une probabilité critique de $2\ 10^{-4}$. La fertilisation, la rotation ou leur interaction ont un effet significatif sur le poids des grains.

On peut remarquer que l'ajustement du modèle n'est pas très satisfaisant. En effet on a $R^2 = 1 - 3052.47/4712.29 = 0.352$. Ceci n'est pas contradictoire avec l'effet très significatif du traitement mais signifie seulement que la variabilité résiduelle reste forte : on ne peut pas espérer prédire précisément le poids des grains en se fondant seulement sur le traitement.

Tests des différents effets

Le dispositif expérimental est équilibré, c'est à dire que n_{ij} est constant et égal à $K = 10$.

Or un dispositif équilibré est un dispositif *orthogonal*, il vérifie en effet la condition énoncée dans la proposition 1.5.1, p. 39 :

$$\forall i, j : \quad n_{ij} = \frac{n_{i+} n_{+j}}{n}.$$

Comme mentionné dans la partie 1.5.1, p. 38, la propriété d'orthogonalité assure une décomposition unique des effets. On pourra donc ici dissocier l'effet de la fertilisation et l'effet de la rotation.

		Sum Sq	Df	F value	Pr(>F)
Type I	Fertilisation	145.70	1	2.58	0.1142
	Rotation	1180.48	2	10.44	0.0001
	Fertilisation :Rotation	333.63	2	2.95	0.0608
	Residuals	3052.47	54		
		Sum Sq	Df	F value	Pr(>F)
Type II	Fertilisation	145.70	1	2.58	0.1142
	Rotation	1180.48	2	10.44	0.0001
	Fertilisation :Rotation	333.63	2	2.95	0.0608
	Residuals	3052.47	54		

TABLE 2.19 – Effets des différents facteurs (type I et type II) dans le modèle d'analyse de la variance à deux facteurs avec interaction.

Le tableau 2.19 donne les valeurs des sommes de carrés ainsi que les tests des effets de chacun de ces facteurs. On retrouve bien ici que grâce à l'orthogonalité les tests de type I et II sont les mêmes et qu'il n'y a aucune confusion d'effets : les valeurs des

réductions ne dépendent que de l'effet considéré et pas du modèle de référence. On a donc :

$$SCA = R(\alpha/\beta, \mu) = R(\alpha/\mu), \quad SCB = R(\beta/\alpha, \mu) = R(\beta/\mu). \tag{2.2}$$

Ainsi, du fait de l'équilibre du plan, la somme des carrés expliqués par l'ajout du facteur fertilisation (resp. rotation) dans un modèle contenant le facteur rotation (resp. fertilisation) et la constante est égale à la somme des carrés expliqués par l'ajout du facteur fertilisation (resp.rotation) dans un modèle contenant uniquement la constante. Il n'y a donc pas d'ambiguité sur la notion d'effet du facteur fertilisation (resp. rotation) et on le note SCA (resp. SCB)

La décomposition des sommes de carrés permet de tester des hypothèses plus précises que l'hypothèse générale testée dans la table d'analyse de la variance du tableau 2.18. Elle permet notamment de tester de façon spécifique l'effet de chacun des facteurs. Les tests présentés correspondent à ceux détaillés dans le chapitre précédent (table 1.1, p. 35 ou de manière équivalente dans la table 1.3, p. 36 puisque le dispositif est orthogonal).

Interprétation des tests des différents effets.

Effet de l'interaction : La probabilité critique associée au test sur l'interaction vaut 0.06. L'effet de l'interaction n'est donc pas significatif au niveau 0.05. et on pourrait accepter un modèle sans interaction. De façon plus nuancée on peut dire que cet effet est faiblement significatif. Une expérience avec plus de répétitions donnerait un test plus puissant qui permettrait peut-être de détecter un effet significatif.

Effet de la fertilisation : La statistique de Fisher F_A vaut 2.58, ce qui n'est pas significatif au niveau 5% (probabilité critique) 11.4%). On peut donc conclure à l'absence d'effet moyen de la fertilisation.
Pourtant, on ne peut pas supprimer l'effet principal du modèle avec interaction. En terme d'interprétation, cela reviendrait à dire que la fertilisation n'a pas d'effet propre sur le poids des grains, mais qu'elle interagit avec la rotation, ce qui n'a pas de sens. Si l'interaction est conservée dans le modèle alors l'effet principal l'est également.

Effet de la rotation : L'effet de la rotation est nettement significatif (probabilité critique $= 10^{-4}$). La comparaison des statistiques de Fisher nous montre qu'il est près de 4 fois plus fort que l'effet de la fertilisation ou de l'interaction. C'est clairement l'effet majoritaire de ce modèle.

Estimation des paramètres

Le modèle n'étant pas identifiable, l'estimation des paramètres est faite en imposant un système de $I + J + 1 = 6$ contraintes linéairement indépendantes. Ce système peut être choisi arbitrairement mais il est nécessaire d'en tenir compte lors des tests sur les paramètres. Deux systèmes de contraintes sont principalement utilisés.

Contraintes naturelles. On impose que la somme pondérée des paramètres associés à chaque effet soit nulle, c'est-à-dire

$$\sum_i n_{i+}\alpha_i = 0, \qquad \sum_j n_{+j}\beta_j = 0, \qquad \forall i : \sum_j n_{ij}\gamma_{ij} = 0 \qquad \forall j : \sum_i n_{ij}\gamma_{ij} = 0.$$

On obtient alors les estimateurs naturels

$$\hat{\mu} = Y_{\bullet\bullet\bullet}, \qquad \hat{\alpha}_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet},$$
$$\hat{\beta}_j = Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet}, \qquad \hat{\gamma}_{ij} = Y_{ij\bullet} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) = Y_{ij\bullet} - Y_{i\bullet\bullet} - Y_{\bullet j\bullet} + Y_{\bullet\bullet\bullet},$$

Les valeurs des estimations se déduisent simplement du tableau 2.17, p. 71 :

		rotation			
		$j=1$	$j=2$	$j=3$	total
fertilisation	$i=1$	$\hat{\gamma}_{11} = +2.59$	$\hat{\gamma}_{12} = +0.52$	$\hat{\gamma}_{13} = -3.11$	$\hat{\alpha}_1 = +1.56$
	$i=2$	$\hat{\gamma}_{21} = -2.59$	$\hat{\gamma}_{22} = -0.52$	$\hat{\gamma}_{23} = +3.11$	$\hat{\alpha}_2 = -1.56$
	total	$\hat{\beta}_1 = -4.07$	$\hat{\beta}_2 = -2.11$	$\hat{\beta}_3 = +6.17$	$\hat{\mu} = +24,03$

On peut d'ailleurs ré-exprimer les sommes de carrés SCA, SCB et SCI en fonction des estimations :

$$SCA = JK \sum_i \hat{\alpha}_i^2, \qquad SCB = IK \sum_j \hat{\beta}_j^2, \qquad SCI = R(\gamma|\mu,\alpha,\beta) = K \sum_j \hat{\gamma}_{ij}^2$$

ce qui montre bien que les sommes de carrés donnent une mesure de l'amplitude de chaque effet.

Contraintes R. Elles annulent tous les paramètres associés au premier niveau de chaque facteur, soit

$$\alpha_1 = 0, \qquad \beta_1 = 0, \qquad \forall i : \gamma_{i1} = 0, \qquad \forall j : \gamma_{1j} = 0.$$

Ce système de contraintes est avantageux du point de vue des calculs, il revient à supprimer des colonnes de la matrice X afin qu'elles soient de plein rang. Les matrices X et θ de la version générale du modèle sont remplacées par :

$$\tilde{X} = \begin{array}{c} \begin{array}{cccccc} \mathbf{1} & A^2 & B^2 & B^3 & \Gamma^{22} & \Gamma^{23} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

Ces contraintes ne sont cependant pas forcément avantageuses du point de vue de l'interprétation. On obtient les estimateurs suivants :

$$\hat{\mu} = Y_{11\bullet}, \quad \hat{\alpha}_i = Y_{i1\bullet} - Y_{11\bullet}, \quad \hat{\beta}_j = Y_{1j\bullet} - Y_{11\bullet}, \quad \hat{\gamma}_{ij} = Y_{ij\bullet} - Y_{i1\bullet} - Y_{1j\bullet} + Y_{11\bullet}.$$

L'essentiel est de bien noter que les premiers niveaux de chacun des facteurs interviennent dans tous les estimateurs. Ainsi tester si α_2 est égale à 0 s'interpréte comme le test portant sur la différence entre le fertilisant 2 et le fertilisant 1, utilisé en combinaison avec le premier système de rotation.

Le tableau 2.20 donne les valeurs des estimations correspondantes. Il faut interpréter les tests associés à chacun de ces paramètres en fonction des contraintes choisies par le logiciel utilisé, ici R. Ainsi, l'hypothèse $H_0 = \{\beta_2 = 0\}$ testée sur la ligne `Rotation B` s'exprime littéralement "*Le poids moyen des grains obtenu avec la rotation B est le même que celui obtenu avec la rotation C pour la fertilisation 1*".

Contraintes SAS. Le système de contraintes adopté par `SAS` est similaire à celui adopté par R, mais le niveau de référence est le dernier niveau des facteurs, ce qui correspond aux contraintes

$$\alpha_I = 0, \qquad \beta_J = 0, \qquad \forall i : \gamma_{iJ} = 0, \qquad \forall j : \gamma_{Ij} = 0.$$

Remarque : On présente souvent sept contraintes dans les systèmes de contraintes mais seulement six sont linéairement indépendantes.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.1100	2.3775	10.14	0.0000
Fertilisation2	-8.3000	3.3624	-2.47	0.0168
RotationB	-0.1100	3.3624	-0.03	0.9740
RotationC	4.5300	3.3624	1.35	0.1835
Fertilisation2 :RotationB	4.1400	4.7551	0.87	0.3878
Fertilisation2 :RotationC	11.4100	4.7551	2.40	0.0199

TABLE 2.20 – Estimations de paramètres du modèle d'analyse de la variance à 2 facteurs avec interaction pour le poids des grains de colza avec le logiciel R.

Prédictions. La théorie du modèle linéaire montre que les prédictions $\hat{Y}_{ijk}$ ne dépendent pas du système de contraintes choisies (1.3.1, p. 21). On vérifie facilement que pour les trois systèmes de contraintes présentés ici, on a

$$\hat{Y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = Y_{ij\bullet},$$

ce qui signifie que le poids des grains prédit pour la fertilisation i et la rotation j est simplement le poids moyen observé dans cette combinaison. Il n'y a donc que 6 valeurs prédites possibles (2 fertilisations $\times$ 3 rotations) ; ces valeurs donnent les abscisses des 6 colonnes observées dans la figure 2.11, p. 74.

Comparaison des traitements La comparaison des traitements est un des objectifs classiques de l'analyse de la variance. On vient de voir qu'il peut être difficile de faire cette comparaison en se fondant sur les estimations des paramètres à cause du choix arbitraire de la contrainte.

Effets principaux. Le test sur la fertilisation nous a montré qu'il n'y avait pas de différence significative entre les différents modes de fertilisation. Au contraire on a mis en évidence un effet du système de rotation.

Pour connaître quels sont les systèmes de rotation effectivement différents, on veut comparer deux à deux tous les modes de rotation. Le tableau 2.21, p. 79 présente des comparaisons de moyennes analogues à celle présentées pour la comparaison des statuts des alisiers dans la partie précédente. Pour tout couple j, j', on teste

$$H_0 = \left\{\mu_{\bullet j} = \mu_{\bullet j'}\right\}.$$

Les résultats sont donnés dans la table 2.21, la correction de Bonferroni pour les tests multiples a déjà été appliquée.

```
 contrast estimate       SE df t.ratio p.value
 A - B      -1.960 2.377546 54  -0.824  1.0000
 A - C     -10.235 2.377546 54  -4.305  0.0002
 B - C      -8.275 2.377546 54  -3.480  0.0030

P value adjustment: bonferroni method for 3 tests
```

TABLE 2.21 – Comparaison des moyennes de poids des grains par rotation.

Les rotations A et B sont similaires, alors que les rotations A et B sont différentes de la rotation C. La présence de 4 années de prairie temporaire serait donc la source principale de l'augmentation des poids des grains.

Remarque. On rappelle que ces comparaisons se font dans le cadre d'un modèle avec interaction. Celle-ci est ici faiblement significative, mais si elle était plus forte, il faudrait interpréter ces comparaisons d'effets moyens avec prudence.
Notamment, la "supériorité" de la rotation C peut ne pas être générale : cette rotation peut donner un poids moyen des grains supérieurs en moyenne sur les fertilisations étudiées, et donner de moins bons résultats pour certaines fertilisations spécifiques.

Comparaison des 6 traitements (combinaisons). Le tableau 2.22, p. 80 présente la comparaison des 6 combinaisons sous une forme légèrement différente.

En appliquant la règle de Bonferroni, on conclut que les seuls couples de combinaisons significativement différents sont les couples (`2A`, `1C`), (`2A`, `2C`) et (`2B`, `2C`).

2.4.4 Conclusion

Seul le mode de rotation des cultures semble avoir une influence significative sur le poids des grains de colza. Cette conclusion doit être nuancée par un effet potentiel d'interaction entre la rotation et le mode de fertilisation des cultures. La probabilité

```
$contrasts
 contrast  estimate       SE df t.ratio p.value
 1,A - 2,A     8.30 3.362358 54   2.469  0.2515
 1,A - 1,B     0.11 3.362358 54   0.033  1.0000
 1,A - 2,B     4.27 3.362358 54   1.270  1.0000
 1,A - 1,C    -4.53 3.362358 54  -1.347  1.0000
 1,A - 2,C    -7.64 3.362358 54  -2.272  0.4062
 2,A - 1,B    -8.19 3.362358 54  -2.436  0.2728
 2,A - 2,B    -4.03 3.362358 54  -1.199  1.0000
 2,A - 1,C   -12.83 3.362358 54  -3.816  0.0053
 2,A - 2,C   -15.94 3.362358 54  -4.741  0.0002
 1,B - 2,B     4.16 3.362358 54   1.237  1.0000
 1,B - 1,C    -4.64 3.362358 54  -1.380  1.0000
 1,B - 2,C    -7.75 3.362358 54  -2.305  0.3756
 2,B - 1,C    -8.80 3.362358 54  -2.617  0.1721
 2,B - 2,C   -11.91 3.362358 54  -3.542  0.0124
 1,C - 2,C    -3.11 3.362358 54  -0.925  1.0000

P value adjustment: bonferroni method for 15 tests
```

TABLE 2.22 – Comparaison des moyennes de poids des grains par combinaison Rotation*Fertilisation. Les écarts de moyennes sont donnés dans la colonne estimate, la probabilité critique (corrigée par la méthode de Bonferoni) associée au test d'égalité de ces deux moyennes est fournie dans la dernière colonne.

critique associée au test de cet effet vaut 6% dans cette étude, une autre expérience avec plus de répétitions permettrait peut être d'avoir une probabilité critique plus faible et donc davantage confiance dans l'existence de cet effet d'interaction.

Le système de rotation C est bien meilleur en terme de rendement pour le colza que les systèmes A ou B, d'autant plus quand la fertilisation utilisée est la fertilisation 2.

2.4.5 Programme R

```
colza<-read.table('Colza.txt',header=T)
head(colza) #les 6 premieres lignes

#changer le type de Fertilisation, pour le considerer comme un facteur
colza$Fertilisation<-as.factor(colza$Fertilisation)

summary(colza)

## boites à moustaches
with(colza, boxplot(PdsGrains~ Fertilisation+Rotation ))

##moyennes par  combinaison Rotation x fertilisation
with(colza,
     by(PdsGrains, INDICES=list(Rotation, Fertilisation),mean)
)
dev.copy2pdf(file="ColzaBoxplot.pdf", out.type = "pdf")
```

```
##Pour obtenir le plan d'expérience
with(colza,
     table(Rotation,Fertilisation)
)


#graphe de l'interaction
interaction.plot(colza$Rotation,colza$Fertilisation,colza$PdsGrains,
                 fixed=TRUE,col=2:3,leg.bty="o",
     xlab="Rotation", ylab="PoidsGrain", trace.label="Fertilisation")
dev.copy2pdf(file="Colza-Interaction.pdf", out.type = "pdf")

modA2I<-lm(PdsGrains ~ Fertilisation * Rotation,
        data=colza)

### Validation des hypothèses du modèle
par(mfrow=c(2,2))
plot(modA2I)
dev.copy2pdf(file="Colza-Residus.pdf", out.type = "pdf")
#somme des carres
#definition du modele nul
mod0=lm(PdsGrains ~ 1,data=colza)
anova(mod0,modA2I)

#analyse des effets
library(car) #necessaire pour la fonction Anova
anova(modA2I)
Anova(modA2I)

#estimation et test des paramètres
summary(modA2I)

#Comparaison des moyennes de Rotation
#Attention pairwise t.test n'utilise pas la bonne SCR !!!
## il ne tire pas profit de l'estimation de la variance
## sur l'ensemble des groupes
library(lsmeans) #package  lsmeans

lsmeans(modA2I,pairwise~Rotation,adjust="bonferroni")
lsmeans(modA2I,pairwise~Fertilisation,adjust="bonferroni")

#comparaison des 6 croisements
lsmeans(modA2I,pairwise~Fertilisation+Rotation,adjust="bonferroni")
```

2.4.6 Programme SAS

```
data COLZA;
     infile 'Colza.don' firstobs=2;
     input Fertilisation$ Rotation$ PdsGrains;
proc Print data=COLZA;
symbol1 i=boxJT l=1 c=black bwidth=1 co=black;
symbol2 i=boxJT l=2 c=red bwidth=3;
proc GPlot data=COLZA;
     plot PdsGrains*Rotation = Fertilisation;
run;

proc Means data=COLZA noprint;
     by Fertilisation Rotation;
     output out=INTER mean=PdsGrains std=EcartType;
proc Print data=INTER;
run;

proc GLM data=Colza;
     class Fertilisation Rotation;
     model PdsGrains = Fertilisation Rotation Fertilisation*Rotation /
     solution SS1 SS2;
     means Fertilisation Rotation / bon;
     lsmeans Fertilisation Rotation Fertilisation*Rotation /
     tdiff pdiff;
     output out=ANOVA2 p=Predite r=Residu;
run;


symbol i=none v=plus;
proc GPlot data=ANOVA2;
     plot Residu * Predite / vref=0;
run;
```

2.5 Analyse de la variance à deux facteurs : cas déséquilibré

2.5.1 Objectif et dispositif

Dans le cadre d'un concours, on souhaite évaluer les copies de 70 étudiants. Le classement est important car les 7 premiers se verront attribuer une bourse d'étude, et seuls les 25 premiers seront admis. La note finale attribuée est entre 0 et 100.

Pour minimiser les biais de correction, chaque copie est corrigée par 3 correcteurs différents. Il y a au total 7 correcteurs nommés A, $B, \ldots, G$. Chaque correcteur se voit attribuer 30 copies.

Pour attribuer une note finale à une copie, la méthode la plus simple est d'attribuer la moyenne des 3 notes obtenues mais les notes sont attribuées par des correcteurs différents

et ne sont pas directement comparables. Les copies évaluées par des correcteurs sévères recevront des notes plus basses.

Le but de l'analyse statistique est d'étudier si il y a des différences entre correcteurs et de proposer un classement final à partir des notes corrigées des effets correcteurs.

Données

Le tableau 2.23, p. 83 présente un extrait des résultats du concours.

	copie	correcteur	note
1	1	C	15.50
2	1	D	18.00
3	1	G	7.50
4	2	A	56.00
5	2	D	50.50
6	2	E	47.50
7	3	B	41.00
8	3	D	47.50
9	3	F	37.00
10	4	A	70.00
11	4	D	65.00
12	4	E	61.00
⋮	⋮	⋮	⋮
210	70	F	65.5

TABLE 2.23 – Notes obtenues par 70 copies, notées par 7 correcteurs.

Dispositif

Le dispositif est donc composé de $I = 70$ copies et de $J = 7$ correcteurs. n_{ij}, comme précédemment désigne le nombre de fois où le correcteur j a corrigé la copie i, n_{ij} vaut donc 0 ou 1. Chaque copie a été corrigée par 3 correcteurs différents donc $n_{i+} = 3$, tandis que chaque correcteur a corrigé 30 copies $n_{+j} = 30$. On a au total $n = 210$ observations.

Le tableau 2.24, p. 84 donne un extrait des valeurs de tous ces effectifs.

Plan en blocs incomplets D'un point de vue de planification, ce dispositif est un dispositif en blocs. Chaque correcteur pouvant être sensible à différents aspects d'une copie, il constitue un bloc. Tous les correcteurs ne notant pas toutes les copies, le dispositif est appelé plan en bloc incomplet. Néanmoins chaque correcteur note le même nombre de copies et chaque copie est notée par le même nombre de correcteurs. Ce dispositif, appelé plan en blocs incomplets équilibré est traité en détail dans le paragraphe 8.2, p. 262 du chapitre 8, p. 259.

Le dispositif n'est pas orthogonal puisqu'il ne vérifie pas la condition donnée dans la proposition 1.5.1, p. 39. En effet, $n_{i+} = 3$ et $n_{+j} = 30$, alors que n_{ij} vaut 1 ou 0 selon que la copie i a été notée par le correcteur j ou non. Le dispositif n'est donc pas orthogonal, ce qui signifie qu'il n'existe pas de décomposition unique des sommes de

Obs	A	B	C	D	E	F	G	n_{i+}
1	0	0	1	1	0	0	1	3
2	1	0	0	1	1	0	0	3
3	0	1	0	1	0	1	0	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
67	0	1	0	0	1	0	1	3
68	0	0	1	0	1	1	0	3
69	1	1	1	0	0	0	0	3
70	0	0	1	0	1	1	0	3
n_{+j}	30	30	30	30	30	30	30	210

TABLE 2.24 – Dispositif expérimental pour la répartition des notes : 70 copies réparties entre 7 correcteurs.

carrés associées à chaque effet et qu'*on ne pourra donc jamais complètement séparer les effets des deux facteurs*. Du fait du dispositif, les deux effets sont en partie confondus.

2.5.2 Analyse descriptive

Les graphiques présentés dans la figure 2.12, montrent les distributions des notes en fonction des correcteurs et des copies.

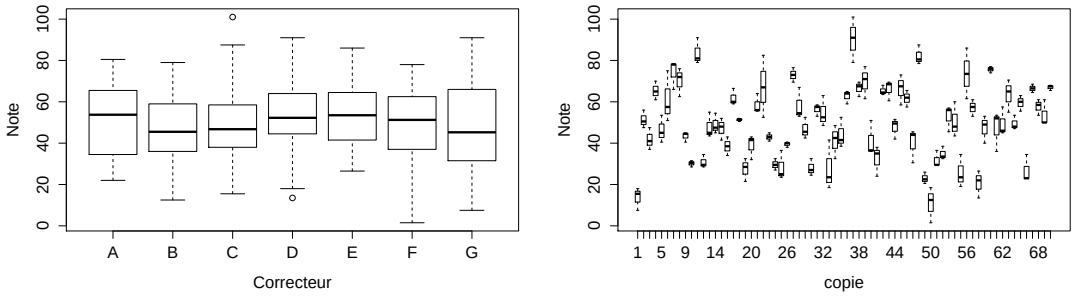


FIGURE 2.12 – Distribution des notes attribuées par les différents correcteurs à gauche et des notes obtenues par les différentes copies à droite.

Il semble qu'il n'y ait pas de différence notable entre les correcteurs et au contraire les notes obtenues par les différentes copies sont très variables, ce qui est attendu. Ici on pourrait être tenté de tester l'existence d'un effet correcteur dans un premier temps, ce qui repose sur le modèle

$$Y_{ij} = \mu + \beta_j + E_{ij}, \qquad E_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{M.2.7}$$

en notant Y_{ij} la note obtenue par la i-ème copie corrigé par le correcteur j.

La table d'analyse de la variance d'un tel modèle donne le résultat présenté dans la table 2.25, p. 85, partie b.

a)

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	52.5000	3.3026	15.90	0.0000
correcteurB	-5.3833	4.6706	-1.15	0.2504
correcteurC	-3.5500	4.6706	-0.76	0.4481
correcteurD	-0.1500	4.6706	-0.03	0.9744
correcteurE	1.1500	4.6706	0.25	0.8058
correcteurF	-4.0833	4.6706	-0.87	0.3830
correcteurG	-4.7333	4.6706	-1.01	0.3121

b)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
correcteur	6	1257.76	209.63	0.64	0.6976
Residuals	203	66425.07	327.22		

TABLE 2.25 – L'estimation des effets est présentée dans la table a) et la table b) présente les résultats de l'analyse de l'effet du facteur correcteur dans une table d'analyse de la variance pour le modèle M_{cor}.

L'effet correcteur n'est pas significatif dans ce cadre mais l'écart type estimé de l'erreur est grand ($\hat{\sigma} = \sqrt{327} = 18.08$) puisqu'il contient également la variabilité associée aux différences de notes entre les copies. Il est possible que l'effet correcteur existe mais qu'il soit masqué par une variabilité résiduelle trop forte, ce qui produit un test peu puissant.

2.5.3 Analyse de la variance à 2 facteurs

Le modèle prend en compte un effet copie et un effet correcteur. Le facteur A est le facteur copie, il a $I = 70$ niveaux, et le facteur B est le facteur correcteur, il a $J = 7$ niveaux.

Interaction Copie*Correcteur. Tous les correcteurs ne sont pas sensibles aux mêmes aspects d'un travail, il y a donc toutes les raisons de penser qu'il peut exister une interaction entre l'effet copie et l'effet correcteur. Néanmoins, en l'absence de répétitions, on ne pourra pas estimer une telle interaction entre ces deux effets puisque dans ce cas le terme d'interaction serait confondu avec le terme d'erreur E_{ij}. Un ajustement par les moindres carrés donnerait un ajustement parfait et des résidus nuls, donc une estimation de la variance résiduelle nulle. Cela ne signifie absolument pas que cette interaction n'existe pas, cela veut seulement dire que nous n'avons pas assez de données pour l'estimer. Cette interaction n'apparaîtra donc pas dans le modèle d'analyse et sera rejetée dans la résiduelle. Si cette interaction existe effectivement, les tests des effets des facteurs seront moins puissants.

Analyse de la variance à deux facteurs sans interaction Le modèle d'analyse est donc un modèle d'analyse de la variance à deux facteurs sans interaction, il s'écrit :

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}, \qquad E_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \tag{M.2.8}$$

où i varie de 1 à 70 et désigne les I copies, tandis que j varie de 1 à 7 et désigne les J niveaux du facteur correcteur. Seuls les couples (i, j) tels que $n_{ij} = 1$ sont présents dans les données. Dans cette version le modèle compte $1+I+J$ paramètres pour décrire l'espérance et un paramètre pour la variance de l'erreur.

Écriture matricielle Le modèle (M.2.8) peut s'écrire sous la forme matricielle

$$E = X\theta + E$$

avec

$$\underset{(n,1)}{Y} = \begin{bmatrix} Y_{1C} \\ Y_{1D} \\ Y_{1G} \\ \vdots \\ Y_{70C} \\ Y_{70E} \\ Y_{70F} \end{bmatrix}, \quad \underset{(1+I+J,1)}{\theta} = \left[\begin{array}{c} \mu \\ \hline \alpha_1 \\ \vdots \\ \alpha_I \\ \hline \beta_1 \\ \vdots \\ \beta_J \end{array}\right], \quad \underset{(n,1)}{E} = \begin{bmatrix} E_{1C} \\ E_{1D} \\ E_{1G} \\ \vdots \\ E_{70C} \\ E_{70E} \\ E_{70F} \end{bmatrix},$$

et [2]

$$\underset{(n,1+I+J)}{X} = \left[\begin{array}{c|ccccccc|ccccccc} 1 & 1 & . & . & . & . & . & . & . & . & 1 & . & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & . & 1 & . & . & . \\ 1 & 1 & . & . & . & . & . & . & . & . & . & . & . & . & 1 \\ \vdots & & & & & & & & & & & & & & \\ 1 & . & . & . & . & . & . & 1 & . & . & 1 & . & . & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & 1 & . & . \\ 1 & . & . & . & . & . & . & 1 & . & . & . & . & . & 1 & . \end{array}\right]$$

Estimation des paramètres Comme dans les modèles d'analyse de la variance à un et deux facteurs vus aux chapitres 2.3, p. 62 et 2.4, p. 70, on doit avoir recours à des contraintes pour obtenir des estimateurs des paramètres puisque la matrice X n'est que de rang $I + J - 1$ (alors qu'elle a $I + J + 1$ colonnes).

D'autre part, le dispositif étudié n'étant pas orthogonal les estimateurs de paramètres n'ont pas forcément une forme explicite, ce qui rend leur interprétation difficile. Nous ne nous attardons donc pas ici sur ce point. On peut néanmoins noter que la note qui serait attribuée par le correcteur A à la copie numéro 1, doit valoir $\mu + \alpha_1 + \beta_1$. Si les contraintes utilisées sont celles par défaut de `R`, alors $\alpha_1 = 0$, $\beta_1 = 0$ et μ serait donc la note obtenue par cette copie. Le correcteur B évaluant cette même copie lui aurait attribué la note $\mu + \alpha_1 + \beta_2$, soit $\mu + \beta_2$ dans le même système de contraintes. Le coefficient β_2 est donc l'écart de notation entre le correcteur B et le correcteur A pour la copie 1 si ils avaient tous les deux évalué cette copie. Plus généralement, l'estimation des paramètres β_j va permettre de comparer l'écart entre la note attribuée par le correcteur du niveau j et celle attribuée par le correcteur A pour la copie 1. Cette interprétation des paramètres n'est valable que si la contrainte utilisée est $\alpha_1 = 0$, $\beta_1 = 0$. Le tableau des

2. Les 0 sont remplacés par des points dans la matrice X pour en faciliter la lecture.

estimations des paramètres est fourni par la table 2.26. Le même type de raisonnement permet d'interpréter α_i comme la différence de notes entre la copie i et la copie 1, si elles avaient été notées par le correcteur A. Ainsi si les copies 1 et 2 avaient été évaluées par le correcteur A, la copie 2 aurait obtenu 33.5 points de plus.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.4381	3.2397	6.00	1.7361e-08
copie2	33.5048	4.2932	7.80	1.5183e-12
copie3	29.6786	4.2932	6.91	1.7549e-10
⋮	⋮	⋮	⋮	⋮
copie68	44.1310	4.2932	10.28	1.2423e-18
copie69	40.3952	4.2932	9.41	1.8786e-16
copie70	53.4643	4.2932	12.45	3.9593e-24
correcteurB	-10.2643	1.5179	-6.76	3.8250e-10
correcteurC	-8.2357	1.5179	-5.43	2.6147e-07
correcteurD	-3.4714	1.5179	-2.29	2.3762e-02
correcteurE	-1.3571	1.5179	-0.89	3.7287e-01
correcteurF	-8.1143	1.5179	-5.35	3.7684e-07
correcteurG	-5.6071	1.5179	-3.69	3.2041e-04

TABLE 2.26 – Estimation des paramètres dans le modèle (AB)

2.5.4 Tests des effets et comparaison des copies

Notations

Les différents modèles que nous aurons à considérer dans cette partie sont présentés dans le tableau 2.27. La notation (BA) (et non (AB)) pour le modèle à deux facteurs est volontaire : nous verrons dans la suite que l'ordre des facteurs dans le modèle à une importance.

Nom	Modèle	Interprétation
(0) :	$Y_{ij} = \mu + E_{ij}$	aucun effet
(A) :	$Y_{ij} = \mu + \alpha_i + E_{ij}$	effet copie seul
(B) :	$Y_{ij} = \mu + \beta_j + E_{ij}$	effet correcteur seul
(AB) :	$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$	effets copie puis correcteur
(BA) :	$Y_{ij} = \mu + \beta_j + \alpha_i + E_{ij}$	effets correcteur puis copie

TABLE 2.27 – Différents modèles d'analyse de la variance pour comparer puis redresser les notes.

Les sommes de carrés obtenues dans chacun de ces modèles seront repérées par le nom du modèle. On notera, par exemple, $SCR(A)$ la sommes des carrés résiduels du modèle (A) ou $SCA(BA)$ la somme des carrés associée au facteur copie (noté A) dans le modèle (BA).

Analyse des résidus

On ajuste le modèle et avant de poursuivre une quelconque analyse, il est important de valider les hypothèses du modèle grâce aux outils de diagnostic graphiques présentés dans la figure 2.13.

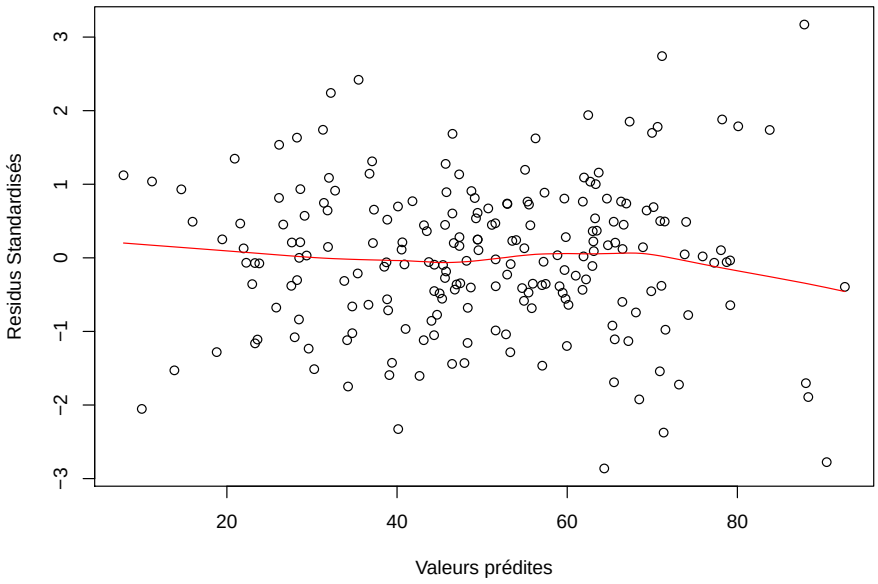


FIGURE 2.13 – Analyse des résidus du modèle (AB).

On peut remarquer que l'hypothèse d'homogénéité de la variance, que l'on peut examiner grâce au graphique présenté dans la figure 2.13 semble raisonnable et on peut poursuivre l'analyse.

Décomposition des sommes de carrés

Modèle complet Le test du modèle complet consiste à comparer le modèle (0) au modèle (AB). Il est présenté dans le tableau 2.28. Il est dit que l'écart-type estimé vaut 5.185. La probabilité critique associée à ce test est inférieure à $2.2\,10^{-16}$, on peut donc rejeter H_0 et affirmer qu'il existe une différence entre les copies ou entre les correcteurs. Ce test n'a que peu d'intérêt dans notre contexte.

```
Residual standard error: 5.185 on 134 degrees of freedom
Multiple R-squared: 0.9468, Adjusted R-squared: 0.917
F-statistic: 31.79 on 75 and 134 DF, p-value: < 2.2e-16
```

TABLE 2.28 – Extrait d'une sortie R. Test de $H_0 = \{\forall i, \alpha_i = 0; \forall j, \beta_j = 0\}$ contre $H_1 = \{\exists i, \alpha_i \neq 0\ \exists j, \beta_j\}$ dans le modèle (AB).

Test des sous modèles Il est plus intéressant de comprendre comment la variabilité totale du jeu de données se répartit entre l'effet copie et l'effet correcteur. On pourra ainsi revenir sur la comparaison des correcteurs.

On note $SCM(AB)$ la somme des carrés associés au modèle (AB). En suivant la notion de réduction déclinée pour l'analyse de la variance dans la section 1.4.3, p. 34, $SCM(AB) = R(\alpha, \beta|\mu)$.

On peut d'abord s'intéresser aux tests de type I qui proposent de décomposer cette réduction de la manière suivante : $R(\alpha, \beta|\mu) = R(\alpha|\mu) + R(\beta|\mu, \alpha)$, c'est à dire dans l'ordre d'introduction des facteurs dans le modèle. Les tests associés à cette décomposition sont les tests de Type I présentés dans la table 2.29. Le test associé à l'effet correcteur porte sur la réduction $R(\beta|\alpha, \mu)$ et teste donc l'existence d'un effet correcteur compte tenu des différences entre copies. Le test sur l'effet copie porte sur la réduction $R(\alpha|\mu)$, il teste donc l'existence d'un effet copie, sans s'ajuster aux différences potentielles entre correcteurs. On remarque ici à nouveau l'asymétrie entre les deux facteurs. Pour tester l'existence d'un effet copie compte tenu des différences entre correcteurs, il faut soit refaire des tests de type I mais sur le modèle (BA) (cf tableau 2.30), soit tester l'effet en type "II".

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
copie	69	62021.67	898.86	33.44	< 2.2e-16
correcteur	6	2059.32	343.22	12.77	2.274e-11
Residuals	134	3601.85	26.88		

TABLE 2.29 – Tests de Type I pour le modèle (AB)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
correcteur	6	1257.76	209.63	7.80	3.294e-07
copie	69	62823.23	910.48	33.87	<2.2e-16
Residuals	134	3601.85	26.88		

TABLE 2.30 – Tests de Type I pour le modèle (BA)

Dans les tests de Type II, les rôles des deux facteurs sont symétrisés. Ainsi l'effet du facteur copie est testé compte tenu d'un potentiel effet correcteur, tandis que l'effet correcteur est testé compte tenu d'un éventuel effet copie. Ces tests sont présentés dans la table 2.31.

	Sum Sq	Df	F value	Pr(>F)
copie	62823.23	69	33.87	<2.2e-16
correcteur	2059.32	6	12.77	2.274e-11
Residuals	3601.85	134		

TABLE 2.31 – Tests de Type II pour le modèle (AB) (ou (BA))

Puisque le plan n'est pas équilibré les tests de type I et II ne donnent pas les mêmes résultats, même si dans cet exemple les deux approches permettent de conclure à l'existence d'un effet correcteur et copie.

Comparaisons des copies

Notre but premier consistait à proposer un classement équitable des copies. Un effet correcteur a été mis en évidence, il semble donc important de le prendre en compte. La première manière d'attribuer une note pour chaque copie consiste à faire la moyenne des notes attribuées par chacun des trois correcteurs l'ayant évaluée. Dans ce cas, la copie i se verra attribuée la note

$$\mu_{i.} = \frac{1}{n_{i+}} \sum_{j=1}^{J} n_{ij}(\mu + \alpha_i + \beta_j) = \mu + \alpha_i + \frac{1}{n_{i+}} \sum_{j=1}^{J} n_{ij}\beta_j$$

On peut remarquer grâce à l'estimation des paramètres présentés dans le tableau 2.26, p. 87 que les correcteurs D et E ne sont pas beaucoup plus sévères que le correcteur A, au contraire B, C et F le sont nettement plus. Si la copie i a été corrigée par A, D et E (comme la copie 2), elle aura une meilleure note que si elle avait été corrigée par les correcteurs C, F et E (copie12). Les deux notes ne sont pas comparables. On va donc définir la note qu'aurait eu chacune de ces copies si elles avaient été notées par tous les correcteurs

$$\tilde{\mu}_{i.} = \mu + \alpha_i + \frac{1}{J} \sum_{j=1}^{J} \beta_j. \tag{2.3}$$

On appelle cette quantité la (note) moyenne ajustée à l'effet correcteur. Cette moyenne ajustée est une combinaison linéaire estimable comme définie en 1.3.1, p. 22. On peut sélectionner les candidats admis soit sur la base de leur note moyenne (système M) soit sur la base de leur note moyenne ajustée (système MA).

Le tableau 2.32, présente les notes de toutes les copies admises selon les systèmes (M) ou (MA). On constate ainsi que la copie 8 est sélectionnée pour avoir une bourse dans le système (MA), mais ne le serait pas dans le système (M) qui ne s'ajustent pas aux différences entre correcteurs. De manière similaire, les copies 46 et 21 échouent au concours dans le système (MA) mais seraient sélectionnées dans les système (M) des moyennes brutes.

Comparaison des correcteurs Dans l'optique d'un nouveau concours, on peut souhaiter garder les correcteurs qui notent de manière similaire, il faut pour cela pouvoir comparer tous les correcteurs les uns aux autres. Si on compare les notes moyennes attribuées par chacun des correcteurs, on omet le fait que les correcteurs ont corrigé des copies différentes. Pour réellement comparer les correcteurs, on va définir les notes moyennes attribuées par chaque correcteur ajustées aux différentes copies. De manière symétrique à la note moyenne ajustée aux correcteurs, on peut définir la moyenne ajustée aux effets copies :

$$\tilde{\mu}_{.j} = \mu + \beta_j + \frac{1}{I} \sum_{i=1}^{I} \alpha_i. \tag{2.4}$$

Ces moyennes ajustées doivent permettre de comparer deux correcteurs, pour cela on veut tester l'hypothèse

$$H_0 = \left\{\tilde{\mu}_{.j} - \tilde{\mu}_{.j'} = 0\right\} \quad \text{contre} \quad H_1 = \left\{\tilde{\mu}_{.j} - \tilde{\mu}_{.j'} \neq 0\right\}$$

	Notes moyennes - (M)			Notes ajustees - (MA)		
copie	note	rang	resultat	note	rang	resultat
37	90.33	1	Bourse	90.81	1	Bourse
11	83.67	2	Bourse	84.12	2	Bourse
48	82.17	3	Bourse	83.04	3	Bourse
56	73.67	6	Bourse	74.28	4	Bourse
27	73.00	7	Bourse	73.87	5	Bourse
7	74.17	5	Bourse	73.45	6	Bourse
8	70.17	8	Admis	72.16	7	Bourse
60	75.67	4	Bourse	71.98	8	Admis
39	69.83	9	Admis	70.28	9	Admis
22	67.33	10	Admis	69.32	10	Admis
45	66.33	14	Admis	68.32	11	Admis
70	67.00	11	Admis	67.61	12	Admis
67	66.50	12	Admis	66.95	13	Admis
43	66.17	15	Admis	66.65	14	Admis
42	65.33	16	Admis	65.94	15	Admis
38	66.50	12	Admis	65.78	16	Admis
36	62.67	19	Admis	64.66	17	Admis
17	61.83	20	Admis	63.82	18	Admis
63	63.50	18	Admis	62.78	19	Admis
6	61.17	22	Admis	62.04	20	Admis
4	65.33	16	Admis	61.65	21	Admis
65	59.50	23	Admis	59.98	22	Admis
28	58.17	25	Admis	58.78	23	Admis
31	56.33	27	–	58.32	24	Admis
68	57.67	26	–	58.28	25	Admis
46	61.67	21	Admis	57.98	26	–
21	58.50	24	Admis	57.78	27	–

TABLE 2.32 – Liste des candidats admis, selon l'un ou l'autre système de notations. La première colonne indique le numéro de copie, les 3 suivantes donnent la note, le classement et la décision obtenue en se basant sur les moyennes brutes, les 3 suivantes donnent les mêmes informations en se basant sur les moyennes ajustées .

Puisque le modèle ne comprend pas de termes d'interaction, tester l'égalité des moyennes ajustées $\tilde{\mu}_{.j}$ et $\tilde{\mu}_{.j'}$ revient à tester $\beta_j = \beta_{j'}$, soit encore à tester une combinaison linéaire de paramètres $C'\theta$, avec C' défini par :

$$C' = \left(\overbrace{0}^{\mu}, \overbrace{0, \ldots, 0}^{\alpha}, \overbrace{0, \ldots, \underset{j}{1}, 0, \ldots, 0, \underset{j'}{-1}, 0, \ldots, 0}^{\beta} \right)$$

Puisque $\sum_k C_k = 0$, cette combinaison linéaire forme un contraste (cf partie sur les combinaisons linéaires 1.3.1, p. 22). Le test portant sur cette combinaison linéaire de paramètres est indépendant du système de contraintes choisi.

Comme présenté dans la proposition 1.4.2, p. 29, sous H_0,

$$\frac{C\widehat{\theta} - a}{\sqrt{C\widehat{\mathbb{V}}(\hat{\theta})C'}} \sim \mathcal{T}(134). \tag{2.5}$$

Lorsque l'on compare tous les correcteurs 2 à 2, on se retrouve dans la situation des tests multiples. Pour contrôler l'erreur de première espèce de l'ensemble des tests, on peut utiliser la correction de Bonferroni et faire chaque test entre deux correcteurs au niveau $2\alpha/(J * (J-1))$, ce qui assure que le niveau global du test est au plus égal à α. L'ensemble des comparaisons deux à deux sont présentées dans le tableau 2.33. Si l'on souhaite faire un test au niveau global de 5%, chaque test doit être fait au niveau $0.05/21 = 0.00238$ A la lecture du tableau, on peut donc conclure que le juge A note différemment des juges B, C, F et G ; le juge B note différemment des juges D, E et G ; etc ...

	contrast	estimate	SE	df	t.ratio	p.value
1	A - B	10	1.5	134	6.8	3.8e-10
2	A - C	8.2	1.5	134	5.4	2.6e-07
3	A - D	3.5	1.5	134	2.3	0.024
4	A - E	1.4	1.5	134	0.89	0.37
5	A - F	8.1	1.5	134	5.3	3.8e-07
6	A - G	5.6	1.5	134	3.7	0.00032
7	B - C	-2	1.5	134	-1.3	0.18
8	B - D	-6.8	1.5	134	-4.5	1.6e-05
9	B - E	-8.9	1.5	134	-5.9	3.3e-08
10	B - F	-2.1	1.5	134	-1.4	0.16
11	B - G	-4.7	1.5	134	-3.1	0.0026
12	C - D	-4.8	1.5	134	-3.1	0.0021
13	C - E	-6.9	1.5	134	-4.5	1.3e-05
14	C - F	-0.12	1.5	134	-0.08	0.94
15	C - G	-2.6	1.5	134	-1.7	0.086
16	D - E	-2.1	1.5	134	-1.4	0.17
= 17	D - F	4.6	1.5	134	3.1	0.0027
18	D - G	2.1	1.5	134	1.4	0.16
19	E - F	6.8	1.5	134	4.5	1.8e-05
20	E - G	4.2	1.5	134	2.8	0.0059
21	F - G	-2.5	1.5	134	-1.7	0.1

TABLE 2.33 – Tests de comparaison de moyennes par correcteur ajustées à l'effet copie. Les probabilités critiques affichées sont brutes.

2.5.5 Programme R

```
# Chargement de packages ----------------------------------


library("lsmeans")
```

```
library("car")
library("xtable")

rm(list=ls())
setwd('modelelineaire/CodesR/LMEx/')
load(file="notes.Rd")
nbourse <- 7
nadmis <- 25
# Chargement données -------------------------------------

table(resultats$correcteur, resultats$copie)



  head(resultats, n=10)
xtable(resultats)


# Anova 1 facteur correcteur ----------------------------

lm1 <- lm(note~correcteur, data= resultats)
anova(lm1)
xtable(anova(lm1))
by(resultats$note, resultats$correcteur, mean)

pdf(file="anova1-bplotCor.pdf", width=12)
boxplot(resultats$note~resultats$correcteur)
dev.off()

xtable(summary(lm1))

# Anova2 ------------------------------------------------


lmAB<- lm(note~ copie +correcteur, data= resultats)

pdf(file="Anova2-Des-Residus.pdf", width=12)
par(mfrow=c(2,2)) ##anlyse des résidus
plot(lmAB)
dev.off()

xtable(summary(lmAB), display=c("s", "f", "f", "f", "e"))
lsmeans(lmAB ,pairwise~correcteur,adjust="none")

anova(lmAB) #tests de type 1 pour le modèle AB
xtable(anova(lmAB))
```

```
Anova(lmAB)#tests de type 2 pour le modèle AB
xtable(Anova(lmAB))

lmBA<- lm(note~ correcteur + copie , data= resultats)
anova(lmBA) #tests de type 1 pour le modèle BA
xtable(anova(lmBA))



# Moyennes ajustées ------------------------------------
notes.moyennes <- as.numeric(with(resultats,
                       by(note, copie, mean)))

moyennes.ajustees <- lsmeans(lmAB ,pairwise~copie,adjust="none")
notes.ajustees <- summary(notes.ajustees)[,2]

ordre.moy <- order(notes.moyennes, decreasing=T)
ordre.moy.ajuste <- order(notes.ajustees, decreasing=T)

selectionnees.bourse <- unique(c(ordre.moy[1:nbourse],
      ordre.moy.ajuste[1:nbourse]))
selectionnees.admis <- unique(c(ordre.moy[1:nadmis],
     ordre.moy.ajuste[1:nadmis]))



final <- data.frame ( copie = selectionnees.admis,
            note.moyenne = notes.moyennes[selectionnees.admis],
            rang.m = rank(-notes.moyennes[selectionnees.admis],
            ties.method="min"),
resultat.moy  = rep(NA, length(selectionnees.admis)),
           note.ajustee = notes.ajustees[selectionnees.admis],
           rang.a =rank(-notes.ajustees[selectionnees.admis],
           ties.method="min"),
resultat.aju  = rep(NA, length(selectionnees.admis)))

final <- final[order(final$note.ajustee, decreasing =T),]

moy.res = sort(notes.moyennes, decreasing=T)[c(nbourse, nadmis)]
final$resultat.moy <- sapply( final$note.moyenne, function(d) ...
                           ... {ifelse(d>=moy.res[1],"Bourse",
                            ifelse(d>=moy.res[2], "Admis","--") )})

aju.res = sort(notes.ajustees, decreasing=T)[c(nbourse, nadmis)]
final$resultat.aju <- sapply( final$note.ajustee, function(d) ...
```

```
                                ... {ifelse(d>=aju.res[1],"Bourse",
                            ifelse(d>=aju.res[2], "Admis", "--") )})


xtable(final)



# Moyennes ajustées correcteurs
-----------------------------------------



# Moyennes ajustées ----------------------------------------
correcteur.moyennes <- as.numeric(with(resultats,
                                    by(note, correcteur, mean)))

correcteur.ajustees <- lsmeans(lmAB ,pairwise~correcteur,
                               adjust="none")
xtable(summary(correcteur.ajustees)$contrast,
          display=c("s","s","g","g","g","g","g"))
```

2.6 Analyse de la covariance

2.6.1 Problème et des données

La RATP a mis en place un système de surveillance de la qualité de l'air dans le métro parisien entre le 1er janvier et le 1er avril 2012 (SQUALES)[3]. Ce système a été installé dans 3 stations (variable `Station`) : Châtelet (ligne 4), Franklin D. Roosevelt (ligne 1) et une gare de RER : Auber (ligne A). Chaque jour (variable `jour`), au pas de temps horaire, sont mesurées deux variables climatiques, la température (variable `Temp`) et l'humidité relative (variable `Humidite`) ainsi que différents indicateurs de pollution. Dans cette étude, nous nous intéressons à la quantité de particules fines en suspension dans l'air et plus spécifiquement aux particules PM10 (variable `PM10`), particules caractérisées par un diamètre aérodynamique inférieur à 10 micromètres. Ces particules ont un effet connu sur la santé, notamment sur les voies respiratoires ([47]) et il est recommandé de ne pas dépasser 50 $\mu g/m^3$ en moyenne journalière. Le dispositif SQUALES doit permettre d'enregistrer si les seuils d'alerte sont dépassés. Dans notre analyse nous fixons l'heure d'étude à 08h00. Un extrait des données est présenté dans la table 2.34.

2.6.2 Station Châtelet

Dans un premier temps, nous concentrons l'étude sur la station Châtelet. La table 2.35 donne les effectifs, la température moyenne et l'humidité moyenne par jour. Les

3. `http://www.ratp.fr/fr/ratp/r_6167/la-qualite-de-lair-dans-les-espaces-souterrains`

```
 PM10 Temp Humidite     jour Station
  294 21.1     53.6 dimanche   Auber
  308 20.7     49.7    lundi   Auber
  231 20.3     47.4    mardi   Auber
  499 20.9     46.9 mercredi   Auber
  357 21.0     49.2    jeudi   Auber
  248 20.3     44.0 vendredi   Auber
  256 20.6     48.8   samedi   Auber
...
...
 PM10 Temp Humidite     jour Station
   37 21.7     37.4 dimanche     FDR
   41 20.7     32.2    lundi     FDR
   52 18.3     33.2    mardi     FDR
   68 18.8     39.6 mercredi     FDR
   58 18.6     38.0    jeudi     FDR
   73 18.2     40.3 vendredi     FDR
   32 17.4     45.2   samedi     FDR
```

TABLE 2.34 – Extrait des données.

températures moyennes entre les jours sont assez proches et quelques différences faibles sont observées sur l'humidité moyenne.

jour	effectifs	Température moyenne	Humidité moyenne
dimanche	13	17.19	38.81
jeudi	12	17.8	40.76
lundi	13	17.3	37.43
mardi	13	17.42	37.13
mercredi	12	17.66	39.19
samedi	13	17.58	42.1
vendredi	11	17.83	38.3

TABLE 2.35 – Effectifs, température moyenne et humidité moyenne par jour.

Effet du jour de la semaine et de la température sur la pollution aux particules fines PM10 à la station châtelet

Modélisation. On cherche à vérifier si la pollution aux PM10 dépend de la température, ou du jour ou potentiellement des deux. Pour tester cette hypothèse, on peut utiliser un modèle d'analyse de la covariance.

Rappelons l'écriture régulière du modèle d'analyse de la covariance :

$$Y_{ik} = \mu_i + \beta_i \text{Temp}_{ik} + E_{ik}, \qquad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \tag{M.2.9}$$

où Y est la concentration en PM10, Temp_{ik} est la $k^{\text{ième}}$ mesure de température le jour i, $i = 1, \ldots, I$ et $k = 1, \ldots, n_i$. Dans notre étude, $I = 7$ jours et $n_1 = n_3 = n_4 = n_6 = 13$, $n_2 = n_5 = 12$ et $n_7 = 11$. Passer de l'écriture régulière à l'écriture singulière (utilisée dans les logiciels) consiste simplement à dissocier les effets des différentes variables, i.e.

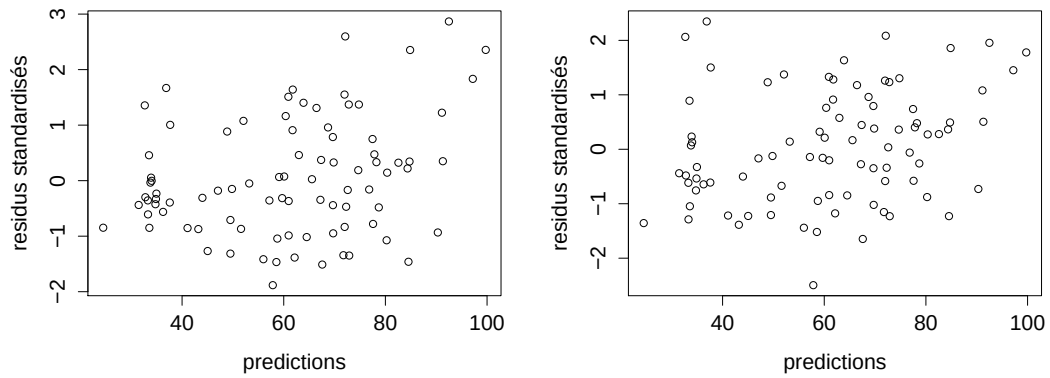


FIGURE 2.14 – Graphe des résidus standardisés du modèle M.2.10 (à gauche) et du modèle où la variable à expliquer a subi une transformation logarithmique (à droite).

à poser : $\mu_i = \mu + \alpha_i$ et $\beta_i = \beta + \gamma_i$. Ainsi, on obtient :

$$Y_{ik} = \mu + \alpha_i + \beta \text{Temp}_{ik} + \gamma_i \text{Temp}_{ik} + E_{ik}, \qquad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2). \tag{M.2.10}$$

En se référant à la partie 1.5.4, p. 40, on peut constater, sur la table 2.35, que la variable température n'est pas centrée au sein de chaque niveau du facteur Jour et ainsi le dispositif n'est pas orthogonal.

La figure 2.14, qui représente les résidus standardisés en fonction des prédictions, nous permet de valider l'hypothèse d'homoscédasticité lorsque l'on travaille sur le logarithme de la concentration en PM10. Dans toute la suite de cette partie, la variable à expliquer Y sera donc le logarithme de la concentration en PM10.

Estimations des paramètres. Comme pour l'analyse de la variance, le modèle écrit sous sa forme régulière est identifiable (matrice X de plein rang) et $\widehat{\theta} = (X'X)^{-1}X'Y$, ce qui n'est pas le cas de l'écriture singulière. Une manière simple d'estimer les différents paramètres consiste à utiliser les estimateurs obtenus dans l'écriture régulière (i) et en déduire ceux de l'écriture singulière en appliquant les contraintes d'identifiabilité (ii).

(i) Expression de $\widehat{\theta}$ dans le modèle régulier. Considérons le modèle écrit sous sa forme matricielle (voir partie 1.2.3, p. 15) en réordonnant les matrices X et θ de la façon suivante :

$$X = \begin{bmatrix} X_{(1)} & 0 & \cdots & 0 \\ 0 & X_{(2)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & X_{(I)} \end{bmatrix}, \theta = \begin{bmatrix} \mu_1 \\ \beta_1 \\ \mu_2 \\ \beta_2 \\ \vdots \\ \mu_I \\ \beta_I \end{bmatrix}$$

où $$X_{(i)} = \begin{bmatrix} 1 & \text{Temp}_{i1} \\ 1 & \text{Temp}_{i2} \\ \vdots & \vdots \\ 1 & \text{Temp}_{in_i} \end{bmatrix}.$$

La matrice X étant diagonale par bloc, il en est de même pour $(X'X)^{-1}$ et $X'Y$:

$$(X'X)^{-1} = \begin{bmatrix} (X'_{(1)}X_{(1)})^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & (X'_{(I)}X_{(I)})^{-1} \end{bmatrix},$$

$$X'Y = \begin{bmatrix} X'_{(1)}Y_{(1)} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \cdots & 0 & X'_{(I)}Y_{(I)} \end{bmatrix},$$

où $Y_{(i)} = (Y_{i1}, \dots, Y_{in_i})'$. On obtient alors que :

$$(X'X)^{-1}X'Y = \begin{bmatrix} (X'_{(1)}X_{(1)})^{-1}X'_{(1)}Y_{(1)} \\ \vdots \\ (X'_{(I)}X_{(I)})^{-1}X'_{(1)}Y_{(1)} \end{bmatrix}.$$

Ainsi les estimateurs des paramètres μ_i et β_i pour $i = 1, \dots, I$ dans un modèle ancova sont exactement les mêmes que ceux de I modèles de régressions séparées.

(ii) Déduction des expressions des estimateurs du modèle singulier. On suppose ici que les contraintes d'identifiabilité sont celles du logiciel R, $\alpha_1 = 0$ et $\gamma_1 = 0$. Ayant posé $\mu_i = \mu + \alpha_i$ et $\beta_i = \beta + \gamma_i$, on obtient facilement que

$$\begin{cases} \widehat{\mu} = \widehat{\mu}_1 & \text{soit l'ordonnée à l'origine de la droite du groupe 1} \\ \widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu}_1 & \forall i, \text{ soit la différence des ordonnées à l'origine des droites} \\ & \text{de régressions entre les groupes } i \text{ et } 1 \\ \widehat{\beta} = \widehat{\beta}_1 & \text{soit la pente du groupe 1} \\ \widehat{\gamma}_i = \widehat{\beta}_i - \widehat{\beta}_1 & \forall i, \text{ soit la différence de pentes des droites} \\ & \text{de régressions entre les groupes } i \text{ et } 1 \end{cases}$$

La table 2.36 donne les estimations ainsi que les résultats des tests de nullité des paramètres obtenus par ce logiciel. On constate que les sorties ne mentionnent jamais le niveau dimanche, qui correspond au niveau 1 et qui sert de catégorie de référence. L'estimation est faite sous la contrainte que l'effet du jour dimanche (μ_1 et γ_1) sont égaux à 0. On peut noter qu'aucun coefficient n'est significativement différent de 0. On peut également remarquer que les écarts types associés à l'erreur d'estimation sont grands (souvent du même ordre de grandeur que l'effet estimé lui-même) et donc les paramètres sont très mal estimés et conduisent à des tests peu puissants.

Si le logiciel SAS est utilisé, la référence est non plus le groupe 1 mais le groupe I, ainsi on obtient : $\widehat{\mu} = \widehat{\mu}_I$, $\widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu}_I$, $\widehat{\beta} = \widehat{\beta}_I$ et $\widehat{\gamma}_i = \widehat{\beta}_i - \widehat{\beta}_I$.

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	3.7964	0.5889	6.45	0.0000
jourjeudi	0.4155	0.8824	0.47	0.6391
jourlundi	0.1717	0.9172	0.19	0.8520
jourmardi	-0.2368	0.9285	-0.26	0.7994
jourmercredi	0.0334	0.8851	0.04	0.9700
joursamedi	-0.7145	0.8626	-0.83	0.4102
jourvendredi	0.2454	0.8791	0.28	0.7809
Temp	-0.0176	0.0338	-0.52	0.6051
jourjeudi :Temp	0.0162	0.0497	0.32	0.7463
jourlundi :Temp	0.0268	0.0526	0.51	0.6116
jourmardi :Temp	0.0518	0.0530	0.98	0.3318
jourmercredi :Temp	0.0413	0.0501	0.82	0.4121
joursamedi :Temp	0.0546	0.0490	1.11	0.2687
jourvendredi :Temp	0.0304	0.0495	0.62	0.5404

TABLE 2.36 – Estimations des paramètres du modèle complet obtenues par logiciel R.

Test du modèle. On souhaite tester les hypothèses suivantes

$$\begin{aligned} H_0 &= \{Y_{ik} = \mu + E_{ik}\} \\ \text{contre} \quad H_1 &= \{Y_{ik} = \mu + \alpha_i + \beta \text{Temp}_{ik} + \gamma_i \text{Temp}_{ik} + E_{ik}\} \end{aligned}$$

La table 2.37 donne les différentes sommes des carrés ainsi que les résultats de ce test (statistique de test et probabilité critique). On rejette l'hypothèse H_0 et on conclut que la température ou le jour de la semaine a un effet significatif sur la pollution aux PM10. Ceci peut paraitre contre intuitif avec le résultat précédent. En fait, dans le test de nullité d'un paramètre, on teste si dans le modèle complet, il est possible de considérer qu'un coefficient est nul (tous les autres étant laissés libres). Ainsi si on teste $\gamma_2 = 0$, ceci revient à comparer le modèle complet à un sous-modèle dans lequel on impose que la pente de la droite liant température et log(PM10) le lundi soit égale à celle du dimanche. Il est possible que tous ces tests apparaissent comme non significatifs et que pour autant le modèle complet soit plus pertinent que le modèle nul.

	DF	Sum of Squares	F value	Pr > F
Model	13	7.03	4.58	1.12e-05
Error	73	8.62		
Total	86	15.65		

TABLE 2.37 – Table d'analyse de la variance du modèle complet.

Tests des différents effets. Le test précédent a mis en évidence un effet significatif soit de la température, soit du jour, soit des deux. Il nous reste à savoir si effectivement ces deux variables/facteurs ont des effets significatifs sur la pollution aux particules fines. Avant d'effectuer ces tests, on rappelle les différentes déclinaisons des tests (tests de type I et de type II).

Les tests de types I et II. On se place ici dans le cas où le facteur n'a que deux modalités afin de faciliter l'interprétation des tests (on parlera de groupes correspondants aux différentes modalités). Dans la table 2.38, on rappelle les réductions utilisées pour les tests en type I et type II ainsi que les modèles théoriques mis en jeu (l'espérance de Y pour chacun des modèles). On voit clairement que le premier test à considérer porte sur l'interaction. Deux cas peuvent alors se présenter :

— l'interaction a un effet significatif. Dans ce cas, les tests sur l'effet du facteur seul (où l'on teste si il existe une différence de réponses moyennes entre les groupes pour une valeur de x égale à 0) et sur l'effet de x seul (où l'on teste si la pente moyenne de référence est nulle) n'ont pas de sens (sauf si le test en $x = 0$ est un test d'intérêt).

— l'interaction n'a pas d'effet significatif. Dans ce cas, le modèle considéré est

$$Y_{ik} = \mu + \alpha_i + \beta \mathrm{Temp}_{ik} + E_{ik}, \tag{M.2.11}$$

i.e. les pentes sont supposées parallèles (l'évolution linéaire est la même quelque soit le groupe). Le test sur l'effet de x à l'aide des type I ou II (qui donnent les mêmes résultats si la variable x est introduite après le facteur dans le modèle) consiste alors à tester si il existe un lien (ici linéaire) entre la réponse y et la variable x. Le test sur l'effet du facteur (ici introduit en premier dans le modèle) est quant à lui différent selon le type utilisé : en type II, ce test revient à savoir si il existe une différence significative entre les réponses moyennes de chaque groupe pour une valeur de x donnée (quelconque puisque les pentes sont supposées parallèles). En type I, ce test revient à comparer les comportements moyens de chaque groupe en faisant abstraction de la variabilité portée par la variable x, ce test n'a donc de sens que si la variable x n'a pas d'effet significatif sur y.

Dans le cadre de l'analyse de la pollution aux particules fines à la station Châtelet, on peut tout d'abord tester si la relation qui lie pollution et température est la même quelque soit le jour. La table 2.40, p. 101 présente les résultats de tests en type II du modèle complet (M.2.10, p. 97). L'interaction n'a pas d'effet significatif, on va donc redéfinir le modèle sans interaction pour estimer les effets principaux du jour et de la température comme défini dans le modèle M.2.11. Le fait de travailler dans le modèle sans interaction, plus parcimonieux, nous permet d'améliorer la précision de nos estimateurs et d'obtenir ainsi des tests plus puissants. Les résultats des tests de type II pour le modèle sans interaction sont donnés dans la partie droite de la table 2.40. Il n'y a pas d'effet de la température mais il y a un effet du jour de la semaine. Ainsi, il semble que la pollution aux particules fines ne dépende pas de la température tandis qu'elle dépend du jour de la semaine (probablement en liaison avec la fréquentation).

Comparaison des groupes - Définition des moyennes brutes et ajustées. Dans le modèle d'analyse de la covariance (M.2.10, p. 97), la moyenne du groupe i

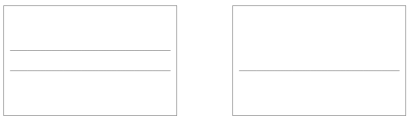
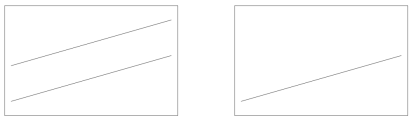
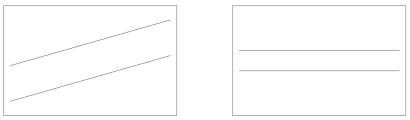
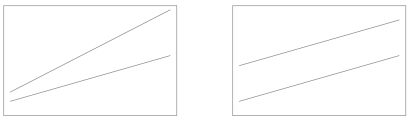
	Type I	Type II
α	$R(\alpha\|\mu)$ = $SCM_{\mu,\alpha} - SCM_{\mu}$ 	$R(\alpha\|\mu,\beta)$ = $SCM_{\mu,\alpha,\beta} - SCM_{\mu,\beta}$ 
β	$R(\beta\|\mu,\alpha)$ = $SCM_{\mu,\alpha,\beta} - SCM_{\mu,\alpha}$ 	$R(\beta\|\mu,\alpha)$
γ	$R(\gamma\|\mu,\alpha,\beta)$ = $SCM_{\mu,\alpha,\beta,\gamma} - SCM_{\mu,\alpha,\beta}$ 	$R(\gamma\|\mu,\alpha,\beta)$

TABLE 2.38 – Réductions associées aux tests de type I et de type II dans une ancova avec une variable quantitative et un facteur à 2 modalités si le facteur est mis en premier dans le modèle. Dans la partie de droite, des blancs indiquent que le résultat est le même que dans la case de gauche.

	Sum Sq	Df	F value	Pr(>F)
jour	6.61	6	9.34	1.3679e-07
Temp	0.10	1	0.86	3.5612e-01
jour :Temp	0.21	6	0.30	9.3526e-01
Residuals	8.62	73		

TABLE 2.39 – Tests de Type II pour le modèle M.2.10

	Sum Sq	Df	F value	Pr(>F)
jour	6.61	6	9.86	4.2300e-08
Temp	0.10	1	0.91	3.4279e-01
Residuals	8.83	79		

TABLE 2.40 – Tests de Type II pour le modèle M.2.11

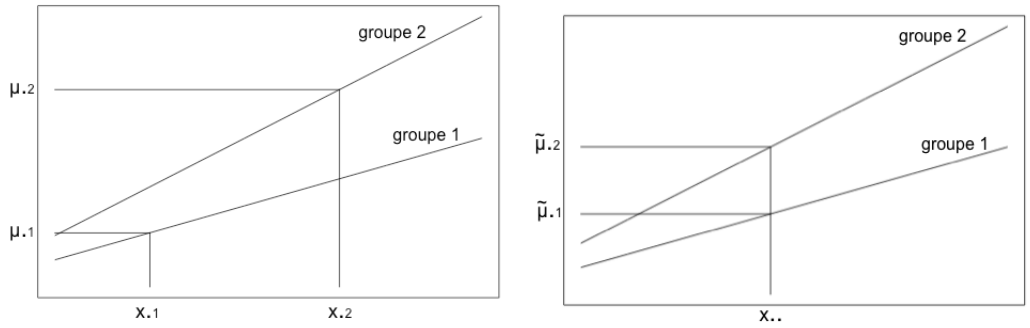


FIGURE 2.15 – Moyennes classiques (à gauche) et ajustées (à droite) pour $I = 2$ groupes.

est :

$$\begin{aligned}\mu_{i\bullet} &= \frac{1}{n_i}\sum_{k=1}^{n_i}\mathbb{E}[Y_{ik}] \\ &= \mu + \alpha_i + (\beta+\gamma_i)\frac{1}{n_i}\sum_{k=1}^{n_i}\text{Temp}_{ik} \\ &= \mu + \alpha_i + (\beta+\gamma_i)\text{Temp}_{i\bullet} \qquad (2.6)\end{aligned}$$

On note $\widehat{\mu}_{i\bullet}$ son estimation (obtenue à partir des estimations des différents paramètres du modèle) qui vaut exactement $y_{i\bullet}$. Pour chaque groupe i, elle correspond donc à la prédiction de la réponse pour une valeur de la variable $x = \text{Temp}_{i\bullet}$ qui représente la valeur moyenne de cette variable pour ce groupe. Une différence observée entre les réponses moyennes, qui pourrait amener à la conclusion d'une différence entre les deux groupes, peut être simplement dûe à une différence importante en abscisse (des $\text{Temp}_{i\bullet}$). Si cet écart en abscisse est le résultat d'un mauvais échantillonnage, il n'est pas pertinent d'utiliser ces moyennes pour comparer les groupes. Une manière de s'en affranchir serait de comparer l'écart de réponses entre les groupes obtenues pour une même abscisse x, comme par exemple la moyenne générale. C'est la définition de la moyenne ajustée (elle s'ajuste à l'effet eventuel de la variable x) qui est pour un groupe i :

$$\tilde{\mu}_{i\bullet} = \mu + \alpha_i + (\beta+\gamma_i)\text{Temp}_{\bullet\bullet}, \qquad (2.7)$$

où $\text{Temp}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}\text{Temp}_{ik}$.

La figure 2.15 illustre les différences entre moyennes classiques et moyennes ajustées.

Remarques :

— Si l'on considère le modèle (M.2.11, p. 100) sans interaction (i.e. pentes parallèles), le test de l'effet de α en type II consiste à tester l'existence de l'effet groupe quelque soit x, ici *Temp*. En effet, les pentes étant supposées égales, l'écart à l'origine (test sur α) est le même que pour n'importe quelle valeur de x. Si ce test a mis en évidence un effet significatif du groupe, alors si $I = 2$ il n'est pas nécessaire d'effectuer la comparaison des moyennes ajustées puisque le test sera le même. En revanche, si $I > 2$, il peut être intéressant de comparer les différents groupes au travers de la comparaison des moyennes ajustées.

— Dans le cas où il y a plus de 2 groupes, effectuer les comparaisons des groupes deux à deux pose un problème de tests multiples et une correction par la méthode de Bonferroni peut être effectuée pour conserver un niveau de test global raisonnable (cf paragraphe 2.3.3, p. 67 de l'exemple d'analyse de la variance à 1 facteur).
— Comparaison des groupes : moyennes ou moyennes ajustées ? Nous avons vu dans le paragraphe précédent qu'une comparaison des groupes au travers de la comparaison des moyen-nes brutes n'était pas pertinente dans le cas où la différence entre les $\text{Temp}_{i\bullet}$ était due à un problème d'échantillonnage. Dans ce cas, considérer les moyennes ajustées a du sens. Cependant, si les gammes de valeurs des $\{\text{Temp}_{ik}\}_k$ sont très différentes et que $\text{Temp}_{\bullet\bullet}$ n'entre pas dans ces gammes, les moyennes ajustées ne sont que des interpolations de ce qui se passe en ce point et le résultat est donc à prendre avec précaution, voire peut être peu pertinent.
Si la différence entre les $\text{Temp}_{i\bullet}$ n'est pas due à un mauvais échantillonnage mais plutôt à un écart qui est réellement observé, la comparaison des groupes au travers des moyennes ajustées perd de son sens et il est plus pertinent d'utiliser les moyennes brutes.

Dans notre exemple, puisque la température n'a pas de réelle influence sur la pollution, les 7 droites à considérer sont quasiment horizontales et les moyennes ou les moyennes ajustées sont sensiblement égales. Les tests de comparaison des moyennes ajustées sont données dans la table 2.41. Puisque la correction pour tests multiples a déjà été faite, si on souhaite que le niveau global du test soit de 5%, on va déclarer deux moyennes significativement différentes si la probabilité critique associée corrigée est inférieure à 5%. A la lecture de cette table, il apparaît que la pollution est significativement différente entre un jour de semaine et un jour du week-end.

Effet du jour de la semaine, de la température et de l'humidité sur le PM10 à la station châtelet

Modèle. Dans cette étude, il y a 2 variables quantitatives et un facteur, le modèle d'analyse de la covariance s'écrit :

$$Y_{ik} = \mu + \alpha_i + \beta_1 \text{Temp}_{ik} + \gamma_{1,i} \text{Temp}_{ik} + \beta_2 \text{Humidite}_{ik} + \gamma_{2,i} \text{Humidite}_{ik} + E_{ik}, \quad \text{(M.2.12)}$$

où $E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, 7$ et $k = 1, \ldots, n_i$ où les effectifs n_i sont donnés dans la table 2.35.

Diagnostic. Comme le montre la figure 2.16 à gauche, qui représente les résidus standardisés en fonction des prédictions, l'hypothèse d'homoscédasticité n'est pas validée. En passant au logarithme de PM10 (i.e. dans le modèle (M.2.12), Y_{ik} représente maintenant le logarithme de la $k^{\text{ième}}$ mesure du PM10 pour le jour i), la variance est stabilisée comme le montre la figure 2.16 à droite. C'est donc ce modèle qui est considéré dans la suite.

Test du modèle. On teste le modèle le plus simple (avec la constante) contre le modèle complet. La table 2.42 donne la table d'analyse de la variance associée. L'hypothèse

	contrast	estimate	SE	df	t.ratio	p.value
1	dimanche - jeudi	-0.68	0.13	79	-5.11	4.68e-05
2	dimanche - lundi	-0.63	0.13	79	-4.82	1.43e-04
3	dimanche - mardi	-0.66	0.13	79	-5.02	6.55e-05
4	dimanche - mercredi	-0.75	0.13	79	-5.59	6.56e-06
5	dimanche - samedi	-0.23	0.13	79	-1.78	1.00e+00
6	dimanche - vendredi	-0.77	0.14	79	-5.60	6.29e-06
7	jeudi - lundi	0.05	0.13	79	0.39	1.00e+00
8	jeudi - mardi	0.03	0.13	79	0.19	1.00e+00
9	jeudi - mercredi	-0.06	0.14	79	-0.47	1.00e+00
10	jeudi - samedi	0.45	0.13	79	3.37	2.44e-02
11	jeudi - vendredi	-0.08	0.14	79	-0.60	1.00e+00
12	lundi - mardi	-0.03	0.13	79	-0.20	1.00e+00
13	lundi - mercredi	-0.12	0.13	79	-0.87	1.00e+00
14	lundi - samedi	0.40	0.13	79	3.04	6.72e-02
15	lundi - vendredi	-0.14	0.14	79	-0.99	1.00e+00
16	mardi - mercredi	-0.09	0.13	79	-0.68	1.00e+00
17	mardi - samedi	0.43	0.13	79	3.24	3.65e-02
18	mardi - vendredi	-0.11	0.14	79	-0.80	1.00e+00
19	mercredi - samedi	0.52	0.13	79	3.85	4.95e-03
20	mercredi - vendredi	-0.02	0.14	79	-0.14	1.00e+00
21	samedi - vendredi	-0.54	0.14	79	-3.91	4.11e-03

TABLE 2.41 – Comparaisons des pollutions moyennes ajustées (à la température). Les probabilités critiques présentées ont été corrigées par la méthode de Bonferroni.

H_0 est rejetée : au moins une/un des variables/facteurs a un effet significatif sur le *log* du PM10.

Test des différents effets. La table 2.43 présente les résultats de tests en type II du modèle complet. Les interactions n'ont pas d'effets significatifs : il existe une relation

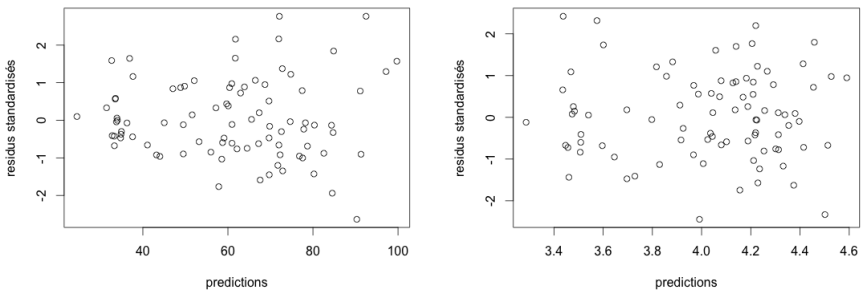


FIGURE 2.16 – Graphe des résidus standardisés du modèle (M.2.12, p. 103) (à gauche) et du modèle où la variable à expliquer a subi une transformation logarithmique (à droite).

	DF	Sum of Squares	F value	Pr > F
Model	20	9.28	4.82	0.0000
Error	66	6.36		
Total	86	15.65		

TABLE 2.42 – Table d'analyse de la variance du modèle complet.

entre les conditions atmosphériques (température et humidité) et le *log* du PM10 (effets Temp et Humidite significatifs) et cette relation est la même quels que soient les jours. Il est logique que la circulation de ces particules en suspension dépende de la condition atmosphérique. L'effet jour significatif indique qu'il existe une différence de *log* du PM10 moyen entre les jours pour une même valeur de température et d'humidité. Le modèle considéré devient alors :

$$Y_{ik} = \mu + \alpha_i + \beta_1 \text{Temp}_{ik} + \beta_2 \text{Humidite}_{ik} + E_{ik}, \qquad \text{(M.2.13)}$$

où Y_{ik} est le logarithme de la $k$ème mesure du PM10 pour le jour i.

	Sum Sq	Df	F value	Pr(>F)
jour	14997.23	6	6.60	0.0000
Temp	6308.18	1	16.67	0.0001
Humidite	8029.67	1	21.22	0.0000
jour :Temp	1263.97	6	0.56	0.7631
jour :Humidite	932.01	6	0.41	0.8696
Residuals	24976.87	66		

TABLE 2.43 – Résultats des tests de type II du modèle complet pour l'exemple du métro.

La table 2.44 donne les estimations des paramètres ainsi que les résultats du test de nullité des paramètres obtenus par le logiciel R (le jour dimanche a été pris comme référence). Pour un jour donné, une humidité donnée, les $PM10$ semblent augmenter avec la température. Pour un jour donné et une température donnée, le logarithme de la concentration en PM10 semble diminuer avec l'humidité.

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	3.2534	0.2212	14.71	0.0000
jourjeudi	0.6957	0.1181	5.89	0.0000
jourlundi	0.5984	0.1157	5.17	0.0000
jourmardi	0.6131	0.1159	5.29	0.0000
jourmercredi	0.7343	0.1180	6.22	0.0000
joursamedi	0.2819	0.1160	2.43	0.0174
jourvendredi	0.7275	0.1212	6.01	0.0000
Temp	0.0602	0.0154	3.91	0.0002
Humidite	-0.0205	0.0042	-4.88	0.0000

TABLE 2.44 – Estimations et tests sur les paramètres du modèle (M.2.13, p. 105) obtenus par le logiciel R.

Comparaison des jours. La table 2.44 ne donne que les comparaisons des *log* du PM10 moyens entre dimanche et les autres jours. Nous souhaitons ici comparer tous les jours et pour une même condition atmosphérique (même température et même humidité). Nous utilisons alors les moyennes ajustées. Les tables 2.45 et 2.46 donnent respectivement les valeurs des moyennes ajustées par jour et les tests de comparaison de ces moyennes deux à deux avec correction par la méthode de Bonferroni (résultats issus du logiciel R). On observe une plus faible quantité de particules le week-end, différence significative par rapport aux jours de la semaine. Ces particules sont en effet essentiellement produites par le roulement et le freinage des rames en circulation, or les métros sont moins fréquents le week-end, en particulier le dimanche.

jour	lsmean	SE	df	lower.CL	upper.CL
dimanche	3.51	0.08	78.00	3.35	3.67
jeudi	4.20	0.09	78.00	4.04	4.37
lundi	4.11	0.08	78.00	3.94	4.27
mardi	4.12	0.08	78.00	3.96	4.29
mercredi	4.24	0.09	78.00	4.07	4.41
samedi	3.79	0.08	78.00	3.63	3.96
vendredi	4.24	0.09	78.00	4.06	4.41

TABLE 2.45 – Moyennes ajustées par jour.

	estimate	SE	df	t.ratio	p.value
dimanche - jeudi	-0.70	0.12	78.00	-5.89	0.00
dimanche - lundi	-0.60	0.12	78.00	-5.17	0.00
dimanche - mardi	-0.61	0.12	78.00	-5.29	0.00
dimanche - mercredi	-0.73	0.12	78.00	-6.22	0.00
dimanche - samedi	-0.28	0.12	78.00	-2.43	0.37
dimanche - vendredi	-0.73	0.12	78.00	-6.01	0.00
jeudi - lundi	0.10	0.12	78.00	0.82	1.00
jeudi - mardi	0.08	0.12	78.00	0.70	1.00
jeudi - mercredi	-0.04	0.12	78.00	-0.32	1.00
jeudi - samedi	0.41	0.12	78.00	3.50	0.02
jeudi - vendredi	-0.03	0.12	78.00	-0.26	1.00
lundi - mardi	-0.01	0.12	78.00	-0.13	1.00
lundi - mercredi	-0.14	0.12	78.00	-1.15	1.00
lundi - samedi	0.32	0.12	78.00	2.71	0.17
lundi - vendredi	-0.13	0.12	78.00	-1.07	1.00
mardi - mercredi	-0.12	0.12	78.00	-1.03	1.00
mardi - samedi	0.33	0.12	78.00	2.83	0.12
mardi - vendredi	-0.11	0.12	78.00	-0.95	1.00
mercredi - samedi	0.45	0.12	78.00	3.81	0.01
mercredi - vendredi	0.01	0.12	78.00	0.05	1.00
samedi - vendredi	-0.45	0.12	78.00	-3.65	0.01

TABLE 2.46 – Résultats des tests de comparaisons des moyennes ajustées avec correction par la méthode de Bonferroni.

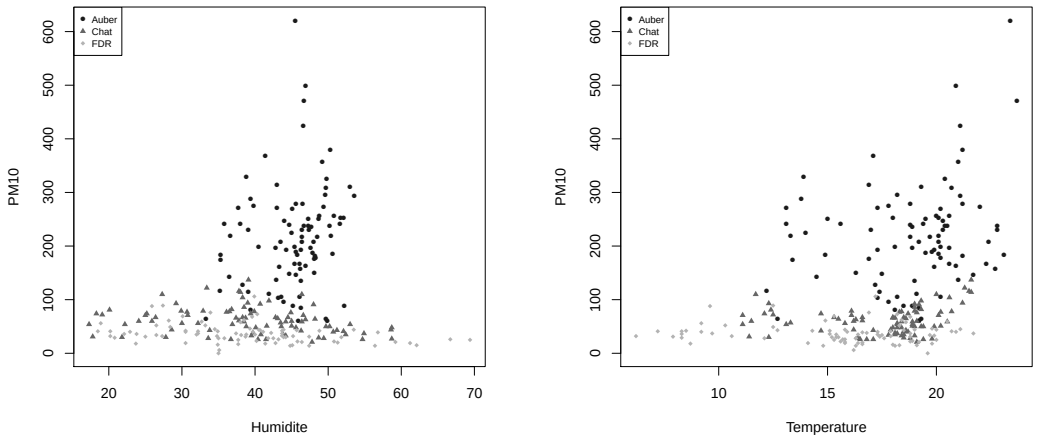


FIGURE 2.17 – Graphe de PM10 en fonction de l'humidité (à gauche) et de la température (à droite) par station.

Ajout du facteur Station

Dans cette nouvelle étude, nous considérons toutes les données, à savoir les données des 3 stations.

Analyse descriptive. La table 2.47 donne la répartition des observations, la température moyenne, l'humidité moyenne et le PM10 moyen par jour et station.

Auber	dimanche	jeudi	lundi	mardi	mercredi	samedi	vendredi	Global
Effectifs	13	13	13	13	13	13	12	90
Temp. moy.	18.6	18.99	18.54	18.68	18.85	19.26	19.21	18.87
Hum. moy.	44.87	46.2	44.52	44.51	45.16	46.10	45.24	45.23
PM10. moy.	158.07	221.38	225	267	244.23	185.69	220.41	217.36
Châtelet	**dimanche**	**jeudi**	**lundi**	**mardi**	**mercredi**	**samedi**	**vendredi**	**Global**
Effectifs	13	12	13	13	12	13	11	87
Temp. moy.	17.19	17.8	17.3	17.42	17.66	17.58	17.83	17.53
Hum. moy.	38.81	40.76	37.43	37.13	39.19	42.1	38.3	39.10
PM10. moy.	34.38	70.16	64.69	67.07	73	44.92	75.45	60.82
FDR	**dimanche**	**jeudi**	**lundi**	**mardi**	**mercredi**	**samedi**	**vendredi**	**Global**
Effectifs	11	13	13	13	13	10	12	85
Temp. moy.	16.45	15.92	15.29	15.10	15.21	16.91	16.07	15.8
Hum. moy.	38.14	41.76	36.43	37.18	39.48	39.61	37.84	38.62
PM10. moy.	21.72	42.23	39.30	36.46	41.07	27.9	45.33	36.82

TABLE 2.47 – Effectifs, Température moyenne, Humidité moyenne et PM10 moyen par croisement jour et station.

La figure 2.17 représente le PM10 en fonction de l'humidité et de la température et de la station.

On observe une grande différence entre la station Auber et les deux autres stations sur la pollution, plus faible sur la température et l'humidité. Pour Auber, les mesures sont effectuées au niveau du RER alors que celles faîtes sur les stations Châtelet et FDR sont faites dans le métro, ce qui peut expliquer cette différence. Le RER semblerait donc produire davantage de pollutions aux particules fines que le métro.

Modèle. Il y a un facteur supplémentaire (Station) par rapport à l'étude précédente. Le modèle s'écrit :

$$\begin{aligned} Y_{ijk} =& \mu + \alpha_i + \beta_j + \gamma_{ij} + \beta_1 \text{Temp}_{ijk} + \beta_2 \text{Humidite}_{ijk} \\ &+ \gamma_{1,i}\text{Temp}_{ijk} + \eta_{1,j}\text{Temp}_{ijk} + \lambda_{1,ij}\text{Temp}_{ijk} + \\ &+ \gamma_{2,i}\text{Humidite}_{ijk} + \eta_{2,j}\text{Humidite}_{ijk} + \lambda_{2,ij}\text{Humidite}_{ijk} + E_{ijk}, \end{aligned} \quad \text{(M.2.14)}$$

où $E_{ijk} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, Y_{ijk} est le logarithme de la $k^{\text{ième}}$ mesure du PM10 pour le jour i à la station j (pour stabiliser la variance, comme précédemment), α représente l'effet principal du jour, β l'effet principal de la station, $i = 1, \ldots, 7$, $j = 1, \ldots, 3$ et $k = 1, \ldots, n_{ij}$. Les valeurs des effectifs pour tous les croisements (n_{ij}) sont donnés dans la table 2.47.

	DF	Sum of Squares	F value	Pr > F
Model	62	172.47	13.61	< 2.2e-16
Error	199	40.675		
Total	261	213.148		

TABLE 2.48 – Table d'analyse de la variance du modèle complet.

	Sum Sq	Df	F value	Pr(>F)
jour	13.32	6	10.86	0.0000
Station	127.77	2	312.55	0.0000
Temp	0.62	1	3.03	0.0832
Humidite	2.91	1	14.21	0.0002
jour :Station	0.95	12	0.39	0.9668
jour :Temp	2.03	6	1.66	0.1335
Station :Temp	0.53	2	1.31	0.2727
jour :Humidite	1.49	6	1.22	0.2986
Station :Humidite	0.67	2	1.64	0.1965
jour :Station :Temp	0.83	12	0.34	0.9812
jour :Station :Humidite	0.37	12	0.15	0.9996
Residuals	40.67	199		

TABLE 2.49 – Résultats des tests de type II du modèle complet.

Test du modèle et test des différents effets. La table 2.48 donne la table d'analyse de la variance du modèle complet. L'hypothèse H_0 est rejetée : au moins une/un des variables/facteurs a un effet significatif sur le *log* du PM10. On cherche maintenant à déterminer quels sont les effets significatifs. La table 2.49 représente les résultats

des tests de type II. Certaines interactions n'ont pas d'effets significatifs, d'autres plus faiblement. On effectue une sélection pas à pas en retirant à chaque itération l'effet le moins significatif. En effet, retirer un effet permet de gagner des degrés de liberté pour l'estimation de la variance σ^2 et ainsi obtenir une estimation plus précise, donc une plus grande puissance de test (rappelons que cette estimation est le dénominateur de chaque statistique de test, ainsi des tests faiblement significatifs peuvent le devenir plus fortement). Le modèle retenu est le suivant :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \beta_1 \text{Temp}_{ijk} + \beta_2 \text{Humidite}_{ijk} + \eta_{2,j} \text{Humidite}_{ijk} + E_{ijk}. \quad \text{(M.2.15)}$$

La table 2.50 donne les résultats des tests de type II associés : il existe une relation entre le *log* du PM10 et les conditions atmosphériques mais seul le lien avec l'humidité dépend de la station. Le lien avec la température est ici beaucoup moins significatif. La table 2.51 donne les estimations des différents paramètres obtenues avec le logiciel R. La station Auber a été prise comme référence. On voit qu'il y a un changement de signe dans la relation entre le *log* du PM10 et l'humidité : elle est positive pour cette station mais négative (0.0126-0.0260 pour Châtelet et 0.0126-0.0298 pour Franklin Roosevelt) pour les deux autres.

	Sum Sq	Df	F value	Pr(>F)
jour	14.07	6	12.49	0.0000
Station	130.60	2	347.97	0.0000
Temp	0.68	1	3.60	0.0590
Humidite	2.83	1	15.06	0.0001
Station :Humidite	1.32	2	3.52	0.0312
Residuals	46.73	249		

TABLE 2.50 – Résultats des tests de type II du modèle retenu.

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	3.8503	0.4626	8.32	0.0000
jourjeudi	0.5820	0.1004	5.79	0.0000
jourlundi	0.5964	0.0995	5.99	0.0000
jourmardi	0.5938	0.0995	5.97	0.0000
jourmercredi	0.6896	0.1002	6.89	0.0000
joursamedi	0.2820	0.1017	2.77	0.0060
jourvendredi	0.6566	0.1023	6.42	0.0000
StationChat	-0.1196	0.5067	-0.24	0.8137
StationFDR	-0.5092	0.4988	-1.02	0.3083
Temp	0.0202	0.0107	1.90	0.0590
Humidite	0.0126	0.0109	1.15	0.2501
StationChat :Humidite	-0.0260	0.0114	-2.28	0.0232
StationFDR :Humidite	-0.0298	0.0113	-2.65	0.0086

TABLE 2.51 – Estimations et tests sur les paramètres du modèle retenu obtenus avec le logiciel R.

Comparaison des jours. La comparaison des jours au moyen des moyennes ajustées (ajustées à la température, à l'humidité mais aussi au fait que les stations peuvent être différentes) donne le même résultat que pour l'étude sur la station Châtelet uniquement, à savoir qu'il existe une différence significative de pollution entre le week-end et la semaine (plus marquée entre le dimanche et la semaine). Les sorties ne sont pas données ici.

Comparaison des stations. Les tables 2.52 et 2.53 donnent respectivement les valeurs des moyennes ajustées (à la température, à l'humidité et au jour) par station et les tests de comparaison de ces moyennes deux à deux avec correction par la méthode de Bonferroni (résultats issus du logiciel R). La pollution est significativement différente entre les 3 stations. La station la plus polluée en PM10 est la station Auber, comme cela avait été mentionné dans l'analyse descriptive.

	Station	lsmean	SE	df	lower.CL	upper.CL
Auber	Auber	5.21	0.06	249.00	5.08	5.33
Chat	Chat	4.02	0.05	249.00	3.93	4.11
FDR	FDR	3.47	0.05	249.00	3.37	3.57

TABLE 2.52 – Moyennes ajustées par station.

	estimate	SE	df	t.ratio	p.value
Auber - Chat	1.19	0.08	249.00	15.11	0.00
Auber - FDR	1.73	0.08	249.00	21.81	0.00
Chat - FDR	0.55	0.07	249.00	7.83	0.00

TABLE 2.53 – Résultats des tests de comparaisons des moyennes ajustées avec correction par la méthode de Bonferroni.

Programme SAS

Le jeu de données a été au préalable mis en forme en ne conservant que les variables d'intérêts.

Lecture des données et extraction du jeu de données concernant la station Châtelet.

```
data RATP;
     infile 'RATP.csv' dlm=";";
     input PM10 Temp Humidite jour$ Station$;
run;
proc sort data=RATP;
by Station;
run;

data RATP_Chatelet;
set RATP;
```

```
set RATP;
where (Station like '"Chat"');
run;
proc print data=RATP_Chatelet;
run;
```

Etude de la station Châtelet.

```
proc sort data=RATP_Chatelet;
by jour;
run;

proc freq data=RATP_Chatelet;
table jour;
run;

proc means data=RATP_Chatelet;
var Temp Humidite;
by jour;
run;

proc glm data=RATP_Chatelet;
class jour;
model PM10=jour Temp jour*Temp Humidite jour*Humidite jour*Temp;
output out=ResPred r=residu p=predite;
run;quit;

proc gplot data=ResPred;
plot residu*predite / vref=0;
run;

data RATP_Chatelet;
set RATP_Chatelet;
logPM10=log(PM10+1);
run;

proc glm data=RATP_Chatelet;
class jour;
model logPM10=jour Temp jour*Temp Humidite jour*Humidite jour*Temp
                / SS1 SS2;
output out=ResPred r=residu p=predite;
run;quit;

proc glm data=RATP_Chatelet;
class jour;
model logPM10=jour Temp Humidite /SS1 SS2 solution;
lsmeans jour / pdiff tdiff ADJUST=BON ;
```

```
run;quit;
```

Etude de toutes les stations.

```
proc freq data=RATP;
table Station*jour;
run;

proc sort data=RATP;
by Station jour;
run;

proc means data=RATP;
var Temp Humidite PM10;
by Station jour;
run;

proc gplot data=RATP;
plot PM10*Humidite=Station;
run;quit;
proc gplot data=RATP;
plot PM10*Temp=Station;
run;quit;

data RATP;
set RATP;
logPM10=log(PM10+1);
run;

proc glm data=RATP;
class Station jour;
model logPM10=jour Station Station*jour Temp jour*Temp Station*Temp
    jour*Station*Temp Humidite jour*Humidite
    Station*Humidite jour*Station*Humidite / SS1 SS2;
run;quit;

proc glm data=RATP;
class Station jour;
model logPM10=jour Station Temp Humidite Station*Humidite
      / SS1 SS2 SS3 solution;
lsmeans jour / pdiff tdiff ADJUST=BON ;
lsmeans Station / pdiff tdiff ADJUST=BON ;
run;quit;
```

Programme R

Lecture des données et extraction du jeu de données concernant la station Châtelet.

```
ratpData=read.table("RATP.csv",sep=";",header=FALSE)
colnames(ratpData)=c("PM10","Temp","Humidite","jour","Station")
chateletData= ratpData[ratpData$Station=="Chat" ,]
```

Etude de la station Châtelet.

```
table(chateletData$jour)
by(chateletData$Temp,chateletData$jour,mean)
by(chateletData$Humidite,chateletData$jour,mean)
lm.chat=lm(PM10~jour+Temp+jour*Temp+Humidite+jour*Humidite+jour*Temp,
    data=chateletData)
par(mfrow=c(2,2))
plot(lm.chat)
par(mfrow=c(1,1))
Residus.standardises=rstandard(lm.chat)
plot(lm.chat$fitted.values,Residus.standardises,
     xlab="predictions",ylab="residus standardisés")
lm.chat.log=lm(log(PM10)~jour+Temp+jour*Temp+Humidite+jour*Humidite,
    data=chateletData)
plot(lm.chat.log$fitted.values,Residus.standardises,
     xlab="predictions",ylab="residus standardisés")
lm.chat.log0=lm(log(PM10)~1, data=chateletData)
anova(lm.chat.log0,lm.chat.log)
anova(lm.chat.log0,lm.chat.log)
library(car)
Anova(lm.chat.log)
lm.chat.log.select=m(log(PM10)~jour+Temp+Humidite, data=chateletData)
Anova(lm.chat.log.select)
summary(lm.chat.log.select)
library(lsmeans)
lsmeans(lm.chat.log.select,pairwise~jour,adjust="bonferroni")
```

Etude de toutes les stations.

```
table(ratpData$Station, ratpData$jour)
by(ratpData$PM10,ratpData$Station,mean)
by(ratpData$Temp,ratpData$Station,mean)
by(ratpData$Humidite,ratpData$Station,mean)
z=ratpData[ratpData$Station=="Chat",]
table(z$jour)
sum(table(z$jour))
mean(z$Temp)
mean(z$Humidite)
mean(z$PM10)
par(mfrow=c(1,1))
plot(ratpData$Humidite,ratpData$PM10,col=ratpData$Station,
    pch=as.numeric(ratpData$Station),cex=0.7,xlab="Humidite",
    ylab="PM10")
```

```
legend("topleft", legend=unique(ratpData$Station),
    pch=1:length(ratpData$Station),col=1:length(ratpData$Station),
    cex=0.7)
plot(ratpData$Temp,ratpData$PM10,col=ratpData$Station,
    pch=as.numeric(ratpData$Station),cex=0.7,xlab="Temperature",
    ylab="PM10")
legend("topleft", legend=unique(ratpData$Station),
    pch=1:length(ratpData$Station),col=1:length(ratpData$Station),
    cex=0.7)
lm.station=lm(log(PM10+1)~jour+Station+Station*jour+
    Temp+jour*Temp+Station*Temp+jour*Station*Temp+
    Humidite+jour*Humidite+Station*Humidite+jour*Station*Humidite,
    data=ratpData)
Anova(lm.station)
lm.station0=lm(log(PM10+1)~1, data=ratpData)
anova(lm.station0,lm.station)
lm.station.select=lm(log(PM10+1)~jour+Station+Temp+
    Humidite+Station*Humidite, data=ratpData)
Anova(lm.station.select)
summary(lm.station.select)
library(lsmeans)
lsmeans(lm.station,pairwise~Station,adjust="bonferroni")
lsmeans(lm.station,pairwise~jour,adjust="bonferroni")
```

Chapitre 3

Modèle linéaire généralisé

3.1 Introduction

Dans beaucoup de cas le modèle linéaire étudié au chapitre précédent suffit pour décrire correctement une relation entre une variable à expliquer et des variables explicatives, pour tester la significativité et comparer les effets des variables explicatives. Néanmoins le recours à un modèle linéaire est soumis au respect d'hypothèses qui dans certaines circonstances ne sont pas vérifiées et dans ces cas il ne représente pas l'outil adapté à l'analyse envisagée.
Ces hypothèses, au nombre de quatre, stipulent que

1. la relation entre l'espérance de la variable à expliquer et les variables explicatives est une relation linéaire. Il existe évidemment beaucoup d'exemples pour lesquels cette relation a une forme non linéaire (courbe de croissance par exemple, où la relation entre le temps et la taille est de forme sigmoïde). Cette contrainte de linéarité entraine en particulier qu'on ne peut pas imposer de borne à l'espérance de la variable à expliquer et si celle-ci est binaire (présence-absence d'une maladie par exemple), ou positive (cumuls de pluie), la prévision par un modèle linéaire peut produire des valeurs en dehors de l'ensemble admissible.
2. Les observations sont distribuées suivant une loi normale. Cette hypothèse est notamment essentielle pour réaliser les tests. Grâce au théorème central limite, le modèle linéaire est robuste aux écarts à la normalité, mais dans certains cas, par exemple si les observations sont issues d'une loi discrète, ou si les écarts à la moyenne présentent une forte dissymétrie, l'hypothèse de normalité n'est plus tenable. Dans ce cas il est nécessaire de modéliser les observations par une loi mieux adaptée.
3. La variance des variables aléatoires représentant les observations est constante. Or il peut arriver que la variance varie en fonction de la moyenne, c'est le cas des variables aléatoires suivant une loi de Poisson, très souvent utilisée pour modéliser des comptages ou bien des variables suivant une loi exponentielle ou Gamma, utilisées pour modéliser des observations qui doivent rester positives.
4. Les variables aléatoires représentant les observations sont non corrélées. Cette hypothèse n'est pas vérifiée dans les cas où les conditions de l'expérience entrainent fatalement des corrélations entre individus : mesures réalisées sur des individus

qui partagent un même ascendant, une même cage, une même parcelle ou des mesures répétées réalisées sur le même individu en des instants différents.

Le modèle linéaire généralisé (GLM) est une extension du modèle linéaire permettant de s'affranchir des trois premières hypothèses et de traiter des observations dont la loi de probabilité appartient à une famille de lois élargie.
Plus précisément, soit $Y = (Y_1, \dots, Y_n)$ le vecteur des observations, X la matrice du plan d'expériences regroupant les variables explicatives, x_i étant le vecteur ligne des variables explicatives associées à l'observation i, θ le vecteur des paramètres. Alors que le modèle linéaire s'écrit

$$Y = X\theta + E$$

avec $Y_i \sim \mathcal{N}(x_i\theta, \sigma^2)$ pour l'observation i, ce qui conduit à $\mathbb{E}[Y_i] = x_i\theta$, le modèle linéaire généralisé est la donnée d'une loi de probabilité pour les Y_i et d'une fonction g appelée fonction de lien telle que

$$g(\mathbb{E}[Y_i]) = x_i\theta$$

Cela permet d'établir une relation non linéaire entre l'espérance de la variable à expliquer et les variables explicatives et d'envisager des observations de nature variée comme des données de présence/absence, des taux de succès pour des traitements, des données de comptage d'espèces, ou encore des durées de vie ou autres variables positives dissymétriques.
Comme dans le cas du modèle linéaire, les variables explicatives peuvent être quantitatives (régression), qualitatives (anova) ou les deux (ancova).

3.2 Modélisation

3.2.1 Famille exponentielle naturelle

La famille exponentielle naturelle est une famille de lois de probabilité qui contient entre autres des lois aussi usuelles que la loi normale, la loi de Bernoulli, la loi binomiale, la loi de Poisson, la loi Gamma ... Ces lois ont en commun une écriture sous forme exponentielle qui va permettre d'unifier la présentation des résultats.

Définition 3.2.1. *Soit f_Y (resp.P_Y) la densité (resp. loi) de probabilité de la variable Y. f_Y (resp. P_Y) appartient à la famille exponentielle naturelle si elle s'écrit sous la forme*

$$f_Y(y)(resp.\ P_Y(Y = y)) = \exp\Big(\frac{1}{\gamma(\phi)}(y\omega - b(\omega)) + c(y, \phi)\Big) \tag{3.1}$$

où c est une fonction dérivable, b est trois fois dérivable et sa dérivée première b' est inversible. Le paramètre ω est appelé paramètre naturel de la loi. ϕ est un paramètre appelé paramètre de nuisance ou de dispersion.

Propriété 3.2.1. *Si la densité f_Y appartient à la famille exponentielle naturelle, alors*
- $\mathbb{E}(Y) = \mu = b'(\omega)$
- $\mathbb{V}(Y) = \gamma(\phi)b''(\omega)$

La démonstration est donnée en section 3.7, p. 130.

Exemples classiques

Pour montrer qu'une loi de probabilité appartient à la famille exponentielle naturelle, il suffit de l'écrire sous la forme d'une exponentielle et d'identifier les termes. C'est le cas pour les lois de probabilité classiques suivantes.

— Loi gaussienne

$$f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) = \frac{e^{-y^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{\sigma^2}(y\mu - \frac{\mu^2}{2})\right)$$

ce qui donne $\omega = \mu$, $b(\omega) = \frac{\omega^2}{2}$, $\gamma(\phi) = \phi = \sigma^2$.

— Loi de Poisson

$$P[Y = y;\lambda] = \frac{\lambda^y}{y!}e^{-\lambda} = \frac{1}{y!}\exp\left(y\log\lambda - \lambda\right)$$

ce qui donne $\omega = \log\lambda$, $b(\omega) = \lambda = \exp(\omega)$, $\gamma(\phi) = 1$.

— Loi de Bernoulli

$$P[Y = y;p] = p^y(1-p)^{1-y} = \exp\left(y\log\frac{p}{1-p} + \log(1-p)\right)$$

ce qui donne $\omega = \log\frac{p}{1-p}$, $b(\omega) = -\log(1-p) = \log(1+\exp\omega)$, $\gamma(\phi) = 1$.

3.2.2 Modèle

La démarche d'écriture d'un modèle linéaire généralisé est constituée de deux étapes.

1. Choix d'une loi de probabilité pour les variables aléatoires Y_i au sein de la famille exponentielle naturelle.
2. Modélisation du lien entre l'espérance des Y_i et les variables explicatives au travers d'une fonction g inversible :
$$g(\mu_i) = x_i\theta$$

La fonction g est appelée fonction de lien du modèle linéaire généralisé.

En notant de manière générale pour un vecteur y, $f(y)$ le vecteur $(f(y_1),\ldots,f(y_n))$ où f est une fonction on peut écrire le modèle linéaire généralisé sous la forme matricielle suivante :

$$g(\mathbb{E}(Y)) = X\,\theta$$

Le choix de la loi de la probabilité appartenant à la famille exponentielle est dicté par la nature des données : loi de Bernoulli pour des observations de type binaire, loi binomiale pour un nombre de succès, une loi de Poisson pour des comptages, loi exponentielle pour des durées de survie, sont des choix standards, mais d'autres possibilités sont envisageables.

3.2.3 Choix de la fonction de lien

En choisissant comme fonction de lien une bijection entre $\mathbb{R}$ et l'ensemble dans lequel est définie l'espérance μ on donne un sens à la relation linéaire $g(\mu) = X\theta$. Très souvent on choisit la fonction de lien g comme étant la fonction qui transforme l'espérance μ

en le paramètre naturel de la loi. D'après l'expression de l'espérance μ donnée dans la propriété 3.2.1, cela revient à choisir $g(\mu) = (b')^{-1}(\mu)$.

Le modèle linéaire entre évidemment dans le cadre du modèle linéaire généralisé. Les variables Y_i suivent une loi normale de moyenne μ_i et de variance σ^2 dont le paramètre naturel est μ_i et la fonction de lien naturel est la fonction identité.

Si les observations sont des comptages modélisés par une loi de Poisson, la fonction de lien naturel est la fonction log.

Si ce sont des données binaires suivant une loi de Bernoulli, la fonction de lien naturel est la fonction logit, définie par $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. Cependant toute fonction bijective de $]0;1[$ dans $\mathbb{R}$ peut être candidate, en particulier les inverses de fonctions de répartition de variables aléatoires continues sur $\mathbb{R}$ et notamment la fonction probit qui est l'inverse de la fonction de répartition de la loi normale centrée réduite.

De manière générale le choix de la fonction de lien est une liberté supplémentaire dans la démarche de modélisation. Néanmoins le choix spécifique de la fonction de lien naturel est motivé par des considérations théoriques, il permet d'assurer la convergence de l'algorithme d'estimation utilisé classiquement (algorithme de Newton-Raphson) vers le maximum de vraisemblance (voir annexe 12, p. 322).

En pratique, si aucune raison de choisir une fonction de lien spécifique ne s'impose, le choix par défaut consiste à choisir la fonction de lien naturel.

3.3 Estimation des paramètres

Le modèle étant posé, il s'agit d'estimer le vecteur de paramètres $\theta = (\theta_1, \ldots, \theta_p)'$ et le paramètre de dispersion ϕ. Notons que ce dernier paramètre n'est le plus souvent pas le paramètre d'intérêt, il n'apparaît pas en effet dans la partie explicative (i.e. l'espérance). Nous utilisons ici la méthode classique d'estimation du maximum de vraisemblance.

3.3.1 Vraisemblance

Considérons un échantillon de n variables aléatoires indépendantes $Y_1, \ldots, Y_n$ dont les densités de probabilité f_{Y_i} sont issues de la famille exponentielle et $y = (y_1, \ldots, y_n)$ une réalisation de cet échantillon. Y_i est la réponse au point $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Si la fonction de lien utilisée est celle du lien naturel, on a $\omega_i = g(\mathbb{E}(Y_i)) = x_i\theta$, la vraisemblance en y s'écrit

$$\mathcal{L}(y;\theta,\phi) = \prod_{i=1}^{n} f(y_i;\omega_i,\phi) = \prod_{i=1}^{n} f(y_i;x_i\theta,\phi)$$

et la log-vraisemblance

$$\ell(y;\theta,\phi) = \sum_{i=1}^{n}\left[\frac{1}{\gamma(\phi)}(y_i x_i\theta - b(x_i\theta)) + c(y_i,\phi)\right]$$

Les valeurs de θ et de ϕ qui rendent maximale cette fonction de log-vraisemblance sont solutions du système d'équations aux dérivées partielles suivant :

$$\begin{cases} \dfrac{\partial \ell(y;\theta,\phi)}{\partial \theta_j} = 0 \quad \text{pour } j = 1,\dots,p \\ \dfrac{\partial \ell(y;\theta,\phi)}{\partial \phi} = 0 \end{cases}$$

D'après l'expression de la log-vraisemblance donnée ci-dessus, on peut remarquer que le paramè-tre ϕ n'apparaîtra pas dans les équations relatives au paramètre θ et la maximisation de la log-vraisemblance en θ ne dépend pas de ϕ. Comme dans le cadre du modèle linéaire, la résolution du système à p équations donne une estimation de θ qu'on injecte ensuite dans la dernière équation pour obtenir une estimation de ϕ (σ^2 dans le cas du modèle linéaire).

Pour simplifier l'écriture, on pose $\gamma(\phi) = 1$ dans cette partie, ce qui ne change pas les résultats obtenus. On a alors

$$\frac{\partial \ell(y;\theta,\phi)}{\partial \theta_j} = \sum_{i=1}^{n} x_i^j \left[y_i - b'(x_i\theta) \right] \qquad j = 1,\dots,p \tag{3.2}$$

Et donc :

$$\sum_{i=1}^{n} x_i \left[y_i - b'(x_i\theta) \right] = 0.$$

Ce système n'est linéaire que si $b'(x) = x$, c'est à dire si f est une densité gaussienne et le modèle est un modèle linéaire. Pour tous les autres modèles linéaires généralisés, ce système à p équations est un système non linéaire en θ et il n'y a pas d' expression explicite pour les estimateurs. Pour obtenir les estimations du maximum de vraisemblance, on a recours à des algorithmes d'optimisation itératifs. Les deux algorithmes les plus utilisés sont l'algorithme de Newton-Raphson et l'algorithme du Fisher-scoring, mais d'autres algorithmes basés sur la descente de gradient, plus stables numériquement comme les quasi-Newton donnent de meilleurs résultats. L'algorithme de Newton-Raphson est présenté en annexe 12, p. 322, les différences avec l'algorithme du Fisher-scoring y sont précisées.

Prédiction A partir d'une estimation $\hat{\theta}$ de θ, on obtient une estimation de $\hat{\omega}_i = x_i\hat{\theta}$. La prédiction par le modèle au point x_i est alors tout simplement l'estimation de la moyenne :

$$\hat{\mu}_i = g^{-1}(\hat{\omega}_i)$$

Dans le cadre du modèle linéaire uniquement cette prédiction est aussi la prédiction de y_i.

3.3.2 Propriétés de l'estimateur du maximum de vraisemblance

Notons $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance (EMV). Cet estimateur vérifie les propriétés suivantes :

Théorème 3.3.1. *Sous certaines conditions de régularité de la densité de probabilité, l'EMV possède les propriétés suivantes*

- *$\hat{\theta}_n$ converge en probabilité vers θ (ce qui implique que $\hat{\theta}_n$ est asymptotiquement sans biais)*

— $\hat{\theta}_n$ *converge en loi vers une loi gaussienne*

$$I_n(\theta,\phi)^{1/2}\left(\hat{\theta}_n-\theta\right)\xrightarrow{loi}\mathcal{N}(0,Id)$$

où $I_n(\theta,\phi)=-\mathbb{E}[\frac{\partial^2\ell(y;\theta,\phi)}{\partial^2\theta}]$ *est la matrice d'information de Fisher évaluée en* θ *et* ϕ *(vraie valeur des paramètres) sur un échantillon de taille* n*.*

Lorsque g *est la fonction de lien naturel, l'information de Fisher vaut*

$$I_n(\theta,\phi)=\frac{1}{\gamma(\phi)}X'\mathbb{V}(Y)X. \tag{3.3}$$

La preuve de ce théorème est donnée dans la partie 5.5 de [56].

La matrice d'information de Fisher $I_n(\theta,\phi)$ dépend des vraies valeurs des paramètres θ et ϕ qui sont inconnues. Classiquement on évalue l'information de Fisher en $\hat{\theta}$ et $\hat{\phi}$.

Ce résultat permet d'établir des intervalles de confiance de niveau asymptotique $1-\alpha$ pour les paramètres θ_j. De $I_n(\theta_j,\phi)_{jj}^{1/2}\left(\hat{\theta}_j-\theta_j\right)\xrightarrow{loi}\mathcal{N}(0,1)$ on déduit en prenant l'information de Fisher au point $(\hat{\theta}_j,\hat{\phi})$

$$IC_{1-\alpha}(\theta_j)=\left[\hat{\theta}_j-u_{1-\alpha/2}I(\hat{\theta},\hat{\phi})_{jj}^{-1/2}\ ;\ \hat{\theta}_j+u_{1-\alpha/2}I(\hat{\theta},\hat{\phi})_{jj}^{-1/2}\right]$$

où $u_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

3.4 Tests d'hypothèses

Contrairement au cas du modèle linéaire, la loi de l'estimateur du maximum vraisemblance n'est connue que de manière asymptotique (pour n suffisamment grand). Cela entraîne que toute la démarche d'analyse des modèles linéaires généralisés (tests de comparaison de modèles, intervalles de confiance des valeurs des paramètres, ...) est conduite dans un cadre asymptotique.

Cette section présente différents tests d'hypothèses qui vont permettre d'examiner les qualités du modèle, de déterminer si les différentes variables explicatives présentes dans le modèle sont pertinentes ou non. Les tests de modèles emboîtés sont les plus généraux et permettent de répondre à la majorité des questions qui se posent. Les tests de chacun des effets ou de combinaisons linéaires des paramètres permettent de reproduire l'approche classique présentée dans le modèle linéaire.

3.4.1 Test de modèles emboîtés

Le test de comparaison des modèles emboîtés permet, comme dans le cas du modèle linéaire (cf section 1.4.2, p. 29), de déterminer si un sous-ensemble des variables explicatives suffit à expliquer la variable Y.

On rappelle que les modèles M_1 et M_0 respectivement définis par $g(\mu)=X_1\theta_1$ et $g(\mu)=X_0\theta_0$ sont dits emboités si le modèle M_0 est un cas particulier du modèle plus général M_1, c'est à dire si le sous-espace engendré par les colonnes de X_0 est inclus dans le sous-espace linéaire engendré par les colonnes de X_1. Le test des hypothèses

$$\mathrm{H}_0=\{M_0\}\ \text{ contre }\ \mathrm{H}_1=\{M_1\},$$

est alors réalisé à l'aide du test du rapport de vraisemblance dont la statistique de test s'écrit :

$$T = -2 \log \frac{\mathcal{L}(y; \hat{\theta}_0)}{\mathcal{L}(y; \hat{\theta}_1)} = -2(\ell(y; \hat{\theta}_0) - \ell(y; \hat{\theta}_1)), \tag{3.4}$$

où $\hat{\theta}_0$ et $\hat{\theta}_1$ sont respectivement les estimateurs du maximum de vraisemblance de θ dans les modèles M_0 et M_1.

Sous certaines hypothèses (voir [49] ou [56]), on peut montrer que cette statistique de test converge en loi vers une loi du χ^2 à $p_1 - p_0$ degrés de liberté, où p_0 et p_1 sont respectivement les dimensions des espaces engendrés par les colonnes de X_0 et X_1. Ainsi si on effectue le test au niveau α, on rejettera H_0 au profit de H_1 si $T \geq \chi^2_{1-\alpha, p_1-p_0}$ où $\chi^2_{1-\alpha,p_1-p_0}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $p_1 - p_0$ degrés de liberté.

Ce même test est parfois présenté sous une forme légèrement différente reposant sur la déviance, qui est l'écart entre le logarithme de la vraisemblance du modèle d'intérêt M et celui du modèle le plus complet possible, appelé *modèle saturé*, noté M_S. Le modèle saturé est le modèle comportant n paramètres, c'est-à-dire autant que d'observations. La déviance du modèle M est alors définie par

$$D(M) = -2 \left(\ell(y; \hat{\theta}) - \ell(y; \hat{\theta}_S) \right).$$

La statistique de test T présentée dans l' équation (3.4) peut être réécrite en terme de déviance sous la forme :

$$T = D(M_0) - D(M_1).$$

Le test global du modèle consiste à tester

$$\mathrm{H}_0 = \{g(\mu_i) = a\} \quad \text{contre} \quad \mathrm{H}_1 = \{g(\mu_i) = x_i\theta\} \tag{3.5}$$

à l'aide du test du rapport de vraisemblance. Il permet de tester si toutes les variables sont inutiles pour expliquer la variable réponse Y.

3.4.2 Tests de $\theta_j = \theta_{0j}$

Si la réponse au test global est positive, la suite logique consiste à tester quels sont les variables ou facteurs qui ont une influence. La connaissance de la loi asymptotique de $\hat{\theta}$ nous permet de construire des tests sur les paramètres θ, sur des combinaisons linéaires de θ ou encore de μ_i ainsi que des intervalles de confiance. Tous ces résultats sont asymptotiques.

On souhaite tester l'hypothèse

$$\mathrm{H}_0 = \{\theta_j = \theta_{0j}\} \quad \text{contre} \quad \mathrm{H}_1 = \{\theta_j \neq \theta_{0j}\} \tag{3.6}$$

où θ_{0j} est une valeur définie a priori. D'après le théorème (3.3.1, p. 119) sous H_0, $T_j = I(\theta, \phi)_{jj} \ (\hat{\theta}_j - \theta_{0j})^2$ converge en loi vers un $\chi^2(1)$ et $P(T_j > t_j)$ donne une p-valeur asymptotique du test. En général on utilise ce test avec $\theta_{0j} = 0$ afin de déterminer si le paramètre θ_j est significativement non nul. Ce test est appelé test de Wald.

En pratique, comme il a été précisé plus haut, l'information de Fisher est calculée non pas en les vrais paramètres qui sont inconnus mais en $\hat{\theta}$ et $\hat{\phi}$. La statistique de test est donc $T_j = I_n(\hat{\theta}, \hat{\phi})_{jj} \ (\hat{\theta}_j - \theta_{0j})^2$.

Un test équivalent, souvent appelé Z-test, est basé sur la statistique de test $Z_j = I(\hat{\theta},\hat{\phi})_{jj}^{1/2}(\hat{\theta}_j - \theta_{0j})$ qui sous l'hypothèse H_0 suit asymptotiquement une loi $\mathcal{N}(0,1)$. La probabilité $P(|Z_j| > |z_j|)$ donne la même p-valeur asymptotique que celle du test de Wald.

3.4.3 Test de $C\theta = 0$

On peut, comme avec le modèle linéaire, être amené à effectuer un test sur une combinaison linéaire des paramètres. Les hypothèses à tester sont

$$\mathrm{H}_0 = \{C\theta = 0\} \quad \text{contre} \quad \mathrm{H}_1 = \{C\theta \neq 0\} \tag{3.7}$$

où C est un vecteur ligne de dimension p (dimension de θ). La construction d'un tel test nécessite de déterminer la loi de $C\hat{\theta}$. Sachant que $\hat{\theta}$ suit asymptotiquement une loi normale, on l'obtient, en utilisant la méthode Delta ([56]) :

$$[C\ I_n(\theta,\phi)^{-1}C']^{-1/2}(C\hat{\theta} - C\theta) \xrightarrow{loi} \mathcal{N}(0, I_p) \tag{3.8}$$

Là encore l'information de Fisher sera évaluée en $\hat{\theta}$ et $\hat{\phi}$.
Le test de région de rejet $\{|T| > u_{1-\alpha/2}\}$ avec

$$T = [C\ I_n(\hat{\theta},\hat{\phi})^{-1}C']^{-1/2}C\hat{\theta},$$

est un test de niveau asymptotique α pour les hypothèses (3.7).
De même, l'intervalle

$$IC_{1-\alpha}(C\theta) = \left[C\hat{\theta} - u_{1-\alpha/2}/\sqrt{C\ I_n(\hat{\theta},\hat{\phi})^{-1}C'}\ ;\ C\hat{\theta} + u_{1-\alpha/2}/\sqrt{C\ I_n(\hat{\theta},\hat{\phi})^{-1}C'}\right]$$

où $u_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$ est un intervalle de confiance de $C\theta$ de niveau asymptotique $1-\alpha$.

3.5 Qualité d'ajustement et choix de modèles

Dans le cas du modèle linéaire, on mesure la qualité d'ajustement du modèle grâce au coefficient de détermination R^2 égal au rapport de la somme des carrés du modèle sur la somme des carrés totale. Dans le cadre du modèle linéaire généralisé, on n'a plus la décomposition de la variabilité en somme des carrés, mais on peut définir des mesures de la qualité d'ajustement du modèle, basées sur le même principe.

3.5.1 Le pseudo R^2

Par analogie avec le coefficient de détermination du modèle R^2 du modèle linéaire, on définit le pseudo-R^2 en associant $D(M_0)$ la déviance du modèle nul (à un seul paramètre) à la somme des carrés totale (SCT) et $D(M_0) - D(M)$ à la somme des carrés du modèle (SCM) :

$$pseudo\, R^2 = \frac{D(M_0) - D(M)}{D(M_0)}$$

Ce pseudo R^2 varie entre 0 et 1, plus il est proche de 1, meilleur est l'ajustement du modèle.

3.5.2 Le χ^2 de Pearson généralisé

Le χ^2 de Pearson généralisé est la statistique définie par

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mathbb{V}(\hat{\mu}_i)}$$

où $\hat{\mu}_i = g^{-1}(x_i\widehat{\theta})$.

Sous l'hypothèse que le modèle étudié est le bon modèle, et si l'approximation asymptotique est valable alors la loi de X^2 peut être approchée par un χ^2 à $n-p$ degrés de liberté (n étant le nombre de données, p le rang de la matrice de design X). On rejettera au niveau α le modèle étudié si la valeur observée de X^2 est supérieure au quantile $\chi^2_{n-p,1-\alpha}$.

3.5.3 Choix de modèle

Quand deux modèles sont emboîtés, le test des modèles emboîtés 3.4.1, p. 120 permet de choisir entre eux.
En présence de plusieurs modèles candidats, non emboîtés, un premier critère de sélection est donné par la déviance. Le modèle qui a la plus mauvaise déviance (la plus forte) est le modèle nul, qui a un seul paramètre. Ce modèle n'a en général aucune utilité car il n'explique rien. Le modèle saturé qui a autant de paramètres que de données possède par définition la meilleure déviance puisqu'elle vaut 0. Ce modèle n'est souvent pas pertinent car il a trop de paramètres. Les déviances de ces deux modèles fournissent les valeurs du pire et du meilleur ajustement possible. Un modèle sera qualifié de bon si sa déviance est proche de celle du modèle saturé (ce qui est équivalent à un pseudo R^2 proche de 1) et s'il est construit avec un faible nombre de paramètres.
Des critères pénalisés permettent de prendre en compte ces deux contraintes antagonistes.

Le plus célèbre d'entre eux est le critère AIC (Akaike Information Criterion) dont une définition est

$$AIC(M(\hat{\theta})) = -2\ell(y;\hat{\theta}) + 2p$$

où p est le rang de la matrice de design X. L'AIC est d'autant plus faible que la log-vraisemblance est élevée et que le nombre de paramètres est petit et permet donc d'établir un ordre sur les modèles en prenant en compte les deux contraintes. Le critère BIC (Bayesian Information Criterion) qui pénalise plus le sur-ajustement est défini par

$$BIC(M(\hat{\theta})) = -2\ell(y;\hat{\theta}) + np$$

3.5.4 Diagnostic, résidus

Les résidus dans le modèle linéaire généralisé

Comme dans le modèle linéaire général, plusieurs types de résidus peuvent être définis. La définition la plus naturelle, consiste à quantifier l'écart entre l'observation y_i et sa prédiction par le modèle $\hat{\mu}_i$. On définit ainsi les *résidus bruts* $\varepsilon_i = y_i - \hat{\mu}_i$, l'inconvénient de ce type de résidus est qu'ils n'ont pas toujours la même variance, il

est donc difficile de les comparer à un comportement type attendu. En normalisant les résidus bruts par une variance estimée, on obtient les *résidus de Pearson* :

$$r_{p_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathbb{V}_{\hat{\theta}}(y_i)}},$$

où $\mathbb{V}_{\hat{\theta}}(y_i)$ désigne la variance théorique de y_i calculée en $\hat{\theta}$. Si Y_i suit une loi de Bernoulli par exemple, on a $\mathbb{V}_{\hat{\theta}}(y_i) = \hat{p}_i(1-\hat{p}_i)$.

Ces résidus de Pearson, ont le même défaut que les résidus similaires du modèle linéaire, leur variance dépend de l'influence de l'observation i. Les *résidus de Pearson standardisés* sont obtenus en renormalisant par l'effet levier :

$$r_{s_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1-h_{ii})\mathbb{V}_{\hat{\theta}}(y_i)}},$$

où h_{ii} désigne le levier, c'est à dire le terme diagonal de la matrice $H = X\,(X'X)^{-1}X'$ dans le cas où la matrice de design X est de rang plein.

Une approche un peu différente consiste à définir les *résidus de déviance*, qui mesurent à quel point la log-vraisemblance pour l'observation i est loin de la log-vraisemblance pour cette même observation dans le cas du modèle saturé. Les résidus de déviance sont définis par

$$r_{d_i} = signe(y_i - \hat{\mu}_i)\sqrt{2\ell(y_i;\hat{\theta}_S,\hat{\phi}) - 2\ell(y_i;\hat{\theta},\hat{\phi})}.$$

où $\ell(y_i;\hat{\theta}_S,\hat{\phi}) - 2\ell(y_i;\hat{\theta},\hat{\phi})$ est la contribution de l'observation i à la déviance totale du modèle. Pour rendre ces résidus comparables entre eux, il faut les corriger pour prendre en compte l'influence de chaque observation, les *résidus de déviance standardisés* sont définis par :

$$r_{ds_i} = signe(y_i - \hat{\mu}_i)\sqrt{\frac{2\ell(y_i;\hat{\theta}_S,\hat{\phi}) - 2\ell(y_i;\hat{\theta},\hat{\phi})}{1-h_{ii}}}.$$

Intuitivement, une observation ayant un résidu de déviance élevé est une observation ayant une grande influence sur l'estimation des paramètres du modèle et doit donc être examinée avec soin.

Dans les deux cas on vérifiera comme pour le modèle linéaire, qu'il n'existe pas de structure inattendue dans les résidus, en moyenne ou en variance. La présence d'une telle structure devrait porter le modélisateur à reprendre le modèle proposé pour identifier la cause de cette structure, par exemple un effet quadratique d'une variable. On peut montrer que les résidus sont asymptotiquement gaussiens si le modèle est adéquat, et cette hypothèse peut être vérifiée à l'aide d'un q-q plot si le nombre de données n est assez grand.

3.6 Régression logistique

Ce modèle est un cas particulier important du modèle linéaire généralisé, avec certaines définitions et outils spécifiques. On étudie ces spécificités dans cette partie.

3.6.1 Régression logistique, cas où la réponse est binaire

Nous nous intéressons maintenant au cas où la variable à expliquer Y est une variable binaire, prenant deux modalités : présence/absence, sain/malade, mort/vivant, en-dessous/au-dessus d'un seuil, en fonction de variables explicatives X quantitatives ou qualitatives. Pour chaque individu i la variable Y_i suit une loi de Bernoulli de paramètre p_i qui est aussi son espérance. Le modèle s'écrit

$$Y_i \sim \mathcal{B}ernoulli(p_i)$$

$$g(\mathbb{E}(Y_i)) = g(p_i) = x_i\theta.$$

On a vu que la fonction de lien naturel est la fonction logit définie par $g(p) = \log(\frac{p}{1-p})$. Lorsque c'est cette fonction de lien qui est choisie, on appelle *régression logistique* le modèle linéaire généralisé associé. Cette dénomination est aussi utilisée dans le cas où une ou plusieurs variables explicatives sont qualitatives alors qu'en toute rigueur on a affaire à une analyse de la variance ou de la covariance plutôt qu'à une régression.
La fonction logit est une bijection de l'intervalle $]0;1[$ dans $\mathbb{R}$, et son graphe est symétrique par rapport au point $(0.5, 0)$.
La probabilité $P(Y = 1|x) = p(x)$, pour un individu pour lequel les variables expli-

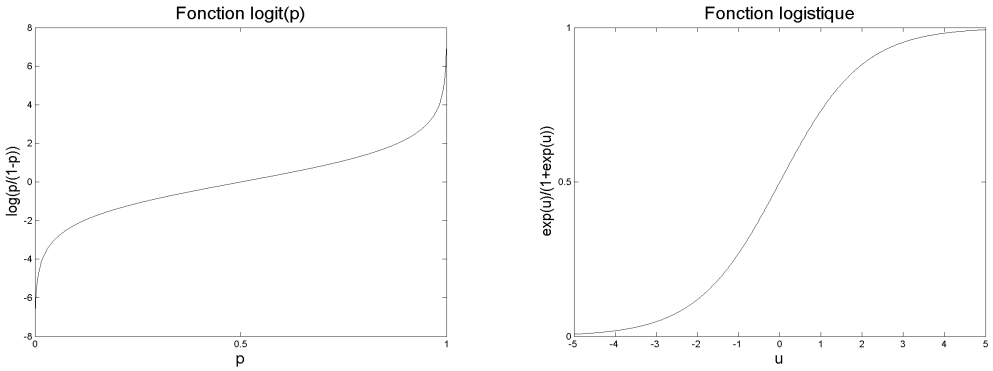


FIGURE 3.1 – Fonctions logit et logistique

catives forment le vecteur ligne x, est $h(x\theta)$ où h est la fonction logistique, la fonction inverse de la fonction logit : $h(u) = \frac{1}{1+e^{-u}}$. Le modèle de régression logistique peut donc s'écrire sous la forme :

$$p(x) = \frac{1}{1 + e^{-x\theta}}. \tag{3.9}$$

Les graphes des fonctions g et h sont donnés dans la figure 3.6.1.

Lorsque toutes les variables explicatives sont qualitatives, les individus présentant les mêmes combinaisons de modalités sont regroupés pour former une seule observation, et la variable réponse Y_i associée à ce groupe (nombre de succès pour cette combinaison de modalités) suit une loi binomiale de paramètres $(n_i,\ p_i)$. Lorsque les variables explicatives sont continues, les valeurs de x_i sont différentes d'un individu à l'autre et aucun regroupement n'est possible.
On a donc deux types de modèles de régression logistique. Dans le premier cas on a des facteurs qualitatifs comme en analyse de variance, et dans le deuxième cas on a un

modèle de type régression. Si les modalités des variables qualitatives sont ordonnées et qu'une relation linéaire entre les variables qualitatives et $log(p/(1-p))$ est plausible, on peut utiliser le modèle de régression comme pour une variable continue. Si ce n'est pas le cas il faut utiliser un modèle du type analyse de variance (avec des effets et éventuellement des interactions) qui modélise plus finement la relation de x avec $log(p/(1-p))$, mais utilise davantage de paramètres dans la modélisation.

Tests exacts

Lorsque le nombre d'observations est faible, on dispose pour la régression logistique d'une méthode exacte pour l'estimation des paramètres et les tests d'hypothèse, qui donne de bien meilleurs résultats que ceux présentés en section 3.4, p. 120 valides asymptotiquement, c'est à dire en pratique quand le nombre d'observations est supérieur à (de l'ordre de) 10 fois le nombre de paramètres. Cette méthode est basée sur l'énumération complète de la distribution de la statistique exhaustive pour un paramètre du modèle, conditionnellement aux estimations des autres paramètres, [33].

Test d'ajustement, test de Hosmer-Lemeshow

Dans le cas où les variables explicatives sont qualitatives, on utilise le test du χ^2 de Pearson défini en section 3.5.2, p. 123 ou le test du rapport de vraisemblance pour tester l'ajustement du modèle aux données. Dans le cas de variables explicatives continues, on ne peut pas l'utiliser parce que les conditions de validité de ces tests ne sont pas remplies. On utilise un autre test dit de Hosmer-Lemeshow, défini de la façon suivante : on ordonne les valeurs de $\hat{p}_i$ par ordre croissant et on considère les classes d'individus définies par les déciles de la distribution des $\hat{p}_i$. Pour chaque classe k on calcule N_k le nombre d'observations, A_k le nombre de cas $Y = 1$ et P_k la moyenne des $\hat{p}_i$ dans la classe k. Le test est basé sur la statistique

$$S = \sum_{k=1}^{10} \frac{(A_k - N_k P_k)^2}{N_k P_k (1 - P_k)} \tag{3.10}$$

Sous l'hypothèse H_0 que le modèle est le vrai modèle, $S \sim \chi^2_8$. Le choix de 10 classes est un compromis entre puissance et validité des conditions asymptotiques. On peut choisir moins de classes si on a peu de données ou plus de classes si on a beaucoup de données. Le nombre de degrés de liberté du χ^2 est égal au nombre de groupes moins 2.

Cote, rapport de cotes et risque relatif

Des outils spécifiques à la régression logistique ont été développés pour l'interprétation des coefficients et la validation du modèle.
La quantité $\frac{p(x)}{1-p(x)}$ est appelée "odds" en anglais et "cote" en français : la phrase "le cheval Toto a une cote de 10 contre 1" signifie qu'il a 10 chances de perdre et une chance de gagner, et le joueur gagne 10 fois sa mise s'il le choisit et que Toto gagne. Le gain net (déduction faite de la mise) du pari est alors égal au produit de la mise et de la cote. [1]

1. Il y a 2 définitions de la cote : celle qui est donnée ici est la "cote anglaise". La "cote française" (et européenne) est égale à la cote anglaise plus 1. Le produit de la cote française et de la mise donne la somme reçue après la course.

Lorsque cette quantité est proche de 1 les deux évènements $(Y = 0)$ et $(Y = 1)$ ont des chances équivalentes de se produire. Lorsqu'elle est proche de 0 c'est l'évènement $(Y = 0)$ qui est le plus probable, tandis que l'évènement $(Y = 1)$ est plus probable lorsque l'odds est grand. Dans le cadre du modèle logistique on a $odds(x) = e^{x\theta}$.
L'odd-ratio ou rapport de cotes est le rapport de deux odds associés à deux valeurs différentes des variables : $OR(x,t) = \frac{p(x)(1-p(t))}{p(t)(1-p(x))}$.
Dans le cas où la variable x^j est quantitative, on obtient en posant $t^j = x^j + 1$ et en laissant inchangées les autres variables, $OR(x,t) = e^{\theta_j}$, ce qui permet d'interpréter les paramètres associés aux variables quantitatives comme le logarithme du rapport de cote lorsque la variable x^j augmente d'une unité. Dans le cas où la variable est qualitative l'odd-ratio permet de comparer les odds entre deux modalités de cette variable.
Le Risque Relatif ou Rapport de Risques entre 2 conditions x et t est $RR(x,t) = \frac{p(x)}{p(t)}$. Il est utilisé en épidémiologie pour comparer les risques de maladie entre 2 conditions, par exemple fumeur et non fumeur ou pour une augmentation de 1 de la variable si elle est continue. OR et RR ne sont pas identiques. Il faut les distinguer d'autant plus que la régression logistique donne directement OR comme sous-produit et qu'il est plus difficile d'interpréter OR que RR dont l'interprétation est naturelle. La facilité conduit souvent à interpréter OR en termes de RR. La relation entre odd-ratio et risque relatif est la suivante : $OR(x,t) = RR(x,t) \times \frac{1-p(t)}{1-p(x)}$. En général p est la probabilité de maladie, donc p est faible et $(1-p) \simeq 1$ dans les 2 conditions. Dans ce cas $OR \simeq RR$, ce qui autorise à interpréter OR en termes de RR en première approximation. Cette approximation est d'autant plus mauvaise que $1 - p(x)$ et $1 - p(t)$ sont différents.

Classement, courbe ROC

On peut utiliser les prédictions fournies par le modèle, $\widehat{p}_i$, pour en déduire des valeurs prédites pour $\widehat{Y}_i$: on pose $\widehat{Y}_i = 0$ si $\widehat{p}_i < s$ et $\widehat{Y}_i = 0$ sinon, ce qui fait de la régression logistique un classifieur. La règle de Bayes consiste à choisir $s = 0.5$, mais une autre valeur peut être mieux appropriée selon le cas.
On rappelle que $p(x) = P(Y = 1|x)$. Si on prédit $Y = 1$ alors que $Y = 0$ on dit que l'on a un "faux positif". Si on prédit $Y = 0$ alors que $Y = 1$ on dit que l'on a un "faux négatif". On juge de la qualité des prédictions par la table de confusion

	$\#(\widehat{Y}_i = 0)$	$\#(\widehat{Y}_i = 1)$
$\#(Y_i = 0)$	a	b
$\#(Y_i = 1)$	c	d

a est le nombre de "vrais négatifs", d est le nombre de "vrais positifs", c est le nombre de "faux négatifs" et b est le nombre de "faux positifs". On a de bonnes prédictions si les valeurs b et c sont petites devant a et d.
Il existe un grand nombre d'indicateurs synthétiques mesurant la qualité des prédictions à partir de la table de confusion : $\frac{b+c}{a+d}$, $\frac{b+c}{a+b+d+c}$, ...
Un autre moyen très populaire en médecine et économie pour évaluer la qualité des prédictions est la courbe ROC (Receiver Operating Characteristic), terme issu de la théorie du signal. Supposons que les prédictions $\widehat{Y}_i$ soient obtenues en utilisant un seuil s variant de 0 à 1. Pour chaque valeur de s, on a une table de confusion qui donne un taux de faux positifs $\tau_-(s) = \frac{b(s)}{a(s)+b(s)}$ et un taux de vrais positifs $\tau_+(s) = \frac{d(s)}{c(s)+d(s)}$. La courbe ROC est la courbe qui relie tous les points $(\tau_-(s), \tau_+(s))$. La figure 3.2 en donne

un exemple. Dans le meilleur des cas il existe un seuil, s_0 qui sépare parfaitement les positifs et les négatifs et $\tau_-(s_0) = 0$ et $\tau_+(s_0) = 1$. Dans ce cas la courbe ROC passe par le point (0,1).
Dans le pire des cas les positifs et les négatifs sont complètement mélangés quel que soit le seuil, et les deux taux sont égaux. Dans ce cas la courbe ROC est proche de la bissectrice du carré. La quantité AUC (Aera Under the Curve) qui est égale à l'aire comprise sous la courbe ROC est un bon indicateur de la qualité des prédictions : si elle est proche de 1, les prédictions sont bonnes, si elle est proche de 0.5 les prédictions sont mauvaises.

Comme pour toute évaluation de la qualité d'une prédiction, il faut séparer le processus d'estimation des paramètres du modèle du processus d'estimation de la qualité de la prédiction. Pour ce faire on peut séparer les données en un échantillon d'apprentissage et un échantillon test, utiliser la validation croisée ou le re-échantillonnage.

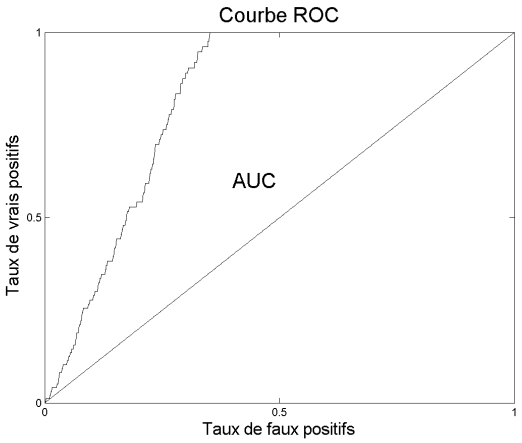


FIGURE 3.2 – Courbe ROC

3.6.2 Régression multilogistique

Considérons maintenant une variable aléatoire Y avec $K > 2$ modalités. Notons $(Y = k)$ la variable qui vaut 1 si l'évènement $(Y = k)$ est vrai et 0 sinon. La loi conjointe des K variables $((Y = 1), \ldots, (Y = K))$ est une loi multinomiale de paramètre $p = (p_1, \ldots, p_K)$ où $p_k = P[Y = k]$:

$$((Y = 1), \ldots, (Y = K)) \sim \mathcal{M}(1, p_1, p_2, \ldots, p_K),$$

tel que $\sum_{k=1}^{K} p_k = 1$. Par abus de notation, on notera Y le vecteur $((Y = 1), \ldots, (Y = K))$, la loi de probabilité s'écrit alors :

$$P[Y = y; p] = \prod_k {p_k}^{y_k} = \exp\left(\sum_k y_k \log p_k\right).$$

On pourrait alors prendre comme fonction de lien, la fonction log et donc poser $\log p_k = x\theta^k$ pour chaque $k = 1, \ldots, K$. Cela reviendrait à traiter les variables $(Y = k)$ de façon

indépendante. Or ces variables, étant liées par le fait que $\sum_k (Y = k) = 1$, ne sont pas indépendantes.

Comme il n'y a que $K-1$ composantes indépendantes, on se ramène à étudier $K-1$ composantes choisies de façon arbitraire en choisissant une modalité de référence, par exemple la première.

On a :

$$\begin{aligned} P[Y = y; p] &= \exp\left(y_1 \log p_1 + y_2 \log p_2 + \ldots y_K \log p_K\right) \\ &= \exp\left(\log p_1 + y_2 \log p_2/p_1 + \ldots y_K \log p_K/p_1\right), \end{aligned}$$

et on suppose que les "odds" sont linéaires en les variables explicatives, en notant $\theta^{(k/1)}$ la reparamétrisation on a :

$$\begin{aligned} \log p_2/p_1 &= x\theta^{(2/1)} \\ \log p_3/p_1 &= x\theta^{(3/1)} \\ \vdots &= \vdots \\ \log p_K/p_1 &= x\theta^{(K/1)} \end{aligned}$$

p_1 est alors déduit de la contrainte $\sum_{k=1}^{K} p_k = 1$ ($p_1 = 1 - \sum_{k>1} p_k$). Ce modèle est appelé le modèle logistique multinomial. Il faut noter que le choix de la référence est important, puisque elle conditionnera l'interprétation des résultats.
On obtient l'expression des différentes probabilités p_k en fonction de x :

$$\begin{aligned} p_1 &= \frac{1}{1 + \sum_{l>1} e^{x\theta^{(l/1)}}} \\ \text{et} & \\ p_k &= \frac{1}{1 + e^{-x\theta^{(k/1)}} + \sum_{l>1, l\neq k} e^{x(\theta^{(l/1)} - \theta^{(k/1)})}}, k = 2, K. \end{aligned}$$

3.6.3 Surdispersion

Les données sont souvent plus variables que ce qui est attendu avec le modèle probabiliste utilisé. Cela peut provenir de différentes causes. Dans le cas de variables de comptage il arrive fréquemment que la distribution des cas soit non aléatoire. Par exemple la répartition des plantes malades dans un champ n'est en général pas uniforme dans le champ mais plutôt aggrégative (les plantes malades sont regroupées). Si on échantillonne plusieurs parcelles de même surface on obtient alors une plus grande variabilité que celle attendue de par la loi de Poisson. On dit alors qu'il y a surdispersion dans les données par rapport au modèle. On a 2 possibilités pour modéliser cette surdispersion

1. on utilise un modèle standard (Poisson, Binomial) avec un paramètre de surdispersion, noté ϕ. Les lois de Poisson et Binomiale avec surdispersion n'existent pas en tant que lois de probabilité, mais on montre ([56] [1]) qu'on peut utiliser une méthode dite de quasi-vraisemblance pour estimer les paramètres θ et ϕ. De plus l'algorithme de Newton-Raphson suffit pour maximiser la quasi-vraisemblance. C'est une solution simple et économique.

2. on utilise un modèle plus complexe qui prend en compte le phénomène de surdispersion. Par exemple on remplace la loi Binomiale par une loi BetaBinomiale et on remplace la loi de Poisson par la loi Binomiale Négative. Cette solution d'une bonne modélisation se heurte au fait que ces nouvelles lois ne font souvent pas partie de la famille exponentielle et qu'on perd alors une partie des bonnes propriétés statistiques du modèle linéaire généralisé standard.

Détection de la surdispersion

Pour détecter la surdispersion il suffit de calculer le rapport entre le χ^2 d'ajustement du modèle aux données défini dans la section 3.5.2, p. 123, et son nombre de degrés de liberté. S'il est nettement supérieur à 1, il y a surdispersion.

3.7 Preuve de la propriété 3.2.1, p. 116

Dans cette annexe, nous démontrons les expressions de l'espérance et de la variance d'une variable aléatoire Y dont la densité de probabilité f_Y appartient à la famille exponentielle naturelle (son expression est donnée par l'équation (3.1, p. 116)). Notons $\mathcal{Y}$ le support de f_Y, f_Y étant une densité de probabilité, elle vérifie :

$$\int_{\mathcal{Y}} f_Y(y;\omega,\phi)dy = 1$$

Pour alléger l'écriture, nous omettons d'écrire le paramètre ϕ dans $f_Y(y;\omega,\phi)$. Sous les conditions de dérivabilité sous le signe somme (la dérivée existe et est dominée uniformément par une fonction intégrable), on a

— en dérivant par rapport à ω,

$$\begin{aligned}
\frac{\partial}{\partial\omega}\int_{\mathcal{Y}} f_Y(y;\omega)dy &= \int_{\mathcal{Y}} \frac{\partial}{\partial\omega} f_Y(y;\omega)dy \\
0 &= \int_{\mathcal{Y}} \frac{1}{f_Y(y;\omega)}\frac{\partial}{\partial\omega} f_Y(y;\omega)\ f_Y(y;\omega)dy \\
0 &= \int_{\mathcal{Y}} \frac{\partial}{\partial\omega}\log f_Y(y;\omega)\ f_Y(y;\omega)dy \\
0 &= \int_{\mathcal{Y}} \frac{1}{\gamma(\phi)}[y - b'(\omega)]\ f_Y(y;\omega)dy \\
0 &= \frac{1}{\gamma(\phi)}(\mathbb{E}[Y] - b'(\omega))
\end{aligned}$$

ce qui donne $\mathbb{E}[Y] = b'(\omega)$,

— en dérivant une seconde fois, on obtient facilement que :

$$\int_{\mathcal{Y}} \frac{\partial^2}{\partial^2\omega}\log f_Y(y;\omega)\ f_Y(y;\omega)dy + \int_{\mathcal{Y}} \frac{\partial}{\partial\omega}\left(\log f_Y(y;\omega)\right)^2\ f_Y(y;\omega)dy = 0$$

Pour obtenir la variance, calculons les deux termes du membre de gauche de l'égalité,

notés respectivement A et B. Le premier terme vaut

$$\begin{aligned} A &= \frac{1}{\gamma(\phi)} \int_{\mathcal{Y}} \frac{\partial}{\partial \omega} [y - b'(\omega)] \; f_Y(y; \omega) dy \\ &= -\frac{1}{\gamma(\phi)} \int_{\mathcal{Y}} b''(\omega) \; f_Y(y; \omega) dy \end{aligned}$$

Le second terme vaut

$$\begin{aligned} B &= \frac{1}{\gamma(\phi)^2} \int_{\mathcal{Y}} [y - b'(\omega)]^2 \; f_Y(y; w) dy \\ &= \frac{1}{\gamma(\phi)^2} \mathbb{V}[Y], \end{aligned}$$

d'après l'expression de $\mathbb{E}[Y] = b'(w)$. La preuve se termine en annulant la somme de A et B.

Chapitre 4

Exemples de modèles linéaires généralisés

4.1 Loi de Bernoulli : pollution par l'ozone

4.1.1 Contexte et données

L'ozone est un polluant secondaire formé à partir des émissions de monoxyde (NO) et dioxyde (NO_2) d'azote issues du trafic routier, en présence de fort rayonnement. Une concentration élevée en ozone peut provoquer des insuffisances respiratoires, des céphalées et autres pathologies. Les agences de qualité de l'air régionales ont parmi d'autres missions, celle de surveiller la concentration des différents polluants et de prévenir la population à risque en cas de concentration trop élevée. Pour l'ozone la procédure suivante a été retenue en Ile de France : si deux stations dépassent simultanément un seuil $s1$, l'alerte 1 est donnée, si deux stations dépassent simultanément un seuil $s2$, l'alerte 2 est donnée. Les mesures accompagnant l'alerte 2 sont d'ordre préventif : circulation alternée et gratuité des transports en commun. La concentration doit donc être anticipée avec plusieurs heures d'avance. La formation d'ozone étant très liée aux conditions météorologiques deux variables climatiques sont pressenties pour prévoir le déclenchement d'une alerte de niveau 2 : la température maximale du jour qui est un bon indicateur pour le rayonnement difficile à mesurer et à prévoir, et la vitesse moyenne du vent au moment où la température est maximale (l'après-midi) qui est un bon indicateur de la dispersion atmosphérique.

On dispose d'une base de données contenant les concentrations maximales mesurées chaque jour en plusieurs stations ainsi que la température maximale et la vitesse moyenne du vent pour les mêmes dates sur une période de plusieurs années. Après avoir éliminé des dates pour lesquelles on a trop de données manquantes on constitue une variable binaire Y qui est égale à 1 si au moins deux stations dépassent le seuil $180\mu/\mathrm{m}^3$ qui représente le seuil d'alerte de niveau 2, et 0 sinon. Le graphique 4.1 montre le lien entre les deux variables météorologiques et la variable indicatrice de l'alerte. Les jours où l'alerte doit être déclenchée la température est élevée et la vitesse du vent est faible : le rayonnement favorable à la formation de l'ozone a lieu lorsque la température est forte, et l'absence de vent empêche la dispersion du polluant. On observe néanmoins des températures élevées ou des vitesses de vent faibles des jours pour lesquels il n'y a pas eu d'alerte et inversement des jours où l'alerte a eu lieu alors que la température

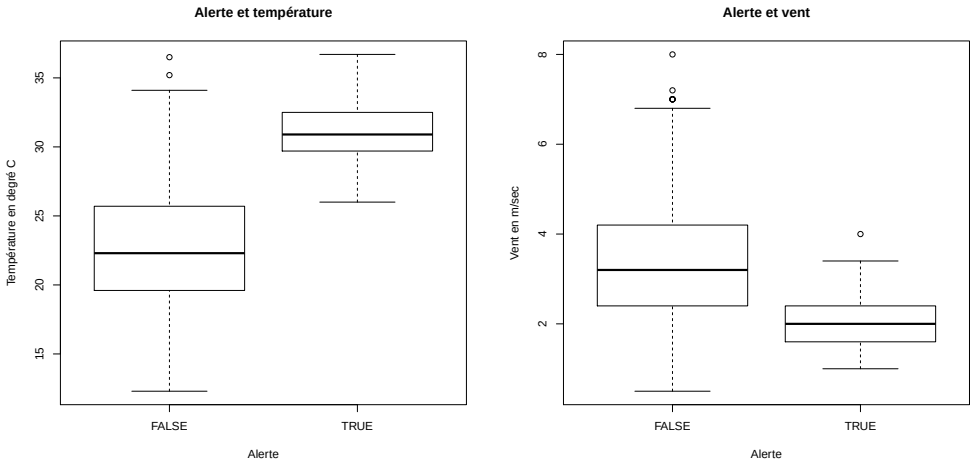


FIGURE 4.1 – Boxplot de la température et du vent sur le déclenchement de l'alerte, FALSE : pas d'alerte, TRUE : alerte. Source : les données de pollution sont disponibles sur le site d'Airparif (`http://www.airparif.asso.fr/`), les données météo ont été fournies gracieusement par Météo-France.

était peu élevée et la vitesse du vent modérée.

4.1.2 Résultats

Soit Y_i la variable aléatoire qui vaut 1 si l'alerte 2 est déclenchée le jour i, 0 sinon. Si on suppose que conditionnellement aux variables météorologiques les déclenchements d'alerte sont indépendants alors il est naturel de modéliser Y_i par une loi de Bernoulli de paramètre p_i.

$$Y_i \overset{ind.}{\sim} Bernoulli(p_i)$$

Les paramètres p_i, dépendent des variables météorologiques. La fonction de lien naturel pour la loi de Bernoulli est la fonction logit. On a alors

$$logit(p_i) = \mu + \alpha Temp_i + \beta Vent_i$$

La table 4.1 donne les estimations des paramètres du modèle, avec les tests de Wald associés. Les paramètres sont significativement non nuls. Comme attendu, le coefficient

	Estimate	Std. Error	z value	Pr(> | z |)
(Intercept)	-16.18890	1.94668	-8.316	< 2e-16 ***
Temperature	0.60921	0.06859	8.881	< 2e-16 ***
Vent	-1.31858	0.23585	-5.591	2.26e-08 ***

TABLE 4.1 – Estimations et tests des coefficients du modèle sur la pollution par l'ozone

de la variable de température est positif, alors que celui de la variable de vitesse de vent est négatif.
La déviance du modèle nul est de 507.51, la déviance résiduelle est de 226.51, l'AIC vaut 232.51. La table d'analyse de la déviance 4.2 donne la décomposition de la déviance.

	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			1035	507.14	
Temperature	1	235.391	1034	271.75	< 2.2e-16 ***
Vent	1	45.243	1033	226.51	1.741e-11 ***

TABLE 4.2 – Décomposition de la déviance pour le modèle sur la pollution

L'introduction de la variable de température fait baisser la déviance de 235.391, alors que l'introduction de la variable de vitesse de vent la fait baisser de 45.243. L'AIC du modèle avec seulement la variable de température est égal à 275.75. On en déduit que les deux variables ont une influence sur le déclenchement de l'alerte.

Le graphique 4.2 donne les résidus $y_i - \widehat{\mu}_i$ en fonction des valeurs ajustées $\widehat{\omega}_i$ du paramètre naturel. La droite $\omega = 0$ sépare les prédiction des jours pollués et non pollués lorsqu'on applique la règle de Bayes : le jour i est prédit pollué si $\widehat{p}_i > 0.5$ ou encore $\widehat{\omega}_i > 0$.

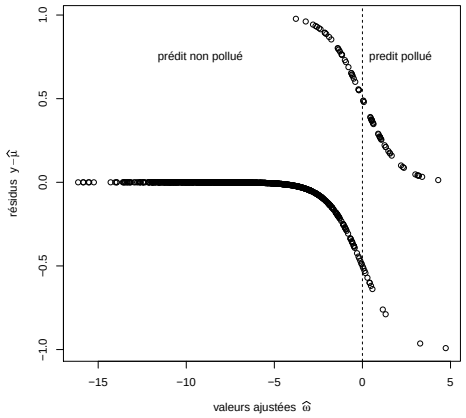


FIGURE 4.2 – Graphe des résidus pour le modèle sur la pollution

La table de confusion associée à la règle de Bayes est :

	$\#(\widehat{Y}_i = 0)$	$\#(\widehat{Y}_i = 1)$
$\#(Y_i = 0)$	954	13
$\#(Y_i = 1)$	32	37

Plus de la moitié des situations d'alerte sont bien prédites mais il y en a presque autant qui ont été manquées. Par contre la très grande majorité des situations de non alerte ont été correctement identifiées. La Figure 4.3 montre l'influence des deux variables météorologiques sur la probabilité d'alerte : La température a un effet croissant sur la probabilité d'alerte alors que la vitesse du vent a un effet décroissant.

Le graphique 4.4 montre les différentes situations dans le plan des variables météorologiques.

On distingue clairement sur ce graphique les situations à problème :

— les triangles noirs représentent les jours à forte température et vent faible et pourtant non pollués ; ce peut être des situations où le rayonnement est faible

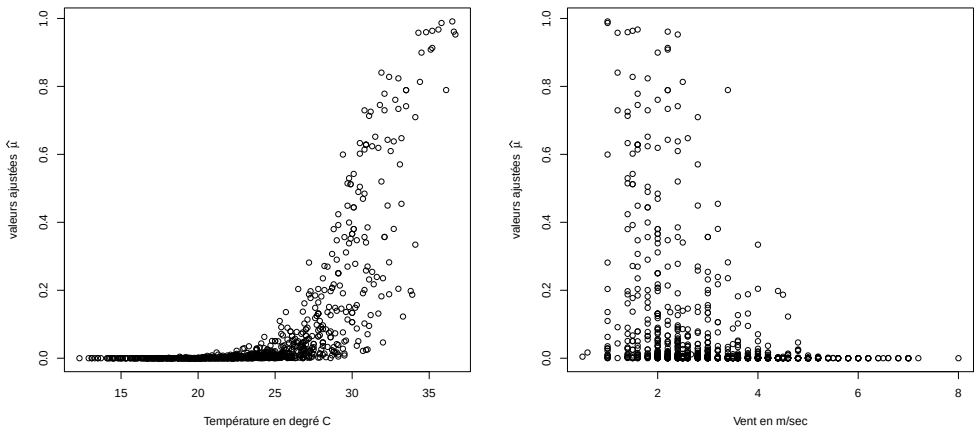


FIGURE 4.3 – Influence de la température et du vent sur la pollution à l'ozone

FIGURE 4.4 – Alertes réalisées et manquées en fonction de la température et du vent pour la pollution à l'ozone

malgré une forte température (couverture nuageuse qui empêche le rayonnement par exemple).

— les disques noirs représentent les situations pour lesquelles le modèle manque une alerte, ce qui arrive lorsque la température est faible ou le vent est élevé et pourtant la concentration en ozone est forte.

4.1.3 Programme R

```
# Pollution.R : regression logistique sur le déclenchement
de la procédure d'alerte
#=================================================================

pollclimat = read.table("ozone.txt",header=TRUE)

# suppression des données manquantes
#-----------------------------------
ete = NULL
for (ian in 1:10){
  ete = c(ete, ((ian-1)*365+152):((ian-1)*365+273))
}

indnapol = which(apply(pollclimat[,1:7],1,function(x)
           sum(is.na(x)))>3)
indnaclim = which(apply(pollclimat[,8:9],1,function(x)
            sum(is.na(x)))>0)
indnan = union(indnapol,indnaclim)
indconserv=setdiff(1:dim(pollclimat)[1],indnan)
```

```
indconservete = intersect(indconserv,ete)

# variable binaire : 1 si deux stations depassent le seuil
#-----------------
depasse = function(x,seuil) y = sum(x>seuil,na.rm=TRUE)>1
seuil = 150
Y = apply(pollclimat[indconservete,1:7],1,depasse,seuil)

alerte=data.frame(Y=Y,Temperature=
                  pollclimat$Temperature[indconservete],
                  Vent = pollclimat$Vent[indconservete])

# etude descriptive
#------------------

boxplot(Temperature~Y,data=alerte,
    main="Alerte et température",ylab="Température en degré C",
    xlab="Alerte")
boxplot(Vent~Y,main="Alerte et vent",data=alerte,
         ylab = "Vent en m/sec",xlab="Alerte")

# Modele avec temperature et vent du jour
#----------------------------------------
glm1 = glm(Y ~ Temperature+Vent,family=binomial,data=alerte)
summary(glm1)
par(mfrow=c(2,2))
plot(glm1)
par(mfrow=c(1,1))
anova(glm1,test="Chisq")
Anova(glm1)
obs = rep(0,dim(alerte)[1])
obs[Y]=1
plot(glm1$fitted,obs-glm1$fitted)
pred = predict(glm1)
plot(pred,obs-glm1$fitted,xlab="valeurs ajustées etachap",
ylab="résidus obs-muchap")
abline(v=0,lty=2)
text(0,0.75,"predit pollué",pos=4)
text(-12,0.75,"prédit non pollué",pos=4)
plot(alerte$Temperature,glm1$fitted,ylab="valeurs ajustées
       muchap", xlab="Température en degré C")
plot(alerte$Vent,glm1$fitted,ylab="valeurs ajustées muchap",
          xlab = "Vent en m/sec")

# tableau d'alerte
#-----------------
probaseuil=0.5
```

```
Ychap = glm1$fitted>0.5
VP = sum(Y & Ychap)
FP = sum(!Y & Ychap)
VN = sum(!Y & !Ychap)
FN = sum(Y & !Ychap)

print(c(VP,FP,VN,FN))
couleur = rep(1,length(Y))
couleur[!Y & Ychap] = 2
couleur[!Y & !Ychap] = 3
couleur[Y & !Ychap] = 4
plot(alerte$Temperature,alerte$Vent,col=couleur,pch=19,
     main="Alertes en fonction de la température et du vent",
     xlab = "Température en degré C",ylab = "Vent en m/sec")
legend("topright",legend=c("Alerte réelle","Fausse alerte",
       "Pas d'alerte","Alerte manquée"),col=1:4,pch=19,cex=0.8)
couleur = rep(1,length(Y))
couleur[glm1$fitted>0.5] = 2
plot(alerte$Temperature,alerte$Vent,col=couleur)
```

4.2 Loi binomiale : équité sociale

4.2.1 Contexte et données

La question s'est posée de savoir si l'accès aux grandes écoles était socialement équitable. Pour tenter de répondre à cette question on a examiné les résultats au concours "A-Bio" (à l'issue de classes préparatoires BCPST) de l'ensemble des candidats l'année 2009. L'information sur l'origine sociale des candidats est donnée par le critère de bourse. On dispose également de l'information sur le genre (féminin, masculin) des candidats.

Le tableau 4.3 donne le nombre de candidats admis et non admis au concours en fonction du genre et de l'attribution ou non d'une bourse.

Genre	Statut	Admis	Non Admis
Féminin	Boursier	111	214
Féminin	Non Boursier	557	625
Masculin	Boursier	61	76
Masculin	Non Boursier	214	256

TABLE 4.3 – Admis et non admis au concours en fonction du Genre et du Statut (source : Service des Concours Agronomiques et Vétérinaires)

4.2.2 Résultats

La variable aléatoire Y_{ij} : nombre de candidats admis dans le groupe du Genre i et du Statut j d'effectif n_{ij}, suit une loi Binomiale $\mathcal{B}(n_{ij}, p_{ij})$. Les paramètres p_{ij}, dépendent du Genre, du Statut et éventuellement de l'interaction entre les deux. En utilisant la

fonction logit qui est la fonction de lien naturel pour la loi Binomiale on a alors :

$$Y_{ij} \overset{ind.}{\sim} \mathcal{B}(n_{ij}, p_{ij})$$

$$logit(p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

La table 4.4 donne l'estimation des paramètres du modèle par maximum de vraisemblance, avec les tests de Wald (cf 3.6, p. 121) associés à chaque paramètre. La déviance

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6564	0.1170	-5.612	2.00e-08 ***
GenreM	0.4366	0.2079	2.100	0.0358 *
StatutNB	0.5413	0.1307	4.142	3.44e-05 ***
GenreM :StatutNB	-0.5006	0.2350	-2.131	0.0331 *

TABLE 4.4 – Estimation et tests sur le paramètres du modèle sur l'équité au concours

du modèle nul est de 17.893. La déviance résiduelle est égale à 0. Le modèle a autant de paramètres que d'observations, 4 ; il est donc saturé et sa vraisemblance est maximale. L'AIC est égal à 33.616.

Les tables 4.5 et 4.6 donnent l'analyse de la déviance lorsqu'on échange la place des deux variables dans le modèle :

	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			3	17.8925	
Genre	1	0.1675	2	17.7250	0.6823375
Statut	1	13.2021	1	4.5229	0.0002796 ***
Genre :Statut	1	4.5229	0	0.0000	0.0334433 *

TABLE 4.5 – Table de déviance pour le modèle sur l'équité au concours

	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			3	17.8925	
Statut	1	13.1667	2	4.7258	0.000285 ***
Genre	1	0.2029	1	4.5229	0.652424
Genre :Statut	1	4.5229	0	0.0000	0.0334433 *

TABLE 4.6 – Table de déviance pour le modèle sur l'équité au concours

Le dispositif n'étant pas orthogonal (voir la section 1.5) les tests sur les effets dépendent de l'ordre dans lequel ils ont été introduits dans le modèle. On note que l'effet de l'interaction est légèrement significatif, l'effet Statut est très significatif et l'effet Genre n'est pas significatif. Si on ne considère plus l'interaction on a

$$Y_{ij} \overset{ind.}{\sim} \mathcal{B}(n_{ij}, p_{ij})$$

$$logit(p_{ij}) = \mu + \alpha_i + \beta_j$$

Ce modèle a une déviance résiduelle égale à 4.5229 et un AIC égal à 36.139. La table d'analyse de la déviance indique que l'effet Genre n'est pas significatif alors que l'effet

Statut l'est. Si on ne prend en compte que le Statut on a

$$Y_i \overset{ind.}{\sim} \mathcal{B}(n_i, p_i)$$

$$logit(p_i) = \mu + \alpha_i$$

Ce modèle a une déviance résiduelle égale à 4.7258 et un AIC égal à 34.341. La table d'analyse de la déviance indique que l'effet Statut est très significatif. Ce modèle a un AIC à peine supérieur à celui du modèle complet.

La significativité de l'effet Statut, et la valeur du coefficient associé à la modalité "non boursier" montre que le fait de ne pas être bénéficiaire d'une bourse est associé à une meilleure réussite au concours. De fait la probabilité de succès pour un candidat boursier est de 0.37 alors qu'elle est de 0.47 pour un candidat non boursier. L'effet Genre n'est pas significatif : la probabilité de succès au concours est de 0.45 pour les candidats masculins et de 0.44 pour les candidates féminines. Par contre l'interaction entre l'effet statut et l'effet genre est statistiquement significative, la probabilité de succès pour les filles est très différente entre celles qui sont titulaires d'une bourse et celles qui ne le sont pas (0.34 et 0.47) alors que cette différence est beaucoup plus faible pour les garçons 0.45 et 0.46.

4.2.3 Programme R

```
# Boursiers.R : modèle linéaire généralisé sur l'équité des concours
#=====================================================================

# Données
#--------
Admis = c(111,557,61,214)
NonAdmis = c(214,625,76,256)
concours = data.frame(Genre = c("F","F","M","M"),
                      Statut=c("B","NB","B","NB"),
                      Admis=Admis,NonAdmis=NonAdmis)

Y = cbind(concours$Admis,concours$NonAdmis)
nb = concours$Admis+concours$NonAdmis
                          # nombre de répétitions par facteur
prop = concours$Admis/nb

# Modeles logistiques
#--------------------

# avec interaction
#-----------------
GLM.1 <- glm(Y ~ Genre * Statut , family=binomial(logit),
            data=concours)
GLM.1b = glm(Y ~ Statut * Genre, family=binomial(logit),
            data=concours)
summary(GLM.1)
```

```
anova(GLM.1,test="Chisq")
anova(GLM.1b,test="Chisq")
predict(GLM.1,type='response')*nb

# sans interaction
#-----------------
GLM.2 <- glm(Y ~ Genre + Statut , family=binomial(logit),
             data=concours)
GLM.2b = glm(prop ~ Genre + Statut , family=binomial(logit),
             data=concours, weight=nb) # ecriture proportion
summary(GLM.2)
anova(GLM.2,test="Chisq")
predict(GLM.2,type='response')*nb

# seulement le statut
#--------------------
GLM.3 <- glm(Y ~ Statut  , family=binomial(logit), data=concours)
summary(GLM.3)
anova(GLM.3)
predict(GLM.3,type='response')*nb

# seulement le genre
#-------------------
GLM.3b <- glm(Y ~ Genre  , family=binomial(logit), data=concours)
summary(GLM.3b)
anova(GLM.3b)
predict(GLM.3b,type='response')*nb
```

4.3 Loi binomiale : courbe dose-réponse

4.3.1 Contexte et données

C.I. Bliss [4] a étudié la survie d'un insecte ravageur de la farine, *Tribolium Confusum* à 5 heures d'exposition au sulfure de carbone gazeux à plusieurs concentrations et a obtenu les résultats de la table 4.7 :

Log dose	Nb d'insectes	Nb d'insectes morts
1.691	59	6
1.724	60	13
1.755	62	18
1.784	56	28
1.811	63	52
1.837	59	53
1.861	62	61
1.884	60	60

TABLE 4.7 – Insecticide

4.3.2 Résultats

La variable réponse est le nombre d'insectes *Tribolium Confusum* tués. Sa loi de probabilité est la loi binomiale et la fonction de lien pourrait être la fonction logit. Mais dans ce cas on préfère la fonction probit parce qu'elle permet une interprétation intéressante en toxicologie comme cela est indiqué dans le paragraphe suivant.

Fonction probit

x est le dosage d'un produit toxique et Y mesure la survie ($Y = 1$: mort, $Y = 0$: survie). Si la dose léthale pour chaque individu, T, varie d'un individu à un autre selon une loi normale $\mathcal{N}(\mu, \sigma)$, on a

$$P(Y = 1|x) = P(T < x) = F\left[(x - \mu)/\sigma\right]$$

où F est la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$. On a donc

$$p(x) = F\left[(x - \mu)/\sigma\right]$$

et donc

$$F^{-1}(p(x)) = (x - \mu)/\sigma = \beta_0 + \beta_1 x$$

Le modèle probit est donc un modèle linéaire généralisé avec F^{-1} comme fonction de lien. On a de plus les relations suivantes entre les paramètres de la loi normale et ceux du modèle probit :

$$\mu = -(\beta_0/\beta_1)$$

et

$$\sigma = 1/\beta_1.$$

Avec les données de Bliss, on obtient $\hat{\beta}_0 = -34.96$ et $\hat{\beta}_1 = 19.74$. On en déduit que la tolérance T de *Tribolium Confusum* au sulfure de carbone est distribuée selon une loi $\mathcal{N}(1.77, 0.05)$.

Informations sur le modèle

Data Set	WORK.BLISS
Distribution	Binomial
Link Function	Probit
Dependent Variable	survie
Frequency Weight Variable	effectif

Number of Observations Read	16
Number of Observations Used	15
Sum of Frequencies Read	481
Sum of Frequencies Used	481
Number of Events	8
Number of Trials	15
Missing Values	1

Profil de réponse

Valeur ordonnée	survie	Fréquence totale
1	0	291
2	1	190

PROC GENMOD is modeling the probability that survie='0'.

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	479	371.2255	0.7750
Scaled Deviance	479	371.2255	0.7750
Pearson Chi-Square	479	434.7189	0.9076
Scaled Pearson X2	479	434.7189	0.9076
Log Likelihood		-185.6128	

Algorithm converged.

Estimation des paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi2	Pr > Khi2
Intercept	1	-34.9561	2.6413	-40.1330	-29.7793	175.15	<.0001
logd	1	19.7410	1.4853	16.8300	22.6521	176.66	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Fonction log-log

Les fonctions probit et logistique sont symétriques par rapport au point $(0, 1/2)$, c'est à dire qu'elles vérifient $h(u) = 1 - h(-u)$. Ces modèles ne conviennent donc pas quand $p(x)$ augmente lentement à partir de 0 mais s'approche de 1 rapidement ou le contraire. Dans ce cas on peut utiliser la fonction de lien log-log complémentaire :

$cloglog(p) = \log[-\log[1 - p(x)]] = \beta_0 + \beta_1 x$

ou la fonction de lien log-log, associée à la distribution des valeurs extrêmes de Gumbel :

$\log[-\log[p(x)]] = \beta_0 + \beta_1 x$

Informations sur le modèle

Data Set	WORK.BLISS
Distribution	Binomial
Link Function	CLL
Dependent Variable	survie
Frequency Weight Variable	effectif

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	479	364.7529	0.7615
Scaled Deviance	479	364.7529	0.7615

```
Pearson Chi-Square        479        437.1062          0.9125
Scaled Pearson X2         479        437.1062          0.9125
Log Likelihood                      -182.3765

  Algorithm converged.

                    Estimation des paramètres

                          Erreur     Wald 95Limites
Paramètre  DF Estimation standard     de confiance %    Khi2  Pr > Khi2
Intercept   1   -39.5222   3.2229  -45.8390  -33.2055  150.38    <.0001
logd        1    22.0147   1.7899   18.5067   25.5228  151.28    <.0001
Scale       0     1.0000   0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.
```

Ajustement du modèle aux données

L'ajustement mesuré par le χ^2 ou la déviance est correct pour les 2 modèles. Cependant la figure 4.5 montre que la fonction de lien "Complementary log-log" s'ajuste mieux aux données aux deux extrémités de la courbe.

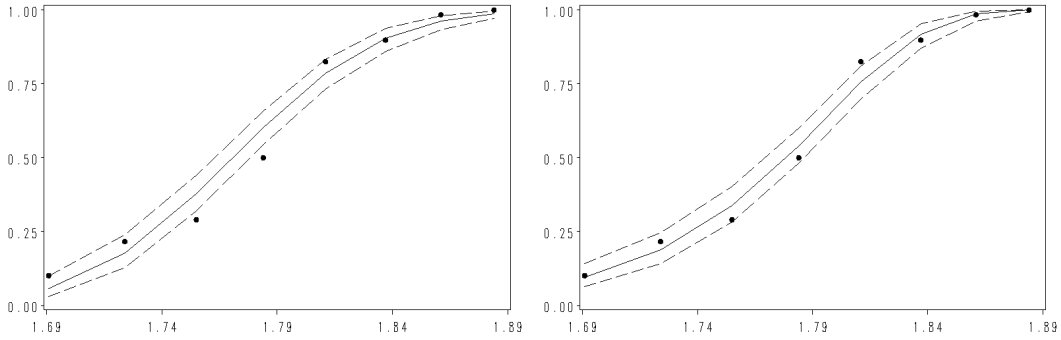


FIGURE 4.5 – Deux modélisations des données de Bliss : Probit (gauche), Cloglog(droite). En abscisse : logarithme de la dose, en ordonnée : taux de mortalité. Les points sont les observations, la ligne continue le taux de mortalité estimé. En pointillés les intervalles de confiance 95%.

4.3.3 programme SAS

```
data bliss; input logd survie effectif @@; lines; 1.691 0 6  1.691
1 53 1.724 0 13 1.724 1 47 1.755 0 18 1.755 1 44 1.784 0 28 1.784
1 28 1.811 0 52 1.811 1 11 1.837 0 53 1.837 1 6 1.861 0 61 1.861 1
1 1.884 0 60 1.884 1 0 ;

/* Modele Probit.Pour le modèle cloglog, il suffit de changer la
                          fonction de lien*/
 proc genmod data=bliss; title'Modele probit';
freq effectif; model survie= logd / link=probit dist=binomial ;
          output out=blissout p=probapredite l=ICinf u=ICsup
```

```
          stdreschi=residuchi2std;

/* preparation de la figure*/
 proc sort data=blissout; by logd;
proc means data=blissout noprint; var effectif;by logd;
           output out=effectiftotal sum=total;
data blissout1;merge blissout effectiftotal;
      by logd;if survie=0;frequenceobs=effectif/total;
     drop _TYPE_ _FREQ_ survie effectif  total; run;
data obs;set blissout1;mortalite=frequenceobs;type='observe';
keep logd mortalite type;
data pred;set blissout1;mortalite=probapredite;type='predit';
keep logd mortalite type;
data ICinf;set blissout1;mortalite=ICinf;type='ICinf';
keep logd mortalite type;
data ICsup;set blissout1;mortalite=ICsup;type='ICsup';
keep logd mortalite type;
data figure;set obs pred ICinf ICsup;

/* axes de la figure*/
axis1 order=(1.69 to 1.89 by 0.05)  value=(height=1.5)
      minor=NONE label=NONE;
axis2 order=(0 to 1 by 0.25) value=(height=1.5) minor=NONE
      label=NONE;
 /* Figure*/
proc gplot data=figure; title ' Modele probit, donnees de Bliss';
plot mortalite*logd=type/ vaxis=axis2 haxis=axis1 legend=legend1;
run;
```

4.4 Loi de Bernoulli, test de Hosmer-Lemeshow : asthme des enfants dans les écoles

4.4.1 Contexte et données

Les connaissances relatives aux effets de la pollution atmosphérique sur la santé respiratoire et allergique sont en augmentation ces dernières années, du fait d'un nombre important d'enquêtes toxicologiques et épidémiologiques sur ce sujet. La plupart des études épidémiologiques se focalisent sur la pollution extérieure, essentiellement celle liée au trafic automobile et montrent un effet de cette pollution sur le développement et l'aggravation des maladies allergiques et respiratoires. La pollution de l'air intérieur est également impliquée dans l'augmentation du risque de phénomènes d'irritation. Cependant, les données sont encore peu nombreuses en comparaison avec celles sur les effets de la pollution de l'extérieur des locaux. Les mécanismes d'action de ces polluants sur les voies respiratoires sont encore peu connus et plusieurs hypothèses ont été proposées. Certains polluants, notamment les particules (PM : particulate matter), le dioxyde d'azote (NO_2) et l'ozone pourraient avoir un rôle dans l'inflammation des bronches. Le but de cette étude est d'analyser les associations pouvant exister entre l'exposition à la pollution intérieure et l'asthme. Les données ont été fournies par Marion Flamant-Hulin, [25]. Elles sont issues de l'enquête épidémiologique E6C (Etude des 6 cités), volet français d'ISAAC (International Study of Asthma and Allergies in Childhood). Cette

étude transversale s'est déroulée de février 1999 à décembre 2000 dans six villes de France. La population de l'enquête se composait de 104 enfants, échantillonnés parmi la population de Clermont-Ferrand de l'E6C. L'échantillon a été établi en sélectionnant un asthmatique pour deux non asthmatiques. Ils étaient issus de 42 classes des 18 écoles de Clermont-Ferrand et sont âgés de 10 ans en moyenne. Le protocole d'enquête consistait en un questionnaire épidémiologique standardisé, un bilan médical centré sur les pathologies allergiques et respiratoires et des mesures de pollution atmosphérique réalisées par des techniciens de l'enquête.

Variables explicatives

Variables mesurant la pollution La pollution a été mesurée en classe pendant une semaine à l'aide de pompes et de capteurs. Les variables créées à partir de la distribution des polluants considérés sont :

- `no2claseco` : le dioxyde d'azote, provenant principalement de sources externes (trafic automobile) mais aussi internes telles que les appareils fonctionnant au gaz ou au fioul
- `acetalclaseco, formalclaseco` : l'acétaldéhyde et le formaldéhyde appartenant à la famille des composés organiques volatiles et provenant de diverses sources internes telles que les produits de construction et de décoration, les sources de combustion (bougies, encens) et les produits d'usage courant (produit d'entretien, cosmétiques)
- `pm2.5claseco` : les particules fines (PM2.5) provenant essentiellement de sources externes (poussières naturels, particules diesel, pollen) mais aussi des appareils de combustion, de bricolage.

Caractéristiques sociodémographiques

- `allfam`, antécédents familiaux d'allergies : un enfant était considéré comme ayant des antécédents d'allergies familiales si son père ou sa mère avaient déjà souffert d'asthme, de rhinite allergique ou d'eczéma (1, 0 sinon)
- `origmere` : origine géographique de la mère selon deux modalités : caucasienne (0) ou autre (1)
- niveauetude : niveau d'éducation parentale, en considérant le parent ayant suivi les études les plus longues. Elle valait 0 lorsque le parent avait suivi un enseignement jusqu'au secondaire ou primaire, et 1 jusqu'au supérieur,
- `age` : age de l'enfant au moment de l'enquête
- `BMI` : indice de masse corporelle de l'enfant
- `sexeb` : 1=garçon, 2=fille

Exposition à la pollution intérieure au domicile

- `tabaenf` : lorsqu'au moins une personne fumait dans le lieu de vie de l'enfant pendant la grossesse et/ou la première année de vie de l'enfant, l'enfant était considéré comme exposé (1, 0 sinon).

Variable Réponse

Pathologie allergique et respiratoire

- `Asthmevrai` : variable binaire, un enfant était considéré comme asthmatique dès qu'il avait une réponse positive à une des deux questions suivantes « Durant les 12 derniers mois, votre enfant a-t-il (elle) eu des sifflements dans la poitrine à un moment quelconque ? », « Votre enfant

a-t-il (elle) déjà eu des crises d'asthme ? » ou lorsqu'il se présentait au moment du bilan médical avec des médicaments contre l'asthme.

4.4.2 Résultats

La variable réponse suit une loi de Bernoulli, une régression logistique est naturelle dans ce cas. Le nombre de variables explicatives étant très élevé, on les sélectionne à l'aide d'une régression logistique descendante qui élimine les variables non significatives à partir du modèle complet.

```
                          The LOGISTIC Procedure
                    Informations sur le modèle
Data Set                              WORK.ASHME
Response Variable                     asthmevrai
Number of Response Levels             2
Model                                 binary logit
Optimization Technique                Fisher's scoring

   Number of Observations Read                104
   Number of Observations Used                 82


                    Profil de réponse
      Valeur                              Fréquence
    ordonnée     asthmevrai                totale
           1              1                    32
           2              0                    50

        Probability modeled is asthmevrai=1.

NOTE: 22 observations were deleted due to missing values for the response
      or explanatory variables.


                                 Backward Elimination Procedure
Step  0. The following effects were entered: Intercept  bmi age sexeb
tabaenf allfam  origmere_2  no2claseco formalclaseco acetalclaseco
       pm2_5claseco

                           Convergence criterion (GCONV=1E-8) satisfied.

                     Statistiques d'ajustement du modèle complet
                                              Coordonnée à l'origine
                     Coordonnée à l'origine                        et
       Critère                  uniquement               covariables
       AIC                         111.693                   103.091
       SC                          114.099                   129.565
       -2 Log L                    109.693                    81.091


Step  1. Effect sexeb is removed:
Step  2. Effect tabaenf is removed:
Step  3. Effect formalclaseco is removed:
```

```
Step  4. Effect acetalclaseco is removed:
Step  5. Effect pm2_5claseco is removed:
Step  6. Effect origmere_2 is removed:
Step  7. Effect bmi is removed:
```

Statistiques d'ajustement du modèle final

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	111.693	94.869
SC	114.099	104.496
-2 Log L	109.693	86.869

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	22.8232	3	<.0001
Score	20.3048	3	0.0001
Wald	15.8405	3	0.0012

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
5.4024	7	0.6110

No (additional) effects met the 0.05 significance level for removal from the model.

Récapitulatif sur l'élimination

Étape	Effet supprimé	DF	Nombre dans	Khi 2 de Wald	Pr > Khi 2
1	sexeb	1	9	0.3311	0.5650
2	tabaenf	1	8	0.3448	0.5571
3	formalclaseco	1	7	0.7783	0.3777
4	acetalclaseco	1	6	0.3570	0.5502
5	pm2_5claseco	1	5	0.9189	0.3378
6	origmere_2	1	4	0.6422	0.4229
7	bmi	1	3	2.2224	0.1360

Estimations des paramètres

Paramètre	DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept	1	-14.9247	4.7427	9.9026	0.0017
age	1	1.0340	0.4078	6.4300	0.0112
allfam	1	1.3733	0.5304	6.7031	0.0096
no2claseco	1	0.1180	0.0490	5.8058	0.0160

Estimations des rapports de cotes

Effet	Point Estimate	95% Limites de confiance de Wald	
age	2.812	1.265	6.254

allfam	3.948	1.396	11.166
no2claseco	1.125	1.022	1.239

Tests de Hosmer et Lemeshow

		asthmevrai = 1		asthmevrai = 0	
Groupe	Total	Observé	Attendu	Observé	Attendu
1	9	0	0.63	9	8.37
2	8	3	1.03	5	6.97
3	8	1	1.31	7	6.69
4	8	1	1.84	7	6.16
5	8	2	2.48	6	5.52
6	8	3	3.18	5	4.82
7	8	5	3.97	3	4.03
8	8	5	4.73	3	3.27
9	9	5	6.27	4	2.73
10	8	7	6.56	1	1.44

Test d'adéquation d'Hosmer et de Lemeshow

Khi 2	DF	Pr > Khi 2
7.3502	8	0.4994

Les variables conservées finalement sont l'age de l'enfant, les antécédents familiaux et le dioxyde d'azote. Les autres variables n'ont pas d'effet statistiquement significatif. Cependant comme le dispositif expérimental comporte peu de données, l'analyse manque de puissance et les variables éliminées ont peut-être un effet qui n'a pas été détecté. Le test de Hosmer-Lemeshow montre une bonne adéquation du modèle au données. Un an de plus pour un enfant multiplie le rapport $\frac{p}{1-p}$ par 2.8, le fait d'avoir des antécédents familiaux par 3.9 et 1 $\mu g/m^3$ de NO_2 en plus par 1.125. Sachant que le taux de NO_2 varie de 15 à 33 dans cette étude, s'il augmente de 10 cela implique une multiplication par $1.125^{10} = 3.2$, ce qui n'est pas négligeable. Cependant l'intervalle de confiance dans ce dernier cas est large : de 1.24 à 8.5. Il faut faire attention à ne pas interpréter les odd-ratios comme des risques relatifs parce $1 - p$ est très différent de 1.

4.4.3 Programme SAS

```
    data ashme;set Tmp1.Tablegnalssfaux;
proc means;run;
ods html;
   ods graphics on;

proc logistic data=ashme descending;model asthmevrai = bmi age
sexeb tabaenf allfam origmere bmi no2claseco formalclaseco
acetalclaseco pm2.5claseco/ selection=backward lackfit ctable
outroc=roc1; run; ods graphics off;
   ods html close;
```

4.5 Loi de Poisson avec offset : biodiversité des fourmis en Guyane

4.5.1 Contexte et données

Le but de l'étude est d'étudier la biodiversité des fourmis en forêt tropicale dans différents milieux, en comparant leur richesse. Le dispositif expérimental se situe sur le site des Nourragues en Guyane Française, quatre types de milieux sont étudiés : la forêt de plateau (`GPWT`), la forêt de liane (`FLWT`), la forêt de transition (`FTWT`) et la forêt d'Inselberg (`INWT`) [29]. Une observation est la donnée du nombre d'espèces de fourmis présentes dans 1 m^2 de litière récolté dans un milieu donné. Les échantillons de litière récoltés sont pesés (variable `Poids` exprimé en kilogramme) et le poids est considéré comme un proxy (un indicateur) de l'épaisseur de la litière. 50 points d'échantillonnage distants d'au moins 10m ont été choisis dans chaque milieu, sauf pour la forêt d'Inselberg, pour laquelle seuls 20 points d'échantillonnage ont été sélectionnés en raison de sa relative petite taille.

Les graphiques proposés sur la figure 4.6 illustrent la variabilité des poids de litière récoltés en fonction du site, ainsi que la variabilité du nombre d'espèces de fourmis présentes dans chaque échantillon pour chacun des sites.

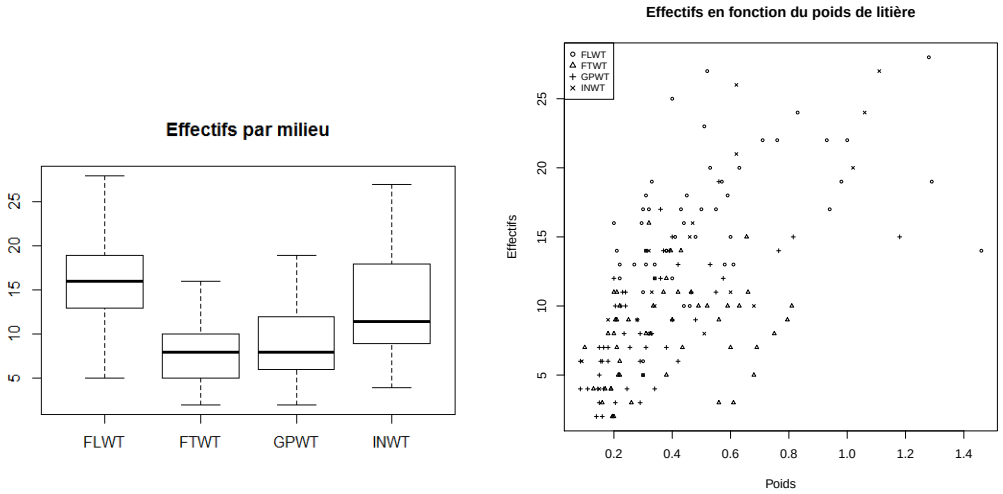


FIGURE 4.6 – Nombre d'espèces en fonction du site et du poids de litière

4.5.2 Résultats

La variable Y que l'on cherche à expliquer est un nombre d'espèces de fourmis présentes dans $1m^2$ de litière sur un site donné.

En supposant que les différentes espèces de fourmis se répartissent au hasard et de façon homogène dans la litière, il est naturel de modéliser le nombre d'espèces de fourmis présentes par une loi de Poisson. Le paramètre de la loi de Poisson représente le nombre moyen d'espèces de fourmis attendu pour chaque observation.

$$Y_{ij} \overset{ind.}{\sim} \mathcal{P}(\lambda_i),$$

$i = 1, \ldots, 4$ désignant le site et $j = 1, \ldots, n_i$, le numéro de l'observation au sein du site i.

Pour que les différents prélèvements soient comparables, il faudrait que les volumes de litière prélevés soient identiques, λ_i désignerait alors un nombre d'espèces de fourmis moyen par unité de volume, c'est-à-dire une densité. Puisque ce n'est pas le cas dans l'expérience, il est nécessaire de prendre en compte la différence de volume des observations. Idéalement, le modèle serait :

$$Y_{ij} \overset{ind.}{\sim} \mathcal{P}(\lambda_i V_{ij}),$$

V_{ij} désignant le volume de litière prélevé pour l'observation j sur le site i. Cette information n'étant pas disponible, le poids de litière va être utilisé comme indicateur de ce volume. Le modèle d'observations proposé est donc

$$Y_{ij} \overset{ind.}{\sim} \mathcal{P}(\lambda_i W_{ij}),$$

W_{ij} désignant le poids de litière prélevée pour l'observation j sur le site i. λ_i désigne donc le nombre d'espèces de fourmis par unité de poids.

En utilisant la fonction de lien naturel de la loi de Poisson, le modèle final est donné par

$$\begin{aligned} Y_{ij} &\overset{ind.}{\sim} \mathcal{P}(\lambda_i W_{ij}), \\ \log \lambda_i &= \mu + \alpha_i \end{aligned} \tag{4.1}$$

En appliquant l'algorithme de Newton-Raphson, on obtient les estimations suivantes pour les paramètres du modèle :

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-3.47289	0.03505	-99.084	< 2e-16
SiteFTWT	-0.39235	0.06127	-6.404	1.51e-10
SiteGPWT	-0.14423	0.05962	-2.419	0.0155
SiteINWT	-0.11043	0.07004	-1.577	0.1149

Les valeurs prédites pour 1 kg de litère sont donc $\widehat{\lambda}_{FLWT} = 31.0$, $\widehat{\lambda}_{FTWT} = 21.0$, $\widehat{\lambda}_{GPWT} = 26.9$, $\widehat{\lambda}_{INWT} = 27.8$. La figure 4.7 donne le nombre d'espèces attendu en fonction du poids de litière et du milieu.

4.5.3 Richesses spécifiques

Pour étudier les différences de richesse entre les différents sites, on est amené à tester l'effet `Site`. On souhaite tester les deux hypothèses suivantes :

$$\begin{aligned} H_0 &= \{\forall i, j \quad \alpha_i = \alpha_j\} \\ H_1 &= \{\text{il existe } i \neq j, \alpha_i \neq \alpha_j\} \end{aligned}$$

La table 4.8 donne les déviances du modèle nul (sous l'hypothèse H_0) et du modèle complet, ainsi que la valeur critique du test de rapport de vraisemblance.

La déviance diminue de 432.2 à 389.3 avec l'introduction de l'effet site et la valeur critique du test du rapport de vraisemblance est très faible : le milieu a un effet significatif sur la richesse spécifique.

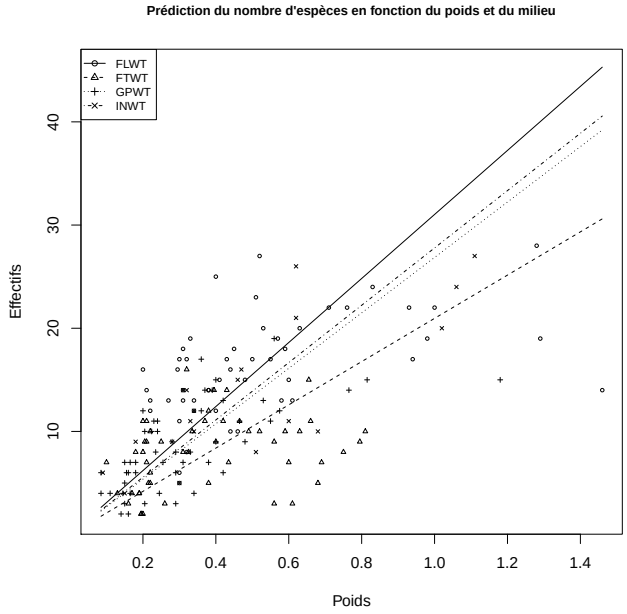


FIGURE 4.7 – Nombre d'espèces prédit en fonction du site et du poids de litière

	Df	Deviance	Resid. Df	Resid. Dev	Pr($>\chi$)
NULL			169	432.22	
Site	3	42.957	166	389.27	2.513e-09

TABLE 4.8 – Table de déviance pour la diversité des fourmis

4.5.4 Comparaison des richesses spécifiques

Pour comparer les richesses de deux sites i et j on va tester l'égalité des densités moyennes des deux sites ou ce qui revient au même l'égalité de leur logarithme, c'est à dire des paramètres naturels. Les hypothèses s'écrivent :

$$H_0 : \omega_i = \omega_j \qquad H_1 : \omega_i \neq \omega_j$$

ou encore

$$H_0 : C\omega \qquad H_1 : C\omega \neq 0$$

avec $C = (c_k)_k$, $c_i = 1$, $c_j = -1$, et $c_\ell = 0$ pour $\ell \notin \{i, j\}$, $\omega = (\omega_1, ...\omega_4)$. D'après le résultat 3.8 la statistique de test $T = [C\ I(\omega)^{-1}C']^{-1/2}(C\hat{\omega})$ suit asymptotiquement une loi normale centrée réduite sous l'hypothèse H_0. La matrice d'information de Fisher est égale d'après l'expression 3.3 à $I(\omega) = \text{diag}(n_i e^{\omega_i} \sum_j W_{ij})$.
La table 4.9 donne pour chaque paire de sites la valeur critique du test d'égalité des richesses spécifiques :

Les différences de richesse spécifique sont donc significatives pour chaque paire de sites au risque de 5%.

4.5.5 Programme R

```
# Fourmis.R : régression Poisson biodiversite des fourmis en Guyane
```

	FTWT	GPWT	INWT
FLWT	$<10^{-14}$	$<10^{-14}$	$<10^{-14}$
FTWT		$<10^{-14}$	$<10^{-14}$
GPWT			0.026

TABLE 4.9 – Valeurs critiques du test d'égalité des richesses spécifiques pour deux sites.

```
#=================================================================

# Lecture des données
#--------------------
Fourmis = read.csv("Fourmis.csv",sep=";",header=TRUE)

# Etude descriptive
#------------------
head(Fourmis)
summary(Fourmis)

with(Fourmis, boxplot(Effectifs~Site,main="Effectifs par milieu"))
with(Fourmis,
     plot(Effectifs~Poids,col=Fourmis$Site,
     pch=as.numeric(Fourmis$Site),
     cex=0.8,main="Effectifs en fonction du poids de litière"))
legend("bottomright",levels(Fourmis$Site),col=1:4,pch=1:4,cex=0.8)

# modele Poisson offset
#----------------------
glmInt = glm(Effectifs~Site,offset=log(Poids), family="poisson",
             data=Fourmis)
summary(glmInt)
anova(glmInt, test="Chisq")

predict(glmInt)
predict(glmInt,type="response")
datanew = data.frame(Site = levels(Fourmis$Site), Poids= rep(1,4))
lambdachap  = predict(glmInt,type="response",new = datanew)

# Graphiques des effectifs prédits en fonction du poids
#------------------------------------------------------
abcisse = seq(min(Fourmis$Poids),max(Fourmis$Poids),length=3)
col1 = c(rep(abcisse[1],4),rep(abcisse[2],4),rep(abcisse[3],4))
col2 = rep(levels(Fourmis$Site),3)
newdata = data.frame(Poids=col1,Site=col2)
tabPred = matrix(predict(glmInt,newdata,type="response"),3,4,
          byrow=TRUE)

matplot(abcisse, tabPred,type="l",col=1:4,lty=1,xlab="Poids",
```

```
        ylab="Effectifs", main="Prédiction du nombre d'espèces en
        fonction du poids et du milieu", cex.main=0.8)
with(Fourmis,
    points(Effectifs~Poids, col=as.numeric(Site),
    pch=as.numeric(Site),
    cex=0.5))
legend("bottomright",levels(Fourmis$Site),col=1:4,pch=1:4,cex=0.75)

# Comparaison des densités
#-------------------------
omegai = coef(glmInt)[1] +c(0,coef(glmInt)[2:4])
exp(omegai)
ni=c(50,50,50,20)
li=exp(omegai)
sumpoids = tapply(Fourmis$Poids,Fourmis$Site,sum)
I=ni*li*sumpoids      # matrice d information de Fisher
Va=diag(1/I) # variance asymptotique de l'estimateur du maximum de
               vraisemblance

tab =  matrix(0,4,4)
for (iforet in 1:3){
  for (jforet in (iforet+1):4){
    C = rep(0,4)
    C[iforet] = 1
    C[jforet] = -1
    C = as.matrix(C)
    num = t(C)%*%omegai
    den = sqrt(t(C)%*%as.matrix(Va)%*%C)
    st = num/den
    tab[iforet,jforet] = 2*pnorm(abs(st),lower=F) #calcul de la
                                                    p.value du test
  }
}
print(tab)
```

4.6 Poisson tronquée en 0 : portées d'agneaux

4.6.1 Contexte et données

On s'intéresse à l'influence du génotype sur la taille des portées chez la brebis. Pour $n = 161$ brebis qui ont mis bas, on a noté la taille de leur portée ainsi que leur génotype : BC pour Berrichon du Cher, BG pour le croisement (Berrichon du Cher) $\times$ Romanov et GG pour 4 générations de croisements BG. La table 4.10 donne les nombres des portées de tailles allant de 1 à 5 petits pour chaque génotype.

De la table précédente on extrait la table 4.11 qui donne pour chaque génotype et pour tous génotypes confondus : le nombre de portées, le nombre total de petits et le nombre moyen de petits par portée.

	Taille de portées				
	1	2	3	4	5
BB	10	17	2	0	0
BG	5	25	52	18	1
GG	1	12	14	4	0

TABLE 4.10 – Taille des portées en fonction du génotype. Source : Ferme de Grignon, 1998.

	Nombre de portées	Nombre de petits	Nombre moyen de petits
BB	29	50	1.72
BG	101	288	2.85
GG	31	83	2.68
tous génotypes confondus	161	421	2.61

TABLE 4.11 – Caractéristiques des portées en fonction du génotype

L'examen de cette table montre qu'il semble y avoir une différence de taille des portées selon le génotype, la question est de savoir si elle est significative.

4.6.2 Résultats

Modèle

La variable Y que l'on cherche à expliquer est le nombre de petits qu'il est naturel de modéliser par une distribution de Poisson. Cependant, ici on ne retient que la taille des portées des brebis qui ont mis bas, et la valeur 0 ne sera jamais observée. Pour prendre en compte cette information dans la modélisation, on va supposer que le nombre de petits est distribué selon une loi de Poisson tronquée en 0. Soit X une variable aléatoire suivant une loi de Poisson de paramètre λ la loi de probabilité de Y est donnée par

$$P[Y=y] \quad = \quad P[X=y|X\neq 0] = \frac{P[X=y; X\neq 0]}{P[X\neq 0]} = \frac{1}{1-e^{-\lambda}}\frac{\lambda^y e^{-\lambda}}{y!} \qquad \text{pour } y>0$$

ou encore

$$P[Y=y] \quad = \quad \exp\left[y\log\lambda - \lambda - \log\left(1-e^{-\lambda}\right) - \log y!\right] \tag{4.2}$$

La fonction de lien naturel est donc comme pour la loi de Poisson la fonction log. L'espérance et la variance de Y sont données par

$$\mathbb{E}(Y) = \frac{\lambda}{1-e^{-\lambda}} \qquad \mathbb{V}(Y) = \frac{\lambda}{1-e^{-\lambda}}\frac{1-e^{-\lambda}(1+\lambda)}{1-e^{-\lambda}} < E(Y)$$

La loi de Poisson tronquée est donc un cas de sous-dispersion.
Soit Y_{ij} le nombre de petits observé pour le génotype i, $i = 1,2,3$ et la portée j,

$j = 1, \ldots, n_i$ ($n_1 = 29$, $n_2 = 101$ et $n_3 = 31$), le modèle s'écrit

$$\begin{aligned} Y_{ij} &\overset{ind.}{\sim} \mathcal{P}_{tronquée}(\lambda_i), \\ g(\lambda_i) &= \log \lambda_i = \omega_i = \mu + \alpha_i \end{aligned} \tag{4.3}$$

où λ_i est le paramètre de la loi de Poisson tronquée du nombre de petits par portée pour les brebis de génotype i.

Estimation

La table 4.12 donne les estimations des paramètres du maximum de vraisemblance obtenues par l'algorithme de Newton-Raphson.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.1906	0.2029	0.940	0.347433
Genotype2	0.7839	0.2133	3.675	0.000238
Genotype3	0.7035	0.2384	2.951	0.003172

TABLE 4.12 – Estimation des paramètres.

On en déduit $\hat{\omega}_1 = \hat{\mu} + \hat{\alpha}_1 = 0.191$, $\hat{\omega}_2 = \hat{\mu} + \hat{\alpha}_2 = 0.975$ et $\hat{\omega}_3 = \hat{\mu} + \hat{\alpha}_3 = 0.894$, d'où $\hat{\lambda}_1 = 1.21$, $\hat{\lambda}_2 = 2.65$ et $\hat{\lambda}_3 = 2.44$ et les valeurs prédites par génotype $\hat{y}_1 = 1.72, \hat{y}_2 = 2.85, \hat{y}_3 = 2.68$.

Effet du génotype.

Pour établir la réalité de l'effet génotype on teste l'hypothèse, écrite en termes de modèle, suivante :

$\mathrm{H}_0 = \{g(\lambda_i) = \mu; \quad i = 1, 2, 3\}$ contre

$\mathrm{H}_1 = \{g(\lambda_i) = \mu + \alpha_i; \exists i \neq j \text{ t.q. } \alpha_i \neq \alpha_j; \quad i = 1, 2, 3\}$.

Ce test peut être réalisé à l'aide du rapport de vraisemblance (cf section 3.4.1, p. 120 du chapitre Modèle Linéaire Généralisé). La statistique de test $T = -2(\ell(y, \widehat{\theta_0}) - \ell(y, \widehat{\theta}))$ est évaluée en calculant la log-vraisemblance en son maximum sous l'hypothèse H_0 et sous H_1. On a $\ell(y, \widehat{\theta_0}) = -236.47$ et $\ell(y, \widehat{\theta}) = -227.62$ ce qui donne $t_{obs} = 17.69$ alors que le quantile d'ordre 0.95 de la loi du χ^2_{3-1} vaut $\chi^2_{0.95,2} = 5.99$ (probabilité critique de 10^{-4}). L'hypothèse H_0 est donc rejetée et on conclut à un effet significatif du génotype sur la taille des portées.

4.6.3 Comparaison des génotypes

On peut remarquer que d'une part il ne semble pas y avoir de différence importante du nombre moyen de petits par génotype entre les générations de croisées (1 génération (BG) ou 4 générations (GG)) mais que par contre le nombre moyen de petits semble plus élevé pour les races croisées (BG ou GG) que pour les races pures (BC). On va chercher ici à affirmer ou infirmer ces hypothèses.

Pour formuler ces hypothèses en termes des paramètres naturels $\omega = (\omega_1, \omega_2, \omega_3)$ on va réécrire la log-vraisemblance en fonction de ces paramètres et calculer la matrice d'information de Fisher associée.

La log-vraisemblance de l'échantillon s'écrit :

$$\begin{aligned}\mathcal{L}(y_{11},\ldots,y_{3n_3};\omega) &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}\left[y_{ij}\log\lambda_i-\lambda_i-\log y_{ij}!-\log\left(1-e^{-\lambda_i}\right)\right]\\ &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}\left[y_{ij}\omega_i-e^{\omega_i}-\log y_{ij}!-\log\left(1-e^{-e^{\omega_i}}\right)\right],\\ &= \sum_{i=1}^{3}\left[y_{i+}\omega_i-n_ie^{\omega_i}-\sum_{j=1}^{n_i}\log y_{ij}!-n_i\log\left(1-e^{-e^{\omega_i}}\right)\right], \quad (4.4)\end{aligned}$$

où $y_{i+}=\sum_{j=1}^{n_i}y_{ij}$.
De

$$\frac{\partial\mathcal{L}(Y;\omega)}{\partial\omega_i}=Y_{i+}-\frac{n_ie^{x_i\omega}}{1-e^{-e^{\omega_i}}} \quad \text{et} \quad \frac{\partial^2\mathcal{L}(Y;\theta)}{\partial\omega_i^2}=-n_ie^{\omega_i}\frac{1-e^{-e^{\omega_i}}(1+e^{\omega_i})}{(1-e^{-e^{\omega_i}})^2}.$$

on déduit la matrice d'information de Fisher $I(\omega)=-E\left[\frac{\partial^2\mathcal{L}(Y;\theta)}{\partial\theta^2}\right]$ de terme I_{ij}, avec $I_{ii}=n_ie^{\omega_i}\frac{1-e^{-e^{\omega_i}}(1+e^{\omega_i})}{(1-e^{-e^{\omega_i}})^2}$ et $I_{ij}=0$ ce qui nous donne la loi asymptotique de $\widehat{\omega_i}$: $I(\omega)^{1/2}\left(\hat{\omega}_n-\omega\right)\xrightarrow{loi}\mathcal{N}(0,Id)$.

Dans un premier temps, on cherche à savoir si il existe un effet du nombre de générations (génotype BG et GG) chez les races croisées. Ce test s'écrit :

$$\mathrm{H}_0=\{\omega_2=\omega_3\} \text{ contre } \mathrm{H}_1=\{\omega_2\neq\omega_3\},$$

ou encore

$$\mathrm{H}_0=\{C\omega=\omega_2-\omega_3=0\} \text{ contre } \mathrm{H}_1=\{C\omega=\omega_2-\omega_3\neq0\},$$

«««< HEAD avec la combinaison linéaire $C=(0,1,-1)$. Pour effectuer ce test (cf section 3.4.3), on ======= On cherche donc à tester la combinaison linéaire des paramètres suivante :

$$\mathrm{H}_0=\{C\omega=0\},$$

où $C=(0,1,-1)$. Pour effectuer ce test (cf section 3.4.3, p. 122 du chapitre Modèle Linéaire Généralisé), on »»»> 035b2577cd7e1d4d05d16db90695406320ee8020 utilise la loi asymptotique de l'estimateur du vecteur des paramètres ω donnée par 3.8, p. 122. La statistique de test va alors s'écrire :

$$T=(1/I_{22}+1/I_{33})^{-1/2}*(\hat{\omega}_2-\hat{\omega}_3),$$

où I_{ii} est la $i$ème composante de la diagonale de l'information de Fisher. La valeur observée est égale à 0.568. Or le quantile d'ordre $1-\alpha/2$ d'une loi normale (avec $\alpha=0.05$) vaut 1.96, (ou la probabilité critique vaut 0.57), donc on ne rejette pas l'hypothèse H_0 : il n'y a pas d'effet du nombre de générations de croisements sur la taille des portées.

Dans un second temps, on cherche à savoir si il y a une différence entre les races pures (BC) ou les races croisées (BG et GG). Le test précédent nous a permis de conclure

qu'il n'y avait pas de différence significative entre les générations de croisements sur la taille des portées. Ainsi, on va regarder s'il existe une différence entre l'effet de BC avec l'effet moyen des croisements, estimé par $\omega_{23} = \frac{n_2\omega_2+n_3\omega_3}{n_2+n_3}$ ce qui se traduit en termes de test d'hypothèses par :

$$\mathrm{H}_0 = \{\omega_1 = \omega_{23}\} \text{ contre } \mathrm{H}_1 = \{\omega_1 \neq \omega_{23}\},$$

ou

$$\mathrm{H}_0 = \{C\omega = 0\} \text{ contre } \mathrm{H}_1 = \{C\omega \neq 0\},$$

avec $C = (1, -n_2/(n_2+n_3), -n_3/(n_2+n_3)$.

La statistique de test $T = [C\ I(\omega)^{-1}C']^{-1/2}(C\hat{\omega})$ a pour valeur observée -3.62. La comparaison de cette statistique (en valeur absolue) au quantile d'ordre 0.975 (probabilité critique 3e-04) de la loi normale nous permet de conclure qu'il existe une différence significative entre les races pures et croisées sur la taille des portées.

4.6.4 Programme R.

```
# Agneaux.R : modèle Poisson tronqué portées d'agneaux
#=====================================================
library(countreg)

# lecture des données
#--------------------
Portees=read.csv("Portees.csv",header=TRUE,sep=";")
Portees$Genotype=as.factor(Portees$Genotype)

# calcul des effectifs
#---------------------
with(Portees,table(Genotype,NbPetits))
with(Portees,tapply(NbPetits,Genotype,sum))
with(Portees,tapply(NbPetits,Genotype,mean))
sum(Portees$NbPetits)
mean(Portees$NbPetits)

# Estimation modèle Poisson tronqué
#----------------------------------
res.mod <- zerotrunc(NbPetits ~ Genotype, data = Portees,
                     dist = "poisson")
summary(res.mod)
esti.para = coef(res.mod)
logLik(res.mod)

# modèle simple
#--------------
res.mod0 <- zerotrunc(NbPetits ~ 1, data = Portees, dist ="poisson")
summary(res.mod0)
logLik(res.mod0)
```

```
# Test de l'effet génotype
#-------------------------
# Statistique de test du rapport de vraisemblance
loglik.mod = res.mod$loglik
loglik.mod0 = res.mod0$loglik
tobs=2*(loglik.mod-loglik.mod0)

qchisq(0.95,2)   # quantile
1-pchisq(tobs,2) # p-value

# Calcul de la matrice d'information de Fisher en l'estimation
des paramètres
#-----------------------------------------------------------------
esti.omegai=
c(esti.para[[1]],esti.para[[1]]+esti.para[[2]],esti.para[[1]]+
      esti.para[[3]])
ni=c(29,101,31)
li = exp(esti.omegai)
I = ni*li*(1-exp(-li)*(1+li))/((1-exp(-li))^2)

# variance asymptotique de l'estimateur du maximum de vraisemblance
Va=diag(1/I)

# Test de l'effet du nombre de générations chez les croisées
#-----------------------------------------------------------
C = as.matrix(c(0 ,1 ,-1))
num = t(C)%*%esti.omegai
den = sqrt(t(C)%*%as.matrix(Va)%*%C)
tobs = num/den
# calcul de la p.value du test
2*pnorm(abs(tobs),lower=F)

# Test de la différence entre les races pures et les races croisées
#------------------------------------------------------------------
C = as.matrix(c(1 ,-ni[2]/sum(ni[2:3]) ,-ni[3]/sum(ni[2:3])))
num = t(C)%*%esti.omegai
den = sqrt(t(C)%*%as.matrix(Va)%*%C)
tobs = num/den
# calcul de la p.value du test
2*pnorm(abs(tobs),lower=F)
```

4.7 Loi Gamma : roulements à billes, durées de vie

4.7.1 Contexte et données

Une expérience a été menée dans l'entreprise SKF pour améliorer la fiabilité des roulements à billes ([30]). La variable réponse est la durée de vie en heures. On soupçonne que 3 facteurs, notés A, B et C, peuvent avoir une effet sur la fiabilité. Un plan factoriel complet 2^3 (cf 7.4.3, p. 239) a été mis en place, et on a obtenu les résultats de la table 4.13 :

A	B	C	y
1	1	1	17
2	1	1	26
1	2	1	25
2	2	1	85
1	1	2	19
2	1	2	16
1	2	2	21
2	2	2	128

TABLE 4.13 – Essai Roulements à billes

4.7.2 Résultats

Un modèle usuel pour ce type de données est la loi de probabilité gamma de densité $f(y) = \frac{y^{k-1}e^{-\frac{y}{\theta}}}{\Gamma(k)\theta^k}$ avec $y \geq 0$ et les paramètres θ et k, avec $\mathbb{E}(Y) = k\theta$ et $\mathbb{V}(Y) = k\theta^2$. Cette loi pour des données positives (comme les durées de vie) permet de prendre en compte une forte dissymétrie fréquente pour ces données. La fonction de lien canonique est la fonction inverse $g(\mu) = 1/\mu$ qui représente le nombre de pannes à l'heure. On utilise aussi parfois la fonction *log*. On présente ci dessous les résultats de l'estimation du modèle par un programme R.

```
Analysis of Deviance Table Model: Gamma, link: inverse Response: y
Terms added sequentially (first to last)
Call: glm(formula = y ~ A + B + C + A:B + A:C + B:C,
          family = Gamma(link = "inverse"), data = RB)

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     7     5.1174
 A     1  2.44711        6     2.6703   3.470e-06 ***
 B     1  2.26437        5     0.4060   8.038e-06 ***
 C     1  0.03994        4     0.3660   0.5533
 A:B   1  0.18304        3     0.1830   0.2044
 A:C   1  0.01074        2     0.1722   0.7585
 B:C   1  0.05813        1     0.1141   0.4744
---

Call: glm(formula = y ~ A + B, family = Gamma(link = "inverse"),
          data = RB)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.128078   0.017471   7.331 0.000741 ***
 A          -0.028050   0.006870  -4.083 0.009511 **
 B          -0.031135   0.007256  -4.291 0.007784 **
---

(Dispersion parameter for Gamma family taken to be 0.07682146)

    Null deviance: 5.11743  on 7  degrees of freedom
Residual deviance: 0.40595  on 5  degrees of freedom AIC: 61.656

Number of Fisher Scoring iterations: 4

Analysis of Deviance Table
Model: Gamma, link: inverse Response: y
 Terms added sequentially

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     7     5.1174
 A     1   2.4471        6     2.6703 1.662e-08 ***
 B     1   2.2644        5     0.4060 5.662e-08 ***
---
       1      2     3     4     5     6     7     8
pred 14.51 24.48 26.48 103.01 14.51 24.48 26.48 103.0
 y   17.00 26.00 25.00  85.0  19.00 16.00 21.00 128.0
```

Le modèle additif avec les facteurs A et B suffit pour prédire correctement la durée de vie. Le facteur C et les interactions ne sont pas significatives. Pour augmenter la durée de vie (diminuer le nombre de pannes à la minute), il faut mettre les facteurs A et B au niveau haut. Si on utilise un lien linéaire, l'effet C et l'interaction AB sont significatives. Il arrive souvent qu'un bon choix de la fonction de lien permette de simplifier le modèle. Ici le choix de la fonction inverse simplifie la modélisation.

4.7.3 Programme R

```
RB=read.table('RoulementsBilles.txt',header=TRUE)
model1<-glm(y~A+B+C+A:B+A:C+B:C,family=Gamma(link="inverse"),data=RB)
anova(model1,test="Chisq")
model2<-glm(y~A+B, family=Gamma(link="inverse"),data=RB)
summary(model2)
anova(model2,test="Chisq")
pred=fitted.values(model2)
rbind(pred,y)
```

4.8 Loi multinomiale : condamnations à mort en Floride

4.8.1 Contexte et données

Les données de la table 4.14 ont été récoltées à partir des jugements sur des meurtres en Floride en 1976 et 77, [46].

		Condamnation à mort	
Couleur du meurtrier	Couleur de la victime	oui	non
Blanc	Blanche	19	132
	Noire	0	9
Noir	Blanche	11	52
	Noire	6	97

TABLE 4.14 – Données de Radelet

4.8.2 Modèles pour les tables de contingence de dimension 3

Ce type de données s'appelle une table de contingence de dimension 3, qui généralise la notion de tableau croisé au cas où il y a plus de 2 variables. La différence essentielle avec les exemples précédents, c'est qu'il n'y a pas de variable explicative, il n'y a que des variables réponses. On cherche à analyser les dépendances entre toutes les variables. On note
X_1 : couleur du meurtrier (blanc ou noir)
X_2 : condamnation à mort (oui ou non)
X_3 : couleur de la victime (blanc ou noir).
Soit

$$Y_{klr} = \#\{i \in (1,n) \ : \ X_1(i) = k, X_2(i) = l, X_3(i) = r\}.$$

On modélise ces données par un modèle linéaire généralisé multinomial avec la fonction de lien log : $Y \sim \mathcal{M}(n,m)$, où $m = (m_{klr}) = \mathbb{E}(Y_{klr})$ est la table de contingence des espérances de Y. On pose un modèle de type analyse de la variance sur $\log(m)$:

$$\log(m_{klr}) = \mu + u_k^1 + u_l^2 + u_r^3 + u_{kl}^{12} + u_{kr}^{13} + u_{lr}^{23} + u_{klr}^{123} \tag{4.5}$$

La plupart des différents modèles possibles ont une interprétation en termes d'indépendance ou indépendance conditionnelle :

- $\log(m_{klr}) = \mu + u_k^1 + u_l^2 + u_r^3 + u_{kl}^{12} + u_{kr}^{13} + u_{lr}^{23} + u_{klr}^{123}$, modèle complet ou saturé, avec $\hat{m}_{klr} = Y_{klr}$
- $\log(m_{klr}) = \mu + u_k^1 + u_l^2 + u_r^3 + u_{kl}^{12} + u_{kr}^{13} + u_{lr}^{23}$, modèle d'absence d'interaction d'ordre 3. Il n'y a pas de forme explicite pour $\hat{m}_{klr}$, ni d'interprétation en termes d'indépendance conditionnelle.
- $\log(m_{klr}) = \mu + u_k^1 + u_l^2 + u_r^3 + u_{kl}^{12} + u_{kr}^{13}$, modèle d'indépendance conditionnelle entre X_2 et X_3 pour X_1 fixé, et $\hat{m}_{klr} = \frac{Y_{kl+}Y_{k+r}}{Y_{k++}}$
- $\log(m_{klr}) = \mu + u_k^1 + u_l^2 + u_r^3 + u_{kl}^{12}$, modèle de l'indépendance entre (X_1,X_2) et X_3, avec $\hat{m}_{klr} = \frac{Y_{kl+}Y_{++r}}{n}$
- $\log(m_{klr}) = \mu + u_k^1 + u_l^2 + u_r^3$, modèle de l'indépendance mutuelle entre X_1,X_2 et X_3, avec $\hat{m}_{klr} = \frac{Y_{k++}Y_{+r+}Y_{+l+}}{n^2}$.

4.8.3 Résultats

```
                    The GENMOD Procedure

                      Model Information
            Data Set                WORK.RADELET
            Distribution                 Poisson
```

Link Function	Log
Dependent Variable	nombre

Number of Observations Read	8
Number of Observations Used	8

Class Level Information

Class	Levels	Values
coulmeur	2	b n
coulvict	2	b n
condamn	2	non oui

Modèle complet

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.7007	0.7007
Scaled Deviance	1	0.7007	0.7007
Pearson Chi-Square	1	0.3755	0.3755
Scaled Pearson X2	1	0.3755	0.3755
Log Likelihood		1080.2379	

Algorithm converged.

LR Statistics For Type 1 Analysis

Source	Deviance(*)	DF	Chi-Square(**)	Pr > ChiSq
Intercept	395.9153			
coulmeur	395.8049	1	0.11	0.7396
coulvict	363.3485	1	32.46	<.0001
coulmeur*coulvict	233.5508	1	129.80	<.0001
condamn	8.1316	1	225.42	<.0001
coulmeur*condamn	7.9102	1	0.22	0.6379
coulvict*condamn	0.7007	1	7.21	0.0073

(*) le modèle est fait de tous les termes au-dessus et sur la ligne en cours
(**) différence entre la déviance du modèle de la ligne au-dessus et celle de la ligne en cours

Modèle réduit

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	1.8819	0.9409
Scaled Deviance	2	1.8819	0.9409
Pearson Chi-Square	2	1.4313	0.7157
Scaled Pearson X2	2	1.4313	0.7157
Log Likelihood		1079.6473	

Algorithm converged.

LR Statistics For Type 1 Analysis

Source	Deviance(*)	DF	Chi-Square(**)	Pr > ChiSq
Intercept	395.9153			
coulmeur	395.8049	1	0.11	0.7396
coulvict	363.3485	1	32.46	<.0001
condamn	137.9294	1	225.42	<.0001
coulmeur*coulvict	8.1316	1	129.80	<.0001
coulvict*condamn	1.8819	1	6.25	0.0124

(*) le modèle est fait de tous les termes au-dessus et sur la ligne en cours
(**) différence entre la déviance du modèle de la ligne au-dessus et celle de la ligne en cours

Le modèle le plus simple qui s'ajuste aux données est le modèle réduit, avec 2 interactions d'ordre 2 , (Couleur de la victime)×Condamnation et (Couleur de la victime)× (Couleur du meurtrier). Cela signifie qu'on peut accepter l'hypothèse d' indépendance conditionnelle entre la condamnation et la couleur du meurtrier à couleur de la victime fixée. On note que la condamnation n'est liée qu'à la couleur de la victime et que la couleur de la victime est liée à la couleur du meurtrier.

4.8.4 Paradoxe de Simpson

A couleur de la victime fixée on note une propension moins forte à condamner un blanc qu'un noir : respectivement 12.6% et 17.5% de condamnés quand la victime est blanche, et 0% et 5.8% quand la victime est noire. Noter cependant que la liaison entre couleur du meurtrier et la condamnation n'est pas significative.

Si on ignore la couleur de la victime, on obtient la table marginale 4.15

TABLE 4.15 – Table marginale

	Condamnation à mort	
Couleur du meurtrier	oui	non
Blanc	19	141
Noir	17	149

On observe que les meurtriers blancs sont davantage condamnés que les meurtriers noirs : respectivement 19/141 et 17/149. Bien que la différence ne soit pas non plus statistiquement significative, on est interpellé par la différence de sens entre l'analyse marginale et l'analyse conditionnelle, alors qu'il s'agit des mêmes données, analysées différemment. Comment expliquer l'apparente contradiction entre les deux affirmations suivantes ?

affirmation 1 conditionnellement à la couleur de la victime, les meurtriers noirs sont plus condamnés que les meurtriers blancs,

affirmation 2 sur l'ensemble des données les meurtriers noirs sont moins condamnés que les meurtriers blancs.

C'est un paradoxe bien connu, appelé paradoxe de Simpson, qui s'explique par la notion de liaison conditionnelle : l'analyse du modèle log-linéaire indique que le modèle

approprié pour ces données est le modèle avec 2 interactions d'ordre 2 , (Couleur de la victime)×Condamnation et (Couleur de la victime)×(Couleur du meurtrier). Ce modèle implique que la peine de mort est liée à la couleur de la victime (les victimes blanches incitent davantage les jurys de Floride à la peine de mort) et que la couleur de la victime est liée à la couleur du meurtrier (les victimes blanches sont plutôt victimes de meurtriers blancs). Pour éviter de tomber dans le piège de ce paradoxe, il faut toujours commencer par analyser la table complète, non marginalisée. Remarquons que toute table de contingence est la marginale de tables plus complètes qui auraient pu être récoltées mais qui, en pratique, ne l'ont pas été. Donc le statisticien analyse toujours une table marginale, il peut donc être victime du paradoxe de Simpson, ce qui doit le rendre prudent dans ses conclusions.

On peut aussi analyser ces données en considérant les couleurs des victimes et des meurtriers comme des variables explicatives fixes et la condamnation à mort comme la variable réponse, avec une régression logistique et loi binomiale. On trouve que le seul facteur explicatif est la couleur de la victime, mais on passe à coté du lien entre la couleur de la victime et celle du meurtrier.

4.8.5 Programme SAS

```
data Radelet; input coulmeur $ coulvict $  condamn $  nombre @@;
lines;
b b oui 19 b b  non 132 b n oui 0 b n non 9
n b oui 11 n b  non 52 n n oui 6 n n non 97
 ;

proc genmod data=Radelet;
class coulmeur coulvict condamn ;
model nombre= coulmeur|coulvict|condamn @2 / dist=poisson type1 ;run;

proc genmod data=Radelet;
class coulmeur coulvict condamn ;
model nombre= coulmeur coulvict condamn coulmeur*coulvict
              coulvict*condamn / dist=poisson type1;
run;
```

Chapitre 5

Modèle mixte, modélisation de la variance

5.1 Modèle

Le modèle mixte permet de distinguer différentes sources de variabilité et de prendre en compte des corrélations entre les observations. De façon générale, il s'agit donc de passer du modèle linéaire gaussien vu au chapitre 1, p. 12 fondé sur les hypothèses d'indépendance et d'homoscédasticité

$$Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n)$$

à un modèle de la forme

$$Y \sim \mathcal{N}_n(X\theta, \Sigma) \tag{5.1}$$

qui s'affranchit de ces hypothèses. La généralisation apportée par le modèle linéaire mixte par rapport au modèle linéaire porte donc sur la matrice de variance Σ. L'espérance du vecteur des réponses Y conserve la même forme $X\theta$ que précédemment.

On peut immédiatement remarquer que si le vecteur des réponses Y est de dimension n, la matrice Σ est de dimensions $n \times n$ et comporte donc, puisqu'elle est symétrique, $n(n+1)/2$ paramètres. Le nombre de données disponibles interdit, par construction, de considérer des modèles dans lesquels ces $n(n+1)/2$ paramètres sont libres. Il faudra donc supposer que la matrice Σ a une certaine structure, gouvernée par un nombre restreint de paramètres appelés dans la suite *paramètres de variance*. On notera ψ le vecteur contenant ces paramètres. Les paramètres du modèle sont donc d'un côté θ pour l'espérance et ψ pour la variance.

La définition de la structure de dépendance constitue ainsi une modélisation de la matrice de variance. On en décrit ici trois modélisations classiques.

5.1.1 Composantes de la variance

On s'intéresse à la variabilité d'un caractère (ou "*trait*") quantitatif (comme la taille ou le poids) au sein d'une population. On souhaite notamment savoir si le caractère est fortement héritable ou non, c'est-à-dire si les individus issus d'un même ascendant sont plus semblables que ceux issus d'ascendants différents. Pour cela on considère un échantillon de m pères numérotés $i = 1, 2, \ldots, m$ chacun ayant eu n_i descendants

numérotés $j = 1, 2, \ldots, n_i$. On note Y_{ij} la valeur du trait pour le j-ème descendant du i-ème père.

Modèle. Les pères inclus dans ce dispositif ne présentent pas d'intérêt individuel, mais seulement en tant que représentants d'une population plus large de pères possibles. Pour prendre en compte le caractère aléatoire de leur sélection dans l'échantillon on propose le modèle suivant :

$$Y_{ij} = \mu + A_i + E_{ij} \tag{5.2}$$

dans lequel les termes A_i et E_{ij} sont tous supposés centrés, gaussiens et indépendants :

$$\{A_i\}_i \text{ i.i.d.}, \quad A_i \sim \mathcal{N}(0, \gamma^2) \qquad \text{et} \qquad \{E_{ij}\}_{i,j} \text{ i.i.d.}, \quad E_{ij} \sim \mathcal{N}(0, \sigma^2).$$

La variabilité entre les pères est donc mesurée par le paramètre de variance due à l'effet père γ^2. Ce modèle est illustré par l'exemple 6.1, p. 196. Un tel modèle est également appelé "modèle à effet aléatoire".

Structure de dépendance. L'effet aléatoire associé à chacun des pères introduit une corrélation entre les traits mesurés sur les descendants d'un même père. En effet, si dans ce modèle l'espérance a une forme très simple :

$$\mathbb{E}(Y_{ij}) = \mu + \mathbb{E}(A_i) + \mathbb{E}(E_{ij}) = \mu,$$

la variance de chaque observation est la somme de la variance due au père et de la variance résiduelle :

$$\mathbb{V}(Y_{ijk}) = \mathbb{V}(A_i) + \mathbb{V}(E_{ij}) = \sigma^2 + \gamma^2, \tag{5.3}$$

et les traits mesurés sur les descendants d'un même père i ne sont plus indépendants

$$\mathbb{C}\text{ov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \gamma^2 & \text{si } i = i', \\ 0 & \text{sinon.} \end{cases} \tag{5.4}$$

Ainsi, ce modèle suppose que les traits mesurés sur des individus non apparentés sont indépendants et que la corrélation au sein des descendants d'un même père est uniforme.

Ce modèle comporte deux paramètres de variance (aussi appelés composantes de la variance) : la variance de l'effet père γ^2 et la variance résiduelle σ^2, soit

$$\psi = \begin{bmatrix} \gamma^2 \\ \sigma^2 \end{bmatrix}.$$

Écriture matricielle. Comme dans le cas du modèle linéaire usuel, le modèle (5.2) peut s'écrire sous forme matricielle. L'espérance du vecteur Y contenant toutes les réponses peut s'écrire sous la forme $X\theta$, où θ est le vecteur contenant simplement le paramètre μ et la variance du vecteur des erreurs E s'écrit toujours $\sigma^2 I_n$ (cf. chapitre 1, p. 12). On complète cette écriture en définissant la matrice Z de dimensions $n \times m$, où m est le nombre de pères et de terme général

$$Z_{a,i} = \begin{cases} 1 & \text{l'individu } a = (i, j) \text{ est le descendant du père } i, \\ 0 & \text{sinon} \end{cases}$$

et le vecteur U de dimension m qui contient les effets aléatoires, le modèle (5.2) s'écrit

$$Y = X\theta + ZU + E. \tag{5.5}$$

En rappelant que les vecteurs E et U sont indépendants, gaussiens, centrés et que la variance du vecteur U vaut $\gamma^2 I_m$, on retrouve

$$\mathbb{E}(Y) = \mathbb{E}(X\theta + ZU + E) = X\theta, \tag{5.6}$$

$$\Sigma = \mathbb{V}(Y) = \mathbb{V}(X\theta + ZU + E) = Z\mathbb{V}(U)Z' + \mathbb{V}(E) = \gamma^2 ZZ' + \sigma^2 I_n. \tag{5.7}$$

On retrouve ainsi l'écriture générale du modèle linéaire mixte donnée en (5.1) où la matrice de variance Σ est diagonale par bloc

$$\Sigma = \begin{bmatrix} R & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & R \end{bmatrix} \quad \text{où} \quad R = \begin{bmatrix} \sigma^2+\gamma^2 & \gamma^2 & \cdots & \gamma^2 \\ \gamma^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma^2 \\ \gamma^2 & \cdots & \gamma^2 & \sigma^2+\gamma^2 \end{bmatrix} \tag{5.8}$$

qui est la représentation matricielle des variances et covariances données aux équations (5.3) et (5.4). Notons que les blocs diagonaux ne sont de mêmes dimensions que si le nombre de descendants n_i est le même pour chaque père. Sinon, ils sont chacun de dimensions $n_i \times n_i$.

5.1.2 Mesures répétées

On considère cette fois une expérience visant à évaluer l'effet d'un régime sur la prise de poids d'un animal au cours du temps. Plusieurs animaux (indicés par j) reçoivent chaque régime (noté i) et un animal ne reçoit qu'un régime au cours de l'expérience. On mesure le poids, noté Y_{ijt}, de chaque animal au bout de t semaines ($t = 1, \ldots, T$). On parle donc de mesures répétées, au cours du temps sur un même individu. De telles mesures sont aussi fréquemment appelées *données longitudinales*.

Pour analyser ces résultats, on veut tenir compte du fait que les mesures faites au cours du temps sur un même animal ne sont pas indépendantes. A cette fin, on pourrait utiliser le modèle

$$\mathbb{E}(Y_{ijt}) = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it}.$$

avec

$$\mathbb{C}\mathrm{ov}(Y_{ijt}, Y_{i'j't'}) = \begin{cases} \sigma^2\rho & \text{si } (i,j) = (i',j'), \\ 0 & \text{sinon.} \end{cases} . \tag{5.9}$$

Dans ce modèle, la covariance entre deux mesures faites aux temps t et t' sur un même modèle est constante, quel que soit l'intervalle de temps qui sépare les deux mesures.

Modèle. On propose ici un modèle qui prend en compte l'aspect cinétique de l'expérience et prévoit que la dépendance entre deux mesures dépend de l'intervalle de temps qui les sépare. Une telle dépendance n'admet pas de représentation simple sous la forme d'un effet aléatoire. On suppose donc que les poids sont gaussiens d'espérances respectives

$$\mathbb{E}(Y_{ijt}) = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it}.$$

Structure de dépendance. On suppose de plus que toutes les mesures ont la même variance

$$\mathbb{V}(Y_{ijt}) = \sigma^2$$

et que la covariance entre elles vaut

$$\mathbb{C}\text{ov}(Y_{ijt}, Y_{i'j't'}) = \begin{cases} \sigma^2\rho^{|t-t'|} & \text{si } (i,j) = (i',j'), \\ 0 & \text{sinon.} \end{cases} . \tag{5.10}$$

Cette structure suppose que les mesures faites sur des animaux différents sont indépendantes. On suppose de plus que $|\rho| < 1$, ce qui implique que celles faites sur un même animal sont d'autant moins corrélées que l'intervalle de temps est grand. Cette forme de covariance correspond à un processus auto-régressif d'ordre 1, généralement noté AR(1).

Ce modèle comporte deux paramètres de variance (aussi appelés composantes de la variance) : la corrélation temporelle ρ et la variance de chaque observation σ^2, soit

$$\psi = \begin{bmatrix} \rho \\ \sigma^2 \end{bmatrix}.$$

Ecriture matricielle. Du fait de l'indépendance entre les mesures obtenues sur des animaux différents, la matrice de variance Σ a la même forme diagonale par blocs que celle donnée en (5.8) mais le bloc R diffère :

$$R = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \cdots & \sigma^2\rho^{T-1} \\ \sigma^2\rho & \ddots & \ddots & \ddots & \vdots \\ \sigma^2\rho^2 & \ddots & \sigma^2 & \ddots & \sigma^2\rho^2 \\ \vdots & \ddots & \ddots & \ddots & \sigma^2\rho \\ \sigma^2\rho^{T-1} & \cdots & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & \ddots & \ddots & \ddots & \vdots \\ \rho^2 & \ddots & 1 & \ddots & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^{T-1} & \cdots & \rho^2 & \rho & 1 \end{bmatrix}.$$

Comme indiqué plus haut, il n'existe pas d'écriture simple de ce modèle sous la forme (5.5) et on recourt à l'écriture générale (5.1).

Il existe d'autres modélisations de la variance dans le cadre des mesures répétées, voir 6.3, p. 208

5.1.3 Spatial

On voudrait expliquer la teneur en matière organique (MO) des sols dans une région agricole, en fonction de caractéristiques physiques et chimiques. Pour cela on effectue des prélèvements en plusieurs sites nus que l'on analyse en laboratoire. On dispose donc de n mesures Y_i acquises aux sites $s_1, s_2, \ldots s_n$ repérés dans l'espace, et pour chaque site une série de covariables rassemblées dans un vecteur ligne x_i.

Modèle. Pour estimer l'effet des covariables, on pose un modèle de régression multiple (cf. chapitre 2, p. 47) :

$$\mathbb{E}(Y_i) = \sum_j \theta_j x_{ij} = x_i\theta.$$

Structure de dépendance. Là encore, on veut prendre en compte la dépendance due à l'éventuelle proximité spatiale entre les sites auxquels ont été effectuées les mesures. Pour cela, on note $d(i, i')$ la distance qui sépare les sites s_i et $s_{i'}$ et on pose

$$\mathbb{C}\text{ov}(Y_i, Y_{i'}) = \gamma^2 e^{-d(i,i')/\rho}, \qquad \text{si } i \neq i'$$

où ρ est appelée *portée*. Cette forme de covariance généralise au cas spatial la covariance définie en (5.10) pour le cas temporel. Plus ρ est grand, plus la portée de la corrélation spatiale est grande. On remarque que cette covariance n'est nulle que pour des sites infiniment distants.

De plus, on pose

$$\mathbb{V}(Y_i) = \sigma^2 + \gamma^2.$$

Le paramètre de variance supplémentaire σ^2 est parfois appelé 'effet pépite'.

Ce modèle comporte, lui, trois paramètres de variance (aussi appelé composantes de la variance) : la portée de la dépendance spatiale ρ, la variance spatiale γ^2 et la variance additionnelle de chaque observation σ^2, soit

$$\psi = \begin{bmatrix} \gamma^2 \\ \rho \\ \sigma^2 \end{bmatrix}.$$

Ecriture matricielle. Comme dans le cas des mesures répétées, il n'existe pas d'écriture simple en termes d'effets aléatoires. De plus, comme toutes les mesures sont dépendantes, la matrice Σ n'est plus diagonale par bloc et s'écrit :

$$\Sigma = \begin{bmatrix} \sigma^2 + \gamma^2 & & & \\ & \ddots & \gamma^2 e^{-d(i,i')/\rho} & \\ & \ddots & & \\ & \gamma^2 e^{-d(i,i')/\rho} & \ddots & \\ & & & \sigma^2 + \gamma^2 \end{bmatrix}.$$

De même, on utilise l'écriture générale (5.1) pour spécifier complètement le modèle.

5.2 Estimation

Il existe de nombreuses méthodes d'estimation des paramètres du le modèle linéaire mixte fondées sur des approches générales comme le maximum de vraisemblance, les moindres carrés ou la méthode des moments. Nous présentons ici les plus utilisées.

Comme on l'a vu à à la section 5.1, p. 166, la matrice de variance Σ possède généralement une structure donnée et sa valeur ne dépend que des paramètres de la variance contenus dans le vecteur ψ. L'estimation de Σ se ramène alors à celle de ψ. Ainsi, la matrice de variance des observations doit être notée $\Sigma(\psi)$ et son estimateur $\Sigma(\widehat{\psi})$. Pour alléger les notations, ces matrices seront notées Σ et $\widehat{\Sigma}$ quand cela ne crée pas de confusion.

5.2.1 Estimation de θ à ψ connu

Lien entre maximum de vraisemblance et moindres carrés généralisés

La vraisemblance associée au modèle (5.1) a la forme générale

$$\begin{aligned}\mathcal{L}(Y;\theta,\psi) &= \log \phi_n(Y; X\theta, \Sigma) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y-X\theta)'\Sigma^{-1}(Y-X\theta) \qquad (5.11)\end{aligned}$$

où $\phi_n(\cdot;\mu,\Sigma)$ désigne la densité de la loi normale n-dimensionnelle de vecteur d'espérance μ et de matrice de variance Σ, en se souvenant ici que $\Sigma = \Sigma(\psi)$.

Propriété 5.2.1. *Pour une matrice de variance Σ fixée, l'estimateur du maximum de vraisemblance de θ vérifie*

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{L}(Y;\theta,\psi) = \arg\min_{\theta}(Y-X\theta)'\Sigma^{-1}(Y-X\theta).$$

Cette propriété, immédiate d'après l'expression 5.11, p. 171, généralise l'équation (1.5, p. 18) vérifiée par l'estimateur de θ dans le modèle linéaire simple : l'estimateur du maximum de vraisemblance est égal à l'estimateur des moindres carrés généralisés. La généralisation provient simplement de l'utilisation de la matrice Σ^{-1} qui remplace la matrice identité qui intervient dans les moindres carrés ordinaires.

Estimation de θ à structure de variance connue

Les situations où la matrice de variance Σ est connue sont rares, mais il peut arriver qu'elle le soit à une constante multiplicative près, prenant la forme

$$\Sigma = \sigma^2 C$$

où C est une matrice de corrélation connue et σ^2 est un paramètre de variance à estimer. Une telle situation se présente par exemple pour l'étude, au sein d'une population d'individus apparentés, des variations d'un caractère dont l'héritabilité est connue ou pour des données organisées spatialement et dont la structure de corrélation spatiale est connue.

Dans ce cas, les propriétés de l'estimateur de θ se déduisent immédiatement de celles obtenues au chapitre 1, p. 12 pour le modèle linéaire simple en utilisant la propriété de linéarité des vecteurs gaussiens rappelée à la section 10.2.3, p. 307. On définit pour cela $C^{1/2}$ le résultat de la décomposition de Cholevsky inférieure de la matrice $C = C^{1/2}(C^{1/2})'$

$$Y \sim \mathcal{N}_n(X\theta, \sigma^2 C) \qquad \Longrightarrow \qquad C^{-1/2}Y \sim \mathcal{N}_n(C^{-1/2}X\theta, \sigma^2 I_n).$$

Ce modèle prend alors la forme du modèle linéaire classique

$$\widetilde{Y} \sim \mathcal{N}_n(\widetilde{X}\theta, \sigma^2 I_n)$$

en notant $\widetilde{Y} = C^{-1/2}Y$ et $\widetilde{X} = C^{-1/2}X$ et toutes les propriétés de l'estimateur

$$\widehat{\theta} = \left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}'\widetilde{Y} = \left(X'C^{-1}X\right)^{-1}X'C^{-1}Y = \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}Y \qquad (5.12)$$

restent valides.

5.2.2 Estimation conjointe de θ et ψ par maximum de vraisemblance

On a

$$\begin{aligned}\widehat{\psi},\widehat{\theta} &= \arg\max_{\psi,\theta} \mathcal{L}(Y;\theta,\psi) \\ &= \arg\min_{\psi,\theta} \left[n\log|\Sigma(\psi)| + (Y-X\theta)'\Sigma^{-1}(\psi)(Y-X\theta)\right]\end{aligned}$$

Donc

$$\frac{\partial\mathcal{L}(Y;\theta,\psi)}{\partial\theta} = 0 \Leftrightarrow X'\Sigma^{-1}(\psi)X\theta = X'\Sigma^{-1}(\psi)Y.$$

qui généralise le système (1.5) vu au chapitre 1, p. 12. Avec les mêmes limitations que celles discutées dans ce même chapitre, si la matrice $X'\Sigma^{-1}(\psi)X$ est inversible, l'estimateur du maximum de vraisemblance de θ vérifie $\theta = h(\psi)$ où h est la fonction

$$h(\psi) = \left(X'\Sigma^{-1}(\psi)X\right)^{-1} X'\Sigma^{-1}(\psi)Y. \tag{5.13}$$

On peut donc décomposer la maximisation de la vraisemblance en 2 étapes :

1. $\widehat{\psi} = \arg\min_{\psi} \left[n\log|\Sigma(\psi)| + (Y-Xh(\psi))'\Sigma^{-1}(\psi)(Y-Xh(\psi))\right]$
2. $\widehat{\theta} = h(\widehat{\psi})$

Le même type de procédure en 2 étapes est aussi utilisée par les autres méthodes d'estimation : on estime d'abord ψ puis on en déduit $\widehat{\theta} = h(\widehat{\psi})$. $\widehat{\theta}$ n'est l'estimateur du maximum de vraisemblance de θ que quand $\widehat{\psi}$ est aussi estimateur du maximum de vraisemblance de ψ.

Les estimateurs du maximum de vraisemblance bénéficient alors des propriétés générales de cette famille d'estimateurs : absence de biais asymptotique, variance asymptotique liée à l'information de Fisher, normalité asymptotique.

Ces propriétés sont cependant soumises à des conditions techniques qui peuvent ne pas être vérifiées. Dans le cas des paramètres de variance, la condition technique la plus souvent violée est que la vraie valeur du paramètre doit se situer à l'intérieur de son ensemble de définition. Cette condition implique notamment que la variance γ^2 d'un effet aléatoire ne peut pas être nulle ou que la corrélation temporelle ou spatiale ρ ne peut pas être égale à 1. Cette remarque a des conséquences théoriques en matière de tests. Elle a également des conséquences pratiques en matière d'estimation puisque l'intervalle de confiance asymptotique d'un paramètre de variance peut, le cas échéant, couvrir des valeurs négatives.

Optimisation de la fonction de vraisemblance. La log-vraisemblance peut être une fonction arbitrairement complexe des paramètres de variance et il n'existe pas de forme générale explicite des estimateurs du maximum de vraisemblance des paramètres de variance. Leur estimation repose donc le plus souvent sur des méthodes numériques. L'algorithme le plus fréquemment utilisé est celui de Newton-Raphson (présenté au chapitre 12, p. 322) qui fournit également une estimation de la matrice de variance asymptotique des paramètres de variance.

La mise en œuvre de cette optimisation demande des développements spécifiques pour chaque modèle, notamment le calcul des dérivées (premières et secondes) de la log-vraisemblance par rapport aux paramètres de variance. Les logiciels disponibles

proposent chacun une gamme plus ou moins large de modèles pour lesquels ces formules sont implémentées. En l'absence d'une telle possibilité, l'utilisateur doit avoir recours lui-même à une routine d'optimisation numérique pour effectuer l'inférence des paramètres de variance.

Paramétrisation du modèle. La fonction de vraisemblance dépend de la paramétrisation du modèle. Ainsi, la matrice de variance du modèle à composantes de la variance de la section 5.1.1, p. 166 peut s'écrire de deux façons équivalentes. Celle donnée à l'équation (5.8) distingue la variance γ^2 de l'effet aléatoire de la variance σ^2 résiduelle. De façon équivalente, on peut écrire

$$R = \sigma^2 \begin{bmatrix} 1 & & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} \qquad \text{avec} \qquad \rho = \frac{\gamma^2}{\sigma^2 + \gamma^2}.$$

Dans cette seconde écriture, les paramètres de variance (σ^2, ρ) remplacent (σ^2, γ^2).

Une telle re-paramétrisation du modèle peut avoir des conséquences pratiques en modifiant la concavité de la log-vraisemblance ou la position de l'optimum dans l'ensemble de définition, ce qui peut changer le comportement de l'algorithme d'optimisation numérique. Elle a aussi des conséquences théoriques puisque l'hypothèse d'absence d'effet aléatoire ($\gamma^2 = 0$) se transforme en absence de corrélation due à l'effet aléatoire ($\rho = 0$) et que la première se situe au bord de l'ensemble de définition du paramètre γ^2 alors que la seconde est au centre de l'ensemble de définition du paramètre ρ. Les hypothèses de la théorie du maximum de vraisemblance sont donc violées pour la première paramétrisation et valides pour la seconde.

Une alternative à l'algorithme de Newton-Raphson : l'algorithme EM

Le modèle à effet aléatoire (5.5) peut être vu comme un modèle à données incomplètes dans la mesure où l'effet aléatoire U est inobservé. Ce point de vue le place dans une catégorie plus large de modèles pour lesquels les estimateurs du maximum de vraisemblance peuvent être obtenus au moyen de l'algorithme "Espérance-Maximisation" (EM, [22]). Le principe général de cet algorithme consiste à définir deux vraisemblances :

- la vraisemblance observée (ou 'incomplète'), c'est-à-dire celle des données observées Y, soit $\mathcal{L}(Y; \theta, \psi)$ et
- la vraisemblance complète, c'est-à-dire la vraisemblance jointe des données observées Y et des effets aléatoires inobservées U, soit $\mathcal{L}(Y, U; \theta, \psi)$.

Ces deux vraisemblances sont liées par la décomposition suivante.

Propriété 5.2.2. *La vraisemblance des données observées se décompose en*

$$\mathcal{L}(Y; \theta, \psi) = \mathbb{E}[\mathcal{L}(Y, U; \theta, \psi)|Y] - \mathbb{E}[\mathcal{L}(U|Y; \theta, \psi)|Y].$$

Démonstration. Il suffit de rappeler que la log-vraisemblance est le logarithme d'une (densité de) probabilité et que donc (en omettant volontairement les paramètres θ et ψ

qui n'interviennent pas)

$$\begin{aligned} & \mathcal{L}(U|Y) &=& \mathcal{L}(Y,U) - \mathcal{L}(Y) \\ \Rightarrow \quad & \mathbb{E}[\mathcal{L}(Y)|Y] &=& \mathbb{E}[\mathcal{L}(Y,U)|Y] - \mathbb{E}[\mathcal{L}(U|Y)|Y] \end{aligned}$$

en prenant l'espérance conditionnellement à Y. Le résultat suit en remarquant que $\mathbb{E}[\mathcal{L}(Y)|Y] = \mathcal{L}(Y)$.

Il est important de noter que cette décomposition s'applique aussi bien à la vraisemblance classique qu'à la vraisemblance restreinte définie plus loin en section 5.2.3, p. 175.

L'intérêt de cette propriété vient de ce que la vraisemblance complète $\mathcal{L}(Y,U)$ est souvent plus facile à manier que la vraisemblance 'incomplète' $\mathcal{L}(Y)$: l'inférence serait plus simple si les effets aléatoires étaient observés.

Application au modèle à effets aléatoires. On considère le modèle (5.5), en notant D la matrice de variance du vecteur des effets aléatoires U et S celle des erreurs E (typiquement, $S = \sigma^2 I_n$). On déduit directement de ce modèle la loi conditionnelle du vecteur des réponses Y sachant le vecteur des effets aléatoires U :

$$Y|U \sim \mathcal{N}_n\left(X\theta + ZU, S\right)$$

On en déduit également la loi jointe de ces deux vecteurs qui est la loi normale

$$\left[\begin{array}{c} Y \\ U \end{array}\right] \sim \mathcal{N}_{n+m}\left(\left[\begin{array}{c} X\theta \\ 0 \end{array}\right], \left[\begin{array}{cc} \Sigma & ZD \\ DZ' & D \end{array}\right]\right), \tag{5.14}$$

en notant m la dimension du vecteur U et $\Sigma = S + ZDZ'$ (cf. eq. (5.6)). En utilisant la propriété des vecteurs gaussiens rappelée à la section 10.2.4, p. 308, on peut écrire la loi conditionnelle du vecteur des effets aléatoires U sachant le vecteur des réponses Y :

$$U|Y \quad \sim \quad \mathcal{N}_m\left(DZ'\Sigma^{-1}(Y - X\theta), D - DZ'\Sigma^{-1}ZD\right) \tag{5.15}$$

On en déduit la forme de la vraisemblance complète

$$\begin{aligned} \mathcal{L}(Y,U;\theta,\psi) &= \mathcal{L}(U;\theta,\psi) + \mathcal{L}(Y|U;\theta,\psi) \\ &= \log\phi(U;0,D) + \log\phi(Y;X\theta+ZU,S) \\ &= \text{ct} - \frac{1}{2}\left\{\log|D| + U'D^{-1}U + \log|S|\right\} \\ &+ \frac{1}{2}\left\{(Y - X\theta - ZU)'S^{-1}(Y - X\theta - ZU)\right\} \\ &= \text{ct} - \frac{1}{2}\left\{\log|D| + \log|S| + (Y - X\theta)'S^{-1}(Y - X\theta)\right\} \\ &\qquad +(Y - X\theta)'S^{-1}ZU - \frac{1}{2}\left\{U'D^{-1}U + U'Z'S^{-1}ZU\right\} \end{aligned}$$

où "ct" désigne une constante qui ne dépend ni de θ, ni de ψ. Cette vraisemblance fait intervenir des termes linéaires et quadratiques en U dont il faut calculer l'espérance

conditionnelle à Y pour pouvoir appliquer la propriété 5.2.2, p. 173. En utilisant les lois données en (5.15), on obtient les quantités nécessaires :

$$\begin{aligned}
\mathbb{E}(U|Y) &= DZ'\Sigma^{-1}(Y - X\theta), \\
\mathbb{E}(U'D^{-1}U|Y) &= (Y - X\theta)'\Sigma^{-1}ZDD^{-1}DZ'\Sigma^{-1}(Y - X\theta) \qquad (5.16)\\
&+ \operatorname{tr}[D^{-1}(D - DZ'\Sigma^{-1}ZD)] \\
&= (Y - X\theta)'\Sigma^{-1}ZDZ'\Sigma^{-1}(Y - X\theta) + \operatorname{tr}(I_n - Z'\Sigma^{-1}ZD), \qquad (5.17)\\
\mathbb{E}[U'(Z'S^{-1}Z)U|Y] &= (Y - X\theta)'\Sigma^{-1}ZD(Z'S^{-1}Z)DZ'\Sigma^{-1}(Y - X\theta) \\
&\quad +\operatorname{tr}[(Z'S^{-1}Z)(D - DZ'\Sigma^{-1}ZD)].
\end{aligned}$$

Dans de nombreux cas, ces formules se calculent facilement du fait que les matrices S et D sont diagonales.

L'algorithme EM consiste à alterner jusqu'à convergence les étapes 'E' et 'M'.

Etape E : Calcul des moments conditionnels des effets aléatoires U sachant les observations Y avec les formules (5.16).

Etape M : Maximisation du premier terme de la décomposition de la propriété 5.2.2, p. 173 par rapport aux paramètres θ et ψ.

Remarque. Il est important de noter que

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{L}(Y; \theta, \widehat{\psi})$$

est l'estimateur du maximum de vraisemblance de θ si et seulement si $\widehat{\psi}$ est lui-même l'estimateur du maximum de vraisemblance de ψ. Nous verrons dans la section suivante que l'estimateur du maximum de vraisemblance de ψ est biaisé. On pourra donc lui préférer un autre estimateur si l'objectif est de bien estimer les paramètres de variance. Cependant, pour pouvoir appliquer les propriétés des estimateurs du maximum de vraisemblance à $\widehat{\theta}$, il faut utiliser le même estimateur pour ψ, même si celui-ci est biaisé.

5.2.3 Autres méthodes d'estimation de Σ

Maximum de vraisemblance restreint

Les estimateurs du maximum de vraisemblance sont asymptotiquement sans biais, mais le biais pour une taille d'échantillon finie peut être important. Ce biais est en partie dû au fait qu'on doit estimer simultanément le paramètre de variance ψ et le paramètre d'espérance θ.

Cas du modèle linéaire classique. Le modèle linéaire étudié au chapitre 1, p. 12 s'écrit $Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n)$ et sa fonction de vraisemblance est

$$\mathcal{L}(Y; \theta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\theta\|^2.$$

En utilisant la forme de $\widehat{\theta}$ donnée en (5.13) avec $\widehat{\Sigma} = \widehat{\sigma}^2 I_n$, on voit que

$$\mathcal{L}(Y; \widehat{\theta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}SCR \qquad \text{avec } SCR = \|Y - X\widehat{\theta}\|^2$$

dont le maximum est atteint en

$$\widehat{\sigma}^2 = \arg\max_{\sigma^2} \mathcal{L}(Y; \widehat{\theta}, \sigma^2) = \frac{1}{n} SCR.$$

Or on a vu au chapitre 1, p. 12 que l'espérance de SCR vaut $\mathbb{E}(SCR) = (n-r)\sigma^2$, en notant r le rang de la matrice X, voir 1.17, p. 26. L'estimateur du maximum de vraisemblance a donc un biais de

$$\mathbb{E}(\widehat{\sigma}^2) - \sigma^2 = \mathbb{E}\left(\frac{SCR}{n}\right) - \sigma^2 = \left(\frac{n-r}{n} - 1\right)\sigma^2 = -\frac{r}{n}\sigma^2$$

et ce biais est proportionnel au nombre de paramètres d'espérance indépendants, à savoir r.

Cette observation conduit à proposer d'estimer les paramètres de variance ψ indépendamment des paramètres d'espérance θ. On trouvera notamment une introduction à cette approche dans [57]. C'est le principe du maximum de vraisemblance restreint. Le principe général consiste à choisir une matrice T (de dimensions $(n-q) \times n$) telle que $TX = 0$ et à mener l'inférence à partir des données transformées

$$\widetilde{Y} = TY$$

dont la loi est centrée :

$$Y \sim \mathcal{N}_n(X\theta, \Sigma) \qquad \Rightarrow \qquad \widetilde{Y} \sim \mathcal{N}_{n-q}(TX\theta, T\Sigma T') = \mathcal{N}_{n-q}(0, \widetilde{\Sigma})$$

en notant $\widetilde{\Sigma} = \widetilde{\Sigma}(\psi) = T\Sigma(\psi)T'$. Ainsi la loi, et donc la vraisemblance, des données transformées $\widetilde{Y}$ dépend de ψ mais plus de θ.

Définition 5.2.1. *Soit T une matrice telle que $TX = 0$ et $\widetilde{Y} = TY$. L'estimateur du maximum de vraisemblance restreint (ReML) de ψ fondé sur T est*

$$\widehat{\psi} = \arg\max_{\psi} \mathcal{L}(\widetilde{Y}; \psi).$$

La vraisemblance $\mathcal{L}(\widetilde{Y}; \psi) = \log \phi_{n-q}(\widetilde{Y}; 0, \widetilde{\Sigma}(\psi))$ est appelée vraisemblance restreinte.

L'estimateur du maximum de vraisemblance restreint est donc un estimateur du maximum de vraisemblance et en partage donc les propriétés. La définition de cet estimateur fait intervenir la matrice T pour laquelle le choix le plus fréquent est le projecteur sur l'espace orthogonal à l'espace engendré par les effets fixes soit, en notant Π le projecteur sur l'espace engendré par les effets fixes $L(X)$:

$$T = I_n - \Pi = I_n - X(X'X)^{-}X'.$$

Cependant, ce choix n'a pas de conséquence sur l'estimateur du fait de la propriété suivante.

Propriété 5.2.3. *L'estimateur du maximum de vraisemblance restreint $\widehat{\psi}$ ne dépend pas du choix de la matrice T, à condition qu'elle soit de rang $n-r$ et que et que $TX = 0$.*

Preuve. Ce résultat provient du fait que la fonction de vraisemblance restreinte $\mathcal{L}(\widetilde{Y}, \psi)$ ne dépend pas de T. En effet, pour toute T de rang $n - r$ telle que $TX = 0$, on a

$$-2\mathcal{L}(\widetilde{Y}, \psi) = \log|\Sigma| + \log|X'\Sigma X| + Y'\left[\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X\Sigma^{-1}\right]Y + \text{cst}.$$

On trouvera une démonstration de ce résultat dans [3], lemme 2.4.1. Ainsi, si la fonction de vraisemblance restreinte est invariante au choix de T, le lieu de son maximum l'est également.

Méthode des moments

La méthode des moments est en fait la méthode la plus ancienne pour définir des estimateurs sans biais des composantes de la variance. Le principe général consiste à projeter le vecteur Y sur une série de sous espaces, de telle façon que l'espérance des normes de ces vecteurs (appelée sommes de carrés) soient des fonctions linéaires des paramètres de variance. Des estimateurs de ces paramètres sont obtenus en résolvant un systèmes d'équations linéaires faisant intervenir les différentes sommes de carrés.

On choisit ici d'illustrer cette méthode sur un exemple. On étudie la variabilité d'une caractéristique Y mesurée sur des feuilles d'arbre, échantillonnés dans des parcelles elle-même réparties dans différents sites. On note $Y_{ijk\ell}$ la mesure obtenue sur la feuille ℓ de l'arbre k de la parcelle j dans le site i et on propose le modèle hiérarchique

$$Y_{ijk\ell} = \mu_i + A_{ij} + B_{ijk} + E_{ijk\ell}. \tag{5.18}$$

La version matricielle de ce modèle est

$$Y = X\theta + Z_A A + Z_B B + E.$$

On suppose de plus que les vecteurs d'effets aléatoires A, B et E sont indépendants de lois respectives

$$A \sim \mathcal{N}_n(0, \gamma_A^2 I_n), \qquad B \sim \mathcal{N}_n(0, \gamma_B^2 I_n), \qquad E \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

où γ_A^2 représente la part de variance due à la parcelle, γ_B^2 celle due à l'arbre et σ^2 la variance résiduelle. On considère ici que l'effet du site est fixe dans l'hypothèse où on souhaite, par exemple, comparer entre eux les sites étudiés.

La structure hiérarchique de ce modèle induit l'emboîtement des espaces engendrés par les différentes matrices :

$$L(X) \subset L(A) \subset L(B) \subset L(I_n) = \mathbb{R}^n.$$

On définit les projecteurs Π_X, Π_A et Π_B sur les sous-espaces respectifs $L(X)$, $L(A)$ et $L(B)$ et les différences entre les projections successives de Y sur ces sous-espaces :

$$\begin{array}{rcccl} Y_1 & = & (I_n - \Pi_B)Y & = & (I_n - \Pi_B)E, \\ Y_2 & = & (\Pi_B - \Pi_A)Y & = & (\Pi_B - \Pi_A)(Z_B B + E), \\ Y_3 & = & (\Pi_A - \Pi_X)Y & = & (\Pi_A - \Pi_X)(Z_A A + Z_B B + E). \end{array} \tag{5.19}$$

Ces trois vecteurs sont d'espérances nulles et de variances

$$\begin{aligned}\mathbb{V}(Y_1) &= \sigma^2(I_n - \Pi_B),\\ \mathbb{V}(Y_2) &= (\Pi_B - \Pi_A)(\gamma_B^2 Z_B' Z_B + \sigma^2 I_n)(\Pi_B - \Pi_A),\\ \mathbb{V}(Y_3) &= (\Pi_A - \Pi_X)(\gamma_A^2 Z_A' Z_A + \gamma_B^2 Z_B' Z_B + \sigma^2 I_n)(\Pi_A - \Pi_X).\end{aligned} \tag{5.20}$$

Le choix des projections sur les différents sous-espaces permet de faire apparaître progressivement les paramètres de variance. En se souvenant que, pour un vecteur centré U, $\mathbb{E}(\|U\|^2) = \mathrm{tr}[\mathbb{V}(U)]$, on déduit des équations ci-dessus le système triangulaire

$$\begin{aligned}\mathbb{E}(\|Y_1\|^2) &= c_{1E}\,\sigma^2,\\ \mathbb{E}(\|Y_2\|^2) &= c_{2B}\,\gamma_B^2 + c_{2E}\,\sigma^2,\\ \mathbb{E}(\|Y_3\|^2) &= c_{3A}\,\gamma_A^2 + c_{3B}\,\gamma_B^2 + c_{3E}\,\sigma^2\end{aligned} \tag{5.21}$$

dont les coefficients sont donnés par les traces des matrices intervenant dans les équations (5.20) :

$$\begin{aligned} c_{1E} &= \mathrm{tr}(I_n - \Pi_B) & c_{2B} &= \mathrm{tr}[(\Pi_B - \Pi_A)Z_B' Z_B(\Pi_B - \Pi_A)]\\ c_{2E} &= \mathrm{tr}(\Pi_B - \Pi_A) & c_{3A} &= \mathrm{tr}[(\Pi_A - \Pi_X)Z_A' Z_A(\Pi_A - \Pi_X)]\\ c_{3B} &= \mathrm{tr}[(\Pi_A - \Pi_X)Z_B' Z_B(\Pi_A - \Pi_X)] & c_{3E} &= \mathrm{tr}(\Pi_A - \Pi_X).\end{aligned}$$

On définit finalement les estimateurs $\widehat{\gamma}_A^2$, $\widehat{\gamma}_B^2$ et $\widehat{\sigma}^2$ comme les solutions du système (5.21) dans lequel les espérances des normes sont remplacées par les sommes de carrés qui constituent leurs versions empiriques :

$$\begin{aligned}\widehat{\sigma}^2 &= \|Y_1\|^2 \,/\, c_{1E},\\ \widehat{\gamma}_B^2 &= \left(\|Y_2\|^2 - c_{2E}\,\widehat{\sigma}^2\right) / c_{2B},\\ \widehat{\gamma}_A^2 &= \left(\|Y_3\|^2 - c_{3B}\,\widehat{\gamma}_B^2 - c_{3E}\,\widehat{\sigma}^2\right) / c_{3A}.\end{aligned} \tag{5.22}$$

Remarques.

— Les estimateurs ainsi obtenus sont sans biais par construction, du fait de l'équation (5.21).

— Ces estimateurs peuvent éventuellement être négatifs alors qu'ils estiment des variances. Une pratique courante consiste à tronquer leur valeur à 0, mais cette troncature induit un biais positif.

Cas d'un dispositif équilibré

Dans le cas particulier d'un dispositif équilibré, la méthode des moments aboutit à des estimateurs explicites. On reprend ici le modèle (5.18), p. 177 pour lequel on considère I sites dans chacun desquels on échantillonne J parcelles, puis K arbres par parcelle et enfin L feuilles par arbre. On dispose donc au total de $n = IJKL$ observations. Les formules des sommes de carrés se fondent alors toutes sur des moyennes ou des différences de moyennes.

En effet, comme on l'a vu à la section 2.4, p. 70, les projections sur les sous-espaces engendrés par des facteurs qualitatifs reviennent à calculer les moyennes au sein de chaque niveau du facteur. Ainsi, la projection sur l'espace $L(B)$ engendré par les arbres fait apparaître les moyennes par arbre $Y_{ijk\bullet}$ et celle sur l'espace $L(X)$ engendré par les

sites fait apparaître les moyennes par sites $Y_{i\bullet\bullet\bullet}$. En reprenant les sommes de carrés définies en (5.19), p. 177, on obtient ainsi

$$\begin{aligned}
\|Y_1\|^2 &= \|(I_n - \Pi_B)Y\|^2 = \sum_{i,j,k,\ell} (Y_{ijk\ell} - Y_{ijk\bullet})^2, \\
\|Y_2\|^2 &= \|(\Pi_B - \Pi_A)Y\|^2 = \sum_{i,j,k,\ell} (Y_{ijk\bullet} - Y_{ij\bullet\bullet})^2 = L\sum_{i,j,k}(Y_{ijk\bullet} - Y_{ij\bullet\bullet})^2, \\
\|Y_3\|^2 &= \|(\Pi_A - \Pi_X)Y\|^2 = \sum_{i,j,k,\ell} (Y_{ij\bullet\bullet} - Y_{i\bullet\bullet\bullet})^2 = KL\sum_{i,j}(Y_{ij\bullet\bullet} - Y_{i\bullet\bullet\bullet})^2.
\end{aligned}$$

Il reste alors à calculer les coefficients c intervenant dans le système triangulaire (5.21), p. 178. Le coefficient c_{1E} est égal à la trace du projecteur $I_n - \Pi_B$ où I_n est la matrice identité de dimension n et Π_B est le projecteur sur l'espace $L(B)$ de dimension IJK engendré par les arbres. $I_n - \Pi_B$ est donc le projecteur sur l'orthogonal de l'espace $L(B)$. La trace d'un projecteur étant égale à la dimension de l'espace sur lequel il projette, on obtient

$$c_{1E} = n - IJK = IJKL - IJK = IJK(L-1).$$

Par des calculs analogues, on obtient

$$c_{2E} = IJK - IJ = IJ(K-1), \qquad c_{3E} = IJ - I = I(J-1).$$

Les autres coefficients font intervenir des matrices de la forme $(\Pi_B - \Pi_A)Z_B'$. La multiplication par la matrice Z_B (resp. Z_A) fait apparaître des blocs de coordonnées égales de longueur L (resp. KL). L'autre terme constitue, en l'occurrence, la différence entre les projections sur les espaces $L(A)$ et $L(B)$ de vecteurs de l'espace $L(B)$. On part ainsi d'un espace à IJK dimensions pour aboutir à un espace de IJ dimension. On obtient au total

$$c_{2B} = L \times (IJK - IJ) = IJL(K-1).$$

Par des calculs analogues, on obtient

$$c_{3A} = KL \times (IJ - I) = IKL(J-1), \qquad c_{3B} = L \times (IJ - I) = IL(J-1).$$

On obtient les estimateurs en appliquant finalement les formules (5.22), p. 178 :

$$\widehat{\sigma}^2 = \frac{\|Y_1\|^2}{IJK(L-1)}, \tag{5.23}$$

$$\widehat{\gamma}_B^2 = \frac{\|Y_2\|^2 - IJ(K-1)\widehat{\sigma}^2}{IJL(K-1)} = \frac{\|Y_2\|^2}{IJL(K-1)} - \frac{\widehat{\sigma}^2}{L}, \tag{5.24}$$

$$\widehat{\gamma}_A^2 = \frac{\|Y_3\|^2 - IL(J-1)\widehat{\gamma}_B^2 - I(J-1)\widehat{\sigma}^2}{IKL(J-1)} = \frac{\|Y_3\|^2}{IKL(J-1)} - \frac{\widehat{\gamma}_B^2}{K} - \frac{\widehat{\sigma}^2}{KL} \tag{5.25}$$

Formulation alternative. On peut établir directement les formules précédentes, sans passer par les projecteurs orthogonaux, en appliquant le modèle (5.18), p. 177 aux moyennes par arbre $Y_{ijk\bullet}$ et par parcelle $Y_{ij\bullet\bullet}$:

$$\begin{array}{rclcrcl} Y_{ijk\bullet} & = & \mu_i + A_{ij} + F_{ijk}, & \text{avec} & F_{ijk} & = & B_{ijk} + E_{ijk\bullet} \\ Y_{ij\bullet\bullet} & = & \mu_i + G_{ij}, & \text{avec} & G_{ij} & = & A_{ij} + B_{ij\bullet} + E_{ij\bullet\bullet}. \end{array}$$

Dans ces modèles, les termes résiduels F_{ijk} et G_{ij} sont indépendants, centrés, gaussiens et de variance respectives $\mathbb{V}(F_{ijk}) = \gamma_B^2 + \sigma^2/L$ et $\mathbb{V}(G_{ij}) = \gamma_A^2 + \gamma_B^2/K + \sigma^2/KL$. Le caractère équilibré du dispositif intervient ici en assurant que les termes résiduels de chacun de ces modèles ont bien une variance constante, ce qui ne serait pas le cas si le nombre de feuilles par arbre ou d'arbres par parcelle variaient.

Le modèle sur les moyennes $Y_{ijk\bullet}$ (resp. $Y_{ij\bullet\bullet}$) porte sur IJK (resp. IJ) 'observations' et la somme de carrés $\|Y_2\|^2/L$ (resp. $\|Y_3\|^2/KL$) constitue sa somme de carrés résiduelle au sens défini au chapitre 1, p. 12. (Il convient de diviser par L et KL car chaque moyenne $Y_{ijk\bullet}$ est 'observée' L fois et chaque $Y_{ij\bullet\bullet}$ l'est KL fois.) On obtient ainsi directement que $\|Y_2\|^2/L(IJK - IJ)$ est un estimateur sans biais de $\gamma_B^2 + \sigma^2/L$ et que $\|Y_2\|^2/KL(IJ - I)$ est un estimateur sans biais de $\gamma_A^2 + \gamma_B^2/K + \sigma^2/KL$, ce qui est équivalent aux formules d'estimation données en (5.23), p. 179.

Estimateur "sandwich"

Il est souvent nécessaire de calculer la variance de l'estimateur $\widehat{\theta}$ donnée à l'équation (5.12) afin de déterminer des intervalles de confiance ou de tester certaines combinaisons de paramètres. En utilisant la propriété 10.1.1, p. 306, on obtient

$$\begin{aligned} \mathbb{V}(\widehat{\theta}) &= \mathbb{V}\left[\left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y\right] \\ &= \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}\mathbb{V}(Y)\Sigma^{-1}X\left(X'\Sigma^{-1}X\right)^{-1}. \end{aligned} \tag{5.26}$$

Selon le modèle (5.1), $\mathbb{V}(Y)$ vaut précisément Σ et la variance de $\widehat{\theta}$ se simplifie donc en

$$\mathbb{V}(\widehat{\theta}) = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}X\left(X'\Sigma^{-1}X\right)^{-1} = \left(X'\Sigma^{-1}X\right)^{-1}. \tag{5.27}$$

La forme de la variance de $\widehat{\theta}$ dépend donc de la matrice Σ qui rend compte de la structure de dépendance spécifiée par le modèle. Elle est donc sensible à une mauvaise spécification de cette structure de dépendance et peut donc donner des résultats biaisés si $\mathbb{V}(Y)$ est différente de Σ.

Lorsque Σ est inconnue, l'estimateur $\widehat{\theta}$ donné à l'équation (5.12) devient

$$\widehat{\theta} = \left(X'\widehat{\Sigma}^{-1}X\right)^{-1} X'\widehat{\Sigma}^{-1}Y,$$

où $\widehat{\Sigma}$ peut être obtenue par une des méthodes décrites précédemment dans cette section. La formule (5.26) devient alors

$$\left(X'\widehat{\Sigma}^{-1}X\right)^{-1} X'\widehat{\Sigma}^{-1}\widehat{\mathbb{V}}(Y)\widehat{\Sigma}^{-1}X\left(X'\widehat{\Sigma}^{-1}X\right)^{-1} \tag{5.28}$$

et, si on choisit d'estimer $\mathbb{V}(Y)$ par $\widehat{\Sigma}$ la formule (5.27) devient $(X'\widehat{\Sigma}^{-1}X)^{-1}$.

Dans le cas de mesures répétées (cf section 5.1.2, p. 168) où la variance de Y est diagonale par bloc (cf (5.8)), on peut obtenir un estimateur direct de $\mathbb{V}(Y)$, indépendant de la forme de dépendance Σ spécifiée par le modèle. Ainsi, en reprenant l'exemple de la section 5.1.2, p. 168 et en notant I le nombre de régimes, n_i le nombre d'animaux ayant reçu le régime i et $N = \sum_{i=1}^{I} n_i$ le nombre total d'animaux, on peut définir un estimateur consistant de la matrice R

$$\widehat{R} = [\widehat{r}_{tt'}] \qquad \text{où} \quad \widehat{r}_{tt'} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ijt} - Y_{i\bullet t})(Y_{ijt'} - Y_{i\bullet t'}).$$

On obtient ainsi un estimateur consistant de $\mathbb{V}(Y)$ en posant

$$\widehat{\mathbb{V}}(Y) = \begin{bmatrix} \widehat{R} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \widehat{R} \end{bmatrix}.$$

On estime alors la variance de $\widehat{\theta}$ par l'estimateur dit "sandwich" donnée à l'équation (5.28). [59] et [28] donnent une interprétation de cet estimateur en terme de pseudo-vraisemblance et montrent qu'il possède de meilleures propriétés que l'estimateur classique $(X'\widehat{\Sigma}^{-1}X)^{-1}$. Une illustration des vertus de cet estimateur dans le cadre du modèle linéaire généralisé est présentée dans [10].

5.2.4 Prédiction des effets aléatoires

On a vu plus haut que les modèles à effets aléatoires de la forme (5.5) permettent de modéliser certaines structures de corrélations entre les observations. Cependant, dans cette modélisation, l'effet aléatoire U peut lui-même donner lieu à une interprétation. Ainsi, dans le modèle (5.2) proposé pour étudier l'héritabilité génétique d'un caractère, l'effet A_i est interprété comme la *valeur génétique (additive)* du père numéro i. On peut donc s'intéresser à la prédiction du vecteur U. On réserve le terme d'*estimation* à un paramètre fixe comme θ. Dans le cas d'un effet aléatoire, on parle de *prédiction*.

Loi conditionnelle de l'effet aléatoire. On se propose généralement de prédire l'effet aléatoire U par son espérance conditionnellement au vecteur des observations Y, c'est-à-dire par $\mathbb{E}(U|Y)$. Le calcul de cette espérance se fonde sur la loi jointe de ces deux vecteurs déjà donnée à l'équation (5.14) et qu'on rappelle ici :

$$\begin{bmatrix} Y \\ U \end{bmatrix} \sim \mathcal{N}_{n+m}\left(\begin{bmatrix} X\theta \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & ZD \\ DZ' & D \end{bmatrix} \right)$$

où $\Sigma = ZDZ' + S$, en notant $D = \mathbb{V}(U)$ et $S = \mathbb{V}(E)$. La loi conditionnelle du vecteur de U donnée à l'équation (5.15) permet d'obtenir

$$\mathbb{E}(U|Y) = DZ'\Sigma^{-1}(Y - X\theta). \tag{5.29}$$

Formule de Henderson. La formule de l'espérance conditionnelle $\mathbb{E}(U|Y)$ donnée en (5.29) nécessite d'inverser la matrice Σ. Cette matrice étant de dimension $n \times n$, le temps de calcul nécessaire à son inversion peut donc être important, voire rédhibitoire dans certaines applications. Une formule due à Henderson [31] (voir aussi [52]) permet de réduire drastiquement la complexité de ce calcul. En effet,

$$W = S^{-1}\left[I_n - Z\left(Z'S^{-1}Z + D^{-1}\right)^{-1}Z'S^{-1}\right]$$

est l'inverse de $\Sigma = ZDZ' + S$. (Il suffit pour s'en convaincre de vérifier que $\Sigma W = W\Sigma = I_n$).
L'intérêt de cette nouvelle version est qu'elle ne requiert que l'inversion des matrices D et S qui ont souvent des formes simples (par exemple diagonales) ou des dimensions plus faibles que Σ. Ainsi, dans le modèle (5.2), D vaut $\gamma^2 I_n$ et S vaut $\sigma^2 I_n$. On obtient alors

$$W = \sigma^{-2}\left[I_n - Z\left(\sigma^{-2}Z'Z + \gamma^{-2}I_n\right)^{-1}Z'\sigma^{-2}\right]$$

qui nécessite seulement l'inversion de $(\sigma^{-2}Z'Z + \gamma^{-2}I_n)$ dont la dimension est celle du vecteur U, soit en l'occurrence, le nombre de pères.

Propriété (BLUP). En pratique, les paramètres d'espérance (θ) et de variance (ψ) doivent être estimés. Soit le prédicteur

$$\widehat{U} = DZ'\Sigma^{-1}(Y - X\widehat{\theta}),$$

où $\widehat{\theta}$ est l'estimateur de θ donné en (5.12) et obtenu à ψ connu. Henderson ([32]) montre que ce prédicteur est le meilleur prédicteur linéaire sans biais (*Best Linear Unbiased Predictor* = 'BLUP'). Plus précisément, parmi tous les estimateurs sans biais d'une combinaison estimable $\phi = c'\theta + d'U$ et de la forme $\widehat{\phi} = b'Y$ (linéaire, donc), le prédicteur $c'\widehat{\theta} + d'\widehat{U}$ est de variance minimale. Ce résultat suppose cependant que les paramètres de variance soient connus. On trouvera une présentation de cet estimateur ainsi que des illustrations dans [57].

5.3 Tests

Comme observé dans les parties précédentes, si l'écriture d'un modèle mixte ne comporte pas de difficulté supplémentaire par rapport à un modèle linéaire simple, l'introduction d'une structure de covariance peut sensiblement compliquer l'inférence statistique. Ces difficultés se retrouvent dans la problématique des tests d'hypothèse, portant sur un effet fixe ou sur un effet aléatoire (c'est-à-dire sur la variance associée à cet effet). Dans le cadre du modèle linéaire classique présenté au chapitre 1, la procédure de test de la Proposition 1.4.2, p. 29 pour une combinaison de paramètres possède deux propriétés remarquables : c'est une procédure de test exacte (la loi de la statistique de test sous H_0 est parfaitement connue) et optimale (au sens de Neyman-Pearson, voir [54] pour plus de détails). Ces deux propriétés garantissent à la fois que le test peut être réalisé en pratique, et qu'il n'existe pas de test préférable à celui proposé. En conséquence, cette procédure de test est (ou devrait être !) universellement utilisée et implémentée par les différents logiciels - en particulier par les deux logiciels employés dans ce livre.

Dans le cadre du modèle mixte, il n'existe pas de procédure de test optimale dans le cas général, bien qu'il soit souvent possible de déterminer des procédures de test optimales pour un modèle, un plan d'expérience et une hypothèse donnés. De même il n'est pas possible de proposer une procédure de test exacte quelle que soit l'hypothèse testée. De cette absence de procédure de test systématiquement optimale, il résulte une multitude de résultats théoriques portant sur l'optimalité locale de procédures de tests, dont la liste est bien trop longue pour figurer dans ce chapitre d'introduction au modèle mixte. Le lecteur intéressé pourra consulter [37]. Par ailleurs, en l'absence de stratégie de test universelle, les procédures de test implémentées diffèrent d'un logiciel à l'autre, et il est possible qu'une même analyse réalisée avec deux logiciels différents amène à tirer des conclusions différentes concernant la significativité de certains paramètres du modèle.

Dans ces conditions, nous nous bornerons dans cette partie à l'étude de quelques procédures de test sur la structure de covariance d'une part, et sur les effets fixes du modèle d'autre part. En particulier, pour les effets fixes, seules quelques procédures de test exactes ou approchées seront présentées. Nous verrons par ailleurs au chapitre 6, p. 196 à travers le traitement de différents exemples que l'utilisation des procédures de test implémentées par défaut dans les logiciels statistiques ne peut être considérée comme une garantie systématique d'une analyse bien menée.

5.3.1 Test sur les paramètres de variance-covariance

Nous considérons premièrement les tests portant sur la structure de covariance, c'est-à-dire portant sur un ou plusieurs paramètres de la matrice de variance. Tester la nullité (ou plus généralement la valeur) d'un paramètre de covariance peut représenter un objectif en soi. Dans l'exemple 5.1.1, p. 166, la question d'intérêt porte sur l'héritabilité d'un caractère quantitatif, que l'on peut définir ici comme la corrélation entre les mesures du caractère d'intérêt réalisées sur deux descendants issus d'un même père. Il est aisé de vérifier que cette corrélation s'écrit

$$\frac{\gamma^2}{\gamma^2 + \sigma^2}.$$

Ainsi, tester $H_0 : \{\gamma^2 = 0\}$ reviendra à tester l'absence d'héritabilité pour le caractère conditionné. Alternativement, le test d'intérêt peut porter sur l'un des paramètres de moyenne du modèle, comme dans l'exemple du paragraphe 5.1.2, p. 168 où l'objectif est de mettre en évidence l'impact du régime sur la prise de poids. L'hypothèse à tester portera soit sur les effets principaux ($H_0 : \{\gamma_t = \gamma_{t'}, \forall t, t'\}$) soit sur l'interaction. Dans le cas du test portant sur les effets principaux, nous verrons (cf exemple 6.3, p. 208) que la statistique de test dépend explicitement de la structure de la matrice de variance. Cette structure doit donc être préalablement choisie pour réaliser les tests sur les effets fixes.

Nous présentons ici quelques procédures permettant de comparer deux modélisations concurrentes de la structure de covariance des données. L'utilisation de critères de sélection de modèles (AIC, BIC, cf paragraphe 3.5.3, p. 123) représentant une stratégie classique de choix de structure, nous présentons aussi brièvement ces critères en fin de chapitre, bien qu'il ne s'agisse pas de procédures de test à proprement parler.

Test exact de Fisher pour l'absence de structure

On considère ici le modèle linéaire mixte proposé dans [43] puis repris dans [21]. Le modèle s'écrit

$$Y_i = X_i\theta + Z_iU_i + E_i, \qquad i = 1, ..., N \tag{5.30}$$
$$U_i \sim \mathcal{N}_m(0, R), \qquad E_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i}) \tag{5.31}$$

où N représente ici le nombre d'unités expérimentales. Le vecteur de mesures Y_i est de taille n_i, on ne suppose donc pas a priori que le nombre de mesures est le même pour toutes les unités. Les effets fixes influant sur les mesures sont les mêmes pour toutes les unités, ces effets sont résumés dans le vecteur θ inconnu. Les effets aléatoires influant sur les mesures sont résumés dans le vecteur U_i, de taille k et propre à chaque unité, mais la matrice de variance de ces effets, notée R, est commune à toutes les unités. Le vecteur E_i est le vecteur des erreurs. On suppose que les variables aléatoires $\{E_1, U_1, ..., E_N, U_N\}$ sont toutes mutuellement indépendantes.

Le modèle générique (5.30) permet d'intégrer et de traiter un large panel de modèles particuliers décrivant les données comme une collection de N unités indépendantes sur lesquelles sont effectuées des mesures. Ce modèle inclut en particulier les exemples des paragraphes 5.1.1, p. 166 et 5.1.2, p. 168. Dans le cas de l'exemple des composantes de la variance, chaque vecteur Y_i est constitué du lot de mesures réalisées sur les descendants d'un même père. Le vecteur aléatoire U_i se résume alors à l'effet aléatoire A_i du i^{eme} père. Le terme aléatoire A_i est donc spécifique à chaque père, mais la loi de ce terme aléatoire est commune à tous les pères, comme spécifié par le modèle 5.30, p. 184.

Le cadre du modèle (5.30) reste toutefois limitant : considérons l'exemple du modèle spatial décrit au paragraphe 5.1.3, p. 169. Dans cet exemple, toutes les mesures sont dépendantes. Il est toujours possible de décrire ce modèle comme un cas particulier du modèle (5.30) où il n'y aurait qu'une seule unité (i.e. $N = 1$) mais nous allons voir qu'alors le test de Fisher ne s'applique pas. Le modèle spatial est donc un exemple où les données ne peuvent pas être décrites comme une collection de *plusieurs* unités indépendantes.

Du point de vue matriciel, le modèle (5.30) peut être réécrit sous les deux formes suivantes :

$$\begin{aligned} Y &= X\theta + ZU + E \\ &= W\nu + E \end{aligned}$$

avec

$$\begin{aligned} Y' &= (Y_1', ..., Y_N'), \quad U' = (U_1', ..., U_N'), \quad E' = (E_1', ..., E_N'), \quad X' = (X_1', ..., X_N'), \\ Z &= \begin{pmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_N \end{pmatrix}, \; W = [X, Z], \; \nu' = (\theta', U') \end{aligned}$$

Le test consiste alors à considérer les effets aléatoires comme des effets fixes et à comparer l'ajustement du modèle sur l'ensemble des facteurs (fixes et aléatoires) à l'ajustement du

modèle ne comportant que les facteurs fixes via une statistique de Fisher. Concrètement, on note P_X et P_W les opérateurs de projection sur les espaces engendrés respectivement par les matrices X et W. L'ajustement d'un modèle est ensuite mesuré via les sommes de carrés résiduelles associées à ce modèle :

$$\begin{aligned} SCR(X) &= Y'(I - P_X)Y \ , \\ SCR(W) &= Y'(I - P_W)Y \ . \end{aligned}$$

Théorème 5.3.1. *Soient r et m les rangs respectifs des matrices X et Y. On a :*

$$\frac{(SCR(X) - SCR(W))/(m-r)}{SCR(X)/(n-m)} \overset{H_0}{\sim} \mathcal{F}_{(r-m,n-m)}$$

où $n = \sum_i n_i$ représente ici le nombre total de mesures.

La démonstration de ce résultat peut être trouvée dans [21].

Le modèle (5.30) est parfois appelé modèle mixte à un facteur, au sens où un seul facteur permet de distinguer les unités indépendantes. C'est le cas du facteur "père" dans l'exemple du modèle de composantes de la variance. Ce modèle peut être généralisé au cas de plusieurs facteurs aléatoires emboîtés. On peut par exemple considérer le cas d'un modèle à deux facteurs aléatoires emboités :

$$\begin{aligned} Y_{ij} = X_{ij}\theta + Z_i U_i + Z_{ij} U_{ij} + E_{ij}, \quad i = 1, ..., M, \ j = 1, ..., N_i, \\ U_i \sim \mathcal{N}(0, R_1), \ U_{ij} \sim \mathcal{N}(0, R_2), \ E_{ij} \sim \mathcal{N}(0, \sigma^2 I_{n_{ij}}) \end{aligned}$$

en reprenant les mêmes notations que pour le modèle (5.30), avec M le nombre de niveaux du premier facteur aléatoire, N_i le nombre de niveaux du second facteur aléatoire au sein de la modalité i du premier facteur, et n_{ij} le nombre de mesures réalisées sur l'unité désignée par le croisement $i \times j$ des facteurs aléatoires. L'exemple des samares traité au paragraphe 6.2, p. 199 correspond à un tel modèle présentant deux facteurs aléatoires emboîtés l'un dans l'autre. Plus généralement, des modèles à K facteurs aléatoires emboîtés peuvent être définis. Il est alors possible de développer des tests de Fisher basés sur la même mécanique que celle présentée dans ce paragraphe pour tester l'absence d'un facteur aléatoire dans ce type de modèle. Le lecteur intéressé consultera [45] pour une étude plus complète des modèles à K facteurs aléatoires emboîtés.

Comme mentionné au début de ce paragraphe, certains modèles ne peuvent pas être mis sous la forme du modèle (5.30) ou de sa généralisation (5.32). Il est alors possible de tester l'absence de structure en réalisant le test du rapport de vraisemblance décrit au paragraphe suivant.

Test du rapport de vraisemblance pour la comparaison de structures de matrice de variance

On s'intéresse à la comparaison de deux modèles m_0 et m_1, ne différant que par la modélisation de la structure de variance covariance, et tels que le modèle m_0 est emboité dans le modèle m_1, i.e. le modèle m_0 peut être vu comme un cas particulier du modèle

m_1 où certains paramètres prennent une valeur fixée (généralement 0). On appelle alors usuellement m_1 le modèle complet et m_0 le modèle réduit. Le test $H_0 : \{m = m_0\}$ vs $H_1 : \{m = m_1\}$ peut être réalisé à partir de la statistique de test

$$L = 2\mathcal{L}_{m_1}(y; \widehat{\psi}_{ReML}(m_1)) - 2\mathcal{L}_{m_0}(y; \widehat{\psi}_{ReML}(m_0)), \tag{5.32}$$

où $\mathcal{L}_m$ représente la log-vraisemblance restreinte des observations calculée pour le modèle m.
Dans le cadre théorique général du test du rapport de vraisemblance, on montre que la statistique de test L suit (le plus souvent asymptotiquement) une loi du χ^2, dont le degré dépend du nombre de contraintes permettant de passer du modèle complet au modèle restreint. Toutefois, la théorie générale ne s'applique pas lorsque le test porte sur un paramètre de variance. Lorsque l'hypothèse testée ne porte que sur un unique paramètre de variance, c'est-à-dire lorsque $H_0 : \{m = m_0\}$ se réécrit en $H_0 : \{\sigma_U^2 = 0\}$, on montre que la loi asymptotique de la statistique de test L est un mélange d'une loi du χ^2 et d'une loi de Dirac $\delta(0)$ centrée en 0 :

$$L \underset{\infty}{\overset{H_0}{\sim}} \frac{1}{2}\delta(0) + \frac{1}{2}\chi^2(1) \quad . \tag{5.33}$$

Il est alors possible de réaliser le test du rapport de vraisemblance en calculant la statistique de test L usuelle, de calculer la probabilité critique comme calculée dans le cadre théorique général puis de diviser par deux la probabilité critique obtenue. Lorsque l'hypothèse H_0 testée est plus complexe, la loi asymptotique de L demeure un mélange de lois du χ^2 mais de proportions inconnues. Plusieurs stratégies peuvent alors être adoptées :

- la difficulté peut être ignorée : on supposera que la loi asymptotique est une loi du χ^2 dont le degré de liberté est déterminé par le nombre de contraintes permettant de passer du modèle complet au modèle restreint, comme dans le cadre théorique général.
- la loi sous H_0 de L peut être approchée par une distribution du χ^2 à r degrés de liberté, où r doit être estimé (cf le paragraphe "Cas du plan BIE à effets blocs aléatoires" de la section 5.3.2, p. 187),
- la loi sous H_0 de L peut être approchée par rééchantillonnage.

Critères de sélection de modèles pour la comparaison de deux structures de covariance dans le cas général

De manière alternative, il est possible de choisir parmi plusieurs modélisations de la structure de covariance en s'appuyant sur des critères de sélection de modèles. La stratégie consiste alors à comparer plusieurs modèles se distinguant par leur structure de covariance mais possédant les mêmes effets fixes sur la base d'un critère de sélection. Deux critères de sélection basés sur la vraisemblance pénalisée sont communément considérés : le critère de Schwarz (BIC) et le critère d'Akaike (AIC).

Considérons un modèle m s'écrivant sous la forme donnée par (5.30). On note $\widehat{R}_{ReML}$ et $\widehat{\sigma}^2_{ReML}$ les estimateurs par vraisemblance restreinte de la structure de covariance et de la variance des erreurs, respectivement. On note par ailleurs $\mathcal{L}_m(y; \widehat{R}, \widehat{\sigma}^2)$ la log-vraisemblance restreinte des observations calculée en les valeurs du maximum. Enfin,

on note D_m le nombre de paramètres décrivant la structure de covariance du modèle m (σ^2 compris). Les critères AIC et BIC s'écrivent

$$\begin{aligned} AIC &= -2\log\left(\mathcal{L}_m(y;\widehat{R}_{ReML},\widehat{\sigma}^2_{ReML})\right) + 2D_m \\ BIC &= -2\log\left(\mathcal{L}_m(y;\widehat{R}_{ReML},\widehat{\sigma}^2_{ReML})\right) + \log(N)D_m \end{aligned}$$

On notera que le nombre d'observations N apparaissant dans le critère BIC correspond au nombre d'unités indépendantes du modèle (5.30). De manière générale, les deux critères évaluent la balance qui s'opère entre l'ajustement du modèle m aux données (via le terme de vraisemblance) et le coût statistique en terme de nombre de paramètres de variance-covariance à estimer du modèle m (via le terme impliquant D_m). La comparaison des modèles se fait par la minimisation de l'un de ces critères. Une application de cette stratégie est présentée au paragraphe 6.3, p. 208.

A l'inverse des procédures de test basées sur la comparaison de deux modèles emboités, les critères permettent la comparaison de deux modèles quelconques m et m' dès lors qu'ils ne diffèrent que par leur structure de covariance. Cette condition est nécessaire puisqu'elle garantit que les log-vraisemblances restreintes associés aux deux modèles sont calculées sur la base des mêmes données transformées $\widetilde{Y}$. Par ailleurs, le choix du critère de sélection reste une question ouverte faisant encore l'objet de développements méthodologiques [20].

5.3.2 Test sur les paramètres de moyenne

Ainsi que nous l'avons déjà mentionné, il n'existe pas de procédure de test statistique universellement optimale dans le cadre du modèle linéaire mixte. Il existe toutefois une multitude de procédures de test approchées, c'est-à-dire basées sur une statistique de test dont la loi sous H_0 n'est pas connue de manière exacte. Ces procédures sont implémentées dans les logiciels et souvent utilisées par défaut, même lorsqu'un test exact existe. Nous présentons donc ici deux exemples de cas où le plan d'expérience n'est pas équilibré, et où il existe néanmoins une procédure de test exacte pour tester l'hypothèse de nullité de l'effet du facteur fixe principal. Nous détaillons aussi sur l'exemple du plan BIE une procédure de test de Fisher approchée pour cette même hypothèse.

Tests de Fisher exacts et approchés

Cas d'un plan orthogonal On considère ici un plan orthogonal à V facteurs sans interaction, c'est-à-dire un plan tel que le modèle statistique s'écrit :

$$Y_i = \mu + \sum_{v=1}^{V}\sum_{j=1}^{K_v} x_{ij}^v \alpha_j^v + E_i, \quad E_i \sim \mathcal{N}(0,\sigma^2)\ \forall i = 1,...n, \tag{5.34}$$

où α_j^v est le paramètre associé à la j_{eme} modalité du facteur v ($j = 1,...,K_v$), et x_{ij}^v la fonction indicatrice valant 1 si l'observation i prend la modalité j pour le facteur v, 0

sinon. Le modèle peut être réécrit sous la forme matricielle :

$$\begin{aligned} Y &= X\theta + E \\ &= \left(X^{(0)} \mid X^{(1)} \mid X^{(2)} \mid ... \mid X^{(V)} \right) \begin{pmatrix} \theta^{(0)} \\ \theta^{(1)} \\ \theta^{(2)} \\ ... \\ \theta^{(V)} \end{pmatrix} + E \end{aligned}$$

où $X^{(v)}$ et $\theta^{(v)}$ représente respectivement le sous-ensemble de colonne de X et le sous-ensemble du vecteur θ associés au $v$ième facteur (avec la convention $v = 0$ pour désigner la constante du modèle). Le plan sera qualifié d'orthogonal si la condition suivante est vérifiée par les facteurs (voir 1.5, p. 37) :

$$\left.\begin{array}{ll} \forall v, v' \in \{1, ..., V\} : & v \neq v', \\ \forall u \in Vect\{X^{(v)}\} : & u'.1 = 0, \\ \forall t \in Vect\{X^{(v')}\} : & t'.1 = 0, \end{array}\right\} \quad u't = 0$$

On rappelle par ailleurs que dans le cas d'un plan orthogonal les sommes de carrés de type I et II sont identiques (cf section 1.4.3, p. 32, p. 32).

Dans l'écriture du modèle 5.34, p. 187, la nature fixe ou aléatoire de chacun des facteurs n'a pas été précisée. On suppose maintenant que le premier facteur est fixe, et que les suivants sont une combinaison de facteurs fixes et aléatoires. On note $SC_I(\theta^{(1)})$ et $SC_{II}(\theta^{(1)})$ les sommes de carrés de type I et II associées au premier facteur Du fait de l'orthogonalité, on a :

$$\begin{aligned} SC_I(\theta^{(1)}) = SC_{II}(\theta^{(1)}) &= \left\| \Pi^{\perp}_{(1)} Y \right\|^2 \\ &= \left\| \Pi^{\perp}_{(1)} (X\theta + E) \right\|^2 \\ &= \left\| \Pi^{\perp}_{(1)} X^{(1)}\theta^{(1)} + \Pi^{\perp}_{(1)} E \right\|^2 \end{aligned}$$

où $\Pi^{\perp}_{(1)}$ est la matrice de projection sur l'orthogonal de $Vect(X^{(2)}, ..., X^{(V)})$. Sous l'hypothèse $H_0 : \{\theta^{(1)} = 0\}$, on obtient :

$$\begin{aligned} SC_I(\theta^{(1)}) &= \left\| \Pi^{\perp}_{(1)} E \right\|^2 \\ \Rightarrow SC_I(\theta^{(1)}) &\sim \sigma^2 \times \chi^2(K_1 - 1) \ . \end{aligned}$$

On peut donc tester l'hypothèse $H_0 : \{\theta^{(1)} = 0\}$ en utilisant la statistique de test

$$F = \frac{SC_I(\theta^{(1)})/(K_1 - 1)}{SCR/(n - P)} \overset{H_0}{\sim} \mathcal{F}(K_1 - 1, n - P) \ .$$

où P est la dimension de l'espace vectoriel engendré par l'ensemble des colonnes de X. On constate ici que dans le cas d'un plan orthogonal sans interaction, il est possible de tester la nullité des paramètres de moyenne associés à un facteur en utilisant la même statistique de test quel que soit la nature (fixe ou aléatoire) des autres facteurs du modèle. Il existe donc toujours un test exact dans ce cadre particulier.

Cas du plan BIE à effets blocs fixes On reprend ici l'analyse de données issues d'un plan en blocs incomplets équilibrés (BIE, exemple de l'évaluation des champagnes, section 8.2, p. 262). Le modèle pour l'analyse d'un tel plan est un modèle ANOVA à 2 facteurs sans interaction :

$$\left\{ \begin{array}{l} Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij} \\ E_{ij} \sim \mathcal{N}(0, \sigma_B^2) \text{ i.i.d.} \end{array} \right. \tag{5.35}$$

où α_i $(i = 1, ..., I)$, β_j $(j = 1, ..., J)$ et E_{ij} désignent respectivement les modalités du facteur d'intérêt appelé traitements (le facteur "champagne"), du facteur bloc ("juge") et l'erreur. Dans l'exemple considéré, $I = 7$ et $J = 14$. On note par ailleurs k le nombre de traitements par bloc, r le nombre de répétitions par traitement et λ le nombre de blocs dans lesquels une même paire de traitement apparaît. On s'intéresse ici au test portant sur l'effet du traitement. L'égalité des modalités de ce facteur peut être testée en se basant sur les carrés moyens de type I ou de type II (voir 1.4.3, p. 34), notées respectivement $MS_I(\alpha)$ et $MS_{II}(\alpha)$. Considérons ces deux carrés moyens. On a :

$$\begin{aligned} \mathbb{E}\left[MS_I(\alpha)\right] &= \mathbb{E}\left[\frac{R(\alpha/\mu)}{I-1}\right] \\ &= \frac{r}{I-1}\sum_{i=1}^{I}\left[\alpha_i - \frac{1}{I}\sum_{i'=1}^{I}\alpha_{i'} + \frac{1}{r}\sum_{j=i}^{J}\left(\beta_j I_{\{i\in j\}} - \frac{1}{I}\sum_{i'=1}^{I}\beta_j I_{\{i'\in j\}}\right)\right]^2 \\ &+ \sigma^2; \\ \mathbb{E}\left[MS_{II}(\alpha)\right] &= \mathbb{E}\left[\frac{R(\alpha/\mu,\beta)}{I-1}\right] = \frac{\lambda I}{k(I-1)}\sum_{i=1}^{I}\left[\alpha_i - \frac{1}{I}\sum_{i'=1}^{I}\alpha_{i'}\right]^2 + \sigma^2 \ , \end{aligned}$$

où $I_{\{A\}}$ représente la fonction indicatrice prenant les valeurs 0 si la condition A n'est pas vérifiée, et 1 si A est vérifiée. Les hypothèses testées sur les paramètres en utilisant ces deux carrés moyens sont respectivement

$$\begin{aligned} H_0^I &: \left\{\alpha_i + \frac{1}{r}\sum_{j=i}^{J}\beta_j I_{\{i\in j\}} \text{ égaux}\right\} \ , \\ \text{et } H_0^{II} &: \{\alpha_i \text{ égaux}\} \ . \end{aligned}$$

On reconnaît ici les hypothèses associées aux tests de type I et II, les statistiques de test sont alors :

$$F_I = \frac{MS_I(\alpha)}{\widehat{\sigma}^2} \tag{5.36}$$

$$\text{et} \qquad F_{II} = \frac{MS_{II}(\alpha)}{\widehat{\sigma}^2} \ , \tag{5.37}$$

les statistiques (5.36) et (5.37) suivant sous l'hypothèse H_0 correspondante une distribution de Fisher $\mathcal{F}_{I-1,n-I-J+1}$.

De cet exemple, on retire certaines caractéristiques importantes du modèle linéaire, toutes devant être reconsidérées dans le cas du modèle linéaire mixte :

— L'écriture de l'hypothèse H_0 pour le facteur α fait intervenir les paramètres α_i associés mais aussi potentiellement les β_j. Ces autres paramètres et leurs coefficients précisent et caractérisent l'hypothèse testée sur le facteur α. Du point de vue géométrique, les sommes des carrés des numérateurs des statistiques de test F_I et F_{II} correspondent à la norme de projections sur des espaces vectoriels différents. On teste ainsi l'effet d'un facteur "corrigé ou non" de l'effet d'un ou plusieurs autres facteurs. Dans le cas d'un modèle mixte, peut-on (et/ou doit-on) corriger un facteur fixe par l'effet d'un facteur aléatoire ? Comment alors formuler l'hypothèse testée, et quel sens lui donner ?

— Au dénominateur des statistiques de test apparaît toujours l'estimation de σ^2. Dans le cadre du modèle linéaire classique, σ^2 peut donc être considérée comme la variabilité de référence, à laquelle est comparée la variabilité associée au facteur testé. Dans un modèle mixte où chaque facteur aléatoire joue le rôle de source de variabilité, qu'appelle-t-on la variabilité de référence ?

— La loi des statistiques de test F_I et F_{II} est connue, en particulier sous H_0, de sorte qu'il est possible d'en calculer exactement les quantiles et donc de calculer la probabilité critique exacte associée à l'hypothèse testée. Une telle situation n'est pas systématique dans le cadre du modèle mixte. Comment alors calculer une probabilité critique approchée ?

Nous traitons chacune de ces questions dans le paragraphe suivant, en dressant le parallèle entre la version du modèle à effets fixes (5.35) et sa version aléatoire.

Cas du plan BIE à effets blocs aléatoires Nous reprenons l'analyse des mêmes données qu'au paragraphe précédent, en considérant maintenant le facteur bloc comme aléatoire. Le modèle est le suivant :

$$\begin{cases} Y_{ij} = \mu + \alpha_i + B_j + E_{ij} \\ B_j \sim \mathcal{N}(0, \sigma_T^2) \text{ i.i.d.}, \quad E_{ij} \sim \mathcal{N}(0, \sigma_B^2) \text{ i.i.d.} \\ B_j \perp\!\!\!\perp E_{ij} \ \forall i, j \end{cases} \tag{5.38}$$

Le modèle (5.38) ne possède qu'un effet fixe. Le test portant sur cet effet fixe se traduit donc directement par l'hypothèse

$$H_0 : \{\alpha_1 = \alpha_2 = ... = \alpha_I\}$$

soit dans sa forme générale une hypothèse de type $H_0 : \{C\theta = 0\}$. Il est parfois fait mention d'hypothèses à tester de forme générale $H_0 : \{C\theta + LU = 0\}$, où U est le vecteur des effets aléatoires [51]. Remarquons tout d'abord qu'une telle hypothèse n'a pas de sens à proprement parler : on ne peut tester qu'une hypothèse portant sur des paramètres, et non portant sur des variables aléatoires. Cette formulation d'hypothèse est donc un abus d'écriture visant à signifier comment l'effet fixe testé va être "orthogonalisé" (i.e. corrigé) pour les différentes variables aléatoires. Afin de comprendre le sens d'une telle orthogonalisation (et de l'utilité de la mentionner), il faut étudier les espérances des deux carrés moyens de la partie précédente dans le cadre du plan BIE

avec effet bloc aléatoire. On a :

$$\begin{aligned}
\mathbb{E}\left[MS_I(\alpha)\right] &= \frac{r}{I-1}\sum_{i=1}^{I}\left[\alpha_i - \frac{1}{I}\sum_{i'=1}^{I}\alpha_{i'}\right]^2 + \frac{J(k-1)}{I-1}\sigma_B^2 + \sigma^2 \\
\mathbb{E}\left[MS_{II}(\alpha)\right] &= \frac{\lambda I}{k(I-1)}\sum_{i=1}^{I}\left[\alpha_i - \frac{1}{I}\sum_{i'=1}^{I}\alpha_{i'}\right]^2 + \sigma^2 \ .
\end{aligned}$$

L'espérance du carré moyen $MS_I(\alpha)$, qui dépendait précédemment des effets fixes des facteurs traitement et bloc ainsi que de la variance résiduelle, dépend maintenant des effets fixes du facteur traitement, de la variance résiduelle et de la variance du facteur aléatoire bloc. L'espérance du carré moyen $MS_{II}(\alpha)$ n'a pas changé. Sous H_0, les effets fixes disparaissent dans les deux carrés moyens et ne restent que les termes de variabilité. On conclut ainsi que les deux carrés moyens peuvent être employées pour tester la même hypothèse H_0 (ce qui n'était pas le cas dans l'analyse du BIE à effets fixes), mais que la variabilité de référence sera différente suivant les carrés moyens que l'on utilisera pour réaliser le test. Certains logiciels feront donc la différence entre

$$\begin{aligned}
H_0^I &: \left\{\alpha_i + \frac{1}{r}\sum_{j=i}^{J} B_j I_{\{i\in j\}} \text{ égaux}\right\} \ , \\
\text{et } H_0^{II} &: \{\alpha_i \text{ égaux}\}
\end{aligned}$$

pour identifier lequel des deux carrés moyens est employé pour réaliser le test.

Deux statistiques de test sont donc à considérer, suivant la réduction qui sera employée comme numérateur. Dans le cas de $MS_{II}(\alpha)$, l'espérance sous H_0 est σ^2, i.e. sous H_0 le carré moyen $MS_{II}(\alpha)$ est un estimateur sans biais de σ^2. La statistique de test sera donc formée du rapport entre ce carré moyen et l'estimateur naturel de σ^2, c'est-à-dire SS_R :

$$F_{II} = \frac{MS_{II}(\alpha)}{SS_R/(n-P)} = \frac{MS_{II}(\alpha)}{\widehat{\sigma}^2} \quad . \tag{5.39}$$

Dans le cas de $MS_I(\alpha)$, l'espérance sous H_0 est $\frac{J(k-1)}{I-1}\sigma_B^2 + \sigma^2$. Il n'y a pas d'estimateur naturel de cette combinaison linéaire des variances σ_B^2 et σ^2. En revanche, on a :

$$\begin{aligned}
& \mathbb{E}\left[MS_{II}(\beta)\right] = \mathbb{E}\left[\frac{R(\beta|\mu,\alpha)}{J-1}\right] = \frac{I(r-1)}{J-1}\sigma_B^2 + \sigma^2 \\
\Rightarrow\ & \mathbb{E}\left[\frac{J-1}{I(r-1)}\left(MS_{II}(\beta) - \frac{SS_R}{n-P}\right)\right] = \sigma_B^2 \\
\Rightarrow\ & \mathbb{E}\left[\frac{J(J-1)(k-1)}{I(I-1)(r-1)}\left(MS_{II}(\beta) - \frac{SS_R}{n-P}\right) + \frac{SS_R}{n-P}\right] = \frac{J(k-1)}{I-1}\sigma_B^2 + \sigma^2 \\
\Rightarrow\ & \mathbb{E}\left[\eta_1 MS_{II}(\beta) + \eta_2 MS_R\right] = \frac{J(k-1)}{I-1}\sigma_B^2 + \sigma^2 \\
& \text{avec } \eta_1 = \frac{J(J-1)(k-1)}{I(I-1)(r-1)} \text{ et } \eta_2 = 1-\eta_1.
\end{aligned}$$

Autrement dit, la combinaison linéaire de carrés moyens $\eta_1 MS_{II}(\beta) + \eta_2 MS_R$ est un estimateur sans biais de la variabilité de référence de $MS_I(\alpha)$. Cet estimateur peut être employé pour obtenir la statistique de test suivante :

$$F_I = \frac{MS_I(\alpha)}{\eta_1 MS_{II}(\beta) + \eta_2 MS_R} \ . \tag{5.40}$$

Constatons que tout comme indiqué en partie 5.2.3, p. 177, l'estimateur $\eta_1 MS_{II}(\beta) + \eta_2 MS_R$ est un estimateur des moments possibles parmi d'autres, basé sur un choix de carrés moyens particuliers. D'autres estimateurs auraient pu être envisagés, il n'existe donc pas de définition unique de la statistique de test F_I.

Considérons tout d'abord la statistique de test F_{II}. Les sommes de carrés du numérateur et du dénominateur étant indépendantes, la statistique de test obtenue suit une distribution de Fisher (cf définition 10.2.5, p. 311). En particulier, sous H_0, on a :

$$F_{II} \overset{H_0}{\sim} \mathcal{F}_{(I-1,n-P)} \ . \tag{5.41}$$

Dans le cas du plan BIE à effet bloc aléatoire, il est donc possible de construire un test exact pour l'effet traitement. Le dénominateur de la statistique de test F_I est constitué non d'un carré moyen mais d'une combinaison linéaire de carrés moyens. Cette combinaison linéaire ne suit donc pas une distribution du χ^2 dans le cas général. Il est toutefois possible d'approcher cette distribution théorique par une distribution du χ^2 à r degrés de liberté, où r doit être choisi de manière à obtenir la meilleure distribution approchée possible. Nous détaillons ici l'approximation de Satterthwaite, basée sur les caractéristiques de la loi du χ^2, rappelées ici :

$$\begin{aligned} Z \sim \chi^2(r) \quad &\Rightarrow \quad \mathbb{E}[Z] = r \text{ et } \mathbb{V}[Z] = 2r \\ &\Rightarrow \quad \mathbb{V}[Z] = \frac{2\left(\mathbb{E}[Z]\right)^2}{r} \end{aligned}$$

On s'intéresse dans le cas général à une combinaison linéaire de termes de variance, estimée à l'aide d'une combinaison linéaire de somme de carrés indépendantes, de la forme

$$Z = \sum_{\ell=1}^{L} \lambda_\ell SC_\ell, \text{ où } SC_\ell \sim \chi^2(r_\ell), \forall \ell \ .$$

En se basant sur les caractéristiques mentionnées précédemment, on a

$$\begin{aligned} \mathbb{V}[Z] &= \frac{2\left(\mathbb{E}\left[\sum_{\ell=1}^{L} \lambda_\ell SC_\ell\right]\right)^2}{r} = \frac{2\left(\sum_{\ell=1}^{L} \lambda_\ell \mathbb{E}[SC_\ell]\right)^2}{r} \\ \text{et} \quad \mathbb{V}[Z] &= \mathbb{V}\left[\sum_{\ell=1}^{L} \lambda_\ell SC_\ell\right] = \sum_{\ell=1}^{L} \lambda_\ell^2 \mathbb{V}[SC_\ell] = \sum_{\ell=1}^{L} \lambda_\ell^2 \frac{2\mathbb{E}[SC_\ell]^2}{r_\ell} \\ \Rightarrow \qquad r &= \frac{\left(\sum_{\ell=1}^{L} \lambda_\ell \mathbb{E}[SC_\ell]\right)^2}{\sum_{\ell=1}^{L} \lambda_\ell^2 \frac{\mathbb{E}[SC_\ell]^2}{r_\ell}}. \end{aligned}$$

Les espérances des sommes de carrés sont ensuite estimées par les valeurs des sommes de carrés observées dans l'échantillon, pour obtenir :

$$\hat{r} \quad = \quad \frac{\left(\sum_{\ell=1}^{L} \lambda_\ell SC_\ell\right)^2}{\sum_{\ell=1}^{L} \lambda_\ell^2 \frac{SC_\ell^2}{r_\ell}}.$$

D'autres approximations ont été proposées pour estimer le degré de liberté d'une combinaison linéaire de sommes de carrés. En particulier, l'approximation de Kenward Rogers (voir [36]) est fréquemment implémentée dans les logiciels statistiques. Notons par ailleurs qu'une stratégie alternative consiste à choisir pour degré de liberté de la combinaison le degré de liberté de l'une des sommes de carrés qui la composent. Cette méthode correspond par exemple à la stratégie appelée "containment" du logiciel SAS.

Test de Wald

A paramètres de variance ψ connus, l'estimateur du maximum de vraisemblance du paramètre θ s'écrit

$$\widehat{\theta} = \left(X'\Sigma^{-1}(\psi)X\right)^{-1} X'\Sigma^{-1}(\psi)Y \ .$$

On en déduit la loi de l'estimateur $\widehat{\theta}$:

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \left(X'\Sigma^{-1}(\psi)X\right)^{-1}\right).$$

On peut alors tester une hypothèse de la forme $H_0 : \{L\theta = C\}$, où L est une matrice correspondant à q combinaisons linéaires indépendantes et C un vecteur de constantes, en se basant sur la statistique de Wald :

$$W = \left(L\widehat{\theta} - C\right)' \left(L\left(X'\Sigma^{-1}(\psi)X\right)^{-1} L'\right)^{-1} \left(L\widehat{\theta} - C\right).$$

Sous H_0, la statistique W suit une distribution du χ^2 à q degrés de libertés. En pratique toutefois, les paramètres de variances sont inconnus, et doivent être préalablement estimés pour calculer la statistique de test

$$\tilde{W} = \left(L\widehat{\theta} - C\right)' \left(L\left(X'\Sigma^{-1}(\widehat{\psi})X\right)^{-1} L'\right)^{-1} \left(L\widehat{\theta} - C\right).$$

La loi exacte de la statistique $\tilde{W}$ est inconnue. Toutefois, en se basant sur la normalité asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\theta}$, on montre que $\tilde{W}$ suit asymptotiquement une distribution du χ^2 à q degrés de libertés.
Ce test asymptotique est généralement désigné sous le nom de test de Wald dans les logiciels statistiques. Notons que dans certains logiciels un test asymptotique de Wald est aussi employé pour tester la nullité d'un ou plusieurs paramètres de variances.

5.4 Analyse de la validité du modèle

Une fois le modèle posé et ses paramètres estimés, avant de l'utiliser pour tirer des conclusions, il faut l'analyser pour garantir sa validité et en particulier vérifier que les

hypothèses qui ont prévalu à sa construction sont raisonnables. Cette démarche n'est bien sûr pas spécifique au modèle mixte, mais dans le cas du modèle mixte elle est compliquée par le fait que pour une même question, suivant la méthode choisie, on risque d'établir des conclusions différentes, et que dans la plupart des cas aucune méthode ne peut être jugée meilleure que toutes les autres. Les hypothèses émises lors de la construction d'un modèle à effets aléatoires concernent outre la normalité des erreurs, la normalité des effets aléatoires et l'indépendance des effets aléatoires et des erreurs. Pour un modèle mixte pour lequel une hypothèse de structure a été émise sur la matrice de variance, les erreurs ne sont pas supposées indépendantes et la vérification portera sur l'adéquation de la matrice de variance estimée à la structure de dépendance des erreurs.

5.4.1 Analyse des résidus

Une première difficulté dans l'analyse des résidus apparaît quant à leur définition. Si on considère le modèle

$$Y = X\theta + ZU + E, \qquad \text{soit} \qquad Y_{ij} = X_i\theta + Z_iU_i + E_{ij},$$

les résidus bruts sont habituellement définis par

$$\widehat{\varepsilon}_{ij} = y_{ij} - X_i\widehat{\theta} - Z_i\widehat{u}_i = y_{ij} - \widehat{y}_{ij},$$

$\widehat{u}_i$ étant la prédiction de l'effet aléatoire par le BLUP, (cf section 5.2.4, p. 181). Mais on peut s'intéresser également aux résidus appelés résidus population, qui représentent l'écart entre les valeurs observées et la prédiction donnée par les effets fixes :

$$\widehat{\varepsilon}_{ij}^{pop} = y_{ij} - X_i\widehat{\theta}.$$

Comme dans le cas du modèle linéaire standard et du modèle linéaire généralisé, on aura tendance à utiliser les résidus transformés pour établir les diagnostics.

Résidus de Pearson

Si on a supposé que la structure de covariance des erreurs est donnée par la matrice Σ, $cov(E) = \Sigma$, les résidus de Pearson ou résidus standardisés sont définis par

$$\widehat{\varepsilon}_i^{Pearson} = \widehat{\Sigma_{ii}}^{-1/2}(y_i - \widehat{y}_i).$$

Ces résidus doivent avoir approximativement une variance constante et être non corrélés. Par contre, si la matrice Σ n'est pas supposée diagonale, ils n'ont évidemment aucune raison d'être non corrélés.

Résidus normalisés

Lorsque la matrice Σ est supposée non diagonale, donc avec une structure de covariance sur les erreurs, on peut examiner les résidus normalisés définis par

$$\widehat{\varepsilon}^{norm} = \widehat{\Sigma}^{-1/2}(y - \widehat{y}),$$

où $\widehat{\Sigma}^{-1/2}\widehat{\Sigma}\widehat{\Sigma}^{-1/2} = I_n$. Ces résidus doivent être approximativement de variance constante et non corrélés si la matrice Σ est bien estimée.
Des représentations graphiques permettent de contrôler visuellement les hypothèses sur les erreurs : des boxplots de résidus par unités expérimentales permettront par exemple de vérifier une hypothèse d'homoscédascticité inter-groupe, et de juger de l'intérêt de considérer des variances d'effets aléatoires non constantes. Lorsque les diagnostics sont négatifs, certains examens complémentaires peuvent aider à enrichir ou modifier le modèle pour avoir une meilleure adéquation. On pourra par exemple en fonction de la nature des données, représenter les résidus en fonction du temps et de l'espace pour juger de la pertinence d'ajouter un terme temporel ou spatial dans la matrice de variance ou changer de modèle de corrélation si ce terme est déjà présent dans la modélisation. Et comme dans le cas du modèle linéaire, la représentation des résidus en fonction d'une covariable non utilisée peut permettre de donner une indication sur l'opportunité de rajouter cette variable dans le modèle, éventuellement en effet aléatoire.

5.4.2 Effets aléatoires

Les hypothèses sur les effets aléatoires seront vérifiées en utilisant leur prédiction BLUP. On contrôlera ainsi visuellement en fonction des modèles, des hypothèses de non corrélation, ou de structure de variance.

La normalité des effets aléatoires et des erreurs sera vérifiée comme d'habitude à l'aide d'un test de normalité (type *q-q* plot ou test non paramétrique) appliqué à leurs estimations. La difficulté dans le cas du modèle mixte vient de ce que les estimations des effets aléatoires et des résidus sont par construction corrélées puisque dépendant chacune des mêmes observations et qu'il n'est pas possible de tester les deux indépendamment. Il faudra donc être très prudent au moment d'énoncer les conclusions.

5.4.3 Adéquation du modèle par simulation

Des méthodes de simulation, qui s'apparentent au bootstrap permettent également de juger de l'adéquation du modèle au données. Une fois le modèle posé et ses paramètres estimés, on tire un grand nombre de fois des pseudo observations que l'on compare aux observations réelles. Dans le cas de mesures répétées par exemple, on comparera une courbe individuelle observée aux courbes simulées pour la même unité expérimentale. Dans d'autres cas, on comparera une statistique donnée sur les observations et la même statistique sur les valeurs simulées. Si les valeurs observées sont proches, dans un sens à définir, des valeurs simulées alors on considère que le modèle est adapté aux données. On parle de VPC (*Visual Predictive Checking*) lorsque la comparaison est simplement visuelle (dans le cas de courbes de réponses par exemple), de PPC (*Posterior Predictive Check*) lorsque une probabilité critique empirique est calculée sur une statistique observée par rapport aux statistiques obtenues par simulation.

Chapitre 6

Modèle mixte : Exemples

6.1 Héritabilité

6.1.1 Données et questions

On a relevé les durées de gestation de 16 filles de 3 taureaux non apparentés et tirés au sort dans une population donnée de bovins. On a obtenu les données de la table 6.1.

Taureau1	Taureau2	Taureau3
292	281	282
289	289	295
299	292	292
296	281	287
302	286	292
289	280	285
288	291	284
293	288	284
293	298	287
294	280	297
284	285	287
286	289	289
287	288	290
291	287	286
301	292	278
290	288	290

TABLE 6.1 – Durées de gestation de 16 filles de 3 taureaux

On veut savoir dans quelle mesure la durée de gestation est un caractère héréditaire se transmettant par les pères.

Modèle

Le modèle est le suivant,

$$Y_{ij} = \mu + A_i + E_{ij} \tag{6.1}$$

A_i et E_{ij} sont des variables aléatoires centrées, gaussiennes et indépendantes :

$$\{A_i\}_i \text{ i.i.d.}, \qquad A_i \sim \mathcal{N}(0, \gamma^2) \qquad \text{et} \qquad \{E_{ij}\}_{i,j} \text{ i.i.d.}, \qquad E_{ij} \sim \mathcal{N}(0, \sigma^2).$$

On justifie le recours à un effet aléatoire par le fait que les taureaux sont aléatoires (ils ont été tirés au sort) et donc leur effet l'est aussi. On ne s'intéresse pas aux taureaux choisis, ils ne sont là qu'en tant que représentants de la population. On s'intéresse à la variation de la durée de gestation due à l'effet génétique des pères. A_i est l'effet génétique du père i, E_{ij} inclut l'effet génétique de la mère, l'interaction entre les facteurs génétiques du père et de la mère et aussi les effets de l'environnement (nourriture, étable, troupeau...). γ^2 représente la variance du caractère due à l'effet génétique additif du père. On parle d'effet génétique additif dans la mesure où E_{ij} a aussi une composante génétique où les gènes du père interviennent de façon non additive. On mesure la caractère héréditaire d'un caractère par le concept d'héritabilité. L'héritabilité au sens étroit est définie par

$$h^2 = \frac{\gamma^2}{\sigma^2 + \gamma^2} \tag{6.2}$$

Elle représente la part de variation du caractère dû à l'effet génétique additif.

Analyse statistique

le programme SAS

```
proc mixed data=taur covtest method=type1 ratio; class taureau;
model duree=;random taureau;run;
```

a donné les résultats suivants :

The Mixed Procedure

Model Information

Data Set	WORK.TAUR
Dependent Variable	duree
Covariance Structure	Variance Components
Estimation Method	Type 1
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
taureau	3	1 2 3

Dimensions

Covariance Parameters	2
Columns in X	1
Columns in Z	3
Subjects	1
Max Obs Per Subject	48

```
                    Number of Observations
Number of Observations Read                  48
Number of Observations Used                  48
Number of Observations Not Used               0
```

```
                    Type 1 Analysis of Variance
             Sum of
Source   DF Squares   M. Sq.  Expected Mean Square           Error Term
taureau   2  231.29   115.64  Var(Residual)+16Var(taureau)  MS(Residual)
Residual 45 1142.62    25.39  Var(Residual)                      .
```

```
                    Type 1 Analysis of Variance
                          Error
Source        DF    F Value    Pr > F
taureau       45       4.55    0.0158
```

```
                Covariance Parameter Estimates

                                      Standard         Z
Cov Parm        Ratio    Estimate       Error     Value        Pr Z
taureau        0.2222      5.6409      7.2356      0.78      0.4356
Residual       1.0000     25.3917      5.3530      4.74      <.0001
```

Les méthodes REML et des moments donnent les mêmes résultats dans ce cas simple et équilibré. La méthode des moments 5.2.3, p. 177 a été utilisée ici (method=type1). Le carré moyen de la table d'analyse de la variance associé à la ligne taureau vaut 115.64 et son espérance est égale à $\sigma^2 + 16\gamma^2$. Connaissant $\hat{\sigma}^2 = 25.39$ on en déduit $\hat{\gamma}^2 = 5.64$. et $\hat{h}^2 = \frac{5.64}{25.39+5.64} = 0.182$. Le test de l'hypothèse $H_0 = \{\gamma = 0\}$ peut se faire de plusieurs façons

Test du rapport de vraisemblance On utilise la méthode du maximum de vraisemblance avec le modèle 6.1, p. 196 puis avec le même modèle mais sans l'effet taureau (programmes SAS non présentés). La log-vraisemblance du modèle 6.1 multipliée par (-2) vaut 294.8 et celle du modèle sans effet taureau vaut 297.2. La différence entre les 2 vaut 2.4 soit une probabilité critique pour un $\chi^2(1)$ égale à 0.121.

Test de Wald Il est calculé à partir des estimateurs du maximum de vraisemblance (non montrés ici) ; on trouve une probabilité critique égale à 0.202. On note que le même test appliqué sur les estimateurs des moments ou ceux du maximum de vraisemblance restreint donne des estimateurs et une probabilité critique différents de ceux du maximum de vraisemblance.

Test F Le test est basé sur le rapport entre le carré moyen associé au facteur taureau et le carré moyen résiduel, rapport qui suit une loi de Fisher à 2 et 45 degrés de liberté sous H_0. La probabilité critique de ce test est égale à 0.0158. Ce test fait partie de la famille des tests F décrits en section 5.3.1, p. 184.

Dans le cas présent simple et équilibré, il est préférable d'utiliser le test F qui est exact alors que les 2 autres ne sont valides qu'asymptotiquement. On conclut alors au rejet

de H_0, ce qui signifie que la durée de gestation est un caractère héritable.

Dans le cas équilibré, on peut de plus construire un intervalle de confiance exact pour h^2 en utilisant le raisonnement suivant (en notant s le nombre de filles et r le nombre de taureaux) :

$Y_{i.} = \mu + A_i + E_{i.}$, donc $\mathbb{V}(Y_{i.}) = \gamma^2 + \sigma^2/s$ et comme les $Y_{i.}$ sont indépendants, on a

$$\frac{\sum_{i=1}^{r}(Y_{i.} - Y_{..})^2}{\gamma^2 + \sigma^2/s} \sim \chi^2(r-1)$$

Comme

$$\frac{\sum_{ij}(Y_{ij} - Y_{i.})^2}{\sigma^2} \sim \chi^2(r(s-1)) \tag{6.3}$$

et que les 2 χ^2 sont indépendants, on en déduit que $\phi = \frac{\sigma^2}{s\gamma^2+\sigma^2}F \sim \mathcal{F}(r-1, r(s-1))$, où

$$F = \frac{\left[\sum_{ij}(Y_{i.} - Y_{..})^2\right]/(r-1)}{\left[\sum_{ij}(Y_{ij} - Y_{i.})^2\right]/[r(s-1)]}.$$

L'encadrement de la loi de Fisher par ses 2 quantiles implique celui de $\frac{\gamma^2}{\sigma^2}$ puis celui de h^2. Les intervalles suivants ont une probabilité égale à $1-\alpha$:

$$\mathcal{F}_{\alpha/2}(r-1, r(s-1)) \leq \phi \leq \mathcal{F}_{1-\alpha/2}(r-1, r(s-1))$$

$$\frac{1}{\mathcal{F}_{1-\alpha/2}(r-1, r(s-1))} \leq \frac{s\gamma^2+\sigma^2}{\sigma^2 F} \leq \frac{1}{\mathcal{F}_{\alpha/2}(r-1, r(s-1))}$$

$$\frac{F/\mathcal{F}_{1-\alpha/2}(r-1, r(s-1)) - 1}{s} \leq \frac{\gamma^2}{\sigma^2} \leq \frac{F/\mathcal{F}_{\alpha/2}(r-1, r(s-1)) - 1}{s}$$

$$\frac{F/\mathcal{F}_{1-\alpha/2}(r-1, r(s-1)) - 1}{F/\mathcal{F}_{1-\alpha/2}(r-1, r(s-1)) + s - 1} \leq h^2 \leq \frac{F/\mathcal{F}_{\alpha/2}(r-1, r(s-1)) - 1}{F/\mathcal{F}_{\alpha/2}(r-1, r(s-1)) + s - 1}$$

La dernière ligne vient du fait que $h^2 = \frac{\gamma^2}{\sigma^2+\gamma^2} = \frac{\gamma^2/\sigma^2}{1+\gamma^2/\sigma^2}$ et que la fonction $f(x) = \frac{x}{1+x}$ est croissante. Dans l'exemple des taureaux on trouve un intervalle de confiance 95% pour h^2 égal à $[0.009; 0.92]$.

6.2 Aptitude à la dispersion des samares du frêne

6.2.1 Présentation de l'expérience

On considère ici une étude écologique visant à étudier la capacité de dispersion des graines de frênes, cette capacité de dispersion pouvant être reliée à la possibilité de coloniser des espaces nouveaux plus ou moins rapidement. Les graines du frênes sont contenues dans des samares dont la forme favorise la dispersion par le vent loin de l'arbre parent. Cette étude est issue des travaux de thèse de S. Brachet [8] et d'O. Ronce [48].

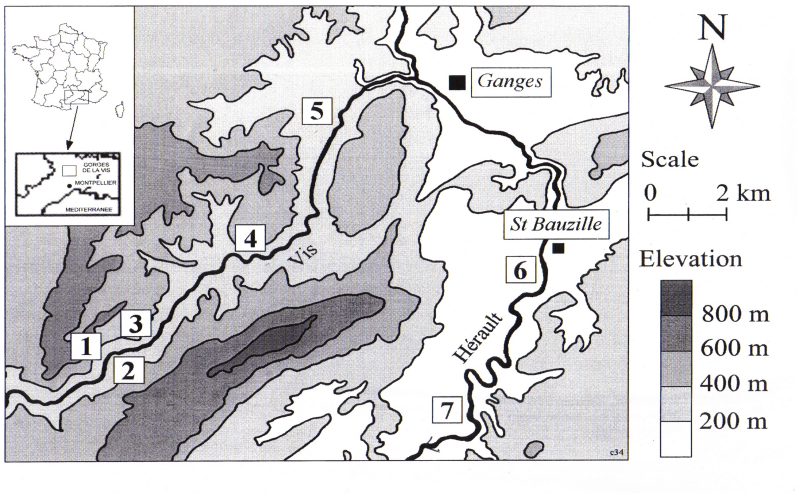


FIGURE 6.1 – Localisation des sept sites de l'étude. (source : [8])

Dispositif. L'étude porte sur $I = 7$ sites répartis le long des cours de la Vis puis de l'Hérault. La figure 6.1 indique leurs localisations respectives. Les sites sont numérotés de l'amont vers l'aval et sont situés respectivement à 0.2, 1.0, 6.3, 11.3, 22.1 et 30.7 km du premier site en suivant le cours de la rivière.

Dans chacun de ces sites, on a échantillonné entre 7 et 29 arbres, soit $m = 118$ arbres au total. Sur chacun de ces arbres, on a prélevé entre 20 et 40 samares, soit $n = 2420$ au total. On a ensuite mesuré le poids (en grammes) et la surface (en cm^2) de chaque samares et calculé l'indice de dispersion

$$Y = 100 \times \text{surface} \,/\, \text{poids}.$$

Cet indice est d'autant plus grand que la samare est susceptible d'être portée loin de son arbre parent. Dans la suite, on notera Y_{ijk} l'indice de dispersion du k-ème samare prélevée sur le j-ème arbre du i-ème site.

Objectifs. On a observé que le cours des deux rivières a été colonisé par les frênes en partant de l'aval. Parmi les objectifs de cette étude, on souhaite savoir

1. s'il existe une grande variabilité de l'indice de dispersion entre les arbres et
2. si les sites les plus récemment colonisés sont peuplés d'arbres produisant des samares plus dispersives ou moins dispersives.

6.2.2 Analyse des composantes de la variance

Modèle. On s'intéresse tout d'abord à la première question pour laquelle on considère le modèle à composantes de la variance qui rend compte de la structure complètement hiérarchique du dispositif (chaque samare est issue d'un seul arbre, lui-même issu d'un seul site). On pose pour cela

$$Y_{ijk} = \mu + A_i + B_{ij} + E_{ijk} \tag{6.4}$$

où μ est le seul paramètre fixe, A_i est l'effet aléatoire associé au $i$ème site, B_{ij} l'effet associé au $j$ème arbre du $i$ème site et E_{ijk} le terme résiduel rendant compte de la variabilité individuel de la $k$ème samare du $j$ème arbre du $i$ème site. Les variables aléatoires $\{A_i\}$, $\{B_{ij}\}$ et $\{E_{ijk}\}$ sont toutes mutuellement indépendantes, normales et centrées, de variances respectives γ_A^2, γ_B^2 et σ^2. On note de plus m_i le nombre d'arbres dans le $i$ème site et n_{ij} le nombre de samares prélevée sur le $j$ème arbre de ce site. Le nombre total d'arbre est noté $m = \sum_i m_i$, le nombre de samares issues du site i $n_{i+} = \sum_j n_{ij}$ et le nombre total de samares $n = \sum_{i,j} n_{ij}$.

Ce modèle s'écrit sous la forme matricielle (5.5) en posant

$$\begin{aligned}
\underset{n\times 1}{Y} &= \begin{bmatrix} Y_{111} \\ \vdots \\ Y_{Im_In_{Im_I}} \end{bmatrix}, \quad \underset{n\times 1}{X} = 1_n, \quad \underset{1\times 1}{\theta} = [\mu], \quad \underset{n\times 1}{E} = \begin{bmatrix} E_{111} \\ \vdots \\ E_{Im_In_{Im_I}} \end{bmatrix}, \\
\underset{I\times 1}{A} &= \begin{bmatrix} A_1 \\ \vdots \\ A_I \end{bmatrix}, \quad \underset{m\times 1}{B} = \begin{bmatrix} B_1 \\ \vdots \\ B_{In_I} \end{bmatrix}, \quad \underset{(I+m)\times 1}{U} = \begin{bmatrix} A \\ B \end{bmatrix}, \qquad (6.5) \\
\underset{n\times I}{Z_A} &= \begin{bmatrix} 1_{n_{1+}} & & \\ & \ddots & \\ & & 1_{n_{I+}} \end{bmatrix}, \quad \underset{n\times m}{Z_B} = \begin{bmatrix} 1_{n_{11}} & & \\ & \ddots & \\ & & 1_{n_{In_I}} \end{bmatrix}, \qquad (6.6) \\
\underset{n\times(I+m)}{Z} &= [Z_A\ Z_B]
\end{aligned}$$

où 1_p représente le vecteur colonne de dimension p composé de uniquement de 1.

Estimation des composantes de la variance.

L'inférence des paramètres du modèle (6.4) peut être menée au moyen du logiciel SAS avec la procédure `Mixed` en utilisant la syntaxe

```
proc Mixed data=SAMARES covtest method=type1;
     class Site Arbre;
     model Dispersion = / ddfm=KenwardRoger;
     random Site Arbre(Site);
```

L'option `covtest` requiert le test de nullité des variances des effets aléatoire au moyen du test de Wald, et l'option `method` indique la méthode d'estimation (ici la méthode des moments fondées sur de sommes de carrés de type 1). L'instruction `class` a la même fonction que dans la procédure `GLM` décrite dans le premier chapitre du livre. L'instruction `model` ne se réfère qu'aux effets fixes et est donc vide ici, la constante μ étant incluse par défaut. L'option `ddfm=KenwardRoger` requiert le test de nullité des effets aléatoires fondées sur les sommes de carrés de type 1, en utilisant l'approximation de Kenward Roger. Les effets aléatoires A_i et B_{ij} sont indiqués dans l'instruction `random`.

On notera que la hiérarchisation des arbres dans les sites est indiquée au moyen de la syntaxe `Arbre(Site)`. Cette instruction est évidemment capitale car l'instruction `Arbre` seule associerait un effet commun à tous les arbres portant le même numéro j dans chaque site, et le modèle résultant serait en fait $Y_{ijk} = \mu + A_i + B_j + E_{ijk}$ qui n'a aucun sens.

La première sortie de cette procédure est donnée au tableau 6.2 qui permet de retrouver les dimensions des vecteurs et matrices. Le modèle comporte trois paramètres de variance $(\gamma_A^2, \gamma_B^2, \sigma^2)$, la matrice X ne comporte qu'une colonne (correspondant à μ), la matrice Z en comporte $m + I = 118 + 7 = 125$, le nombre total de samares étudiées est $n = 2420$. La notion de sujet est liée aux modèles pour mesures répétées qui seront illustrés dans les sections suivantes.

Dimensions

Covariance Parameters	3
Columns in X	1
Columns in Z	125
Subjects	1
Max Obs Per Subject	2420

TABLE 6.2 – Dimension des vecteurs et matrices

Méthode des moments. L'estimation des paramètres de variance γ_A^2, γ_B^2 et σ^2 peut être menée selon différentes méthodes. Nous commençons ici par la méthode de moments décrites à section 5.2.3, p. 177. On présente ici les estimateurs fondés sur les projections de type I décrites dans cette même section. Le tableau 6.3 donne les valeurs des sommes de carrés $\|Y_1\|^2$, $\|Y_2\|^2$, $\|Y_3\|^2$ où les vecteurs Y_1, Y_2 et Y_3 sont définis à l'équation (5.19). Les degrés de libertés associés à ces sommes de carrés sont respectivement $\nu_1 = I - 1 = 6$, $\nu_2 = m - I = 111$ et $\nu_3 = n - m = 2302$.

Source	DF	Sum of Squares	Mean Square
Site	6	171.681492	28.613582
Arbre(Site)	111	1058.063151	9.532100
Residual	2302	1035.191888	0.449692

TABLE 6.3 – Sommes de carrés et carrés moyens de type I

La méthode des moments se fonde sur le système d'équation liant l'espérance des sommes de carrés et les paramètres de variance. Le tableau 6.4 donne l'équivalent du système (5.21) en remplaçant les sommes de carrés $\|Y_k\|^2$ par les carrés moyens $\|Y_k\|^2/\nu_k$. Les estimations des paramètres (donnés dans la colonne 'Moments' du tableau 6.5) sont les solutions de ce système d'équation.

Méthodes du maximum de vraisemblance. Les paramètres de variance peuvent également être estimés par maximum de vraisemblance (éventuellement restreint). Le tableau 6.5 réunit les estimations obtenues avec les trois méthodes vues aux sections 5.2.2, p. 172 et 5.2.3, p. 175. On observe que les estimations sont proches, mais que celles fournies par le maximum de vraisemblance sont plus basses, ce qui rappelle la tendance à la sous-estimation des paramètres de variance de cette méthode.

```
Source       Expected Mean Square

Site         Var(Residual) + 20.918 Var(Arbre(Site)) + 332.62 Var(Site)
Arbre(Site)  Var(Residual) + 20.482 Var(Arbre(Site))
Residual     Var(Residual)
```

TABLE 6.4 – Système d'équations d'estimation des paramètres

Paramètre de variance	Méthode des moments	Maximum de vraisemblance		Maximum de vraisemblance restreint	
γ_A^2	0.05679	0.04786	(0.04889)	0.06557	(0.06279)
γ_B^2	0.4434	0.4425	(0.06294)	0.4412	(0.06255)
σ^2	0.4497	0.4497	(0.01325)	0.4497	(0.01325)

TABLE 6.5 – Estimation des paramètres de variance selon trois méthodes (écart-type asymptotique).

Test sur les composantes de la variance.

La structure hiérarchique de l'échantillonnage suggère fortement le recours à un modèle à effets aléatoires. On peut néanmoins souhaiter tester la significativité de chacun d'eux, ce qui revient à tester la nullité des paramètres de variance γ_A^2 et γ_B^2. Comme on l'a vu à la section 5.3.1, p. 185, on peut utiliser un test du rapport de vraisemblance afin de comparer un modèle incluant l'effet à tester et un modèle ne l'incluant pas.
La table 6.6 donne les vraisemblances de quatre modèles incluant respectivement les deux effets aléatoires dus au site et à l'arbre, au site seul, à l'arbre seul et sans aucun de ces deux effets. Ainsi, si on souhaite tester la nullité de l'effet du site, il convient de comparer les modèles

$$\begin{array}{rrcl} H_0: & \{\gamma_B^2 = 0\} & = & \{Y_{ijk} = \mu + A_i + E_{ijk}\} \\ \text{et } H_1: & \{\gamma_B^2 > 0\} & = & \{Y_{ijk} = \mu + A_i + B_{ij} + E_{ijk}\}. \end{array}$$

Le test se construit de la façon suivante. La log-vraisemblance prise à son maximum vaut $\widehat{\mathcal{L}}_0 = -3274.5$ pour le modèle H_0 et $\widehat{\mathcal{L}}_1 = -2649.9$ pour le modèle H_0. La statistique de test du rapport de vraisemblance vaut alors

$$LRT = 2(\widehat{\mathcal{L}}_1 - \widehat{\mathcal{L}}_0) = 1177.2$$

et doit être comparé au mélange de lois du χ^2 donné à l'équation (5.33), p. 186 (dont l'espérance vaut 1/2 et la variance 5/4). Le test est donc très significatif et on rejette l'hypothèse nulle. Un test analogue peut être construit à partir des vraisemblances restreintes (cf section 5.2.3, p. 175).

Tous les tests présentés dans la table 6.6 aboutissent à rejeter les hypothèses nulles, démontrant la significativité des effets aléatoires prévus dans le modèle (6.4).

On peut également tester la nullité des paramètres de variances des différents effets aléatoires en utilisant les test de Fisher présentés dans la section 5.3.1, p. 184. Le dispositif étudié ici n'étant pas équilibré, on ne dispose pas de la loi exacte de la statistique de Fisher sous H_0 et on doit donc recourir à une loi de Fisher approchée. On choisit ici d'utiliser la méthode de Kenward-Rogers pour en estimer les degrés de liberté. La sortie

Modèle	$ReML$	ML
$\mu + A_i + B_{ij} + E_{ijk}$	-2651.1	-2649.9
$\mu + A_i + E_{ijk}$	-3275.6	-3274.5
$\mu + B_{ij} + E_{ijk}$	-3240.2	-3238.5
$\mu + E_{ijk}$	-3356.7	-3353.7

H_0	$ReML$	ML
'$\gamma_B = 0$'	1249.0	1249.2
'$\gamma_A = 0$'	1178.2	1177.2
'$\gamma_A = \gamma_B = 0$'	1411.1	1407.6

TABLE 6.6 – Test sur les paramètres de variance. Gauche : Log-vraisemblance des différents modèles emboîtés. Droite : Statistique du rapport de vraisemblance entre chacun des modèles et le modèle complet. ML : maximum de vraisemblance classique, $ReML$: maximum de vraisemblance classique.

SAS correspondant aux tests des hypothèses '$\gamma_A = 0$' et '$\gamma_B = 0$' est présentée dans la table 6.7. Là encore on conclue à la significativité des effets aléatoires.

Source	Error Term	Error DF	F Value	Pr > F
Site	1.0213 MS(Arbre(Site)) - 0.0213 MS(Residual)	110.78	2.94	0.0106
Arbre(Site)	MS(Residual)	2302	21.20	<.0001
Residual	.	.	.	.

TABLE 6.7 – Test des paramètres de variance par la méthode des moments (sommes de carrés de type I) en utilisant l'approximation de Kenward-Roger pour les degrés de liberté.

6.2.3 Répartition le long du cours

On cherche maintenant à savoir si la répartition des sites le long du cours de la rivière est liée à la capacité de dispersion des arbres qui y poussent. On pose donc un modèle dans lequel l'effet du site est fixe :

$$Y_{ijk} = \mu_i + B_{ij} + E_{ijk} \tag{6.7}$$

où μ_i est le paramètre fixe associé au site i et les B_{ij} et E_{ijk} ont la même définition et suivent la même loi que dans le modèle (6.4). Le caractère fixe des paramètres μ_i nous permettra de tester des hypothèses sur leurs valeurs respectives. L'objectif est notamment d'étudier comment les paramètres μ_i évoluent en fonction de la localisation du site.

La forme matricielle (5.5) de ce modèle s'obtient en reprenant les matrices définies en (6.5) et en posant

$$\underset{n\times I}{X} = Z_A, \qquad \underset{I\times 1}{\theta} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_I \end{bmatrix}, \qquad \underset{m\times 1}{U} = B, \qquad \underset{n\times m}{Z} = Z_B.$$

L'inférence de ce modèle peut être menée avec le logiciel SAS avec la syntaxe suivante

```
proc Mixed data=SAMARES method=ReML;
     class Site Arbre;
     model Dispersion = Site / solution noint;
     random Arbre(Site);
```

L'option '`noint`' élimine la constante du modèle (conformément à la définition de θ donnée ci-dessus), le rendant ainsi identifiable. L'option '`solution`' permet d'obtenir les estimations des paramètres associés aux effets fixes. On choisit ici d'utiliser le maximum de vraisemblance restreint pour estimer les composantes de la variance (option '`method=ReML`'). Les estimations obtenues sont $\widehat{\gamma}_B^2 = 0.4387$ et $\widehat{\sigma}^2 = 0.4497$ dont on note qu'elles sont proches de celles obtenues dans le modèle (6.4).

Effet du site. On souhaite ensuite tester l'existence d'une effet du site, c'est-à-dire tester l'hypothèse

$$H_0 = \{\mu_1 = \mu_2 = \cdots = \mu_I\}.$$

Le tableau 6.8 présente la sortie associée à ce test fournie par le logiciel SAS. On rappelle que, puisque le dispositif n'est pas équilibré, on ne dispose pas de test exact et que, sous H_0, la statistique ne suit qu'approximativement une loi de Fisher. On remarque que, comme le test porte sur les sites, le nombre de degrés de liberté du dénominateur est bien celui associé à l'effet de l'arbre, hiérarchisé dans le site, soit $\nu_2 = m - I = 111$. Ce test permet de conclure à une forte disparité des capacités de dispersion moyennes entre les différents sites.

Effect	Num DF	Den DF	F Value	Pr > F
Site	7	111	403.03	<.0001

TABLE 6.8 – Test de l'effet du site dans le modèle (6.7).

Lien avec la position le long du cours d'eau. Le tableau 6.9 donne les estimations des paramètres μ_i associés à chacun des sites. On remarque que la valeur de ces estimations décroît avec le numéro du site, c'est-à-dire avec sa distance position en descendant le cours du fleuve. Ainsi les arbres présents dans les sites les plus en aval, plus récemment colonisés, ont une capacité de dispersion plus faible. Des informations externes (portant notamment sur la structure d'âge des population d'arbres dans les différents sites) suggèrent que la colonisation a eu lieu de l'amont vers l'aval. Les estimations obtenues sont alors cohérentes avec l'hypothèse selon laquelle les arbres croissant au sein de populations anciennes sont soumis à une plus grande compétition et doivent donc développer une plus grande capacité de dispersion.
Le test de nullité de chacun des μ_i constitue une sortie standard du logiciel mais ne présente pas d'intérêt ici : une capacité de dispersion nulle n'aurait aucun sens biologique.

On souhaite maintenant préciser le lien entre la localisation du site et la capacité de dispersion moyenne des arbres. Une hypothèse simple consiste à supposer que les paramètres μ_i évoluent linéairement avec la distance d_i séparant le site i du premier

Effect	Site	Estimate	Standard Error	DF	t Value	Pr > |t|
Site	1	3.6922	0.1815	111	20.34	<.0001
Site	2	3.3209	0.1752	111	18.95	<.0001
Site	3	3.8121	0.2148	111	17.75	<.0001
Site	4	3.3266	0.1261	111	26.39	<.0001
Site	5	3.2102	0.1261	111	25.46	<.0001
Site	6	2.9272	0.1813	111	16.14	<.0001
Site	7	2.9034	0.2567	111	11.31	<.0001

TABLE 6.9 – Estimations des effets fixes de chacun des sites

site (en posant $d_1 = 0$), soit

$$H_0 = \left\{ \text{pour } \ 2 \leq i \leq I-1 : \frac{\mu_{i+1} - \mu_i}{d_{i+1} - d_i} = \frac{\mu_i - \mu_{i-1}}{d_i - d_{i-1}} \right\} \tag{6.8}$$

ou, de manière équivalente,

$$\begin{aligned} H_0 &= \{(d_{i+1} - d_i)\mu_{i-1} - (d_{i+1} - d_{i-1})\mu_i + (d_i - d_{i-1})\mu_{i+1} = 0,\ 2 \leq i \leq I-1\} \\ &= \{\delta_i \mu_{i-1} - (\delta_i + \delta_{i-1})\mu_i + \delta_{i-1}\mu_{i+1} = 0,\ 2 \leq i \leq I-1\}. \end{aligned}$$

en notant $\delta_i = d_{i+1} - d_i$ la distance qui sépare les sites i et $i+1$. Cette hypothèse peut s'exprimer sous forme matricielle $H_0 = \{C\theta = 0\}$ où la matrice de contraste C est donnée par

$$\begin{aligned} C &= \begin{bmatrix} \delta_2 & -\delta_1 - \delta_2 & \delta_1 & & & \\ & \delta_3 & -\delta_2 - \delta_3 & \delta_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \delta_6 & -\delta_5 - \delta_6 & \delta_5 \end{bmatrix} \\ &= \begin{bmatrix} 0.8 & -0.6 & 0.2 & & & \\ & 5.3 & -4.5 & 0.8 & & \\ & & 5.0 & 0.3 & 5.3 & \\ & & & 10.8 & -5.8 & 5.0 \\ & & & & 8.6 & 2.2 & 10.8 \end{bmatrix} \end{aligned}$$

L'hypothèse H_0 définie en (6.8) peut ainsi être testée au moyen de la commande SAS suivante, inclue dans la procédure `Mixed`,

```
contrast 'distance' Site 0.8 -0.6  0.2,
                    Site  0   5.3 -4.5  0.8,
                    Site  0    0   5.0  0.3  5.3,
                    Site  0    0    0  10.8 -5.8 5.0,
                    Site  0    0    0    0   8.6 2.2 10.8;
```

Le résultat de ce test est donné dans le tableau 6.10. L'hypothèse H_0 d'un lien linéaire entre la capacité de dispersion et la position le long du cours de la rivière n'est

pas rejetée par ce test. Le caractère linéaire de ce lien n'a pas d'intérêt particulier dans ce cas précis, mais le résultat de ce test va dans le sens d'une relation monotone entre les μ_i et les d_i. Le sens décroissant de cette relation est cohérent avec l'hypothèse selon laquelle les arbres des populations les plus anciennement installées sont soumis à une plus grande compétition et développent donc une plus grande capacité de dispersion.

Label	Num DF	Den DF	F Value	Pr > F
distance	5	111	0.96	0.4464

TABLE 6.10 – Test de l'hypothèse H_0 définie en (6.8).

6.3 Mesures répétées, nutrition humaine

6.3.1 Problématique et données

Les études de nutrition humaine évaluent les réponses métaboliques à des régimes et des repas de composition variable. Nous avons extrait une petite partie d'une étude fournie par C Gaudichon [26] . 17 sujets, hommes et femmes, sont habitués, pendant deux périodes consécutives d'une semaine, soit à un régime *normoprotéique* (NP, apportant 1 g/kg/j de protéines) soit *hyperprotéique* (HP, apportant 2 g/kg/j de protéines). A la fin de chaque période de régime, ils ingèrent un repas test contenant des biscottes soit à dose élevée (ce qui correspond à 30 g de protéines, n=8) soit à dose faible (ce qui correspond à 15 g de protéines, n=9). On mesure alors la cinétique de la glycémie pendant 8 heures.

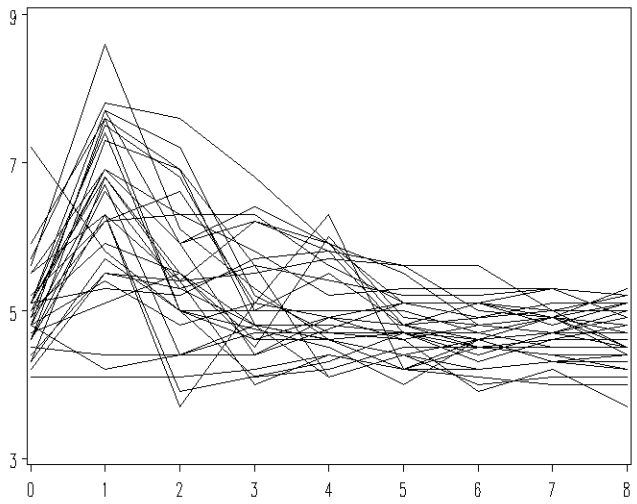


FIGURE 6.2 – Cinétiques de la glycémie pour 34 individus. En abscisse le temps en heures, en ordonnée la glycémie.

Les facteurs sont les suivants : sujet (de B1 à B17), régime (NP/HP), dose (15/30), sexe, temps (0, 1, 2, 3, 4, 5, 6, 7, 8 , en h). La variable réponse est la cinétique de la glycémie aux temps 0, 1, 2, 3, 4, 5, 6, 7, 8h , en mmol/l. D'autres cinétiques ont été mesurées mais ne sont pas analysées ici. On veut connaître l'effet du régime et de la dose ingérée sur la cinétique de la glycémie. Les 11 premières données (sur 298) se présentent de la façon suivante

Obs	sujet	sexe	regime	dose	temps	glycemie
1	B6	f	HP	15	0	4.8
2	B6	f	HP	15	1	7.6
3	B6	f	HP	15	2	5
4	B6	f	HP	15	3	4.1
5	B6	f	HP	15	4	4.2
6	B6	f	HP	15	5	4.5
7	B6	f	HP	15	6	4.5
8	B6	f	HP	15	7	4.5
9	B6	f	HP	15	8	4.5

10	B8	f	HP	15	0	4.6
11	B8	f	HP	15	1	6.6

La figure 6.2 représente les 34 cinétiques associées aux 17 sujets pour les 2 régimes. Elle montre que la glycémie passe par un maximum une heure après l'ingestion avant de redescendre progressivement vers la valeur basale. La figure 6.3 permet d'imaginer des effets potentiels du régime et de la dose. Pour prouver leur existence il faut faire une analyse statistique avec le modèle mixte.

6.3.2 Analyse statistique

Il s'agit de données répétées : la glycémie a été mesurée toutes les heures entre 0 et 8h sur le même individu. Chaque individu n'a eu qu'une seule dose mais il a participé à 2 expériences : une fois après avoir subi le régime HP et une autre fois après le régime NP. Pour ne pas compliquer l'analyse on considère ici qu'il s'agit alors de 2 individus indépendants. Le modèle est le suivant :

$$Y_{ijklt} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_t + \theta_{ij} + \eta_{ik} + ... + E_{ijklt} \tag{6.9}$$

où i est l'indice du régime, j celui de la dose, k celui du sexe, t celui du temps et l celui de l'individu. Toutes les interactions ne sont pas indiquées pour alléger l'écriture. De plus E_{ijklt} et $E_{i'j'k'l't'}$ sont indépendantes sauf si $l = l'$ et $j = j'$ (et donc $i = i'$ et $k = k'$). Il faut donc poser un modèle sur la matrice de variance de E_{ijklt}, $(t = 0...8)$.

Modèle pour la variance

On compare les 5 modèles suivants à l'aide des critères d'Akaike et de Schwartz dans la table 6.11. On choisit le modèle ARH(1) qui est le meilleur au sens du critère BIC et très près du meilleur au sens du critère AIC.

Test des effets fixes

Le programme SAS

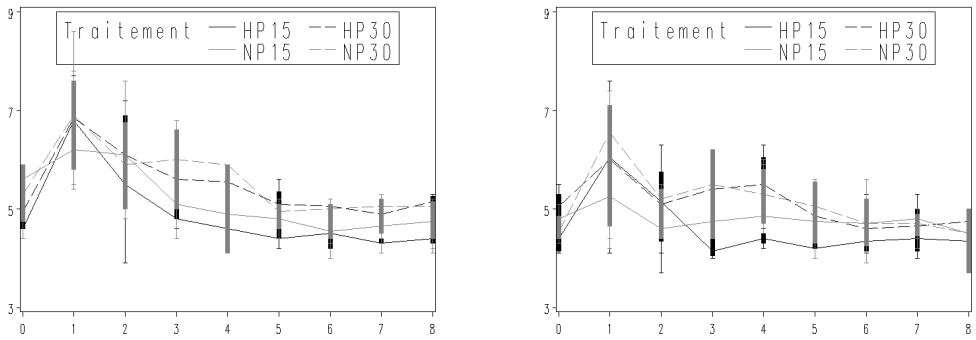


FIGURE 6.3 – Cinétiques de la glycémie regroupées par traitement (figure de gauche : hommes, figure de droite : femmes). En abscisse le temps en heures, en ordonnée la glycémie.

Modèle	Notation	$\mathbb{C}\text{ov}(E_{ijklt}, E_{ijklt'})$	AIC	BIC
Compound Symmetry	CS	$\sigma^2 + \delta(t,t')\gamma^2$	515	518
Autoregressif ordre 1	AR(1)	$\sigma^2\rho^{\lvert t-t'\rvert}$	519	522
CS heterogène	CSH	$\sigma_t\sigma_{t'}(1+\rho(1-\delta(t,t'))$	413	429
AR(1) heterogène	ARH(1)	$\sigma_t\sigma_{t'}\rho^{\lvert t-t'\rvert}$	401	416
Corrélations non structurées	UNR	$\sigma_t\sigma_{t'}\rho_{tt'}$	397	466

TABLE 6.11 – Comparaison des modèles de variance. AIC est le critère d'Akaike et BIC est le critère de Schwartz. Le meilleur modèle est celui pour lequel le critère est le plus faible

```
proc mixed data=biscottes1 covtest; class sexe regime dose temps sujet;
model glycemie=temps sexe regime dose temps*regime temps*dose sexe*regime
sexe*dose dose*regime temps*sexe sexe*dose*regime;
repeated /sub=sujet*regime type=ARH(1)  rcorr;
```

donne les résultats suivants. Les variances sont très différentes d'un temps à un autre, (élevée au temps 1 et faible aux temps 5 à 8) ce qu'on avait pu observer sur les graphiques. De plus il y a une corrélation entre les mesures proches dans le temps ($\hat{\rho} = 0.6$).

```
                    Model Information

Data Set                      WORK.BISCOTTES1
Dependent Variable            glycemie
Covariance Structure          Heterogeneous
                              Autoregressive
Subject Effect                regime*sujet
Estimation Method             REML
Residual Variance Method      None
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Between-Within


                Class Level Information
  Class     Levels    Values
  sexe           2    f h
  regime         2    HP NP
  dose           2    15 30
  temps          9    0 1 2 3 4 5 6 7 8
  sujet         17    B1 B10 B11 B12 B13 B14 B15 B16
                      B17 B2 B3 B4 B5 B6 B7 B8 B9


                  Dimensions
        Covariance Parameters            10
        Columns in X                     90
        Columns in Z                      0
        Subjects                         34
        Max Obs Per Subject               9
```

Number of Observations

Number of Observations Read	298
Number of Observations Used	298
Number of Observations Not Used	0

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Var(1)	regime*sujet	0.3177	0.08512	3.73	<.0001
Var(2)	regime*sujet	1.4658	0.3819	3.84	<.0001
Var(3)	regime*sujet	0.8876	0.2282	3.89	<.0001
Var(4)	regime*sujet	0.4482	0.1172	3.82	<.0001
Var(5)	regime*sujet	0.3377	0.08899	3.79	<.0001
Var(6)	regime*sujet	0.1121	0.02923	3.83	<.0001
Var(7)	regime*sujet	0.1032	0.02235	4.62	<.0001
Var(8)	regime*sujet	0.07365	0.01574	4.68	<.0001
Var(9)	regime*sujet	0.08227	0.01957	4.20	<.0001
ARH(1)	regime*sujet	0.5984	0.05514	10.85	<.0001

Fit Statistics

-2 Res Log Likelihood	381.0
AIC (smaller is better)	401.0
AICC (smaller is better)	401.9
BIC (smaller is better)	416.2

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
9	180.03	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
temps	8	232	10.00	<.0001
sexe	1	26	7.20	0.0125
regime	1	26	0.62	0.4399
dose	1	26	7.48	0.0111
regime*temps	8	232	1.19	0.3076
dose*temps	8	232	1.51	0.1562
sexe*regime	1	26	0.06	0.8012
sexe*dose	1	26	0.12	0.7304
regime*dose	1	26	12.28	0.0017
sexe*temps	8	232	2.18	0.0295
sexe*regime*dose	1	26	0.33	0.5730

Après élimination des interactions non significatives, on trouve qu'il y a un effet combiné du temps et du sexe et un effet combiné du régime et de la dose. Cela signifie que le sexe influence la forme de la courbe de cinétique de la glycémie. L'influence combinée de la dose et du régime ne s'exerce pas sur la forme de la courbe mais seulement sur son niveau moyen. Ces interactions sont illustrées dans la figure 6.4.

Tests des effets fixes après élimination des termes non significatifs

Effect	Num DF	Den DF	F Value	Pr > F
temps	8	248	10.27	<.0001
sexe	1	29	7.69	0.0096
regime	1	29	3.36	0.0770
dose	1	29	12.21	0.0015
regime*dose	1	29	12.05	0.0016
sexe*temps	8	248	2.32	0.0202

Les box-plots sont faits à l'aide des intervalles de confiance 95% issus du modèle mixte. On voit que le régime hyperprotéique induit une glycémie moyenne plus faible que le régime normoprotéique quand il est suivi d'une dose faible. Par contre la glycémie moyenne associée au régime hyperprotéique suivi d'une dose forte est équivalente à celle du régime normoprotéique. D'autre part les femmes ont une glycémie plus faible que les hommes dans les 3 premières heures, mais une glycémie équivalente à partir de la quatrième heure. Ces analyses ont été faites à l'aide du programme SAS suivant :

```
proc mixed data=biscottes1 covtest; class sexe regime dose temps sujet;
model glycemie=temps sexe regime dose dose*regime temps*sexe ;
lsmeans regime*dose temps*sexe /cl; ods output data1;
repeated /sub=sujet*regime type=ARH(1)  ;run;


/* Preparation du jeux de donnees pour faire le graphique regime dose */
legend1 label=(h=3 'Regime') value=(h=3 'HP' 'NP') across=2
        position=(inside top) frame;
axis5 order=(4.5 to 5.5 by 0.5) value=(h=1.5) minor=none label=NONE;
axis6 order=(15 to 30 by 15)  value=(h=1.5) minor=none label=NONE;

data regimedose;set data1;if effect='regime*dose';
data rd1;set regimedose ;y=lower;keep regime dose y;
```

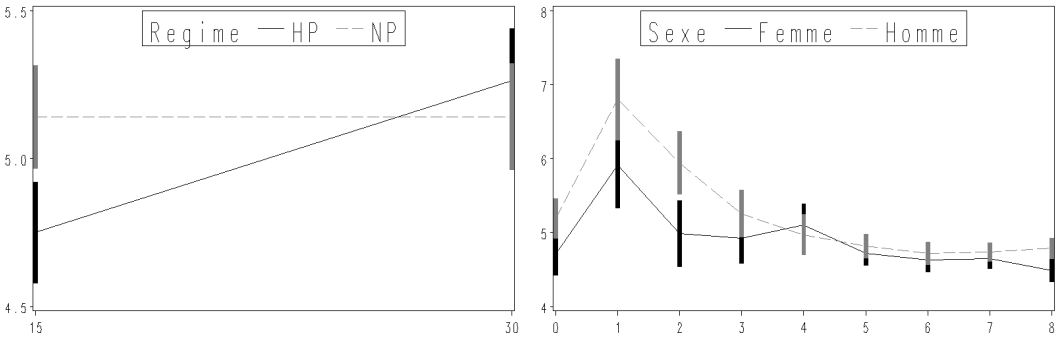


FIGURE 6.4 – Illustration des interactions regime*dose et temps*sexe. En abscisse la dose (figure de gauche) et le temps en heures (figure de droite), en ordonnée la glycémie.

```
data rd2;set regimedose ;y=estimate;keep regime dose y;
data rd3;set regimedose ;y=upper;keep regime dose y;
data rd;set rd1 rd2 rd3;

symbol1 i=boxFJT c=black;symbol2 i=boxFJT c=grey;
proc gplot data=rd;plot y*dose=regime/vaxis=axis5 haxis=axis6
            legend=legend1;run;

/* Preparation du jeux de donnees pour faire le graphique temps sexe */
legend2 label=(h=3 'Sexe') value=(h=3 'Femme' 'Homme')across=2
        position=(inside top) frame;
axis7 order=(4 to 8 by 1) value=(h=1.5) minor=none label=NONE;
axis8 order=(0 to 8 by 1) value=(h=1.5) minor=none label=NONE;

data tempssexe;set data1;if effect='sexe*temps';
data ts1;set tempssexe;y=lower;keep temps sexe y;
data ts2;set tempssexe;y=estimate;keep temps sexe y;
data ts3;set tempssexe;y=upper;keep temps sexe y;
data ts;set ts1 ts2 ts3;

proc gplot data=ts;plot y*temps=sexe/legend=legend2 vaxis=axis7
          haxis=axis8;run;
```

Degrés de liberté pour les tests des effets fixes

La section 5.3.2, p. 187 décrit les méthodes pour calculer les degrés de liberté du dénominateur des tests des effets fixes. La table 6.12 donne les valeurs des degrés de liberté pour 5 méthodes, (cf [51]). On observe des différences non-négligeables entre les différentes méthodes pour le degré de liberté, et même pour la valeur du F calculé par la méthode de *Satterthwaite*. Par contre *containment* et *residual* donnent les mêmes résultats dans ce cas. Il n'y a pas de meilleure méthode en général, et très peu de comparaisons publiées. Dans le cas présent les 5 méthodes conduisent aux mêmes conclusions.

Complément d'analyse

Nous avons admis pour simplifier que les 34 individus analysés étaient indépendants. Ce n'est pas exact puisqu' il n'y avait que 17 individus, et que chacun a eu les 2 régimes. Dans la section précédente nous avons trouvé une interaction significative entre régime et dose. Pour vérifier ce résultat de façon plus précise on peut utiliser la variable différence entre les 2 régimes pour un même individu :

$$D_{jklt} = Y_{1jklt} - Y_{2jklt} = \mu' + \beta'_j + \gamma'_k + \delta'_t + \theta'_{jk} + \eta'_{jt} + ... + E'_{jklt} \tag{6.10}$$

avec les mêmes notations que dans le modèle 6.9. Dans ce modèle les individus sont tous différents et indépendants et l'interaction regime*dose devient un simple effet dose. Pour calculer les différences et faire des graphiques des courbes de différences de la figure 6.5 on a utilisé le programme suivant :

```
/* Fabrication des differences entre les 2 régimes HP-NP */
data np;set biscottes1;if regime='NP';glynp=glycemie;
proc sort data=np;by sujet temps;
```

TABLE 6.12 – Comparaison des tests des effets fixes pour 5 méthodes

Facteur		*between-within*	*residual*	*Satterthwaite*	*Kenward-Roger*
temps	df	248	277	88.6	103
	F	10.27	10.27	10.27	9.84
	Pvalue	$< .0001$	$< .0001$	$< .0001$	$< .0001$
sexe	df	29	277	38.1	38.1
	F	7.69	7.69	7.69	7.72
	Pvalue	0.0096	0.0059	0.0085	0.0084
regime	df	29	277	66.8	66.8
	F	3.36	3.36	3.36	2.96
	Pvalue	0.077	0.0678	0.0712	0.0901
dose	df	29	277	66.8	66.8
	F	12.21	12.21	12.21	10.75
	Pvalue	0.0015	0.0006	0.0008	0.0017
regime*dose	df	29	277	66.9	66.9
	F	12.05	12.05	12.05	10.62
	Pvalue	0.0016	0.0006	0.0009	0.0018
sexe*temps	df	248	277	88.6	103
	F	2.32	2.32	2.32	2.20
	Pvalue	0.0202	0.0198	0.0259	0.0335

```
data hp;set biscottes1;if regime='HP';glyhp=glycemie;
proc sort data=hp;by sujet temps;
data diffgly;merge np hp;by sujet temps;difgly=glyhp-glynp;
     keep sujet sexe dose temps difgly;

/* Definitions pour les graphiques */
axis3 order=(0 to 8 by 1) value=(h=1.5) minor=none label=NONE;
symbol1 i=j v=none c=black l=1 r=10;
axis2 order=(-3 to 3 by 1) value=(h=1.5) minor=none label=NONE;

/* Graphique des différences au cours du temps*/
proc sort data=diffgly;by dose sujet temps;
data diffgly30; set diffgly; if dose=30;
proc gplot data=diffgly30;plot difgly*temps=sujet
/haxis=axis3 vaxis=axis2 vref=( 0 ) nolegend; run;
```

Le modèle CSH est le meilleur au sens des critères AIC et BIC et le seul effet fixe significatif est la dose (programmes SAS et résultats non présentés). Pour répondre à la question de l'existence d'un effet dose on a utilisé le programme SAS suivant :

```
/* CS est mieux et seule la dose est utile*/
proc mixed data=diffgly covtest; class sexe  dose temps sujet;
model difgly= dose  ; lsmeans dose /cl;
ods output LSMeans; repeated /sub=sujet type=CSH rcorr ;run;
```

On a obtenu les résultats suivants qui confirment l'analyse faite en section 6.3.2 : il y a un écart de -0.44 entre la glycémie du régime HP et celle du régime NP pour les individus ayant ingéré la dose de 15, et un écart de +0.16 pour ceux qui ont ingéré la dose 30. Dans le premier cas l'écart est significativement différent de 0.

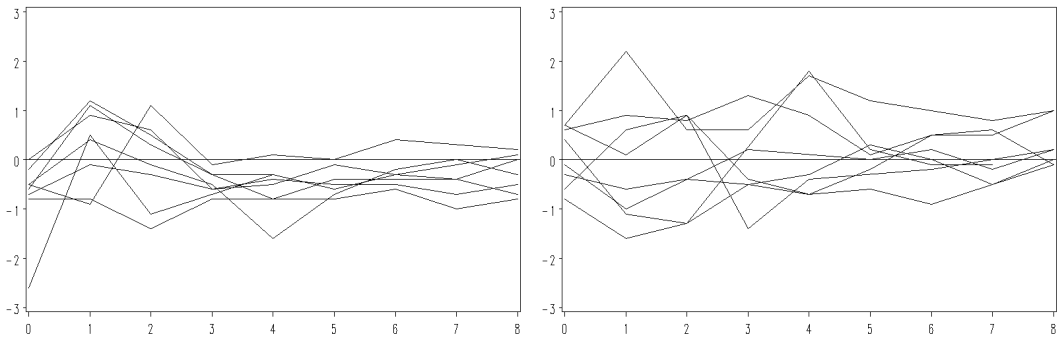


FIGURE 6.5 – Cinétiques des différences de glycémie entre les 2 régimes (figure de gauche : dose=15, figure de droite : dose=30). En abscisse le temps en heures, en ordonnée la différence de glycémie.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
dose	1	15	17.13	0.0009

Least Squares Means

Effect	dose	Estimate	Standard Error	DF	tValue	Pr>|t|	Alpha	Lower	Upper
dose	15	-0.4462	0.1025	15	-4.35	0.0006	0.05	-0.6646	-0.2277
dose	30	0.1604	0.1047	15	1.53	0.1464	0.05	-0.0628	0.3836

6.4 Avalanches

6.4.1 Problématique et données

L'analyse des données d'avalanches est indispensable pour l'évaluation des risques pour les personnes et les biens et pour permettre de prendre les mesures de prévention nécessaires, telles que la délimitation des zones constructibles, ou le dimensionnement d'ouvrages de protection. L'enquête Permanente des Avalanches (EPA) recense depuis 1945 les avalanches survenues dans un certain nombre de couloirs surveillés des Alpes et des Pyrénnées. Les évolutions climatiques observées pendant cette période laissent à penser que le nombre annuel moyen d'avalanches varie et ce différemment suivant les régions. On a extrait de la base EPA le nombre d'avalanches répertoriées dans 105 communes des Alpes sur la période allant de 1988 à 2009. Ce nombre est corrigé en fonction du nombre de couloirs dans la commune. Les évolutions climatiques étant différentes au Nord et au Sud des Alpes, on a également noté la région "AlpesSud" ou "AlpesNord". On s'intéresse à l'évolution temporelle du nombre d'avalanches et de savoir si celle-ci dépend de la région. La figure 6.6 montre que les données comportent un très grand nombre de zéros. Les profils moyens par région ne montrent pas d'évolution

particulièrement différenciée.

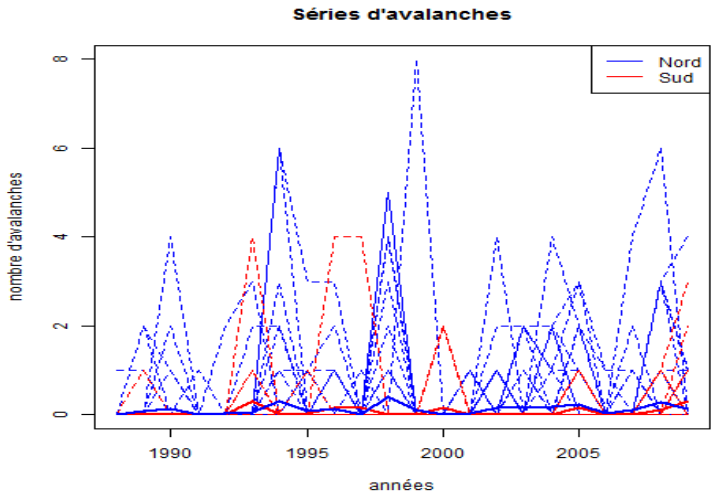


FIGURE 6.6 – Evolution temporelle du nombre d'avalanches par commune (traits pointillés). Les courbes en trait plein donnent les évolutions moyennes par région : bleu Alpes du Nord, rouge : Alpes du Sud

6.4.2 Analyse statistique

Si on ne considère dans un premier temps que des effets fixes on peut commencer par étudier le modèle suivant :

$$N(s,t) \sim \mathcal{P}(\lambda(s,t))$$

$$\log \lambda(s,t) = \mu + \alpha_{R(s)} + (\beta + \gamma_{R(s)})t$$

où $N(s,t)$ est le nombre d'avalanches dans la commune s l'année t. La table 6.13 donne les estimations des paramètres du modèle, avec les tests de Wald associés.

	Estimate	Std. Error	z value	Pr(> | z |)
(Intercept)	-1.250093	0.923919	-1.353	0.176
region	-0.686453	0.425588	-1.613	0.107
temps	0.024377	0.062821	0.388	0.698
region :temps	0.008185	0.028870	0.284	0.777

TABLE 6.13 – Estimations et tests des coefficients du modèle sur le nombre d'avalanches

La table 6.14 donne les résultats des tests du rapport de vraisemblance pour les modèles emboités successifs : $M_0 : \log \lambda(s,t) = \mu$, $M_1 : \log \lambda(s,t) = \mu + \alpha_{R(s)}$, $M_2 : \log \lambda(s,t) = \mu + \alpha_{R(s)} + \beta t$, $M_3 : \log \lambda(s,t) = \mu + \alpha_{R(s)} + (\beta + \gamma_{R(s)})t$.

La déviance du modèle nul M_0 est égale à 1507.2, et la déviance résiduelle du modèle complet est encore de 1478.0, l'AIC est de 1808.2. D'après la table 6.14 l'effet de l'interaction n'est pas significatif et on peut retenir le modèle

$$N(s,t) \sim \mathcal{P}(\lambda(s,t))$$

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2309	1507.2	
region	1	11.9837	2308	1495.2	0.0005367 ***
temps	1	17.2076	2307	1478.0	3.351e-05 ***
region :temps	1	0.0807	2306	1478.0	0.7763154

TABLE 6.14 – Estimations et tests des coefficients du modèle sur le nombre d'avalanches

$$\log \lambda(s,t) = \mu + \alpha_{R(s)} + \beta t$$

La table 6.15 donne les estimations des paramètres du modèle, avec les tests de Wald associés.

	Estimate	Std. Error	z value	Pr(> | z |)
(Intercept)	-1.48335	0.41094	-3.610	0.000307 ***
region	-0.57779	0.17823	-3.242	0.001188 **
temps	0.04196	0.01022	4.105	4.05e-05 ***

TABLE 6.15 – Estimations et tests des coefficients du modèle sur le nombre d'avalanches

La déviance résiduelle est identique et l'AIC est passé à 1806.3. Ce modèle semble donc plus intéressant quoique n'expliquant pas grand chose non plus. Le coefficient β est estimé à une valeur positive, l'intensité du nombre d'avalanches est donc supposée croître avec le temps. Le terme constant est plus faible pour la région Sud que pour la région Nord, l'intensité du nombre d'avalanches est donc plus élevée au Nord qu'au Sud. La figure 6.7 illustre les différences d'intensité entre les deux régions.

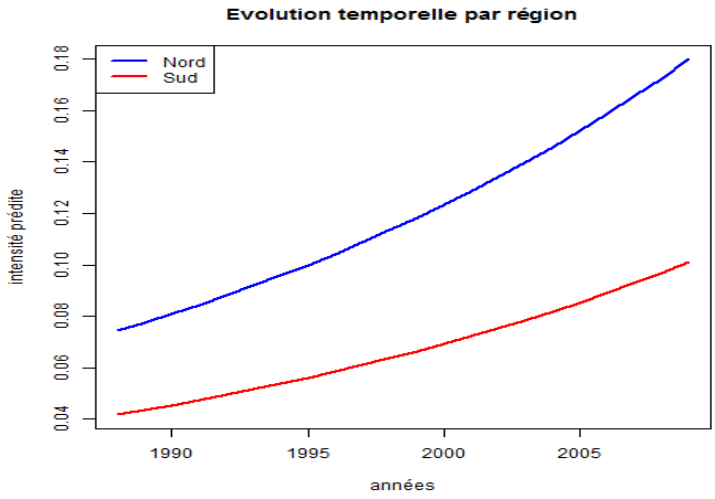


FIGURE 6.7 – Intensité estimée du nombre d'avalanches par commune suivant la région : bleu Alpes du Nord, rouge : Alpes du Sud

6.4.3 Prise en compte d'effets aléatoires

Il est légitime de penser que si peu de déviance est expliquée par le modèle, c'est que les profils temporels sont en réalité plus liés au site qu'à la simple appartenance à une région Nord ou Sud. On introduit alors un effet aléatoire "site" dans le modèle qui peut porter sur l'ordonnée à l'origine et sur la pente. Si on part du modèle complet, avec interaction on a

$$N(s,t) \sim \mathcal{P}(\lambda(s,t))$$

$$\log \lambda(s,t) = \mu + \alpha_{R(s)} + A(s) + (\beta + \gamma_{R(s)} + B(s))t$$

où $A \sim \mathcal{N}(0, \sigma_A)$, $B \sim \mathcal{N}(0, \sigma_B)$ sont les effets aléatoires.

Les écarts-type des effets aléatoires sont estimées à

$$\sigma_A = 2.3670 \qquad \sigma_B = 0.0988$$

La variabilité sur l'ordonnée à l'origine est donc beaucoup plus forte que celle sur la pente. La table 6.16 donne les estimations des effets fixes avec les tests de Wald associés. L'AIC est égal à 1467.8 et la déviance résiduelle à 1453.8, ce qui est un peu meilleur

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.40201	1.96553	-0.713	0.476
region	-1.45979	0.89976	-1.622	0.105
temps	-0.04768	0.10456	-0.456	0.648
region :temps	0.05615	0.04758	1.180	0.238

TABLE 6.16 – Estimations et tests des coefficients du modèle sur le nombre d'avalanches avec effets aléatoires

que ce qui était obtenu pour le modèle à effets fixes.

La table 6.17 donne les AIC de plusieurs modèles où l'on ajoute successivement l'effet aléatoire de la pente puis l'interaction. La table 6.18 donne les tests du rapport de vraisemblance des termes ajoutés. Les p-valeurs associées sont obtenues par des approximations.

modèle $\log(\lambda(s,t)) =$	Df	AIC	logLik	deviance
$\mu + \alpha_{R(s)} + A(s) + \beta t$	4	1491.9	-741.96	1483.9
$\mu + \alpha_{R(s)} + A(s) + (\beta + B(s))t$	6	1467.3	-727.64	1455.3
$\mu + \alpha_{R(s)} + A(s) + (\beta + \gamma_{R(s)} + B(s))t$	7	1467.8	-726.91	1453.8

TABLE 6.17 – AIC log-vraisemblance et deviance des modèles sur le nombre d'avalanches avec effets aléatoires.

modèle $\log(\lambda(s,t)) =$	Chisq	Chi Df	Pr(>Chisq)
$\mu + \alpha_{R(s)} + A(s) + \beta t$			
$\mu + \alpha_{R(s)} + A(s) + (\beta + B(s))t$	28.642	2	6.031e-07 ***
$\mu + \alpha_{R(s)} + A(s) + (\beta + \gamma_{R(s)} + B(s))t$	30.013	2	3.04e-07 ***

TABLE 6.18 – Tests du rapport de vraisemblance de différences de modèle sur le nombre d'avalanches avec effets aléatoires.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.23874	1.27412	-2.542	0.0110 *
region	-0.59761	0.49619	-1.204	0.2284
temps	0.07222	0.03204	2.254	0.0242 *

TABLE 6.19 – Estimations et tests des coefficients du modèle sur le nombre d'avalanches avec effets aléatoires

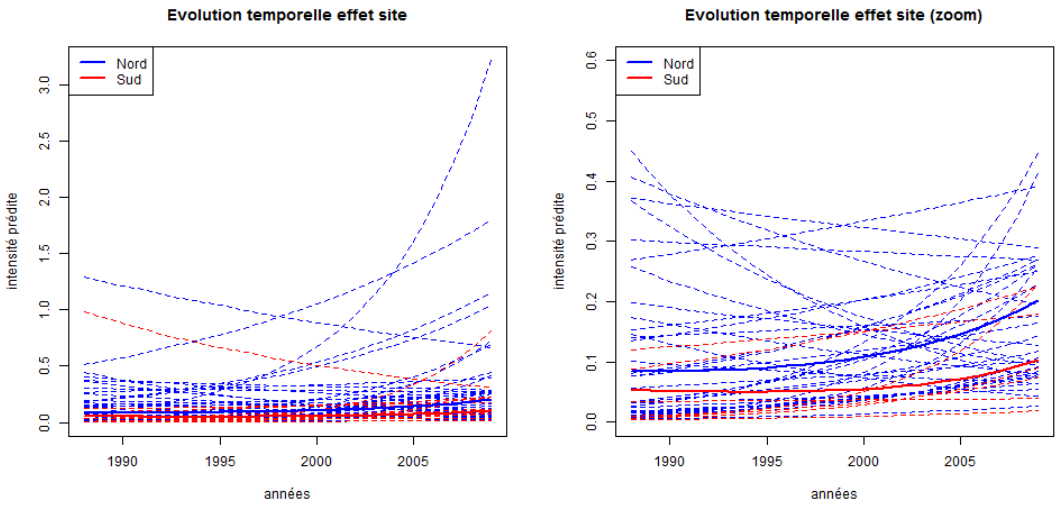


FIGURE 6.8 – Intensité estimée du nombre d'avalanches par commune (traits pointillés). Les courbes en trait plein donnent les évolutions moyennes par région : bleu Alpes du Nord, rouge : Alpes du Sud. A droite : tous les sites, à gauche : sans les sites d'intensité exceptionnelle.

Le modèle le plus pertinent semble donc être le modèle sans interaction, avec un effet aléatoire sur l'ordonnée à l'origine et sur la pente. Pour ce modèle les écarts-type des effets aléatoires sont maintenant estimés à

$$\sigma_A = 2.33276 \qquad \sigma_B = 0.09827$$

qui sont sensiblement les mêmes valeurs que celles obtenues pour le modèle complet.

La table 6.19 donne les estimations des effets fixes avec les tests de Wald associés.

Les valeurs obtenues donnent un écart plus fort entre les deux régions pour l'ordonnée à l'origine que pour le modèle à effets fixes, et une pente également légèrement plus forte. Le fait de prendre en compte des effets aléatoires permet donc de mieux différencier l'effet des régions et l'évolution temporelle. La figure 6.8 montre les évolutions temporelles prédites par site.

La figure 6.9 montre la valeur de l'intensité estimée pour 2008, en fonction de la localisation des sites.

6.4.4 Programme R

```
library(lme4)
```

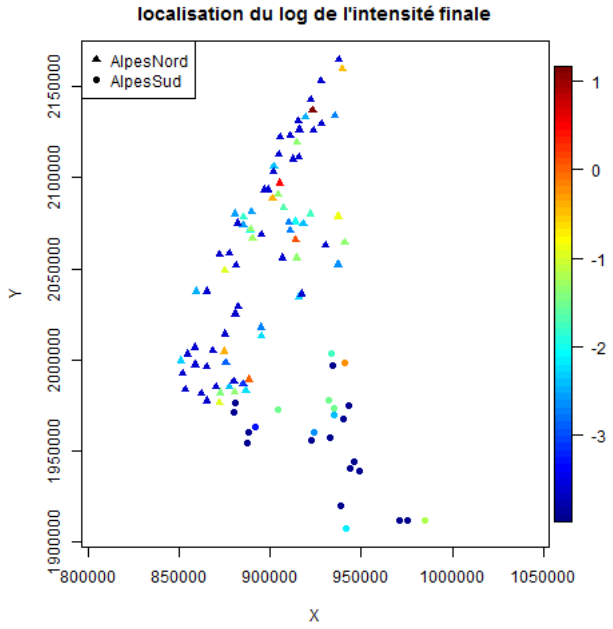


FIGURE 6.9 – log de l'intensité en fonction de la localisation et de la région, triangle : Alpes du Nord, disque : Alpes du Sud.

```
# lecture des donnees
#====================
tab = read.table("avalanches.csv",sep=";",header=TRUE)
Station = unique(tab$site)
ns = length(Station)    # nombre de stations
ls = dim(tab)[1]/ns     # longueur d'une serie
Region = tab$region[seq(1,dim(tab)[1],ls)]
tabMat = matrix(tab$nbav,ls,ns)

# analyse descriptive
#====================
coul = rep("blue",ns)
coul[Region=="AlpesSud"]="red"

matplot(1988:2009,tabMat,type="l",main="Séries d'avalanches",col=coul,
        xlab="années",ylab="nombre d'avalanches",lty=2)
lines(1988:2009,apply(tabMat,1,mean),col=2,lwd=2)
lines(1988:2009,apply(tabMat,1,mean),col=4,lwd=2)
legend("topright",legend=c("Nord","Sud"),lty=1,col=c(4,2))

# modele glm
#===========
# modele complet
modav1 = glm(nbav ~ region*temps, family = poisson,data=tab)
summary(modav1)
anova(modav1,test="Chisq")

# sans interaction
modav2 = glm(nbav ~ region + temps, family = poisson,data=tab)
summary(modav2)
```

```
pred = predict(modav3)
Matpred = matrix(pred,ls,ns)

# representation des intensites par region
matplot(1988:2009,exp(Matpred),type="l",main="Evolution temporelle
          par région",col=coul,
        xlab="années",ylab="intensité prédite",lty=1,lwd=2)
legend("topleft",legend=c("Nord","Sud"),lty=1,col=c(4,2),lwd=2)

# modele avec effets aléatoires
#==============================
# modele complet
modav.glmer1 = glmer(nbav ~ region*temps+ (temps|site),
                      family = poisson, data = tab)
summary(modav.glmer1)

# sans effet aleatoire sur la pente
modav.glmer2 = glmer(nbav ~ region*temps+ (1|site),
              family = poisson, data = tab)

# sans interaction
modav.glmer3 = glmer(nbav ~ region+temps+ (temps|site),
                     family = poisson, data = tab)

# sans effet aleatoire sur la pente et sans interaction
modav.glmer4 = glmer(nbav ~ region+temps+ (1|site),
                     family = poisson, data = tab)
summary(modav.glmer4)

# comparaison des modeles
anova(modav.glmer1,modav.glmer2,modav.glmer4)
anova(modav.glmer1,modav.glmer3,modav.glmer4)

# modele conserve
summary(modav.glmer3)

# prediction
pred.glmer = predict(modav.glmer3)
Matpred = matrix(pred.glmer,ls,ns)

matplot(1988:2009,exp(Matpred),type="l",main="Evolution
                      temporelle effet site",col=coul,
        xlab="années",ylab="intensité prédite",lty=2)
lines(1988:2009,apply(exp(Matpred)[,80:105],1,mean),col=2,lwd=2)
lines(1988:2009,apply(exp(Matpred)[,1:79],1,mean),col=4,lwd=2)
legend("topleft",legend=c("Nord","Sud"),lty=1,col=c(4,2),lwd=2)
```

Chapitre 7

Plans d'expériences

7.1 Pourquoi des plans d'expériences ? vocabulaire de base.

Dans toute étude quantitative, la première étape consiste à récolter des données sur la question étudiée. Il y a 2 façons de le faire.

— On peut constituer un échantillon de la population concernée et mesurer, sur chaque unité de l'échantillon, les variables en relation avec la question posée. Dans ce cas, appelé échantillonnage ou enquête, on se contente d'observer les valeurs de toutes les variables.

— On peut aussi fixer a priori des valeurs de certaines variables et mesurer les valeurs des autres variables. Dans ce cas c'est l'expérimentateur qui fixe les conditions de l'expérience, conditions qui peuvent être très différentes de celles observées dans "la réalité". On dit que l'on fait une expérimentation. Cette deuxième façon de procéder est beaucoup plus efficace pour approfondir une question et établir des relations de causalité. Les Sciences Expérimentales (Physique, Chimie, Biologie) procèdent de cette façon. Les sciences qui ne peuvent pas le faire ont plus de difficultés pour construire des théories scientifiques solides (Sciences Sociales, Economie, Cosmologie...) parce qu'elles ne peuvent pas être validées ou réfutées par des expériences conçues pour cela.

Les plans d'expériences sont utilisés dans la recherche scientifique pour mieux comprendre le fonctionnement de la nature mais aussi dans l'industrie pour déterminer les facteurs qui influencent une caractéristique d'un produit ou d'un processus de production. Ils ont été créés dans les années 1920 pour l'agronomie et la génétique, puis se sont beaucoup développés en Pharmacie et en Médecine. Depuis 1980 ils sont utilisés de façon intensive pour la conception de nouveaux produits dans l'industrie, en particulier l'automobile et les industries agro-alimentaires, et commencent à être utilisés en Psychologie et Economie Expérimentales ainsi que dans le domaine du Marketing. Quand on choisit de faire de l'expérimentation, il faut définir plusieurs éléments :

— le nombre de variables dont on fixe les valeurs

— les valeurs pour chaque variable

— le nombre de répétitions

— l'organisation des expériences en fonction des contraintes (de temps, d'espace...)

On appelle plan d'expériences (en anglais *Design of Experiments* ou *Experimental design*) l'organisation d'ensemble des expériences et en particulier les 4 éléments ci-dessus. Dans ce chapitre nous présentons les principes gouvernant le choix d'un plan d'expé-

riences et nous définissons les dispositifs expérimentaux les plus utilisés. Les principaux critères de qualité d'un plan d'expériences sont présentés en section 7.2, p. 225. L'objectif de l'expérience est généralement l'identification des variables ayant une influence sur la variable réponse étudiée, et la modélisation de cette influence. L'influence d'une variable peut être quantifiée à travers les paramètres du modèle statistique associés à cette variable (les coefficients associés à cette variable dans une régression, par exemple). Les critères de qualité d'un plan peuvent être alors la variance des estimateurs, la puissance d'un test ou la taille d'un intervalle de confiance portant sur les paramètres associés. Pour avoir une bonne puissance des tests (ou des intervalles de confiance petits) il faut diminuer la variance de l'erreur du modèle linéaire. Pour atteindre cet objectif on utilise des facteurs dits "de contrôle de l'hétérogénéité". Ce sont des facteurs qui ne font pas partie de l'étude scientifique mais qui doivent être pris en compte parce qu'ils ont une influence forte sur la variable réponse. On les appelle parfois "facteurs non étudiés", "facteurs contrôlés" ou "facteurs blocs". Comme ces trois appellations prêtent à confusion, nous utiliserons le terme de "facteur de contrôle d'hétérogénéité" dans la suite de cet ouvrage. La section 7.3, p. 229 décrit les différentes formes de prise en compte de l'hétérogénéité.

Une fois qu'on a défini les facteurs étudiés, les facteurs de contrôle d'hétérogénéité et le nombre total d'expériences, il reste à définir la manière dont on répartit les expériences entre tous ces facteurs. La section 7.4, p. 231 traite de cette répartition. La notion de confusion entre effets est un concept clé pour définir cette répartition. Pour que l' analyse soit la plus simple et claire possible il faut limiter la confusion possible entre les effets des facteurs, afin de distinguer clairement l'impact de chacun sur la variable réponse. Certains plans d'expériences éliminent toute possibilité de confusion, d'autres en tolèrent une forme limitée. Dans cette section, les plans d'expériences usuels sont décrits en utilisant le concept de confusion d'effets comme fil conducteur .

Dans certains cas le modèle linéaire n'est pas suffisant pour analyser les expériences et on a besoin du modèle mixte. La section 7.5, p. 247 considère le problème de planification expérimentale dans le cas de ce modèle.

Enfin, la section 7.6, p. 250 est consacrée aux plans d'expériences faits pour estimer une surface de réponse en fonction de variables continues.

Pour conclure cette courte introduction il faut insister sur un point important. La construction d'un plan d'expériences est une étape essentielle dans la démarche expérimentale. Il arrive malheureusement encore souvent que les objectifs scientifiques de l'expérience ne soient pas atteints à cause d'un plan d'expériences mal conçu au départ. Ce chapitre ne vise pas à être exhaustif sur le sujet. Il existe des ouvrages généraux beaucoup plus complets, comme par exemple [16] et [50].

Vocabulaire de base

Nous rappelons ici les principaux termes de vocabulaire associés à la planification expérimentale, en renvoyant aux chapitres précédents pour certains d'entre eux :

Variable réponse, notée Y. Variable qui mesure le phénomène auquel on s'intéresse. On cherche à connaître l'effet des autres variables sur Y. En pratique on ne mesure pas une seule variable et donc il y a plusieurs variables réponses. Cependant le raisonnement fait pour une variable réponse peut être généralisé à plusieurs variables. Pour des raisons de simplification de notations, on ne traitera dans ce

chapitre que le cas d'une variable réponse. Le cas de plusieurs variables est abordé dans le chapitre 8, p. 259.

Variable explicative ou facteur, notée X_j, pour $j = 1, ..., p$. Variable dont on cherche à étudier l'effet sur Y. X_j peut être qualitative ou continue. Les valeurs de ces variables sont fixées à priori par le plan d'expériences. De ce fait, une variable X_j continue ne prend qu'un petit nombre de valeurs et peut donc être considérée soit comme une variable qualitative soit comme une variable continue dans la modélisation.

Niveau d'un facteur. Valeur possible de ce facteur dans le plan d'expériences. On note m_j le nombre de niveaux de la variable X_j.

Facteur étudié. Facteur dont l'influence sur Y est un des objets de l'étude. On le désigne aussi sous le nom de facteur d'intérêt.

Facteur de contrôle de l'hétérogénéité. Facteur dont on admet qu'il puisse avoir un effet sur Y mais qui n'est pas un objet d'étude. Par exemple un effet "cage" dans une expérience sur des souris.

Bloc. Sous-ensemble d'unités expérimentales ayant une caractéristique commune. Par exemple les souris d'une même cage, les essais dans un même champ ou le même jour ou réalisés par le même opérateur. Le facteur bloc est un facteur de contrôle d'hétérogénéité.

Cellule. Combinaison des niveaux de tous les facteurs (facteurs étudiés et facteurs de contrôle d'hétérogénéité).

Unité de traitement. Combinaison de niveaux de tous les facteurs d'intérêt. Par exemple si X_1 = A ou B, et X_2 = a ou b, il y a 4 unités de traitement, Aa, Ab, Ba et Bb. Contrairement à la cellule une telle combinaison ne prend pas en compte les facteurs de contrôle d'hétérogénéité.

Covariable. Variable explicative comme X_j, mais dont les valeurs ne sont pas fixées par le plan d'expériences.

Unité expérimentale. Unité de base sur laquelle sera réalisée une expérience. Elle peut être représentée par un patient en médecine, une parcelle de terrain en agriculture...

Répétition. On dit qu'il y a r répétitions si la même unité de traitement est réalisée sur r unités expérimentales. Une répétition ($r = 1$) signifie qu'on ne répète pas l'expérience.

Randomisation. Procédure d'affectation aléatoire des unités expérimentales aux cellules.

Facteurs orthogonaux. Voir la définition dans la section 1.5, p. 37

Modèle de l'expérience. Modèle statistique adopté pour l'analyse des données de l'expérience, en général un modèle linéaire $Y = X\theta + E$ (ou un modèle mixte ou un modèle linéaire généralisé). Le plan d'expériences est défini par X. C'est dans le cadre de ce modèle que l'on réalise les tests d'hypothèses.

Confusion d'effets. Il y a confusion d'effets entre 2 facteurs non orthogonaux. Dans ce cas il est difficile de séparer la part d'influence sur la réponse de chacun des 2 facteurs. En général la confusion d'effets est partielle. Elle est totale si le sous espace de $\mathbb{R}^n$ engendré par les colonnes de X associées à un des deux facteurs est inclus dans l'espace associé à l'autre facteur.

7.2 Critères de qualité d'un plan

Un "bon" plan d'expériences permet d'obtenir des résultats intéressants pour un coût réduit. Le coût est directement lié aux nombre d'essais du plan et à la facilité matérielle de réalisation, certains plans étant plus difficiles que d'autres à réaliser en pratique. Pour définir un "bon" plan d'expériences, il faut préciser la notion de "résultat intéressant". Prenons le cas de la comparaison de 2 traitements A et B.

— Si l'objectif est de déterminer s'il existe un impact significativement différent des 2 traitements sur la réponse, la méthodologie à mettre en place est celle des tests d'hypothèses. Le critère pertinent pour juger de la qualité du plan est la puissance du test à réaliser.

— Si l'objectif est de quantifier l'écart de réponse qui existe entre les patients traités avec A et ceux traités avec B, la méthodologie à utiliser est l'estimation, et les intervalles de confiance. Le critère pertinent est la variance de l'estimateur du paramètre d'intérêt, ou de manière équivalente la taille de l'intervalle de confiance.

Lorsque la question porte sur un unique paramètre d'intérêt, les 2 approches sont concordantes. Nous détaillons le critère de la puissance du test dans le paragraphe qui suit, et le même raisonnement peut être appliqué à la variance de l'estimateur du paramètre correspondant. En revanche, lorsque les questions portent sur plusieurs paramètres d'intérêt, le plan d'expériences doit permettre d'optimiser plusieurs critères (puissances de plusieurs tests, ou variances de plusieurs estimateurs) en même temps. L'optimisation multicritère et les différentes stratégies qui s'y rapportent sont esquissées au paragraphe 7.6.6, p. 257.

7.2.1 Puissance d'un test

On rappelle ici quelques notions concernant les tests d'hypothèses. Dans cette méthodologie, une hypothèse H_0 et une alternative H_1 sont confrontées. Le test doit permettre de décider si l'hypothèse H_0 doit être conservée ou au contraire rejetée en faveur de H_1. Il y a donc deux types d'erreur possibles :

— Erreur de type 1, on rejette à tort l'hypothèse H_0,

— Erreur de type 2, on accepte à tort l'hypothèse H_0.

Le risque de première (respectivement deuxième) espèce est la probabilité de réaliser une erreur de type 1 (respectivement de type 2). Par définition, la puissance d'un test, notée π, est définie comme la probabilité de détecter que l'hypothèse H_0 est fausse. On a donc

$$\pi = 1 - \beta \quad ,$$

où β est la probabilité de l'erreur de deuxième espèce. Quel que soit le plan d'expériences, les procédures de test statistique sont conçues pour contrôler l'erreur de type 1. On ne choisit donc pas un plan plutôt qu'un autre selon ce critère. Par contre certains plans conduisent à une plus forte erreur de type 2 (i.e. une puissance plus faible). C'est donc un critère de choix essentiel. Le calcul de la puissance est présenté dans cette section dans le détail dans le cas du test de comparaison de 2 populations, puis généralisé au cas des tests de combinaisons de paramètres dans le cadre du modèle linéaire.

Comparaison de 2 traitements

On considère une expérimentation ayant pour objectif de comparer la réponse des patients à 2 traitements différents. Pour cela, on note Y_i la réponse du patient i issu du groupe associé au traitement 1, et Z_j la réponse du patient j issu du groupe associé au traitement 2. $(Y_1, ...Y_{n_1})$ et $(Z_1, ...Z_{n_2})$ sont donc deux échantillons (de tailles respectives n_1 et n_2) issus de 2 populations différentes. On pose le modèle statistique suivant :

$$\begin{array}{rcl} Y_i & \sim & \mathcal{N}(\mu_1, \sigma^2) \ \ \forall i, \\ Z_i & \sim & \mathcal{N}(\mu_2, \sigma^2) \ \ \forall j, \\ \text{toutes} & \text{les} & \text{observations sont indépendantes 2 à 2.} \end{array}$$

On veut réaliser le test unilatéral confrontant l' hypothèse $H_0 = \{\mu_1 \leq \mu_2\}$ à $H_1 = \{\mu_1 > \mu_2\}$. La statistique de test considérée est la suivante

$$T = \frac{\overline{Y} - \overline{Z}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} \mathcal{T}_{(n_1+n_2-2)}$$

où

— $S^2 = \frac{\sum(Y_i - \overline{Y})^2 + \sum(Z_i - \overline{Z})^2}{n_1+n_2-2}$ est l'estimateur de la variance commune σ^2,

— $\mathcal{T}_{(n_1+n_2-2)}$ représente la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Pour un test de niveau α, on rejettera l'hypothèse H_0 dès lors que la valeur observée de la statistique de test t_{obs} est supérieure au quantile d'ordre $1 - \alpha$ de la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Lorsque l'hypothèse H_0 est fausse, la statistique de test T a pour distribution

$$\begin{array}{rcl} T & \overset{H_1}{\sim} & \mathcal{T}_{(n_1+n_2-2,\Delta)} \\ \text{où } \Delta & = & \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \times \frac{\mu_1 - \mu_2}{\sigma} \ , \end{array}$$

$\mathcal{T}_{(n_1+n_2-2,\Delta)}$ représente la loi de Student décentrée à $n_1 + n_2 - 2$ degrés de liberté et de paramètre de décentrage Δ. La puissance du test $\pi(\Delta)$ est directement reliée à ce paramètre de décentrage. Plus précisément, on a :

$$\begin{array}{rcl} \pi(\Delta) & = & P\left[H_0 \text{ rejetée} \mid \mu_1 - \mu_2 = \Delta\sigma\sqrt{(n_1+n_2)/(n_1 n_2)}\right] \\ & = & P\left[T > t_{n_1+n_2-2,1-\alpha} \mid \mu_1 - \mu_2 = \Delta\sigma\sqrt{(n_1+n_2)/(n_1 n_2)}\right] \\ & = & P(\mathcal{T}_{(n_1+n_2-2,\Delta)} > t_{n_1+n_2-2,1-\alpha}) \end{array}$$

La puissance, à travers le coefficient de décentrage, dépend des paramètres μ_1, μ_2, qui sont des données du problème sur lesquelles l'expérimentateur n'a pas la possibilité d'agir et de σ. Dans ce paragraphe on suppose pour simplifier le raisonnement que σ est donné. On verra plus loin que l'expérimentateur peut jouer sur sa valeur et qu'il a intérêt à le diminuer le plus possible. La puissance dépend aussi des valeurs de n_1 et n_2, à la fois à travers le coefficient de décentrage et à travers les degrés de liberté de la loi de Student. Le choix des effectifs n_1 et n_2 étant à l'initiative de l'expérimentateur, on s'intéressera à la valeur de ces effectifs permettant de maximiser la puissance.

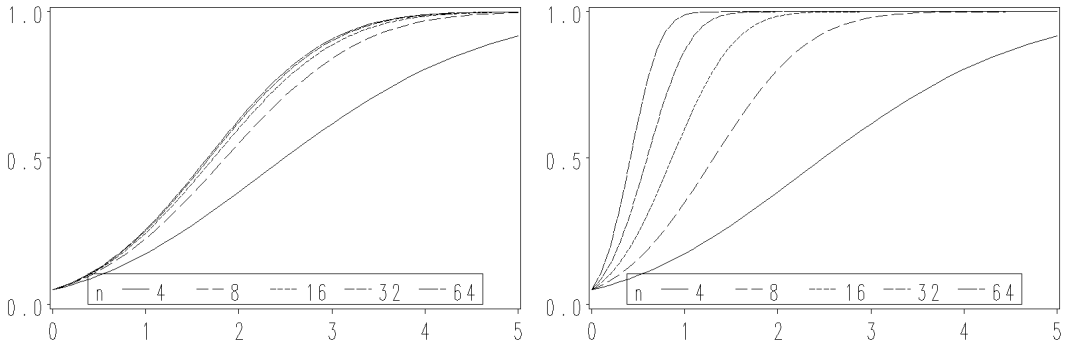


FIGURE 7.1 – Puissance du test de Student de comparaison de 2 traitements pour plusieurs tailles d'échantillon, n. En ordonnée, la puissance du test. Dans la figure de gauche l'abscisse est égale à $\Delta = \sqrt{n/4}\frac{\delta}{\sigma}$, dans la figure de droite l'abscisse est égale à $\frac{\delta}{\sigma}$.

Considérons tout d'abord le nombre de degrés de libertés de la loi de Student. Ce nombre augmente avec la somme $n_1 + n_2$, ce qui implique que le quantile $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ diminue et donc on rejette plus facilement H_0 quand cette somme augmente . Le nombre de degrés de liberté de la loi de Student est égal au nombre de degrés de liberté dont on dispose pour estimer σ. Plus il est élevé meilleure est cette estimation. Ainsi, on choisira $n_1 + n_2$ aussi grand que possible.

Par ailleurs, à somme $n_1 + n_2 = n$ fixée, le coefficient de décentrage est une fonction croissante de $\frac{n_1 n_2}{n_1+n_2}$. Afin de maximiser la puissance, on cherchera donc les valeurs

$$\begin{aligned}(n_1^*, n_2^*) &= \arg\max_{n_1,n_2} \frac{n_1 n_2}{n_1 + n_2} \quad \text{sous contrainte} \quad n_1 + n_2 = n \\ \Leftrightarrow n_1^* &= \arg\max_{n_1} \frac{n_1(n - n_1)}{n} \text{ et } n_2^* = n - n_1^*.\end{aligned}$$

Il est aisé de montrer que les valeurs optimales sont $n_1^* = n_2^* = n/2$. On choisira donc un plan où les effectifs des échantillons sont égaux.

La figure 7.1, p. 227 représente la puissance en fonction de n (avec $n_1 = n_2 = n/2$), et du rapport δ/σ, où $\delta = \mu_1 - \mu_2$. On constate que la puissance est une fonction croissante de n. Cela provient de 2 effets distincts :

- Le nombre de degrés de liberté de la loi de Student $(2n-2)$ augmente avec n, ce qui améliore la puissance pour toute valeur de Δ, comme on le voit sur la partie gauche de la figure 7.1.
- n influe sur l'écart-type de la différence des moyennes empiriques, $\bar{Y} - \bar{Z}$. Cet écart-type est $\sigma\sqrt{\frac{4}{n}}$, quantité qui apparaît dans le coefficient de décentrage. Cet effet (cumulé avec le précédent) est illustré par la partie droite de la figure 7.1.

Il est possible de généraliser le résultat obtenu au cas de la comparaison de K traitements. On montre qu'un plan équilibré permet de maximiser la puissance du test de Fisher associé aux hypothèses $H_0 = \{\mu_1 = \mu_2 = ... = \mu_K\}$ contre $H_1 = \{\exists i, i' \text{ t.q. } \mu_i \neq \mu_{i'}\}$ ([42], p.59).

Pour résumer, la puissance de détection de différences entre traitements est une fonction croissante de la (ou des) vraie(s) différence(s) de moyennes des traitements, du nombre d'expériences n et décroissante de la variance résiduelle σ^2. Elle dépend également des proportions d'expériences attribuées à chaque traitement.

Puissance d'un test dans le modèle linéaire

Le comportement de la puissance du test de Student est similaire quels que soient le modèle linéaire utilisé et le test effectué. Le plus souvent le modèle linéaire utilisé est plus complexe que dans une simple comparaison de moyennes, mais la démarche reste la même. Soit $Y = X\theta + E$ avec $E \sim N(0, \sigma^2 I)$, un modèle linéaire. On a fait l'expérimentation pour tester une certaine hypothèse scientifique. A cette hypothèse on associe une combinaison linéaire des paramètres du modèle (pour simplifier on ne considère ici qu'une seule combinaison linéaire). Soit L le vecteur des coefficients (connus) de cette combinaison linéaire. L'hypothèse s'écrit alors $H_0 = \{L'\theta = L'\theta_0\}$, où θ_0 est connu. Le test de H_0 se fait avec la statistique de test suivante :

$$\frac{\mid L'\widehat{\theta} - L'\theta_0 \mid}{\sqrt{\widehat{\mathbb{V}}(\widehat{\theta})}} = \frac{\mid L'\widehat{\theta} - L'\theta_0 \mid}{S\sqrt{L'(X'X)^{-1}L}}.$$

On a donc un test similaire au test de comparaison de 2 moyennes. Le terme $\sqrt{L'(X'X)^{-1}L}$ remplace $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, et le nombre de degrés de libertés de la loi de Student est celui de la variance résiduelle du modèle linéaire. Le calcul de puissance est le même que précédemment.

On doit choisir un plan d'expériences de façon à avoir la plus grande puissance possible pour la question posée. Le choix du plan n'a bien sûr aucun effet sur δ qui est donné par "la nature". Par contre il a un impact sur σ. On doit donc choisir un plan d'expériences X tel que $\sigma\sqrt{L'(X'X)^{-1}L}$ soit le plus faible possible.

7.2.2 Cas de plusieurs questions, plans optimaux

Bien souvent, une même expérience a pour objectif de répondre à plusieurs questions scientifiques. A chaque question est associée une hypothèse concernant une combinaison linéaire des paramètres du modèle. Il y a alors p combinaisons linéaires d'intérêt, $L_1, ...L_p$, où p est le nombre de questions. Il n'y a aucune raison pour que le plan d'expériences qui maximise la puissance du test portant sur L_1 maximise aussi la puissance du test portant sur L_2, ou celle de toute autre combinaison linéaire d'intérêt. Il peut même arriver que le plan optimal pour la puissance du test portant sur L_1 ait une faible puissance pour le test portant sur L_2. Il faut donc que l'expérimentateur *hiérarchise* ses objectifs, c'est-à-dire ses questions scientifiques, par ordre d'importance. Si l'on suppose que les combinaisons $L_1, ...L_p$ sont rangées par ordre d'importance (de la plus à la moins importante), on favorisera un plan d'expériences ayant la meilleure puissance pour le test de L_1, puis parmi tous les plans ayant cette qualité, on choisira le plan qui a la meilleure puissance pour le test de L_2 et ainsi de suite.

Il est quelquefois difficile de hiérarchiser les questions scientifiques que se pose l'expérimentateur, qui ne veut pas sacrifier un objectif à un autre. Dans ce cas on fait en

sorte que tous les tests potentiels aient à peu près la même puissance. Nous avons vu que la puissance d'un test portant sur θ est liée à la matrice de variance-covariance de l'estimateur $\widehat{\theta}$. Une façon de poser le problème consiste donc à choisir le plan d'expériences qui minimise dans leur ensemble les coefficients de cette dernière. Pour cela, il faut choisir une manière de "résumer" la matrice de covariance par un indice, qui permettra de hiérarchiser les différents plans d'expériences. Plusieurs choix d'indice sont possibles : le déterminant, la trace ou la première valeur propre de la matrice. On considère le modèle linéaire

$$Y = X\theta + E$$

On sait que $\widehat{\theta} = (X'X)^{-1}X'Y$ et $V(\widehat{\theta}) = \sigma^2(X'X)^{-1}$. On remarque que cette variance ne dépend que du plan d'expériences (et de σ) et pas de Y. On sait que X est la matrice du plan d'expériences : connaissant X on connaît le plan d'expériences et réciproquement la connaissance du plan d'expériences implique celle de X. Donc rechercher un plan d'expériences revient à rechercher X. Il existe plusieurs critères :

- D-optimalité : on cherche X qui maximise le déterminant de $X'X$
- A-optimalité : on cherche X qui minimise la trace de $(X'X)^{-1}$
- E-optimalité : on cherche X qui maximise la plus petite valeur propre de $X'X$

Une variante de la D-optimalité consiste à minimiser le determinant de la matrice de variance des combinaisons linéaires d'intérêt ([27], p 151). On peut aussi choisir le plan d'expériences qui minimise la variance des prédictions fournies par le modèle linéaire et non pas celle des estimateurs des paramètres. C'est la stratégie souvent choisie pour les plans d'expériences pour les surfaces de réponse, voir la section 7.6, p. 250.

Pour en savoir davantage sur les critères d'optimalité et les algorithmes pour trouver un plan optimal le lecteur pourra se référer à [27].

7.3 Contrôle de l'hétérogénéité des unités expérimentales

La variabilité expérimentale est une sorte de brouillard qui cache ou déforme les effets des traitements : si elle est importante on ne sait pas si la différence de réponse observée entre 2 groupes d'unités expérimentales ayant subi 2 traitements provient de la différence des effets des 2 traitements ou des différences entre les unités expérimentales elles-mêmes. La réduction de la variabilité expérimentale, mesurée par σ, est très efficace pour augmenter la puissance des tests. La puissance dépend du terme $r\frac{\delta}{\sigma}$. Donc diviser σ par 2 a le même effet sur la puissance que de multiplier r (et donc n) par 4. Compte tenu du coût de chaque unité expérimentale, on comprend que les expérimentateurs aient fait preuve d'inventivité pour diminuer cette variabilité. Réduire cette variabilité revient à considérer des unités expérimentales qui soient très proches de façon à ce que la différence de la variable réponse entre 2 traitements provienne essentiellement de la différence d'effet des traitements et non pas de la différence entre les unités expérimentales ayant eu les 2 traitements. On peut par exemple considérer des unités expérimentales similaires voire identiques sur le plan génétique, ayant eu la même histoire et "vivant" dans les mêmes conditions. Il y a cependant une limite pour le nombre d'unités expérimentales quasi identiques : les portées d'animaux (ayant les mêmes parents) dépassent rarement 10 animaux, le nombre d'animaux appartenant à la même cage ou étable est

aussi limité à 10. Le nombre de périodes successives disponibles pour un même animal est aussi limité à quelques unités et il en est de même du nombre de produits différents que l'on peut fabriquer le même jour sur une même chaîne de production et du nombre de patients dans le même service d'un hôpital. En résumé il est très difficile de trouver ou de fabriquer 30 unités expérimentales qui soient identiques à tout point de vue. Par contre il est possible de constituer des paquets ou blocs de 2 à 10 unités quasi identiques. C'est l'origine du dispositif dit en "blocs". Une solution plus générale (incluant les dispositifs en blocs) consiste à classer les unités expérimentales selon certains critères qualitatifs appelés facteurs de contrôle d'hétérogénéité et d'intégrer ces facteurs dans le modèle linéaire, ce qui permet de diminuer σ. Dans ce paragraphe nous présentons 2 plans d'expériences types, le "bloc complet" et le "carré latin".

7.3.1 Bloc complet

On considère b ensembles (appelés blocs) d'unités expérimentales. Les unités d'un même bloc sont semblables et tous les blocs ont le même nombre d'unités expérimentales, t. De plus il y a t unités de traitement (par exemple un facteur traitement avec t traitements). Le dispositif en *blocs complets randomisé* consiste à affecter (par tirage au sort) les t unités de traitement aux t unités expérimentales de chaque bloc. Les t unités de traitement sont donc toutes présentes dans un même bloc. Le modèle d'analyse statistique est un modèle d'analyse de variance à 2 facteurs sans interaction :

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}, \quad E_{ij} \hookrightarrow \mathcal{N}(0, \sigma^2) \ i.i.d. \ ,$$

où Y_{ij} est la mesure de la variable réponse réalisée dans le bloc i pour l'unité de traitement j.[1] Le dispositif en blocs permet de réduire la valeur de σ parce que la variabilité expérimentale prise en compte est celle qui existe entre unités expérimentales d'un même bloc et non pas celle existant entre unités de blocs différents.

7.3.2 Carré Latin

On considère t^2 unités expérimentales réparties dans un carré ayant t lignes et t colonnes. Les unités d'une même colonne ont une caractéristique commune et les unités d'une même ligne ont aussi une caractéristique commune. De plus il y a t unités de traitement (par exemple un facteur traitement avec t traitements). Le dispositif en *carré latin* consiste à affecter les t unités de traitement aux t^2 unités expérimentales de telle sorte que chaque unité de traitement se trouve une fois et une seule dans chaque ligne et dans chaque colonne. Le modèle d'analyse statistique est un modèle d'analyse de variance à 3 facteurs sans interaction :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + E_{ijk}, \quad E_{ijk} \hookrightarrow \mathcal{N}(0, \sigma^2) \ i.i.d. \ ,$$

où Y_{ijk} est la mesure de la variable réponse réalisée dans la ligne i et la colonne j pour l'unité de traitement k, α_i est l'effet de la ligne i, β_j est l'effet de la colonne j et δ_k est l'effet du traitement k. Il y a plusieurs carrés latins possibles et la randomisation

1. Ce modèle suppose qu'il n'y a pas d'interaction bloc*traitement car on ne dispose pas de suffisamment de données pour l'estimer. Si on soupçonne que cette interaction existe il faut répéter chaque unité de traitement dans un même bloc et inclure l'interaction dans le modèle.

consiste à tirer au sort l'un d'entre eux. [2] Ce dispositif permet de réduire la valeur de σ parce que la prise en compte des effets ligne et colonne permet de diminuer la variabilité résiduelle du modèle.

7.4 Plans d'expériences utilisant la confusion d'effets

Dans le dispositif en blocs il n'y a pas de confusion d'effets entre les facteurs bloc et traitement. Dans le carré latin il n'y a pas de confusion d'effets entre le traitement et les facteurs ligne et colonne : les effets du traitement et des facteurs de contrôle d'hétérogénéité sont estimés de façon indépendante. Mais ces dispositifs ne peuvent pas toujours être utilisés pour des raisons de contraintes pratiques. On est alors amené à accepter certaines confusions d'effets partielles ou totales pour pouvoir respecter ces contraintes. C'est ce que nous décrivons dans cette section. En section 7.2, p. 225, nous avons vu que la recherche d'un plan d'expériences revient à optimiser un critère (précision d'un ou plusieurs contrastes, puissance d'un test) sous contrainte du nombre limité d'expériences qu'il est possible de réaliser. En particulier, l'analyse de l'exemple de la comparaison de deux traitements a montré qu'il ne suffit pas de définir le nombre total n d'expériences à réaliser, mais que le plan d'expériences doit aussi définir :

- une règle de répartition en proportion, qui détermine comment le nombre total d'expériences doit être réparti, c'est-à-dire en quelles proportions les unités expérimentales doivent être allouées aux différentes cellules,
- une règle d'affectation individuelle, qui détermine comment l'affectation des unités expérimentales doit être réalisée en pratique pour respecter la répartition en proportion qui a été définie ci-dessus. Par exemple dans le cas d'une comparaison de 2 traitements, on peut affecter les patients aux différents traitements suivant leur ordre d'arrivée, leur âge ou aléatoirement...

Ces deux notions sont définies en section 7.4.1, p. 231, et la stratégie de randomisation pour la règle d'affectation individuelle est brièvement décrite. Certaines règles de répartition en proportion génèrent des confusions d' effets. Les stratégies classiques visant à gérer la confusion sont présentées.

7.4.1 Définitions

Règle de répartition en proportion

Définition 7.4.1. *On appelle règle de répartition en proportion la donnée, pour chaque cellule c_ℓ, $\ell = 1, ..., L$, de la proportion d'unités expérimentales p_ℓ allouées à c_ℓ.*

A toute règle de répartition en proportion correspond une matrice $X'X$ du plan d'expériences et donc la matrice de variance des estimateurs des paramètres, $\sigma^2 X'X$. Il est donc possible de comparer les règles de répartition en proportion, voire de déterminer la ou les règles optimales selon un des critères vus en section 7.2.2, p. 228. Dans l'exemple 7.2.1, p. 226, il existe 2 cellules correspondant aux 2 niveaux du facteur traitement. Nous avons vu que pour cet exemple un plan d'expériences avec effectifs égaux garantit la maximisation de la puissance du test de comparaison des 2 moyennes. Ainsi, la règle

2. Ce modèle suppose qu'il n'y a pas d'interaction ligne*traitement ni colonne*traitement ni ligne*colonne car on ne dispose pas de suffisamment de données pour les estimer. Si on soupçonne que l'une de ces interactions existe il vaut mieux ne pas utiliser ce dispositif.

de répartition en proportion *optimale* au sens du critère choisi est la règle allouant aux cellules c_1 et c_2 les proportions $p_1 = p_2 = 1/2$.

Règle d'affectation individuelle : randomisation

Dans l'exemple de la section 7.2.1, p. 225, on veut comparer les moyennes de la variable réponse entre les 2 traitements, l'impact du traitement k étant quantifié sur un groupe de n_k unités expérimentales. Une hypothèse implicite est que toute différence observée entre groupes est imputable à la différence de traitements. Autrement dit il n'existe pas de différence initiale entre groupes avant traitement due à une autre cause que le hasard de l'affectation des unités expérimentales. Si cette hypothèse n'est pas vérifiée, il ne sera pas possible de conclure si la différence finale (après traitement) est due à l'administration des traitements ou à la répercussion de la différence initiale.

Le choix de la règle de répartition en proportion, réalisé avant le début de l'expérience, ne peut créer une différence à priori entre groupes d'unités expérimentales. En revanche, la règle d'affectation individuelle des unités expérimentales aux différentes cellules peut générer des *biais de sélection*, en particulier lorsque l'affectation des unités expérimentales est réalisée en fonction de leurs caractéristiques. On conçoit aisément que lorsque l'objectif d'une étude est de comparer deux traitements médicaux, il faut veiller à ne pas affecter les unités expérimentales aux groupes de traitement en fonction de la gravité des symptômes ! Une stratégie pertinente pour l'affectation individuelle est alors le tirage au sort appelé *randomisation*. On distingue deux cas, suivant que l'expérimentation nécessite ou non que l'on utilise un ou des facteur(s) de contrôle d'hétérogénéité.

Randomisation en l'absence de facteur de contrôle d'hétérogénéité Dans ce cas, chaque cellule est assimilable à une unité de traitement, et la randomisation se définit comme suit :

Définition 7.4.2. *Une procédure d'affectation est dite randomisée si toutes les unités expérimentales ont la même probabilité* a priori p_t *d'être affectées à l'unité de traitement* u_t, p_t *étant donnée par la règle de répartition en proportion.*

Autrement dit, si l'on considère la règle de répartition en proportion optimale pour l'exemple 1, la randomisation résultera en une probabilité a priori pour chaque patient de 1/2 d'être affecté au traitement k.

Randomisation en présence de facteur de contrôle d'hétérogénéité Dans ce cas les unités expérimentales ne peuvent souvent pas être affectées aléatoirement aux facteurs de contrôle d'hétérogénéité parce que chaque unité expérimentale a des caractéristiques précises concernant des facteurs de contrôle d'hétérogénéité. Par exemple la classe de productivité d'un animal est fixée à priori et ne peut pas être modifiée. En général il en est de même pour le parc. Il peut arriver cependant que la modalité ne soit pas fixée à priori. Par exemple si on est libre d'attribuer n'importe quel parc à n'importe quel animal, il faut faire cette affectation par tirage au sort. Une fois cette affectation effectuée, il reste à attribuer les animaux aux unités de traitement. Si la règle de répartition en proportion prévoit plusieurs unités de traitement pour un croisement

donné des facteurs de contrôle d'hétérogénéité alors l'affectation est faite par tirage au sort.

Exemple (Production Laitière) On s'intéresse à l'impact de 4 traitements sur la production laitière journalière des vaches. Cet impact pouvant varier suivant la période de lactation, trois périodes de lactation sont considérées. Il y a donc deux facteurs d'intérêt, notés dans la suite A et B. On sait par ailleurs que la production dépend des conditions d'élevage de l'animal (3 parcs considérés) ainsi que de la productivité moyenne de la vache considérée, ces deux derniers facteurs jouant ici le rôle de facteurs de contrôle d'hétérogénéité (facteurs C et D). Les facteurs Traitement, Lactation, Parc et Productivité ont respectivement 4, 3, 3 et 4 modalités. Chaque cellule peut donc être représentée par une série de 4 chiffres $ijk\ell$, $i = 1, ..., 4$, $j = 1, ..., 3$, $k = 1, ..., 3$, $\ell = 1, ..., 4$. Il y a donc en tout $4 \times 3 \times 3 \times 4 = 144$ cellules et 12 unités de traitement possibles. On suppose par ailleurs que l'on dispose d'un troupeau suffisamment grand pour inclure 144 vaches dans l'expérimentation.

On considère un plan complet (toutes les cellules sont observées), avec effectifs égaux et randomisé, ce qui signifie que :

— chaque cellule est représentée avec une proportion 1/144 (règle de répartition en proportion), il y a donc exactement 1 animal alloué à chaque cellule ;
— La règle de répartition en proportion impose que chaque croisement des facteurs de contrôle d'hétérogénéité contienne 12 unités expérimentales. Il y a 12 unités de traitement (i, j). L'attribution des 12 unités de traitements aux 12 unités expérimentales de chaque croisement des facteurs de contrôle d'hétérogénéité se fait par tirage au sort.

Ce type de plan est appelé plan en blocs complets randomisés.

La randomisation est une procédure incontournable en planification expérimentale, garantissant l'absence de biais de sélection. Dans la suite et sauf mention du contraire, nous supposerons systématiquement qu'une fois la règle de répartition en proportion choisie, l'affectation des unités expérimentales se fait par randomisation respectant la règle de répartition en proportion. Nous nous intéresserons donc dans la suite uniquement au choix d'une règle de répartition en proportion.

Confusion d'effets

Dans le cas simple d'une expérience ne prenant en compte qu'un facteur de variabilité, le plan optimal est le plan randomisé avec effectifs égaux (cf. [42], p.59). Lorsqu'il existe plusieurs facteurs de variabilité, le plan optimal dépend du nombre et de la hiérarchisation des questions auxquelles l'expérience doit permettre de répondre, et peut être difficile à identifier. En revanche, il est en général aisé de disqualifier certains plans du fait qu'ils ne permettront pas de répondre à l'ensemble des questions posées. En effet, dès lors qu'un plan n'est pas orthogonal, il existe une confusion, partielle ou totale, entre certains facteurs étudiés. Le niveau de confusion d'un plan se déduit directement de sa règle de répartition en proportion. Plus précisément, si les proportions associées à chaque cellule ne sont pas bien choisies, certains effets seront (au moins partiellement) confondus. Cette confusion peut être fatale à l'analyse.

Le plan introduit en section 7.4.1, p. 232 a des cellules de même effectif, et garantit

donc l'orthogonalité entre les facteurs (et leurs interactions). Il semble donc que ce plan soit toujours la meilleure option. Toutefois, un tel plan n'est pas toujours réalisable voire pertinent, et ce pour plusieurs raisons :

— Les contraintes techniques de l'expérience peuvent rendre impossible en pratique l'observation de l'ensemble des croisements des niveaux des facteurs. Par exemple, la taille des blocs peut être trop petite pour faire apparaître tous les traitements au sein d'un même bloc ;

— Lorsque le nombre de facteurs ou le nombre de modalités par facteur est élevé, il n'est pas envisageable de réaliser un plan ayant des cellules de même effectif du fait du coût prohibitif (en terme de nombre d'unités expérimentales requises) d'un tel plan ;

— Il n'est pas toujours nécessaire de garantir l'absence de confusion d'effets entre toutes les interactions de tous les facteurs. Dans bien des cas, il est vraisemblable que seules les interactions d'ordre faible aient un impact déterminant sur la variable réponse. Il est alors possible de tolérer une confusion partielle, voire totale, entre les interactions d'ordre élevé et les interactions d'ordre faible.

En pratique l'expérimentateur ne pourra donc pas toujours se prémunir complètement de la confusion d'effets. Il lui appartient en revanche de choisir la confusion d'effets et non de la subir, c'est-à-dire d'organiser l'expérience de manière à minimiser l'impact de la confusion sur l'analyse à mener et garantir l'estimabilité des combinaisons linéaires d'intérêt.

Avant de considérer différentes stratégies pour l'organisation de cette confusion, remarquons que la confusion d'effets n'est définie que pour des facteurs fixes. Il peut arriver que la connaissance du niveau d'un facteur aléatoire soit informative sur le niveau d'un facteur fixe, mais cela n'entraînera pas de confusion, au sens où l'effet du facteur aléatoire ne "masquera pas" l'effet du facteur fixe. La superposition d'un effet aléatoire à un effet fixe aura cependant des conséquences sur la variabilité des mesures effectuées sur l'effet fixe (et donc sur la précision avec laquelle l'impact de l'effet fixe est mesuré). Le plan d'expériences a aussi un impact sur la précision des estimateurs des variances des effets aléatoires. Cette question est abordée en section 7.5, p. 247 de ce chapitre.

Organiser la confusion d'effets

Dans tout ce qui suit, on fera l'hypothèse que, pour chaque facteur de contrôle d'hétérogénéité, les nombres d'unités expérimentales ayant une même modalité sont égaux, et que les unités de traitement sont observées un même nombre de fois. Ceci exclut de ce chapitre le cas où l'on ne s'intéresse qu'à un petit sous-ensemble de contrastes entre traitements, ou les cas où ces contrastes sont hiérarchisés, certains étant plus importants pour l'expérimentateur que d'autres. Le lecteur intéressé par des cas où ces deux contraintes ne sont pas respectées peut se référer à [19].

Il existe deux configurations caractéristiques où la confusion d'effets est inévitable. La première est celle où la taille des blocs est fixée (c'est une contrainte de l'expérience) et trop petite pour que toute les unités de traitement soient représentées au moins une fois dans chaque bloc. Ainsi, dans l'exemple présenté en section 7.4.1, p. 232, si la contenance des parcs est inférieure à 12 animaux, il ne sera pas possible d'observer chaque croisement *traitement* × *période* au sein de chaque parc. Les blocs ne sont donc plus complets. La seconde est la configuration où les facteurs d'intérêt (ou leurs

modalités) sont trop nombreux. Il n'est donc pas possible d'observer toutes les unités de traitement.

Ces deux cas sont présentés ici.

7.4.2 Confusion due au nombre trop petit de niveaux d'un ou de plusieurs facteurs de contrôle d'hétérogénéité

Nous commençons par étudier le cas d'une expérience concernant un facteur d'intérêt A et un facteur de contrôle d'hétérogénéité B. Dans ce qui suit, le nombre de modalités du facteur A est noté $m_a = t$, le nombre de niveaux du facteur de contrôle d'hétérogénéité est noté $m_b = b$, la taille des blocs ℓ et le nombre de répétitions par traitement r. Enfin, une unité de traitement correspondant à un niveau du facteur A, ces niveaux seront simplement désignés dans la suite par le terme "traitement".

Un facteur bloc

Lorsque $\ell < t$, il n'est plus possible de mettre toutes les unités de traitement dans un même bloc. Chaque traitement est présent au plus une fois dans un bloc donné, et certains sont absents dans certains blocs, créant ainsi une non-orthogonalité entre les effets blocs et traitement. Le modèle d'analyse statistique des résultats est le même que pour un bloc complet mais les estimateurs des effets blocs et traitements ne sont plus indépendants :

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}, \quad E_{ij} \hookrightarrow \mathcal{N}(0, \sigma^2) \ i.i.d. \ ,$$

où Y_{ij} est la mesure réalisée dans le bloc i pour la modalité j du facteur A.

Exemple (Production laitière, suite) On considère ici une version simplifiée de l'exemple de la section 7.4.1, p. 232 où l'on suppose que
⋆ seuls le facteur traitement (facteur A) et le facteur parc (facteur B) ont une influence sur la production laitière,
⋆ chaque parc ne peut contenir que 2 animaux (blocs de taille 2),
⋆ on dispose de 4 parcs.
Considérons les 2 plans suivants :

Parc	1	2	3	4
Trt	1, 2	2, 1	3, 4	4, 3

Plan 1

Parc	1	2	3	4
Trt	1, 2	2, 3	3, 4	4, 1

Plan 2

On remarque que dans aucun des deux plans les traitements 1 et 3 ne sont comparés au sein d'un même bloc. On peut alors s'interroger sur la possibilité de comparer ces deux traitements, c'est-à-dire d'obtenir un estimateur sans biais de l'écart $\beta_1 - \beta_3$. On montre aisément qu'un tel estimateur est accessible dans le plan 2. En effet, on a :

$$\left.\begin{array}{l} Y_{11} = \mu + \alpha_1 + \beta_1 + E_{11} \\ Y_{12} = \mu + \alpha_1 + \beta_2 + E_{12} \\ Y_{22} = \mu + \alpha_2 + \beta_2 + E_{22} \\ Y_{23} = \mu + \alpha_2 + \beta_3 + E_{23} \end{array}\right\} \Rightarrow \quad E\left(Y_{11} - Y_{12} + Y_{22} - Y_{23}\right) = \beta_1 - \beta_3 \ .$$

Bloc	Traitements				Bloc	Traitements			
I	1	2	3	4	VIII	2	8	5	3
II	5	2	1	6	IX	6	3	2	7
III	2	1	7	8	X	2	4	7	5
IV	7	1	3	5	XI	4	6	2	8
V	8	1	3	6	XII	6	5	4	3
VI	8	1	4	5	XIII	8	4	3	7
VII	1	6	4	7	XIV	5	6	7	8

TABLE 7.1 – Exemple de plan en blocs incomplets équilibrés (8 traitements, 14 blocs).

La combinaison linéaire $Y_{11} - Y_{12} + Y_{22} - Y_{23}$ est bien un estimateur sans biais de l'écart $\beta_1 - \beta_3$. Il est ainsi possible de comparer des unités de traitement qui n'ont jamais été observées dans les mêmes conditions, c'est-à-dire au sein d'un même bloc. Plus généralement, on peut montrer que pour le plan 2 toutes les comparaisons 2 à 2 entre traitements sont estimables. A l'inverse, dans le cas du plan 1 les groupes de traitements $1, 2$ et $3, 4$ sont complètement séparés, dans le sens où jamais un traitement du premier n'est comparé à un traitement du deuxième dans les mêmes conditions. Du fait de cette complète séparation, seules les comparaisons $1-2$ et $3-4$ sont estimables dans le plan 1. Cet exemple illustre une notion plus générale appelée "connexité" d'un plan d'expériences, le lecteur intéressé se référera à [12]. Dans ce livre, nous nous bornerons à présenter des plans d'expérience garantissant l'estimabilité des comparaisons de traitements 2 à 2.

Nous avons vu que le plan 2 est préférable au plan 1 car il garantit l'estimabilité de toutes les comparaisons de traitements 2 à 2. En revanche, chacune de ces comparaisons ne se fait pas avec la même précision : une différence entre traitements comparés au sein d'un même bloc sera estimée avec plus de précision qu'une différence entre traitements comparés au sein de blocs différents. De ce fait la puissance associée au test $H_0 : \{\beta_j = \beta_{j'}\}$ contre $H_1 : \{\beta_j \neq \beta_{j'}\}$ dépend des traitements j et j' à comparer. Si toutes les comparaisons présentent pour l'expérimentateur un même intérêt, ces différences de puissance ne sont pas satisfaisantes. Dans la suite, on cherchera donc dans la mesure du possible à garantir une puissance identique pour l'ensemble des comparaisons 2 à 2.

Plan BIE Le plan en Blocs Incomplets Equilibrés (BIE) constitue un exemple de plan dont les cellules ne sont pas de même effectif (certaines cellules sont vides) mais garantissant l'estimabilité ainsi qu'une puissance identique pour chacune des comparaisons. Un exemple de BIE est traité en détail en section 8.2, p. 262.

Définition 7.4.3. *On appelle plan BIE un plan vérifiant les deux conditions suivantes :*
(i) chaque traitement est présent au plus une fois dans chaque bloc,
(ii) tout couple de traitements apparaît exactement λ fois dans λ blocs différents.

Un exemple de plan BIE pour une comparaison de $t = 8$ traitements répartis dans des blocs de taille $\ell = 4$ est donnée en Table 7.1. Dans cet exemple, on a $r = 7$ et $\lambda = 3$.

L'estimabilité des comparaisons de traitements 2 à 2 est garantie puisque tout couple est comparé au moins une fois au sein d'un même bloc. Il est plus difficile de démontrer que la variance associée à chacune de ces comparaisons est identique. La démonstration

peut être trouvée dans ([19] chapitre 11), et repose sur les propriétés (i) et (ii), qui associées garantissent que le nombre de comparaisons directes (dans un même bloc) et indirectes (via une combinaison connexe) pour deux traitements i et i' sont les mêmes pour tous les couples (i, i').

En pratique, l'existence d'un plan BIE n'est pas garantie, et dépend des quantités t, b, ℓ et r. On peut montrer que pour des valeurs données de t et ℓ (qui sont en général les données du problème), un plan BIE existe si les conditions (nécessaires mais non suffisantes) suivantes sont vérifiées :

$$\begin{aligned} rt &= \ell b \\ r(\ell - 1) &= \lambda(t - 1) \end{aligned}$$

où t, b, l et r sont des nombres entiers.

Lorsqu'il n'existe pas de plan BIE, des plans alternatifs dits en blocs incomplets partiellement équilibrés peuvent être envisagés. Un exemple d'un tel plan est donné en section 8.5, p. 275. L'équilibre n'étant pas garanti (au sens de la définition 7.4.3, p. 236), ces plans ont des propriétés plus pauvres que celles du plan BIE : l'estimabilité des comparaisons entre traitements est garantie, mais le nombre de comparaisons directes varie d'un couple de traitements à l'autre. Ces plans, moins contraignants, ne garantissent donc pas une précision identique pour toutes les comparaisons.

Les contraintes données par la définition 7.4.3, p. 236 et vérifiées par un plan BIE sont équivalentes à des contraintes directement imposées sur la règle de répartition en proportion. Par exemple, la règle de répartition en proportion correspondant à un plan BIE doit entre autres vérifier :

$$p_{+j} = \sum_{i=1}^{b} p_{ij} \quad = \quad r/n \quad ,$$

où p_{ij} est la proportion d'unités expérimentales associée à la cellule c_{ij}. Toutefois, il est souvent plus simple d'exprimer les contraintes portant sur le plan d'expériences (comme celles données en Définition 7.4.3, p. 236) plutôt que celles portant directement sur la règle de répartition en proportion. Dans la suite, on gardera donc à l'esprit l'équivalence entre les deux, et seules les contraintes portant sur le plan d'expériences seront explicitées.

Plusieurs facteurs de contrôle d'hétérogénéité

L'exemple de la production laitière de la section 7.4.1, p. 231 contient deux facteurs de contrôle d'hétérogénéité : le parc (facteur B, m_b modalités) et la productivité moyenne (facteur C, m_c modalités). On ne considère toujours que le facteur d'intérêt Traitement. Supposons que les quantités m_b et m_c n'imposent pas de contrainte sur l'analyse et peuvent être choisies librement. Une première stratégie est alors de considérer qu'il existe un facteur bloc "général" G, dont chacun des $m_g = m_b \times m_c$ niveaux correspond au croisement d'un niveau de facteur B avec un niveau du facteur C. On

peut ainsi se ramener au cas d'un seul facteur bloc, cas traité dans la section précédente, et employer le modèle suivant pour l'analyse des données du plan d'expériences :

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}, \quad E_{ij} \hookrightarrow \mathcal{N}(0, \sigma^2) \ i.i.d. \ . \tag{7.1}$$

Les notations sont identiques à celles du paragraphe précédent, à l'exception de l'indice i qui désigne maintenant les différents niveaux du facteur G. Le facteur G ayant $m_b \times m_c$ modalités, l'estimation des paramètres associés à ce facteur nécessitent $m_b \times m_c - 1$ degrés de libertés, incluant à la fois l'estimation des effets principaux et de l'interaction entre les facteurs B et C. Lorsque m_b et m_c sont grands, un nombre conséquent de degrés de liberté est donc affecté à l'estimation de paramètres ne présentant pas d'intérêt pour l'expérimentateur. Ce plan est rarement utilisé car trop coûteux en nombre d'unités expérimentales ($n = m_b \times m_c \times m_A$). En l'absence d'interaction entre les facteurs de contrôle d'hétérogénéité (ou lorsque cette interaction est supposée négligeable), une partie de ces degrés de liberté est utilisée inutilement, et pourrait être réaffectée à l'estimation de la variance résiduelle. Plutôt que de considérer un facteur bloc général il faut alors modifier l'écriture du modèle pour prendre en compte à la fois la présence de deux facteurs de contrôle d'hétérogénéité et l'absence d'interaction entre ces 2 facteurs :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + E_{ijk}, \quad E_{ijk} \hookrightarrow \mathcal{N}(0, \sigma^2) \ i.i.d. \ , \tag{7.2}$$

où Y_{ijk} est la réponse au traitement k mesurée dans les blocs i et j des facteurs de contrôle d'hétérogénéités B et C. Dans cette écriture, on constate que $m_b + m_c - 1$ plutôt que $m_b \times m_c - 1$ degrés de libertés seront consacrés à l'estimation des paramètres associés aux facteurs B et C. D'où la nécessité - du moins la pertinence - de développer des plans d'expérience ad hoc pour l'analyse des données en présence de 2 facteurs de contrôle d'hétérogénéité.

Carrés Latin La particularité de ce plan, déja vu en section 7.3.2, p. 230 est qu'il n'admet qu'une unité expérimentale par croisement de modalités des facteurs de contrôle d'hétérogénéité et il impose que $m_a = m_b = m_c$.

Définition 7.4.4. *On appelle plan en carré latin un plan vérifiant les conditions suivantes :*
(i) chaque croisement des facteurs de contrôle d'hétérogénéité B et C est représenté exactement une fois,
(ii) chaque traitement est présent exactement une fois dans chaque modalité du facteur B,
(iii) chaque traitement est présent exactement une fois dans chaque modalité du facteur C.

Pour une comparaison de m_a traitements, un plan en Carré Latin comporte donc m_a^2 unités expérimentales, et peut être représenté sous la forme d'un tableau croisé, où colonnes et lignes correspondent respectivement aux modalités des facteurs de contrôle d'hétérogénéité B et C, et chaque cellule du tableau indique quelle unité de traitement a été affectée au croisement. Dans ce plan l' effet de chaque facteur est partiellement confondu avec l'interaction des 2 autres facteurs. Les interactions sont supposées nulles on les utilise pour estimer la variance résiduelle. Un exemple de carré latin pour $m_a = 4$ traitements est donné dans la table 7.2.

		Facteur C			
		I	II	III	IV
Facteur B	I	1	2	3	4
	II	2	3	4	1
	III	3	4	1	2
	IV	4	1	2	3

		Facteur C		
		I	II	III
Facteur B	I	1	2	3
	II	2	3	4
	III	3	4	1
	IV	4	1	2

TABLE 7.2 – Exemples d'un plan en carré latin (gauche) et d'un plan de Youden (droite) pour une expérimentation comparant $m_a = 4$ traitements. Les facteurs de contrôle d'hétérogénéité sont indicés en numérotation romaine, les traitements en numérotation arabe.

L'analyse des données du plan se fait via le modèle (7.2), et on vérifie que

$$E\left(Y_{\cdot\cdot k} - Y_{\cdot\cdot l}\right) = \beta_k - \beta_l \quad ,$$

ce qui démontre que toutes les comparaisons entre traitements sont estimables avec la même précision. De plus il y a orthogonalité entre les facteurs ligne, colonne et traitement.

Plan de Youden En l'absence de contrainte pratique sur le nombre de niveaux des facteurs de contrôle d'hétérogénéité, il est toujours possible de construire un carré latin. Lorsque la taille de l'un des deux facteurs de contrôle d'hétérogénéité est contrainte ($m_b < m_a = t$ ou $m_c < m_a$), on utilise un plan de Youden :

Définition 7.4.5. *On appelle plan de Youden un plan vérifiant les conditions suivantes :*
(i) les facteurs B et C ont respectivement $m_b = m_a$ et $m_c < m_a$ modalités,
(ii) chaque traitement est présent exactement une fois dans chaque modalité du facteur B,
(iii) chaque traitement est présent au plus une fois dans chaque modalité du facteur C,
(iv) les couples de traitements apparaissent un même nombre de fois ensemble pour un même niveau du facteur C.

On a donc un dispositif en blocs complets (pour le facteur B) croisé avec un dispositif en blocs incomplets équilibrés (pour le facteur C). Comme les plans en carré latin, les plans de Youden garantissent l'estimabilité des comparaisons de traitements avec même variance. En revanche, il n'y a plus orthogonalité entre les facteurs et donc les estimations des effets des facteurs A et C ne sont plus indépendantes. Par ailleurs, notons qu'à cause de la condition (iv) il n'existe pas de plan de Youden pour toutes les configurations (m_b, m_c). Comme pour les plans BIE, il est possible de lever cette condition, et de considérer des plans "en ligne-colonne" quelconques, voir [16]. L'estimabilité des contrastes d'intérêt n'est alors pas systématiquement garantie, et la précision de l'estimation peut être sensiblement différente d'un contraste à l'autre.

7.4.3 Confusion d'effets pour traiter le cas d'un grand nombre d'unités de traitement, plans fractionnaires

On suppose ici que l'expérimentation à mener a une visée exploratoire : p facteurs d'intérêt sont considérés, et plutôt que de modéliser finement la relation entre ces facteurs et la variable réponse, on cherche à discriminer en peu d'expériences les facteurs

(ou les interactions) susceptibles d'avoir un impact sur la réponse de ceux qui n'en ont pas. Cette démarche est illustrée dans l'exemple 8.6, p. 281. Nous restreignons ici notre étude au cas particulier où tous les facteurs d'intérêt considérés ont deux modalités. Il y a donc 2^p unités de traitement, mais le budget étant limité (ou chaque expérience coûteuse), seule une fraction $1/2^m$ de ces unités sont effectivement incluses dans le plan, une seule fois (pas de répétition). Un tel plan est appelé *plan fractionnaire* et se note 2^{p-m}, où 2 représente le nombre de modalités par facteur, p le nombre de facteurs d'intérêt et m désigne l'ordre de la fraction. La valeur 2^{p-m} représente donc le nombre d'expériences qui seront effectivement réalisées.

Notons que comparer toutes les unités de traitement 2 à 2 représente un ensemble de $2^p - 1$ contrastes et nécessite donc autant de degrés de liberté. Il n'y a donc pas assez d'observations pour estimer l'ensemble des contrastes. Le nombre de contrastes estimables étant limité, on part de l'hypothèse que plus une interaction a un ordre élevé, plus elle est susceptible d'être négligeable : un effet principal a plus de chance d'avoir un réel impact sur la réponse qu'une interaction d'ordre 4. De ce fait, il est souhaitable d'organiser la confusion entre les facteurs de manière à garantir que les contrastes portant sur les interactions d'ordre faible soient estimables, quitte à confondre les interactions d'ordre élevé.

Exemple

On présente ici le principe de mise au point d'un plan fractionnaire sur un exemple simple comportant 3 facteurs d'intérêt A, B et C. Le modèle d'analyse d'un tel plan peut s'écrire sous les formes régulière ou singulière :

$$Y_{ijk} = \mu_{ijk} + E_{ijk} \tag{7.3}$$

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + (\alpha\beta\delta)_{ijk} + E_{ijk} \quad . \tag{7.4}$$

La première écriture fait apparaître un paramètre par unité de traitement, tandis que la seconde explicite l'impact de chacun des facteurs A, B et C et de leurs interactions sur la variable réponse. Rappelons qu'une fois les contraintes d'identifiabilité choisies, les deux écritures sont équivalentes. Nous travaillerons donc ici avec l'écriture régulière. Observons par ailleurs qu'il n'y a pas d'indice de répétition. Le plan d'expériences qui réalise toutes les 2^p expériences possibles s'appelle un plan factoriel complet (tout croisement de modalités $A \times B \times C$ est observé exactement une fois). Si on utilise ce plan, et si tous les paramètres de moyenne du modèle doivent être estimés, il ne reste aucun degré de liberté disponible pour l'estimation de la variance. Le modèle peut s'écrire de manière matricielle

$$Y = X\theta + E.$$

Dans cette dernière écriture, l'espace vectoriel engendré par les colonnes de X est de dimension 8. L'ensemble des combinaisons linéaires estimables est donc de dimension 8, autrement dit il existe un système de 8 combinaisons des paramètres engendrant l'ensemble des combinaisons linéaires estimables. Un tel système est donné dans la table 7.3. Chaque colonne de cette table représente une combinaison linéaire estimable, tandis que chaque ligne représente l'une des 8 expériences possibles. On remarque que les combinaisons linéaires sont associées à la constante (colonne I pour *Intercept*) aux facteurs ou à leurs interactions : la colonne 3 est la colonne associée à B. Cela est dû au fait que

	I	A	B	C	AB	AC	BC	ABC
μ_{111}	+1	+1	+1	+1	+1	+1	+1	+1
μ_{112}	+1	+1	+1	-1	+1	-1	-1	-1
μ_{121}	+1	+1	-1	+1	-1	+1	-1	-1
μ_{122}	+1	+1	-1	-1	-1	-1	+1	+1
μ_{211}	+1	-1	+1	+1	-1	-1	+1	-1
μ_{212}	+1	-1	+1	-1	-1	+1	-1	+1
μ_{221}	+1	-1	-1	+1	+1	-1	-1	+1
μ_{222}	+1	-1	-1	-1	+1	+1	+1	-1

TABLE 7.3 – Contrastes associés aux tests sur les facteurs et leur interactions, dans un modèle ANOVA à 3 facteurs. Chaque colonne donne la liste des coefficients associés aux moyennes des unités de traitement.

la combinaison linéaire 3 est exactement la combinaison linéaire qui serait utilisée pour estimer l'effet du facteur B dans le cas d'un plan à cellules de même effectif. Remarquons qu'il est aisé de construire une telle table de contraste. Il suffit pour cela de commencer par remplir les colonnes associées à la constante (qui prend toujours le même niveau : 1), puis de détailler les différentes possibilités de combinaisons de modalités des facteurs, ligne par ligne. Une fois les colonnes des effets principaux construites, les interactions d'ordre 2 ou supérieures sont obtenues par simple multiplication des colonnes des facteurs intervenants dans l'interaction. Par exemple, la colonne de l'interaction AB est obtenue en multipliant les valeurs de la colonne de A par les valeurs de la colonne de B, ligne par ligne. Chacune des colonnes ainsi obtenues est appelée "contraste", car pour chaque combinaison linéaire les coefficients somment à 0 [3]. Enfin, on constate aisément que ces différents contrastes sont orthogonaux entre eux : le produit scalaire de deux colonnes du tableau vaut 0.

Si l'on réalise l'ensemble des 8 expériences correspondant au 8 lignes de la table, c'est-à-dire si le plan est complet et avec des cellules de même effectif, tous les contrastes sont estimables et peuvent être estimés indépendamment les uns des autres (puisque les colonnes sont orthogonales les unes aux autres). Supposons maintenant que les deux premières expériences ne soient pas réalisées faute de budget, ce qui revient à supprimer les 2 premières lignes de la table 7.3. On vérifie alors que les contrastes ne sont plus orthogonaux. De ce fait, il n'est par exemple plus possible d'estimer l'effet du facteur A indépendamment de l'effet du facteur B : on retrouve ici la confusion d'effet, conséquence du plan qui n'est plus complet. Lorsque le choix des expériences (des lignes) à retirer est arbitraire, les confusions d'effets ne sont pas maîtrisées, et risquent de porter sur l'ensemble des facteurs indistinctement. A l'inverse, un choix raisonné des expériences va permettre de structurer la confusion d'effets.

Considérons maintenant un plan fractionnaire 2^{3-1} avec, par définition, 4 expériences. Puisque par hypothèse il est plus important d'estimer les contrastes portant sur les interactions d'ordre faible, la stratégie consiste à confondre complètement le contraste associé à l'interaction d'ordre 3 avec la constante. Cela signifie que les 4 ex-

3. On remarque que l'utilisation du mot contraste pour désigner la combinaison linéaire portant sur la constante est fausse : les coefficients ne somment pas à 0. Toutefois cet abus de langage étant couramment utilisé dans la bibliographie, nous le reproduisons ici.

périences sélectionnées dans le plan (i.e. pour les lignes de la table 7.3) doivent être tels que les colonnes I et ABC résultantes soient colinéaires. Quelles que soient les 4 lignes sélectionnées, la colonne I s'écrira après sélection $(1, 1, 1, 1)$. Il faut donc sélectionner 4 lignes de la table 7.3 de manière à ce que le vecteur des coefficients associé au contraste ABC vale $(1, 1, 1, 1)$ ou $(-1, -1, -1, -1)$.

Comme le montre le tableau 7.4, il apparaît que seuls deux choix sont possibles : la sélection des observations correspondant aux traitements $\{111, 122, 212, 221\}$, ou celle correspondant aux traitements $\{112, 121, 211, 222\}$. En désignant la constante par la lettre I, le premier choix est parfois nommé "fraction principale" et la notation associée est $I = ABC$ (les contrastes associés à I et ABC sont identiques), tandis que le second choix est nommé "fraction complémentaire, de notation $I = -ABC$ (les contrastes sont identiques au signe près).

Considérons le plan correspondant à la fraction principale. L'analyse de l'ensemble des colon-nes montre que le choix des 4 observations correspondant aux traitements $\{111, 122, 212, 221\}$ n'implique pas uniquement la confusion entre I et ABC. En effet, les colonnes correspondant aux contrastes associés à A et BC sont identiques. Il y a donc aussi confusion entre le facteur principal A et l'interaction BC. Cette confusion, observée à partir de la table 7.4, peut aussi être déduite de la relation $I = ABC$. En effet, en utilisant les règles de calcul $A \times A = I$ et $I \times A = A \times I = A$, la commutativité et l'associativité, on a :

$$ABC = I \Rightarrow A \times ABC = A \times I \Rightarrow BC = A \quad .$$

Des calculs analogues montrent qu'il y a aussi confusion totale entre les facteurs B et AC, ainsi qu'entre les facteurs C et AB.

Le plan considéré ici définit donc 4 groupes de facteurs : $\{A, BC\}$, $\{B, AC\}$, $\{C, AB\}$ et $\{I, ABC\}$, tels que les facteurs issus d'un même groupe sont complètement confondus, et les groupes entre eux sont orthogonaux (ceci se déduit encore de l'observation de la table 7.4). Remarquons que la stratégie consistant à confondre l'interaction ABC avec la constante est pertinente car elle préserve l'orthogonalité entre deux contrastes portant sur deux effets principaux différents, ce qui correspond à l'objectif qui était fixé. La stratégie développée dans les plans fractionnaires consiste donc à confondre volontairement entre eux des facteurs d'intérêt. On parle alors non plus de confusion mais "d'aliasage" entre facteurs. Dans le plan considéré, l'alias $I = ABC$ a été choisi, et nous avons vu qu'il est possible de déduire l'ensemble des relations (confusion totale ou orthogonalité) entre facteurs à partir de cette première relation. Les relations ainsi obtenues sont équivalentes, et il aurait été possible d'obtenir le même plan en partant de l'objectif de confondre les effets A et BC. Toutefois, la dénomination des plans fractionnaires se fait en fonction de la ou les relations d'aliasage impliquant la constante I. La relation $I = ABC$ est appelée *relation génératrice* du plan, les effets impliqués dans cette relation génératrice sont appelés les *mots* du plan, et la *résolution* du plan est définie comme étant la taille du plus petit mot de la relation génératrice. Le plan proposé ici est donc de résolution 3, puisqu'ABC est le seul mot apparaissant dans la relation génératrice, et le plan est noté 2^{3-1}_{III}. Remarquons enfin que dans l'exemple traité ici tous les degrés de liberté disponibles ont été alloués à l'estimation de contrastes. On ne dispose donc d'aucun degré de liberté pour l'estimation de la variance résiduelle, qui est supposée connue (ou estimée) par ailleurs.

	I	A	B	C	AB	AC	BC	ABC
μ_{111}	**+1**	+1	+1	+1	+1	+1	+1	**+1**
μ_{122}	**+1**	+1	-1	-1	-1	-1	+1	**+1**
μ_{212}	**+1**	-1	+1	-1	-1	+1	-1	**+1**
μ_{221}	**+1**	-1	-1	+1	+1	-1	-1	**+1**
μ_{211}	**+1**	-1	+1	+1	-1	-1	+1	**-1**
μ_{112}	**+1**	+1	+1	-1	+1	-1	-1	**-1**
μ_{121}	**+1**	+1	-1	+1	-1	+1	-1	**-1**
μ_{222}	**+1**	-1	-1	-1	+1	+1	+1	**-1**

TABLE 7.4 – Les deux plans 2^{3-1} amenant à confondre l'interaction ABC avec la constante. Les colonnes associées à I et ABC sont identiques dans les deux cas.

Quelques résultats sur les plans fractionnaires

Fractions d'ordre supérieur Comment se généralise la stratégie présentée au paragraphe précédent lorsque l'ordre m de la fraction est supérieur à 1 ? On observe tout d'abord que l'ordre de la fraction correspond au nombre d'alias qu'il faudra fixer pour réaliser le plan. Ainsi une fraction 2^{p-2} nécessitera l'aliasage de deux interactions avec la constante I. Il faut ensuite choisir les aliasages. Dans l'exemple précédent, le premier alias choisi était celui de I avec l'interaction ABC, c'est-à-dire avec l'interaction d'ordre le plus élevé du modèle. Cette stratégie garantissait l'indépendance des contrastes portant sur les effets principaux A, B et C. Cette stratégie peut être systématisée quel que soit le nombre de facteurs lorsque l'ordre de la fraction est 1. En revanche, lorsque l'ordre de la fraction est plus élevé, la stratégie n'est plus pertinente.

Considérons un plan 2^{5-2} portant sur 5 facteurs. Supposons que l'on fixe comme premier alias

$$I = ABCDE \quad .$$

Un deuxième alias doit être ensuite sélectionné, la fraction étant d'ordre 2. Ce deuxième alias ne peut pas être de la forme $I = A$ (ou $I = B$, les rôles des facteurs sont ici interchangeables), car cela confondrait un effet principal avec la constante. Cela nécessiterait de prendre toujours la même modalité du facteur A pour toutes les observations du plan, ce qui empêcherait de quantifier l'impact dû aux variations du facteur A. Pour la même raison, il n'est pas possible d'aliaser I avec une interaction d'ordre 4. En effet, si l'on choisit $I = ABCD$ par exemple, en multipliant ce deuxième alias avec le premier avec les règles de calcul présentées précédemment, on obtient

$$I \times I = ABCDE \times ABCD \Leftrightarrow I = (A \times A)(B \times B)(C \times C)(D \times D)E \Leftrightarrow I = E \quad .$$

Enfin, un alias de type $I = ABC$ n'est pas admissible (et donc par complémentarité un alias de type $I = AB$ non plus) puisqu'il résulterait en une confusion entre deux effets principaux. Par exemple, si l'on choisit $I = ABC$, on a

$$I = ABC = ABCDE \Rightarrow DE = I \Rightarrow D = E \quad .$$

Ainsi, si le choix du premier alias est $I = ABCDE$, quel que soit le choix du second alias il y aura confusion entre deux facteurs principaux, ou entre un facteur principal et la constante.

I	ADE	ABC	$BCDE$
A	BC	DE	$ABCDE$
B	AC	CDE	$ABDE$
C	AB	BDE	$ACDE$
D	AE	BCE	$ABCD$
E	AD	BCD	$ABCE$
BD	CE	ACD	ABE
BE	CD	ACE	ABD

TABLE 7.5 – Liste des alias pour un plan 2^{5-2} de relation génératrice $I = ABC = ADE$. Les effets d'ordre inférieur à 2 sont indépendants.

Quel choix réaliser alors ? On montre que le choix de la relation génératrice $I = ABC = ADE$ garantit l'orthogonalité entre les contrastes portant sur la constante et les effets principaux. La liste des effets confondus pour un plan 2^{5-2} ayant pour relation génératrice $I = ABC = ADE$ est donnée en table 7.5. Il est aisé de vérifier que le choix proposé pour la relation génératrice n'est pas unique. Comme on le voit, les garanties d'estimabilité des contrastes dépendent des alias choisis et de la fraction du plan. Dans le cas général, il n'existe pas de stratégie systématique pour déterminer la relation génératrice "optimale" en terme d'estimabilité. Toutefois de nombreux cas particuliers ont été décrits dans la littérature, et le lecteur pourra se référer à [19] pour obtenir la liste des relations génératrices associées à plusieurs plans fractionnaires usuels.

Résolution L'exemple précédent montre qu'il est possible pour une fraction 2^{5-2} de construire un plan garantissant l'indépendance des effets principaux. Était-il possible de construire un plan garantissant l'indépendance entre effets d'ordre inférieurs ou égal à 2 ? Cette question est directement reliée à la résolution du plan fractionnaire considéré. La résolution d'un plan a été définie comme l'ordre de l'interaction la plus faible (hormis la constante) apparaissant dans la relation génératrice. Remarquons que pour obtenir cet ordre le plus faible, il faut expliciter l'ensemble des alias associés à la constante, et non seulement les alias proposés. Ainsi, dans l'exemple du plan fractionnaire 2^{5-2}, le choix des alias $I = ABCDE = CDE$ ne résulte pas en un plan de résolution 3. En effet

$$ABCDE = CDE \Rightarrow AB = I \quad ,$$

l'interaction la plus faible confondue avec la constante est d'ordre 2, et le plan est de résolution 2.

Une fois la résolution du plan établie, il est possible d'apporter des garanties concernant l'orthogonalité des contrastes via la proposition suivante [50] [4] :

Propriété 7.4.1. *Un plan de résolution R garantit l'orthogonalité entre tous les contrastes portant sur des interactions d'ordre inférieur à $E[(R-1)/2]$, où $E[u]$ désigne la partie entière de u.*

En résumé, l'orthogonalité entre contrastes portant sur les effets d'ordres faibles est garantie par la résolution du plan. Pour un plan donné il est donc aisé de déterminer à

4. Notons que dans cet ouvrage, définition et caractérisation de la résolution sont inversées par rapport à la présentation qui en est faite ici : un plan de résolution R est *caractérisé* par le fait que les effets confondus avec la constante sont au moins d'ordre R.

partir de la relation génératrice si les effets d'intérêt seront confondus entre eux ou non. Par ailleurs, la résolution dépendant du nombre de facteurs p et de l'ordre de la fraction m, il est clair que plus la fraction sera faible plus la résolution sera faible, ce qui signifie que pour garantir un niveau de résolution donné pour un plan factoriel fixé, il faut être en mesure de réaliser suffisamment d'expériences (i.e. d'éviter un ordre de la fraction trop élevé). Le facteur limitant des expérimentations étant souvent le nombre d'unités expérimentales, il peut être utile de déterminer, pour un nombre d'unités expérimentales fixé, quel est le nombre de facteurs maximum p_{max} qu'il est possible d'intégrer à l'expérimentation pour garantir un niveau de résolution donné. Il n'existe pas de formule exacte pour calculer p_{max}, mais une majoration de ce nombre est proposée dans [50] dans les cas de plans de résolution 3, 4 ou 5.

Plans fractionnaires en présence d'un facteur bloc

Lorsqu'il est impossible de maintenir les conditions expérimentales identiques pour toutes les expériences, on réalise un plan en bloc randomisé : les conditions expérimentales sont homogènes au sein d'un même bloc et différentes d'un bloc à l'autre. L'introduction d'un facteur bloc a plusieurs conséquences. D'une part, elle va nécessiter l'allocation d'un degré de liberté à l'estimation du contraste associé à ce facteur bloc. Ce degré de liberté est donc perdu pour l'estimation des contrastes portant sur les facteurs d'intérêt. D'autre part, il va falloir organiser d'une part l'aliasage entre facteurs d'intérêt, et d'autre part la confusion entre les facteurs d'intérêt et le facteur bloc, de manière à conserver l'estimabilité du plus grand nombre possible de contrastes d'intérêt. Comme dans tout dispositif en blocs, on supposera que les interactions entre le facteur bloc et les facteurs d'intérêt sont nulles ou négligeables.

Est-il possible de gérer séparément l'aliasage et la confusion ? Autrement dit, peut-on choisir indépendamment la relation génératrice et la répartition des traitements dans les différents blocs ? Reprenons l'exemple d'un plan à 5 facteurs, et supposons maintenant que la fraction est 1/2, et que l'expérience nécessite deux blocs de taille 2^3. On notera ici Z le facteur bloc. Si seule la fraction du plan est prise en compte pour choisir la relation génératrice, la fraction étant d'ordre 1, on choisit la relation $I = ABCDE$, qui garantit en l'absence de blocs un plan de résolution 5, et donc l'orthogonalité des contrastes portant sur les effets principaux et les interactions d'ordre 2. Il s'agit ensuite de confondre le facteur bloc avec une autre interaction. Comme pour le cas d'une fraction d'ordre supérieur, les choix $Z = A$ ou $Z = ABCD$ résultent en une confusion entre le facteur bloc et un effet principal (A dans le premier cas, E dans le second) et ne sont donc pas pertinents. Il faut donc choisir une relation de la forme $Z = AB$ ou $Z = ABC$, ce qui dans les deux cas résulte en une confusion complète entre le facteur bloc et une interaction d'ordre 2 (AB dans le premier cas, DE dans le second). Considérons maintenant le choix de $I = ABCD$ pour la relation génératrice et $Z = CDE$. La table 7.6 donne la liste des effets confondus pour ce plan et pour le plan défini par $I = ABCDE$ et $Z = ABC$. Comme on le voit, le premier plan considéré induit une confusion totale entre l'interaction AB et le facteur Z, mais toutes les autres interactions d'ordre 2 sont associées à des contrastes orthogonaux. On perd donc complètement l'information portant sur une interaction d'ordre 2, mais celle portant sur les autres interactions du même ordre est préservée. Autrement dit, exception faite de l'interaction AB qui n'est

Plan $I = ABCDE$, $Z = AB$	
$I = ABCDE$	$AD = BCE$
$A = BCDE$	$AE = BCD$
$B = ACDE$	$BC = ADE$
$C = ABDE$	$BD = ACE$
$D = ABCE$	$BE = ACD$
$E = ABCD$	$CD = ABE$
$Z = AB = CDE$	$CE = ABD$
$AC = BDE$	$DE = ABC$

Plan $I = ABCD$, $Z = CDE$	
$I = ABCD$	$AC = BD$
$A = BCD$	$AD = BC$
$B = ACD$	$AE = BCDE$
$C = ABD$	$BC = AD$
$D = ABC$	$BD = AC$
$E = ABCDE$	$BE = ACDE$
$Z = CDE = ABE$	$CE = ABDE$
$AB = CD$	$DE = ABCE$

TABLE 7.6 – Deux plans 2^{5-1} avec un facteur bloc Z de taille 2.

plus estimable, le plan "reste" pour les autres facteurs de résolution 5. D'un autre coté, le deuxième plan considéré est de résolution 4 pour tous les facteurs. En particulier, on ne pourra pas distinguer l'effet de l'interaction AB de celui de l'interaction CD, mais il sera possible de conclure si l'un de ces deux effets a un effet significatif ou non. On ne perd donc pas complètement l'information sur AB, en revanche on dégrade de manière générale la résolution, puisque les contrastes associés aux interactions d'ordre 2 ne sont plus orthogonaux. Le choix entre ces deux plans se fera donc en fonction de la connaissance a priori dont on dispose sur les facteurs d'intérêt : si l'interaction entre 2 facteurs d'intérêt donnés est supposée négligeable (ou non intéressante) a priori, le choix se portera sur le premier plan. Si aucune connaissance n'est disponible a priori, on réalisera le second plan. De manière générale, les choix portant sur la confusion et l'aliasage doivent donc être pensés simultanément. Il existe un paquet R appelé *planor* [38] pour construire de tels plans .

Analyse d'un plan fractionnaire

Comme pour les plans précédents, le modèle utilisé pour l'analyse des données issues d'un plan fractionnaire est le modèle d'analyse de la variance. Il est toutefois nécessaire de prendre quelques précautions particulières voire de faire appel à une méthodologie spécifique pour réaliser l'analyse d'un plan fractionnaire. Une première spécificité réside dans la liste des facteurs à faire apparaître dans le modèle, qui doit respecter les contraintes de confusion et d'aliasage choisies. Concrètement, si l'on considère par exemple le plan fractionnaire avec blocs défini par $I = ABCD$ et $Z = CDE$ (table 7.6, droite), seul l'un des effets apparaissant dans chacune des lignes peut être rentré dans le modèle d'analyse. Si deux effets aliasés sont inclus dans le modèle, les sommes de carrés de type II ou III associées à ces effets seront automatiquement égales à 0, la confusion entre deux effets aliasés étant totale. Par ailleurs, en l'absence d'estimation de la variance, il n'est pas possible de réaliser un test, pour déterminer quels sont les alias ayant un effet significatif.

Dans le cas le plus général, un plan factoriel comporte un nombre d'unités expérimentales inférieur au nombre d'unités de traitement possible. C'est pourquoi même en l'absence de facteur bloc aucun degré de liberté n'est disponible pour l'estimation de la variance si on inclut dans le modèle tous les termes possibles. Il faut donc avoir une procédure pour estimer la variance résiduelle. Il existe plusieurs possibilités :

1. Le plus simple consiste à répéter quelques expériences en quelques points

2. Il est possible de récupérer des degrés de libertés pour l'estimation de la variance lorsque l'on a la connaissance a priori que certaines interactions sont nulles ou négligeables. Dans le cas d'un modèle à 5 facteurs d'intérêt et sans facteur bloc, de fraction 1/2 et de résolution 4, l'une des relations d'aliasage porte sur des interactions d'ordre 3 uniquement (on peut l'observer en table 7.6, en oubliant la contrainte portant sur le bloc : on a $CDE = ABE$). Si l'on suppose que toutes les interactions d'ordre supérieur à 2 sont nulles, cette relation disparaît. Il y a donc un contraste en moins à estimer, ce qui libère un degré de liberté pour la somme des carrés résiduelles (et donc pour l'estimation de la variance).
3. On peut aller plus loin dans cette stratégie en supprimant du modèle tous les termes qui ne semblent pas exister. Il faut cependant utiliser une procédure qui ne soit pas arbitraire. Il existe des procédures graphiques pour déterminer les effets dits actifs, le plus populaire étant le graphique des quantiles demi-normaux [50]. L'idée de base sous-jacente à cette procédure est la suivante. Si un effet est nul son estimateur est distribué selon une loi normale d'espérance nulle. Donc les estimateurs des effets nuls sont des variables aléatoires d'espérance nulle et de même variance. On construit le graphique dit "qqplot" des estimateurs des effets et on décide que la partie linéaire contient les effets qui sont nuls. On peut alors les supprimer du modèle et les inclure dans la somme des carrés résiduelle. Il existe d'autres méthodes comme par exemple celle définie par [39].

Les méthodes 2 et 3 comportent la possibilité d'inclure dans la résiduelle des effets qui ne sont pas nuls en réalité. Cela implique que l'estimateur de σ^2 surestime ce dernier ce qui conduit à une diminution de la puissance. Par contre le niveau des tests est inférieur ou égal au niveau nominal, ce qui signifie qu'il n'y a pas plus de faux positifs qu'attendu parmi les effets détectés.

La présentation faite ici pour les plans fractionnaires est limitée à l'étude de facteurs binaires. L'ensemble des concepts présentés peut être appliqué à l'étude de facteurs à 3 niveaux, voire à l'étude de plans considérant un mélange de facteurs à 2 ou 3 niveaux, au prix de difficultés (essentiellement techniques) supplémentaires. D'autres travaux ont par ailleurs été réalisés sur l'analyse de plans fractionnaires avec plus d'un facteur de contrôle d'hétérogénéité, ou avec un facteur bloc à plus de deux modalités. Le lecteur intéressé se reportera à l'ouvrage [40], qui traite de nombreux cas de plans fractionnaires irréguliers (tous les facteurs n'ont pas le même nombre de modalités) ou comportant plus d'un facteur de contrôle d'hétérogénéité. En pratique on peut utiliser le paquet R *Planor* pour construire un plan factoriel régulier correspondant à ses besoins.

7.5 Plans d'expériences et modèle mixte

7.5.1 Expériences à facteurs aléatoires

Lorsque l'un des facteurs de l'expérience est aléatoire cela n'a plus de sens de chercher à répartir les niveaux de ce facteur parmi les unités expérimentales. L'objectif de la planification expérimentale est alors de faire en sorte de pouvoir estimer correctement la variabilité du facteur aléatoire et de la différencier de la variabilité individuelle.

Considèrons une expérience avec un facteur aléatoire A pour lequel I niveaux ont été tirés, et lors de laquelle on a observé n_i répétitions par niveau pour une variable réponse Y. Y peut être par exemple la quantité de graisse de la viande de porcs issus

de I pères différents, l'efficacité de produits issus de I fabriques différentes, ... (cf 5.1.1, p. 166). Le modèle s'écrit :

$$Y_{ij} = \mu + A_i + E_{ij} \qquad i = 1\ldots, I \quad j = 1, \ldots, n_i$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \qquad A_i \sim \mathcal{N}(0, \sigma_A^2)$$

et les E_{ij} et les A_i sont indépendants.
Les estimateurs de σ^2 et σ_A^2 sont donnés par (cf 5.2.3, p. 177)

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n-I}\sum_{i,j}(Y_{ij} - Y_{i.})^2 = SCR/(n-I)\\ \widehat{\sigma_A}^2 &= \frac{1}{c}\left(\frac{1}{I-1}\sum_i n_i(Y_{i.} - Y_{..})^2 - \widehat{\sigma}^2\right) = \frac{1}{c}\left(SCM/(I-1) - \widehat{\sigma}^2\right)\end{aligned}$$

avec $c = \frac{n^2 - \sum n_i^2}{n(I-1)}$ ($c = J$ si $n_i = J$ pour tout i).
On peut se demander quel doit être le nombre I de niveaux du facteur aléatoire à tirer et le nombre de répétitions à effectuer pour avoir une précision suffisante de l'estimation des composantes de la variance. Comme dans le cas d'un effet fixe, la question peut être formulée en terme de puissance du test permettant de détecter un écart Δ entre deux traitements, ou sur la longueur de l'intervalle de confiance pour un ou plusieurs contrastes sur les traitements.
Plaçons nous dans le cas équilibré ($n_i = J$ pour tout i), on peut montrer que l'estimateur de σ_A^2 divisé par σ_A^2 suit approximativement une loi du χ^2 divisée par son nombre de degrés de liberté ν estimé par ($\widehat{\nu} = \frac{(SCM/(I-1) - SCR/(n-I))^2}{SCM^2/(I-1)^3 + SCR^2/(n-I)^3}$) (5.3.2, p. 187). Un intervalle de confiance de niveau $1-\alpha$ pour σ_A^2 est donc donné par

$$I_{\sigma_A^2,\alpha} = [\frac{\widehat{\nu}\widehat{\sigma_A}^2}{\chi^2_{\widehat{\nu},1-\alpha/2}}; \frac{\widehat{\nu}\widehat{\sigma_A}^2}{\chi^2_{\widehat{\nu},\alpha/2}}]$$

Par ailleurs la variance de l'estimateur de σ_A^2 est donnée par

$$\tau^2 = \frac{2n^2}{I^2}\left(\frac{(n\sigma_A^2/I + \sigma^2)^2}{I-1} + \frac{\sigma^4}{n-I}\right)$$

pour $n > I$.
Supposons que $n = JI$ soit fixé, déterminé par des contraintes budgétaires. Pour avoir un intervalle de confiance le plus petit possible il faut que la variance τ^2 soit la plus faible possible ce qui sera le cas si I est le plus grand possible. Dans le cas extrême où $n = I$ et $J = 1$ alors $\widehat{\sigma_A}^2 = \frac{SCM}{I-1}$ et $\tau^2 = 2\sigma_A^4/(I-1)$ est la plus petite possible.
Si maintenant on demande un intervalle de confiance le plus précis possible pour le rapport $\frac{\sigma_A^2}{\sigma^2}$, un estimateur non biaisé de ce rapport est donné par

$$U = \frac{1}{J}\left[\frac{(n-I-2)SCM/(I-1)}{(n-I)SCR/(n-I)} - 1\right]$$

On déduit de la définition de U et de l'expression de la variance d'une loi de Fisher de degrés de liberté $(I-1, n-I)$ que

$$Var(U) = \left(\frac{\sigma_A^2}{\sigma^2} + \frac{1}{J}\right)^2\left(\frac{2(n-3)}{(I-1)(n-I-4)}\right)$$

Si le nombre d'observations n est fixé et si on pense que $\sigma_A^2 \geq \sigma^2$ alors pour minimiser la variance de U il faut minimiser le deuxième terme du produit ce qui peut être réalisé en prenant $I = n/2$ et donc $J = 2$. Si par contre dans le cas inhabituel où l'on pense que σ_A^2 est beaucoup plus petit que σ^2 alors il faut prendre I le plus petit possible.
Cherchons maintenant à déterminer combien d'observations $n = JI$ sont nécessaires pour assurer une certaine puissance du test $H_0 : \frac{\sigma_A^2}{\sigma^2} \leq \gamma$ contre $H_1 : \frac{\sigma_A^2}{\sigma^2} > \gamma$. L'hypothèse H_0 est rejetée si $\frac{SCM/(I-1)}{SCR/(n-I)} > (c\gamma + 1)f_{I-1,n-I,1-\alpha} = k$. Lorsque la vraie valeur du rapport $\frac{\sigma_A^2}{\sigma^2}$ est égale Δ la statistique $F = \frac{SCM/(I-1)}{J\Delta SCR/(n-I)}$ suit une loi de Fisher de degrés de liberté $(I-1, n-I)$ et la probabilité de rejeter H_0 est la probabilité que F dépasse la quantité $\frac{J\gamma+1}{J\Delta+1}f_{I-1,n-I,1-\alpha}$. Si l'on souhaite avoir une puissance au moins égale à π pour la valeur Δ alors il faut que $\frac{J\gamma+1}{J\Delta+1}f_{I-1,n-I,1-\alpha} < f_{I-1,n-I,1-\pi}$. Il reste à déterminer J et I par essais-erreurs de façon à avoir

$$(f_{I-1,n-I,1-\alpha})(f_{n-I,I-1,1-\pi}) \leq \frac{J\Delta + 1}{J\gamma + 1}$$

Dans le cas où on a plusieurs facteurs aléatoires, il faut adapter cette formule en considérant les différents estimateurs des variances des effets aléatoires et les distributions de Fisher associées.

7.5.2 Plans d'expériences avec effets bloc aléatoires : split-plot

On a vu que l'idéal lors d'une expérience à deux ou plusieurs facteurs traitements, est d'orthogonaliser et randomiser les unités expérimentales pour obtenir un plan en blocs complets. Cela implique que deux unités expérimentales contigues (dans l'espace ou le temps par exemple) se voient attribuées une combinaison de traitements aléatoire et qu'on soit amené à changer les modalités des traitements à chaque changement d'unité expérimentale. Cela peut se révéler très contraignant dans certaines situations et on préfère souvent dans ces cas là répartir les différents traitements de manière séquentielle dans les blocs.

Cette stratégie de planification est appelée split-plots ou plans à parcelles partagéees, ou encore plans en unités subdivisées.

Supposons que lors d'une expérience impliquant deux traitements A à I modalités et B à J modalités, les modalités de A sont plus difficiles à changer que celles de B et que l'on dispose pour cette expérience de K blocs. On va commencer par répartir les modalités de A dans les blocs en grandes unités suivant un dispositif en bloc complet randomisé (7.3.1, p. 230). Les effets du traitement A vont donc être testés à l'aide de la résiduelle à (I-1)(K-1) degrés de liberté obtenue dans le modèle

$$Y_{ik} = \mu + a_i + B_k + F_{ik}$$

Il n'est en général pas utile d'estimer l'effet bloc B_k et on peut la plupart du temps le considérer comme aléatoire.

On divise ensuite chaque grande unité correspondant à une modalité de A en J petites unités, auxquelles on attribue aléatoirement une modalité du traitement B.

On a donc divisé chaque bloc en IJ petites unités et les effets du traitement B ainsi que ceux de l'interaction entre A et B peuvent être testés à l'aide de la résiduelle à

$I(J-1)(K-1)$ degrés de liberté obtenue dans le modèle

$$Y_{ik} = \mu + a_i + B_k + F_{ik} + b_j + c_{ij} + E_{ijk}$$

où maintenant F_{ik} est vu comme un effet aléatoire de l'interaction entre le facteur A et les blocs. Il est clair que les effets du traitement B et de l'interaction $A * B$ sont estimés avec une plus grande précision que ceux du traitement A, et outre les contraintes expérimentales il faudra en tenir compte au moment de choisir le traitement qui sera attribué aux grandes unités.

7.6 Plans pour les surfaces de réponse

7.6.1 Introduction

Ces plans sont souvent utilisés pour optimiser un processus industriel ou un nouveau produit. Les processus industriels ou de service ne sont pas complètement maîtrisés et prévisibles parce qu'on ne sait pas décrire, comprendre ni modéliser complètement leur fonctionnement. Cette incapacité peut être due à la complexité, à l'absence de connaissance de certains aspects particuliers ou des relations entre certaines parties du processus. De plus un certain nombre de phénomènes internes comme l'usure, les facteurs humains ou externes comme les conditions de fonctionnement, température, humidité ambiante, peuvent créer une variabilité qui fait que dans des conditions apparemment identiques, le processus ne produit pas exactement le même résultat.

Cette situation de connaissance imparfaite est très courante et acceptable dans beaucoup de situations sauf si elle cause des nuisances comme par exemple des produits de mauvaise qualité. On est alors placé devant un problème à résoudre sans disposer de tous les éléments de connaissance nécessaires.

On dispose alors de deux stratégies possibles :

— on établit un programme de recherche dont l'objectif est d'avoir une connaissance approfondie de chaque mécanisme du processus et des relations entre eux.

— On refuse de se lancer dans un tel programme qui risque d'être coûteux et dont le succès n'est pas certain. On cherche à étudier et prédire le comportement du processus en le faisant fonctionner dans certaines conditions et avec certaines valeurs des paramètres de commande et en observant les résultats. On accepte la situation de connaissance limitée dans laquelle on se trouve et on s'en accommode.

La seconde solution est assez difficile à accepter pour l'ingénieur ou le chercheur puisqu'il s'agit d'une sorte de renoncement à comprendre réellement. C'est pourquoi la méthode des plans d'expériences a longtemps été mal acceptée dans les écoles d'ingénieurs sauf dans le domaine biologique où la voie de la compréhension totale et complète était clairement fermée à cause de processus trop complexes.

Par contre elle plaît beaucoup au gestionnaire industriel ou de la recherche-développement car elle est économique, rapide et sûre.

Les plans d'expériences décrits dans cette section sont donc très utilisés dans la recherche-développement. L'objectif est de trouver une combinaison des variables de commande du processus $(X_1, ...X_p)$ qui donne une réponse Y optimale. Si X_j représente la part de l'ingrédient j dans un produit, on parle de mélange, traité en section 7.6.6, p. 257. Dans le cas général on parle de surface de réponse. Ces plans d'expériences sont

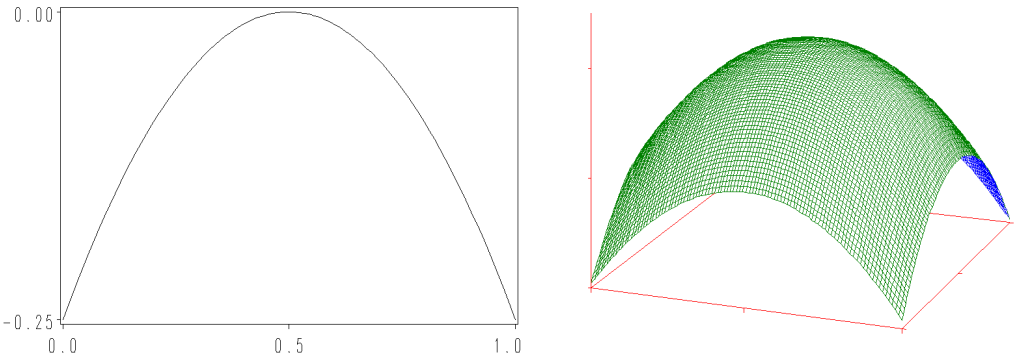


FIGURE 7.2 – Surfaces de réponse. Dans la figure de gauche $y = -(x - \frac{1}{2})^2$, dans la figure de droite $y = -(x_1 - \frac{1}{2})^2 - (x_2 - \frac{1}{2})^2$. Dans les 2 figures y est sur l'axe vertical. x est en abscisse dans la figure de gauche et z_1 et z_2 sont les axes horizontaux dans la figure de droite.

assez coûteux en nombre d'essais si p est grand. Ils ne sont utilisés en pratique que pour un petit nombre de facteurs (en général moins de 6). Souvent on procède en 2 étapes : la première étape consiste à sélectionner les facteurs influents avec un plan fractionnaire vu en section 7.4.3, p. 239. La deuxième étape consiste à analyser plus finement avec un modèle de surface de réponse les facteurs retenus dans l'étape précédente.

Quand les facteurs sont par nature des variables continues, il est logique d'en tenir compte dans le modèle qui devient un modèle de régression linéaire. Un modèle polynomial de degré 1 (purement linéaire) n'est en général pas suffisant. Par exemple la courbe de réponse de la qualité d'un gâteau en fonction de la durée de cuisson passe par un maximum : si la durée est trop courte il n'est pas cuit, si elle est trop longue il est brûlé. Il y a donc une durée optimale qui correspond à une qualité de cuisson maximale. La courbe la plus simple qui passe par un optimum est un polynôme de degré 2, $y = a + bx + cx^2$. La régression polynomiale, qui fait partie du modèle linéaire, permet de prendre en compte des polynômes des facteurs. Le modèle le plus utilisé est le polynôme de degré 2 à p variables : indexrégression polynomiale

$$Y = \theta_0 + \sum_{j=1}^{p} \theta_j X_j + \sum_{j,j'=1}^{p} \theta_{jj'} X_j X_{j'} + E \tag{7.5}$$

La Figure 7.2 donne un exemple de courbe de réponse parabolique et un exemple de surface de réponse avec un polynôme de degré 2 de 2 variables. Comme la surface de réponse n'est pas toujours facile à interpréter, on préfère souvent les lignes de niveaux données dans la Figure 7.3. Dans cette représentation, il est plus facile de visualiser où se trouve le point (x_1, x_2) pour lequel la fonction est optimale.

On peut augmenter le nombre de niveaux des facteurs : un plan factoriel ou fractionnaire avec 2 niveaux par facteur ne permet pas d'estimer les paramètres du modèle 7.5. Il faut un minimum de 3 niveaux, et il est préférable d'avoir 5 niveaux, ce qui permet de tester l'adéquation du modèle. La classe de plans d'expériences la plus utilisée pour les surfaces de réponse est la classe des plans composites.

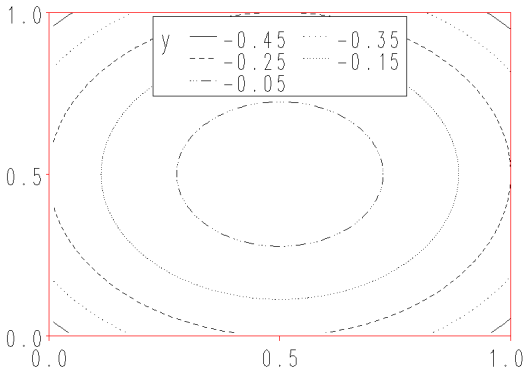


FIGURE 7.3 – Lignes de niveau de la surface de réponse $y = -(x_1 - \frac{1}{2})^2 - (x_2 - \frac{1}{2})^2$

7.6.2 Plan composite centré

Un plan composite centré est un plan comportant 3 types d'expériences :

un plan fractionnaire, avec $N_f = 2^{p-k}$ expériences. La valeur des variables X_j est égale à 1 ou -1. Ce plan doit être au moins de résolution 5 pour éviter de confondre les interactions d'ordre 2 entre elles.

N_0 **expériences au centre,** expériences au même point $x_1 = x_2 = ... = x_p = 0$.

$2p$ **expériences en "étoile",** expériences du type suivant : pour chaque j, on a 2 essais avec $X_j = \pm\alpha$ et $X'_j = 0$ pour tout $j' \neq j$.

Pour avoir un plan "isovariant par rotation", c'est à dire tel que l'erreur de prédiction faite par le modèle linéaire avec les paramètres estimés ne dépende que de la distance au centre et pas de la direction d'éloignement, on montre qu'il faut choisir $\alpha = N_f^{\frac{1}{4}}$. La Figure 7.5 montre la différence entre un plan invariant par rotation et un plan non invariant.

Essai	X_1	X_2
1	-1	-1
2	-1	1
3	1	-1
4	1	1
5	0	0
6	0	0
7	$-\alpha$	0
8	α	0
9	0	$-\alpha$
10	0	α

TABLE 7.7 – Plan central composite pour 2 facteurs

Par exemple avec $p = 2$ et $k = 0$, on obtient le plan de la Table 7.7 avec $N_0 = 2$ points au centre. Ce plan est représenté dans la Figure 7.4.

Le choix du nombre d'essais au centre se fait selon les critères suivants

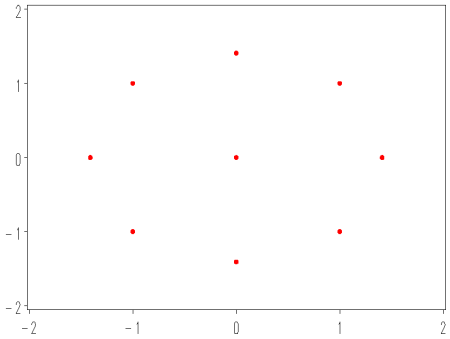


FIGURE 7.4 – Représentation graphique du plan composite de la Table 7.7. x_1 est en abscisse, x_2 en ordonnée.

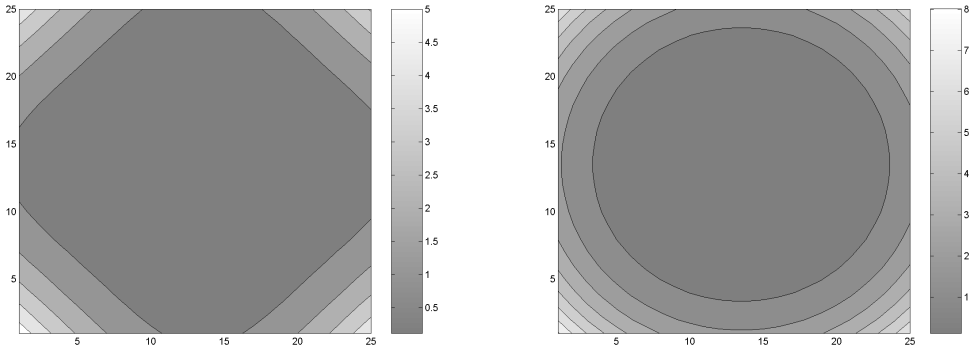


FIGURE 7.5 – Variance de la prédiction pour un plan central composite avec $\alpha = 2$ (figure de gauche) et pour le plan central composite de la Table 7.7 avec $\alpha = \sqrt{2}$ (figure de droite)

— il est utile d'avoir au moins 2 essais au même point pour estimer la variabilité expérimentale. La variabilité résiduelle du modèle (7.5) contient 2 composantes, la variabilité expérimentale et l'erreur de modèle due au fait que le vrai modèle n'est pas un polynôme de degré 2 mais une autre fonction. Comme on ne connait pas le vrai modèle, on ne peut pas calculer directement l'erreur de modèle. L'existence de 2 ou plus répétitions au même point permet d'estimer séparément σ^2 et donc aussi l'erreur de modèle, comme cela est précisé dans le paragraphe 7.6.3, p. 254.

— Si on calcule la corrélation entre les variables de régression de ce plan on voit que certaines corrélations (par exemple entre x_2^2 et x_1^2) sont non nulles. Les plans pour des surfaces de réponses sont donc non-orthogonaux. Par exemple le plan central composite avec $N_0 = 2$ donne la matrice de variance des estimateurs des paramètres de la Table 7.8. On peut cependant assurer une orthogonalité partielle en choisissant la valeur de N_0. Si on choisit $N_0 = 4 + 4\sqrt{N_f} - 2p$ on a orthogonalité entre toutes les colonnes de X sauf la colonne de 1 liée à l'ordonnée à l'origine. Par exemple le plan central composite avec $N_0 = 8$ donne la matrice de variance des estimateurs des paramètres de la Table 7.9. En pratique l'expression

	θ_0	θ_1	θ_2	θ_{12}	θ_{11}	θ_{22}
θ_0	0.5	0	0	0	-0.25	-0.25
θ_1	0	0.125	0	0	0	0
θ_2	0	0	0.125	0	0	0
θ_{12}	0	0	0	0.25	0	0
θ_{11}	-0.25	0	0	0	0.219	0.094
θ_{22}	-0.25	0	0	0	0.094	0.219

TABLE 7.8 – Matrice de variance des estimateurs, plan central composite avec $N_0 = 2$

	θ_0	θ_1	θ_2	θ_{12}	θ_{11}	θ_{22}
θ_0	0.125	0	0	0	-0.0625	-0.0625
θ_1	0	0.125	0	0	0	0
θ_2	0	0	0.125	0	0	0
θ_{12}	0	0	0	0.25	0	0
θ_{11}	-0.0625	0	0	0	0.125	0
θ_{22}	-0.0625	0	0	0	0	0.125

TABLE 7.9 – Matrice de variance des estimateurs, plan central composite partiellement orthogonal avec $N_0 = 8$

ci-dessus ne donne pas nécessairement un nombre entier, donc on choisit l'entier le plus proche, ce qui donne presque l'orthogonalité. Cependant cette contrainte supplémentaire augmente le nombre d'essais : pour $p = 2$, on passe de 10 essais à 16. Pour $p = 5$ le plan central composite invariant par rotation avec $k = 0$ demande 44 essais et le plan invariant par rotation et orthogonalisé demande 59 essais.

7.6.3 Validation du modèle

Il faut faire attention au fait que le modèle polynomial n'est qu'une approximation. Il faut distinguer le vrai modèle inconnu et le modèle linéaire qui n'est qu'une approximation.

Supposons que la vraie fonction qui exprime la liaison entre Y et $(X_1, ...X_p)$ soit $f(X_1, ...X_p)$. On a $Y = f(X_1, ...X_p) + F$, où $F \sim \mathcal{N}(0, \sigma^2)$. On peut écrire le modèle (7.5, p. 251) sous la forme suivante :

$$Y = \theta_0 + \sum_{j=1}^{p} \theta_j X_j + \sum_{jj'=1}^{p} \theta_{jj'} X_j X_{j'} + r(X_1, ...X_p) + F \tag{7.6}$$

où $r(X_1, ...X_p) = f(X_1, ...X_p) - (\theta_0 + \sum_{j=1}^{p} \theta_j X_j + \sum_{jj'=1}^{p} \theta_{jj'} X_j X_{j'})$ est l'écart entre la vraie fonction et le polynôme. La figure 7.6 illustre cette différence. L'erreur de prédiction du modèle (7.5) est

$$E = Y - (\theta_0 + \sum_{j=1}^{p} \theta_j X_j + \sum_{jj'=1}^{p} \theta_{jj'} X_j X_{j'}).$$

On voit que $E = F + r(X_1, ...X_p)$, c'est à dire que l'erreur de prédiction du modèle linéaire cumule 2 quantités différentes : l'erreur de modèle $R(X_1, ...X_p)$ et la variabilité

expérimentale F. Dans l'analyse des résultats statistiques il faudra donc séparer les 2 sources, ce qui impose des contraintes aux plans d'expériences proposés.

Certains plans ne le permettent pas et il est préférable d'utiliser un plan d'expériences qui permet de tester l'erreur de modèle.

Pour faire un test de validation ou d'adéquation au modèle on procède de la façon suivante. On calcule 3 sommes de carrés :

- la somme de carrés (au sens usuel) des erreurs du modèle linéaire (7.5, p. 251) $SC_1 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$, avec $n - \frac{(p+1)(p+2)}{2}$ degrés de liberté (il y a $\frac{(p+1)(p+2)}{2}$ paramètres dans le modèle (7.5)),
- la somme de carrés de variabilité expérimentale "pure", SC_0, calculée sur les répétitions du point central, avec $N_0 - 1$ degrés de liberté
- la somme de carrés des erreurs d'ajustement, $SC_2 = SC_1 - SC_0$, avec $n - \frac{(p+1)(p+2)}{2} - N_0 + 1$ degrés de liberté.

Sous l'hypothèse $H_0 = \{$Le modèle (7.5) est le vrai modèle$\}$, les 2 sommes de carrés SC_2 et SC_0 suivent des lois du χ^2 indépendantes et donc on a

$$\frac{\frac{SC_2}{n-\frac{(p+1)(p+2)}{2}-N_0+1}}{\frac{SC_0}{N_0-1}} \sim \mathcal{F}(n - \frac{(p+1)(p+2)}{2} - N_0 + 1, N_0 - 1), \tag{7.7}$$

ce qui permet de tester H_0. Si on a d'autres points du plan d'expériences avec 2 répétitions ou plus, ce test s'étend en cumulant les sommes de carrés de type SC_0. Pour avoir une puissance raisonnable il faut au moins 5 degrés de liberté pour la somme des carrés du dénominateur.

7.6.4 Solution optimale

Quand on a estimé les paramètres du modèle (7.5) et validé le modèle, le choix des valeurs des facteurs qui optimisent la réponse est immédiat : si $p = 2$ on peut utiliser la Figure 7.3 pour trouver graphiquement les valeurs de x_1 et x_2 optimales. Dans le cas général où $p \geq 1$ on obtient analytiquement les valeurs de $(x_1...x_p)$ en annulant les dérivées partielles de (7.5) par rapport à chaque variable et en résolvant le système linéaire à p équations et p inconnues. Attention, le point obtenu n'est pas forcément un maximum ou un minimum. Ce peut être un "point selle" qui est un maximum dans une direction et un minimum dans une autre. Pour vérifier ce point on calcule la matrice des dérivées secondes. Si elle est définie positive, c'est un minimum, si elle est définie négative, c'est un maximum, sinon c'est un "point selle". Les logiciels calculent la solution qui annule les dérivées et donnent les valeurs propres de la matrice des dérivées secondes. Cette partie de l'analyse est parfois nommée "Analyse canonique" dans certains logiciels.

Il arrive que le choix des valeurs des facteurs $(x_1, ...x_p)$ soit éloigné de la solution optimale. Dans ce cas la réponse est linéaire et non quadratique. Cela se manifeste par le fait que si on teste la nullité des coefficients $\theta_{jj'}$, on accepte l'hypothèse nulle. Si on cherche à optimiser la réponse il faut alors faire de nouvelles expériences dans une zone plus favorable. La direction dans laquelle il faut prendre les nouvelles valeurs des facteurs est indiquée par "le chemin de plus grande pente" donné par les coefficients $\widehat{\theta}_j$: on se déplace dans la direction du vecteur $(\widehat{\theta}_1, ...\widehat{\theta}_p)'$.

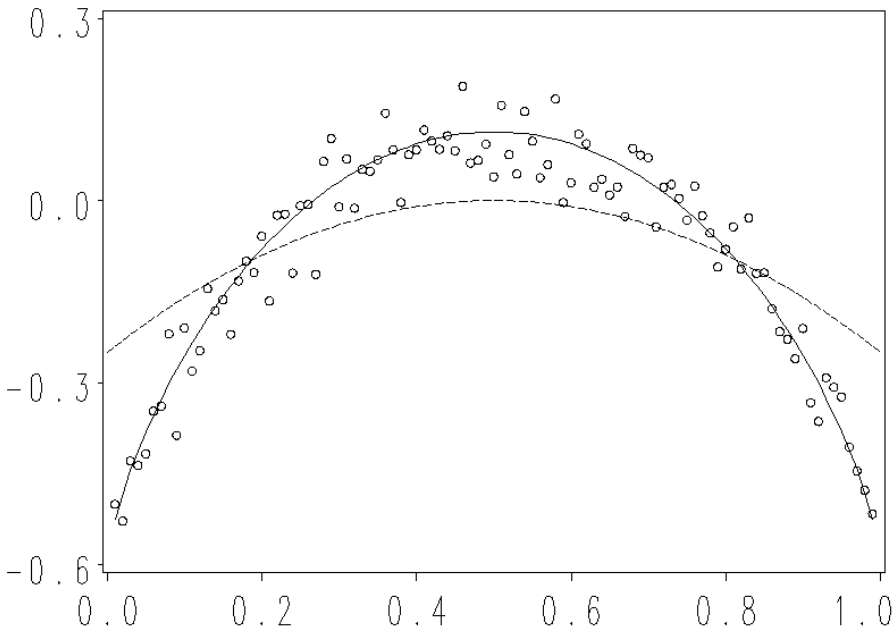


FIGURE 7.6 – Les 2 sources d'erreur : variabilité expérimentale autour du vrai modèle et erreur de modèle. Le modèle vrai est $f(x) = -(x \log x + (1-x)\log(1-x)) - 0.58$ (ligne continue) et le modèle ajusté est un polynôme de degré 2 (ligne pointillée). L'écart entre les 2 courbes est artificiellement exagéré pour mieux visualiser la différence. Les cercles sont les données observées.

7.6.5 Plans Box-Behnken

Le plan central composite comporte 5 niveaux pour chaque facteur. On peut ajuster un polynôme de degré 2 avec seulement 3 niveaux par facteur, ce qui est plus économique en nombre d'essais que si on choisit de prendre 5 niveaux par facteur comme dans le plan composite. On pourrait utiliser un plan fractionnaire 3^{p-l} de résolution V pour pouvoir estimer séparément les interactions d'ordre 2, mais cette solution est trop coûteuse en nombre d'essais. Les plans de type "Box-Behnken" sont fabriqués de la façon suivante :

- on construit un plan en b blocs fictifs incomplets équilibrés de taille k, chacun ayant $p' < p$ facteurs.
- pour chaque bloc on construit un plan factoriel $2^{p'}$, on a donc $bk = b2^{p'}$ essais pour les b blocs,
- on ajoute des points au centre.

Par exemple avec $p = 4$, $p' = 2$, $k = 4$, on construit un plan de 6 blocs incomplets équilibrés contenant chacun 2 facteurs. Chaque bloc est constitué d'un plan factoriel 2^2. Avec 3 points au centre cela donne 6*4+3=27 essais.

On peut dans certains cas obtenir un plan invariant par rotation et un plan partiellement orthogonal.

7.6.6 Pour aller plus loin

Plans de Doehlert

Il existe d'autres plans d'expériences pour les surfaces de réponse. Les plans de Doehlert sont constitués de N_0 essais au centre et d'essais répartis uniformément sur la sphère de rayon 1 dans l'espace $\mathbb{R}^p$ des facteurs. Ces plans sont moins coûteux que le plan central composite mais ont une moins bonne précision. Par contre ils s'adaptent très bien à une construction séquentielle. On fait des essais dans une zone donnée de $\mathbb{R}^p$, puis on étend le plan dans une direction intéressante avec peu de points supplémentaires tout en conservant la structure du plan d'expériencess. Voir [50], chapitre 4 et [42].

Plusieurs variables réponse

Quand il y a plusieurs variables réponses à optimiser on se trouve devant un sérieux problème, car le meilleur plan d'expériences pour une variable peut être médiocre pour une autre. Dans ce cas, on cherche à trouver un compromis acceptable pour tous les critères importants, voir l'étude de cas 8.10, p. 294.

Plans pour les mélanges

Il existe aussi des plans d'expériences pour les mélanges ("mixture design"). Beaucoup de produits de l'industrie chimique, pétrolière, textile ou agro-alimentaire sont un mélange de plusieurs constituants : l'essence, les crèmes solaires, la mayonnaise... On cherche à optimiser la composition du mélange de sorte que le produit possède certaines propriétés. On procède alors à une expérimentation avec différentes compositions. Soit $(x_1, ...x_p)$ avec $\sum_{j=1}^{p} x_j = 1$ et $x_j \geq 0, j = 1, p$ les proportions des composants 1 à p, et Y la propriété du produit (résistance d'un tissu, qualité gustative d'un aliment..) qui est la variable réponse. On approche la fonction $y = f(x_1, ...x_p)$ par un polynome de degré d. Il existe des plans d'expériences privilégiés pour cette situation. Ces plans dépendent du degré d, du nombre de constituants p, mais aussi des contraintes sur les proportions. Une particularité de ce type de plans est qu'il y a une relation linéaire entre les facteurs $(\sum_{j=1}^{p} x_j = 1)$. Voir [50] Chapitre 5 ou [15].

Plans pour les expériences numériques

Un nombre de plus en plus important de disciplines font appel à des modèles complexes qui sont construits pour représenter le plus fidèlement possible la réalité. Par exemple des modèles météorologiques permettent de prédire l'évolution des masses d'air autour du globe terrestre, ce qui permet d'obtenir des prévisions météorologiques. Des modèles de croissance des plantes prennent en compte le sol, les racines, les tiges et les feuilles, l'air, l'ensoleillement etc, et permettent de prédire la croissance des plantes et les récoltes des prochains mois. Des modèles de diffusion des radionucléides suite à un événement majeur sur une centrale nucléaire prennent en compte les transferts dans le sol, l'air, les eaux, les plantes, les animaux et l'être humain. On peut ainsi prévoir l'impact sur la santé humaine de ce type d'incident. Ces modèles sont simulés avec des codes de calcul. Cependant ces codes sont si complexes qu'ils peuvent mettre plusieurs jours pour donner un résultat. Par ailleurs leur complexité est telle qu'on a parfois besoin de travailler avec une version simplifiée pour obtenir une prédiction rapide de nouvelles

réponses, une exploration du modèle pour une meilleure compréhension de son comportement, des effets de ses variables d'entrée et de leurs interactions, ou l'estimation de la probabilité d'occurrence d'événements rares en sortie du modèle numérique. On veut donc obtenir un modèle mathématique représentatif du code étudié en termes de qualité d'approximation, ayant de bonnes capacités de prédiction, et dont le temps de calcul pour évaluer une réponse soit négligeable. Ce modèle simplifié est construit et ajusté à partir de quelques simulations du code (correspondant à différents jeux de valeurs des paramètres). Le nombre de simulations nécessaire dépend de la complexité du code et du scénario qu'il modélise, du nombre de facteurs et de la qualité d'approximation souhaitée. On appelle expériences numériques ce type d'expérimentation. On utilise parfois pour ce faire des plans d'expériences pour les surfaces de réponses. Pour en savoir plus sur ce sujet voir [35].

Chapitre 8

Exemples de plans d'expériences

8.1 Blocs complets : 2 variétés de mais dans 8 lieux

8.1.1 Plans d'expériences et données

Pour déterminer l'influence de la variété sur le rendement de 2 variétés de maïs, le plan d'expériences suivant a été utilisé avec 8 lieux. Dans chaque lieu on a cultivé les 2 variétés de maïs sur 2 parcelles contigües.

Lieu	1	2	3	4	5	6	7	8
Variété 1	110	90	92	75	100	89	72	90
Variété 2	109	92	90	70	92	82	67	81

On observe 2 phénomènes sur la figure 8.1 créée par le programme SAS 7.1.a : les rendements sont très différents d'un lieu à un autre et la variété 1 a un meilleur rendement que la variété 2 pour 7 lieux sur 8.

```
 PG-SAS 1.a
data mais; input lieu variete rendement @@;
lines;
1 1 110 1 2 109 2 1 90 2 2 92 3 1 92 3 2 90 4 1 75 4 2 70 5 1 100
5 2 92 6 1 89 6 2 82 7 1 72 7 2 67 8 1 90 8 2 81 ;
symbol1 h=2 v='circle' color='red'; symbol2 h=2 v='dot'color='blue';
         axis1 minor=none;
proc gplot data=mais;plot rendement*lieu=variete /haxis = axis1;
title 'rendement de 2 varietes de maïs dans 8 lieux';
run;
```

Il y a un facteur d'intérêt (la variété) et un facteur de contrôle d'hétérogénéité (le lieu). Il y a 2 unités de traitement, 16 cellules et 16 unités expérimentales. A chaque cellule est associée une unité expérimentale. Ce dispositif est un plan en blocs complets randomisé : complet parce qu'il y a 2 parcelles dans chaque bloc, et les 2 variétés sont présentes dans chaque bloc. Randomisé parce que l'attribution de la parcelle semée avec la variété 1 a été tirée au sort dans chaque bloc.

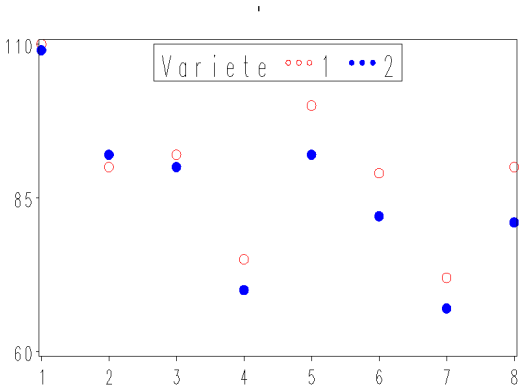


FIGURE 8.1 – Rendement de 2 variétés de maïs dans 8 lieux. En abscisse le lieu, en ordonnée le rendement.

8.1.2 Analyse de l'expérience

Le modèle de l'expérience (correspondant au programme SAS PG-SAS 1.b) est un modèle d'analyse de variance à 2 facteurs sans interaction (avec les notations usuelles du chapitre 1, i l'indice du traitement et j celui du lieu) :

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

La première analyse de variance (avec effet lieu, programme PG-SAS 1.b) permet de conclure qu' il y a un effet variété significatif et que le lieu est un facteur de contrôle d'hétérogénéité important. La différence de rendement entre les 2 variétés est estimée à 4.375 avec un intervalle de confiance 95% égal à [1.22 7.53].

La prise en compte du facteur de contrôle d'hétérogénéité du lieu est essentielle pour avoir une variance résiduelle réduite et une forte puissance. Si on ne l'avait pas pris en compte on aurait eu le résultat de la deuxième analyse de variance (modèle $Y_{ij} = \mu + \alpha_i + E_{ij}$, programme PG-SAS 1.c). L'écart-type résiduel qui mesure la variabilité naturelle du matériel expérimental est de 12.9 dans la deuxième analyse, contre 2.7 dans la première analyse ce qui augmente le "bruit de fond" de l'expérience. De ce fait la différence de rendement entre les 2 variétés n'est plus significative. La différence moyenne entre les 2 variétés est inchangée, mais l'intervalle de confiance de cette différence est trop grand pour conclure quoique ce soit. Bien sûr si on a construit un dispositif en blocs il serait stupide de ne pas prendre en compte l'effet du bloc dans le modèle quand ce dernier est important comme c'est le cas ici. Cependant on aurait obtenu un résultat semblable à celui de la deuxième analyse de variance si on avait tiré dans l'ensemble des 16 unités expérimentales les 8 unités à qui serait attribuée la variété 1, sans tenir compte du lieu. Dans ce cas la même variété aurait pu être cultivée 2 fois dans certains lieux. Cet exemple illustre parfaitement l'intérêt de contrôler l'hétérogénéité des parcelles expérimentales pour améliorer la puissance des tests, comme cela a été exposé en section 7.3, p. 229.

```
PG-SAS1.b /* Analyse de variance avec le facteur bloc */
proc glm data=mais; class lieu variete; model rendement=lieu variete;
lsmeans  variete/stderr cl pdiff;quit;run;
```

```
PG-SAS 1.c  /* Tableau de l'analyse de variance sans le facteur bloc */
proc glm data=mais; class lieu variete; model rendement= variete;
lsmeans  variete/stderr cl pdiff;quit;run;
```

Analyse de variance avec effet bloc (PG-SAS 1.b)

Informations sur le niveau de classe

Classe	Niveaux	Valeurs
lieu	8	1 2 3 4 5 6 7 8
variete	2	1 2

Number of Observations Read	16
Number of Observations Used	16

Dependent Variable: rendement

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	8	2352.000000	294.000000	41.21	<.0001
Error	7	49.937500	7.133929		
Corrected Total	15	2401.937500			

R-carré	Coeff Var	Racine MSE	rendement Moyen
0.979209	3.050326	2.670942	87.56250

Source	DF	Type I SS	Carré moyen	F	Pr > F
lieu	7	2275.437500	325.062500	45.57	<.0001
variete	1	76.562500	76.562500	10.73	0.0136

Moyennes des moindres carrés pour l'effet variete

i	j	Différence entre les moyennes	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	4.375000	1.217113	7.532887

Analyse de variance sans effet bloc (PG-SAS 1.c)

Source	DF	Somme des carrés	Carré moyen	F	Pr > F
Model	1	76.562500	76.562500	0.46	0.5083
Error	14	2325.375000	166.098214		
Corrected Tota	15	2401.937500			

R-carré	Coeff Var	Racine MSE	rendement Moyenn
0.031875	14.71853	12.88791	87.56250

Source	DF	Type I SS	Carré moyen	F	Pr > F
variete	1	76.56250000	76.56250000	0.46	0.5083

Moyennes des moindres carrés pour l'effet variete

```
                 Différence
                  entre les    95% Confidence Limits for
i    j             moyennes       LSMean(i)-LSMean(j)
1    2             4.37            -9.44      18.19
```

8.2 Bloc incomplet équilibré, Champagnes

8.2.1 Présentation

Objectif

On veut comparer la qualité globale de $I = 7$ champagnes (A, B, C, D, E, F, G). La qualité est mesurée par une note comprise entre 1 et 50 attribuée par un juge.

Il s'agit d'un problème d'*analyse sensorielle* : la notation fait appel à une évaluation subjective fournie par un juge entraîné.

La gamme assez large de notes (de 1 à 50) rend raisonnable le recours au modèle linéaire qui suppose la mesure continue. Une notation sur une gamme plus étroite (de 1 à 5, par exemple) rendrait cette hypothèse plus discutable.

Dispositif

$J = 14$ juges (numérotés de 1 à 14) interviennent dans l'évaluation des champagnes. Comme souvent dans ce type d'analyse, on ne demande pas à tous les juges d'évaluer tous les champagnes pour éviter que leurs facultés sensorielles ne s'émoussent.

Soit n_{ij} le nombre de fois (0 ou 1) où le juge j a noté le champagne i, $n_{i+} = \sum_j n_{ij}$ et $n_{+j} = \sum_i n_{ij}$. Ici chaque juge note $n_{+j} = 3$ champagnes parmi les 7 et chaque champagne est noté par $n_{i+} = 6$ juges. On a donc au total

$$n = I \times n_{i+} = J \times n_{+j} = 42$$

observations. Le tableau 8.1, p. 263 donne les valeurs de tous ces effectifs.

Ce dispositif est un dispositif en blocs. On peut en effet considérer chaque juge comme un appareil de mesure ; toutes les mesures ne sont pas faites avec le même appareil et il faut tenir compte de l'hétérogénéité de ce matériel expérimental (particulièrement forte dans les problèmes d'analyse sensorielle). Chaque juge constitue donc un bloc. Chaque juge ne notant qu'une partie des champagnes mais tous les juges en notant le même nombre, c'est un plan en *Blocs Incomplets Équilibrés* (BIE), défini en 7.4.3, p. 236.

On peut remarquer que, pour que le plan puisse être équilibré, il faut pouvoir trouver deux entiers n_{i+} et n_{+j} tels que $I \times n_{i+} = J \times n_{+j}$. Le nombre $J = 14$ de juges n'a donc pas été choisi au hasard.

Une caractéristique principale de ce dispositif est qu'il n'est pas orthogonal . On rappelle (cf 1.5, p. 37) qu'un dispositif à deux facteurs est orthogonal si et seulement si on a

$$\forall i, j : n_{ij} = \frac{n_{i+} n_{+j}}{n}.$$

Cette condition n'est pas vérifiée ici puisque $n_{i+} = 6$ pour tous les champagnes, $n_{+j} = 3$ pour tous les juges et donc $n_{i+}n_{+j}/n = 18/42 = 3/7$. Or l'effectif n_{ij} ne peut prendre

Table of Juge by Champagne

Juge	Champagne							
Frequency	A	B	C	D	E	F	G	Total
1	0	0	1	1	0	0	1	3
2	1	0	0	1	1	0	0	3
3	0	1	0	1	0	1	0	3
4	1	0	0	1	1	0	0	3
5	0	1	0	0	1	0	1	3
6	1	1	1	0	0	0	0	3
7	1	0	0	0	0	1	1	3
8	0	1	0	1	0	1	0	3
9	0	0	1	1	0	0	1	3
10	1	0	0	0	0	1	1	3
11	0	1	0	0	1	0	1	3
12	0	0	1	0	1	1	0	3
13	1	1	1	0	0	0	0	3
14	0	0	1	0	1	1	0	3
Total	6	6	6	6	6	6	6	42

TABLE 8.1 – Dispositif en BIE : répartition des 7 champagnes entre les 14 juges.

que deux valeurs :

$$n_{ij} = 1 \quad \text{si le champagne } i \text{ est noté par le juge } j,$$
$$n_{ij} = 0 \quad \text{si le champagne } i \text{ n'est pas noté par le juge } j.$$

Le dispositif n'est donc pas orthogonal, ce qui signifie qu'il n'existe pas de décomposition unique des sommes de carrés associées à chaque effet et qu'*on ne pourra donc jamais complètement séparer les effets des deux facteurs.* Du fait du dispositif, les deux effets sont en partie confondus.

Il existe un grand nombre de répartitions possibles des 7 champagnes entre les 14 juges. Le tableau 8.1, p. 263 présente celle qui a été utilisée dans cette expérience. Dans les plans BIE, on veille à ce que chaque couple de champagne soit noté le même nombre de fois par le même juge.

On observe ici que chaque couple est noté 2 fois par le même juge : le champagne A est noté en même temps que le champagne B par les juges 6 et 13, le champagne C est noté en même temps que le G par les juges 1 et 9, *etc.*

Cet équilibre dans la répartition des couples fait que toutes les comparaisons des champagnes 2 à 2 (contrastes) auront la même variance et donc que les tests les concernant auront la même puissance (cf. 7.4.2, p. 235).

Bilan sur le plan d'expériences Il y a 7 champagnes à comparer, 14 juges, et chaque juge analyse 3 champagnes. Il y a un facteur d'intérêt (champagne), un facteur de contrôle d'hétérogénéité (juge), 7 unités de traitement, 42 unités expérimentales et 98 cellules qui ne sont donc pas toutes remplies. En effet, le juge est un facteur bloc, mais il y a plus de traitements (7) que de places dans le bloc (3). On a donc un bloc incomplet. Comme chaque couple de champagnes est analysé par 2 juges, c'est un bloc incomplet équilibré. Le dispositif est donc un BIE avec $t = 7$, $r = 6$, $b = 14$, $l = 3$ et $\lambda = 2$.

Données

Le tableau 8.2, p. 264 présente les résultats de l'expérience. On observe qu'il n'y a aucune données manquante, ce qui est une condition *sine qua non* pour conserver l'équilibre du BIE.

Obs	Juge	Champagne	Note	Obs	Juge	Champagne	Note
1	1	C	36	22	8	B	28
2	1	D	16	23	8	D	21
3	1	G	19	24	8	F	33
4	2	A	22	25	9	C	26
5	2	D	25	26	9	D	16
6	2	E	29	27	9	G	33
7	3	B	24	28	10	A	30
8	3	D	19	29	10	F	25
9	3	F	26	30	10	G	20
10	4	A	27	31	11	B	32
11	4	D	22	32	11	E	37
12	4	E	19	33	11	G	27
13	5	B	29	34	12	C	36
14	5	E	33	35	12	E	19
15	5	G	26	36	12	F	38
16	6	A	29	37	13	A	33
17	6	B	34	38	13	B	25
18	6	C	35	39	13	C	35
19	7	A	37	40	14	C	29
20	7	F	32	41	14	E	20
21	7	G	28	42	14	F	31

TABLE 8.2 – Notes de 7 champagnes mesurés par 14 juges.

8.2.2 Analyse des résultats

Modèle

Le modèle permettant d'analyser les sources de variabilité de la note Y doit prendre en compte l'effet champagne et l'effet juge, même si la comparaison des juges entre eux ne nous intéresse pas directement. L'effet juge doit être pris en compte pour analyser

précisément l'effet champagne. Nous devons donc considérer un modèle d'analyse de la variance à deux facteurs :

$$\begin{aligned} A &: \text{facteur Champagne à } I = 7 \text{ niveaux,} \\ B &: \text{facteur Juge à } J = 12 \text{ niveaux.} \end{aligned}$$

Interaction Champagne*Juge. Il n'y a *a priori* aucune raison de penser que le champagne et le juge n'interagissent pas sur la note. On devrait donc considérer le modèle d'analyse de la variance à deux facteurs avec interaction :

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

en notant α_i l'effet champagne, β_j l'effet juge et γ_{ij} l'interaction.

On voit cependant que, dans ce modèle, le terme d'interaction γ_{ij} est confondu avec le terme résiduel E_{ij} puisque ces deux termes portent exactement les mêmes indices. Ceci vient de l'absence de répétition : aucun champagne n'est noté plusieurs fois par le même juge.

Cela ne signifie pas que cette interaction n'existe pas, cela veut seulement dire que nous n'avons pas assez de données pour l'estimer. Cette interaction n'apparaîtra donc pas dans le modèle d'analyse et sera incluse dans la résiduelle. Si cette interaction existe effectivement, les tests des effets des facteurs seront moins puissants. On va donc utiliser le modèle

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}. \tag{8.1}$$

On doit spécifier des contraintes sur les paramètres pour que le modèle soit identifiable, par exemple $\alpha_I = \beta_J = 0$.

8.2.3 Résultats de l'analyse

Analyse de la variance, Tests des effets Champagne et Juge

Le tableau 8.3 présente la table d'analyse de la variance du modèle (8.1). On y lit que la somme des carrés dus au modèle vaut 1014.6 et représente $R^2 = 64\%$ de la somme des carrés totaux $SCT = 1585.6$. Même si les effets juges et champagne prennent en compte 64% de la variabilité des notes, il existe donc d'autres sources de variabilité : l'interaction, la variabilité intra-juge et celle entre 2 verres du même champagne.

Analyse des résidus. Comme toujours dans le modèle linéaire, l'analyse des résidus est nécessaire pour s'assurer que les hypothèses portant sur l'indépendance, la normalité et l'homoscédasticité des résidus sont vérifiées. Le graphique croisant les résidus et les valeurs prédites (cf. figure 8.2) ne montre pas de structure particulière, notamment concernant les variances des différents champagnes.

On obtient une première décomposition des sommes de carrés en introduisant les facteurs *dans un ordre donné* dans le modèle. On obtient ainsi les sommes de carrés de type I, présentées dans le tableau 8.4.

Les résultats obtenus avec les sommes de carrés de type I dépendent de l'ordre dans lequel les facteurs sont introduits. Cette caractéristique n'est pas toujours souhaitable car il n'existe pas toujours un ordre naturel entre les facteurs.
Ici, on peut cependant considérer qu'il est préférable d'introduire l'effet juge (effet bloc)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	1014.595238	53.399749	2.06	0.0530
Error	22	571.047619	25.956710		
Corrected Total	41	1585.642857			

R-Square	Coeff Var	Root MSE	Note Mean
0.639864	18.43070	5.094773	27.64286

TABLE 8.3 – Table d'analyse de la variance des notes des champagnes.

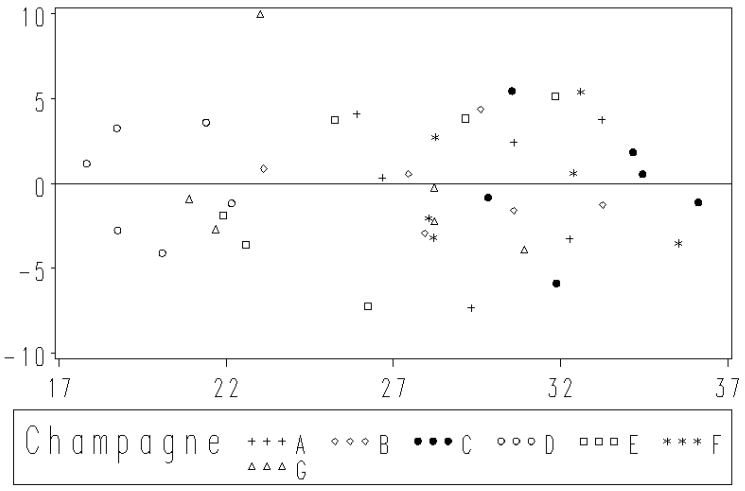


FIGURE 8.2 – Graphe des résidus du modèle (8.1). $\widehat{Y}$ en abscisse, $\widehat{E}$ en ordonnée.

en premier afin de prendre en compte, avant tout, l'hétérogénéité du matériel expérimental. Le test de l'effet champagne (effet d'intérêt) se fait ainsi après correction de l'effet bloc.

Les tests de type I et de type II (Tableau 8.5) montrent que l'effet champagne est statistiquement significatif au niveau 5% et l'effet des juges ne l'est pas.

Les résultats en types I et II montrent l'absence d'effet juge. On serait donc tenté finalement de retenir le modèle dans lequel l'effet champagne apparaît seul. Ce modèle présente l'avantage d'accorder plus de degrés de liberté aux résidus et offre donc des tests plus puissants.

La pratique consiste pourtant, dans un dispositif comme celui-ci, à maintenir l'effet bloc (juge) dans le modèle pour s'assurer que les notes des champagnes sont bien comparables. Cette pratique est commandée par la prudence : la puissance du test de l'effet juge n'est pas parfaite et un écart opposant 1 seul juge aux 13 autres pourrait ne pas être détecté par le test de Fisher. Les champagnes notés par ce juge se trouveraient systématiquement avantagés (ou pénalisés) si on effectuait les comparaisons des moyennes

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Juge	13	522.3095238	40.1776557	1.55	0.1772
Champagne	6	492.2857143	82.0476190	3.16	0.0217

TABLE 8.4 – Sommes de carrés de type I et tests des effets des facteurs.

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Juge	13	354.4523810	27.2655678	1.05	0.4440
Champagne	6	492.2857143	82.0476190	3.16	0.0217

TABLE 8.5 – Sommes de carrés de type II et tests des effets des facteurs.

classiques.

Les tests portant sur l'effet juge permettent malgré tout de montrer que le matériel expérimental (*i.e.* l'ensemble des juges) est globalement homogène, ce qui constitue un gage de qualité de l'expérience.

Comparaisons des champagnes

Moyennes classiques Une première comparaison des champagnes peut se fonder sur les moyennes empiriques

$$\hat{\mu}_{i\bullet} = Y_{i\bullet}$$

données dans le tableau 8.6. Le champagne obtenant la meilleure note moyenne est le champagne C.

Level of Champagne	N	Notes Mean	Notes Std Dev
A	6	29.6666667	5.12510163
B	6	28.6666667	3.88158043
C	6	32.8333333	4.26223728
D	6	19.8333333	3.54494946
E	6	26.1666667	7.90990940
F	6	30.8333333	4.79235502
G	6	25.5000000	5.24404424

TABLE 8.6 – Moyennes classiques des notes par champagne.

Ces moyennes ne sont pas comparables à cause de l'hétérogénéité potentielle des juges. On doit leur préférer les moyennes ajustées pour l'effet juge.

Moyennes ajustées Les moyennes ajustées permettent de rendre comparables les moyennes par juges. Le principe consiste à estimer la note moyenne qu'aurait obtenu un champagne donné *s'il avait été noté par tous les juges* : on parle donc de moyenne ajustée sur l'effet juge (ou corrigée de l'effet juge).

L'espérance de la note obtenue par le champagne i avec le juge j est :

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j.$$

Il est important de noter que cette espérance est définie dans le cadre du modèle *même si le champagne i n'a pas été noté par le juge j*. La moyenne ajustée est définie par

$$\tilde{\mu}_{i\bullet} = \frac{1}{J}\sum_j \mathbb{E}(Y_{ij}) = \mu + \alpha_i + \frac{1}{J}\sum_j \beta_j \tag{8.2}$$

et estimée par

$$\hat{\tilde{\mu}}_{i\bullet} = \hat{\mu} + \hat{\alpha}_i + \frac{1}{J}\sum_j \hat{\beta}_j.$$

Les valeurs de ces moyennes ajustées estimées sont données dans le tableau 8.7. On remarque que l'ordre des champagnes est le même qu'on considère les moyennes empiriques ou les moyennes ajustées :

$$\mathrm{D} < \mathrm{G} < \mathrm{E} < \mathrm{B} < \mathrm{A} < \mathrm{F} < \mathrm{C}.$$

Ceci est cohérent avec l'absence d'effet juge démontré à la subsection 8.2.3. En présence d'un effet juge fort, on pourrait tout à fait observer des inversions de classement.
Les écarts entre les champagnes sont cependant modifiés : l'écart de note entre les champagnes E et B diminue de 2.50 à 1.43 quand on passe des moyennes empiriques aux ajustées, alors que l'écart entre A et B augmente de 1.00 à 2.64. Ces changements modifient évidemment la conclusion des tests de comparaison des champagnes entre eux.

	Note	LSMEAN
Champagne	LSMEAN	Number
A	29.5714286	1
B	26.9285714	2
C	33.4285714	3
D	21.6428571	4
E	25.5000000	5
F	31.8571429	6
G	24.5714286	7

TABLE 8.7 – Moyennes des notes par champagne ajustée sur l'effet juge.

Les moyennes ajustées définies en (8.2) sont des combinaisons linéaires (estimables et invariantes) des paramètres μ, α_i et β_j. On peut donc calculer leur variances et covariances et ainsi effectuer des tests d'hypothèses de la forme

$$H_0 = \{\tilde{\mu}_{i\bullet} = \tilde{\mu}_{j\bullet}\} \qquad \text{contre} \qquad H_1 = \{\tilde{\mu}_{i\bullet} \neq \tilde{\mu}_{j\bullet}\}.$$

Le tableau 8.8 donne les valeurs des statistiques de test et les probabilités critiques associées pour toutes les comparaisons. On observe des différences de notes significatives (au niveau $\alpha = 5\%$) entre les couples de champagnes (A, D), (B, C), (C, D), (C, E), (C, G), (D, F) et (F, G).

Correction pour les tests multiples. Le niveau $\alpha = 5\%$ choisi pour chaque test ne prend pas en compte l'effet des tests multiples déjà vu en section 2.4.3, p. 76. On effectue ici $7 \times 6/2 = 21$ comparaisons. La correction de Bonferroni amène donc à tester

i/j	1	2	3	4	5	6	7
1		0.792387	-1.15646	2.377161	1.220704	-0.68531	1.49911
		0.4366	0.2599	0.0266	0.2351	0.5003	0.1481
2	-0.79239		-1.94884	1.584774	0.428317	-1.47769	0.706723
	0.4366		0.0642	0.1273	0.6726	0.1537	0.4872
3	1.156456	1.948843		3.533617	2.377161	0.471149	2.655567
	0.2599	0.0642		0.0019	0.0266	0.6422	0.0144
4	-2.37716	-1.58477	-3.53362		-1.15646	-3.06247	-0.87805
	0.0266	0.1273	0.0019		0.2599	0.0057	0.3894
5	-1.2207	-0.42832	-2.37716	1.156456		-1.90601	0.278406
	0.2351	0.6726	0.0266	0.2599		0.0698	0.7833
6	0.685308	1.477694	-0.47115	3.062468	1.906012		2.184418
	0.5003	0.1537	0.6422	0.0057	0.0698		0.0399
7	-1.49911	-0.70672	-2.65557	0.87805	-0.27841	-2.18442	
	0.1481	0.4872	0.0144	0.3894	0.7833	0.0399	

TABLE 8.8 – Comparaison des moyennes ajustées des 7 champagnes. Pour chaque comparaison : valeur supérieure = statistique de test, valeur inférieure = probabilité critique.

chaque comparaison au niveau $\alpha = 5\%/21 = 0.24\%$.
A ce niveau, seul les champagnes C (le meilleur) et D (le plus mauvais) sont significativement différents. La méthode de Bonferroni est une méthode conservative, c'est-à-dire que les tests qui en résultent sont peu puissants.

8.2.4 Programme SAS

Données. Le tableau 8.2, p. 264 est produit par les instructions suivantes.

```
data CHAMPAGNE;
     infile 'champagne/Champagne.don' firstobs=2;
     input Juge Champagne$ Note;
proc Print data=CHAMPAGNE;
run;
```

Dispositif en Blocs Incomplets Equilibres (BIE). Le tableau 8.1, p. 263 est produit par les instructions suivantes.

```
proc Freq data=CHAMPAGNE;
     tables Juge * Champagne / nocol norow nopct;
run;
```

proc Freq permet d'obtenir les fréquences croisées des champagnes et des juges. Les options `nocol`, `norow` et `nopct` suppriment les calculs des pourcentages en ligne, en colonne et totaux, inutiles ici.

Anova sur le champagne. Le tableau 8.6, p. 267 est produit par les instructions suivantes.

```
proc Anova data=CHAMPAGNE;
     class Champagne;
     model Note = Champagne;
```

```
     means Champagne;
run;
```

Anova globale. Les tableaux 8.3, p. 266, 8.4, p. 267, 8.7, p. 268, 8.8, p. 269 et la figure 8.2 sont produits par les instructions suivantes.

```
proc GLM data=CHAMPAGNE;
     class Juge Champagne;
     model Note = Juge Champagne;
     lsmeans Champagne / tdiff pdiff cov;
     output out=GLM p=Predite r=Residu;
     symbol1 v=plus c=black;
     symbol2 v=diamond c=black;
     symbol3 v=dot c=black;
     symbol4 v=circle c=black;
     symbol5 v=square c=black;
     symbol6 v=star c=black;
     symbol7 v=triangle c=black;
axis1 order=(17 to 37 by 5) value=(h=2) label=none minor=none;
axis2 order=(-10 to 10 by 5) value=(h=2) label=none minor=none;
legend1 label=(h=3 'Champagne') value=(h=2) across=6
       position=(outside bottom) frame;
proc GPlot data=GLM;
      plot Residu *  Predite = Champagne / vref=0;
run;
```

8.3 *Change-over design*, croissance de moutons

Donnée issues du livre [17].

8.3.1 Plan d'expériences et données

Pour déterminer l'influence d'un régime renforcé en protéines sur la croissance de moutons d'un an, le plan d'expérience suivant a été utilisé avec 12 animaux (notés de 53 à 97) trois régimes (A, B et C) d'une durée de 3 mois chacun et trois périodes de 3 mois (notées 1, 2 et 3). Le gain de poids est donné dans chaque cellule du tableau 8.9 accompagné du régime.

animal	53	54	58	75	81	97	72	79	96	84	89	70
Pér. 1	A 72	B 75	C 75	A 64	B 80	C 74	A 58	B 64	C 72	B 76	A 61	C 71
Pér. 2	B 73	C 78	A 77	C 68	A 72	B 76	C 62	A 56	B 69	C 79	B 50	A 72
Pér. 3	C 77	A 70	B 73	B 71	C 80	A 70	B 67	C 60	A 66	A 65	C 60	B 75

TABLE 8.9 – Change-over croissance de moutons

Il y a un facteur étudié (régime), 2 facteurs de contrôle d'hétérogénéité (animal et période), 3 unités de traitement (A, B et C) et 36 unités expérimentales (animal × période). Il y a 12 × 3 × 3=108 cellules, mais seulement 36 d'entre elles ont été

réalisées. C'est un plan, appelé *cross-over* ou *change-over*, constitué de 4 carré-latins de dimension 3, séparés par une double barre verticale dans le tableau 8.9. On reconnait un carré latin au fait que le nombre de lignes est égal au nombre de colonnes et au nombre de traitements, et que chaque traitement apparaît une et une seule fois dans chaque ligne et chaque colonne. Il n'y a pas orthogonalité entre les facteurs régime et régime précédent parce qu'un régime ne peut pas être précédé par lui-même, ni entre période et régime précédent parce que la période 1 est toujours précédée d'une absence de régime (c'est à dire d'un régime de base), ni enfin entre animal et régime précédent parce qu'animal n'a que 3 périodes d'observations et qu'il y a 4 régimes précédents possibles. Si on supprime le facteur régime précédent, les effets des 3 facteurs restants sont orthogonaux.

8.3.2 Analyse des résultats

Le modèle de l'expérience correspondant au programme PG-SAS 2 est le suivant :

$$Y_{ijk} = \mu + \alpha_i + \beta_{a(jk)} + \gamma_j + \delta_k + E_{ijk} \tag{8.3}$$

où α_i est l'effet du régime i, $a(jk)$ est le régime donné à la période $j-1$ à l'animal k, $\beta_{a(jk)}$ est l'arrière-effet du régime donné à la période $j-1$ sur le résultat de la période j, γ_j est l'effet de la période j et δ_k est l'effet de l'animal k. Ce dernier effet peut aussi être considéré comme aléatoire, ce qui n'est pas fait ici. L'analyse de variance montre que le régime a un effet direct, mais l'arrière effet sur la période suivante n'est pas statistiquement significatif.

```
/* Donnees */
data sheep;
input periode traitement \$ animal poids @@;
cards;
1 A 53 72 1 B 54 75 1 C 58 75 1 A 75 64 1 B 81 80 1 C 97 74 2 B 53 73
2 C 54 78 2 A 58 77 2 C 75 68 2 A 81 72 2 B 97 76 3 C 53 77 3 A 54 70
3 B 58 73 3 B 75 71 3 C 81 80 3 A 97 70 1 A 72 58 1 B 79 64 1 C 96 72
1 B 84 76 1 A 89 61 1 C 70 71 2 C 72 62 2 A 79 56 2 B 96 69 2 C 84 79
2 B 89 50 2 A 70 72 3 B 72 67 3 C 79 60 3 A 96 66 3 A 84 65 3 C 89 60
3 B 70 75
;
proc sort data=sheep; by animal periode;
data sheep1;set sheep;traitprec=lag(traitement);
if periode=1 then traitprec='.';

PG-SAS 2/* Analyse de variance cross-over */
proc glm data=sheep1;
class periode animal traitement traitprec;
model poids=periode animal traitement traitprec;
means traitement /regwq;run;
```

```
                                The GLM Procedure

                        Informations sur le niveau de classe
```

```
            Classe          Niveaux    Valeurs
            periode               3    1 2 3
            animal               12    53 54 58 70 72 75 79 81 84 89 96 97
            traitement            3    A B C
            traitprec             4    . A B C

                      Number of Observations Read          36
                      Number of Observations Used          36

Dependent Variable: poids
                              Somme des
Source              DF          carrés   Carré moyen          F    Pr > F
Model               17     1625.666667     95.627451       7.67    <.0001
Error               18      224.333333     12.462963
Corrected Total     35     1850.000000

                 R-carré     Coeff Var    Racine MSE    poids Moyenne
                0.878739      5.067405      3.530292         69.66667


Source              DF       Type I SS     Carré moyen          F    Pr > F
periode              2        4.666667        2.333333       0.19    0.8309
animal              11     1427.333333      129.757576      10.41    <.0001
traitement           2      138.166667       69.083333       5.54    0.0133
traitprec            2       55.500000       27.750000       2.23    0.1368

Source              DF     Type III SS     Carré moyen          F    Pr > F
periode              1        0.166667        0.166667       0.01    0.9092
animal              11     1328.800000      120.800000       9.69    <.0001
traitement           2      191.633333       95.816667       7.69    0.0039
traitprec            2       55.500000       27.750000       2.23    0.1368

        Test d'intervalle multiple de Ryan-Einot-Gabriel-Welsch pour poids
        NOTE: Ce test contrôle le taux d'erreur par expérience de Type I.

           Les moyennes avec la même lettre ne sont pas très différentes.
          REGWQ Groupement       Moyenne     Nb    traitement
                        A        71.333     12    C
                        A
                        A        70.750     12    B

                        B        66.917     12    A
```

8.4 Plan fractionnaire, fabrication de biscuit

8.4.1 Plan d'expériences et données

La fabrication de biscuits peut être schématisée de la façon suivante : Matières premières → Pétrissage → Repos → Laminage-Découpage → Cuisson → Conditionne-

ment. A l'aide d'un procédé pilote de biscuiterie, les étapes de pétrissage et laminage sont étudiées afin de maîtriser la taille des biscuits. Cette étude concerne 8 facteurs à 2 niveaux : la farine (noté Facteur A), le degré d'hydratation (B), la vitesse (C) et la durée de pétrissage (D), la température du bain-marie au repos (E), le temps de repos (F), la vitesse de laminage (G) et l'écartement des cylindres (H). Le facteur Farine (2 variétés de blé) est qualitatif, les autres sont quantitatifs. Les deux premiers facteurs (Farine, Hydratation) correspondent aux caractéristiques de la matière première, les autres facteurs correspondent aux principaux paramètres technologiques du pétrissage et du laminage. On définit un niveau "haut" et un niveau "bas" pour chaque facteur. Une des réponses étudiées est la compressibilité de la pâte à 50 grammes (C50). On cherche à la rendre la plus faible possible de sorte que le découpage se passe bien (il est difficile de découper du caoutchouc qui est très compressible)

On veut faire une expérimentation pour mettre en évidence les facteurs influents sur la compressibilité de la pate.

Un plan factoriel complet nécessiterait $2^8 = 256$ essais. Sachant qu'on ne peut faire qu'une fabrication par jour sur le pilote, un tel plan demanderait près d'une année, ce qui est exclu d'autant plus que le pilote doit être libéré pour d'autres expériences. Le responsable du pilote limite le nombre d'essais à 16 expériences au maximum. On choisit un plan fractionnaire 2^{8-4}, c'est à dire qu'on réalise un seizième des essais du plan factoriel complet. C'est un plan de résolution 4, composé des essais qui sont tels que E=BCD, F=ACD, G=ABC et H=ABD. Ce choix est guidé par le fait qu'on pense a priori que les interactions susceptibles d'exister sont AB, AC, AD , BD et CD. On ne veut donc pas confondre ces interactions entre elles, ce qui est le cas avec le plan choisi. Par exemple les interactions d'ordre 2 confondues avec AB sont GC (obtenu en multipliant les 2 membres de l'égalité G=ABC par ABG), HD et EF (obtenu en multipliant les 2 premiers générateurs d'alias entre eux). Dans ce plan il y a 8 facteurs étudiés (dont un qualitatif, la farine et 7 continus), aucun facteur de contrôle de l'hétérogénéité, 256 cellules et 16 unités expérimentales. Ce plan peut-être construit avec le programme SAS suivant (inclus dans le paquet SAS-QC) :

```
proc factex;
factors A B C D E F G H;
size design=minimum;
model estimate=(A B C D E F G H) nonneg=(A*B A*C A*D) ;
*nonneg=() signifie que les éléments qui suivent ne sont pas
               négligeables;
examine design confounding aliasing(3);
run;
```

Le plan et les résultats sont donnés dans le tableau 8.10

8.4.2 Analyse des résultats

Trois facteurs sont statistiquement significatifs au seuil de 5%, farine, hydratation et vitesse. Pour avoir une faible compressibilité il faut donc choisir la farine codée 1, et les niveaux faibles pour les facteurs hydratation et vitesse. Les autres facteurs peuvent être mis au niveau le plus économique car ils n'ont pas d'impact prouvé sur la compressibilité.

	Farine	Hydrat	Vitesse	Durée	TBM	Repos	Débit	Ecart	C50
	A	B	C	D	E	F	G	H	
1	-1	-1	-1	-1	-1	-1	-1	-1	364
2	1	-1	-1	-1	-1	1	1	1	310
3	-1	1	-1	-1	1	-1	1	1	565
4	1	1	-1	-1	1	1	-1	-1	417
5	-1	-1	1	-1	1	1	1	-1	495
6	1	-1	1	-1	1	-1	-1	1	476
7	-1	1	1	-1	-1	1	-1	1	554
8	1	1	1	-1	-1	-1	1	-1	502
9	-1	-1	-1	1	1	1	-1	1	489
10	1	-1	-1	1	1	-1	1	-1	440
11	-1	1	-1	1	-1	1	1	-1	531
12	1	1	-1	1	-1	-1	-1	1	466
13	-1	-1	1	1	-1	-1	1	1	532
14	1	-1	1	1	-1	1	-1	-1	485
15	-1	1	1	1	1	-1	-1	-1	546
16	1	1	1	1	1	1	1	1	504

TABLE 8.10 – Plan d'expériences Biscuits

```
 data biscuits;
input farine hydrat vitesse duree TBM repos debit ecart C50;
cards;
  -1 -1 -1 -1 -1 -1 -1 -1 364
  1 -1 -1 -1 -1 1 1 1 310
  -1 1 -1 -1 1 -1 1 1 565
  1 1 -1 -1 1 1 -1 -1 417
  -1 -1 1 -1 1 1 1 -1 495
  1 -1 1 -1 1 -1 -1 1 476
  -1 1 1 -1 -1 1 -1 1 554
  1 1 1 -1 -1 -1 1 -1 502
  -1 -1 -1 1 1 1 -1 1 489
  1 -1 -1 1 1 -1 1 -1 440
  -1 1 -1 1 -1 1 1 -1 531
  1 1 -1 1 -1 -1 -1 1 466
  -1 -1 1 1 -1 -1 1 1 532
  1 -1 1 1 -1 1 -1 -1 485
  -1 1 1 1 1 -1 -1 -1 546
  1 1 1 1 1 1 1 1 504
  ;
 proc glm;  class farine;
model  C50= farine  hydrat vitesse duree TBM repos debit ecart
            /solution;
run;

                              Classe       Niveaux    Valeurs
```

```
                          farine             2    -1 1

                    Number of Observations Read          16
                    Number of Observations Used          16

Dependent Variable: C50
                           Somme des
Source             DF        carrés     Carré moyen      F   Pr > F
Model               8   55976.00000      6997.00000   3.08   0.0782
Error               7   15913.00000      2273.28571
Corrected Total    15   71889.00000

              R-carré     Coeff Var    Racine MSE    C50 Moyenne
             0.778645      9.938298      47.67899       479.7500


Source      DF       Type I SS     Carré moyen          F    Pr > F
farine       1     14161.00000     14161.00000       6.23    0.0412
hydrat       1     15252.25000     15252.25000       6.71    0.0359
vitesse      1     16384.00000     16384.00000       7.21    0.0313
duree        1      6006.25000      6006.25000       2.64    0.1481
TBM          1      2209.00000      2209.00000       0.97    0.3571
repos        1       702.25000       702.25000       0.31    0.5957
debit        1       420.25000       420.25000       0.18    0.6802
ecart        1       841.00000       841.00000       0.37    0.5622


                                      Erreur
Paramètre       Estimation           standard   test t    Pr > |t|
Intercept      450.0000000 B   16.85706719    26.70      <.0001
farine    -1    59.5000000 B   23.83949304     2.50      0.0412
farine    1      0.0000000 B     .               .         .
hydrat          30.8750000     11.91974652     2.59      0.0359
vitesse         32.0000000     11.91974652     2.68      0.0313
duree           19.3750000     11.91974652     1.63      0.1481
TBM             11.7500000     11.91974652     0.99      0.3571
repos           -6.6250000     11.91974652    -0.56      0.5957
debit            5.1250000     11.91974652     0.43      0.6802
ecart            7.2500000     11.91974652     0.61      0.5622
```

8.5 Blocs incomplets partiellement équilibrés, expression du génome de *Teleost* fish

Les données sont issues de [44].

8.5.1 Plan d'expériences et données

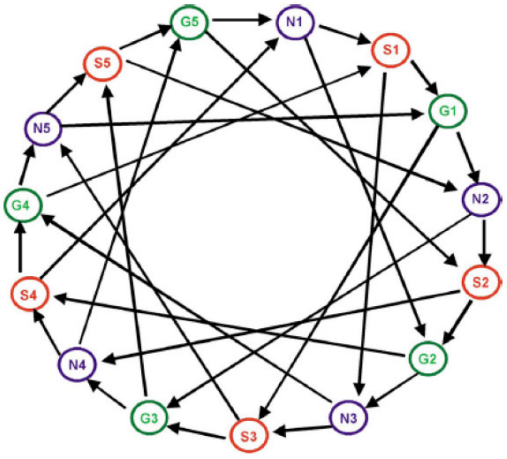


FIGURE 8.3 – Plan d'expériences en boucle Teleost, figure issue de [44]

L'analyse de données d'expression du génome peut se faire grâce à des microarrays (ou lames) à 2 couleurs. Sur un support rectangulaire de verre de quelques cm^2 sont disposés les gènes d'un organisme vivant (en l'occurrence un poisson appelé *Teleost*). On plonge la plaque de verre dans une préparation faite à partir de cellules de cet organisme. On est alors capable de mesurer, par hybridation, l'expression de chacun de ces gènes, c'est à dire la quantité d'ARN fabriqué à partir de ce gène. On peut faire cette mesure pour des cellules placées dans différentes conditions de milieu, ou issues de différents individus ou espèces, et on cherche à comparer l'expression des gènes dans ces différentes conditions. De plus, on peut comparer 2 conditions sur le même microarray, qui joue le rôle de bloc contenant 2 unités expérimentales. Une condition est hybridée avec un marqueur rouge et l'autre avec un marqueur vert. Ces marqueurs ont un effet sur le résultat, et il faut donc prendre en compte ce facteur de contrôle d'hétérogénéité. On veut comparer l'expression de 907 gènes de 3 espèces de poissons *Teleost*. On a 5 poissons de chaque espèce, et pour chaque poisson, on dispose de 4 préparations pour l'hybridation. On ne dispose donc pas d'un nombre suffisant de préparations pour faire toutes les comparaisons 2 à 2 entre les individus d'espèces différentes, dans un même microarray. Il faudrait $(5+5) \times 2 = 20$ préparations pour chaque individu. Les auteurs décident donc d'utiliser le plan en boucle représenté par la figure 8.3. Dans cette figure les noeuds du graphe sont les individus. Comme il y a 3 espèces et 5 individus de chaque espèce, ces derniers sont numérotés de N1 à N5 pour l'espèce *Northern Fundulus heteroclitus*, S1 à S5 pour l'espèce *Southern Fundulus heteroclitus* et G1 à G5 pour l'espèce *Fundulus grandis*. Il y a un arc entre 2 noeuds si les 2 individus ont été hybridés sur le même microarray. L'arc est orienté : il part de l'individu marqué en rouge et va vers l'individu marqué en vert. Comme on le voit sur le dessin il y a un équilibre dans ce dispositif : chaque poisson est analysé 4 fois, 2 fois en rouge et 2 fois en vert. Chaque poisson est dans un même microarray avec 4 poissons d'espèces différentes, 2 fois avec une espèce et 2 fois avec l'autre.

Il y a un facteur d'intérêt, l'espèce, 3 facteurs de contrôle d'hétérogénéité, l'individu,

le microarray et la couleur de marquage. Il y a 60 microarrays (autant que d'arcs). Le plan d'expériences est un plan en blocs incomplets partiellement équilibré.

8.5.2 Analyse des résultats

Pour chaque gène on dispose de $15 \times 4 = 60$ données. Le modèle statistique pour chaque gène est le suivant :

$$Y_{ijkl} = \mu + A_i + \delta_j + \beta_k + E_{kl} + F_{ijkl}, \tag{8.4}$$

où Y_{ijkl} est l'expression mesurée sur le microarray i avec la couleur j pour l'échantillon l de l'espèce k. A est l'effet aléatoire du microarray, δ est l'effet fixe de la couleur. Le paramètre d'intérêt est l'effet de l'espèce k, β_k, $\beta_k - \beta_{k'}$ est la différence d'expression pour le gène considéré entre les espèces k et k'. E_{kl} est l'effet aléatoire de l'individu l pour l'espèce k et F_{ijkl} représente l'erreur de mesure pour l'expression du gène considéré sur le microarray i avec la couleur j et l'individu l de l'espèce k, avec $\mathbb{V}(E_{kl}) = \sigma_E^2$ et $\mathbb{V}(F_{ijkl}) = \sigma_F^2$.

Il y a 907 gènes, ce qui implique d'utiliser le modèle mixte 907 fois. Une autre possibilité proposée par [41] consiste à utiliser les différences $Y_{ijkl} - Y_{ij'k'l'}$. On a en effet une simplification car

$$\begin{aligned} Y_{ijkl} - Y_{ij'k'l'} &= \delta_j - \delta_{j'} + \beta_k - \beta_{k'} + E_{kl} + F_{ijkl} - E_{k'l'} - F_{ij'k'l'} \\ D_{klk'l'} &= \alpha + \beta_k - \beta_{k'} + G_{klk'l'} \end{aligned}$$

avec $\mathbb{V}(G_{klk'l'}) = 2\sigma_E^2 + 2\sigma_F^2$. On peut estimer les variances σ_E^2 et σ_F^2 à l'aide d'un estimateur sans biais facile à calculer. Avec cette méthode et compte tenu du contrôle du risque de première espèce lié à la multiplicité des tests (il y a 3 fois 907 tests) on identifie 9 gènes pour lesquels il y a des différences d'expression entre les espèces. Les lecteurs intéressés par l'analyse de ces données peuvent trouver 2 méthodes d'analyse dans respectivement [44] et dans [41].

Voici un exemple de traitement, pour un gène sur les 907, on a les 60 valeurs donnés dans le programme SAS pour le gène "KIAA0240" :

```
    data KIAA0240;
input N1_36_Cy3 S1_36_Cy5 S1_11_Cy3 G1_11_Cy5 G1_25_Cy3  N2_25_Cy5
N2_04_Cy3   S2_04_Cy5 S2_14_Cy3   G2_14_Cy5   G2_17_Cy3  N3_17_Cy5
 N3_13_Cy3  S3_13_Cy5  S3_47_Cy3   G3_47_Cy5 G3_22_Cy3   N4_22_Cy5
 N4_42_Cy3  S4_42_Cy5  S4_80_Cy3   G4_80_Cy5  G4_32_Cy3  N5_32_Cy5
 N5_21_Cy3  S5_21_Cy5  S5_60_Cy3   G5_60_Cy5  G5_30_Cy3  N1_30_Cy5
 N1_10_Cy3  G2_10_Cy5 G2_24_Cy3   S4_24_Cy5   S4_26_Cy3  N1_26_Cy5
 S1_45_Cy3  N3_45_Cy5  N3_90_Cy3   G4_90_Cy5 G4_70_Cy3   S1_70_Cy5
 G1_28_Cy3  S3_28_Cy5  S3_34_Cy3   N5_34_Cy5  N5_27_Cy3  G1_27_Cy5
 N2_12_Cy3  G3_12_Cy5  G3_23_Cy3   S5_23_Cy5  S5_19_Cy3  N2_19_Cy5
 S2_16_Cy3  N4_16_Cy5  N4_40_Cy3   G5_40_Cy5  G5_43_Cy3  S2_43_Cy5;
lines;
10013 14229 16315 12017 6178 5594 6421 3844 10982 17825 11002 13186
5813 4693 6899 12214 7822 9882 9666 9790 5863 8807 7257 8331 8554
```

```
12513 3857 5413 13908 14523 9112 9713 11223 18757 12968 12491 10816
12476 16846 25298 5476 5354 20101 21834 9278 11792 10085 12668 10203
10631 9535 10691 5173 5592 6463 19845 12804 21341 9550 11373
;
proc transpose data=KIAA0240 out=TKIAA0240;
datat TKIAA0240;set TKIAA0240;Espece=substr(_NAME_,1,1);
individu=substr(_NAME_,2,1);
Microarray=substr(_NAME_,4,2);Couleur=substr(_NAME_,7,3);
LogExpress=log2(Col1);drop _NAME_ Col1;
proc print;
Proc mixed data=TKIAA0240 covtest;
class Espece Couleur Microarray ;
model logExpress = Espece Couleur;
random Microarray ;
lsmeans Espece /diff;
run;quit;
```

et on obtient les résultats suivants en utilisant le modèle mixte 8.4. La variance due à l'individu (non significative) a été supprimée dans le modèle. On observe qu'il n'y a pas de différence entre les 3 espèces pour ce gène. Il faut répéter cette analyse pour les 906 autres gènes. Il faut noter que comme il y a 907 gènes, un contrôle pour les tests multiples doit être effectué, ce qui implique que pour être jugée significative, une différence entre 2 conditions pour un gène doit avoir une *Pvalue* nettement plus faible que 5%.

Obs	Espece	individu	Microarray	Couleur	Log Express
1	N	1	36	Cy3	13.2896
2	S	1	36	Cy5	13.7965
3	S	1	11	Cy3	13.9939
4	G	1	11	Cy5	13.5528
5	G	1	25	Cy3	12.5929
6	N	2	25	Cy5	12.4497
7	N	2	04	Cy3	12.6486
8	S	2	04	Cy5	11.9084
9	S	2	14	Cy3	13.4229
10	G	2	14	Cy5	14.1216
11	G	2	17	Cy3	13.4255
12	N	3	17	Cy5	13.6867
13	N	3	13	Cy3	12.5051
14	S	3	13	Cy5	12.1963
15	S	3	47	Cy3	12.7522
16	G	3	47	Cy5	13.5762
17	G	3	22	Cy3	12.9333
18	N	4	22	Cy5	13.2706
19	N	4	42	Cy3	13.2387
20	S	4	42	Cy5	13.2571
21	S	4	80	Cy3	12.5174
22	G	4	80	Cy5	13.1044
23	G	4	32	Cy3	12.8252

24	N	5	32	Cy5	13.0243
25	N	5	21	Cy3	13.0624
26	S	5	21	Cy5	13.6111
27	S	5	60	Cy3	11.9133
28	G	5	60	Cy5	12.4022
29	G	5	30	Cy3	13.7636
30	N	1	30	Cy5	13.8261
31	N	1	10	Cy3	13.1536
32	G	2	10	Cy5	13.2457
33	G	2	24	Cy3	13.4542
34	S	4	24	Cy5	14.1951
35	S	4	26	Cy3	13.6627
36	N	1	26	Cy5	13.6086
37	S	1	45	Cy3	13.4009
38	N	3	45	Cy5	13.6069
39	N	3	90	Cy3	14.0401
40	G	4	90	Cy5	14.6267
41	G	4	70	Cy3	12.4189
42	S	1	70	Cy5	12.3864
43	G	1	28	Cy3	14.2950
44	S	3	28	Cy5	14.4143
45	S	3	34	Cy3	13.1796
46	N	5	34	Cy5	13.5255
47	N	5	27	Cy3	13.2999
48	G	1	27	Cy5	13.6289
49	N	2	12	Cy3	13.3167
50	G	3	12	Cy5	13.3760
51	G	3	23	Cy3	13.2190
52	S	5	23	Cy5	13.3841
53	S	5	19	Cy3	12.3368
54	N	2	19	Cy5	12.4491
55	S	2	16	Cy3	12.6580
56	N	4	16	Cy5	14.2765
57	N	4	40	Cy3	13.6443
58	G	5	40	Cy5	14.3813
59	G	5	43	Cy3	13.2213
60	S	2	43	Cy5	13.4733

The Mixed Procedure

Informations sur le modèle

Data Set	WORK.TKIAA0240
Dependent Variable	LogExpress
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Informations sur les niveaux de classe

Classe	Niveaux	Valeurs
Espece	3	G N S
Couleur	2	Cy3 Cy5
Microarray	30	04 10 11 12 13 14 16 17 19 21 22 23 24 25 26 27 28 30 32 34 36 40 42 43 45 47 60 70 80 90

Dimensions

Covariance Parameters	2
Columns in X	6
Columns in Z	30
Subjects	1
Max Obs Per Subject	60

Nombre d'observations

Number of Observations Read	60
Number of Observations Used	60
Number of Observations Not Used	0

Historique des itérations

Itération	Evaluations	-2 Log-vrais. des rés.	Critère
0	1	117.82319198	
1	2	94.84931770	0.00018911
2	1	94.84854674	0.00000004
3	1	94.84854656	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Parm cov	Estimation	Erreur type	Valeur Z	Pr Z
Microarray	0.2872	0.08889	3.23	0.0006
Residual	0.09473	0.02573	3.68	0.0001

Statistiques d'ajustement

-2 Res Log Likelihood	94.8
AIC (smaller is better)	98.8
AICC (smaller is better)	99.1
BIC (smaller is better)	101.7

The Mixed Procedure

Type 3 Tests des effets fixes

Effet	Num DF	Den DDL	Valeur F	Pr > F
Espece	2	27	1.66	0.2088
Couleur	1	27	11.76	0.0020

Moyennes des moindres carrés

Effet	Espece	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
Espece	G	13.3543	0.1232	27	108.42	<.0001
Espece	N	13.3100	0.1232	27	108.06	<.0001
Espece	S	13.1631	0.1232	27	106.87	<.0001

Différences des moyennes des moindres carrés

Effet	Espece	_Espece	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
Espece	G	N	0.04436	0.1098	27	0.40	0.6895
Espece	G	S	0.1912	0.1098	27	1.74	0.0930
Espece	N	S	0.1469	0.1098	27	1.34	0.1923

8.6 Plan fractionnaire Thermostat

Données issues de [9] avec l'aimable autorisation de Bill Woodall.

Eaton Corp. fabrique des thermostats industriels. Un défaut rare mais grave et inexpliqué, apparaît de façon épisodique. La proportion de thermostats défectueux, qui est seulement de 200 pour un million en temps normal, augmente brutalement à 70 000 pour un million de façon sporadique. Les thermostats défectueux ont une durée de vie raccourcie, de l'ordre de 30 000 cycles au lieu de 500 000 cycles pour un thermostat normal. Un cycle est constitué d'un changement de mode du thermostat de ON à OFF puis de OFF à ON, du à des changements de température. Le processus de fabrication est complexe et comporte de nombreuses étapes. Un groupe de travail interdisciplinaire comprenant des physiciens, des chimistes et des ingénieurs de production, mais aussi des techniciens opérateurs sur le site de production s'est réuni sur une durée de 150 hommes×heures, et a identifié 50 causes possibles, puis a choisi de tester 11 d'entre elles et choisi un niveau haut et un niveau bas pour chacun des facteurs. Les 11 facteurs sont les suivants :

8.6.1 Plan d'expériences et données

L'usine a été fermée à la production pour effectuer les essais. Un plan Plackett-Burman avec 12 essais a été choisi. C'est un plan fractionnaire différent de ceux définis

Lettre	Facteur
A	Diaphragm Platine Rinse
B	Current Density
C	Sulfuric Acid Cleaning
D	Diaphragm electroclean
E	Beryllium Copper Grain Size
F	Stress Orientation
G	Humidity
H	Heat Treatment
I	Brazing Machine
J	Power Element Electroclean
K	Power Element Platine Rinse

TABLE 8.11 – Liste des facteurs du plan d'expériences

dans la section 7.4.3, p. 239, de fraction 12/512 et de résolution 3, c'est à dire que les effets principaux sont confondus avec des interactions d'ordre 2. Lors de chaque essai on a fabriqué et testé 10 thermostats. On a obtenu les moyennes des logarithmes des 10 durées de vie :

Essai	A	B	C	D	E	F	G	H	I	J	K	log(Nbcycles)
1	-	-	-	-	-	-	-	-	-	-	-	4.673
2	-	-	-	-	-	+	+	+	+	+	+	2.530
3	-	-	+	+	+	-	-	-	+	+	+	2.233
4	-	+	-	+	+	-	+	+	-	-	+	2.409
5	-	+	+	-	+	+	-	+	-	+	-	2.489
6	-	+	+	+	-	+	+	-	+	-	-	4.059
7	+	-	+	+	-	-	+	+	-	+	-	2.623
8	+	-	+	-	+	+	+	-	-	-	+	2.106
9	+	-	-	+	+	+	-	+	+	-	-	2.502
10	+	+	+	-	-	-	-	+	+	-	+	2.552
11	+	+	-	+	-	+	-	-	-	+	+	5.352
12	+	+	-	-	+	-	+	-	+	+	-	2.151

8.6.2 Analyse des résultats et conclusions

On aurait pu analyser les 10 répétitions séparément au lieu de prendre la moyenne. Cependant les 10 valeurs pour chaque essai ne sont pas de réelles répétitions puisqu'il ne s'agit pas d'essais différents mais de produits obtenus lors du même essai. La variabilité entre ces répétitions est certainement inférieure à celle qu'on aurait obtenue avec 10 essais différents et on ne peut pas la prendre en compte pour estimer la variabilité résiduelle.

Si on considère le modèle avec tous les facteurs, on ne peut pas estimer la variance résiduelle par manque de degrés de liberté. Cela arrive fréquemment dans ce contexte industriel où on ne peut pas faire beaucoup d'essais. Dans ce cas on considère a priori que 60% des facteurs n'ont pas d'effet. On ordonne donc les effets par ordre croissant de somme de carrés et on élimine les 7 premiers du modèle. On obtient les résultats de

la dernière procédure GLM. Cette méthode peut perdre de la puissance dans la mesure où on intègre dans la somme de carrés résiduelle des effets de facteurs qui peuvent ne pas être négligeables. Par contre le risque de première espèce est bien contrôlé.

```
data thermostat;
input Essai   A   B   C   D   E   F   G   H   I   J   K
LNbcycles;
lines;
1   -1   -1   -1   -1   -1   -1   -1   -1   -1   -1   -1   4.673
    2   -1   -1   -1   -1   -1   1   1   1   1   1   1   2.530
    3   -1   -1   1   1   1   -1   -1   -1   1   1   1   2.233
    4   -1   1   -1   1   1   -1   1   1   -1   -1   1   2.409
    5   -1   1   1   -1   1   1   -1   1   -1   1   -1   2.489
    6   -1   1   1   1   -1   1   1   -1   1   -1   -1   4.059
    7   1   -1   1   1   -1   -1   1   1   -1   1   -1   2.623
    8   1   -1   1   -1   1   1   1   -1   -1   -1   1   2.106
    9   1   -1   -1   1   1   1   -1   1   1   -1   -1   2.502
    10  1   1   1   -1   -1   -1   -1   1   1   -1   1   2.552
    11  1   1   -1   1   -1   1   -1   -1   -1   1   1   5.352
    12  1   1   -1   -1   1   -1   1   -1   1   1   -1   2.151
;
proc glm data=thermostat outstat=ssout ;
model LNbcycles= A B C D E F G H I J K/ss1 ;run;quit;
data ss;set ssout; keep _SOURCE_ ss;if df >0;
proc sort data=ss;by ss;
proc print;run;
proc glm data=thermostat ;model LNbcycles=  E  G H I /solution;run;quit;
```

```
                                The GLM Procedure

                          Number of Observations Read          12
                          Number of Observations Used          12

Dependent Variable: LNbcycles

                                 Somme des                       Valeur
Source                    DF        carrés     Carré moyen          F    Pr > F

Model                     11     12.97419225      1.17947202        .       .
Error                      0      0.00000000       .
Corrected Total           11     12.97419225


                      R-carré     Coeff Var    Racine MSE    LNbcycles Moyenne
                     1.000000           .              .             2.973250


                                                                     Valeur
Source                    DF       Type I SS     Carré moyen          F    Pr > F

A                          1      0.10212075      0.10212075        .       .
B                          1      0.45825208      0.45825208        .       .
```

C	1	1.05316875	1.05316875	.	.
D	1	0.59719408	0.59719408	.	.
E	1	5.19951675	5.19951675	.	.
F	1	0.47880075	0.47880075	.	.
G	1	1.28249408	1.28249408	.	.
H	1	2.49249675	2.49249675	.	.
I	1	1.09505208	1.09505208	.	.
J	1	0.07099408	0.07099408	.	.
K	1	0.14410208	0.14410208	.	.

Paramètre	Estimation	Erreur standard	Valeur du test t	Pr > |t|
Intercept	2.973250000	.	.	.
A	-0.092250000	.	.	.
B	0.195416667	.	.	.
C	-0.296250000	.	.	.
D	0.223083333	.	.	.
E	-0.658250000	.	.	.
F	0.199750000	.	.	.
G	-0.326916667	.	.	.
H	-0.455750000	.	.	.
I	-0.302083333	.	.	.
J	-0.076916667	.	.	.
K	-0.109583333	.	.	.

Obs	_SOURCE_	SS
1	J	0.07099
2	A	0.10212
3	K	0.14410
4	B	0.45825
5	F	0.47880
6	D	0.59719
7	C	1.05317
8	I	1.09505
9	G	1.28249
10	H	2.49250
11	E	5.19952

The GLM Procedure

Dependent Variable: LNbcycles

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	4	10.06955967	2.51738992	6.07	0.0197
Error	7	2.90463258	0.41494751		
Corrected Total	11	12.97419225			

R-carré	Coeff Var	Racine MSE	LNbcycles Moyenne
0.776122	21.66532	0.644164	2.973250

Source	DF	Type I SS	Carré moyen	Valeur F	Pr > F
E	1	5.19951675	5.19951675	12.53	0.0095
G	1	1.28249408	1.28249408	3.09	0.1221
H	1	2.49249675	2.49249675	6.01	0.0441
I	1	1.09505208	1.09505208	2.64	0.1483

Paramètre	Estimation	Erreur standard	Valeur du test t	Pr > |t|
Intercept	2.973250000	0.18595419	15.99	<.0001
E	-0.658250000	0.18595419	-3.54	0.0095
G	-0.326916667	0.18595419	-1.76	0.1221
H	-0.455750000	0.18595419	-2.45	0.0441
I	-0.302083333	0.18595419	-1.62	0.1483

Deux facteurs influents ont été identifiés : La taille du grain de Beryllium (E) et le traitement à la chaleur (H). Ces 2 facteurs ont été fixés à leur niveau faible qui donne la durée de cycle la plus élevée. Les autres facteurs ont été fixés à leur niveau le plus économique. Avec ces nouveaux réglages, le processus de fabrication est supérieur à l'ancien sur 4 critères :

1. Il produit des thermostats de meilleure qualité : le problème sporadique est supprimé et le taux de thermostats défectueux est inférieur à 200 pour un million.
2. La durée de vie moyenne des thermostats est supérieure à 7 000 000 cycles
3. Les thermostats peuvent fonctionner dans un environnement hostile, plus humide et plus acide, ce qui permet de gagner de nouveaux marchés.
4. Le processus de production est moins coûteux et a une meilleure productivité.

Les membres du groupe de travail ont été convaincus par l'efficacité de cette méthode expérimentale. Ils ont apprécié le fait qu'elle se fonde sur les faits expérimentaux et non pas sur le poids de la hiérarchie ou de ceux qui parlent le plus fort dans un groupe de travail.

8.7 Conception de produits robustes

C'est un exemple fictif et simple qui illustre la notion de conception robuste : il s'agit de faire des produits qui résistent aux utilisateurs. Dans l'exemple qui suit on produit et l'utilisateur stocke des sacs de plâtre, de la colle ou des biscottes... Il y a deux types de facteurs

1. Les facteurs de production dont l'industriel possède la maîtrise et qu'il convient de choisir au mieux. Dans cet exemple ce sont la température et la durée du processus industriel de cuisson du produit.
2. Les facteurs d'usage qui caractérisent les modes d'utilisation possibles du produit par les clients. Dans cet exemple ce sont les conditions de stockage (durée et humidité) avant son utilisation proprement dite.

Un bon produit est un produit qui fonctionne quelle que soient les conditions de stockage et d'utilisation par le client (dans des limites raisonnables). On cherche à produire un objet robuste c'est à dire qui fonctionne correctement dans tous les cas envisagés y compris dans le pire des cas. On cherche donc une bonne "pire performance" et non pas la meilleure performance moyenne. Le plan d'expériences simpliste ci-dessous illustre la problématique. Dans ce cas c'est le couple (température, durée)=(+, -) qui est le meilleur choix car il donne la meilleure "pire performance", (14 contre 10 ou 9 pour les autres produits).

	age	+	+	-	-
	hygrométrie	+	-	+	-
Température	durée				
+	+	11	17	10	18
+	-	15	14	15	14
-	+	10	12	10	12
-	-	13	12	11	9

Performances d'un produit selon les facteurs de production et d'utilisation

En général on construit 2 plans fractionnaires, un pour les facteurs de production et un pour les facteurs d'utilisation et on les croise. Ensuite on considère une "mesure de performance robuste". Il existe plusieurs définitions possibles de cette mesure. On peut considérer la pire valeur de chaque ligne, mais cela a l'inconvénient de reposer sur une seule valeur de la ligne (la pire) et ne pas utiliser les autres valeurs. Certains ont proposé de combiner la moyenne et le logarithme de la variance de chaque ligne avec une pondération ad hoc de ces 2 éléments. Plusieurs possibilités existent sans consensus des spécialistes, voir [42].

8.8 Plan fractionnaire puis surface de réponse, mesure de polluants

Données issues de [6] avec l'aimable autorisation de Marjolaine Bourdat-Deschamps Les hydrocarbures aromatiques polycycliques (PAHs) sont des molécules présentes dans l'atmosphère et donc dans la pluie et les poussières. Certaines d'entre elles sont potentiellement cancérigènes comme le benzo(a)pyrene, et 16 d'entre elles ont été placées par l'agence US *Environmental Protection Agency* sur une liste de polluants prioritaires dont la présence dans l'environnement doit être surveillée. Elles sont extraites d'échantillons

d'eau, puis analysées par une méthode d'analyse chimique dite de "chromatographie en phase liquide à haute performance", HPLC en anglais. Le processus de mesure est difficile parce que ces molécules sont présentes dans l'eau de pluie en faible concentration et que l'étape d'extraction n'est pas bien maîtrisée et dépend de nombreux facteurs. L'objectif est de proposer des valeurs optimales de ces facteurs pour cette étape, valeurs qui pourront ensuite être utilisées en routine.

8.8.1 Recherche de facteurs influents

Après une discussion dans l'équipe de chimistes, 6 facteurs ont été jugés potentiellement importants. Ils sont donnés dans la Table 8.12. Le dernier facteur est un additif acetonitrile qui est utilisé pour éviter l'adsorption de molécules de PAHs dans les parois du récipient. Un plan fractionnaire a été mis au point pour identifier les facteurs influents, puis une analyse en surface de réponse a permis de proposer des valeurs optimales pour ces paramètres. Nous détaillons maintenant ces 2 plans d'expériences.

	Concentration PAH (ng/L) F	Vitesse d'agitation (rpm) C	Temperature (C) D	Temps d'extraction (min) B	Volume de l'échantillon (mL) A	% MeCN E
Low level -1	8	600	23	20	5	2
High level 1	80	1000	40	180	20	12
1	1	1	-1	-1	-1	1
2	1	1	-1	-1	1	-1
3	1	1	1	1	-1	-1
4	1	1	1	1	1	1
5	1	-1	1	-1	-1	-1
6	1	-1	1	-1	1	1
7	1	-1	-1	1	-1	1
8	1	-1	-1	1	1	-1
9	-1	-1	-1	-1	-1	-1
10	-1	-1	-1	-1	1	1
11	-1	-1	1	1	-1	1
12	-1	-1	1	1	1	-1
13	-1	1	1	-1	-1	1
14	-1	1	1	-1	1	-1
15	-1	1	-1	1	-1	-1
16	-1	1	-1	1	1	1
17	1	1	-1	-1	1	-1
18	1	1	-1	-1	-1	-1

TABLE 8.12 – Plan fractionnaire, mesure des hydrocarbures aromatiques polycycliques

La variable réponse est la proportion (moyenne sur les 15 molécules PAH) de molécules de PAH retrouvées. C'est un plan fractionnaire 2^{6-2} de résolution 4, c'est à dire que les effets principaux sont confondus avec les interactions d'ordre 3 et les interactions d'ordre 2 sont confondues avec d'autres interactions d'ordre 2.

La figure 8.4 donne le diagramme de Pareto des effets standardisés (c'est à dire l'effet estimé divisé par son écart-type estimé). Les termes qui dépassent le trait vertical sont significatifs pour un test de niveau 5%. Le temps d'extraction et le volume de l'échantillon

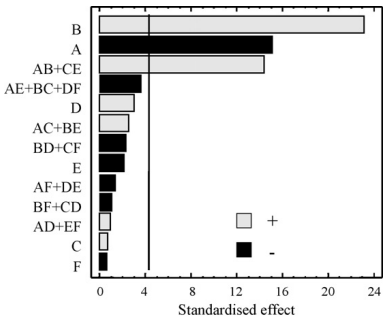


FIGURE 8.4 – Résultats de l'analyse du plan fractionnaire

ont une influence statistiquement significative sur la proportion de molécules retrouvées. Les autres facteurs ne sont pas significatifs au seuil de 5%. La somme des interactions (Temps d'extraction)×(Volume de l'échantillon) et (Vitesse d'agitation)×(% MeCN) est significative. Il faut donc séparer les effets de chaque interaction pour en savoir plus. Les 2 dernières expériences (17 et 18) ont été ajoutées pour supprimer la confusion de ces interactions entre elles. Seule la première est statistiquement significative.

8.8.2 Optimisation des valeurs des facteurs

L'expérience avec le plan fractionnaire a permis d'identifier 2 facteurs influents. On étudie ensuite plus finement leur effet sur la réponse en utilisant une modélisation polynomiale de degré 2. Pour cela on construit un plan 3^2 avec deux répétitions au centre, décrit dans la Table 8.13. On note que le point "au centre" est légèrement décalé pour des raisons pratiques.

	Sample volume (mL)	Extraction time (min)
Low level -1	5	20
High level 1	20	180
Centre 0	10	100
1	-1	1
2	1	0
3	0	0
4	-1	-1
5	0	1
6	1	1
7	0	1
8	-1	1
9	0	0
10	1	-1

TABLE 8.13 – Plan 3^2

Grâce aux résultats de cette expérience le réglage de la procédure de mesure a pu être défini de façon optimale. Pour obtenir un bon résultat tout en gardant un temps d'extraction raisonnable, les auteurs de l'étude concluent qu'une bonne solution consite à faire durer l'extraction pendant 140 minutes en utilisant un volume de 10mL. La

méthodologie expérimentale utilisée dans cette étude a permis d'évaluer l'influence de 6 facteurs sur la qualité de la mesure. Auparavant les chimistes utilisaient des méthodes pas à pas en modifiant la valeur des facteurs un par un, ce qui ne permettait pas de tirer le parti des expériences faites, qui étaient pourtant nombreuses et coûteuses.

8.9 Split-plot : effet secondaire d'un fongicide

Le fongicide est un produit appliqué en champ ou directement sur les semences de céréales, qui élimine les champignons, protège la récolte et améliore le rendement. C'est son effet attendu. A côté de cet effet attendu, on soupçonne qu'il existe un effet secondaire non attendu qui se manifesterait sur le métabolisme de l'azote et par conséquent sur le rendement en grain. Dans cette expérience on s'intéresse à cet effet secondaire. Les effets intéressant le plus l'expérimentateur sont l'effet du fongicide, mais surtout l'interaction entre le fongicide et la dose d'azote, car si cette interaction existe, cela tendrait à démontrer un effet "non attendu" du fongicide sur l'utilisation de l'azote. En revanche, l'estimation de l'effet moyen des doses d'azote sur le rendement est secondaire dans l'essai. Le dispositif split-plot (Cf section 7.5.2, p. 249) organise une première répartition des niveaux d'azote dans de grandes unités, ce qui permet une économie expérimentale, car cela évite de changer très souvent le réglage du distributeur d'azote en champ. Il établit également une répartition des niveaux du fongicide dans les petites unités, ce qui permettra d'estimer l'interaction avec une grande précision. Il a donc été choisi en raison de cette hiérarchie a priori entre les effets étudiés. Compte tenu de la problématique, l'effet du fongicide a été testé pour différents niveaux d'apports azotés.

8.9.1 Plan d'expérience et données

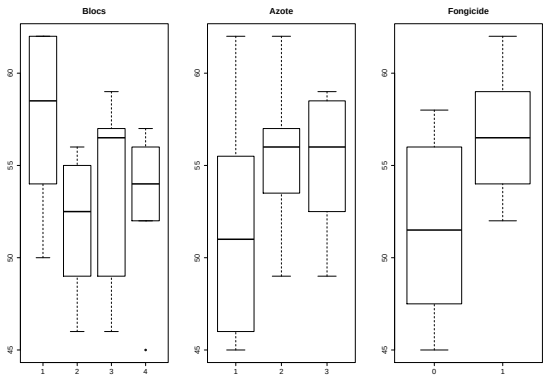
La variable réponse est le rendement du blé (q/ha). On dispose de 4 parcelles (blocs) divisées chacune en 3 grandes unités sur lesquelles sont réparties aléatoirement 3 niveaux d'azote différents. Chacune des grandes unités est divisée en deux petites unités, l'une traitée avec le fongicide et l'autre non (témoin), l'unité traitée étant déterminée aléatoirement. Il y a au total 12 grandes unités, $2 \times 3 \times 12 = 72$ cellules et 24 unités expérimentales (les petites unités), ce qui signifie que toutes les cellules ne sont pas remplies. Il y a 2 facteurs étudiés (Azote et Fongicide), 6 unités de traitement et 2 facteurs de contrôle d'hétérogénéité (bloc et grande unité). Le dispositif est un Split-Plot et le "facteur sacrifié" est la dose d'azote.

Fongicide des semences	F	2 niveaux : 0 1	témoin/traité
Dose d'azote	N	3 niveaux : 1 2 3	100, 140 et 180 kg N/ha
Bloc	B	4 niveaux : 1 2 3 4	

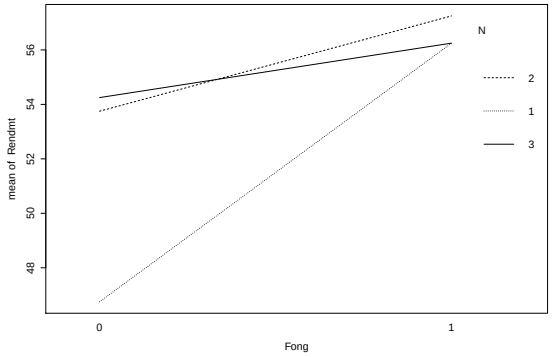
Répartition des traitements dans l'essai et rendement(q/ha)

Bloc 1	F0-N3 58	F1-N3 59	F1-N1 62	F0-N1 50	F0-N2 54	F1-N2 62
Bloc 2	F0-N2 56	F1-N2 53	F0-N3 49	F1-N3 52	F1-N1 55	F0-N1 46
Bloc 3	F0-N1 46	F1-N1 56	F0-N2 49	F1-N2 57	F0-N3 57	F1-N3 59
Bloc 4	F1-N1 52	F0-N1 45	F0-N3 53	F1-N3 55	F1-N2 57	F0-N2 56

8.9.2 Analyse descriptive des données



L'examen des boxplots montre qu'il semble y avoir un effet du fongicide sur le rendement, qu'il y a une certaine variabilité liée aux blocs et que les doses de 140 et 180 kg N/ha entrainent un rendement supérieur à la dose 100 kg N/ha, mais que l'effet des deux doses semble équivalent.



L'examen du graphique d'interaction montre que le fongide associé à la dose de 100 kg N/ha d'azote produit un rendement élevé, sensiblement équivalent à celui obtenu avec les doses de 140 et 180 kg N/ha, et que pour ces deux dernières doses le traitement par le fongicide n'augmente pas le rendement de façon aussi nette.

8.9.3 Analyse des résultats

L'analyse de la variance du modèle décrivant l'effet de l'azote sur le rendement à partir des grandes unités donne un effet de l'azote peu significatif (p-value = 0.094). Dans ce modèle (en considérant l'effet bloc comme fixe), la somme des carrés résiduelle vaut 67 et a 6 degrés de liberté :

```
Error: Azote:Bloc
             Df Sum Sq Mean Sq F value Pr(>F)
Azote         2  80.33   40.17   3.597  0.094
Bloc          3 107.50   35.83   3.209  0.104
Residuals     6  67.00   11.17
```

L'analyse de la variance pour les effets du fongicide et de l'interaction entre azote et fongicide donne

```
Error: Within
                   Df Sum Sq Mean Sq F value   Pr(>F)
Fongicide           1    150  150.00  25.962 0.000649 ***
Azote:Fongicide     2     63   31.50   5.452 0.028111 *
Residuals           9     52    5.78
```

Cette fois ci la résiduelle vaut 52 et a 9 degrés de liberté. L'effet du fongicide est très significatif (p-value = 0.00064) et celui de l'interaction est significatif (p-value = 0.028). L'effet secondaire suspecté est donc confirmé par cette analyse.

Il est à noter que si on avait testé l'effet de l'azote contre cette résiduelle on aurait abouti à la conclusion (erronée) d'un effet significatif avec une statistique de Fisher égale à 6.9519 et une p-value de 0.0149453.

Si maintenant on se place dans le cadre du modèle mixte avec l' effet bloc et l'interaction entre l'azote et le bloc aléatoires, on obtient les mêmes tests F :

```
Analysis of Variance Table
                Df  Sum Sq Mean Sq F value
Azote            2  41.565  20.783  3.5970
Fongicide        1 150.000 150.000 25.9615
Azote:Fongicide  2  63.000  31.500  5.4519
```

avec les estimations suivantes des variances :

```
Random effects:
 Groups        Name        Variance Std.Dev.
 Azote:Bloc   (Intercept) 2.6944   1.6415
 Bloc         (Intercept) 4.1111   2.0276
 Residual                 5.7778   2.4037
```

On remarque que les sommes de carrés et les statistiques de Fisher sont identiques pour les 2 modèles (effets fixes ou effets mixtes), sauf pour la somme de carrés du facteur azote qui a été "renormalisée" dans cette sortie R, de sorte que, divisée par le même carré moyen résiduel que les autres effets, elle donne le F correct. Les estimations des variances des effets aléatoires donnent des indications sur la variabilité expérimentale due aux blocs, celle due aux grandes parcelles et la variabilité résiduelle. Il faut remarquer que la colonne Std.Dev n'est pas l'écart-type estimé de l'estimateur de la variance correspondante, comme dans certains logiciels comme SAS. C'est simplement la racine de la colonne Variance.

Le programme SAS suivant :

```
data SplitPlot;
input Obs Fongicide Azote Bloc Rendement @@;
```

```
lines;
 1 0 1 1 50  2 0 1 2 46  3 0 1 3 46  4 0 1 4 45  5 0 2 1 54  6 0 2 2 56
 7 0 2 3 49  8 0 2 4 56  9 0 3 1 58 10 0 3 2 49 11 0 3 3 57 12 0 3 4 53
13 1 1 1 62 14 1 1 2 55 15 1 1 3 56 16 1 1 4 52 17 1 2 1 62 18 1 2 2 53
19 1 2 3 57 20 1 2 4 57 21 1 3 1 59 22 1 3 2 52 23 1 3 3 59 24 1 3 4 55
;
Proc mixed data=SplitPlot covtest method=reml;
class Fongicide Azote Bloc;
model rendement =  Fongicide|Azote @2/ddfm=satterth;
random  Bloc Bloc*Azote ;
lsmeans Fongicide*Azote /diff;
run;quit;
```

donne les résultats suivants :

The Mixed Procedure

Informations sur le modèle

Data Set	WORK.SPLITPLOT
Dependent Variable	Rendement
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Informations sur les niveaux de classe

Classe	Niveaux	Valeurs
Fongicide	2	0 1
Azote	3	1 2 3
Bloc	4	1 2 3 4

Dimensions

Covariance Parameters	3
Columns in X	12
Columns in Z	16
Subjects	1
Max Obs Per Subject	24

Nombre d'observations

Number of Observations Read	24
Number of Observations Used	24
Number of Observations Not Used	0

Historique des itérations

Itération	Evaluations	-2 Log-vrais. des rés.	Critère
0	1	104.98227073	
1	1	100.39996021	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Parm cov	Estimation	Erreur type	Valeur Z	Pr Z
Bloc	4.1111	4.9933	0.82	0.2052
Azote*Bloc	2.6944	3.4994	0.77	0.2207
Residual	5.7778	2.7237	2.12	0.0169

Statistiques d'ajustement

-2 Res Log Likelihood	100.4
AIC (smaller is better)	106.4
AICC (smaller is better)	108.1
BIC (smaller is better)	104.6

Type 3 Tests des effets fixes

Effet	Num DF	Den DDL	Valeur F	Pr > F
Fongicide	1	9	25.96	0.0006
Azote	2	6	3.60	0.0940
Fongicide*Azote	2	9	5.45	0.0281

Moyennes des moindres carrés

Effet	Fongicide	Azote	Estimation	Erreur type	DDL	test t	Pr > |t|
Fongicide*Azote	0	1	46.7500	1.7736	10.5	26.36	<.0001
Fongicide*Azote	0	2	53.7500	1.7736	10.5	30.30	<.0001
Fongicide*Azote	0	3	54.2500	1.7736	10.5	30.59	<.0001
Fongicide*Azote	1	1	56.2500	1.7736	10.5	31.71	<.0001
Fongicide*Azote	1	2	57.2500	1.7736	10.5	32.28	<.0001
Fongicide*Azote	1	3	56.2500	1.7736	10.5	31.71	<.0001

Effet	Fongicide	Azote	Fongicide	Azote	Estimation	e.t.	DDL	t	Pr > |t|
Fong*Azote	0	1	0	2	-7.0000	2.0582	11.7	-3.40	0.0054
Fong*Azote	0	1	0	3	-7.5000	2.0582	11.7	-3.64	0.0035

```
Fong*Azote  0   1   1   1   -9.5000   1.6997     9  -5.59  0.0003
Fong*Azote  0   1   1   2  -10.5000   2.0582  11.7  -5.10  0.0003
Fong*Azote  0   1   1   3   -9.5000   2.0582  11.7  -4.62  0.0006
Fong*Azote  0   2   0   3   -0.5000   2.0582  11.7  -0.24  0.8123
Fong*Azote  0   2   1   1   -2.5000   2.0582  11.7  -1.21  0.2484
Fong*Azote  0   2   1   2   -3.5000   1.6997     9  -2.06  0.0696
Fong*Azote  0   2   1   3   -2.5000   2.0582  11.7  -1.21  0.2484
Fong*Azote  0   3   1   1   -2.0000   2.0582  11.7  -0.97  0.3508
Fong*Azote  0   3   1   2   -3.0000   2.0582  11.7  -1.46  0.1712
Fong*Azote  0   3   1   3   -2.0000   1.6997     9  -1.18  0.2695
Fong*Azote  1   1   1   2   -1.0000   2.0582  11.7  -0.49  0.6360
Fong*Azote  1   1   1   3 -8,24E-16   2.0582  11.7  -0.00  1.0000
Fong*Azote  1   2   1   3    1.0000   2.0582  11.7   0.49  0.6360
```

Les trois dernières lignes du tableau montrent qu'il n'y a pas de différence statistiquement significative entre les trois doses d'azote quand le fongicide est présent. On trouve les mêmes tests F qu'avec R. Il faut noter que les degrés de liberté des tests de comparaisons multiples entre les unités de traitement fongicide*azote dépendent de l'option *ddfm* choisie dans la ligne *model* . Si on prend l'option par défaut, tous les tests ont 9 degrés de liberté, alors que si on choisit l'option *ddfm=satterth* les tests ont 11.7 degrés de liberté, sauf ceux internes à une grande parcelle qui en ont 9.

8.10 Plusieurs variables réponses

Données issues de [7]

Il s'agit de maximiser le rendement d'une réaction chimique en fonction de la température et de la durée de réaction et de déterminer la position du maximum. Les meilleures conditions connues sont une durée de 90' et une température égale à 145°. On pose $x_1 = (duree - 90)/10$ et $x_2 = (temperature - 145)/5$

Le programme SAS suivant permet d'analyser la surface de réponse.

```
    data box;
input essai duree temperature x1  x2    rendement;
lines;
11          100      150         1          1         78.8
12          100      140         1         -1         84.5
13           80      150        -1          1         91.2
14           80      140        -1         -1         77.4
15           90      145         0          0         89.7
16           90      145         0          0         86.8
17          104      145         1.414      0         83.3
18           76      145        -1.414      0         81.2
19           90      152         0         1.414      81.2
20           90      138         0        -1.414      79.5
21           90      145         0          0         87.0
22           90      145         0          0         86.0
;
proc sort data=box;
      by duree Temperature;
   run;
```

```
proc rsreg data=box ;
model rendement= duree temperature/lackfit ;
run;

data plus;
do x1=-2 to 2 by 0.1;
 do x2= -2 to 2 by 0.1;
 duree=90+10*x1;
 temperature=145+5*x2;
 rendement=.;
 output;
 end;
 end;
 data tout;set box plus;
proc rsreg data=tout out=pred noprint;
     model rendement= duree temperature / predict;
   run;

axis2 order=(140 to 155 by 5) value=(h=2.5) minor=none label=none;
axis1 order=(70 to 110 by 20) value=(h=2.5) minor=none label=none;

 symbol1  color=black line=1 h=2;symbol2  color=black line=2 h=2;
 symbol3  color=black line=3 h=2;
 symbol4  color=black line=4 h=2;symbol5  color=black line=5 h=2;
 symbol6  color=black line=6 h=2;
 symbol7  color=black line=7 h=2;

proc gcontour data=pred;
    plot duree*Temperature=rendement
      / levels = 70 80 85 to 89 by 1 vaxis=axis1  haxis=axis2
          nolegend autolabel;
  run;
```

On obtient les résultats suivants :

```
                                The RSREG Procedure

          Codage des coefficients pour les variables indépendantes

                Facteur         Soutrait(s)    Divisé(s) par

                duree             90.000000        14.000000
                temperature      145.000000         7.000000


                  Response Surface for Variable rendement

                  Response Mean                  83.883333
                  Root MSE                        2.106039
                  R-Square                          0.8746
                  Coefficient of Variation          2.5107


                  Type I Somme                Valeur
```

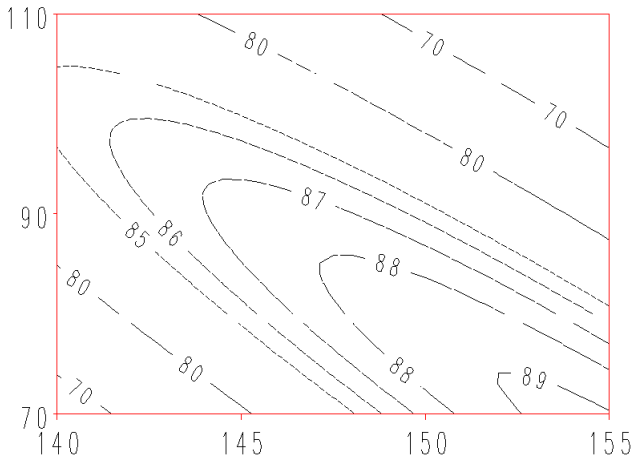


FIGURE 8.5 – Lignes de niveau du rendement d'une réaction chimique. Température en abscisse et durée en ordonnée.

Régression	DF	des carrés	R-carré	F	Pr > F
Linear	2	14.570707	0.0686	1.64	0.2698
Quadratic	2	76.031053	0.3582	8.57	0.0174
Crossproduct	1	95.062500	0.4478	21.43	0.0036
Total Model	5	185.664260	0.8746	8.37	0.0112

Résidus	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Lack of Fit	3	18.844907	6.281636	2.43	0.2429
Pure Error	3	7.767500	2.589167		
Total Error	6	26.612407	4.435401		

Paramètre	DF	Estimation	Erreur standard	test t	Pr > |t|	Paramètres estimés de des données codées
Intercept	1	-4015.393038	776.0727	-5.17	0.0021	87.357018
duree	1	17.972095	3.4118	5.27	0.0019	-0.417172
temperature	1	45.188411	9.9762	4.53	0.0040	1.852525
duree*duree	1	-0.021469	0.0084	-2.54	0.0439	-4.207895
temp*duree	1	-0.097500	0.0210	-4.63	0.0036	-9.555000
temp*temp	1	-0.124651	0.0337	-3.69	0.0102	-6.107895

Facteur	DF	Somme des carrés	Carré moyen	F	Pr > F

duree	3	124.446993	41.482331	9.35	0.0111
temperature	3	169.359835	56.453278	12.73	0.0052

The RSREG Procedure
Canonical Analysis of Response Surface Based on Coded Data

Facteur	Valeur critique Codé	Non codé
duree	-1.981055	62.265236
temperature	1.701200	156.908400

Valeur prédite au point stationnaire : 89.345996

Valeurs propres	Vecteurs propres duree	temperature
-0.286857	0.772991	-0.634417
-10.028932	0.634417	0.772991

Stationary point is a maximum.

Le test d'ajustement de la surface de réponse permet de ne pas rejeter l'hypothèse que cette réponse est un polynôme de degré 2. On observe qu'il y a des effets durée et température, que les effets quadratiques sont significatifs et que l'optimum est en dehors de la zone expérimentée, avec une durée de 62.2' et une température de 156.9.

Cependant le rendement n'est pas le seul critère pour choisir les paramètres du process. La couleur et la viscosité sont également des variables réponses importantes. L'optimisation multicritères est un problème souvent rencontré en pratique. Il n'est pas possible de maximiser du même coup ces trois variables réponses mais on s'en sort en remplaçant l'optimisation par l'instauration de contraintes pour deux d'entre elles.

On réécrit le problème sous la forme suivante : Maximiser le rendement sous les contraintes :

— $70 <$ viscosité< 80
— index de couleur > 8

On procède pour la couleur et la viscosité comme on l'a fait pour le rendement. On obtient des courbes de niveau qui sont inscrites en surimpression sur celles du rendement comme cela apparaît sur la figure 8.6. Les courbes de niveau pour la couleur et le rendement sont des droites parce que les coefficients quadratiques de ces surfaces de réponses sont statistiquement non significatifs et ont donc été mis à zéro. On obtient alors une zone admissible, qui correspond aux contraintes poées sur la viscosité et la couleur. Il ne reste plus alors qu'à choisir la solution qui appartient à la zone admissible et qui maximise le rendement.

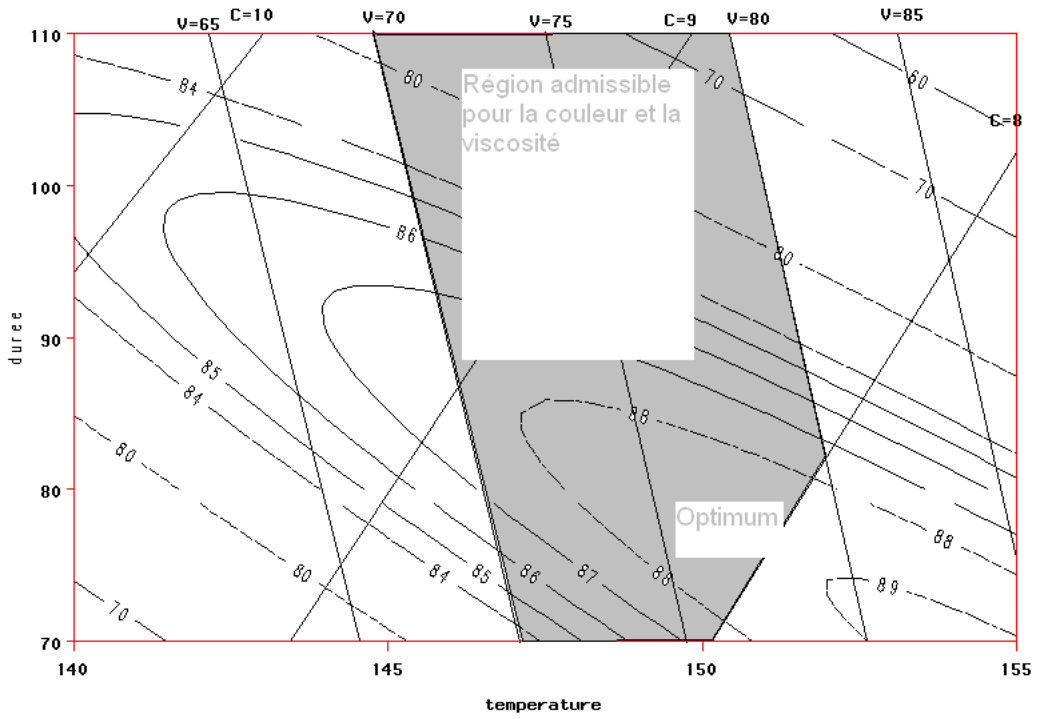


FIGURE 8.6 – Lignes de niveau du rendement d'une réaction chimique, définition de la région admissible et de la position optimale

8.11 Mélange, résistance de tissu

Données issues de [15]

8.11.1 Plan d'expériences et données

Trois constituants (polyéthylène, polystirène, polypropylène) entrent dans la fabrication de tissus. Les proportions respectives des constituants sont notées x_1, x_2, x_3. On mesure la résistance du tissu en exerçant une force croissante d'extension. La variable réponse est la résistance à l'élongation (mesurée en kg), c'est à dire la force la plus faible qui déchire le tissu. Pour chaque mélange il y a 2 ou 3 mesures de résistance. Le plan d'expériences et les résultats de la variable réponse sont donnés dans la table 8.14. Le plan d'expériences est un réseau de Scheffé de type (p,2) qui contient tous les corps purs ($x_i = 1$) et tous les mélanges 2 à 2 ($x_i = x_j = 1/2,\ i \neq j$). Ce plan permet d'estimer les coefficients du polynôme (8.5), [15] :

$$Y(x_1, x_2, x_3) = \sum \beta_i x_i + \sum_{i \neq j} \beta_{ij} x_i x_j + E \tag{8.5}$$

où $E \sim \mathcal{N}0, \sigma^2)$.

Dans le cas des mélanges on ne met pas de terme constant dans le modèle car comme la somme des composants est égale à 1, les variables x_1, x_2, x_3 sont linéairement liées. Pour la même raison, les termes x_1^2, x_2^2, x_3^2 n'apparaissent pas.

essai	x_1	x_2	x_3	y
1	1	0	0	11,12.4
2	0.5	0.5	0	15, 14.8, 16.1
3	0	1	0	8.8, 10
4	0	0.5	0.5	10, 9.7, 11.8
5	0	0	1	16.8, 16
6	0.5	0	0.5	17.7, 16.4, 16.6

TABLE 8.14 – Résistance de tissus

8.11.2 Analyse des résultats

le programme R suivant (qui utilise le package mixexp)

```
install.packages("mixexp")
 library("mixexp")
 ## Exemple des tissus dat = data.frame(
  "x1"=c(1, 1, 0.5, 0.5, 0.5, 0, 0, 0,   0,   0,   0, 0, 0.5, 0.5,0.5),
  "x2"=c(0, 0, 0.5, 0.5, 0.5, 1, 1, 0.5, 0.5, 0.5, 0, 0,  0,   0,  0 ),
  "x3"=c(0, 0,  0,   0,   0,   0,  0, 0, 0.5, 0.5, 0.5, 1, 1, 0.5, 0.5, 0.5),
  "y"=c(11,12.4,15,14.8,16.1,8.8,10,10,9.7,11.8,16.8,16,17.7,16.4,16.6)
)
polynome=lm(y~0+(x1+x2+x3)^2,dat)
summary(polynome)

MixturePlot(dat$x3,dat$x2,dat$x1,dat$y, x3lab="Fraction x3",
            x2lab="Fraction x2", x1lab="Fraction x1",
            789
            corner.labs=c("x3","x2","x1"),
            constrts=FALSE,contrs=TRUE,cols=TRUE, mod=2,n.breaks=9)
```

a donné le résultat suivant

```
Call:
lm(formula = y ~ 0 + (x1 + x2 + x3)^2, data = dat)
Residuals:
   Min     1Q Median     3Q    Max
 -0.80  -0.50  -0.30   0.65   1.30

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
x1     11.7000     0.6037  19.381 1.20e-08 ***
x2      9.4000     0.6037  15.571 8.15e-08 ***
x3     16.4000     0.6037  27.166 6.01e-10 ***
x1:x2  19.0000     2.6082   7.285 4.64e-05 ***
x1:x3  11.4000     2.6082   4.371  0.00180 **
x2:x3  -9.6000     2.6082  -3.681  0.00507 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .

Residual standard error: 0.8537 on 9 degrees of freedom
Multiple R-squared:  0.9977,    Adjusted R-squared:  0.9962
F-statistic: 658.1 on 6 and 9 DF,  p-value: 2.271e-11
```

et la figure 8.7. Tous les termes du modèle sont statistiquement significatifs. La valeur du R^2 est élevée. Cependant pour tester le bon ajustement du modèle il est utile de séparer l'erreur de modèle de l'erreur due à la variabilité pure. On ne peut pas le faire dans ce cas car, par construction, l'erreur de modèle est nulle avec ce plan d'expériences. En effet il y a 6 points expérimentaux et 6 paramètres. L'écart-type résiduel estimé (0.8537) mesure donc seulement la variabilité du matériel expérimental. Il aurait fallu ajouter au moins une expérience en un autre point pour tester l'erreur de modèle.

Finalement le choix optimal consiste à choisir un mélange avec entre 20 et 40 % de polyéthylè-ne et le complément de polypropylène, le polystirène n'entrant pas dans la composition optimale.

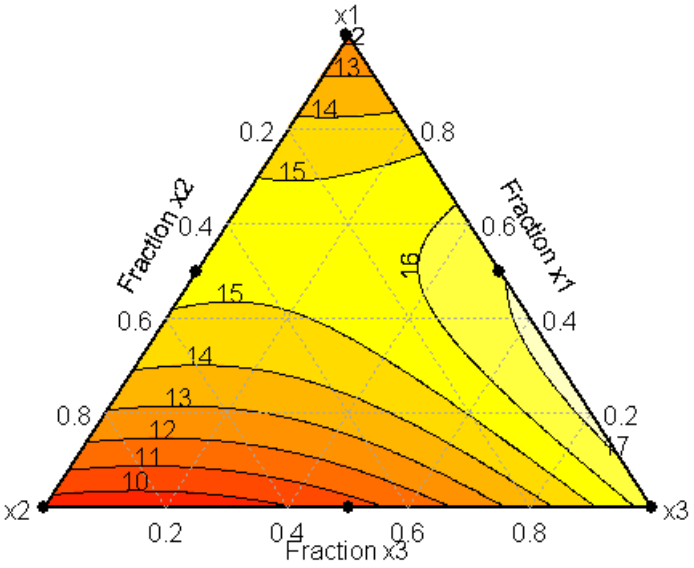


FIGURE 8.7 – Lignes de niveau de la résistance de tissus à l'étirement

8.12 programmes SAS et packages R (fin 2013)

Il existe des programmes SAS adaptés à l'analyse de plans d'expériences types (blocs, BIE, carrés latins, split-plots...) :

`http://statistics.ats.ucla.edu/stat/sas/library/SASExpDes_os.htm#IBD)`

Un survol (2013) des packages R pour les plans d'expériences est dans

`http://cran.r-project.org/web/views/ExperimentalDesign.html`

Chapitre 9

Annexe1 : espaces euclidiens de dimension finie

9.1 Introduction

L'objectif de cette annexe est de rassembler les propriétés des espaces euclidiens de dimension finie utiles pour la lecture de cet ouvrage. La plupart des propriétés et propositions sont démontrées. On suppose connues les définitions et propriétés des espaces vectoriels, des applications linéaires ainsi que le calcul matriciel.

On travaille sur des sous espaces vectoriels de dimension finie de $\mathbb{R}^n$ muni du produit scalaire usuel qui lui confère une structure d'espace euclidien. Rappelons que

- pour tout $u = (x_1, ..., x_n)$ et $v = (y_1, ..., y_n)$ de $\mathbb{R}^n$, $\langle u, v\rangle = \sum_{i=1}^{n} x_i y_i$,
- $\|u\| = \sqrt{\langle u, u\rangle}$
- écriture matricielle : soit X le vecteur des coordonnées de u et Y celui de v dans la base canonique de $\mathbb{R}^n$, $\langle u, v\rangle = X'Y = Y'X$ où X' est la transposée du vecteur X.
- On dit que u est orthogonal à v si le produit scalaire $\langle u, v\rangle$ est nul.

9.2 Sous espaces d'un espace vectoriel euclidien

9.2.1 Sous espaces orthogonaux d'un espace euclidien

Définition 9.2.1. *Soit E un espace vectoriel euclidien (pour notre étude $E = \mathbb{R}^n$), F et G deux sous espaces vectoriels de E. On dit que F et G sont orthogonaux si*

$$\forall u \in F, \ \ \forall v \in G \quad \langle u, v\rangle = 0$$

9.2.2 Supplémentaire orthogonal d'un sous espace vectoriel F

Définition 9.2.2. *Soit F un sous espace vectoriel de E, le supplémentaire orthogonal de F, noté $F^{\perp}$, est l'ensemble des éléments de E orthogonaux à tous les éléments de F.*

- E est somme directe de F et $F^{\perp}$, c'est-à-dire que tout élément de E s'écrit de façon unique comme somme de 2 éléments appartenant l'un à F et l'autre à $F^{\perp}$. On note $E = F \oplus F^{\perp}$.

— $(F^\perp)^\perp = F$.

Le théorème de Pythagore se généralise à une somme directe de plusieurs sous espaces orthogonaux :

Théorème 9.2.1. *Si E est somme directe de sous espaces deux à deux orthogonaux F_1,..., F_p, et si $x = \sum_{i=1}^{p} x_i$ est la décomposition sur les F_i d'un élément quelconque x de E, on a $\|x\|^2 = \sum_{i=1}^{p} \|x_i\|^2$.*

9.3 Base orthonormée de $\mathbb{R}^n$

On dit que des vecteurs $(e_1,...e_n)$ de l'espace euclidien $E = \mathbb{R}^n$ forment une base orthonormée s'ils sont deux à deux orthogonaux, et si leur norme est égale à 1.
Ce système est automatiquement libre et générateur. Tout système orthonormé d'un espace vectoriel E peut être complété en une base orthonormée, on utilise pour cela le procédé d'orthonormalisation de Schmidt.

9.3.1 Procédé d'orthonormalisation de Schmidt (ou Gram-Schmidt)

Ce procédé est un algorithme qui permet de construire une base orthonormée de E à partir d'une base quelconque.

Proposition 9.3.1. *Soit $(v_1, ...v_n)$ une base quelconque de E. Il existe une base orthonormée $(e_1, ...e_n)$ unique de E, vérifiant les conditions suivantes :*

— $\forall p \in (1, n), \langle e_p, v_p\rangle > 0$

— *Pour tout entier $p \in (1, n)$, les espaces vectoriels engendrés par $(e_1, ..e_p)$ et par $(v_1, ..., v_p)$ sont identiques.*

Démonstration. Nous construisons par récurrence sur p, des vecteurs $e_1, ..., e_n$ tels que $(e_1, ...e_p)$ soit un système orthonormé pour tout p, que $\langle e_k, v_k\rangle > 0$ pour $k \leq p$ et que les espaces engendrés par $(e_1, ..e_p)$ et $(v_1, ..., v_p)$ soient identiques.

Soit $e_1 = \frac{v_1}{\|v_1\|}$ et

$$w_{p+1} = v_{p+1} - \sum_{i=1}^{p} \langle v_{p+1}, e_i\rangle e_i.$$

w_{p+1} possède les propriétés suivantes

— $\langle w_{p+1}, e_i\rangle = 0,\ i \in (1, n)$

— $\|w_{p+1}\| > 0$

La première propriété est évidente par construction de w_{p+1}. On ne peut pas avoir $\|w_{p+1}\| = 0$, car alors $v_{p+1} = \sum_{i=1}^{p} \langle v_{p+1}, e_i\rangle e_i$, ce qui impliquerait que v_{p+1} appartient à l'espace engendré par $(v_1, ...v_p)$, ce qui est en contradiction avec le fait que $(v_1, ...v_n)$ soit une base de E.

Le vecteur

$$e_{p+1} = \frac{w_{p+1}}{\|w_{p+1}\|}$$

vérifie donc les conditions demandées. $-e_{p+1}$ vérifie les mêmes propriétés que e_{p+1}. On choisit celui des 2 dont le produit scalaire avec v_{p+1} est positif.

Enfin s'il existe une base orthonomée $(f_1, ...f_n)$ ayant les mêmes propriétés que $(e_1, ...e_n)$, alors

$$\forall p \in (1, n-1), f_{p+1} \perp Vect(f_1, ...f_p) = Vect(v_1, ...v_p) = Vect(e_1, ...e_p)$$

et

$$f_{p+1} \in Vect(v_1, ...v_{p+1}) = Vect(e_1, ...e_{p+1})$$

implique que

$$f_{p+1} = \sum_{i=1}^{p+1} \langle f_{p+1}, e_i \rangle e_i = a e_{p+1},$$

avec $|a| > 0$. Comme f_{p+1} et e_{p+1} sont unitaires, $a = 1$ et $f_{p+1} = e_{p+1}$, ce qui prouve l'unicité de $(e_1, ...e_p)$. □

9.3.2 Matrice orthogonale

Une matrice carrée, de format (n, n) à coefficients dans $\mathbb{R}$ est orthogonale, si c'est la matrice de passage d'une base orthonormée à une autre base orthonormée.
Soit M une matrice orthogonale, alors $M'M = MM' = I_n$ où M' est la transposée de M.

Propriété 9.3.1. *Le procédé d'orthonormalisation de Schmidt s'exprime matriciellement de la façon suivante : toute matrice inversible M peut se mettre sous la forme $M = AT$, où A est une matrice orthogonale et où T est une matrice triangulaire supérieure à coefficients diagonaux strictement positifs.*

9.4 Projecteur orthogonal

9.4.1 Projecteur

Il existe deux définitions équivalentes :

Définition 9.4.1. *Soit E_1 et E_2 2 sous espaces supplémentaires de $\mathbb{R}^n$, ($E_1 \oplus E_2 = \mathbb{R}^n$). Soit $u \in \mathbb{R}^n$. La décomposition $u = u_1 + u_2$ avec $u_1 \in E_1$ et $u_2 \in E_2$ est unique. On appelle projecteur sur E_1 parallèlement à E_2 , l'application linéaire p de $\mathbb{R}^n$ dans $\mathbb{R}^n$, telle que $\forall u \in \mathbb{R}^n$, $p(u) = u_1$.*

On parle parfois de projection bien que le projecteur soit une application linéaire et non affine.

On a alors $Ker(p) = E_2$ et $Im(p) = E_1$ et donc $Im(p) \oplus Ker(p) = \mathbb{R}^n$.

La matrice P de p dans une base adaptée à la décomposition $E_1 \oplus E_2 = \mathbb{R}^n$ est de la forme

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

avec un bloc correspondant à l'identité de E_1 et des 0 ailleurs.
On en déduit que la trace d'un projecteur p est égale à la dimension de $Im(p)$ soit le rang de p. (la trace d'une application linéaire est égale à la trace de sa matrice dans une base quelconque, c'est-à-dire la somme de ses termes diagonaux, somme qui est invariante par changement de base).

Définition 9.4.2. *On appelle projecteur, une application linéaire p de $\mathbb{R}^n$ dans $\mathbb{R}^n$ telle que $p \circ p = p$.*

Si P est la matrice de p dans une base de $\mathbb{R}^n$, ceci est équivalent à dire que $P^2 = P$.

Proposition 9.4.1. *Les définitions 9.4.1 et 9.4.2 sont équivalentes.*

Démonstration. La démonstration de l'implication "définition 9.4.1 $\Rightarrow$ définition 9.4.2" est immédiate. La démonstration de l'implication "définition 9.4.2 $\Rightarrow$ définition 9.4.1" est la suivante :
Pour toute application p vérifiant $p \circ p = p$, on a $Im(p) \oplus Ker(p) = \mathbb{R}^n$.
En effet $\forall v \in Im(p) \cap Ker(p), \exists u \in \mathbb{R}^n, p(u) = v$ car $v \in Im(p)$ et $p(v) = 0$ car $v \in Ker(p)$.
Par suite, $v = p(u) = p^2(u) = p(v) = 0 \Rightarrow v = 0 \Rightarrow Im(p) \cap Ker(p) = 0$. La somme est donc directe et vaut $\mathbb{R}^n$ par un argument de dimension.
De plus la restriction de p à $Im(p)$ est l'identité car $p \circ p = p$. □

9.4.2 Projecteur orthogonal

Définition 9.4.3. *Un projecteur est projecteur orthogonal, si $Im(p) \perp Ker(p)$.*

Proposition 9.4.2. *Un projecteur, de matrice P dans une base orthonormée de $\mathbb{R}^n$, est un projecteur orthogonal si et seulement si sa matrice, vérifie $P^2 = P$ et $P' = P$*

Démonstration. **Projecteur orthogonal $\Rightarrow P^2 = P$ et $P' = P$**

$Im(p)$ et $Ker(p)$ sont des sous espaces orthogonaux. Donc il existe une base adaptée orthonormée $(e_1, ...e_n)$ de $\mathbb{R}^n$, telle que $Vect(e_1, ..e_k) = Im(p)$ et $Vect(e_{k+1}, ..e_n) = Ker(p)$ avec $k = rang(p)$ et $Vect(e_1...e_s)$ est l'espace vectoriel engendré par $(e_1...e_s)$.

Dans cette base la matrice P de p est

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

avec un bloc correspondant à l'identité de $Im(p)$ et des 0 ailleurs, elle vérifie $P^2 = P$, et $P' = P$.

$P^2 = P$ et $P' = P \Rightarrow$ Projecteur orthogonal

$P^2 = P$ donc c'est la matrice d'un projecteur. Vérifions que $Im(P) \perp Ker(P)$.

$\forall X \in Im(P), \exists Z$ vecteur de $\mathbb{R}^n$, tel que $X = PZ$.

$\forall Y \in Ker(P)$, $X'Y = (PZ)'Y = Z'P'Y$ d'après les propriétés de la transposition. $P = P' \Rightarrow X'Y = Z'PY = 0$ car $PY = 0$ puisque $Y \in Ker(P)$. Donc $Im(P)$ et $Ker(P)$ sont bien des sous espaces orthogonaux, et P est une projection orthogonale. □

Proposition 9.4.3. *La projection orthogonale sur $Im(p)$ minimise les distances : Soient u et v deux vecteurs de $\mathbb{R}^n$, u fixé et $v \in Im(p)$, $\|u - v\|$ est minimale pour $v = p(u)$.*

Démonstration. $\forall v \in Im(p)$ on a l'égalité :

$$\|u - v\|^2 = \|u - p(u)\|^2 + \|p(u) - v\|^2$$

d'après le théorème de Pythagore car $u - p(u) \in Ker(p)$ et $p(u) - v \in Im(p)$ et ces deux espaces sont orthogonaux, donc $\|u - v\|$ est bien minimale pour $v = p(u)$. □

Chapitre 10

Annexe2 : lois normale multidimensionnelle, χ^2, Student, Fisher

10.1 Vecteurs aléatoires

Notations : On note A' la transposée d'une matrice A à valeurs dans $\mathbb{R}$. Un élément de $\mathbb{R}^p$, c'est à dire une matrice à une ligne et p colonnes est noté $(x_1, x_2, ...x_p)$ les x_i étant des réels. Un vecteur de $\mathbb{R}^p$, c'est-à-dire une matrice à p lignes et une colonne, est noté $(x_1, x_2, ...x_p)'$.

Définition 10.1.1. *X est un vecteur aléatoire de $\mathbb{R}^p$, où $X = (X_1, X_2, ...X_p)'$, si ses composantes X_i sont des variables aléatoires à valeurs dans $\mathbb{R}$.*

Son espérance est un vecteur de $\mathbb{R}^p$, $E(X) = (\mu_1, ..., \mu_p)'$ avec $\mu_i = E(X_i), i = 1, p$
Sa matrice de variance covariance Σ_X, est définie par :

$$\Sigma_X = [E(X - E(X))(X - E(X))'],$$

notée aussi $\Sigma_X = V(X)$. Cette matrice contient les variances sur la diagonale et les covariances en dehors de la diagonale :

$$\Sigma_X = \begin{pmatrix} V(X_1) & ... & cov(X_1, X_j) & ... & cov(X_1, X_p) \\ ... & ... & ... & ... & ... \\ cov(X_1, X_i) & ... & cov(X_i, X_j) & ... & cov(X_i, X_p) \\ ... & ... & ... & ... & ... \\ cov(X_1, X_p) & ... & cov(X_p, X_j) & ... & V(X_p) \end{pmatrix}$$

Les propriétés de linéarité de l'espérance et de bilinéarité de la covariance se conservent :

Propriété 10.1.1. *Soient X et Y deux vecteurs aléatoires, A et B deux matrices $p \times p$ à coefficients réels non aléatoires, alors $E(AX) = AE(X)$, $cov(AX, BY) = Acov(X, Y)B'$, et $V(AX) = \Sigma_{AX} = AV(X)A' = A\Sigma_X A'$*

Cas particulier : Si a est un vecteur de $\mathbb{R}^p$, $V(a'X) = a'V(X)a = a'\Sigma_X a$

Proposition 10.1.1. *Soit X un vecteur aléatoire de densité f_X, soit A une matrice $p \times p$, inversible, alors $Y = AX$ est un vecteur aléatoire de densité f_Y, avec pour tout vecteur y de $\mathbb{R}^p$, $f_Y(y) = det(A)^{-1} f_X(A^{-1}y)$*

Cette proposition est une application des changements de variables dans les intégrales.

10.2 Lois gaussiennes multidimensionnelles

10.2.1 Loi normale unidimensionnelle

Définition 10.2.1. *On dit qu'une variable aléatoire à valeurs réelles X, suit une loi normale d'espérance μ et de variance σ^2, si sa densité f est définie pour tout x réel par* $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

on note : $X \backsim \mathcal{N}(\mu, \sigma^2)$.

10.2.2 Cas particulier d'un échantillon gaussien

Soit X un vecteur aléatoire de $\mathbb{R}^p$, où $X = (X_1, X_2, ...X_p)'$, dont les composantes X_i sont indépendantes et de même loi (i.i.d.) , $X_i \backsim \mathcal{N}(\mu; \sigma^2)$, alors la densité f de X est définie pour tout p-uplet $(x_1, ..., x_p)$ de $\mathbb{R}^p$ par :

$$f(x_1, ..., x_p) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^p \prod_{i=1}^{p} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}$$

Soit :

$$f(x_1, ..., x_p) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^p e^{-\frac{1}{2}\sum_{i=1}^{p}(\frac{x_i-\mu}{\sigma})^2}$$

Dans ce cas $E(X) = (\mu, ..., \mu)'$ et $V(X) = \sigma^2 I_p$ où I_p est l'identité de $\mathbb{R}^p$.

10.2.3 Loi gaussienne multidimensionnelle, vecteur gaussien

Définition 10.2.2. *Soit X un vecteur aléatoire de $\mathbb{R}^p$, d'espérance μ, vecteur de $\mathbb{R}^p$, et de matrice de variance covariance Σ. On dit que X suit une loi gaussienne multidimensionnelle, si sa densité f est définie pour tout p-uplet $(x_1, ..., x_p)$ de $\mathbb{R}^p$ par :*

$$f(x_1, ..., x_p) = \frac{1}{[2\pi det(\Sigma)]^{p/2}} \, exp\left\{-\frac{1}{2}[(x-\mu)'\Sigma^{-1}(x-\mu)]\right\}$$

où $x = (x_1, ..., x_p)'$.

On note : $X \backsim \mathcal{N}_p(\mu, \Sigma)$. Cette loi est entièrement caractérisée par son espérance et sa variance. X est appellé **vecteur gaussien**

L'échantillon gaussien de la section 10.2.2, p. 307 est un exemple de vecteur gaussien, avec les composantes de μ identiques et $\Sigma = \sigma^2 I_p$.

Propriété 10.2.1. *Soit X un vecteur gaussien $X \backsim \mathcal{N}_p(\mu;\Sigma)$ et $X = (X_1, X_2)$ une décomposition de X en deux vecteurs aléatoires à respectivement p_1 et p_2 composantes, avec $p_1 + p_2 = p$.*
Soit $\mu_1 = E(X_1)$ et $\mu_2 = E(X_2)$, $\Sigma_{11} = V(X_1)$, $\Sigma_{22} = V(X_2)$, $\Sigma_{12} = cov(X_1, X_2)$, de telle sorte que

$$\Sigma_X = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

on a les propriétés suivantes :

- *si A est la matrice (p,p) d'un endomorphisme de $\mathbb{R}^p$, alors $AX \backsim \mathcal{N}_p(A\mu, A\Sigma A')$. En particulier, toute combinaison linéaire des composantes de X suit une loi normale.*
- *X_1 et X_2 suivent des lois normales $X_1 \backsim \mathcal{N}_{p_1}(\mu_1, \Sigma_{11})$ et $X_2 \backsim \mathcal{N}_{p_2}(\mu_2, \Sigma_{22})$*
- *X_1 et X_2 sont indépendantes si et seulement si elles sont non corrélées, c'est-à-dire si la matrice $\Sigma_{12} = cov(X_1, X_2)$ est la matrice nulle. En particulier pour tout $(i,j) \in (1,p)^2$ tel que $i \neq j$, X_i et X_j sont indépendantes si et seulement si $cov(X_i, X_j) = 0$.*

Démonstration. La première propriété n'est pas démontrée ici. Si A est inversible on peut la démontrer en utilisant la proposition 10.1.1, p. 307.

La deuxième propriété est un cas particulier de la première avec $A = \begin{bmatrix} I_{p_1} & 0 \\ 0 & 0 \end{bmatrix}$.

La troisième propriété est immédiate car si $\Sigma = \begin{bmatrix} \Sigma_{X_1X_1} & 0 \\ 0 & \Sigma_{X_2X_2} \end{bmatrix}$, la densité de X se factorise en un produit de la densité de X_1 et de la densité de X_2. □

Cas particulier des combinaisons linéaires

Proposition 10.2.1. *Soit $a = (a_1, ..., a_p)'$ un vecteur quelconque de $\mathbb{R}^p$, $a'X = \sum a_i X_i$ est une combinaison linéaire des coordonnées de X et $a'X \backsim \mathcal{N}(a'\mu; a\Sigma a')$.*

Réciproquement, si toute combinaison linéaire de X est gaussienne alors X est un vecteur gaussien.

La première partie de la proposition est un cas particulier de la propriété 10.2.1, p. 308. La démonstration de la réciproque n'est pas faite ici. Elle utilise la fonction caractéristique du vecteur X.

Les composantes d'un vecteur gaussien sont des variables aléatoires gaussiennes, mais la réciproque est fausse. En voici un contre-exemple :

Soit $X \backsim \mathcal{N}(0,1)$, et ε une variable de Bernoulli de paramètre $\frac{1}{2}$, indépendante de X. Alors $X_1 = X$ et $X_2 = (2\varepsilon - 1)X$ sont des variables gaussiennes mais $(X_1, X_2)'$ n'est pas un vecteur gaussien. Dans cet exemple, $cov(X_1, X_2) = 0$ mais X_1 et X_2 ne sont pas indépendantes.

10.2.4 Loi conditionnelle

Propriété 10.2.2. *Soit $(X', Y')'$ un vecteur gaussien de loi*

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0_X \\ 0_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \right),$$

où 0_X (resp. 0_Y) désigne le vecteur nul de même dimension que X (resp Y) et $\Sigma_{YX} = \Sigma'_{XY}$. La loi conditionnelle de Y sachant X est la loi normale

$$Y|X \sim \mathcal{N}\left(\Sigma_{YX}\Sigma_{XX}^{-1}X, \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right).$$

Démonstration. La démonstration de cette propriété repose notamment sur l'identité suivante portant sur l'inversion par bloc d'une matrice :

$$S = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad \Rightarrow \quad S^{-1} = \begin{bmatrix} A^{-1}(I + BU^{-1}CA^{-1}) & -A^{-1}BU^{-1} \\ -U^{-1}CA^{-1} & U^{-1} \end{bmatrix} \tag{10.1}$$

où $U = D - CA^{-1}B$. On montre facilement cette identité en vérifiant que les produits SS^{-1} et $S^{-1}S$ sont égaux à la matrice identité.

Notons Ω_{XX} l'inverse de la matrice de variance du vecteur $(X', Y')'$:

$$\Omega = \begin{bmatrix} \Omega_{XX} & \Omega_{XY} \\ \Omega_{YX} & \Omega_{YY} \end{bmatrix} := \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}^{-1}$$

en notant bien que $\Omega_{XX} \neq \Sigma_{XX}^{-1}$. Notons de plus $f_{XY}(x, y)$ la densité du vecteur (X', Y') et $f_X(x)$ celle du vecteur X. La densité conditionnelle $f_{Y|X}(y; x)$ de Y sachant X s'écrit

$$\begin{aligned} f_{Y|X}(y;x) &= f_{XY}(x,y)/f_X(x) \\ &\propto \exp\left(-\frac{1}{2}[x'y']\ \Omega\ [x'y']'\right) \Big/ \exp\left(-\frac{1}{2}x'\ \Sigma_{XX}\ x\right) \\ &= \exp\left\{-\frac{1}{2}\left(x'\Omega_{XX}x + 2x'\Omega_{XY}y + y'\Omega_{YY}y - x'\Sigma_{XX}^{-1}x\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(y + \Omega_{YY}^{-1}\Omega_{YX}x\right)'\Omega_{YY}\left(y + \Omega_{YY}^{-1}\Omega_{YX}x\right)\right\} \end{aligned}$$

où la première relation de proportionnalité permet d'omettre les constantes de normalisation des deux densités et la seconde permet d'omettre des termes ne dépendant que de x (et pas de y). On reconnaît alors le terme situé dans l'exponentielle de la densité d'un loi normale

$$\mathcal{N}\left(-\Omega_{YY}^{-1}\Omega_{YX}x, \Omega_{YY}^{-1}\right). \tag{10.2}$$

Le reste de la démonstration repose sur l'identité donnée à l'équation 10.1, p. 309. Pour la variance, on reconnaît d'abord

$$\Omega_{YY}^{-1} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

Pour l'espérance, on a $\Omega_{YX} = -\left(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right)^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}$ donc

$$-\Omega_{YY}^{-1}\Omega_{YX} = \left(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right)\left(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right)^{-1}\Sigma_{YX}\Sigma_{XX}^{-1},$$

ce qui termine la démonstration. □

Corollaire 10.2.1. *Si le vecteur $(X', Y')'$ n'est pas centré :*

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right),$$

on a

$$Y|X \sim \mathcal{N}\left(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right).$$

Démonstration. Il suffit d'appliquer la propriété précédente aux vecteurs $U = X - \mu_X$ et $V = Y - \mu_Y$. □

On observe ici une propriété particulière de la loi normale : l'espérance conditionnelle $\mathbb{E}(Y|X)$ dépend de X, mais pas la variance conditionnelle $\mathbb{V}(Y|X)$.

10.2.5 Loi du χ^2

Définition 10.2.3. *La loi de $Z = \sum_{i=1}^{p} X_i^2$ où $X_i \backsim \mathcal{N}(0,1)$ et les X_i sont indépendantes est la loi du χ^2 à p degrés de liberté.*

On note $Z \backsim \chi_p^2$.

Propriété 10.2.3. *La densité f_Z est celle d'une loi Gamma de paramètres $p/2$ et $1/2$, définie pour $x > 0$ par*

$$f_Z(x) = K_p \, x^{\frac{p-2}{2}} e^{-\frac{x}{2}}$$

où K_p est une constante. $\mathbb{E}(Z) = p$ et $\mathbb{V}(Z) = 2p$.

Loi de la variance empirique d'un échantillon gaussien

Soit $X = (X_1, X_2, ...X_n)'$, le vecteur des résultats d'un n-échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Alors $\overline{X} \backsim \mathcal{N}(\mu, \sigma^2/n)$, $\frac{(n-1)S^2}{\sigma^2} \backsim \chi_{n-1}^2$, et $\overline{X}$ et S^2 sont des variables aléatoires indépendantes.

Ces résultats sont utiles pour construire la loi de Student 10.2.6, p. 310.

Lien avec les vecteurs gaussiens

Soit $X \backsim \mathcal{N}_p(\mu, \Sigma)$ un vecteur gaussien avec Σ inversible. Alors :

$$[(X - \mu)'\Sigma^{-1}(X - \mu)] \backsim \chi_p^2$$

Cette propriété est à la base des tests de Wald.

10.2.6 Loi de Student

Définition 10.2.4. *Soit $X \backsim \mathcal{N}(0,1)$ et $Z \backsim \chi_p^2$, X et Z étant des variables aléatoires indépendantes. Le quotient $T = \frac{X}{\sqrt{Z/p}}$ suit une loi de Student à p degrés de libertés.*

On note $T \backsim T_p$

Propriété 10.2.4. *Sa densité f_T est définie pour tout réel x par :*

$$f_T(x) = H_p \, \frac{1}{(1 + \frac{x^2}{p})^{\frac{p+1}{2}}}$$

où H_p est une constante telle que l'intégrale de la densité sur $\mathbb{R}$ égale 1.

$\mathbb{E}(T) = 0$ si $p > 1$, et $\mathbb{V}(T) = \frac{p}{p-2}$ si $p > 2$.

Les quantiles de la loi de Student sont tabulés. Pour n grand, la distribution de la loi de Student s'approche de celle de la loi normale centrée et réduite. En pratique, on utilise la loi normale à la place de la loi de Student pour $n > 30$, pour des quantiles pas trop extrêmes.

Cas d'un échantillon gaussien

Soit $X = (X_1, X_2, ...X_n)'$ vecteur des résultats d'un n-échantillon d'une loi $\mathcal{N}(\mu, \sigma^2)$, Soit $\overline{X}$ la moyenne des X_i, et $S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$ la variance empirique de l'échantillon. Alors les résultats précédents permettent d'affirmer :

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \backsim T_{n-1}$$

Car $\overline{X}$ et S^2 sont des variables aléatoires indépendantes.

10.2.7 Loi de Fisher

Définition 10.2.5. *Soit $Z_1 \backsim \chi^2_p$,et $Z_2 \backsim \chi^2_q$, Z_1 et Z_2 étant des variables aléatoires indépendantes. Le quotient $F = \frac{Z_1/p}{Z_2/q}$ suit une loi de Fisher à p et q degrés de liberté*

On note $F \backsim F_{p,q}$. On dit que p est le degré de liberté du numérateur et q celui du dénominateur. On a $\frac{1}{F} \backsim F_{q,p}$.

Propriété 10.2.5. *Sa densité f_F est définie pour tout réel x par :*

$$f_F(x) = H_{p,q} \frac{x^{(p-2)/2}}{(1 + \frac{p}{q}x)^{\frac{p+q}{2}}}$$

où $H_{p,q}$ est une constante telle que l'intégrale de la densité sur $\mathbb{R}^+$ égale 1. $\mathbb{E}(F) = \frac{q}{q-2}$ si $q > 2$

$\mathbb{E}(F)$ ne dépend pas de p car les lois du χ^2 de Z_1 et Z_2 ont été normalisées en les divisant par leur degré de liberté, afin de ramener leur espérance à 1.

La distribution de la loi de Fisher est tabulée à partir de ses quantiles. Notons $F_{p,q,\alpha}$, le α-quantile de $F_{p,q}$, alors $F_{p,q,1-\alpha} = \frac{1}{F_{q,p,1-\alpha}}$.

10.2.8 Loi du χ^2 décentrée

Définition 10.2.6. *Soit $(X_i \backsim \mathcal{N}(\mu_i, 1), i = 1, p)$, p variables aléatoires gaussiennes indépendantes et $Z = \sum_{i=1}^p X_i^2$. Z suit une loi du χ^2 décentrée à p degrés de liberté, et de paramètre de non centralité $\lambda^2 = \sum_{i=1}^p \mu_i^2$.*

On note $Z \backsim \chi^2_{p,\lambda^2}$.
Quand u et p sont fixés, $P(Z < u)$ est une fonction croissante de λ. Cette loi est utilisée pour calculer la fonction puissance d'un test du χ^2.

Si, au numérateur d'une loi de Fisher $F_{p,q}$, la loi du χ^2 à p degrés de liberté est remplacée par une loi du χ^2 décentrée à p degrés de liberté de paramètre de non centralité

λ^2, la loi du quotient devient une loi de Fisher décentrée F_{p,q,λ^2}. Cette distribution sert à faire des calculs de puissance pour les tests de Fisher.

Il existe aussi une loi de Student décentrée.

Chapitre 11

Annexe3 : loi des estimateurs

11.1 Théorème de Cochran

Théorème 11.1.1. *On considère n variables aléatoires, $Y_1, Y_2, ...Y_n$, indépendantes, de même loi normale $\mathcal{N}(0,1)$. Soit $Y = (Y_1, Y_2, ...Y_n)$ et $\|Y\|^2 = \sum_i Y_i^2$. On sait que $\|Y\|^2 \sim \chi^2_n$. Soit L un sous espace de $\mathbb{R}^n$, de dimension r. On note $P_L(Y)$ le projeté orthogonal de Y sur L. On a la décomposition orthogonale suivante :*

$$Y = P_L(Y) + (Y - P_L(Y)).$$

Alors les variables aléatoires $\|P_L(Y)\|^2$ et $\|Y - P_L(Y)\|^2$ sont indépendantes et de plus $\|P_L(Y)\|^2 \sim \chi^2_r$ et $\|Y - P_L(Y)\|^2 \sim \chi^2_{n-r}$.

Ce théorème s'étend immédiatement à un nombre quelconque de sous espaces orthogonaux de $\mathbb{R}^n$.

La démonstration utilise les 2 lemmes suivants :

Lemme 11.1.2. *Soit $u \in \mathbb{R}^n$ un vecteur unitaire et Y un vecteur gaussien de $\mathbb{R}^n$. On note $P_u(Y)$ le projeté orthogonal de Y sur la droite vectorielle portée par u. Ce vecteur aléatoire est colinéaire à u, et on note $c_u(Y)$ le coefficient de linéarité : $P_u(Y) = c_u(Y)u$, et $c_u(Y) = \langle Y, u\rangle = \sum_{i=1}^n u_i Y_i$. On a alors $c_u(Y) \sim \mathcal{N}(0,1)$*

Démonstration. $c_u(Y) = u'Y$ suit une loi normale comme combinaison linéaire des composantes de Y grâce à la propriété 10.2.1, p. 308.

De plus $\mathbb{E}(c_u(Y)) = u'\mu = 0$ et $\mathbb{V}(c_u(Y)) = u'I_n u = 1$. □

Lemme 11.1.3.

$$\mathbb{C}ov(c_u(Y), c_v(Y)) = 0 \Leftrightarrow \langle u, v\rangle = 0$$

En utilisant la propriété 10.1.1, p. 306 on a $\mathbb{C}\text{ov}(c_u(Y), c_v(Y)) = \mathbb{C}\text{ov}(u'Y, v'Y) = u'I_n v = u'v$ □

Démonstration. **Première étape**
On construit une base orthonormée $(l_1, ..., l_r)$ du sous-espace L (ceci est toujours possible grâce au procédé d'orthonormalisation de Gram-Schmidt défini par la proposition 9.3.1, p. 302). Le projeté orthogonal de Y sur L est la somme des projetés orthogonaux sur les axes l_i :

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

Où $c_{l_i}(Y) \sim N(0,1)$, $i = 1...n$
Les l_i formant un système orthonormé, les $c_{l_i}(Y)$ sont non corrélés grâce au lemme 11.1.3, p. 313, donc indépendants grâce à la normalité du vecteur aléatoire Y et à la propriété 10.2.1, p. 308. Par conséquent :

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

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

11.2 Loi des sommes de carrés du modèle linéaire

11.2.1 Décompositions de $\mathbb{R}^n$ associées au modèle linéaire

On rappelle le modèle utilisé et les notations :

$$Y = X\theta + E$$

avec $E \sim \mathcal{N}(0, \sigma^2 I_n)$.

$X = (X_0, ...X_s)$. X_0 est la colonne de 1 correspondant au terme constant. Soit $L_0 = vect(X_0)$, l'espace engendré par la colonne X_0. Soit $L = vect(X)$, l'espace engendré par les colonnes de X et L^* le sous espace de L orthogonal à L_0.

On a $dim L = p + 1$, $p = rg(X) - 1$, et $dim L^* = p$.

On a les décompositions

$$\begin{aligned} \mathbb{R}^n &= L_0 \oplus L^* \oplus L^\perp \\ L &= L_0 \oplus L^* \\ \mathbb{R}^n &= L \oplus L^\perp \end{aligned}$$

Par projection orthogonale de Y sur L_0, on obtient le vecteur $P_{L_0}(Y) = \overline{Y} X_0$ de L_0.
Par projection orthogonale de Y sur L, on obtient le vecteur $P_L(Y) = \widehat{Y}$ de L,
et par projection orthogonale sur $L^\perp$, on obtient (I-P)(Y)=$Y - \widehat{Y} \in L^\perp$
La décomposition de $Y \in \mathbb{R}^n$ sur $L_0 \oplus L^* \oplus L^\perp$ s'écrit

$$Y = \overline{Y} X_0 + (\widehat{Y} - \overline{Y} X_0) + (Y - \widehat{Y})$$

On a $E = Y - X\theta$ et $\widehat{E} = Y - \widehat{Y}$.

11.2.2 Loi de la somme des carrés résiduelle

La somme des carrés résiduelle est $SCR = \left\|\widehat{E}\right\|^2$.

Proposition 11.2.1. $\frac{SCR}{\sigma^2} \sim \chi^2_{n-p}$

Démonstration. $\frac{E}{\sigma} \sim \mathcal{N}_n(0, I_n)$. La décomposition de $\frac{E}{\sigma} \in \mathbb{R}^n$ sur $L \oplus L^\perp$ s'écrit

$$\frac{E}{\sigma} = \frac{E - \widehat{E}}{\sigma} + \frac{\widehat{E}}{\sigma}$$

Le théorème 11.1, p. 313 implique que $\frac{\left\|\widehat{E}\right\|^2}{\sigma^2} \sim \chi^2_{n-p}$. On a aussi que $\frac{\left\|E-\widehat{E}\right\|^2}{\sigma^2} \sim \chi^2_p$, mais ce résultat reste théorique car on ne connait pas E. □

11.2.3 Loi de la somme des carrés du modèle sous $H_0 = (\theta_1 = ... = \theta_s = 0)$

L'hypothèse H_0 stipule que tous les paramètres du modèle sauf celui associé à X_0 (noté μ) sont nuls.

La somme des carrés du modèle est $SCM = \left\|\widehat{Y} - \overline{Y}X_0\right\|^2$.

Proposition 11.2.2. *Sous l'hypothèse H_0, $\frac{SCM}{\sigma^2} \sim \chi^2_p$, et est indépendant de SCR.*

Démonstration. Sous H_0, $E = Y - \mu X_0$. La décomposition de $\frac{E}{\sigma}$ sur $L_0 \oplus L^* \oplus L^\perp$ s'écrit

$$\frac{Y - \mu X_0}{\sigma} = \frac{\overline{Y} - \mu}{\sigma} X_0 + \frac{\widehat{Y} - \overline{Y}X_0}{\sigma} + \frac{Y - \widehat{Y}}{\sigma}$$

Le théorème 11.1, p. 313 implique que $\frac{\left\|(\overline{Y}-\mu)X_0\right\|^2}{\sigma^2} \sim \chi^2_1$, $\frac{\left\|\widehat{Y}-\overline{Y}X_0\right\|^2}{\sigma^2} \sim \chi^2_p$ et $\frac{\left\|Y-\widehat{Y}\right\|^2}{\sigma^2} \sim \chi^2_{n-p}$, et que ces 3 variables aléatoires sont indépendantes. Les 2 dernières variables sont respectivement la somme des carrés du modèle et la somme des carrés résiduelle. La première variable reste purement théorique, on ne peut pas la calculer car on ne connait pas μ. □

11.2.4 Modèles emboîtés : loi de la différence entre les sommes de carrés des modèles.

On considère deux modèles emboîtés $M_1 \subset M_0$, $\mathbb{E}(Y) = X_1\theta_1$ et $\mathbb{E}(Y) = X_0\theta_0$.

Les sous espaces vectoriels associés, L_0 et L_1, vérifient $L_1 \subset L_0$, $dim L_1 = p_1 + 1$, $dim L_0 = p_0 + 1$, avec $p_0 > p_1$.

Soit SCM_0 et SCM_1 les sommes des carrés des modèles M_0 et M_1. Soit η les paramètres du modèle M_0 associés aux colonnes de X_0 qui ne sont pas contenues dans L_1.

Proposition 11.2.3. *Sous l'hypothèse $H_0 = \{\eta = 0\}$, $\frac{SCM_0 - SCM_1}{\sigma^2} \sim \chi^2_{p_0-p_1}$ et est indépendante de la somme de carrés résiduelle du modèle M_0.*

Démonstration. Soit $L_0 \cap L_1^\perp$ le sous espace vectoriel de L_0 qui est orthogonal à L_1. Soit P_i le projecteur orthogonal sur L_i, $i = 0, 1$.

On utilise la décomposition $\mathbb{R}^n = L_1 \oplus (L_0 \cap L_1^\perp) \oplus L_0^\perp$ appliquée à $\frac{E_1}{\sigma} \sim \mathcal{N}(0, I_n)$.

On écrit $E_1 = (P_1 + (P_0 - P_1) + (I - P_0))(E_1)$.

$$\frac{E_1}{\sigma} = \frac{E_1 - \widehat{E_1}}{\sigma} + \frac{\widehat{E_1} - \widehat{E_0}}{\sigma} + \frac{\widehat{E_0}}{\sigma} \tag{11.1}$$

en effet

$$\begin{aligned} P_1(E_1) &= P_1(Y - X_1\theta_1) \\ &= P_1(Y) - X_1\theta_1 \\ &= P_1(Y) - Y + (Y - X_1\theta_1) \\ &= E_1 - \widehat{E_1} \end{aligned}$$

$$\begin{aligned} (P_0 - P_1)(E_1) &= (P_0 - P_1)(Y - X_1\theta_1) \\ &= (P_0 - P_1)(Y) \\ &= (I - P_1)(Y) - (I - P_0)Y \\ &= \widehat{E_1} - \widehat{E_0} \end{aligned}$$

$$\begin{aligned} (I - P_0)(E_1) &= (I - P_0)(Y - X_1\theta_1) \\ &= (I - P_0)(Y) \\ &= \widehat{E_0} \end{aligned}$$

Il est facile de vérifier que les 3 composantes sont orthogonales en utilisant la propriété $P_1 \circ P_0 = P_0 \circ P_1 = P_1$.

Le théorème 11.1, p. 313 implique que $\frac{\left\|E_1 - \widehat{E_1}\right\|^2}{\sigma^2} \sim \chi^2_{p_1}$, $\frac{\left\|\widehat{E_1} - \widehat{E_0}\right\|^2}{\sigma^2} \sim \chi^2_{p_0 - p_1}$ et $\frac{\left\|\widehat{E_0}\right\|^2}{\sigma^2} \sim \chi^2_{n-p_0}$, et que ces 3 variables aléatoires sont indépendantes. Les 2 dernières variables sont respectivement $\frac{SCM_0 - SCM_1}{\sigma^2}$ et $\frac{SCR_0}{\sigma^2}$ où SCR_0 est la somme des carrés résiduelle. La première variable reste purement théorique car on ne connait pas E_1. □

11.3 Espérance et matrice de variance-covariance de $\widehat{\theta}$

11.3.1 Espérance de $\widehat{\theta}$

Proposition 11.3.1. $\mathbb{E}(\widehat{\theta}) = \theta$.

Démonstration. Soit r le rang de X.

Cas $r = p + 1$. L'estimateur des paramètres θ est : $\widehat{\theta} = (X'X)^{-1}X'Y$

Le seul terme aléatoire de cette expression étant Y, on a :

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

Cas $r < p+1$. L'estimateur de θ est $\widehat{\theta} = (G'G)^{-1}X'Y$ et son espérance est égale à :

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

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

□

11.3.2 Matrice de variance-covariance de $\widehat{\theta}$

Proposition 11.3.2. *Si* $r = p+1$, $\mathbb{V}(\widehat{\theta}) = \sigma^2(X'X)^{-1}$,
et si $r < p+1$, $\mathbb{V}(\widehat{\theta}) = \sigma^2(X'X)^{-}X'X(X'X)^{-}$.

Démonstration. On traite les 2 cas séparément :

Cas $r = p+1$. D'après la forme de l'estimateur donné ci-dessus et en utilisant la propriété 10.1.1, p. 306 avec $A = (X'X)^{-1}X'$, on obtient que

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

$\mathbb{V}(Y) = \sigma^2 I_n$, donc

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

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

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

Si on choisit une inverse généralisée réflexive, c'est-à-dire qui satisfait aussi l'égalité : $(X'X)^{-}(X'X)(X'X)^{-} = (X'X)^{-}$, on a alors

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

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

□

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

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

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

11.4 Espérance de la somme des carrés résiduelle

Proposition 11.4.1. $\mathbb{E}(SCR) = \sigma^2 (n - r)$.

Démonstration. ce résultat vient de la proposition 11.2.1, p. 315 et du fait que l'espérance d'un χ^2 est son nombre de degrés de liberté.

On peut aussi faire une démonstration directe :
Par définition $SCR = (Y - \widehat{Y})'(Y - \widehat{Y}) = (Y - PY)'(Y - PY) = Y'(I_n - P)Y$, où P est le projecteur orthogonal sur le sous-espace $L(X)$, et I_n est la matrice identité d'ordre n. Le résultat a été obtenu en utilisant les propriétés d'idempotence et de symétrie de P et de I_n ($P^2 = P$ et $P' = P$).

On a donc $\mathbb{E}(SCR) = \mathbb{E}\left[Y'(I_n - P)Y\right]$.
La matrice $\mathbb{E}(SCR)$ est égale à sa trace car elle est de dimension 1×1. Cela permet d'utiliser la propriété de commutativité de la trace d'un produit de matrices. De plus les opérateurs trace et espérance sont commutatifs ce qui donne :

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

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

$$\mathbb{E}(SCR) = \sigma^2 \text{trace}\left[(I_n - P)\right] = \sigma^2 \left[n - \text{trace}P\right].$$

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

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

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

□

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

11.5.1 Loi sous H_0

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

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

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

11.5.2 Loi sous H_1

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

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

Quand on remplace $\mathbb{V}(\widehat{\theta}_1)$ par $\widehat{\mathbb{V}(\widehat{\theta}_1)}$, la loi $\mathcal{N}(\theta_1 - a, 1)$ devient une loi de Student **décentrée** dont le nombre de degrés de liberté est ν_1 et le paramètre de décentrement est $\theta_1 - a$. Ceci permet de calculer la puissance du test, mais la loi de Student décentrée n'est pas facile à manipuler.

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

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

11.6.1 Loi sous H_0

Sous H_0, la proposition 11.2.3, p. 315 indique que $SCM_1 - SCM_0$ et SCR_1 sont des variables aléatoires indépendantes et distribuées selon des lois du χ^2 à respectivement $p_1 - p_0$ et ν_1 degrés de liberté. Par suite la statistique de test $\frac{(SCM_1 - SCM_0)/(p_1 - p_0)}{SCR_1/\nu_1}$, où $\nu_1 = n - p_1$ est distribuée selon une loi de Fisher, $\mathcal{F}(p_1 - p_0, \nu_1)$ sous H_0.

11.6.2 Loi sous H_1

$SCM_1 - SCM_0$ est une somme de carrés de lois normales non centrées. Elle est distribuée selon une loi du χ^2 décentrée, $\chi^2(p_1 - p_0, \lambda)$, où $p_1 - p_0$ est le nombre de degrés de liberté et λ le paramètre de décentrement. Ce dernier est égal à la somme des carré des paramètres de θ qui ont été annulés dans θ_0, divisée par σ^2. Par suite, la statistique $\frac{(SCM_1 - SCM_0)/(p_1 - p_0)}{SCR_1/\nu_1}$ est distribuée selon une loi de Fisher décentrée, $F(p_1 - p_0, \nu_1, \lambda)$ sous H_1. Cette distribution permet de calculer la puissance du test.

11.7 Démonstration du théorème de Gauss-Markov

On rappelle ce théorème :

Théorème 11.7.1. *(Gauss-Markov) $\widehat{\theta}$ est le meilleur estimateur linéaire sans biais de θ au sens suivant :*

$$\forall \tilde{\theta},\ \forall C,\ \mathbb{V}(C\widehat{\theta}) \leq \mathbb{V}(C\tilde{\theta}),$$

où C est un vecteur $(1, p+1)$ et $\tilde{\theta}$ un estimateur linéaire sans biais de θ.

Démonstration. L'estimateur des moindres carrés dans le cas où $r < p+1$ est $\widehat{\theta} = (G'G)^{-1}X'Y$. Si $r = p+1$, on remplace G par X et la démonstration reste valide. Tout d'abord, on voit que $\widehat{\theta}$ est un estimateur linéaire de Y puisqu'il existe une matrice B tel que $\widehat{\theta} = BY$. De plus, il est sans biais d'après la proposition 11.3.1, p. 316.
Soit $\tilde{\theta}$ un estimateur linéaire sans biais de θ. Comme il est linéaire en Y, il existe A tel que $\tilde{\theta} = AY$. De plus on a $\mathbb{E}(\tilde{\theta}) = \mathbb{E}(AY) = A\mathbb{E}(Y) = AX\theta$. Comme $\tilde{\theta}$ est sans biais,

on a pour tout θ, $AX\theta = \theta$, ce qui implique que $AX = I_{p+1}$.

On a :

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}(\tilde{\theta} + \widehat{\theta} - \widehat{\theta}) = \mathbb{V}(\tilde{\theta} - \widehat{\theta}) + \mathbb{V}(\widehat{\theta}) - 2\,\mathbb{C}\mathrm{ov}(\tilde{\theta} - \widehat{\theta}, \widehat{\theta}).$$

De plus

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\tilde{\theta} - \widehat{\theta}, \widehat{\theta}) &= \mathbb{C}\mathrm{ov}(AY, (G'G)^{-1}X'Y) - \mathbb{V}(\widehat{\theta}) \\
&= A\mathbb{V}(Y)[(G'G)^{-1}X']' - \mathbb{V}(\widehat{\theta}) \\
&= \sigma^2 AX(G'G)^{-1} - \mathbb{V}(\widehat{\theta}) \\
&= \sigma^2(G'G)^{-1} - \mathbb{V}(\widehat{\theta}) \\
&= 0.
\end{aligned}$$

Donc

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}(\tilde{\theta} - \widehat{\theta}) + \mathbb{V}(\widehat{\theta}).$$

Finalement

$$\begin{aligned}
\mathbb{V}(C\tilde{\theta}) &= C\mathbb{V}(\tilde{\theta})C' \\
&= C\left[\mathbb{V}(\tilde{\theta} - \widehat{\theta}) + \mathbb{V}(\widehat{\theta})\right]C' \\
&= \mathbb{V}[C(\tilde{\theta} - \widehat{\theta})] + \mathbb{V}(C\widehat{\theta}) \\
&\geq \mathbb{V}(C\widehat{\theta}).
\end{aligned}$$

□

11.8 Démonstration du résultat (1.14, p. 24)

On veut démontrer que

$$cor(\widehat{\beta}_i, \widehat{\beta}_j) = -\widehat{cor}(X^{(i)}, X^{(j)}|X_{\setminus(i,j)}),$$

Les corrélations simples ou partielles entre les colonnes de X ne sont pas des corrélations entre variables aléatoires mais simplement des corrélations empiriques entre colonnes et notées (improprement) $\widehat{cor}$. Soit $X = (X^{(p)}, X^{(p-1)}, \ldots, X^{(3)}, X^{(1)}, X^{(2)})$ la matrice du modèle linéaire $Y = X\theta + E$, $H = Vect(X^{(p)}, \ldots, X^{(3)})$ le sous espace de $\mathbb{R}^n$ engendré par les colonnes $X^{(3)}, \ldots, X^{(p)}$, et P_H le projecteur ortogonal sur H. On a changé l'ordre des variables pour des raisons pratiques de notation.

On suppose, pour simplifier l'écriture mais sans perte de généralité, que les $X^{(i)}$ sont centrées et normées. On note $T = (X^{(1)}, X^{(2)})$, $W = (X^{(p)}, \ldots, X^{(3)})$, $V = X'X = \begin{bmatrix} V_{WW} & V_{WT} \\ V_{TW} & V_{TT} \end{bmatrix}$ et

$\Omega = V^{-1} = \begin{bmatrix} \Omega_{WW} & \Omega_{WT} \\ \Omega_{TW} & \Omega_{TT} \end{bmatrix}$.

Le coefficient de corrélation partielle empirique entre $X^{(1)}$ et $X^{(2)}$ conditionnellement à $X^{(3)}, \ldots, X^{(p)}$ est défini par

$$\widehat{cor}(X^{(1)}, X^{(2)}|X_{\setminus(1,2)}) = \frac{\langle (I - P_H)(X^{(1)}), (I - P_H)(X^{(2)}) \rangle}{\left\|(I - P_H)(X^{(1)})\right\| \left\|(I - P_H)(X^{(2)})\right\|} \tag{11.2}$$

On note $I - P_H$ la matrice de $I - P_H$. On a

$$\begin{aligned}((I-P_H)T)'(I-P_H)T &= T'(I-P_H)T \\ &= T'T - T'W(W'W)^{-1}W'T \\ &= V_{TT} - V_{TW}V_{WW}^{-1}V_{WT} \\ &= \Omega_{TT}^{-1}\end{aligned}$$

La dernière égalité provient de la relation (10.1, p. 309) appliquée à V. On a

$$\Omega_{TT}^{-1} = \begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}^{-1} = \frac{1}{\omega_{11}\omega_{22} - \omega_{12}^2}\begin{bmatrix}\omega_{22} & -\omega_{21} \\ -\omega_{12} & \omega_{11}\end{bmatrix},$$

Donc la corrélation partielle entre X_1 et X_2 étant donné l'ensemble des autres variables, vaut

$$\widehat{cor}(X^{(1)}, X^{(2)}|X_{\setminus(1,2)}) = \frac{-\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}}$$

Ω est, à la constante σ^2 près, la matrice de variance-covariance des estimateurs des paramètres de régression $\beta_0, \beta_1, \ldots, \beta_p$. On a donc

$$cov(\widehat{\beta}_1, \widehat{\beta}_2) = \omega_{12}\sigma^2,$$

et donc

$$cor(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} = -\widehat{cor}(X^{(1)}, X^{(2)}|X_{\setminus(1,2)})$$

Le choix des variables $X^{(1)}$ et $X^{(2)}$ étant arbitraire, la propriété (11.2, p. 320) est vraie pour tout couple $(i, j), \in (1, p)^2$.

Chapitre 12

Annexe4 : algorithme de Newton Raphson

Par souci de simplification, l'algorithme de Newton-Raphson est présenté ici dans sa version scalaire, i.e le vecteur θ n'a qu'une composante et sera noté θ. De plus, pour alléger l'écriture, nous omettons d'écrire le paramètre ϕ dans $\ell(y;\theta,\phi)$.

Soit $\hat{\theta}$ la solution de $\frac{\partial \ell(y;\theta)}{\partial \theta} = 0$. L'algorithme de Newton-Raphson est un algorithme itératif basé sur le développement de Taylor à l'ordre 1 de la fonction $v(\theta) = \frac{\partial \ell(y;\theta)}{\partial \theta}$ en un point θ autour de $\hat{\theta}$, i.e $v(\hat{\theta}) = v(\theta) + (\hat{\theta} - \theta)v'(\theta) + o(|\theta - \hat{\theta}|)$. Par définition de $\hat{\theta}$, $v(\hat{\theta}) = 0$, donc

$$0 = \frac{\partial \ell(y;\theta)}{\partial \theta}\Big|_\theta + (\hat{\theta} - \theta)\frac{\partial^2 \ell(y;\theta)}{\partial^2 \theta}\Big|_\theta + o(|\theta - \hat{\theta}|),$$

d'où

$$\hat{\theta} = \theta - \left(\frac{\partial^2 \ell(y;\theta)}{\partial^2 \theta}\Big|_\theta\right)^{-1} \frac{\partial \ell(y;\theta)}{\partial \theta}\Big|_\theta + o\left(\left|\left(\frac{\partial^2 \ell(y;\theta)}{\partial^2 \theta}\Big|_\theta\right)^{-1} (\hat{\theta} - \theta)\right|\right).$$

12.1 Description de l'algorithme

La figure 12.1, p. 323 montre le principe de l'algorithme consiste à approcher localement la fonction h par sa tangente. Soit θ_{h-1} la valeur du paramètre θ à l'itération $(h-1)$. A l'itération (h), l'algorithme actualise la valeur de θ de la façon suivante :

$$\theta_h = \theta_{h-1} - \left(\frac{\partial^2 \ell(y;\theta)}{\partial^2 \theta}\Big|_{\theta=\theta_{h-1}}\right)^{-1} \frac{\partial \ell(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_{h-1}} \tag{12.1}$$

θ_h est simplement l'abscisse qui annule la fonction tangente de $v(\theta)$ en θ_{h-1}, d'équation $y = v(\theta_{h-1}) + v'(\theta_{h-1})(\theta - \theta_{h-1})$.

Dans le cas où θ est un vecteur de p paramètres, l'équation (12.1, p. 322) reste la même mais $\frac{\partial^2 \ell(y;\theta)}{\partial^2 \theta}\Big|_{\theta=\theta_{h-1}}$ est une matrice (p,p) et $\frac{\partial \ell(y;\theta)}{\partial \theta}\Big|_{\theta=\theta_{h-1}}$ est un vecteur $(p,1)$.

Le Fisher-scoring est un algorithme similaire à celui de Newton-Raphson mais au lieu de rechercher la nouvelle valeur de θ suivant la direction $-\frac{\partial^2 \ell(y;\theta)}{\partial \theta^2}$, il la recherche

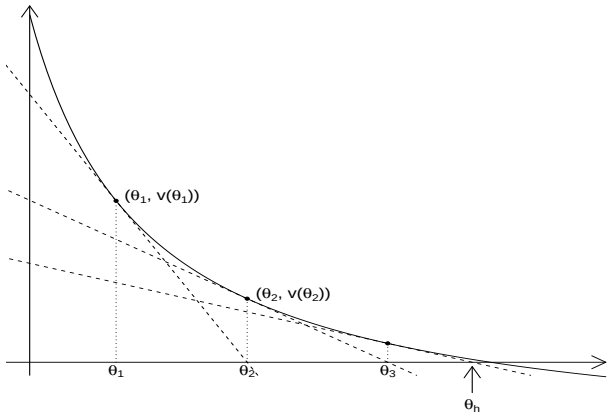


FIGURE 12.1 – Illustation de l'algorithme de Newton-Raphson recherche de la racine de la fonction $v(\theta) = \frac{\partial \ell(y;\theta)}{\partial \theta}$

suivant son espérance $-\mathbb{E}\left[\frac{\partial^2 \ell(y;\theta)}{\partial \theta^2}\right]$, qui n'est autre que l'information de Fisher. Cette solution peut permettre d'éviter les problèmes de non inversibilité de la dérivée seconde.

Ces 2 algorithmes peuvent donner des maximums locaux si la vraisemblance n'est pas une fonction concave des paramètres et ne pas converger si la matrice des dérivées secondes n'est pas inversible.

Un autre cas de non-convergence se produit pour la régression logistique, paradoxalement quand les 2 sous-populations des observations pour lesquelles la variable réponse vaut respectivement 1 et 0 sont parfaitement séparées. Cela se produit s'il existe θ tel que pour toute observation i, $Y_i = 1 \Leftrightarrow x_i\theta > 0$. Dans ce cas les paramètres estimés sont de plus en plus grands en valeur absolue à chaque itération et les logiciels donnent un avertissement.

12.2 Cas de la fonction de lien naturel

D'après la relation 3.3, p. 120, on a $I_n(\theta, \phi) = \frac{1}{\gamma(\phi)} X'\mathbb{V}(Y)X$. Si X est de plein rang, la matrice $X'\mathbb{V}(Y)X$ est définie positive et la matrice $\dfrac{\partial^2 \ell(y;\theta)}{\partial \theta^2}$ est alors définie négative. La log-vraisemblance est alors une fonction strictement concave de θ, et le maximum de cette fonction existe et est unique. Cette propriété assure que l'algorithme converge bien vers le maximum.

Index

AIC, 38, 117, 126, 131, 174
algorithme de Newton-Raphson, 114, 162, 302
alias, 226
analyse de la covariance, 11, 14, 17, 37, 93
analyse de la variance, 12, 253, 259
  à deux facteurs, 31, 36, 69, 82, 220, 222, 243, 248
  à un facteur, 20, 61

base orthonormée, 283
BIC, 118, 174
bloc, 210, 220, 229
blocs complets, 215, 218, 242
blocs incomplets équilibrés(BIE), 176, 220, 221, 245
blocs incomplets partiellement équilibrés, 258
BLUP, 170

carré latin, 216, 223
cellule, 210
change-over design, 253
coefficient de détermination, 38
coefficient de détermination ajusté, 38
comparaison de modèles, 38
comparaisons multiples, 65, 77, 88, 99, 101, 104, 251
composantes de la variance, 156, 187
confusion d'effets, 209, 210, 216, 218
contrôle de l'hétérogénéité, 209, 215
contraste, 20
corrélation spatiale, 159
cote, 120
courbe ROC, 121
covariable, 210

déviance, 115, 126, 131, 142
diagnostic, 39–41, 46, 47, 52, 61, 84, 99, 182

estimable, 20–22, 221, 224
exemple de plan d'expériences
  Teleost, 258
  champagnes, 245
  croissance de moutons, 253
  fabrication de biscuits, 255
  fongicide, 270
  mesure de polluants, 268
  production laitière, 218, 220
  réaction chimique, 275
  rendement de maïs, 242
  textile, 279
  thermostat, 263
exemple modèle linéaire
  alisiers, 59
  brochets, 43
  chenilles processionnaires, 49
  colza, 67
  notes, 80
  pollution dans le métro, 91
exemple modèle mixte
  avalanches, 201
  héritabilité, 183
  nutrition humaine, 194
  samares du frêne, 186
exemple modele linéaire généralisé
  Bernoulli
    asthme des enfants, 137
    ozone, 125
  binomiale
    équité sociale, 130
    dose-réponse, 133
  gamma
    roulements à bille, 149
  multinomiale
    condamnations à mort, 151
  Poisson
    brebis, 145
    fourmis en Guyane, 141

facteur étudié, 210
facteur de contrôle de l'hétérogénéité, 210, 219, 222, 243
famille exponentielle, 111
fonction de lien, 112
fonction log-log, 135
formule de Henderson, 170

indépendance conditionnelle, 152
information de Fisher, 114, 143, 147
interaction, 27, 32, 33, 37
inverse généralisée, 297
inversion par blocs d'une matrice, 289

levier, 39, 71
logit, 119, 126, 130
loi conditionnelle(vecteur gaussien), 170, 289
loi de $\widehat{\theta}$, 296
loi de Bernoulli, 112
loi de Fisher, 292
loi de Fisher décentrée, 292
loi de Poisson, 112
loi de Student, 291
loi des sommes de carrés, 294–296
loi du χ^2, 291
loi du χ^2 décentrée, 292
loi normale multidimensionnelle, 288
loi normale unidimensionnelle, 288

matrice de variance d'un vecteur aléatoire, 287
matrice orthogonale, 284
mesures répétées, 158, 194
modèle hiérarchique, 187
modèle linéaire généralisé, 110
modèle linéaire généralisé mixte, 203
modèle mixte
  estimation
    sandwitch, 169
    EM, 163, 164
    moments, 166, 168, 188
    REML, 165, 189
  modèle de Laird et Ware, 172
  test des paramètres de moyenne
    test approché, 179, 191, 199
    test de Wald, 180
    test exact, 175, 179
  test des paramètres de variance
    test de Wald, 185
    test du rapport de vraisemblance, 173, 185, 189
    test exact, 172, 185
modèles emboîtés, 26, 27, 29, 31, 53, 115
moyenne ajustée, 87, 88, 98, 99, 101

niveau d'un facteur, 209

orthogonalité, 16, 35, 37, 73, 81, 131, 228, 245
orthonormalisation de Gram-Schmidt, 283

paradoxe de Simpson, 154
plan Box-Behnken, 240
plan complet, 218
plan composite centré, 236
plan d'expériences optimal, 214
plan de Doehlert, 240
plan de Youden, 223
plan fractionnaire, 224, 255, 263, 268
plan incomplet partiellement équilibré, 222
plan Plackett-Burman, 264
plan pour la conception robuste, 267
plans pour les expériences numériques, 241
plans pour les mélanges, 240, 279
probit, 133
projecteur, 284, 285
projecteur orthogonal, 286
pseudo R^2, 117
puissance, 209, 211–213, 243

quantiles demi-normaux, 231

régression logistique, 119, 138
  classement, 121
  test exact, 120
régression multilogistique, 122
régression multiple, 17, 21, 22, 26, 31, 32, 37, 52
régression polynomiale, 17, 44, 279
Régression simple, 304
régression simple, 12, 43, 44
répétition, 210
résidus, 23, 24, 39, 40, 44–47, 71, 84, 118, 126, 181, 248
résidus standardisés, 39, 40, 44, 45, 52, 118, 181

résolution, 228
règle d'affectation, 217
règle de répartition en proportion, 217
randomisation, 210, 217
rapport de cotes, 121, 140
risque relatif, 121, 140

sélection de variables, 54, 138
somme de carrés de type I, 31, 35
somme de carrés de type II, 31, 32, 35
somme des carrés résiduelle, 24
somme directe de sous-espaces, 283
sous-espaces orthogonaux, 283
split-plot, 233, 270
supplémentaire orthogonal, 283
surdispersion, 123, 146
surface de réponse, 233, 269, 275

table de contingence, 151
test de Hosmer-Lemeshow, 120, 140
test de Wald, 126, 131
théorème de Cochran, 293
théorème de Pythagore, 283
trace, 285
transposée(notation), 287

unité de traitement, 210
unité expérimentale, 210

validation du modèle, 238
variabilité expérimentale, 215
variable explicative, 209
variable réponse, 209
vecteur aléatoire, 287
vecteur gaussien, 288, 289

Bibliographie

[1] Alan Agresti. *Categorical Data Analysis, 3rd Edition.* Wiley, 2013.

[2] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F Caski, editors, *Proceeding of the Second International Symposium on Information Theory, Budapest*, pages 267–281, 1973.

[3] Alnosaier. *Kenward-Roger Approximate F Test for Fixed Effects in Mixed Linear Models.* PhD thesis, Oregon State University, 2007.

[4] C.I. Bliss. The calculation of the dosage-mortality curve. *Annals of Applied BIology*, 22 :134–167, 1935.

[5] Benjamin M. Bolker, Beth Gardner, Mark Maunder, Casper W. Berg, Mollie Brooks, Liza Comita, Elizabeth Crone, Sarah Cubaynes, Trevor Davies, Perry de Valpine, Jessica Ford, Olivier Gimenez, Marc Kéry, Eun Jung Kim, Cleridy Lennert-Cody, Arni Magnusson, Steve Martell, John Nash, Anders Nielsen, Jim Regetz, Hans Skaug, and Elise Zipkin. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, 4(6) :501–512, 2013.

[6] Marjolaine Bourdat-Deschamps, Jean-Jacques Daudin, and Enrique Barriuso. An experimental design approach to optimise the determination of polycyclic aromatic hydrocarbons from rainfall water using stir bar sorptive extraction and high performance liquid chromatography-fluorescence detection. *Journal of Chromatography A*, 1167(2) :143–153, 2007.

[7] G. E. P. Box, W. G. Hunter, and J.S. Hunter. *Statistics for Experimenters : An Introduction to Design, Data Analysis, and Model Building.* Wiley, 1978.

[8] S. Brachet. *La dispersion : déterminisme et conséquences. Approche théorique et experimentale chez le frêne.* PhD thesis, ENGREF, 1999.

[9] R.G. Bullington, S. Lovin, D. Miller, and W. Woodall. Improvement of an industrial thermostat using designed experiments. *Journal of Quality Technology*, 25(4) :262–270, 1993.

[10] M Chavance and S Escolano. Misspecification of the covariance structure in generalized linear mixed models. *Statistical methods in medical research*, page 0962280212462859, 2012.

[11] William G. Cochran. The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3) :315–345, 09 1952.

[12] D. Collombier. *Plans d'expérience factoriels : construction et propriétés des fractions de plans.* Springer, 1996.

[13] William J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, New York, 1971.

[14] R. Dennis ; Cook and Sanford Weisberg. *Residuals and influence in regression*. Chapman and Hall, New York, 1982.

[15] J.A. Cornell. *Experiments with mixtures*. Wiley, 1990.

[16] P Dagnelie. *Principes d'expérimentation : planification des expériences et analyse de leurs résultats*. Edition electronique, 2012. www.dagnelie.be.

[17] M.N. Dasand and N.C. Giri. *Design and Analysis of Experiments*. John Wiley and Sons Inc, 1987.

[18] JJ. Daudin, S. Robin, and C. Vuillet. *Statistique inférentielle*. PUR, 2001.

[19] A.M. Dean and D. Voss. *Design and Analysis of Experiments*. Springer, 1999.

[20] M. Delattre. *Inférence statistique dans les modèles mixtes à dynamique Markovienne*. PhD thesis, Ecole doctorale Mathématiques de la région Paris-Sud, 2012.

[21] E. Demidenko. *Mixed Models : Theory and Applications*. Wiley Series in Probability and Statistics, 2004.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. 39 :1–38, 1977.

[23] James Durbin and Geoffrey S. Watson. Testing for serial correlation in least squares regression, i. *Biometrika*, 37 :409–428, 1950.

[24] James Durbin and Geoffrey S. Watson. Testing for serial correlation in least squares regression, ii. *Biometrika*, 38 :159–179, 1951.

[25] M. Flamant-Hulin. *Pollution intérieure et santé respiratoire : données issues des milieux urbain et rural*. PhD thesis, Ecole doctorale Paris6, 2010. `http://www.theses.fr/2010PA066721`.

[26] Hélène Fouillet, Barbara Juillet, Cécile Bos, François Mariotti, Claire Gaudichon, Robert Benamouzig, and Daniel Tomé. Urea-nitrogen production and salvage are modulated by protein intake in fed humans : results of an oral stable-isotope-tracer protocol and compartmental modeling. *The American journal of clinical nutrition*, 87(6) :1702–1714, 2008.

[27] JP. Gauchi. Plans d'expériences optimaux : un exposé didactique. *Modulad*, 33 :139–162, 2005. `www-rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-gauchi-33/gauchi-33-tutorial.pdf`.

[28] Christian Gourieroux, Alain Monfort, and Alain Trognon. Pseudo maximum likelihood methods : Theory. *Econometrica : Journal of the Econometric Society*, pages 681–700, 1984.

[29] Groc Sarah, Delabie Jacques H.C., Fernández Fernando, Leponce Maurice, Orivel Jérôme, Silvestre Rogerio, Vasconcelos Heraldo L., and Dejean Alain. Leaf-litter ant communities (Hymenoptera : Formicidae) in a pristine Guianese rainforest : stable functional structure versus high species turnover. *Myrmecological News*, 19 :43–51, 2013.

[30] Michael Hamadan and John A. Nelder. Generalized linear models for quality-improvement experiments. *Journal of Quality Technology*, 29(3) :292–304, 1997.

[31] Charles R Henderson. Estimation of genetic parameters. In *Biometrics*, volume 6-2, pages 186–187. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, 1950.

[32] Charles R Henderson. Selection index and expected genetic advance. *Statistical genetics and plant breeding*, 982 :141–163, 1963.

[33] Karim Hirji, F. Mehta, R. Cyrus, Patel, and R Nitin. Computing distributions for exact logistic regression. *JASA*, 82(6) :1110–1117, 1987.

[34] Sylvie Huet, Anne Bouvier, and Marie-Anne Poursat. *Statistical Tools for Nonlinear Regression. A Practical Guide With S-PLUS and R Examples.* Springer Series in Statistics. New York, NY : Springer, 2010.

[35] Bertrand Iooss. Revue sur l'analyse de sensibilité globale de modèles numériques. *Journal de la Société FranÃ§aise de Statistique*, 152(1), 2011. `http://smf4.emath.fr/Publications/JSFdS/152_1/pdf/sfds_jsfds_152_1_3-25.pdf`.

[36] M.G. Kenward and J.H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53 :983–997, 1997.

[37] A.I. Khuri, T. Mathew, and B.K. Sinha. *Statistical Tests for Mixed Linear Models.* Wiley Series in Probability and Statistics, 1998.

[38] A. Kobilinsky, H. Monod, and O. David. Generation of regular fractional factorial designs. *web*, 2012. `www.jouy.inra.fr/unites/miaj/public/logiciels/planor`.

[39] R. V. Lenth. Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4) :469–473, 1989.

[40] R. Linder. *Les plans d'expériences - Un outil indispensable à l'expérimentateur.* Presses Ponts Et Chaussées, 2005.

[41] T. Mary-Huard, J. Aubert, N. Mansouri-Attia, O. Sandra, and J.-J. Daudin. Statistical methodology for the analysis of dye-switch microarray experiments. *BMC Bioinformatics*, 9(98), 2008.

[42] D. C. Montgomery. *Design and Analysis of Experiments.* Wiley, 1976.

[43] N.M. Nan M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4) :963–974, 1982.

[44] M.F. Oleksiak, G.A. Churchill, and D.L. Crawford. Variation in gene expression within and among natural populations. *Nature Genetics*, 32 :261–266, 2002.

[45] J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS.* Statistics and Computing. Springer, 2000.

[46] Michael L Radelet. Racial characteristics and the imposition of the death penalty. *American Sociological Review*, 46(6) :918–927, 1981.

[47] Bureau régional de l'Organisation mondiale de la Santé pour l'Europe. *Air quality guidelines for Europe.* Number 23 in OMS, Publications rÃ©gionales, SÃ©rie europÃ©enne. Publications régionales de l'OMS, Série européenne, 1987.

[48] O. Ronce. *Histoires de vie dans un habitat fragmenté : étude théorique de l'évolution de la dispersion et d'autres traits.* PhD thesis, Montpellier II, 1999.

[49] Wilks S. S. *Mathematical Statistics.* Princeton University Press, 1943.

[50] G. Saporta, J.-J Droesbeke, and J. Fine. *Plans d'expériences, Applications Ã l'entreprise*. Editions Technip, 1997.

[51] SAS Institute Inc. *SAS/STAT(R) 9.2 User's Guide, Second edition.*

[52] S.R. Searle. *Linear models*. Wiley, 1st edition, 1971.

[53] Samuel Sanford Shapiro Shapiro and Martin Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52 :591–611, 1965.

[54] P. Tassi. *Méthodes statistiques*. Economica, 1985.

[55] R. Tomassone, S. Audrain, E. Lesquoy-de Turckeim, and Millier C. *La régression, nouveaux regards sur une ancienne méthode statistique*. Masson, 1992.

[56] van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[57] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer, 2009.

[58] S.N Wood. *Generalized Additive Models : An Introduction with R*. Chapman and Hall/CRC, 2006.

[59] Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.