



The Generalized Linear Model (GLM)

C. Marteau

Année universitaire 2018-2019

Clément Marteau
Bâtiment Braconnier, Bureau 109b
marteau@math.univ-lyon1.fr

Table des matières

1	Introduction et rappels	1
1.1	La régression linéaire	1
1.2	L'analyse de variance	3
1.3	La régression logistique	4
1.4	Similitudes	5
1.5	Objectifs du cours	7
1.6	Exercices	7
2	Le modèle linéaire	9
2.1	Définition	9
2.2	Estimation	10
2.3	Validation (graphique) des hypothèses du modèle	11
2.4	Tests	13
2.4.1	Test sur la présence d'un sous-modèle	13
2.4.2	Test sur la nullité d'une combinaison linéaire	14
2.4.3	Test sur la nullité jointe de plusieurs combinaisons linéaires	15
2.4.4	Méthodologie	15
2.4.5	La problématique des tests multiples	15
2.5	Modèles singuliers	16
2.6	Exercices	18
3	La sélection de modèle	21
3.1	Cadre général	21
3.2	Une première approche	22
3.2.1	Les coefficients d'ajustement	22
3.2.2	Stratégies de régression ascendantes et descendantes	22
3.3	Trois critères de sélection de modèle	23
3.3.1	Risque quadratique	23
3.3.2	Le critère C_p de Mallows	25
3.3.3	Le critère AIC	26
3.3.4	Le critère BIC	28
3.4	Exercices	28
4	Statistique en grande dimension	31
4.1	Motivations	31
4.2	Principe de la méthode LASSO	32
4.3	Performances théoriques	34
4.3.1	Un premier résultat	34
4.3.2	Vitesse rapide avec contrainte structurelle sur X	36
5	Le modèle linéaire généralisé	39
5.1	Introduction	39
5.2	Caractérisation d'un modèle	40
5.2.1	Familles exponentielles	40
5.2.2	Fonctions de liens	41

5.2.3	Modèles avec réponse binaire	42
5.3	Estimation	42
5.3.1	Estimation par maximum de vraisemblance	42
5.3.2	Calculs pour des familles exponentielles	44
5.3.3	Retour au modèle linéaire	46
5.4	Inférence pour le modèle linéaire généralisé	46
5.4.1	Loi asymptotique de l'EMV	46
5.4.2	Test du rapport de vraisemblance	47
5.4.3	Sélection de modèle	48
5.5	Exercices	48
6	Analyse de variance et plans d'expériences	51
6.1	Pourquoi planifier l'expérience ?	51
6.2	Contraintes et décomposition	52
6.2.1	Orthogonalité	52
6.2.2	Modèle croisé et modèle additif	53
6.3	Exemples de plans	55
6.3.1	Plans en randomisation totale	55
6.3.2	Plans en blocs complet	55
6.3.3	Plans en blocs incomplets	56
6.3.4	Plans en carré latin	57
6.4	Critères d'optimalité	58
A	Rappels et compléments	61
A.1	Théorème de Cochran.	61
A.2	La méthode de Newton-Raphson	61
A.3	Théorème central limite : condition de Lindeberg	62
B	Exams given in previous sessions	63

Chapitre 1

Introduction et rappels

1.1 La régression linéaire

Les travaux sur la régression linéaire ont été initiés vers le début du XIX^{ème} siècle parallèlement par Gauss (1777-1855) et Legendre (1753-1833). Cet outil est aujourd'hui très largement utilisé dans un grand nombre de domaines (biologie, économie, etc...) et retient encore l'attention d'une importante partie de la communauté scientifique.

Le modèle de régression linéaire s'inscrit dans la problématique suivante : on dispose de deux variables réelles Y et Z , Y étant communément appelée variable à expliquer et Z variable explicative. A partir d'un échantillon d'observations $(Y_i, Z_i)_{i=1\dots n}$, la situation idéale consiste à pouvoir 'expliquer' Y en fonction de Z , c'est-à-dire trouver une fonction f telle que :

$$Y = f(Z).$$

Ce problème s'avère souvent extrêmement compliqué à résoudre de part la richesse des solutions possibles. Il est donc courant d'émettre des hypothèses sur la fonction f et de se concentrer (au moins dans un premier temps) sur le modèle de régression linéaire simple : on suppose qu'il existe a et b tels que :

$$Y_i = aZ_i + b + \epsilon_i, \quad i = 1, \dots, n. \quad (1.1)$$

Les ϵ_i désignent l'erreur associée à chaque observation¹. En statistique, chaque erreur ϵ_i est une variable aléatoire. L'erreur du modèle $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ pouvant par exemple être due à des erreurs de mesure, ou des fluctuations d'un individu à l'autre. Le modèle (1.1) désigne sous cette forme le modèle de dépendance linéaire.

EXEMPLE 1.1 On s'intéresse dans cet exemple à la chenille processionnaire, un parasite du pin. Le but est d'essayer d'expliquer et éventuellement de prédire le nombre de nids de cet animal sur une parcelle donnée. Pour cela, on peut par exemple mettre en relation le (log du) nombre de nids et l'altitude de la parcelle. La Figure 1.1 ci-dessous présente une représentation cartésienne d'observations tirée d'un jeu de données de 1973 sur ces deux variables. On constate que plus l'altitude augmente, moins les chenilles sont présentes. Cette dépendance semble linéaire, il est raisonnable au premier abord de chercher à utiliser le modèle (1.1) pour expliquer ces données.

Dans la mesure où plusieurs variables $Z^{(1)}, \dots, Z^{(p)}$ sont susceptibles de pouvoir expliquer les observations, on utilisera le modèle :

$$Y_i = a_1 Z_i^{(1)} + a_2 Z_i^{(2)} + \dots + a_p Z_i^{(p)} + \epsilon_i, \quad i = 1, \dots, n, \quad (1.2)$$

où p désigne le nombre de variables explicatives. On parle dans ce cas de modèle de régression multiple.

EXEMPLE 1.2 Reprenons l'exemple 1.1. Afin de gagner en pouvoir explicatif, il est possible de

1. Suivant le contexte, on pourra par exemple supposer que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

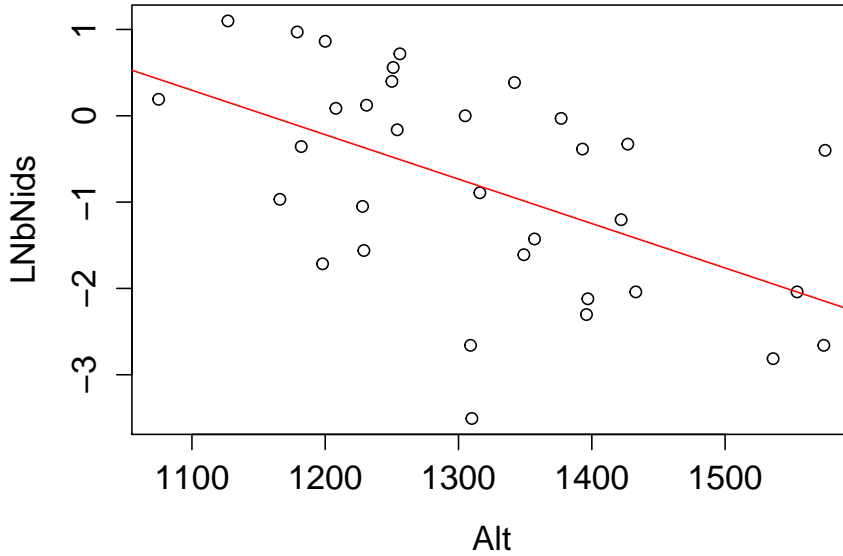


FIGURE 1.1 – Densité de peuplement et altitude

prendre en compte comme prédicteurs l'altitude de la parcelle, le degré de la pente, la densité de pins, l'orientation, etc... On a clairement affaire dans cette situation à un modèle de régression multiple : il y a plusieurs variables explicatives.

Afin de pouvoir travailler plus facilement, il est courant d'imposer une espérance nulle, une variance constante et une indépendance deux-à-deux pour les termes d'erreurs ϵ_i . Ces hypothèses sont peu restrictives et conviennent à un très grand nombre de situations.

Les variables explicatives peuvent être tout aussi bien déterministes qu'aléatoires. Dans cette dernière situation, il est cependant recommandé de travailler avec un design indépendant des erreurs. L'ensemble des résultats présentés dans ce cours peuvent être utilisés pour un design aléatoire. Cependant, par soucis de simplicité et de concision, nous supposons par la suite les Z_i déterministes.

Intéressons-nous à présent à l'estimation des paramètres a et b dans le cadre de la régression linéaire simple. Nous utiliserons pour cela la méthode des moindres carrés. On choisit pour estimer a et b , le couple (\hat{a}, \hat{b}) vérifiant :

$$(\hat{a}, \hat{b}) = \underset{\alpha, \beta}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \alpha Z_i - \beta)^2. \quad (1.3)$$

Proposition 1.1 *Etant donné un échantillon $(Y_i)_{i=1\dots n}$, les estimateurs de a et b par la méthode des moindres carrés sont donnés par :*

$$\begin{cases} \hat{a} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^n (Z_i - \bar{Z})^2}, \\ \hat{b} = \bar{Y} - \hat{a}\bar{Z}, \end{cases}$$

où \bar{Z} et \bar{Y} désignent respectivement la moyenne empirique des vecteurs $(Z_i)_{i=1\dots n}$ et $(Y_i)_{i=1\dots n}$.

PREUVE. En exercice. On pourra s'intéresser à la fonction à deux variables :

$$f : (\alpha, \beta) \mapsto f(\alpha, \beta) = \sum_{i=1}^n (Y_i - \alpha Z_i - \beta)^2.$$

□

Exercice : Vérifier que

$$\mathbb{E}[\hat{a}] = a, \quad \mathbb{E}[\hat{b}] = b,$$

et

$$\text{Var}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^n (Z_i - \bar{Z})^2}, \quad \text{Var}(\hat{b}) = \frac{\sigma^2}{n} \frac{\sum_{i=1}^n Z_i^2}{\sum_{i=1}^n (Z_i - \bar{Z})^2}.$$

Ces résultats préliminaires ne donnent qu'une approximation du modèle linéaire sous jacent. Dans bien des situations, il reste à mener une étude approfondie permettant dans un premier temps de 'valider' le modèle, puis d'exploiter ce dernier : construction de tests, intervalles de confiances, etc... Nous reviendrons plus en détail sur ces notions dans les chapitres suivants.

1.2 L'analyse de variance

L'analyse de variance (ou encore **aov** pour *analyse of variance*) intervient dans la situation suivante : on dispose d'un échantillon d'observations $(Y_i)_{i=1\dots n}$ et d'une ou plusieurs variables qualitatives (ou facteurs). Toute la problématique consiste à essayer de déterminer si ces facteurs ont une influence sur la variable Y .

Considérons dans un premier temps l'analyse de variance à un seul facteur. Supposons par exemple que le facteur en question possède I modalités. On note respectivement μ_i la valeur moyenne de la quantité Y et n_i le nombre d'observations disponibles pour chaque modalité. On considère alors le modèle :

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, n_i, \quad (1.4)$$

avec par exemple $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

EXEMPLE 1.3 Sur 21 parcelles de même aire, on mesure les rendements de trois variétés différentes de blé. Les résultats sont rassemblés dans le tableau suivant :

variété 1	3,4	14	18,3	15,5	26,9	12,4	20,8	20
variété 2	5,1	22	17	21,6	30,4	17,4		
variété 3	40,1	39,6	31,6	32,2	26,4	27,0	27,8	

On retrouve le modèle (1.4) avec $I = 3$, $n_1 = 8$, $n_2 = 6$ et $n_3 = 7$. Problématique : à partir de quand peut-on affirmer que $\mu_1 = \mu_2 = \mu_3$? Autrement dit, existe-t-il un effet 'variété' ? Cette information a-t-elle une importance ?

Déterminer si les facteurs ont une influence ou non revient donc (dans le cas de l'**aov** à un facteur et I modalités) à tester l'hypothèse :

$$\begin{aligned} H_0 : & \text{ le modèle vérifie } \mu_1 = \dots = \mu_I, \\ \text{contre } H_1 : & \text{ les moyennes sont différentes.} \end{aligned}$$

Avant de trancher entre ces deux hypothèses, il nous faut estimer les moyennes. Il est assez naturel d'utiliser :

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

Sous l'hypothèse H_0 , il ne nous reste qu'un seul 'grand' modèle :

$$Y_{ij} = \mu + \epsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, n_i. \quad (1.5)$$

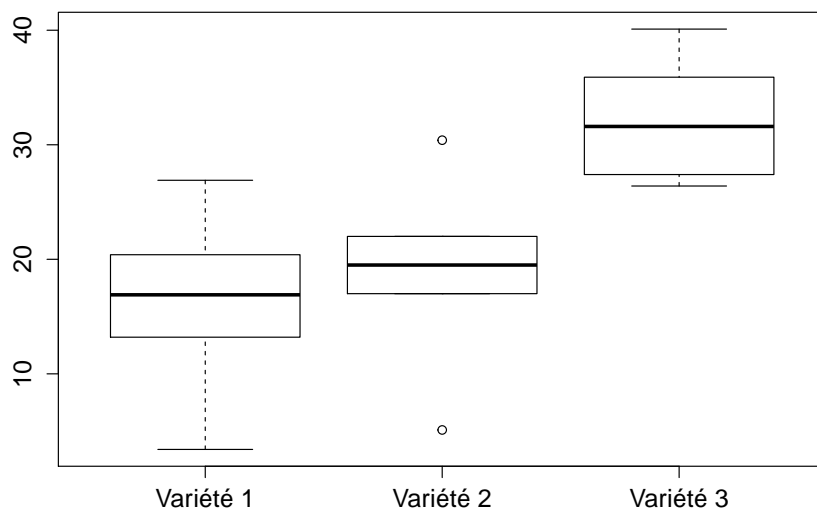


FIGURE 1.2 – Rendement en fonction de la variété de blé considérée.

Accepter ou rejeter H_0 revient donc à choisir un modèle parmi (1.4) et (1.5). On compare pour cela la somme des carrés résiduels : c'est une quantification de la part des observations non-prédites par chacun des modèles. Nous aborderons plus en détail la construction d'un test dans le Chapitre 2.

Remarque : l'analyse de variance à un facteur avec deux modalités correspond exactement au cadre de la comparaison de moyennes. Utiliser l'analyse de variance permet cependant d'aborder des situations bien plus variées.

Bien que le modèle soit plus compliqué à écrire, il est tout à fait envisageable suivant les situations de mettre en place une analyse de variance à plusieurs facteurs. Des secteurs comme l'agronomie sont très friands de ce type de modèle (cf Chapitre 5 pour plus de détails).

1.3 La régression logistique

Le modèle de régression logistique permet en quelque sorte de généraliser le principe de la régression linéaire à des variables binaires. Plus formellement, on dispose d'un vecteurs d'observations $(Y_1, \dots, Y_n)'$ où $Y_i \sim \text{Ber}(\pi_i)$ pour tout $i \in \{1, \dots, n\}$. L'idée est d'essayer de modéliser la variable Y en fonction d'un certain nombre de régresseurs $Z^{(1)}, \dots, Z^{(p)}$, chacun des $Z^{(i)}$ correspondant à un vecteur de taille n .

EXEMPLE. On s'intéresse à la vitesse de sédimentation des globules rouges dans le plasma sanguin. Cette vitesse est désignée sous l'acronyme ESR pour *erythrocytes sedimentation rate*. Pour un individu 'sain', cette vitesse est inférieure à 20 mm/h. Généralement, on ne s'intéresse pas à la valeur de l'ESR en elle-même. Il s'agit plutôt de savoir si cette dernière est inférieure à 20 (la 'réponse' sera 0 dans ce cas) ou supérieure à 20 (réponse égale à 1). Pour chaque individu, on met en relation cette variable qualitative avec le taux de *fibrinogen* qui est une protéine présente dans le plasma.

Dans ce contexte, rien n'empêche de réutiliser le principe de la régression linéaire introduit au

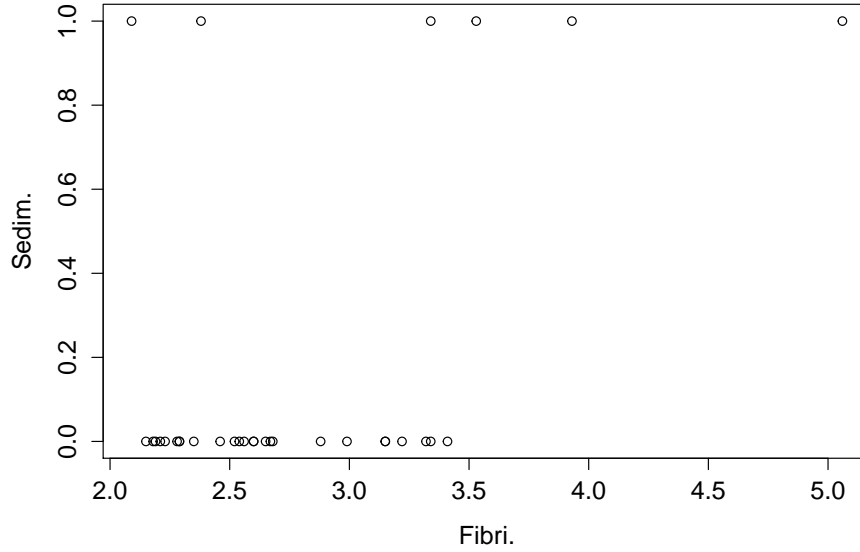


FIGURE 1.3 – Sédimentation en fonction de la concentration en fibrinogène du plasma sanguin.

tout début de ce chapitre, et d'utiliser le modèle

$$\mathbb{E}[Y_i] = \pi_i = a_1 Z_i^{(1)} + a_2 Z_i^{(2)} + \cdots + a_k Z_i^{(k)}, \quad i = 1, \dots, n,$$

Cependant, dans la mesure où l'on cherche à modéliser et prédire des probabilités, cette approche semble peu recommandée : certaines valeurs prédites pourraient en effet ne pas appartenir à l'intervalle $[0, 1]$.

Le modèle de **régression logistique** est lié à la relation

$$g(\pi_i) = a_1 Z_i^1 + \cdots + a_m Z_i^m, \quad \forall i \in \{1, \dots, n\},$$

où

$$g(x) = \log \left(\frac{x}{1-x} \right), \quad \forall x \in]0, 1[.$$

La fonction $g : (0, 1) \rightarrow \mathbb{R}$ est appelée fonction de lien.

De manière plus générale, il est possible d'envisager de considérer d'autres distributions pour la variable Y et d'autres fonctions de lien. À ce titre, on pourra remarquer que le modèle de régression abordé au début de ce chapitre correspond à une distribution gaussienne et une fonction de lien canonique ($g(x) = x$ pour tout $x \in \mathbb{R}$). Nous verrons qu'il est possible d'étudier tous ces modèles à travers un même cheminement.

1.4 Similitudes

Les différents modèles présentés ci-dessus semblent assez éloignés de part les objectifs visés et les méthodes employées. Nous allons voir qu'il est cependant possible de trouver des similitudes et d'unifier les deux formalismes.

Revenons en premier lieu sur le modèle de régression linéaire simple. Le modèle (1.1) peut être présenté sous la forme :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

ce qui peut s'écrire plus synthétiquement de la manière suivante :

$$Y = X\theta + \epsilon, \quad (1.6)$$

où

$$X = \begin{pmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{pmatrix}, \quad \theta = \begin{pmatrix} b \\ a \end{pmatrix},$$

et $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ représente le vecteur d'erreurs.

Un traitement analogue est possible pour la regression linéaire multiple en présence de m variables explicatives. Le vecteur d'observations $(Y_i)_{i=1\dots n}$ est alors représenté par :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1^{(1)} & \dots & Z_1^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & \dots & Z_n^{(m)} \end{pmatrix} \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

cette formule pouvant à nouveau être écrite sous la forme :

$$Y = X\theta + \epsilon, \quad (1.7)$$

avec

$$X = \begin{pmatrix} 1 & Z_1^{(1)} & \dots & Z_1^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & \dots & Z_n^{(m)} \end{pmatrix}, \quad \theta = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_m \end{pmatrix}.$$

On observe donc de grandes similitudes entre ces deux modèles de régression, il suffit pour cela de regarder (1.6) et (1.7).

Intéressons-nous à présent à l'analyse de variance. Afin de simplifier les notations, on se restreindra au cas de l'**aov** à un facteur et deux modalités. En utilisant (1.4), le vecteur d'observations peut, pour ce modèle, être ré-écrit sous la forme :

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}.$$

Sous forme matricielle, on obtient :

$$Y = X\theta + \epsilon, \quad (1.8)$$

avec cette fois-ci :

$$X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad \theta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

Une représentation similaire est envisageable pour l'analyse de variance à plusieurs facteurs.

On remarquera enfin que dans le cadre de la régression logistique, il est possible d'écrire que

$$g(\pi_i) = (X\theta)_i,$$

ou encore

$$\pi_i = \frac{\exp((X\theta)_i)}{1 + \exp((X\theta)_i)}.$$

1.5 Objectifs du cours

On voit donc bien que ces deux modèles de régression linéaire (simple ou multiple) et d'analyse de variance peuvent être rassemblés sous un même formalisme : on parle alors de **modèle linéaire**. Un cran au-dessus, on peut encore rassembler le modèle linéaire et par exemple la régression logistique sous une même bannière : le **modèle linéaire généralisé**. Les possibilités offertes par ces différentes modélisations ne s'arrêtent pas à une simple écriture commune. C'est en fait tout le traitement et l'exploitation des données qui peut être abordé de manière unifiée.

Au travers de ces modèles, nous aurons l'occasion de nous poser les questions suivantes :

- Comment choisir les variables les plus explicatives et mettre de côtés celles qui n'interviennent que très peu (voire pas du tout) dans l'explication des données ?
- Comment croiser variables qualitatives et quantitatives ?
- Est-il possible d'agir à la source sur la collecte des données et espérer ainsi améliorer la qualité d'estimation ?

1.6 Exercices

Exercice 1. We consider a classical Gaussian linear regression model :

$$Y_i = aZ_i + b + \epsilon_i, \quad i = 1, \dots, n.$$

Compute the maximum likelihood estimators of a , b and σ^2 . Compare these estimators to those obtained via the least square method (compute in particular the variance of the estimators of σ^2).

Exercice 2. We consider an ANOVA model with a single factor having two different modes, the error term being assumed to be Gaussian, with variance σ^2 .

1. Write the model in a matrix form.
2. Compute the maximum likelihood estimator.
3. Prove the efficiency of this estimator, i.e. that the associated covariance matrix is equal to the inverse of the Fisher information matrix.

Exercice 3. We consider a general framework of hypothesis testing, where one would like to test H_0 against H_1 from an i.i.d. sample $\mathbf{X} = (X_1, \dots, X_n)'$. For all $\alpha \in]0, 1[$, we consider a test of the form $\Psi_\alpha = \mathbf{1}_{\{T(\mathbf{X}) \in \mathcal{R}_\alpha\}}$, where $T : \mathbb{R}^n \rightarrow \mathbb{R}^p$ denotes a measurable function and $\mathcal{R}_\alpha \subset \mathbb{R}^p$ is such that $\mathbb{P}_{H_0}(\Psi_\alpha = 1) = \alpha$.

Définition 1.1 The p -value p_{val} associated to the familys $(\Psi_\alpha)_{\alpha \in]0, 1[}$ is defined as

$$p_{val} = \inf\{\alpha \in]0, 1[: T(\mathbf{X}) \in \mathcal{R}_\alpha\}.$$

It corresponds to the smallest level at which it is possible to reject H_0 using the sample \mathbf{X} .

The goal of the following questions is to provide a practical intuition to this p -value concept in a classical parametric setting.

1. We consider the case where $H_0 : \theta \in \Theta_0$ and where for all $\alpha \in]0, 1[$ the rejection region is of the form $\mathcal{R}_\alpha = [c_\alpha, +\infty[$, where $c_\alpha \in \mathbb{R}$. Prove that in such a case

$$p_{val} = \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta, \tilde{\mathbf{X}}} (T(\tilde{\mathbf{X}}) \geq T(\mathbf{x})),$$

where $\tilde{\mathbf{X}}$ denotes an i.i.d. copy of \mathbf{X} and $\mathbb{P}_{\theta, \tilde{\mathbf{X}}}$ is the associated measure.

2. Now, assume that $X_i \sim \mathcal{N}(\theta, 1)$. Our aim is to test the positivity of the parameter θ ($H_0 : \theta = 0$ et $H_1 : \theta > 0$). Soit $\alpha > 0$ et t_α tel que $\mathbb{P}_{H_0}(\bar{X}_n > t_\alpha) = \alpha$. Prove that

$$p_{val} \leq \alpha \Leftrightarrow \bar{X}_n \geq t_\alpha.$$

Deduce a decision rule based on the p -value.

3. What happens with the p -value in bilateral case?

Chapitre 2

Le modèle linéaire

2.1 Définition

Nous avons pu voir dans le chapitre précédent qu'il était possible d'utiliser la même représentation pour à la fois la régression linéaire et l'analyse de variance.

Définition 2.1 Une variable Y constituée de n observations Y_i suit un modèle linéaire statistique si Y peut être écrite sous la forme :

$$Y = X\theta + \epsilon, \quad (2.1)$$

où

- X est une matrice réelle à n lignes et p colonnes **avec $p < n$** ,
- θ est un vecteur réel inconnu de taille p ,
- le vecteur ϵ représente l'erreur du modèle.

Cette définition est très générale et dépasse largement le cadre de la régression et de l'analyse de variance. L'hypothèse $p < n$ signifie que le nombre d'observations doit être supérieur au nombre de paramètres à estimer. C'est en quelque sorte une hypothèse d'identifiabilité. Cette hypothèse sera relaxée dans le cours de *Statistique pour la grande dimension en génomique*.

Définition 2.2 Le modèle linéaire (2.1) est dit régulier si la matrice X est régulière, c'est-à-dire de rang p . Dans le cas contraire où X est de rang $r < p$, on parle de modèle singulier.

En particulier, si X est régulière alors $XC = 0 \Rightarrow C = 0$ pour tout $C \in \mathbb{R}^p$. Cette propriété assure que les colonnes de X sont linéairement indépendantes dans \mathbb{R}^n et garantit l'unicité de θ . Dans certaines situations, la matrice considérée X ne pourra être régulière. Nous verrons cependant (cf Section 2.4) qu'il est parfois possible de pallier à ce problème en rajoutant des contraintes dites d'identifiabilité sur les paramètres à estimer. A moins que cela ne soit mentionné explicitement, la matrice X sera supposée par la suite régulière.

Afin de pouvoir travailler plus simplement et d'aller plus loin dans l'étude de ce modèle, nous allons maintenant imposer quelques restrictions concernant le vecteur ϵ .

Hypothèse H1 : Les erreurs sont centrées, c.à.d. $\mathbb{E}[\epsilon] = \mathbf{0}$.

Remarquons que l'on note ici $\mathbf{0}$ pour le vecteur nul. Cette hypothèse est relativement importante et assure que le modèle est correctement défini. En effet, s'il s'avérait que $\mathbb{E}[\epsilon] \neq 0$, cela pourrait signifier qu'une partie de l'information n'a pas été prise en compte.

Hypothèse H2 : La variance des erreurs est constante, c.à.d. $\mathbb{E}[\epsilon_i^2] = \sigma^2, \forall i = 1, \dots, n$.

Il est souvent raisonnable de supposer que **H2** est bien vérifiée. Dans la situation où ce ne serait pas le cas, il est possible de mettre en place un traitement statistique du modèle linéaire... cela nécessite cependant bien plus de travail.

Hypothèse H3 : Les variables ϵ_i sont indépendantes.

Il existe un certain nombre de cas où ce postulat ne peut s'appliquer. On pourra par exemple penser aux séries temporelles : l'erreur du passé peut avoir une influence sur l'erreur future. Ces dernières font appel à un traitement statistique particulier (processus ARMA par exemple).

Dans la littérature statistique, un certain nombre de méthodes, souvent graphiques, sont proposées afin de vérifier la satisfaction des hypothèses **H1-H3**. Ces dernières seront abordées de manière très succincte en Section 2.3.

Afin de simplifier l'ensemble des calculs et résultats, nous supposerons parfois que les erreurs ϵ_i suivent une loi gaussienne centrée, de variance σ^2 . Les résultats associés à cette hypothèse seront dans ce cas signalés par une étoile. Encore une fois il est souvent possible de se passer de ce postulat... au prix d'un peu plus de travail.

2.2 Estimation

Nous allons à présent nous intéresser à l'estimation du vecteur θ . Comme dans le cadre de la régression linéaire simple, nous allons utiliser la méthode des moindres carrés. Il s'agit ici de trouver le vecteur $\hat{\theta}$ qui va minimiser la distance entre l'image de la matrice X et les observations Y . Autrement dit, l'estimateur de θ par la méthode des moindres carrés est défini par :

$$\hat{\theta} = \arg \min_{\vartheta \in \mathbb{R}^p} \|Y - X\vartheta\|^2. \quad (2.2)$$

La norme $\|\cdot\|$ est issue du produit scalaire usuel dans \mathbb{R}^n , i.e.

$$\|x\|^2 = \langle x, x \rangle = \sum_{k=1}^n x_k^2, \quad \forall x \in \mathbb{R}^n.$$

Sous forme matricielle, il est possible d'écrire :

$$\hat{\theta} = \arg \min_{\vartheta \in \mathbb{R}^p} (Y - X\vartheta)'(Y - X\vartheta). \quad (2.3)$$

Dans ce paragraphe, nous nous intéressons aux propriétés de cet estimateur.

Théorème 2.1 *Soit Y suivant un modèle linéaire. L'estimateur $\hat{\theta}$ obtenu par la méthode des moindres carrés est*

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

où A' désigne la transposée d'une matrice A quelconque.

PREUVE. On cherche dans un premier temps le vecteur $X\hat{\theta}$ appartenant au sous-espace vectoriel de \mathbb{R}^n engendré par les vecteurs colonnes de la matrice X (cet espace vectoriel sera noté par la suite $[X]$). On a :

$$\min_{\vartheta \in \mathbb{R}^p} \|Y - X\vartheta\|^2 = \min_{u \in [X]} \|Y - u\|^2 = \|Y - P_{[X]}Y\|^2,$$

où $P_{[X]}$ désigne la projection orthogonale de \mathbb{R}^n sur $[X]$. Ainsi, pour tout $i \in \{1, \dots, n\}$, en désignant par $X^{(i)}$ le i -ème vecteur colonne de la matrice X , on obtient :

$$\langle X^{(i)}, Y \rangle = \langle X^{(i)}, P_{[X]}Y \rangle = \langle X^{(i)}, X\hat{\theta} \rangle.$$

Ceci revient à écrire :

$$X'Y = X'X\hat{\theta} = P_{[X]}Y \quad (2.4)$$

et donc

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

□

Ce premier théorème nous donne donc une formule explicite pour l'estimateur du vecteur θ par la méthode des moindres carrés. Il est intéressant de noter que cette dernière est purement géométrique et ne demande aucune connaissance de la loi des erreurs.

Remarque : Dans le cas particulier où les erreurs sont gaussiennes, l'estimateur des moindres carrés $\hat{\theta}$ correspond exactement à l'estimateur du maximum de vraisemblance.

Le résultat suivant explicite les performances de l'estimateur des moindres carrés.

Théorème 2.2 *Soit Y suivant un modèle linéaire et $\hat{\theta}$ l'estimateur par la méthode des moindres carrés défini en (2.2). Alors*

$$\mathbb{E}[\hat{\theta}] = \theta, \quad \text{Var}(\hat{\theta}) = \sigma^2(X'X)^{-1}, \quad \text{et} \quad \mathbb{E}\|\hat{\theta} - \theta\|^2 = \sigma^2 \text{Trace}((X'X)^{-1}).$$

De plus, si les variables ϵ_i sont i.i.d, gaussiennes centrées réduites, $\hat{\theta}$ est le meilleur estimateur parmi tous les estimateurs sans biais de θ , i.e.

$$\text{Var}(C'\tilde{\theta}) \geq \text{Var}(C'\hat{\theta}),$$

pour tout estimateur $\tilde{\theta}$ et tout vecteur $C'\theta$.

PREUVE. Les égalités sur l'espérance et la variance sont obtenues à partir de la définition de l'estimateur. Le résultat d'optimalité requiert un peu plus de travail et ne sera pas développé ici.

□

Le dernier résultat de ce paragraphe s'intéresse à l'estimation du 'niveau de bruit' σ^2 .

Théorème 2.3* *Soit $\hat{\theta}$ l'estimateur de θ par la méthode des moindres carrés. Si les erreurs sont gaussiennes,*

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\theta}\|^2}{n - p} = \frac{\|Y - \bar{Y}\|^2}{n - p}.$$

est un estimateur sans biais (optimal) de σ^2 .

PREUVE. Remarquons dans un premier temps que :

$$\|Y - X\hat{\theta}\|^2 = \|Y - P_{[X]}Y\|^2 = \|X\theta + \epsilon - X\theta - P_{[X]}\epsilon\|^2 = \|P_{[X^\perp]}\epsilon\|^2.$$

La matrice X étant supposée régulière, le sous espace vectoriel $[X^\perp]$ est donc de dimension $n - p$. Le vecteur ϵ étant composé de variables aléatoires gaussiennes indépendantes, de moyenne 0 et de variance σ^2 , la variable $\sigma^{-2}\|Y - X\hat{\theta}\|^2$ suit donc une loi du χ^2 à $n - p$ degrés de liberté (voir Théorème de Cochran en Annexe). Il en découle naturellement que $\hat{\sigma}^2$ est un estimateur sans biais de σ^2 .

□

Remarque : Le terme $\|Y - X\hat{\theta}\|^2$ s'appelle somme des carrés résiduelles. Dans le cas particulier où ϵ désigne un vecteur gaussien, nous avons pu voir dans la preuve que cette variable suit (à une constante σ^2 près) une loi du χ^2 à $n - p$ degrés de liberté.

2.3 Validation (graphique) des hypothèses du modèle

Les résultats présentés ci-dessus, mais surtout les suivants, dépendent énormément de la validité des hypothèses introduites dans la Section 2.1. Avant d'entamer toute démarche additionnelle dans la compréhension du modèle utilisé, il convient de s'assurer que ces hypothèses sont bien satisfaites. On utilisera pour cela les résidus $(E_i)_{i \in 1, \dots, n}$ définis comme

$$E_i = Y_i - \hat{Y}_i, \quad \forall i \in \{1, \dots, n\}.$$

Les résidus 'mesurent' l'écart entre nos prédictions et les observations et peuvent s'avérer précieux pour vérifier la qualité du modèle. On pourra vérifier (en exercice) que la variance de ces résidus n'est pas constante. Afin de pouvoir les comparer entre eux, il est parfois préférable d'utiliser les résidus standardisés :

$$R_i = \frac{E_i}{\sqrt{\text{Var}(E_i)}}, \quad \forall i \in \{1, \dots, n\}.$$

Dans la suite de cette section, on donne une liste (non-exhaustive) de quelques outils permettant un premier diagnostic concernant la pertinence d'un modèle.

- Erreurs centrée (**H1**) : on trace communément le graphique des résidus (standardisés) en fonction des prédictions. Si **H1** est vérifiée, aucune tendance particulière ne doit apparaître. Si ce n'est pas le cas, c'est probablement que la composante déterministe du modèle n'est pas appropriée. On pourra remédier à ce problème en modifiant cette dernière (voir Figure 2.3 pour une illustration sur données simulée).

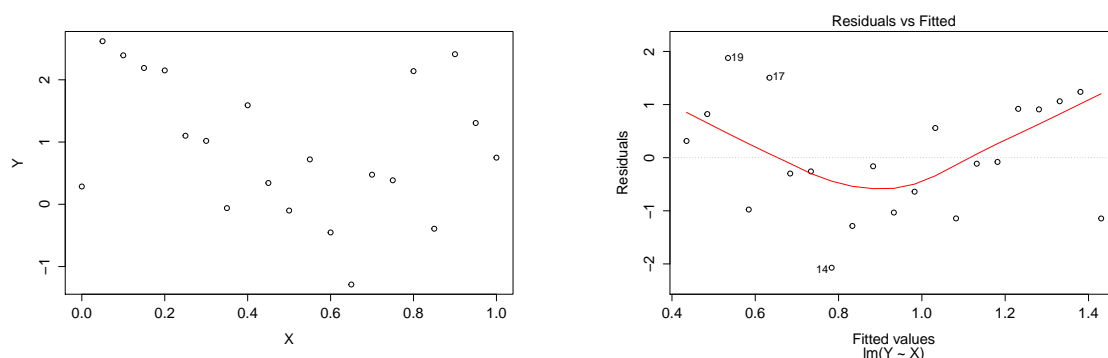


FIGURE 2.1 – Regression linéaire d'une variable Y en fonction d'un prédicteur X . Une tendance quadratique (ou autre) a probablement été manquée dans la modélisation.

- Variance des erreurs constante - homoscedasticité (**H2**) : on peut à nouveau se tourner vers le graphe des résidus afin de détecter une évolution de la variance des erreurs, comme illustré dans la Figure 2.3 ci-dessous.

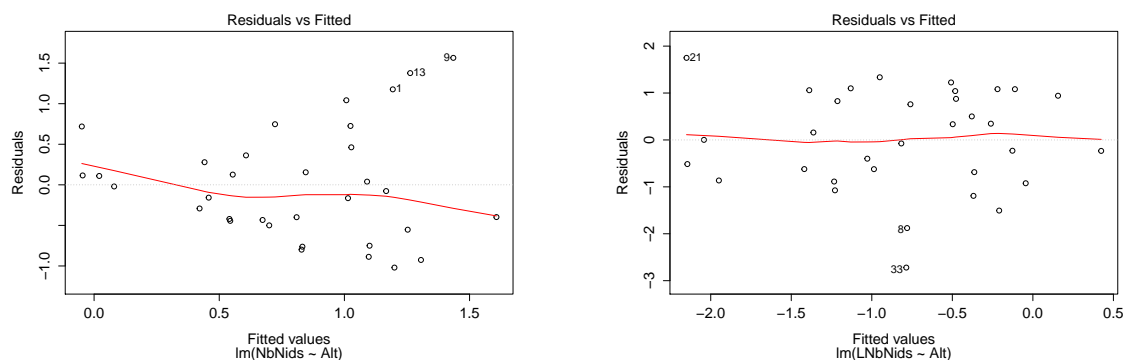


FIGURE 2.2 – Retour aux chenilles. Sur la figure de gauche, on explique le l'ombre de nids par l'altitude, ce qui donne lieu à une forte hétéroscédasticité. A droite, ce 'problème' est réglé en passant au log du nombre de nids.

- Indépendance des erreurs (**H3**) : Même si il existe quelques tests permettant de poser cette question d'un point de vue statistique, cette hypothèse est la plupart du temps liée au plan d'expérience utilisé, i.e. à la manière dont ont été collectées les données.

- Caractère gaussien des erreurs : Il existe de nombreux tests permettant de décider si un échantillon donné suit une loi gaussienne (on pourra par exemple penser au test de Kolmogorov-Smirnov)... sauf que la plupart du temps, les données de l'échantillon considéré sont supposées indépendantes. Or, même si **(H3)** est vérifiée, les erreurs n'étant pas observables seront approximées par les résidus... qui sont tout sauf indépendants. Bien souvent, on se contente donc de comparer les quantiles empiriques des résidus aux quantiles théoriques de la loi de Student, en utilisant la représentation de la droite de Henri.

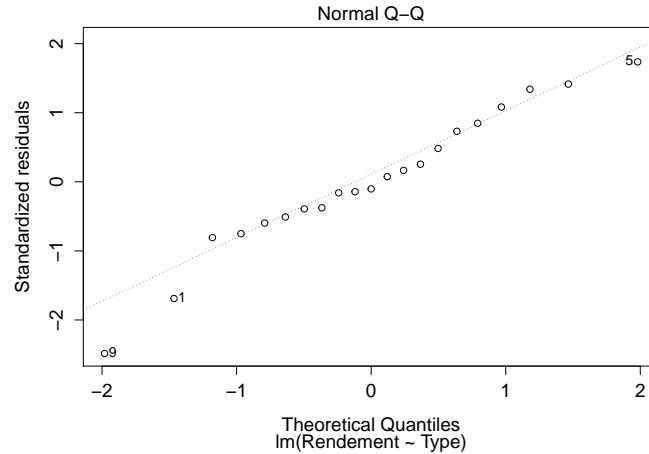


FIGURE 2.3 – Droite de Henri pour un modèle d'analyse de variance à un facteur : données sur le rendement du blé présentée dans la Section 1.2.

2.4 Tests

Nous allons voir dans ce paragraphe un certain nombre de tests pouvant être mis en oeuvre sur le modèle linéaire. Ces tests peuvent être utilisés aussi bien en régression linéaire qu'en analyse de variance. Nous supposerons cependant pendant toute cette partie que les ϵ_i sont des variables i.i.d. gaussiennes centrées réduites. Les tests présentés ci-dessous ne peuvent être utilisés si cette contrainte n'est pas satisfaite.

2.4.1 Test sur la présence d'un sous-modèle

Cette approche vise à déterminer si le modèle utilisé peut être oui ou non simplifié. Plus formellement, soit X_0 une matrice de rang $p_0 < p$, telle que $[X_0] \subset [X]$. Le modèle :

$$Y = X_0 \theta_0 + \epsilon, \quad (2.5)$$

est appelé **sous-modèle** issu du modèle linéaire défini en (2.1). Il peut être parfois intéressant d'essayer de savoir si les observations sont issues du modèle (2.1) ou (2.5). Soit le modèle défini par :

$$Y = S + \epsilon.$$

Tester la présence d'un sous-modèle revient donc à tester :

$$\begin{aligned} H_0 &: S \in [X_0] \\ \text{contre } H_1 &: S \in [X] \end{aligned}$$

EXEMPLE 2.1 Dans le modèle de régression linéaire simple, pour tester la nullité de la pente, on peut considérer le sous-modèle $Y_i = b + \epsilon_i$, $i = 1 \dots n$ avec $\theta_0 = b$ et $X_0 = (1 \dots 1)'$.

EXEMPLE 2.2 Reprenons l'exemple 1.3. Dans la situation où le facteur "cours d'eau" n'a aucune

influence sur la taille des poissons, le modèle devient $Y_{ij} = \mu + \epsilon_{ij}$ i.e. $\theta_0 = \mu$ et $X = (1 \dots 1)'$.

Avant de proposer un test permettant de répondre à ces problématiques, quelques notations. On note $\hat{\theta}_0$ l'estimateur des moindres carrés issu du modèle (2.5). On définit également

$$SCR = \|Y - X\hat{\theta}\|^2, \text{ et } SCR_0 = \|Y - X_0\hat{\theta}_0\|^2,$$

les sommes des carrés résiduels associées à chaque modèle. Dans la mesure où $[X_0] \subset [X]$ et par définition de $\hat{\theta}$, on a $SCR_0 \geq SCR$.

Lemme 2.1* *Sous l'hypothèse nulle H_0 , la variable*

$$\hat{F} = \frac{n-p}{p-p_0} \frac{SCR_0 - SCR}{SCR},$$

suit une loi de Fisher de paramètres $(p-p_0, n-p)$.

PREUVE. En utilisant le Théorème de Pythagore, on a

$$SCR_0 = SCR + \|P_{[X_0]}Y - P_{[X]}Y\|^2, \text{ et } SCR_0 - SCR = \|P_{[X_0]}Y - P_{[X]}Y\|^2.$$

Par le Théorème de Cochran (cf Annexe), les variables $SCR_0 - SCR$ et SCR sont indépendantes et sous l'hypothèse H_0 , $SCR_0 - SCR$ suit (à une constante σ^2 près) une loi du χ^2 à $k - k_0$ degrés de liberté.

□

La quantité d'importance dans notre construction du test est $SCR_0 - SCR$. Intuitivement, si la valeur observée de $SCR_0 - SCR$ est très grande, il y a peu de chance que les observations Y soient 'issues' du sous-modèle. A l'opposé, si la valeur observée $SCR_0 - SCR$ est petite, il est fort possible que le modèle initial puisse être simplifié : le sous-modèle explique aussi bien les observations dans la mesure où SCR_0 est comparable à SCR . Ceci nous conduit à la règle de décision suivante :

- si $\hat{F} > t$, on rejette H_0 ,
- si $\hat{F} \leq t$, on accepte H_0 .

Il reste alors à choisir t de telle sorte à contrôler l'erreur de première espèce. Plus directement, on peut calculer la p-valeur :

$$P_{val} = \mathbb{P}(\hat{F} > \hat{f}),$$

où \hat{f} désigne la valeur observée de \hat{F} . Si la p-valeur est inférieure à l'erreur de première espèce fixée par le commanditaire du test, on rejette H_0 , et inversement.

2.4.2 Test sur la nullité d'une combinaison linéaire

Soit C un vecteur donné de \mathbb{R}^k . On s'intéresse ici à la nullité de la combinaison linéaire $C'\theta$. On souhaite donc tester l'hypothèse :

$$\begin{array}{ll} H_0 & : C'\theta = 0 \\ \text{contre } H_1 & : C'\theta \neq 0. \end{array}$$

Lemme 2.2* *Sous l'hypothèse nulle H_0 , la variable \hat{T} définie par :*

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

suit une loi de Student de paramètre $n-p$.

PREUVE. La démonstration utilise le postulat de gaussianité des erreurs.

□

Ce résultat nous permet de proposer un test adapté à la situation. Si $|C'\theta|$ (ou de manière équivalente $|\hat{T}|$) prend de grandes valeurs, il y a peu de chances que cette combinaison linéaire soit nulle (et réciproquement). Comme pour le test sur la présence d'un sous-modèle, on pourra donc s'intéresser à la p-valeur :

$$P_{val} = \mathbb{P}(|\hat{T}| \geq |\hat{t}|),$$

où \hat{t} représente la valeur observée de \hat{T} .

2.4.3 Test sur la nullité jointe de plusieurs combinaisons linéaires

Soit maintenant C une matrice de taille $k \times m$ pour $m \in \mathbb{N}$. La quantité m correspond au nombre de combinaisons linéaires à tester. On souhaite maintenant tester les hypothèses :

$$\begin{array}{ll} H_0 & : C'\theta = 0 \\ \text{contre } H_1 & : C'\theta \neq 0 \end{array}$$

où 0 désigne ici le vecteur nul.

EXEMPLE : Ce type de test est particulièrement utile en analyse de variance et analyse de covariance (chapitre suivant). Considérons par exemple l'analyse de variance en présence d'un facteur à quatre modalités μ_1, μ_2, μ_3 et μ_4 . Si on souhaite tester l'hypothèse ' $\mu_1 = \mu_2$ et $\mu_3 = \mu_4$ ', on utilise :

$$C' = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \text{ et ainsi } C'\theta = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_3 - \mu_4 \end{pmatrix}.$$

Nous aurons besoin du lemme suivant pour mettre en place un test.

Lemme 2.3* *Sous l'hypothèse nulle H_0 , la variable \hat{F} définie par :*

$$\hat{F} = \frac{\hat{\theta}' C (C' (X' X)^{-1} C)^{-1} C' \hat{\theta}}{m \hat{\sigma}^2},$$

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

La mise en place d'un test permettant de trancher entre les deux hypothèses H_0 et H_1 suit les mêmes grandes lignes que pour les paragraphes précédents.

2.4.4 Méthodologie

A ce stade, il semble important de rappeler quelques éléments méthodologiques associés à la théorie des tests.

De manière générale, le statisticien aime bien rejeter l'hypothèse H_0 : cela signifie que non seulement on a mis de l'information en évidence mais également que la probabilité de se tromper est assez faible (contrôlée par l'erreur de première espèce). L'hypothèse H_0 est en effet souvent conservative au sens où elle fait appel à des modèles assez simples.

Le fait de ne pas rejeter H_0 est beaucoup plus problématique dans la mesure où l'erreur de seconde espèce n'est presque jamais calculable : seule la fonction puissance est éventuellement disponible. La question naturelle à se poser est donc de savoir si le non rejet de H_0 est dû à la réalité de cette hypothèse... ou l'absence insuffisante d'informations pour pouvoir mettre en évidence H_1 . Dans ce contexte, de nombreux statisticiens préfèrent *ne pas rejeter H_0* plutôt que *l'accepter*.

Un autre point important concerne l'interprétation éventuelle pouvant être mise en place suite au non-rejet de H_0 . Par exemple, si on se place dans un contexte de régression linéaire simple et on s'intéresse à la valeur de la pente $H_0 : a = 0$. Accepter l'alternative revient à dire que la pente a de la droite n'est pas nulle. La variable de réponse et celle de régression sont donc corrélées. Il ne faut cependant pas confondre *corrélacion* et *relation de cause à effet*.

2.4.5 La problématique des tests multiples

Dans un certain nombre d'applications, il est assez courant de devoir effectuer en parallèle une série de tests. On peut par exemple penser aux domaines suivants

- Génomique : on étudie conjointement l'activation éventuelle d'un ensemble de gènes pour une condition expérimentale donnée.
- Médecine : on test si un panel donnée de substances sont nocives pour l'organisme.
- Neurologie : on s'intéresse à l'activation de différentes zones du cerveau pour une tâche donnée.
- ...

Formellement, on considère que l'on travaille avec une série d'échantillons $(\mathcal{S}_{j=1}^N)$ où pour tout $j \in \{1, \dots, N\}$, $\mathcal{S}_j = (X_1^j, \dots, X_{n_j}^j)$ avec $n_j \in \mathbb{N}$. Pour chaque échantillon, on souhaite tester H_0^j contre H_1^j . On suppose que l'on dispose pour chaque j d'un test $\psi_{\alpha,j}$ avec la convention

- Si $\psi_{\alpha,j} = 1$, on rejette H_0^j ,
- Si $\psi_{\alpha,j} = 0$, on ne rejette pas H_0^j .

Si les tests considérés sont de niveau α , alors pour un j fixé, l'erreur de première espèce, i.e. la probabilité de rejeter H_0^j à tort est inférieure à α . La probabilité de prendre une mauvaise décision (en décidant de rejeter H_0^j est donc parfaitement contrôlée. Les choses sont cependant beaucoup plus mitigées si on se place dans un cadre global.

Notons

$$H_0 = \bigcap_{j=1}^N H_0^j,$$

l'hypothèse pour laquelle tous les H_0^j sont vraies. Avec cette notation, en supposant que les événements $\{\psi_{\alpha,j} = 1\}$ sont disjoints deux-à-deux, et que les tests considérés sont exactement de niveau α , on obtient

$$\mathbb{P}_{H_0} \left(\bigcup_{j=1}^N \{\psi_{\alpha,j} = 1\} \right) = \sum_{j=1}^N \mathbb{P}_{H_0^j}(\psi_{\alpha,j} = 1) = \alpha N.$$

Autrement dit, si toutes les hypothèses H_0 sont vraies, la probabilité de rejeter au moins une de ces dernières à tort n'est non plus égale à α mais αN . Cette erreur peut-être extrêmement grande dès lors que N est grand.

Il existe une manière simple de contourner ce problème. On pose

$$\psi_{\alpha}^* = \max_{j=1..N} \psi_{\alpha/N,j}.$$

D'une certaine manière, on agrège les tests dont nous disposons, mais en modifiant le niveau : α est remplacé par α/N . On parle de **correction de Bonferroni**. Il est dans ce cas facile de voir que

$$\mathbb{P}_{H_0}(\psi_{\alpha}^* = 1) = \mathbb{P}_{H_0} \left(\bigcup_{j=1}^N \{\psi_{\alpha/N,j} = 1\} \right) \leq \sum_{j=1}^N \mathbb{P}_{H_0^j}(\psi_{\alpha/N,j} = 1) \leq \sum_{j=1}^N \frac{\alpha}{N} = \alpha.$$

Cette correction est très facile à mettre en oeuvre et peut être utilisée dans un très grand nombre de situations. D'un point de vue pratique, en particulier pour les applications mentionnées ci-dessus, elle est s'avère parfois trop conservatrice. Dans ce cas, d'autres stratégies doivent être mises en oeuvre.

2.5 Modèles singuliers

Nous nous sommes jusqu'à présent cantonnés à l'étude des modèles linéaires réguliers. Il existe cependant un certain nombre de situations pour lesquelles la matrice X est de rang inférieur à p .

EXEMPLE. On considère un modèle d'analyse de variance à 1 facteur et J modalités, pour lequel une seule observation est disponible pour chaque modalité :

$$Y_i = \mu_i + \epsilon_i, \quad i = 1 \dots J. \quad (2.6)$$

Sous cette forme, la matrice X correspondante est régulière. Supposons maintenant que l'on modifie légèrement le modèle ci-dessus en introduisant un effet fixe μ :

$$Y_i = \mu + \alpha_i + \epsilon_i, \quad i = 1 \dots I. \quad (2.7)$$

La suite $(\alpha_i)_i$ correspond à l'effet différentiel, propre à chaque modalité. Les différences entre les deux modèles (2.6) et (2.7) semblent minimes, pourtant, dans ce dernier cas, la matrice X n'est

plus régulière... Le modèle est en fait **surparamétré** : nous avons $I + 1$ paramètres inconnus pour seulement I observations.

La matrice $X'X$ n'est pas inversible si X n'est pas régulière. Pour contourner ce problème, on définit alors un inverse généralisé de $(X'X)$, i.e une application $(X'X)^-$ telle que

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

où I désigne la matrice identité. Cette construction est toujours possible. En effet, $(X'X)$ définit une application bijective de $\text{Ker}(X)^\perp$ sur lui-même. Il suffit donc simplement de négliger la partie contenue dans le noyau : on prend l'inverse sur $\text{Ker}(X)^\perp$, complété arbitrairement sur $\text{Ker}(X)$. La définition de $(X'X)^-$ est donc loin d'être unique ! Il est alors possible de généraliser les résultats du cas régulier en utilisant :

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

Cet estimateur n'est pas unique et dépend de la définition choisie pour $(X'X)^-$.

Remarque : Le vecteur $X\hat{\theta}$ reste par contre unique, même si la matrice X est singulière. Ce vecteur correspond en effet à la projection orthogonale de Y sur $[X]$.

En règle générale, on préfère lever l'indétermination sur $\hat{\theta}$ en fixant des contraintes, souvent afin de donner un sens plus intuitif à θ .

Proposition 2.1 *Supposons la matrice X singulière, de rang $r < p$. Soit H une matrice à $p - r$ lignes et p colonnes, supposée de rang $p - r$ et telle que : $\text{Ker}(H) \cap \text{Ker}(X) = \{0\}$. Alors, la matrice $(X'X + H'H)$ est inversible, et l'estimateur $\hat{\theta} = ((X'X + H'H)^{-1} X'Y)$ est l'unique solution du système :*

$$\begin{cases} X\theta &= P_{[X]}Y \\ H\theta &= 0. \end{cases}$$

PREUVE. Soit G la matrice $(n + p - r) \times p$ définie comme

$$G = \begin{pmatrix} X \\ H \end{pmatrix}.$$

Par construction, cette matrice est de rang $r + p - r$ et donc la matrice

$$G'G = (X' \ H') \begin{pmatrix} X \\ H \end{pmatrix} = X'X + H'H,$$

est inversible. Considérons le système d'équations

$$\begin{pmatrix} X \\ H \end{pmatrix} \theta = G\theta = \begin{pmatrix} P_{[X]}Y \\ 0 \end{pmatrix}$$

La solution $\hat{\theta}$ de ce dernier vérifie

$$G'G\theta = G' \begin{pmatrix} P_{[X]}Y \\ 0 \end{pmatrix} = X'P_{[X]}Y = X'Y,$$

où pour la dernière égalité, on a utilisé (2.4). La matrice $G'G$ étant inversible, on obtient comme solution du système $\hat{\theta} = (G'G)^{-1} X'Y = (X'X + H'H)^{-1} X'Y$.

□

Le choix de la contrainte n'est pas toujours évident. Par ailleurs, pour chaque contrainte H , on aura un estimateur et des résultats de tests correspondants : ceci peut s'avérer gênant d'un point de vue pratique. A noter que ces contraintes ne sont pas toujours exprimées de manière transparente sur les logiciels couramment utilisés en statistique, et peuvent différer d'un logiciel à l'autre (comme par exemple entre R et SAS).

EXEMPLE. Reprenons l'exemple d'analyse de variance à un facteur avec effet différentiel : on suppose pour simplifier que $I = 4$. La matrice X associée au modèle est :

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

Il nous faut poser une contrainte (dite d'identifiabilité) sur le vecteur θ au travers du choix d'une matrice à 1 ligne et k colonnes. Une possibilité est de considérer $H = (0 \ 1 \ 1 \ 1 \ 1)'$. La contrainte correspondante est :

$$H\theta = 0 \Leftrightarrow \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0.$$

on impose donc que la somme des effets différentiels soit nulle. On peut vérifier que les conditions de la proposition précédente sont bien satisfaites : l'estimateur suggéré peut alors être construit.

Exercice : construire l'estimateur $\hat{\theta}$ obtenu par la méthode des moindres carrés sous la contrainte $H\theta = 0$.

En présence de matrice singulières, il est donc toujours possible de construire un estimateur. Qu'en est-il des tests ? En particulier, ces contraintes sont-elles systématiquement nécessaires ?

La plupart des quantités que nous avons voulu tester ne sont pas 'directement' liées à l'estimateur $\hat{\theta}$: elles sont indépendantes des contraintes d'identifiabilité choisies. La définition suivante formalise cette notion.

Définition 2.3 Une combinaison linéaire $C'\theta$ est dite estimable (de paramètre θ) si il existe une matrice D ($n \times k$) de rang k telle que $C'\theta = D'X\theta$.

Autrement dit, une fonction estimable est indépendante des contraintes posées sur θ .

Définition 2.4 On appelle contraste une fonction estimable $C'\theta$ telle que $C'1 = 0$, où 1 désigne le vecteur unité.

En analyse de variance, la plupart des combinaisons linéaires que l'on teste sont en fait des contrastes.

2.6 Exercises

Exercise 1. A psychologist conducts a study over 4 children and is interested in the link between the time spends watching television and concentration abilities. The results are the following

TV hours number	Concentration score
5	8
20	12
8	6
2	4

1. We consider a classical linear regression model. Compute the value of the MLE on this sample.
2. Draw an approximative picture and the associated regression curve.
3. According to this sample, one would like to know if the attention score might be correlated to the time spend in front of a TV. What hypotheses should be considered ? Construct a test that will answer to this question.
4. What is the decision associated to this test ? What can we conclude ?

Exercise 2. We deal here with a dataset containing data on the pine's caterpillar. Conduct an analysis of this sample (we are interested in the link between the elevation and the number of nests. Provide a confidence band on the predicted values (for instance thanks to the command `predict`).

Exercise 3. Provide an analysis on the data contained in the file `television.dat`¹ linking life expectation and number of TV per resident. For further investigations, one might take into account the gender of the residents (using, e.g., the commands `anova`, `Anova` and `summary`).

Exercise 4. One would like to analyse the PH values of 4 different ecological fields in which 5 different measures have been obtained :

Milieu 1	4,94	3,92	3,36	3,93	4,50
Milieu 2	4,13	4,18	4,37	5,41	4,82
Milieu 3	3,11	2,35	2,90	3,75	3,82
Milieu 4	6,06	6,65	7,57	7,58	7,79

Construct a boxplot for each field. Provide an analysis of these data. Can we say that all these fields have a similar PH?

Exercise 5. One would like to compare the results of three different groups to a given test. Before this test, a pretest has been proposed to the participants. In particular, we have in mind that the results to the pretest might predict the result to the test of interest. The data, available in a book by J.Kennedy and A.Bush (1985), are gathered below

Group 1	Group 2	Group 3
Prétest Test	Prétest Test	Prétest Test
12 34	18 35	10 28
6 26	8 30	4 22
9 33	16 37	10 24
13 35	5 28	17 29
12 34	9 31	9 27
10 33	8 30	7 22

1. Draw an approximative picture of this dataset (distinguish the group using for instance different symbols).
2. For each group, provide an estimation of the parameters of the linear regression curve.
3. Propose a model for this experience and make precise the structure of the parameter θ . What kind of question might be interesting in this context?
4. Provide a complete analysis of this dataset.

1. a description of this dataset is provided in the file `television.txt`.

Chapitre 3

La sélection de modèle

Dans les chapitres précédents, nous nous sommes intéressés à l'estimation du vecteur inconnu θ ainsi qu'à quelques tests permettant d'en savoir un peu plus sur cette quantité. Nous allons maintenant adopter une démarche un peu différente et nous concentrer sur l'étude de la matrice X autrement dit sur les variables explicatives elles-mêmes.

Dans ce chapitre, nous allons voir comment choisir le modèle le plus en adéquation avec nos données et éliminer certaines variables peu explicatives.

3.1 Cadre général

Une des premières questions qui vient à l'esprit lorsque l'on essaye de définir un modèle est : les variables explicatives retenues sont elles réellement explicatives ? Nous amènent-elles à mieux comprendre le modèle que nous sommes en train d'étudier ? Ces questions prennent tout leur sens lorsque l'on étudie des modèles complexes pour lesquels un grand nombre de facteurs potentiels est disponible (génétique, épidémiologie, météo, etc...). Nous allons voir ici un certain nombre d'approches permettant d'affiner (de sélectionner) un modèle parmi d'autres, c'est-à-dire de déterminer quelles sont les variables les plus 'explicatives', importantes dans notre compréhension du modèle.

Par soucis de simplicité, on se place dans ce chapitre dans le cadre de la régression linéaire multiple. Les outils présentés ici peuvent bien sûr être utilisés dans un cadre plus général (bien souvent sans travail supplémentaire).

On dispose d'un échantillon de taille n représentant des observations sur une variable à expliquer Y et de p variables explicatives réelles $Z^{(i)}$. On se donne une famille de modèles \mathcal{M} représentant formellement une famille de sous-ensembles de $\{1, \dots, p\}$. Ce choix est fait a priori et peut ne pas être exhaustif.

EXEMPLES :

- famille exhaustive : $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$ i.e. la famille de tous les sous-ensembles de $\{1, \dots, p\}$,
- famille croissante : $\mathcal{M} = (\{1, \dots, m\})_{m=1..p}$.

Pour tout $i \in \{1, \dots, p\}$, on note $Z^{(i)}$ le vecteur constitué des n réalisations de la i -ème variable explicative et pour tout $m \in \mathcal{M}$, $X_{(m)}$ représente la matrice constituée des vecteurs $Z^{(i)}$ pour $i \in m$:

$$X_{(m)} = \begin{pmatrix} 1 & Z_1^{(i_1)} & Z_1^{(i_2)} & \dots & Z_1^{(i_m)} \\ 1 & Z_2^{(i_1)} & Z_2^{(i_2)} & \dots & Z_2^{(i_m)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_m^{(i_1)} & Z_m^{(i_2)} & \dots & Z_m^{(i_m)} \end{pmatrix},$$

si m est de la forme $m = \{i_1, \dots, i_m\}$.

Il nous faut donc choisir $m \in \mathcal{M}$, ou de manière équivalente sélectionner parmi tous les modèles :

$$Y = X_{(m)}\theta_{(m)} + \epsilon, \quad m \in \mathcal{M}, \quad (3.1)$$

avec $\theta_{(m)} \in \mathbb{R}^{|m|+1}$ et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, celui qui s'approche le plus (dans un sens à préciser) du modèle sous-jacent, i.e. celui qui est à l'origine de ce que nous observons. On écrira ce modèle :

$$Y = X_{(m^*)}\theta_{(m^*)} + \epsilon, \quad (3.2)$$

avec $m^* \in \mathcal{M}$, $\theta_{(m^*)} \in \mathbb{R}^{|m^*|+1}$.

Remarque : Les coordonnées de $\theta_{(m^*)}$ sont toutes **supposées non-nulles**. Ceci permet de rendre le modèle m^* unique.

Nous allons voir dans ce chapitre diverses approches permettant, non pas de retrouver m^* , mais au moins de s'en approcher. Ceci correspond aux bases de la **sélection de modèle**.

3.2 Une première approche

3.2.1 Les coefficients d'ajustement

Dans la situation où seul un petit nombre de régresseurs est en jeu, il existe déjà un certain nombre d'approches s'inspirant plus ou moins directement des outils étudiés dans les Chapitres 1 et 2. Pour 'tester' la validité d'un sous-modèle m par rapport à un modèle plus grand, il existe deux indices (ou coefficients) dont le calcul et l'interprétation sont assez immédiats.

Une première possibilité consiste à s'intéresser au coefficient de détermination :

$$R_m^2 = \frac{SCT - SCR}{SCT}, \quad \text{avec } SCT = \|Y - \bar{Y}\|^2 \text{ et } SCR = \|Y - X_m \hat{\theta}_{(m)}\|^2.$$

On peut en particulier écrire que :

$$R_m^2 = 1 - \frac{\|Y - X_m \hat{\theta}_{(m)}\|^2}{\|Y - \bar{Y}\|^2}.$$

Cet indice compare donc les valeurs prédites de Y aux valeurs observées par l'intermédiaire de $\|\hat{Y}_{(m)} - Y\|^2$, le dénominateur correspondant en quelque sorte à une renormalisation. Plus le coefficient R^2 sera proche de 1, plus l'adéquation aux données du modèle retenu sera importante. Si on est amené à choisir entre deux modèles explicatifs, on est donc facilement tentés de retenir celui possédant le coefficient de détermination le plus important.

Il est cependant important d'apporter un petit bémol à ce type de raisonnement. Utiliser ce type de critère favorise en effet l'utilisation de modèles très paramétrés : plus on utilise de variables explicatives, plus l'adéquation augmentera. Ceci peut être intéressant d'un point de vue mathématique mais poser un problème pour une éventuelle interprétation du modèle retenu.

Le coefficient de détermination ajusté (noté \tilde{R}^2) permet de tenir compte du nombre de régresseurs retenus et propose donc un compromis entre l'adéquation et le paramétrage du modèle. Cet indice est défini par :

$$\tilde{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SCR}{SCT} = 1 - \frac{n-1}{n-k-1} \cdot \frac{\|Y - X_m \hat{\theta}_{(m)}\|^2}{\|Y - \bar{Y}\|^2}.$$

L'interprétation est similaire à celle du R^2 .

3.2.2 Stratégies de régression ascendantes et descendantes

Ce type d'outils peut être utilisé en présence d'un petit nombre de modèles. Dans le cas contraire, on pourra utiliser une stratégie dite de **régression descendante** faisant appel au test

de Fisher sur la présence d'un sous modèle. La méthodologie est la suivante : on part du modèle utilisant tous les régresseurs possibles. A chaque étape, on calcule la statistique de Fisher correspondant au retrait de chacune des variables encore présentes. On retire alors la variable possédant la plus petite valeur. On réitère ensuite ce processus jusqu'à ce que toutes les statistiques soient supérieures à un seuil pré-déterminé, par exemple 5%.

Ce type de stratégie peut-être extrêmement lourd à mettre en place suivant le nombre de variables en question (on peut aller jusqu'à $k!$ tests de Fisher).

La sélection de modèle par régression ascendante reprend exactement les mêmes arguments, sauf que l'on part du modèle vide (sans régresseur) et que l'on rajoute au fur à mesure les variables les plus significatives (au sens du test de Fisher), jusqu'au dépassement par les p-valeurs d'un seuil fixé préalablement.

3.3 Trois critères de sélection de modèle

3.3.1 Risque quadratique

Avant d'essayer de sélectionner le modèle le plus approprié, il faut pouvoir avoir une idée de la 'distance' entre chaque élément de \mathcal{M} et m^* . Il existe de nombreuses manières d'aborder cette notion.

Définition 3.1 Soit $m \in \mathcal{M}$. Le risque quadratique (de prédiction) entre les modèles m et m^* est défini par :

$$R(m, m^*) = \mathbb{E} \left(\left\| X_{(m^*)} \theta_{(m^*)} - X_{(m)} \hat{\theta}_{(m)} \right\|^2 \right).$$

A noter qu'il est également possible de s'intéresser par exemple au risque quadratique d'estimation défini comme

$$\tilde{R}(m, m^*) = \mathbb{E} \left(\left\| \theta_{(m^*)} - \hat{\theta}_{(m)} \right\|^2 \right).$$

Les analyses statistiques associées sont par contre légèrement plus compliquées. Par soucis de concision, nous nous concentrerons par la suite sur le risque en prédiction.

Par la suite, pour tout $m \in \mathcal{M}$, on définit

$$\mu = X_{(m^*)} \theta_{(m^*)}, \text{ et } \mu_{(m)} = P_{[X_{(m)}]} \mu,$$

le projeté orthogonal de μ sur l'espace vectoriel $[X_{(m)}]$. Il est alors possible, sous les hypothèses formulées au Chapitre 2, de calculer explicitement ce risque quadratique.

Proposition 3.1 Pour tout $m \in \mathcal{M}$, on a :

$$R(m, m^*) = \sigma^2(|m| + 1) + \|\mu_{(m)} - \mu\|^2. \quad (3.3)$$

PREUVE. A l'aide de calculs élémentaires :

$$\begin{aligned} R(m, m^*) &= \mathbb{E} \left(\|X_{(m)} \hat{\theta}_{(m)} - \mu\|^2 \right), \\ &= \mathbb{E} \left(\|X_{(m)} \hat{\theta}_{(m)} - \mu_{(m)} + \mu_{(m)} - \mu\|^2 \right), \\ &= \mathbb{E} \left(\|X_{(m)} \hat{\theta}_{(m)} - \mu_{(m)}\|^2 \right) + \mathbb{E} \left(\|\mu_{(m)} - \mu\|^2 \right). \end{aligned}$$

En effet, $(\mu_{(m)} - \mu) \in [X_{(m)}]^\perp$. A l'aide du Théorème de Cochran, on obtient alors :

$$\begin{aligned} R(m, m^*) &= \mathbb{E} \left(\|\hat{Y}_{(m)} - \mu_{(m)}\|^2 \right) + \mathbb{E} \left(\|\mu_{(m)} - \mu\|^2 \right), \\ &= \mathbb{E} \left(\|P_{[X_{(m)}]} \epsilon\|^2 \right) + \mathbb{E} \left(\|\mu_{(m)} - \mu\|^2 \right), \\ &= \sigma^2(|m| + 1) + \|\mu_{(m)} - \mu\|^2. \end{aligned}$$

□

Afin de minimiser la distance entre m et m^* , il y a donc un compromis à trouver. Si m est petit, il en sera de même pour le terme de variance $\sigma^2(|m| + 1)$, au dépend du terme de biais $\|\mu_{(m)} - \mu\|^2$. Au contraire, pour de grandes valeurs de m , on peut espérer avoir un petit biais, mais au risque d'avoir une erreur plus importante, ce qui se traduit par une augmentation du terme $\sigma^2(|m| + 1)$. Ce compromis biais-variance est très classique dans ce cadre de sélection de modèle et se retrouve dans un grand nombre de thématiques.

Remarque : A partir du moment où $m^* \subset m$, on a $\|\mu_{(m)} - \mu\|^2 = 0$, puisque $\mu_{(m)}$ correspond au projeté orthogonal de $\mu = X_{m^*}\theta_{m^*}$ sur $[X_{(m)}]$. A noter également que le meilleur modèle au sens du risque quadratique n'est pas forcément m^* .

EXEMPLE 1 : On simule un jeu de données en posant

$$X_i = \frac{i-1}{n} \quad \forall i \in \{1, \dots, n+1\},$$

et

$$Y_i = 10 \times (X_i - 0.5)^2 + \epsilon_i, \quad (3.4)$$

les ϵ_i désignant des erreurs i.i.d. gaussiennes. On utilise le modèle suivant :

$$Y_i = \sum_{k=1}^M [a_i \times \cos(2\pi k) + b_i \times \sin(2\pi k)] + \epsilon_i, \quad i = 1, \dots, n,$$

pour différentes valeurs de M .

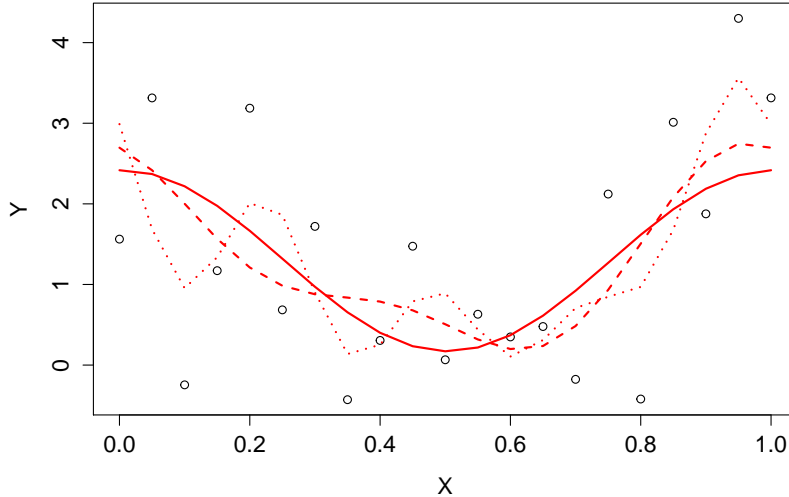


FIGURE 3.1 – Interpolation du nuage de points pour différentes fréquences.

On s'aperçoit que plus la valeur de M (fréquence maximale considérée) augmente, plus l'adéquation aux données est grande. En augmentant très fortement ce paramètre, on arriverait à une interpolation complète de ce nuage de points... ce qui n'est bien entendu pas le plus intéressant d'un point de vue statistique, surtout à la lumière du vrai modèle (3.4).

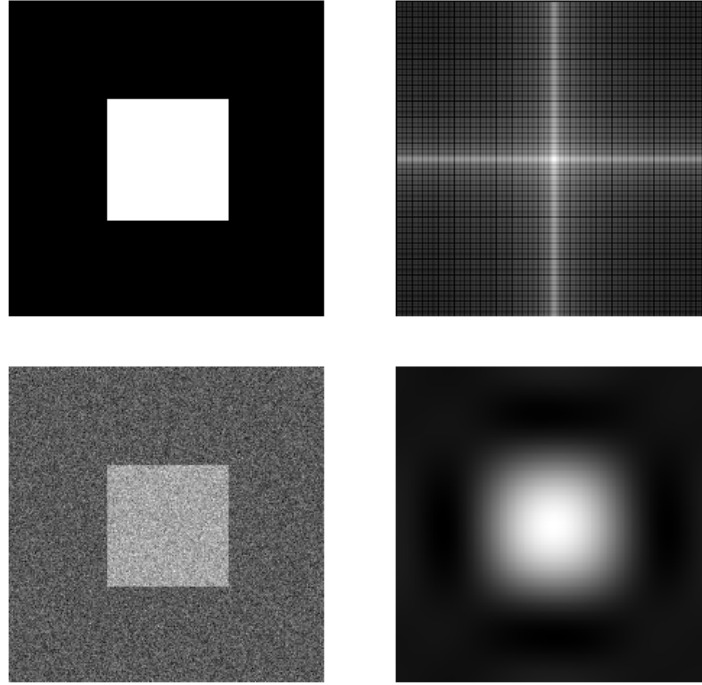


FIGURE 3.2 – De gauche à droite et de haut en bas : image originale, transformée de Fourier, image bruitée (modèle complet), image estimée avec 4 coefficients de Fourier.

EXEMPLE 2 : On s'intéresse dans cet exemple à du débruitage d'image. On considère ici une image de taille 256×256 . Après bruitage, et calcul des coefficients de Fourier¹, on obtient les observations

$$y_{kl} = \theta_{kl} + \sigma \xi_{kl}, \quad k, l \in \{1, \dots, 256\},$$

où les y_{kl} désignent les variables observées (coefficients de Fourier de l'image bruité), θ_{kl} les coefficients de Fourier de la vraie image, σ le niveau de bruit et les ξ_{kl} des variables i.i.d. centrées (gaussiennes). Sous une forme vectorisée, le modèle peut donc s'écrire sous la forme $Y = X\theta + \epsilon\xi$, où $X = I_d$ est égale dans cette configuration à la matrice identité. La figure 3.3.1 donne deux estimateurs possibles, l'un avec le modèle complet (sur-apprentissage), et l'autre conservant peu de coefficients de Fourier (sous-apprentissage).

La question qui se pose à présent est : comment approcher le modèle qui va minimiser le risque quadratique ? Clairement, trouver le meilleur modèle possible nécessite la connaissance de $\mu \dots$ que l'on cherche justement à estimer !

3.3.2 Le critère C_p de Mallows

L'idée proposée par Mallows dans les années 60 consiste à estimer ce risque quadratique à partir des données elles-mêmes et de prendre ensuite une décision à partir de cette estimation.

Soit $m \in \mathcal{M}$ fixé. On note par la suite $\hat{Y}_{(m)} = X_{(m)}\hat{\theta}_{(m)}$. Commençons par essayer d'estimer le terme de biais. A l'aide de calculs élémentaires :

$$\mathbb{E}\|Y - \mu_{(m)}\|^2 = \mathbb{E}\|Y - \hat{Y}_{(m)}\|^2 + \mathbb{E}\|\hat{Y}_{(m)} - \mu_{(m)}\|^2,$$

1. L'utilisation des coefficients de Fourier est peu recommandée en traitement d'image. Cet exemple n'est présenté qu'à des fins pédagogiques

FIGURE 3.3 – Estimation d'image par la méthode C_p .

ce qui revient à écrire,

$$\begin{aligned}\mathbb{E}\|Y - \hat{Y}_{(m)}\|^2 &= \mathbb{E}\|Y - \mu_{(m)}\|^2 - \mathbb{E}\|\hat{Y}_{(m)} - \mu_{(m)}\|^2, \\ &= \mathbb{E}\|Y - \mu\|^2 + \|\mu - \mu_{(m)}\|^2 - (|m| + 1)\sigma^2, \\ &= \|\mu - \mu_{(m)}\|^2 + n\sigma^2 - (|m| + 1)\sigma^2,\end{aligned}$$

ou encore

$$\|\mu - \mu_{(m)}\|^2 = \mathbb{E}\|Y - \hat{Y}_{(m)}\|^2 + (|m| + 1)\sigma^2 - n\sigma^2. \quad (3.5)$$

D'après (3.5), le biais $\|\mu - \mu_{(m)}\|^2$ peut donc être estimé par $\|Y - \hat{Y}_{(m)}\|^2 + (|m| + 1)\sigma^2$ (on néglige le terme en $n\sigma^2$ puisque ce dernier ne dépend pas de m et n'interviendra donc pas dans la minimisation).

Si la variance est connue, on obtient alors le critère :

$$C_p(m) = \|Y - \hat{Y}_{(m)}\|^2 + 2|m|\sigma^2.$$

On retiendra alors le modèle \hat{m}_{CP} vérifiant :

$$\hat{m}_{CP} = \arg \min_{m \in \mathcal{M}} C_p(m).$$

Dans le cas où la variance est inconnue, on utilisera l'estimateur $\hat{\sigma}^2 = \hat{\sigma}_{(m_p)}^2$ où $m_p = \{1, \dots, p\}$ est le modèle prenant en compte tous les régresseurs.

Remarque : Par définition $\hat{\sigma}_{(m)}^2$ est un estimateur de $n^{-1}\mathbb{E}\|Y - \hat{Y}_{(m)}\|^2$ (en supposant que le nombre p de régresseurs ne dépende pas de m . Il arrive donc parfois de voir le critère C_p écrit sous la forme :

$$\tilde{C}_p(m) = n\hat{\sigma}_{(m)}^2 + 2|m|\hat{\sigma}_{(m_p)}^2.$$

3.3.3 Le critère AIC

Le critère C_p de Mallows était basé sur une volonté de minimiser la distance entre \mathcal{M} et le vrai modèle au sens du risque quadratique. Le critère AIC s'appuie quand à lui sur la notion de vraisemblance. Par la suite, on note $l_n(Y, m, \theta)$ la vraisemblance associée au modèle m pour $\theta \in \mathbb{R}^{|m|}$. En particulier, dans un cadre gaussien, on a

$$\hat{\theta}_{(m)} = \arg \max_{\theta \in \mathbb{R}^{|m|}} l_n(Y, m, \theta).$$

On se souvient que la vraisemblance donne une mesure de l'adéquation de nos données par rapport au modèle proposé. Au passage, on rappelle la notion de vraisemblance est basée sur la notion de dissemblance de Kullback.

Définition 3.2 Soient \mathbb{P} et \mathbb{P}^* deux mesures de probabilités dominées par une même mesure (dans notre cas la mesure de Lebesgue). La dissemblance de Kullback entre ces deux mesures est donnée par :

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

En particulier, si deux mesures \mathbb{P}_θ et \mathbb{P}^* admettent une densité w.r.t. la mesure de Lebesgue, on obtient

$$K(\mathbb{P}, \mathbb{P}^*) = \int \ln \left(\frac{f^*}{f_\theta} \right) f^* = \int \ln(f^*) f^* - \int \ln(f_\theta) f^*.$$

Dans un modèle paramétrique ($\theta \in \Theta$ inconnu), le meilleur paramètre, au sens de la dissemblance de Kullback vérifie

$$\theta^0 = \arg \max_{\theta \in \Theta} \int \ln(f_\theta) f^*.$$

Cette intégrale n'étant pas calculable en pratique (f^* inconnue), on utilise sa version empirique, à savoir

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ln(f_\theta(Y_i)) = \arg \max_{\theta \in \Theta} l_n(Y, \theta) = \hat{\theta}_{MV}.$$

Dans le cadre de sélection de modèle qui nous intéresse ici, il est possible de baser notre choix sur une minimisation de la dissemblance de Kullback plutôt que sur la notion de risque quadratique. Une approche de sélection de modèle naïve consisterait à choisir le modèle $\hat{m} \in \mathcal{M}$ ayant la plus grande vraisemblance. Cependant, ce type d'approche ne tient pas compte de la taille de modèle. En effet, la vraisemblance a tendance à naturellement augmenter avec la complexité du modèle, et il y aurait donc un vrai risque de sur-ajustement aux données en utilisant une telle approche. Afin de corriger ce défaut, le critère *AIC* est alors défini de la manière suivante

$$AIC(m) = -2l_n(Y, m, \hat{\theta}_{(m)}) + 2|m| \quad \forall m \in \mathcal{M}.$$

En particulier, on choisira le modèle

$$\hat{m}_{AIC} = \arg \min_{m \in \mathcal{M}} AIC(m).$$

Autrement dit, le modèle retenu par l'intermédiaire de ce critère va maximiser la vraisemblance (qui peut être vue comme une mesure de l'adéquation au modèle) tout en proposant un contrôle de la dimension du modèle, le terme $2|m|$ pouvant être vu comme un terme de pénalisation.

La forme de cette pénalisation peut être justifiée de manière heuristique par le théorème de Wilks. Ce dernier indique, que sous certaines conditions de régularité,

$$2 \left\{ l_n(Y, \hat{\theta}_{(m)}, m) - l_n(Y, \theta^0) \right\} \xrightarrow{\mathcal{L}} \chi^2(|m|) \quad \text{quand } n \rightarrow +\infty.$$

En particulier, pour de grandes valeurs de n ,

$$-2 \left\{ l_n(Y, \hat{\theta}_{(m)}, m) - l_n(Y, \theta^0) \right\} \simeq S_m^2$$

avec $S_m^2 \sim -\chi_{|m|}^2$ et $\text{Var}[S_m^2] = 2|m| \rightarrow \infty$ as $|m| \rightarrow +\infty$. La pénalité de utilisée dans le critère AIC peut donc être vue comme un moyen de contrebalancer la variabilité du critère utilisé.

3.3.4 Le critère BIC

Le dernier critère présenté dans ce cours utilise le point de vue bayésien : on ne considère non plus le paramètre inconnu θ comme un vecteur de \mathbb{R}^k mais plutôt comme une variable aléatoire à valeurs dans \mathbb{R}^k . Une loi a priori est alors placée sur le 'paramètre' à estimer. La démarche consiste ensuite à essayer d'exploiter cette information pour l'estimation. Ce type d'approche apporte en théorie plus de richesse puisque l'on étend l'éventail des solutions possibles.

Cette approche conduit au critère BIC (pour *Bayesian Information Criterion*) défini par :

$$BIC(m) = n \log(\hat{\sigma}_{(m)}^2) + \log n \times |m|. \quad (3.6)$$

Le modèle correspondant \hat{m}_{BIC} est obtenu en posant :

$$\hat{m}_{BIC} = \arg \min_{m \in \mathcal{M}} BIC(m). \quad (3.7)$$

Nous ne nous étendrons pas sur les détails permettant d'arriver à la construction de ce critère.

Remarque : De nombreuses études ont été menées sur les performances théoriques et pratiques de ces trois critères. D'autres approches ont été également proposées dans des contextes similaires. Nous ne nous attarderons pas sur cet aspect dans ce cours.

3.4 Exercises

Exercise 1. Let $(\phi_k)_{k \in \mathbb{N}}$ be an orthonormal basis of $L^2([0, 1])$. We deal with the observations

$$Y_i = \langle f, \phi_k \rangle + \epsilon_i, \quad i = 1, \dots, M,$$

where the ϵ_i are i.i.d. random variables with distribution $\mathcal{N}(0, 1)$ and $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in $L^2(0, 1)$. We assume that the unknown function f is a linear function of the first M terms of the basis $(\phi_k)_{k \in \mathbb{N}}$.

1. Write the model in a matrix form.
2. Compute the quadratic risk of the MLE.

We assume now that the following information is available :

$$f \in \Theta(\alpha, Q) = \left\{ h : \sum_{k=1}^M \langle h, \phi_k \rangle^2 k^{2\alpha} \leq Q \right\}.$$

We consider the family $\mathcal{M} = (\{1, \dots, m\})_{m=1 \dots M}$.

3. Compute the quadratic risk associated to each model.
4. Propose an upper bound on this risk for each model under the assumption $f \in \Theta(\alpha, Q)$. What is the model that provides the best performance?
5. What strategy could we follow when the parameter α is unknown?

Exercise 2. The goal of this exercise is to obtain an expression of the AIC criterion in a Gaussian setting.

1. Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Compute the associated Kullback divergence (in a first time, one can assume that $\sigma_1 = \sigma_2 = \sigma$).
2. Prove that when the variance σ^2 is known, strategies based on the minimization of C_p and AIC are equivalent.
3. What could we do when σ^2 is unknown? What happens with the AIC criterium?

Exercise 3. This exercise is based on the dataset `proc-pin.dat`. It gathers data describing the population density of pine's caterpillar. Measures have been conducted on 33 different places. The potential explanatory variables are : elevation, slope, pine's density on the site, size of the trees, diameter, vegetation density, orientation, size of the dominants trees, number of vegetation strata and a mixing coefficient. Each column of the data set corresponds to the observation of each explanatory variable for each measure. The 11th column gathers the number of nests. Provide a complete analysis of this dataset.

Exercise 4. (Construction of a protein molecular thermometer)

The optimal development temperature of eucariotic cells strongly depends on its amino acids composition. If confirmed, we expect a linear relationship. Our aim is to characterize this dependency from a given sample and then to identify the most important amino acids, i.e. those who can explain the temperature.

The data are gathered in the file `procaryotes.dat`. For each of the 730 investigated cells, the optimal growing temperature is indicated, jointly with the frequency of each amino acid.

Objectif : Provide a complete analysis of this sample (do not hesitate to include interactions schemes). RÉFÉRENCES

[1] J.R. Lobry et A. Necsulea. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* 385 (2006) pp. 128-136.

[2] J.R. Lobry. Sélection de modèle pour la création d'un thermomètre moléculaire protéique. Document pédagogique (2008).

Chapitre 4

Statistique en grande dimension

4.1 Motivations

On considère à nouveau le modèle linéaire

$$Y = X\theta + \epsilon, \quad (4.1)$$

où X désigne la matrice de design de taille $n \times p$, $\theta \in \mathbb{R}^p$ le paramètre à estimer et ϵ le vecteur d'erreur. Dans le chapitre précédent, nous nous sommes intéressés à la situation où $p < n$, la matrice X étant régulière. Cependant, il arrive dans de nombreuses applications (récentes) que le nombre de paramètres à estimer p soit beaucoup plus grand que le nombre d'observations disponibles. On donne ci-dessous une liste non-exhaustive d'exemples faisant intervenir de telles situations.

Médecine. On souhaite quantifier la présence d'une pathologie dans une population à l'aide d'un marqueur moléculaire pré-défini. On réunit pour cela un panel d'individus (de patients) pour lequel on va déterminer la quantité présente de ce type de marqueur dans le sang, ainsi que d'un certain nombre de caractéristiques biologiques. Pour des raisons de coût, le nombre de patients (ici n) est souvent peu élevé (typiquement entre 20 et 200) comparé à la quantité d'informations disponible pour chaque individu (parfois jusqu'à plusieurs milliers).

Régression non-paramétrique. On dispose d'une série d'observations indépendantes $(Y_1, Z_1), \dots, (Y_n, Z_n)$, les Y_i, Z_i désignant des variables réelles. On cherche à modéliser le lien existant entre Y et Z par l'intermédiaire du modèle

$$Y_i = f(Z_i) + \epsilon_i, \quad i = 1 \dots n,$$

les ϵ_i correspondant à une erreur de mesure, et la fonction f étant inconnue. On se munit pour cela d'un dictionnaire de fonctions $(\Psi_j)_{j=1 \dots p}$, les Ψ_j pouvant être extraites d'une base de Fourier, ondelettes, ou autre. En supposant que f puisse être écrite comme une combinaison linéaire des Ψ_j , on peut alors écrire, pour tout $i \in \{1, \dots, n\}$

$$Y_i = \sum_{j=1}^p \theta_j \Psi_j(Z_i) + \epsilon_i,$$

ce qui donne la forme matricielle (4.1) avec $X_{ij} = \Psi_j(Z_i)$ et $\theta_j = \langle \Psi_j, f \rangle$. Dans ce contexte, le nombre d'observations n est potentiellement limité comparé à la taille du dictionnaire p .

Clairement, dans le contexte où $p > n$ (voire $p \gg n$), la matrice X n'est plus régulière. Par conséquent, l'estimateur des moindres carrés n'est plus unique et la consistance n'est plus assurée : on parle souvent dans ce cas de sur-ajustement par rapport aux données.

Question : *Peut-on malgré tout espérer pouvoir reconstruire le vecteur θ dans certaines situations ?*

On va voir au cours de ce chapitre que la réponse est 'oui', à condition de faire quelques hypothèses structurelles sur le signal θ et la matrice de design X . À ce titre, la notion de parcimonie (sparsité) jouera un rôle central.

Définition 4.1 *L'indice de parcimonie du vecteur θ , noté s correspond au nombre de coefficients non-nuls*

$$s = \#\{j : \theta_j \neq 0\} = \sum_{j=1}^p \mathbf{1}_{\{\theta_j \neq 0\}}.$$

Le support de θ est quand à lui défini comme

$$J(\theta) = \{j : \theta_j \neq 0\}.$$

Dans le cas où le support de θ est connu, il est possible de proposer une estimation consistante de θ même dans la situation où $p \gg n$. On définit pour cela

$$\hat{\theta}^0 := \arg \min_{\nu \in \mathbb{R}^p : J(\nu) = J(\theta)} \|Y - X\nu\|^2.$$

Proposition 4.1 *Soit $\hat{\theta}^0$ l'estimateur défini ci-dessus. On a alors*

$$R(\theta^0, \theta) = \frac{1}{n} \mathbb{E} \|X\hat{\theta}^0 - X\theta\|^2 = \frac{s\sigma^2}{n}.$$

PREUVE. En posant X^0 la matrice X restreinte à l'ensemble $J(\theta)$, on obtient

$$\mathbb{E} \|X\hat{\theta}^0 - X\theta\|^2 = \mathbb{E} \|P_{[X^0]} Y - X\theta\|^2 = \mathbb{E} \|P_{[X^0]} \epsilon\|^2 = s\sigma^2,$$

en utilisant le théorème de Cochran.

□

À la lumière de la Proposition 4.1, on voit donc qu'il est possible d'obtenir une estimation consistante de θ dès lors que

$$\frac{s}{n} = o(1) \quad \text{quand } n \rightarrow +\infty.$$

Malheureusement, l'estimateur θ^0 n'est pas disponible en pratique puisque le vecteur θ , et donc son support, sont inconnus. Malgré tout, il existe un certain nombre de méthodes permettant de contourner ce problème.

Dans la suite de ce chapitre, on va donc implicitement supposer que le vecteur d'intérêt θ est parcimonieux. Ce type de restriction se justifie tout à fait dans de nombreux domaines applications. En médecine, cela revient à supposer que la pathologie d'intérêt peut-être expliquée par un petit nombre de co-variables. En régression non-paramétrique, on supposera que la fonction f peut-être représentée comme une combinaison linéaire d'un petit nombre d'éléments du dictionnaire (hypothèse tout à fait réaliste si on utilise par exemple une base d'ondelettes).

4.2 Principe de la méthode LASSO

Sachant que le paramètre θ d'intérêt est parcimonieux (avec indice de sparsité s inconnu), il semble 'naturel' de rechercher un estimateur $\tilde{\theta}$ du type

$$\tilde{\theta} = \arg \min_{\nu \in \mathbb{R}^p \text{ t.q. } J(\nu) \leq R} \|Y - X\nu\|^2,$$

ou $R \in \mathbb{R}^+$ est une constante à choisir par l'utilisateur. Ce problème d'optimisation peut être ré-écrit sous la forme

$$\tilde{\theta} = \arg \min_{\nu \in \mathbb{R}^p} \left[\frac{1}{n} \|Y - X\nu\|^2 + \lambda \|\nu\|_0 \right], \quad (4.2)$$

ou $\|\nu\|_0 = |J(\nu)|$ désigne en fait la 'norme' l_0 du vecteur ν . D'une certaine manière, on pénalise dans ce problème d'optimisation, le terme d'attachement aux données par une contrainte sur la

structure du vecteur recherche. D'un point de vue théorique, cet estimateur pourrait s'avérer tout à fait compétitif. Malheureusement, les performances pratiques ne sont pas au rendez-vous, en particulier en ce qui concerne la calculabilité de $\hat{\theta}$, le problème d'optimisation (4.2) n'étant pas convexe.

Dans ce contexte, une alternative possible est d'utiliser une pénalité différente, qui va continuer à forcer la solution à être parcimonieuse, tout en rendant le problème convexe. L'algorithme du LASSO (pour Least Absolute Shrinkage and Selection Operator) consiste à utiliser une régularisation (pénalité) de type l_1 . Autrement dit, on va s'intéresser à l'estimateur $\hat{\theta}^L$ défini comme

$$\hat{\theta}^L = \arg \min_{\nu \in \mathbb{R}^p} \left[\frac{1}{n} \|Y - X\nu\|^2 + \lambda \|\nu\|_1 \right], \quad (4.3)$$

ou

$$\|x\|_1 = \sum_{j=1}^p |x_j|, \quad \forall x \in \mathbb{R}^p.$$

D'un point de vue heuristique, la pénalité l_1 offre un bon compromis entre les régularisations l_0 (absence de convexité) et l_2 (contrainte faible sur la parcimonie de la solution). Au passage, une contrainte de type l_2 donne lieu à la méthode de régression 'ridge', dont l'estimateur associé est défini comme

$$\hat{\theta}^R = \arg \min_{\nu \in \mathbb{R}^p} \left[\frac{1}{n} \|Y - X\nu\|^2 + \lambda \|\nu\|_2 \right] = (X'X + \lambda I_p)^{-1} X'Y.$$

Calcul de l'estimateur LASSO

Le problème d'optimisation étant convexe, on peut alterner les étapes de minimisation sur chacune des coordonnées de θ . Pour tout $\beta \in \mathbb{R}^p$, soit

$$G(\beta) := \frac{1}{n} \|Y - X\theta\|^2 + \lambda \|\theta\|_1.$$

On peut remarquer que

$$\frac{\partial G}{\partial \theta_j}(\theta) = -\frac{2}{n} X_j'(Y - X\theta) + \lambda \frac{\theta_j}{|\theta_j|} \quad \forall j \in \{1, \dots, p\}.$$

On peut alors voir, après quelques calculs, que pour $j \in \{1, \dots, p\}$ la fonction $\theta_j \mapsto G(\theta_1, \dots, \theta_j, \dots, \theta_p)$ atteint son minimum pour

$$\theta_j^* = R_j \left(1 - \frac{\lambda}{2|R_j|} \right)_+ \quad \text{avec} \quad R_j = \frac{1}{n} X_j'(Y - \sum_{k \neq j} \theta_k X_k).$$

On obtient l'algorithme suivant

Algorithme pour le LASSO

- Initialiser θ de manière arbitraire,
- Itérer jusqu'à obtention de la convergence

$$\theta_j = R_j \left(1 - \frac{\lambda}{2|R_j|} \right)_+ \quad \forall j \in \{1, \dots, p\}.$$

- Conserver θ .
-

4.3 Performances théoriques

4.3.1 Un premier résultat

On s'intéresse à présent aux performances théoriques de la méthode LASSO. On supposera par souci de simplicité que les erreurs sont gaussiennes. Sans perte de généralité, on travaillera également avec une matrice de design X re-normalisée, i.e.

$$\frac{1}{n} \sum_{i=1}^n X_{ij} = 1 \quad \forall j \in \{1, \dots, p\}.$$

Dans ce contexte, il est possible d'obtenir l'inégalité oracle suivante.

Théorème 4.1 *On suppose les erreurs gaussiennes et la matrice X normalisée. Dans ce cas, en choisissant*

$$\lambda \geq \frac{4\sigma}{\sqrt{n}} \sqrt{2 \ln(p/\delta)},$$

on obtient, avec probabilité supérieure à $1 - \delta$,

$$\frac{1}{n} \|X(\hat{\theta} - \theta)\|^2 + \lambda \|\hat{\theta}\|_1 \leq 3 \inf_{\theta \in \mathbb{R}^p} \left[\frac{1}{n} \|X(\theta - \theta^*)\|^2 + \lambda \|\theta\|_1 \right]. \quad (4.4)$$

PREUVE. Par construction de $\hat{\theta}^L = \hat{\theta}$, on a, pour tout $\theta \in \mathbb{R}^p$,

$$\frac{1}{n} \|Y - X\hat{\theta}\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|Y - X\theta\|^2 + \lambda \|\theta\|_1. \quad (4.5)$$

Pour tout $\gamma \in \mathbb{R}^p$, on peut remarquer que

$$\begin{aligned} \|Y - X\gamma\|^2 &= \|Y - X\theta^* + X\theta^* - X\gamma\|^2, \\ &= \|Y - X\theta^*\|^2 + \|X(\gamma - \theta^*)\|^2 + 2\langle Y - X\theta^*, X\theta^* - X\gamma \rangle. \end{aligned}$$

En injectant cette égalité de chaque cote de (4.5), on obtient

$$\begin{aligned} &\frac{1}{n} \|Y - X\theta^*\|^2 + \frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \frac{2}{n} \langle Y - X\theta^*, X\theta^* - X\hat{\theta} \rangle + \lambda \|\hat{\theta}\|_1 \\ &\leq \frac{1}{n} \|Y - X\theta^*\|^2 + \frac{1}{n} \|X(\theta - \theta^*)\|^2 + \frac{2}{n} \langle Y - X\theta^*, X\theta^* - X\theta \rangle + \lambda \|\theta\|_1, \end{aligned}$$

ce qui implique

$$\begin{aligned} &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + \frac{2}{n} \langle Y - X\theta^*, X(\hat{\theta} - \theta) \rangle + \lambda \|\theta\|_1, \\ \Leftrightarrow &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + \frac{2}{n} \langle \epsilon, X(\hat{\theta} - \theta) \rangle + \lambda \|\theta\|_1. \end{aligned} \quad (4.6)$$

Par la suite, il convient de contrôler le terme aléatoire

$$T := \frac{1}{n} \langle \epsilon, X(\hat{\theta} - \theta) \rangle.$$

Dans un premier temps,

$$\begin{aligned} T &= \frac{1}{n} \sum_{i=1}^n \epsilon_i \left[X(\hat{\theta} - \theta) \right]_i, \\ &= \frac{1}{n} \sum_{i=1}^n \epsilon_i \sum_{j=1}^p X_{ij} (\hat{\theta}_j - \theta_j), \\ &= \sum_{j=1}^p (\hat{\theta}_j - \theta_j) \times \frac{1}{n} \sum_{i=1}^n X_{ij} \epsilon_i = \sum_{j=1}^p (\hat{\theta}_j - \theta_j) V_j, \end{aligned}$$

ou

$$V_j := \frac{1}{n} \sum_{i=1}^n X_{ij} \epsilon_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \quad \forall j \in \{1, \dots, p\}.$$

Pour tout $x \in \mathbb{R}_+$, on introduit l'événement

$$\mathcal{A}_x = \bigcap_{j=1}^p \{|V_j| \leq x\}.$$

Par la suite, on cherche à déterminer les valeurs de x pour lesquelles l'événement \mathcal{A}_x se réalise avec 'grande' probabilité. On commence par remarquer que pour tout $j \in \{1, \dots, p\}$,

$$\begin{aligned} \mathbb{P}(|V_j| > x) &= 2 \int_x^{+\infty} \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-\frac{nu^2}{2\sigma^2}} du, \\ &\leq \frac{2}{x} \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} \int_x^{+\infty} ue^{-\frac{nu^2}{2\sigma^2}} du, \\ &= \frac{2\sigma}{x\sqrt{2\pi n}} \left[-e^{-\frac{nu^2}{2\sigma^2}}\right]_x^{+\infty} = \frac{2}{x} \frac{\sigma}{\sqrt{2\pi n}} e^{-\frac{nx^2}{2\sigma^2}}. \end{aligned}$$

A l'aide d'inégalités élémentaires, on arrive à

$$\mathbb{P}(\mathcal{A}_x^c) = \mathbb{P}\left(\bigcup_{j=1}^p \{|V_j| > x\}\right) \leq \sum_{j=1}^p \mathbb{P}(|V_j| > x) \leq \frac{2p}{x} \frac{\sigma}{\sqrt{2\pi n}} e^{-\frac{nx^2}{2\sigma^2}} \leq \delta,$$

pour tout x t.q.

$$x \geq x_0 := \sigma \sqrt{\frac{2}{n} \ln(p/\delta)}.$$

En utilisant ce contrôle dans l'inégalité (4.6), on obtient qu'avec probabilité supérieure à $1 - \delta$,

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + 2x_0 \|\hat{\theta} - \theta\|_1 + \lambda \|\theta\|_1. \quad (4.7)$$

En choisissant¹ alors le paramètre de régularisation λ de telle sorte que

$$2x_0 \leq \frac{\lambda}{2} \Leftrightarrow \lambda = 4x_0 = 4\sigma \sqrt{\frac{2}{n} \ln(p/\delta)},$$

on obtient d'après l'inégalité (4.7)

$$\begin{aligned} &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + \frac{\lambda}{2} \|\hat{\theta} - \theta\|_1 + \lambda \|\theta\|_1, \\ \Rightarrow &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + \frac{\lambda}{2} \|\hat{\theta}\|_1 + \frac{\lambda}{2} \|\theta\|_1 + \lambda \|\theta\|_1, \\ \Rightarrow &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \frac{\lambda}{2} \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + \frac{3\lambda}{2} \|\theta\|_1. \end{aligned}$$

Cette dernière inégalité étant valable pour tout $\theta \in \mathbb{R}^p$, on obtient en particulier que

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq 3 \inf_{\theta \in \mathbb{R}^p} \left[\frac{1}{n} \|X(\theta - \theta^*)\|^2 + \lambda \|\theta\|_1 \right],$$

ce qui correspond exactement au résultat souhaité. □

L'inégalité (4.4) est appelée *inégalité oracle* au sens où elle permet de comparer les performances de l'estimateur lasso à la perte (pénalisée) minimale envisageable, atteinte par l'oracle qui dépend explicitement de la solution inconnue du problème. Ce type de résultat permet entre autre d'obtenir des vitesses de convergence, comme le montre le corollaire suivant.

1. On se rend compte à ce stade de l'importance de la pénalité : sans elle, la seule borne disponible est de l'ordre de n/p

Corollaire 4.1 Soit $\hat{\theta}^L$ l'estimateur défini en (4.3) avec

$$\lambda = 4\sigma\sqrt{\frac{2}{n}\ln(p/\delta)}, \quad (4.8)$$

Avec probabilité supérieure à $1 - \delta$, on a

$$\frac{1}{n}\|X(\hat{\theta} - \theta^*)\|^2 \leq C \frac{\sigma s}{\sqrt{n}} \sqrt{\ln(p/\delta)},$$

ou C designe une constante positive.

PREUVE. En reprenant l'inégalité (4.4) avec $\theta = \theta^*$, on obtient

$$\frac{1}{n}\|X(\hat{\theta} - \theta^*)\|^2 \leq 3\lambda\|\theta^*\|_1.$$

On remarque alors que

$$\|\theta^*\|_1 = \sum_{j=1}^p |\theta_j^*| \leq \|\theta^*\|_\infty s.$$

Le choix de λ (4.8) permet alors de conclure. □

Le principal enseignement des résultats obtenus ci-dessus, et en particulier du Corollaire 4.1 est que la méthode Lasso produit une prediction consistante des lors que

$$\frac{\sigma s}{\sqrt{n}} \sqrt{\ln(p)} = o(1) \quad \text{quand } n \rightarrow +\infty.$$

Cette condition est beaucoup plus faible que dans le cadre classique de regression par méthode des moindres carres et permet en particulier de considerer les cas ou $p \gg n$.

4.3.2 Vitesse rapide avec contrainte structurelle sur X

La borne supérieure obtenue dans le corollaire 4.1 ci-dessus fait intervenir un terme de l'ordre de $1/\sqrt{n}$. Il est légitime de se demander si cette vitesse peut être améliorée. Nous allons voir que la réponse est *oui*... en étant par exemple un peu plus exigeant sur la structure de la matrice X .

Hypothèse de compatibilité. Soit $\hat{\Sigma}$ la matrice de covariance définie comme $\hat{\Sigma} = \frac{1}{n}X'X$. Il existe une constante positive ϕ_0 telle que pour tout $\gamma \in \mathbb{R}^p$ satisfaisant

$$\|\gamma_{J_0^c}\|_1 \leq 3\|\gamma_{J_0}\|_1,$$

alors

$$\|\gamma_{J_0}\|_1^2 \leq \phi_0^{-2}(\gamma' \hat{\Sigma} \gamma)s.$$

Moralement,, cette hypotheses demande que les co-variables impliquées dans le 'vrai' modèle, i.e. celle associées au vecteur θ^* ne soit pas trop colinéaire. Une telle propriété garantit alors d'une certaine manière que le signal present dans les observations sera visible malgré la grand dimension du modèle. Le lemme présenté ci-dessous donne une condition suffisante pour la validation d'une telle condition.

Lemme 4.1 Si la matrice de Gram satisfait

$$\hat{\Sigma}_{ii} = 1 \quad \text{et} \quad |\hat{\Sigma}_{ij}| \leq \frac{1}{7\alpha s} \quad \forall i, j \in \{1, \dots, m\},$$

pour une constante $\alpha > 1$, alors l'hypothèse de compatibilité est satisfaite avec $\phi_0 = \sqrt{1 - 1/\alpha}$.

PREUVE. Pour tout $J \subset \{1, \dots, M\}$ tel que $|J| \leq s$ et $\gamma \in \mathbb{R}^n$, on a

$$\begin{aligned} \frac{\gamma'_J \hat{\Sigma} \gamma_J}{\|\gamma_J\|^2} &= 1 + \frac{\gamma'_J (\hat{\Sigma} - I_m) \gamma_J}{\|\gamma_J\|^2}, \\ &\geq 1 - \frac{1}{7\alpha s} \frac{\left(\sum_{j \in J} |\gamma_j|\right)^2}{\|\gamma_J\|^2}, \\ &\geq 1 - \frac{1}{7\alpha}. \end{aligned}$$

Ensuite, en utilisant l'inégalité $\|\gamma_{J^c}\|_1 \leq 3\|\gamma_J\|_1$, on obtient

$$\begin{aligned} \frac{|\gamma'_{J^c} \hat{\Sigma} \gamma_J|}{\|\gamma_J\|^2} &\leq \frac{1}{7\alpha s} \frac{\sum_{i \in J} \sum_{j \in J^c} |\gamma_i| \times |\gamma_j|}{\|\gamma_J\|^2}, \\ &= \frac{1}{7\alpha s} \frac{\|\gamma_{J^c}\|_1 \times \|\gamma_J\|_1}{\|\gamma_J\|^2}, \\ &\leq \frac{3}{7\alpha s} \frac{\|\gamma_J\|_1}{\|\gamma_J\|^2}, \\ &\leq \frac{3}{7\alpha}. \end{aligned}$$

En combinant les deux inégalités précédentes, on obtient

$$\frac{\gamma' \hat{\Sigma} \gamma}{\|\gamma_J\|^2} \geq \frac{\gamma'_J \hat{\Sigma} \gamma_J}{\|\gamma_J\|^2} + 2 \frac{\gamma'_{J^c} \hat{\Sigma} \gamma_J}{\|\gamma_J\|^2} \geq 1 - \frac{1}{\alpha} > 0.$$

□

Grace a cette hypothèse de compatibilité, il est possible d'améliorer l'inégalité oracle obtenue dans la section précédente.

Théorème 4.2 *On suppose les erreurs gaussiennes, la matrice X normalisée et l'hypothèse de compatibilité satisfaite. Dans ce cas, en choisissant*

$$\lambda \geq \frac{4\sigma}{\sqrt{n}} \sqrt{2 \ln(p/\delta)},$$

on obtient, avec probabilité supérieure a $1 - \delta$,

$$\frac{1}{n} \|X(\hat{\theta} - \theta)\|^2 \leq 3 \left(\frac{1 + \alpha^{-1}}{1 + \alpha} \right) \frac{\lambda^2}{\phi_0^2} s. \quad (4.9)$$

avec $\alpha > 0$.

PREUVE. On rappelle l'inégalité (4.7) obtenue ci dessus

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{n} \|X(\theta - \theta^*)\|^2 + 2x_0 \|\hat{\theta} - \theta\|_1 + \lambda \|\theta\|_1.$$

On peut alors remarquer que

- $\|\hat{\theta}\|_1 = \|\hat{\theta}_{J_0}\|_1 + \|\hat{\theta}_{J_0^c}\|_1,$
- $\|\hat{\theta} - \theta^*\|_1 = \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \|(\hat{\theta} - \theta^*)_{J_0^c}\|_1 = \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \|\hat{\theta}_{J_0^c}\|_1$

En insérant ces égalites dans (4.7), et en choisissant $\lambda \geq 4x_0$ et $\theta = \theta^*$, on obtient

$$\begin{aligned} &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}_{J_0}\|_1 + \lambda \|\hat{\theta}_{J_0^c}\|_1 \leq \frac{\lambda}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \frac{\lambda}{2} \|\hat{\theta}_{J_0^c}\|_1 + \lambda \|\theta^*\|_1, \\ \Rightarrow &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}_{J_0}\|_1 \leq \frac{\lambda}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \lambda \|\theta^*\|_1, \\ \Rightarrow &\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}_{J_0}\|_1 \leq \frac{\lambda}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \lambda \|\theta_{J_0}^*\|_1. \end{aligned}$$

A l'aide d'une inégalité triangulaire, on peut écrire

$$\|\theta_{J_0}^*\|_1 \leq \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \|\hat{\theta}_{J_0}\|_1.$$

En mettant ces inégalités bout-a-bout, on arrive alors à

$$\begin{aligned} \frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 + \lambda \|\hat{\theta}_{J_0}\|_1 &\leq \frac{\lambda}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \lambda \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \lambda \|\hat{\theta}_{J_0}\|_1, \\ \Rightarrow \frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 &\leq \frac{3\lambda}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1. \end{aligned}$$

Admettons pour l'instant qu'avec probabilité supérieure à $1 - \delta$, on ait

$$\|(\hat{\theta} - \theta^*)_{J_0^c}\|_1 \leq 3 \|(\hat{\theta} - \theta^*)_{J_0}\|_1. \quad (4.10)$$

Alors, en utilisant l'hypothèse de compatibilité, on obtient que

$$\|(\hat{\theta} - \theta^*)_{J_0}\|_1 \leq \phi_0^{-1} \sqrt{s} \sqrt{(\hat{\theta} - \theta^*)' \hat{\Sigma} (\hat{\theta} - \theta^*)} = \phi_0^{-1} \frac{1}{\sqrt{n}} \|X(\hat{\theta} - \theta^*)\| \sqrt{s}.$$

En remarquant que pour tout $a, b \in \mathbb{R}$ et $\alpha > 0$,

$$2ab \leq \alpha a^2 + \alpha^{-1} b^2,$$

on obtient

$$\lambda \|(\hat{\theta} - \theta^*)_{J_0}\|_1 \leq \alpha \|X(\hat{\theta} - \theta^*)\|^2 + \alpha^{-1} \frac{\lambda^2}{\phi_0^2} s.$$

Finalement

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 \leq 3 \left(\frac{1 + \alpha^{-1}}{1 + \alpha} \right) \frac{\lambda^2}{\phi_0^2} s,$$

ce qui permet de conclure. Il reste pour finir cette preuve à vérifier que l'inégalité (4.10) est valable.

En repartant à nouveau de l'inégalité (4.7) avec $\theta = \theta^*$, et $2x_0 \leq \lambda$, on obtient

$$\begin{aligned} \lambda \|\hat{\theta}\|_1 &\leq \frac{\lambda}{2} \|\hat{\theta} - \theta^*\|_1 + \lambda \|\theta^*\|_1, \\ \Leftrightarrow \|\hat{\theta}_{J_0}\|_1 + \|\hat{\theta}_{J_0^c}\|_1 &\leq \frac{1}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \frac{1}{2} \|\hat{\theta}_{J_0^c}\|_1 + \|\theta_{J_0}^*\|_1, \\ \Rightarrow \|\hat{\theta}_{J_0}\|_1 + \|\hat{\theta}_{J_0^c}\|_1 &\leq \frac{1}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \frac{1}{2} \|\hat{\theta}_{J_0^c}\|_1 + \|(\theta^* - \theta)_{J_0}\|_1 + \|\hat{\theta}_{J_0}\|_1, \\ \Leftrightarrow \|\hat{\theta}_{J_0^c}\|_1 &\leq \frac{1}{2} \|(\hat{\theta} - \theta^*)_{J_0}\|_1 + \frac{1}{2} \|\hat{\theta}_{J_0^c}\|_1 + \|(\theta^* - \theta)_{J_0}\|_1, \\ \Leftrightarrow \|(\hat{\theta} - \theta^*)_{J_0^c}\|_1 &\leq 3 \|(\hat{\theta} - \theta^*)_{J_0}\|_1. \end{aligned}$$

This concludes the proof. □

Une conséquence immédiate de ce résultat est qu'en choisissant

$$\lambda = \frac{4\sigma}{\sqrt{n}} \sqrt{2 \ln(p/\delta)},$$

il est possible de trouver une constante positive C telle qu'avec probabilité supérieure à $1 - \delta$, on ait

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|^2 \leq C \frac{\sigma s}{n} \ln(p).$$

Cette vitesse est bien plus rapide que celle obtenue dans la section précédente. En particulier, la méthode Lasso produit un estimateur consistant des que

$$\frac{s}{n} \ln(p) = o(1) \quad \text{quand } n \rightarrow +\infty.$$

On voit également que l'estimateur ainsi produit possède un comportement similaire à l'estimateur proposé en tout début de chapitre... sauf que ce dernier supposait la connaissance explicite du support de θ^* , ce qui n'est plus le cas ici. Le prix à payer pour la non-connaissance de ce support est un terme additionnel en $\ln(p)$, ainsi que la nécessité de manipuler une matrice de design ayant de 'bonnes' propriétés.

Chapitre 5

Le modèle linéaire généralisé

5.1 Introduction

On observe un vecteur Y de taille n , réalisation d'une variable aléatoire de moyenne μ et dont les composants sont indépendants. Dans le cadre du modèle linéaire, on a $\mu = X\theta$ où X est une matrice de design $n \times p$. Le vecteur θ est inconnu et modélise l'influence des régresseurs sur la réponse aléatoire Y .

Le modèle linéaire tel que nous l'avons vu peut donc être caractérisé de la manière suivante :

1. une **composante aléatoire** : le vecteur Y est une variable aléatoire de moyenne μ ,
2. une **composante 'systématique'**, les régresseurs X_1, X_2, \dots, X_p définissent un prédicteur linéaire : $\eta = X\theta$,
3. **la relation liant** μ et η , $\mu = \eta$ pour le modèle linéaire.

Imposer une dépendance linéaire entre les régresseurs et $\mathbb{E}[Y]$ permet une étude approfondie mais peut être parfois trop restrictive. Une généralisation possible du modèle linéaire consiste donc à supposer que la relation liant μ à η n'est pas l'identité, mais plutôt un lien du type :

$$\eta_i = g(\mu_i), \text{ pour } \eta = (\eta_1, \dots, \eta_n)' \text{ et } \mu = (\mu_1, \dots, \mu_n).$$

La fonction g modélise donc le lien entre ces deux vecteurs. Cette formulation permet de modéliser un panel plus riche d'expériences.

EXEMPLE 3.1 : Dans une expérience clinique, on cherche à comparer deux modes opératoires pour une opération chirurgicale donnée. L'expérience est menée sur deux hôpitaux différents. On dispose donc ici de deux facteurs à deux modalités : *mode opératoire* et *hôpital*. La variable réponse correspond pour chaque patient au succès (ou à l'échec) de l'intervention : c'est donc bien une variable binaire.

EXEMPLE 3.2 : Dans le cadre d'une étude sociologique, on cherche à mettre en relation l'absentéisme scolaire avec différentes variables : groupe, sexe des individus, âge et performances scolaires. On compte pour chaque individu le nombre de jours d'absence sur une année : il s'agit de la variable à expliquer. Dans la mesure où cette variable est de type *comptage*, il semble naturel de la modéliser par l'intermédiaire d'une loi de Poisson.

Dans le cas où la fonction de lien est de type canonique (i.e. $g(x) = x$), rien n'interdit d'utiliser la méthode des moindres carrés introduite au Chapitre 2. Cette dernière est en effet purement géométrique et peut donc tout à fait s'appliquer à des réponses de type 'binaire', même si l'interprétation qui en découle peut s'avérer délicate (Figure 5.2). La partie inférentielle traitée dans ce cours nécessite quand à elle des hypothèses très fortes sur la distribution des observations. Pour des modèles alternatifs, il faut donc complètement repenser la construction des tests et des intervalles de confiance. Par ailleurs, une relation de type canonique est relativement restrictive : il convient donc de se placer dans un cadre plus général afin de pouvoir faire face à des problèmes plus variés.

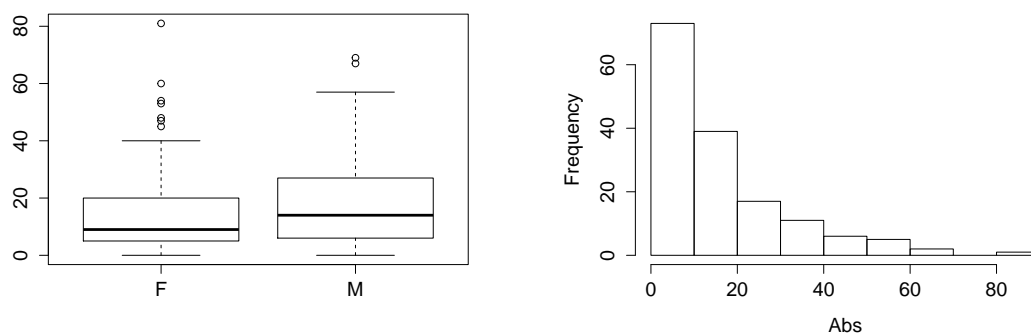


FIGURE 5.1 – Nombre de jours d’absence (par valeurs entières) en fonction du sexe des individus (à gauche) et répartition empirique (à droite).

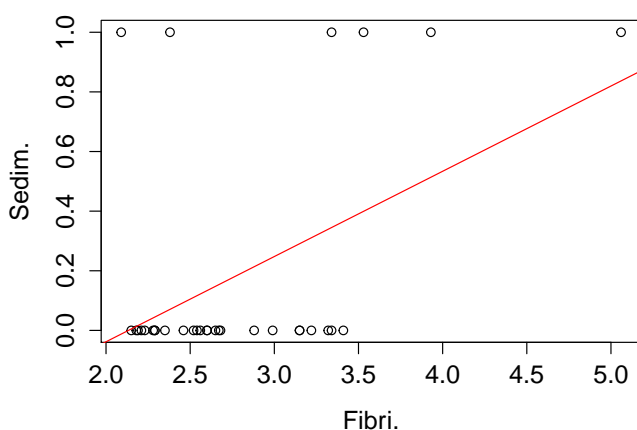


FIGURE 5.2 – Vitesse de sédimentation dans le plasma sanguin et interpolation par regression linéaire.

Etant donnée une fonction de lien, la méthode des moindres carrés introduite au Chapitre 2 ne peut pas toujours être implémentée. Bien souvent, le problème d’optimisation associé n’est en effet pas convexe. Une première ‘parade’ consiste à utiliser l’estimateur du maximum de vraisemblance... mais dans la plupart des cas, ce dernier n’est pas calculable analytiquement. Il est cependant possible d’utiliser un algorithme itératif inspiré de la méthode de Newton-Raphson permettant d’approcher le maximum de la vraisemblance. Sous certaines conditions, cet algorithme propose des résultats tout à fait satisfaisants.

5.2 Caractérisation d’un modèle

La notion de modèle linéaire généralisé regroupe un grand nombre de modèles différents (en apparence). Ce paragraphe propose un tour d’horizon des plus utilisés.

5.2.1 Familles exponentielles

Comme nous l’avons vu ci-dessus, un modèle de type ‘linéaire généralisé’ est en partie caractérisé par la distribution de la variable de réponse.

Afin de pouvoir mener à bien une inférence pertinente sur les données, nous nous restreindrons dans ce chapitre à des distributions bien particulières : les familles exponentielles. Soit Y une variable unidimensionnelle. On dit que la loi de Y appartient à une famille exponentielle de paramètre ρ si :

$$f_Y(y, \rho) = \exp [a(y)b(\rho) + c(\rho) + d(y)], \quad (5.1)$$

où a, b, c et d sont des fonctions réelles connues, le terme $f_Y(., \rho)$ désignant la densité de Y dans le cas continu, ou $\mathbb{P}_\rho(Y = y)$ dans le cas discret. Dans le cas particulier où $a(y) = y$, la distribution de Y est dite canonique. Le tableau ci-dessous présente les trois distributions de ce type les plus utilisées.

Loi	b	c	d	ρ
Normale	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$	μ
Binomiale	$\log\left(\frac{\pi}{1-\pi}\right)$	$n \log(1 - \pi)$	$\log(C_n^y)$	π
Poisson	$\log(\lambda)$	$-\lambda$	$-\log(y!)$	λ

FIGURE 5.3 – Valeurs des paramètres pour quelques distributions (canoniques) appartenant à la famille exponentielle.

Cette liste n'est bien sûr pas exhaustive. La proposition ci-dessous présente quelques propriétés des familles exponentielles.

Proposition 5.1 *Soit Y une variable aléatoire dont la densité $f(y, \rho)$ appartient à une famille exponentielle avec a, b, c, d fonctions réelles. On a :*

$$\mathbb{E}[a(Y)] = -\frac{c'(\rho)}{b'(\rho)}, \text{ et } \text{Var}[a(Y)] = \frac{b''(\rho)c'(\rho) - c''(\rho)b'(\rho)}{(b'(\rho))^3}. \quad (5.2)$$

PREUVE. Seul le calcul de l'espérance sera effectué. La variance s'obtient en utilisant des arguments similaires. Par définition de $f(y, \rho)$, on a :

$$\int f(y, \rho) dy = 1 \Rightarrow \frac{d}{d\rho} \int f(y, \rho) dy = 0,$$

l'intégrale étant prise sur l'ensemble des valeurs possibles de y . Dans le cas particulier où Y est une variable discrète, on pourra remplacer l'intégrale par une somme. Par ailleurs, on peut remarquer que

$$\frac{d}{d\rho} \int f(y, \rho) dy = \int \frac{d}{d\rho} f(y, \rho) dy = 0.$$

On a de plus :

$$\frac{d}{d\rho} f(y, \rho) = [a(y)b'(\rho) + c'(\rho)]f(y, \rho).$$

Finalement

$$\int \frac{d}{d\rho} f(y, \rho) dy = \int [a(y)b'(\rho) + c'(\rho)]f(y, \rho) dy = 0 \Rightarrow \mathbb{E}[a(Y)] = -\frac{c'(\rho)}{b'(\rho)}.$$

□

On pourra vérifier que cette formulation 'convient' bien aux lois évoquées ci-dessus.

5.2.2 Fonctions de liens

On s'intéresse à présent aux fonctions de lien. Ces dernières relient l'espérance des observations aux variables explicatives. Certaines fonctions de lien sont naturellement utilisées avec certaines distributions comme le montre le tableau ci-dessous.

Distribution	Fonction lien	$g(\mu)$
Normale	Identité	μ
Binomiale	Logit	$\log(\mu/(1-\mu))$
Poisson	Log	$\log(\mu)$

Le choix de cette fonction lors de l'étape de modélisation est parfois effectué de manière *ad hoc* et est le plus souvent relié à la nature du problème considéré. Historiquement, il est effectué afin de faciliter la mise en place de la procédure d'estimation. On pourra en particulier remarquer que les fonctions de liens évoquées dans le tableau ci-dessus correspondent à chaque fois à la fonction b de la famille exponentielle.

5.2.3 Modèles avec réponse binaire

De nombreuses études pratiques et théoriques s'intéressent au cas particulier où la variable de réponse Y ne prend que deux valeurs 0 et 1. Pour tout $i \in \{1, \dots, n\}$:

$$\mathbb{P}(Y_i = 1) = \pi_i \text{ et } \mathbb{P}(Y_i = 0) = 1 - \pi_i.$$

On dispose également d'un certain nombre de régresseurs ou facteurs suivant si l'on a affaire à des variables qualitatives ou quantitatives ($k-1$). Toute la problématique consiste à essayer de quantifier l'influence des différents régresseurs sur le vecteur π .

Il existe différents modèles permettant de modéliser ce type de situation. Chacun d'entre eux est encore une fois caractérisé par la fonction g représentant le lien entre la partie linéaire et la moyenne des observations. Les deux principaux sont

— le modèle **logistique** ou **logit**

$$g_1(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \quad \text{ou encore} \quad \pi_i = \frac{e^{(X\theta)_i}}{1 + e^{(X\theta)_i}}.$$

— le modèle **probit** :

$$g_2(\pi) = \Phi^{-1}(\pi) \quad \text{ou encore} \quad \pi_i = \Phi((X\theta)_i).$$

où Φ désigne la fonction de répartition d'une loi gaussienne centrée réduite.

Il est parfois utile d'utiliser la fonction de lien canonique, bien que cette dernière soit peu recommandée pour la modélisation de probabilité. Il arrive en effet que les valeurs prédites ne soient pas contenues dans l'intervalle $(0, 1)$. D'autres modèles font intervenir des fonctions plus 'exotiques'

$$g_3(\pi) = \log(-\log(1-\pi)) \quad \text{ou encore} \quad g_4(\pi) = -\log(-\log(\pi)).$$

Remarque : Ce type de modélisation est une manière intéressante de résoudre des problèmes de classification supervisée. En particulier, il est possible de donner un sens à la règle de classification à l'aide des variables explicatives. Par exemple, si on s'intéresse à la vitesse de sédimentation du plasma sanguin, on s'aperçoit (voir Figure 5.4), que cette dernière est plutôt (avec probabilité supérieure à 0.5) de type '1' dès lors que le taux de Fibrinogen est supérieur à 3.7.

5.3 Estimation

5.3.1 Estimation par maximum de vraisemblance

Commençons par rappeler le cadre de travail proposé jusqu'à présent :

- on observe un échantillon de variables de réponse Y_1, \dots, Y_n et des régresseurs $Z^{(1)}, \dots, Z^{(p)}$ associés,

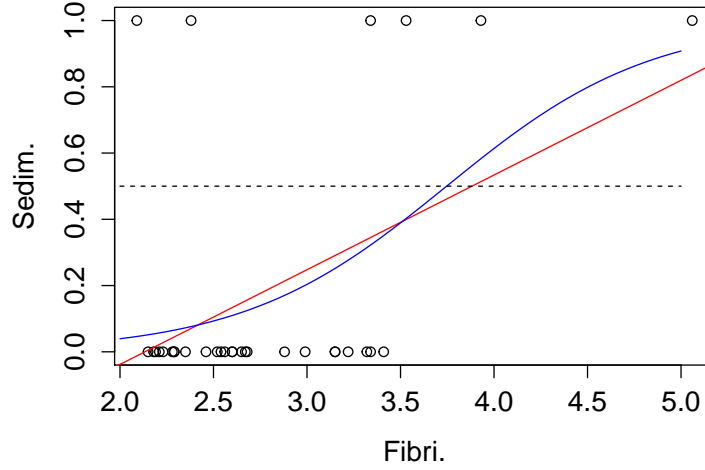


FIGURE 5.4 – Fonction de lien identité et logistique pour un modèle de type binaire

- pour tout $i \in \{1, \dots, n\}$, la variable Y_i admet pour densité

$$f_{Y_i}(y, \rho_i) = \exp[a(y)b(\rho) + c(\rho) + d(y)],$$

les ρ_i correspondant au paramétrage de la famille exponentielle (supposée ici canonique),

- pour tout $i \in \{1, \dots, n\}$, $\mathbb{E}[Y_i] = \mu_i$, les μ_i étant relié aux ρ_i par la formule (5.2),
- chaque μ_i vérifie $g(\mu_i) = \eta_i$,
- les η_i sont liés aux régresseurs par la relation $\eta_i = (X\theta)_i$, le paramètre $\theta \in \mathbb{R}^p$ étant inconnu et la matrice X correspondant à la matrice de design.

L'objectif de cette section est de proposer une méthode d'estimation pour le paramètre θ . Il se trouve que la méthode des moindres carrés n'est pas applicable dans un grand nombre de situations (excepté pour des fonctions de lien d'identité). De part la complexité du modèle, aucune formule analytique n'est disponible. Le problème d'optimisation associé n'est pas convexe et est donc difficilement implémentable en toute circonstance. Nous allons donc revenir au principe d'estimation par maximum de vraisemblance.

Dans notre cadre, la vraisemblance s'écrit

$$L(Y, \theta) = \prod_{i=1}^n f_{Y_i}(Y_i, \rho_i),$$

et l'EMV associé vérifie donc

$$\hat{\theta}_{MV} = \arg \max_{\theta \in \mathbb{R}^p} L(Y, \theta) = \arg \max_{\theta \in \mathbb{R}^p} l(Y, \theta).$$

Afin d'obtenir une expression de ce dernier, il est naturel dans un premier temps de s'intéresser au score

$$S(Y, \theta) = \frac{d}{d\theta} l(Y, \theta) = \left(\frac{\partial}{\partial \theta_1} l(Y, \theta), \dots, \frac{\partial}{\partial \theta_p} l(Y, \theta) \right)'.$$

Dans l'absolu, l'estimateur du maximum de vraisemblance vérifie

$$S(Y, \hat{\theta}_{MV}) = 0.$$

Cependant, à part pour quelques situations bien précises (typiquement une fonction de lien de type canonique), il n'existe pas de solution analytique pour cette équation. Il est cependant possible de

montrer que le problème associé à la détermination de $\hat{\theta}_{MV}$ est un problème d'optimisation convexe qui peut donc être traité par un algorithme de type Newton-Raphson, adapté à un cadre statistique.

On va avoir pour cela besoin de la matrice d'information de Fisher \mathcal{J} définie comme

$$\mathcal{J} = \left(-\mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(Y, \theta) \right] \right)_{i,j}.$$

Il est important de noter que bien souvent, \mathcal{J} dépend de θ . Lors de l'utilisation de notre algorithme (de type itératif), il convient donc de mettre à jour cette matrice à chaque étape.

Algorithme d'approximation du maximum de vraisemblance

- Initialisation : $t^{(0)}$.
- Pour tout entier m

$$t^{(m)} = t^{(m-1)} + [\mathcal{J}^{(m-1)}]^{-1} S^{(m-1)}. \quad (5.3)$$

- Arrêt quand

$$|t^{(m)} - t^{(m-1)}| \leq \Delta.$$

- on pose $\hat{\theta}_{MV} = t^{(m)}$.

Cet algorithme est implémenté dans la plupart des logiciels statistiques (SAS et R en particulier). Le paramètre Δ , permettant un arrêt de l'algorithme, est calibré par défaut dans la plupart des logiciels.

5.3.2 Calculs pour des familles exponentielles

Il est possible d'obtenir une formulation explicite pour le score et l'information de Fisher en présence de familles de distributions exponentielles.

Proposition 5.2 *Soit $S(Y, \theta) := (S_1 \dots S_p)'$ le vecteur score. Pour tout $j \in \{1, \dots, p\}$, la variable S_j est centrée et*

$$S_j = \sum_{i=1}^n \left[\frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} Z_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right].$$

PREUVE. Pour chaque Y_i , la log-vraisemblance s'écrit ici

$$l_i := l(Y_i, \rho_i) = Y_i b(\rho_i) + c(\rho_i) + d(Y_i),$$

et la log-vraisemblance de l'ensemble de l'échantillon est donc donnée par

$$l := l(\mathbf{Y}, \theta) = \sum_{i=1}^n l_i = \sum_{i=1}^n Y_i b(\rho_i) + \sum_{i=1}^n c(\rho_i) + \sum_{i=1}^n d(Y_i).$$

Pour chaque paramètre θ_j , $j \in \{1, \dots, p\}$, on a donc

$$\frac{\partial l}{\partial \theta_j} := S_j = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_j} = \sum_{i=1}^n \frac{\partial l_i}{\partial \rho_i} \cdot \frac{\partial \rho_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \theta_j}. \quad (5.4)$$

Il nous reste à calculer les trois dérivées partielles présentes dans (5.4). Soit $i \in \{1, \dots, n\}$ fixé. On a dans un premier temps

$$\frac{\partial l_i}{\partial \rho_i} = Y_i b'(\rho_i) + c'(\rho_i) = b'(\rho_i)(Y_i - \mu_i),$$

où on a utilisé la Proposition 5.1. Ensuite

$$\frac{\partial \rho_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \rho_i} \right)^{-1}$$

et

$$\frac{\partial \mu_i}{\partial \rho_i} = \frac{\partial}{\partial \rho_i} \left(-\frac{c'(\rho_i)}{b'(\rho_i)} \right) = \frac{-c''(\rho_i)}{b'(\rho_i)} + \frac{c'(\rho_i)b''(\rho_i)}{[b'(\rho_i)]^2} = b'(\rho_i)\text{Var}(Y_i).$$

Pour finir

$$\frac{\partial \mu_i}{\partial \theta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \theta_j} = \frac{\partial \mu_i}{\partial \eta_i} Z_{ij}.$$

En compilant les résultats précédents, on obtient donc

$$S_j = \sum_{i=1}^n \left[\frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} Z_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right].$$

A la vue de cette formule, il apparait clairement que les variables S_j sont centrées.

□

Une expression générale est également disponible pour la matrice d'information \mathcal{J} .

Proposition 5.3 *La matrice \mathcal{J} peut être écrite sous la forme*

$$\mathcal{J} = X'WX,$$

où W est une matrice diagonale $n \times n$ ayant pour entrées

$$w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \quad \forall i \in \{1, \dots, n\}.$$

PREUVE. Soient j, k fixés. On a

$$\begin{aligned} \mathcal{J}_{jk} &= \mathbb{E} \left\{ \sum_{i=1}^n \left[\frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} Z_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^n \left[\frac{(Y_l - \mu_l)}{\text{Var}(Y_l)} Z_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\}, \\ &= \mathbb{E} \sum_{i=1}^n \frac{(Y_i - \mu_i)^2}{[\text{Var}(Y_i)]^2} Z_{ij} Z_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \end{aligned}$$

puisque les variables Y_i sont indépendantes. Il reste alors à remarquer que $\mathbb{E}[(Y_i - \mu_i)^2] = \text{Var}(Y_i)$.

□

La Proposition 5.3 permet d'écrire l'équation (5.3) sous une forme plus parlante. En effet, cette dernière peut être exprimée sous la forme :

$$\mathcal{J}^{(m-1)} t^{(m)} = \mathcal{J}^{(m-1)} t^{(m-1)} + S^{(m-1)} \Leftrightarrow X'W_{(m)} X t^{(m)} = \mathcal{J}^{(m-1)} t^{(m-1)} + S^{(m-1)}.$$

Le membre de droite de cette équation est quand à lui un vecteur ayant pour entrées :

$$\sum_{k=1}^p \sum_{i=1}^N \frac{Z_{ij} Z_{ik}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 t_k^{(m-1)} + \sum_{i=1}^n \frac{(Y_i - \mu_i) Z_{ij}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right).$$

Ce vecteur peut donc être écrit sous la forme

$$X'Wz \text{ avec } z_i = \sum_{k=1}^p Z_{ik} b_k^{(m-1)} + (Y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^{-1},$$

où $\mu_i = \mu_i^{(m-1)}$ et $\partial \mu_i / \partial \eta_i$ sont évaluées en $b^{(m-1)}$.

Finalement, on obtient donc l'équation :

$$X'W_{(m)} X t^{(m)} = X'W_{(m)} z_{(m)}. \quad (5.5)$$

Cette équation généralise l'équation normale (2.4) établie pour le modèle linéaire.

Remarque : Dans le cas particulier où ni W ni z ne dépendent de β , l'itération est inutile. Une solution explicite est donc disponible pour l'estimateur du maximum de vraisemblance.

5.3.3 Retour au modèle linéaire

Le modèle linéaire avec erreurs gaussiennes est un cas particulier du modèle linéaire généralisé (fonction de lien de type canonique). Pour tout $i \in \{1, \dots, n\}$, on a

$$\mathbb{E}[Y_i] = \mu_i = (X\theta)_i = X_i\theta \text{ et } Y_i \sim \mathcal{N}(\mu_i, \sigma^2).$$

Les entrées de la matrice W deviennent

$$w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sigma^{-2}, \quad \forall i \in \{1, \dots, n\}.$$

En effet, on a $\eta_i = \mu_i$. La matrice W correspond donc à l'identité multipliée par la constante σ^{-2} et ne dépend donc pas du vecteur b . On s'intéresse à présent au vecteur z . Par définition de ce dernier :

$$z_i = \sum_{k=1}^p X_{ik} t_k^{(m-1)} + (Y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = \sum_{k=1}^p x_{ik} t_k^{(m-1)} + (Y_i - \mu_i) = \mu_i + (Y_i - \mu_i) = Y_i.$$

On obtient donc $z = Y$ et l'équation (5.5) devient :

$$X'Xt = X'Y.$$

On retrouve donc l'estimateur tel qu'on l'a construit dans le Chapitre 2. Ceci est complètement logique dans la mesure où l'on a vu que dans cette situation particulière, les estimateurs des moindres carrés et du maximum de vraisemblance coïncidaient.

5.4 Inférence pour le modèle linéaire généralisé

5.4.1 Loi asymptotique de l'EMV

De part la complexité du modèle linéaire généralisé, l'obtention d'un intervalle de confiance va nécessiter un peu plus de travail que dans un cadre de statistique paramétrique usuel. Encore une fois, la fonction score U va jouer un rôle prépondérant dans l'obtention de ces intervalles. Une grande partie de la construction est basée sur le résultat suivant.

Proposition 5.4 *Sous les conditions énoncées dans ce chapitre, le vecteur score $S = (S_1, \dots, S_p)'$ vérifie*

$$S' \mathcal{J}^{-1} S \xrightarrow{\mathcal{L}} \chi_p^2 \quad \text{quand } n \rightarrow +\infty.$$

En particulier,

$$S \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \mathcal{J}),$$

pour de grandes valeurs de n , où $\mathcal{N}_p(0, \Sigma)$ désigne une loi normale multivariée de moyenne O et de matrice de covariance Σ .

PREUVE. Par soucis de simplicité, on effectue simplement la preuve en dimension 1. Dans ce cas, on rappelle que

$$S = \sum_{i=1}^n D_i, \quad \text{avec } D_i := \left[\frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} Z_i \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right], \quad \forall i \in \{1, \dots, n\}.$$

Dans ces conditions, il est facile de voir que les variables D_i sont centrées et indépendantes. Par définition de l'information de Fisher, et par une application directe du TCL version Lindeberg (cf Appendice B), on a

$$\frac{S}{\sqrt{\mathcal{J}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{ou encore} \quad \frac{S^2}{\mathcal{J}} \xrightarrow{\mathcal{L}} \chi_1^2.$$

La généralisation au cas multidimensionnel est (presque) évidente.

□

Le résultat précédent nous donne donc la normalité asymptotique du score. A ce stade, comment exploiter ce résultat afin de produire une inférence sur le vecteur d'intérêt θ . Afin de donner un premier élément de réponse, on va pour cela à nouveau se placer dans un cadre unidimensionnel.

Par la suite, on note $S = S(\theta)$ pour θ fixé, et ce afin d'insister sur la dépendance en θ du vecteur score. Soit $\tilde{\theta}$ un vecteur 'proche' de θ (dans un sens qu'il conviendrait de préciser). A l'aide d'un développement de Taylor, on a

$$S(\theta) \simeq S(\tilde{\theta}) + (\theta - \tilde{\theta})S'(\tilde{\theta}).$$

Si on applique cette formule à $\tilde{\theta} = \hat{\theta}_{MV}$, on obtient

$$S(\theta) \simeq (\theta - \hat{\theta}_{MV})S'(\theta),$$

dans la mesure où, par définition, $S(\hat{\theta}_{MV}) = 0$. En se rappelant que $\mathcal{J} = -\mathbb{E}[S'(\theta)]$, on obtient l'approximation en loi suivante

$$\theta - \hat{\theta}_{MV} \stackrel{\mathcal{L}}{\simeq} \mathcal{N}(0, \mathcal{J}^{-1}) \Leftrightarrow \hat{\theta}_{MV} \stackrel{\mathcal{L}}{\simeq} \mathcal{N}_m(\theta, \mathcal{J}^{-1}), \text{ quand } n \rightarrow +\infty. \quad (5.6)$$

Dans un cadre plus général, la proposition suivante décrit la loi asymptotique de l'estimateur du maximum de vraisemblance.

Proposition 5.5 *Sous les conditions énoncées dans ce chapitre, on a*

$$\mathcal{W} := (\hat{\theta}_{MV} - \theta)' \mathcal{J}(\hat{\theta}_{MV} - \theta) \stackrel{\mathcal{L}}{\rightarrow} \chi_p^2, \text{ quand } n \rightarrow +\infty.$$

La variable est appelée **statistique de Wald**.

Remarque : Dans le cas particulier où la distribution des Y_i est gaussiennes et la fonction de lien canonique, il est possible de montrer que l'estimateur du maximum de vraisemblance est lui aussi gaussien et ce sans avoir recours à l'approximation (5.6). Si maintenant les erreurs ne sont pas gaussiennes, le résultat précédent propose une alternative intéressante aux tests de Fisher.

5.4.2 Test du rapport de vraisemblance

Comme pour le modèle linéaire classique, il peut être pertinent de vouloir simplifier le modèle, ou dit autrement de comparer d'un point de vue statistique des modèles emboîtés. Plus formellement, on note

$$S_i = g(\mu_i) \quad \forall i \in \{1, \dots, n\}.$$

Etant donnée une matrice X_0 telle que $[X_0] \subset [X]$, on souhaite alors tester

$$H_0 : S \in [X_0] \quad \text{contre} \quad H_1 : S \in [X].$$

Ce type d'approche permet en particulier de tester globalement l'intérêt de la prise en compte des régresseurs sur la réponse Y , i.e. de tester

$$H_0 : g(\mu_i) = a \quad \text{contre} \quad H_1 : g(\mu_i) = (X\theta)_i.$$

Le test de Fisher ne permet absolument pas, sauf cas particulier, de répondre à cette thématique. La construction de ce dernier est en effet complètement conditionnée par le caractère gaussien des observations. Une alternative possible consiste à utiliser à nouveau la vraisemblance pour accompagner notre prise de décision.

Par la suite, étant donné un modèle M_0 (resp. M_1) ou de manière équivalente un ensemble de valeurs possibles Θ_0 (resp. Θ_1) pour θ , on désigne par $L(\mathbf{Y}, \theta)$ la vraisemblance pour θ fixé et on appelle la quantité

$$T = \frac{\sup_{\theta \in \Theta_1} L(\mathbf{Y}, \theta)}{\sup_{\theta \in \Theta_0} L(\mathbf{Y}, \theta)},$$

le rapport de vraisemblance entre les 'modèles' Θ_0 et Θ_1 . Par construction, dès lors que $\Theta_0 \subset \Theta_1$, on aura $T \geq 1$. Si T prend de grandes valeurs, c'est que le modèle Θ_0 est trop simpliste, ce qui conduira à un rejet de l'hypothèse nulle. La construction d'une règle de décision rigoureuse repose sur la loi asymptotique de ce rapport de vraisemblance. Sous un certain nombre d'hypothèses sur le modèle, il est possible de montrer que

$$T^* := 2 \log(T) \xrightarrow{\mathcal{L}} \chi_{p-p_0}^2 \quad \text{quand } n \rightarrow +\infty,$$

où p (resp. p_0) désigne la dimension de Θ (resp. Θ_0).

5.4.3 Sélection de modèle

On termine ce chapitre par quelques éléments de sélection de modèle. La littérature sur ce sujet est un peu moins riche que pour le modèle linéaire classique. On ne s'attardera ici que sur les stratégies existantes, sans forcément s'intéresser aux performances théoriques associées.

Une première approche naïve consiste à construire un indice mesurant l'adéquation entre le modèle et les données observées, sur le principe du coefficient d'ajustement R^2 dans le modèle linéaire. Pour tout modèle $\Theta_m \subset \Theta$, on définit la déviance $D(m)$ comme

$$D(m) = 2 \log \left(\frac{\sup_{\theta \in \Theta} L(\mathbf{Y}, \theta)}{\sup_{\theta \in \Theta_m} L(\mathbf{Y}, \theta)} \right),$$

cette quantité mesurant l'écart entre le modèle complet (ou saturé) et le modèle m considéré. On pose ensuite

$$\tilde{R}_m^2 = \frac{D(m_0) - D(m)}{D(m_0)},$$

où m_0 désigne le modèle réduit à un effet constant (on parle de modèle nul). Le terme \tilde{R}_m^2 est appelé *pseudo- R^2* . Il reprend en partie les propriétés du coefficient d'ajustement construit pour le modèle linéaire. En particulier, cette quantité est comprise entre 0 et 1, et plus \tilde{R}_m^2 est proche de 1, plus l'adéquation aux données est 'bonne'.

De manière un peu plus rigoureuse, il est possible d'étendre la construction des critères AIC et BIC au cadre considéré dans ce chapitre. On se souvient en effet que ces deux critères peuvent être formulés à partir de la notion de vraisemblance. Sur la base de ce qui a été discuté précédemment, on posera donc pour un modèle m donné

$$AIC(m) = -2l(\mathbf{Y}, \hat{\theta}_m) + 2|m|,$$

et

$$BIC(m) = -2l(\mathbf{Y}, \hat{\theta}_m) + 2|m| \log(n).$$

On retiendra alors le modèle \hat{m} minimisant un de ces deux critères.

5.5 Exercices

Exercise 1. The goal of this exercise is to explain the mortality rate of a given ladybird specie by the degree of exposure to a chemical product, here carbon disulfide. The experimental design is the following : each ladybird is exposed during 4 hours to a given concentration of this gas. At the end, we count the number of death. The results are gathered below :

Concentration ($\log_{10} CS_2 mg l^{-1}$)	Ladybird number	Dead number
1.69	59	6
1.78	56	28
1.88	60	60

We use a logistic regression model to explain the experiment described above.

1. Provide a sharp description of the model.
2. Draw (approximately) on a picture the mortality rate in terms on the carbon disulfide concentration.
3. Compute (by hands!) the log-likelihood, score functions, and the Fisher information matrix.
4. Construct an estimator of the unknown parameters and provide the estimation associated to this sample. For the sake of simplicity, one can only use the first iteration of the iterative algorithm.

Exercise 2. We are interested here in the life expectation of 17 patients suffering from leukaemia. The following table links the life expectation (Y_i variable, in weeks since the detection of the disease) and the X_i variable corresponding to the \log_{10} of the initial number of white blood cells.

y_i	65	156	100	134	16	108	121	4	39
x_i	3.36	2.88	3.63	3.41	3.78	4.02	4	4.23	3.73

y_i	143	56	26	22	1	1	5	65
x_i	3.36	2.88	3.63	3.41	3.78	4.02	4	4.23

1. We decide to model the relationship between the X_i and the Y_i through the equality $\mathbb{E}[Y_i] = e^{-\beta_1 - \beta_2 X_i}$. What is the associated link function?
2. We assume that the Y_i are distributed according to an exponential distribution. Compute the associated score and Fisher information matrix.
3. Provide an analysis of this dataset.

Exercise 3. Provide a complete analysis of the dataset concerning the sedimentation rate of the blood plasma contained in the file `plasma.dat`. For each patient, we measure the *fibrinogen* (1st column) and γ -*globulin* (2nd column) rates, and the sedimentation rate for the blood plasma (we indicate 1 if this rate is greater than 20mm/h, and 0 otherwise).

Exercise 4. The dataset `ozkidsm.dat` gathers data regarding Australian scholar absenteeism. For each individual of this sample, we indicate in the first column its ethnical group (A for arborigean, N otherwise) - its gender, 2nd column - study level, 3rd column - and its performances (SL for *slow learner* and AL for *average learner* in the 5th column). The first column provide all the possible combinations of the explanatory variables. One would like to explain the absence number during a reference period (last column).

Chapitre 6

Analyse de variance et plans d'expériences

6.1 Pourquoi planifier l'expérience ?

Une très grande partie de la théorie statistique est orientée vers le traitement de l'information. A partir d'un certain nombre de données (récupérées sous diverses formes), le but est d'essayer d'obtenir le plus d'informations possibles sur le modèle sous-jacent à travers des estimations de paramètres, tests, résultats de convergence et une interprétation éventuelle des résultats obtenus.

Cependant, n'est-il pas possible de préparer la collecte des données, i.e. de faire un travail en amont pour améliorer les prévisions futures et rendre l'interprétation plus claire ? C'est tout l'objet de la théorie des plans d'expériences.

Avant toutes choses, quelques notions (rappels) de vocabulaire :

Définition 6.1

- On appelle **facteur** une cause assignable susceptible d'affecter les valeurs observées. Les facteurs sont en général qualitatifs.
- Les valeurs prises par un facteur sont appelées **modalités**.
- Une **unité expérimentale**, ou encore **parcelle**, est une partie expérimentale sur laquelle on applique un traitement donné.
- On appelle **répétition** l'ensemble des unités recevant la même traitement. Une expérience est dite **factorielle** si on effectue toutes les combinaisons possibles.

Historiquement, les plans ont été développés pour le domaine de l'agronomie. Le vocabulaire utilisé a ensuite été généralisé aux autres domaines d'études.

Parmi les facteurs, on appellera **facteurs contrôlés** les facteurs fixés par l'expérimentateur. La principale motivation des plans d'expériences est de fixer ces facteurs de façon à obtenir les meilleurs résultats statistiques possibles. Les **facteurs incontrôlés** sont ceux pouvant être observés mais non déterminés par l'expérimentateur. Ils sont alors pris en compte dans l'étude sans pour autant pouvoir les planifier.

EXEMPLE 4.1 (Agronomie) On désire étudier le rendement d'une variété de blé suivant le type d'engrais utilisé. Les facteurs correspondent donc à ces différents engrais. Ils est également envisageable suivant les situations de faire intervenir les facteurs ensoleillement, pluviométrie, etc... Ces derniers facteurs sont souvent incontrôlés.

EXEMPLE 4.2 (Comparatif) Une association de consommateurs désire comparer les propriétés gustatives de plusieurs marques de fromage (ce sont les facteurs). Comment comparer ces dernières sachant que chaque testeur ne peut raisonnablement goûter tous les fromages ?

Dans la suite de ce chapitre, nous allons présenter les plans d'expériences les plus classiques. Chacun de ces plans peut être mis en place dans une situation bien précise : tout dépend du nombre de facteurs, de répétitions et de contraintes sur les parcelles. Nous allons en particulier nous intéresser aux configurations suivantes :

- 1 facteur et t modalités (plan en randomisation totale),
- 2 facteurs : 'traitement' et 'bloc' (plan en blocs complets),
- même situation mais le nombre d'unités disponibles pour chaque modalité du facteur 'bloc' et limité par rapport au facteur traitement (plans en blocs incomplets),
- 3 facteurs (plans en ligne et colonnes, carrés latins).

Les performances statistiques associées à ces différents plans seront étudiées en TD.

Par soucis de simplicité et de concision, on se placera au cours de ce chapitre dans un contexte d'analyse de variance : le modèle est linéaire et tous les facteurs sont qualitatifs.

6.2 Contraintes et décomposition

Avant d'aller plus en avant dans la présentation et l'étude des plans d'expérience, il convient d'étudier de manière plus précise le modèle d'analyse de variance. Certaines propriétés et difficultés ont en effet été occultées dans les chapitres précédents. Nous allons introduire la notion d'orthogonalité et présenter une brève méthodologie de test.

6.2.1 Orthogonalité

La notion d'orthogonalité apporte beaucoup de facilités dans l'étude du modèle linéaire.

Définition 6.2 Soient E_1, \dots, E_p des sous-espaces vectoriels de \mathbb{R}^k , de dimensions l_1, \dots, l_p . Notons $R = X(\mathbb{R}^k)$ et $R_i = X(E_i)$, P_i et P_0 les projecteurs orthogonaux de \mathbb{R}^k sur R_i et R .

Le modèle linéaire est dit orthogonal pour E_1, \dots, E_p si les sous espaces vectoriels R_i sont deux à deux orthogonaux.

Par la suite, on note $\theta = (\theta_1, \dots, \theta_p)'$ la partition de θ de taille (l_1, \dots, l_p) et $(\hat{\theta}_1, \dots, \hat{\theta}_p)'$ les estimateurs associés.

Proposition 6.1 On considère un modèle linéaire orthogonal pour une partition E_1, \dots, E_p et $\theta = (\theta_1, \dots, \theta_p)'$ la partition de θ associée. Dans ce cas, les estimateurs des moindres carrés $\hat{\theta}_1, \dots, \hat{\theta}_p$ sont indépendants (non-corrélés sans hypothèse gaussienne). Par ailleurs, l'estimateur $\hat{\theta}_i$ n'est pas affecté par une hypothèse faite sur les paramètres θ_j pour $j \neq i$.

PREUVE. D'après les hypothèses, le modèle peut être écrit sous la forme :

$$Y = X_1\theta_1 + \dots + X_p\theta_p + \epsilon.$$

De plus, l'orthogonalité implique que

$$P_{[X]}Y = P_{[X_1]}Y + \dots + P_{[X_p]}Y.$$

Pour tout $i \in \{1, \dots, p\}$, on a donc

$$X_i\hat{\theta}_i = P_{[X_i]}Y.$$

Le théorème de Cochran nous donne l'indépendance. Par ailleurs, on

$$\hat{\theta}_i = (X_i'X_i)^{-1}X_i'Y, \quad \forall i \in \{1, \dots, p\}.$$

Chaque estimateur ne dépend donc que des 'coordonnées' correspondantes.

□

La notion d'orthogonalité est fondamentale car elle permet d'obtenir des estimateurs $\hat{\theta}_i$ non-corrélés et donc indépendant dans le cas gaussien. L'orthogonalité est ainsi une des notions les plus intéressantes pour un plan d'expériences X .

Lorsque le modèle est singulier, il est nécessaire de rajouter des contraintes. Il est alors raisonnable d'effectuer cette démarche en tenant compte de la partition, soit $C_i(\theta_i) = 0$ où $X_i/Ker(C_i)$ sont injectives.

Définition 6.3 *On dit que le modèle est orthogonal pour la partition considérée, muni des contraintes C_i si les images de $Ker(C_i)$ par $X_i/Ker(C_i)$ sont orthogonales deux à deux.*

Cette notion est proche du cas régulier. Cependant, la notion d'orthogonalité dépend des contraintes choisies. L'idée sera en général de choisir des contraintes qui rendent le modèle orthogonal.

6.2.2 Modèle croisé et modèle additif

Dans cette section, nous nous intéressons à l'analyse de variance en présence de deux facteurs. Avant d'en dire plus sur les cas qui vont nous intéresser, il convient d'explorer les propriétés du modèle linéaire dans ce type de situation.

Nous supposons dans ce paragraphe que nos deux facteurs (parfois appelés A et B) possèdent respectivement I et J modalités. Les modèles présentés ici pourront aisément être étendus à des études impliquant un plus grand nombre de facteurs.

Soit n_{ij} le nombre de parcelles où on a utilisé respectivement la i ème et la j ème modalité des facteurs A et B . Le **modèle croisé** s'exprime alors sous la forme :

$$Y_{ijk} = \theta_{ij} + \epsilon_{ijk}, \quad i = 1 \dots I, \quad j = 1 \dots J \text{ et } k = 1 \dots n_{ij}, \quad (6.1)$$

ou encore

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad i = 1 \dots I, \quad j = 1 \dots J \text{ et } k = 1 \dots n_{ij}, \quad (6.2)$$

si l'on souhaite faire intervenir des effets différentiels. Le modèle (6.1) ou alternativement (6.2) est appelé modèle croisé (à deux facteurs). Le modèle est dit **complet** si $n_{ij} > 0$ pour tout i, j et **avec répétition** si il existe i, j tels que $n_{ij} > 1$.

Le modèle (6.2) est bien plus 'expressif' que (6.1). Les termes γ_{ij} (appelés effets croisés de A et B) sont particulièrement intéressants. Ils permettent de modéliser les interactions possibles entre les deux facteurs A et B . Les termes α_i et β_j désignent respectivement l'effet différentiel additif des deux facteurs. En revanche, ce modèle est souvent singulier : on dispose de $I \times J$ observations pour $1 + I + J + IJ$ paramètres à estimer. Il est donc nécessaire dans ce cas d'imposer des contraintes sur les paramètres. Ceci pose un réel problème dans la mesure où la modélisation... et donc les résultats ne seront pas uniques. Il y a donc une part d'arbitraire qu'il faut comprendre pour pouvoir prendre les 'bonnes' décisions.

Décomposition de type I

Cette configuration est associée aux contraintes suivantes :

$$\sum_{i=1}^I n_{i.} \alpha_i = \sum_{j=1}^J n_{.j} \beta_j = \sum_{i=1}^I n_{ij} \alpha_i \beta_j = \sum_{ij} n_{ij} \gamma_{ij} = 0, \quad (6.3)$$

avec

$$n_{i.} = \sum_{j=1}^J n_{ij}, \text{ et } n_{.j} = \sum_{i=1}^I n_{ij}.$$

Avec les contraintes (6.3), en raison de l'orthogonalité, les résultats vont être assez faciles à obtenir. On peut dans un premier temps remarquer que $\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = Y_{ij.}$ pour tout couple i, j . De

plus, les estimateurs obtenus sont non corrélés et ne dépendent pas des contraintes sur les autres éléments de la partition. On estime alors μ en supposant que $\alpha_i = \beta_j = \gamma_{ij} = 0$ pour tout i, j . L'EMC dans le modèle $Y_{ijk} = \mu + \epsilon_{ijk}$ nous donne $\hat{\mu} = \bar{Y} = Y_{...}$. De la même manière, en supposant $\beta = \gamma = 0$, on obtient $\hat{\alpha}_i = Y_{i..} - \bar{Y}$, etc... Au final, pour tout couple (i, j) , on obtient donc les estimateurs :

$$\hat{\mu} = Y_{...}, \hat{\alpha}_i = Y_{i..} - \bar{Y}, \hat{\beta}_j = Y_{.j.} - \bar{Y}, \text{ et } \hat{\gamma}_{ij} = Y_{ij.} - \bar{\mu} - \hat{\alpha}_i - \hat{\beta}_j.$$

En appelant A et B les effets associés respectivement aux vecteurs α et β , les tests sous cette décomposition considèrent les hypothèses suivantes :


H_1	H_0
A	aucun effet
A,B	A
A,B,AB	A,B

Sous , la décomposition de Type I est obtenue grâce à la commande `anova`.

Décomposition de Type II

Un des principaux défauts de la décomposition précédente est l'absence de symétrie entre les différents facteurs. La décomposition de Type II permet de pallier à ce problème. Les hypothèses testées sont les suivantes :

H_1	H_0
A,B	B
A,B	A
A,B,AB	A,B

Sous , ce type de décomposition peut s'obtenir grâce à la commande `Anova` (librairie `car`).


Décomposition de type III

Dans ce cadre, on impose les contraintes

$$\sum_{i=1}^I \alpha_i = \sum_{j=1}^J \beta_j = \sum_{i=1}^I \gamma_{ir} = \sum_{j=1}^J \gamma_{sj} = 0 \quad \forall r, s \in \{1, \dots, I\} \times \{1, \dots, J\}. \quad (6.4)$$

Cette décomposition ne rend pas toujours le modèle orthogonal. Il est même possible de montrer que cette propriété n'est vérifiée que si $n_{ij} = cte$: le modèle est alors dit équiréparté. La décomposition de type III est la plus utilisée d'un point de vue pratique. Elle est d'ailleurs implémentée par défaut dans des logiciels comme SAS. Les hypothèses testées sont :

H_1	H_0
A,B,AB	B,AB
A,B,AB	A,AB
A,B,AB	A,B

Il est possible de remarquer qu'utiliser une décomposition de Type III revient à effectuer un test de Student (ou Fisher) pour chacun des paramètres. Sous , ce type de manipulation peut par exemple s'effectuer par l'intermédiaire de la commande `summary`.

Une question naturelle se pose donc : quelle approche utiliser ? Clairement, pour un modèle équiréparté, il y a unicité de la table d'analyse de variance. Les deux décompositions coïncident et les sommes de carrés résiduels sont les mêmes. Ce n'est plus le cas pour un modèle non-équiréparté. Il faut dans ce cas regarder précisément la question posée par chacun des test et choisir la décomposition qui collera le mieux à nos attentes.

Il est souvent judicieux, lorsque l'on amorce une étude sur ce type de modèle de vérifier si l'interaction entre les deux facteurs est significative, i.e. tester l'hypothèse " $H_\gamma : \gamma = 0$ " (voir

Chapitre 2 pour la mise en place du test). Si l'hypothèse H_γ est acceptée, on obtient un **modèle additif**, pouvant s'exprimer sous la forme :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \quad i = 1 \dots I, \quad j = 1 \dots J \text{ et } k = 1 \dots n_{ij}, \quad (6.5)$$

Ce modèle ne suppose pas d'interaction entre les différents facteurs.

Remarque : Le modèle croisé complet sans répétition peut s'avérer délicat à manipuler. Il est en effet difficile de mesurer les interactions γ_{ij} dans ce cas puisque une seule observation est disponible pour chaque couple i, j . Il est dans ce cas recommandé de travailler directement avec un modèle additif.

6.3 Exemples de plans

6.3.1 Plans en randomisation totale

On se place ici dans un cadre d'analyse de variance à un facteur et t modalités. Chaque modalité sera répétée r fois. On utilise donc le modèle :

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, t \quad j = 1, \dots, r.$$

Le principe de randomisation totale permet d'attribuer chaque modalité à une parcelle précise, le tout de manière aléatoire. L'avantage est de pouvoir mettre en place une analyse sans que cette dernière soit perturbée par un intervenant extérieur (a priori sur un produit, interprétation hâtive des résultats, etc...). On tire donc au hasard r parcelles sans remise pour la première modalité, puis encore r parcelles distinctes pour la seconde, et ainsi de suite.

Ce type de plan est particulièrement utilisé dans le milieu médical afin de lutter contre l'effet placebo. Il est mieux connu sous le nom d'étude en double aveugle.

EXEMPLE 4.2. Un laboratoire pharmaceutique souhaite mettre en place un nouveau traitement pour une maladie donnée. Avant de valider cette formule et éventuellement de commercialiser ce produit, il est nécessaire de s'assurer que ce médicament a réellement un effet. On dispose pour cela de 10 patients atteints par la maladie en question. Ces patients jouent le rôle de 'parcelles' et on s'intéresse au facteur 'traitement'. Le médicament (modalité 1) est distribué de manière aléatoire à la moitié des sujets, le reste se voyant administrer une solution sans aucun effet (modalité 0). Le tableau suivant présente une réalisation possible de ce plan en randomisation totale :

Patient	1	2	3	4	5	6	7	8	9	10
Traitement	0	0	1	0	1	1	1	1	0	0

Ni les patients, ni les médecins ne doivent être au courant de la répartition du facteur traitement.

6.3.2 Plans en blocs complet

Un groupe important de plans d'expérience est rangé dans la catégorie des plans en blocs. C'est aussi par ce type d'approche que la théorie des plans d'expérience a pris historiquement son envol. Le principe est de regrouper les unités expérimentales en b blocs, de préférence aussi homogènes que possible. Cette idée provient de l'agronomie. On peut par exemple souhaiter étudier la production de maïs par rapport à un nouvel engrais. On met en place ce type d'expérience sur différentes parcelles, les conditions climatiques pouvant ne pas être les mêmes.

Cette approche permet en général de réduire la variance des observations. Le modèle utilisé fait intervenir un facteur 'traitement' et un facteur supplémentaire 'bloc'. Ce dernier permet en général de contrôler les conditions expérimentales, sans pour autant être l'objet principal de l'étude : c'est au final le facteur 'traitement' qui nous intéresse.

EXEMPLE. Reprenons l'exemple 4.2. Le laboratoire pharmaceutique en question pense que le sexe des patients est susceptible d'avoir une influence sur l'efficacité du médicament. On peut dans ce cas construire deux blocs : 'homme' et 'femme'. L'étude est alors menée en tenant compte de ces

deux facteurs. Le facteur qui nous intéresse est toujours le 'traitement' mais on autorise une dépendance par rapport au sexe des patients : le modèle est plus riche. Le même type d'expérience est envisageable en considérant cette fois-ci des blocs 'poids'.

Lorsque l'on utilise des plans en blocs complets, il est courant d'imposer un modèle additif : on considère qu'il n'y a pas d'interaction 'bloc/traitement'. Le modèle est donc en général :

$$Y_{ijk} = \alpha_i + \beta_j + \epsilon_{ijk}, \quad i = 1 \dots I, \quad j = 1 \dots J, \quad k = 1 \dots n_{ij},$$

où I représente le nombre de traitements, J le nombre de blocs et $n_{ij} > 1$ (le plan est complet). Il est cependant aussi envisageable d'utiliser des modèles croisés.

EXEMPLE. Une usine fabriquant des crochets en métal se voit proposer trois types de machines, appelées A, B et C. Chaque machine requiert un ouvrier pour la faire fonctionner. La direction souhaite déterminer quelle machine est la plus efficace, i.e. laquelle permet de produire le plus grand nombre de crochets pour une période fixée. Trois ouvriers sont disponibles (1, 2 et 3). Il semble assez naturel chaque ouvrier réalise trois séquences de fabrication de crochets. Ce type d'approche peut mener au plan suivant :

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

Ce type de plan peut poser problème. Si il s'avère que la machine C est plus performante, cela provient-il de l'ouvrier ou de la machine ?

Afin d'éviter ce type de conflit, chaque ouvrier va utiliser chacune des machines à tour de rôle. On obtient par exemple le plan :

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

Il est enfin toujours recommandé d'utiliser une part de randomisation : l'ordre dans lequel les ouvriers les machines peut perturber les résultats (effet placebo, fatigue,...). On peut dans ce cas, randomiser les lignes, i.e. l'ordre d'utilisation des machines. On arrive alors par exemple au plan :

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

Il est tout à fait envisageable d'étendre ce type de plan au cas où plusieurs blocs sont impliqués dans l'étude.

6.3.3 Plans en blocs incomplets

Nous avons présenté deux types de plans dit factoriels, au sens où toutes les combinaisons de modalité sont représentées. Nous allons à présent nous intéresser dans les deux paragraphes suivants aux plans incomplets (ou non-factoriels). On dispose toujours de deux facteurs : un facteur 'traitement' et un facteur 'bloc'. Certaines expériences sont compliquées au sens où chaque facteur implique un très grand nombre de modalités : il est alors difficile d'envisager de pouvoir utiliser toutes les combinaisons de modalités. D'autres expériences imposent des contraintes sur les modalités du facteur 'bloc' : ces dernières sont moins nombreuses que les modalités du facteur 'traitement'.

Dans ce type de configuration, l'idée consiste donc à construire des plans plus réduits, mais qui vont idéalement conserver de bonnes propriétés statistiques. Il est raisonnable de considérer un modèle additif :

$$Y_{ijk} = t_i + b_j + \epsilon_{ijk}, \quad i = 1, \dots, t \quad j = 1, \dots, b \quad k \leq n_{ij}.$$

Le cas $n_{ij} = 0$ signifie qu'aucune observation n'est disponible pour le couple (t_i, b_j) .

Définition 6.4 *Un plan en blocs complets équilibré est un plan en blocs de mêmes dimensions k tel que :*

- *chaque traitement apparait au plus une fois par bloc ($n_{ij} = 0$ ou 1),*
- *la taille des blocs k est inférieure au nombre de traitements t ,*
- *chaque traitement admet le même nombre de répétitions r ,*
- *le nombre de blocs où une paire quelconque de traitement apparait est égal à λ .*

Par ailleurs, les paramètres t, b, k, r, λ vérifient les équations :

$$r.t = k.b \text{ et } r(k-1) = \lambda(t-1).$$

Utiliser un plan équilibré présente de nombreux avantages. En particulier, la qualité d'estimation sera identique pour tout les contraste, ce qui permet de rendre une éventuelle analyse plus claire. Suivant la valeur des différents paramètres du plan, il n'est cependant pas toujours possible de construire un plan équilibré.

Plans circulants

Voici maintenant un exemple simple de plan en blocs incomplets équilibré, lorsque l'on dispose de t traitement répartis en t blocs de taille $t-1$. Il s'agit des plans circulants. Le principe est le suivant : pour construire les blocs, on élimine tour à tour chacun de traitements. Il est possible de montrer qu'un plan circulant est toujours équilibré et possède un indice de concurrence λ égal à $t-2$.

EXEMPLE. On souhaite comparer les performances de 4 engrais différents (E_1, E_2, E_3 et E_4). Pour effectuer les tests, on dispose de 4 parcelles, chacune étant découpée en trois zones de même taille. Voici alors un exemple de plan circulant pour cette configuration :

Unité 1	Unité 2	Unité 3
E_1	E_2	E_3
E_1	E_2	E_4
E_1	E_3	E_4
E_2	E_3	E_4

Une randomisation des lignes permettra ensuite de conserver les propriétés du plan, tout en évitant un effet 'placebo'.

Plans lattices

Nous allons voir maintenant une méthode permettant de construire des plans dans la situation où l'on dispose de p^2 traitement (avec p nombre premier ou puissance d'un nombre premier) pour p blocs contenant chacun p unités.

Aux p^2 traitements sont associés deux facteurs A et B à p niveaux, ces niveaux étant numérotés par des éléments du corps de Galois $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. Un traitement est donc en fait indicé par une paire (i, j) à valeur dans \mathbb{F}_p^2 . Les répétitions sont indicées par $\mathbb{F}_{r-1} \cup \{\infty\}$, les blocs dans chaque répétition sont indicés par $k \in \mathbb{F}_p$ de la manière suivante :

- lors de la répétition ∞ , on affecte au block $k \in \mathbb{F}_p$ les traitements (i, j) tels que $j = k$.
- lors de la répétition $\bar{0}$, on affecte au block $k \in \mathbb{F}_p$ les traitements (i, j) tels que $i = k$.
- lors de la répétition $l \in \mathbb{F}_{r-1} \setminus \{\bar{0}\}$, on affecte au bloc $k \in \mathbb{F}_p$ les traitements (i, j) tels que $i + l.j = k$.

Le plan ainsi construit est appelé (r, p^2) lattice.

EXEMPLE. Un $(3, 2^2)$ lattice est un plan construit pour $2^2 = 4$ traitement, répartis sur 2 blocs contenant chacun 2 unités. Chaque blocs est répété trois fois, ce qui nous laisse donc 6 blocs de 2 unités disponibles.

Tous les plan lattices ne sont pas équilibrés. Il conviendra donc de vérifier les equations de la définition.

6.3.4 Plans en carré latin

Nous allons maintenant considérer la situation où deux facteurs blocs interviennent dans l'expérience pour un nombre de traitements supérieur ou égal à 2. On suppose par ailleurs que chaque

facteur possède le même nombre de modalités p .

Définition 6.5 *Un plan en carré latin implique p traitements, répartis suivant les deux facteurs 'blocs' appelés 'ligne' et 'colonne'. Chaque traitement doit apparaître exactement une fois dans chaque ligne et chaque colonne.*

EXEMPLE. Un fabricant de pneu de voiture s'intéresse à l'usure de ses produits. Ils dispose de quatre modes opératoires pour la fabrication et souhaite sélectionner celui offrant la meilleur résistance à l'usure. Il utilise pour cela quatre voitures test V_1, V_2, V_3 et V_4 , chacune ayant des caractéristiques différentes. Par ailleurs, on sait que la position du pneu sur la voiture affecte son usure. On dispose donc d'un facteur 'traitement' correspondant au mode de fabrication de chaque pneu et de deux facteurs blocs : 'voiture' et 'position'. Chaque facteur possède quatre modalités. Le plan en carré latin est donc construit de la manière suivante :

	Avant g.	Avant d.	Arrière g.	Arrière d.
V_1	A	B	C	D
V_2	D	A	B	C
V_3	C	D	A	B
V_4	B	C	D	A

Le modèle considéré pour ce type de plan est additif : le peu d'observations disponibles ne permettent pas de modéliser d'éventuelles interactions. Le modèle est le suivant :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk},$$

où les i, j, k ne sont pas tous observés. Pour avoir l'orthogonalité, il nous faut en particulier :

$$\langle X(0 \ \alpha \ 0 \ 0)', X(0 \ 0 \ \beta \ 0)' \rangle = 0 \Leftrightarrow \langle \alpha, \beta \rangle = 0 \Leftrightarrow \left(\sum_{i=1}^n \alpha_i \right) \left(\sum_{j=1}^n \beta_j \right) = 0.$$

De manière générale, on va donc imposer la contrainte :

$$\sum_{i=1}^n \alpha_i = \sum_{j=1}^n \beta_j = \sum_{k=1}^n \gamma_k = 0,$$

qui rend donc le modèle orthogonal. On obtient alors les estimateurs :

$$\hat{\mu} = Y_{...}, \hat{\alpha}_i = Y_{i..} - \hat{\mu}, \hat{\beta}_j = Y_{.j.} - \hat{\mu}, \text{ et } \hat{\gamma}_k = Y_{..k} - \hat{\mu},$$

pour tout $i, j, k \in \{1, \dots, n\}$.

Lorsque l'on est en présence de trois facteurs 'blocs', il est possible de généraliser ce principe sur le même modèle : on parle alors de plan greco-romain.

6.4 Critères d'optimalité

Nous avons vu dans cette partie quelques exemples de plan d'expériences correspondant aux approches les plus utilisées. Une question naturelle vient à l'esprit : pourquoi utiliser ceux-là plutôt que d'autres ? La plupart du temps, le bon sens permet rapidement de trancher. Il est quand même utile, voire nécessaire de se doter d'outils rigoureux permettant de trancher entre plusieurs plans d'expériences donnés.

Avant de parler d'optimalité, il faut s'intéresser à la qualité des estimateurs des moindres carrés. Le risque quadratique est donné par

$$\mathbb{E} \|\hat{\theta}_{MC} - \theta\|^2 = \sigma^2 \text{Trace}((X'X)^{-1}).$$

On mesure à travers cette quantité la distance entre notre estimateur et la cible θ . On sent donc bien que plus la matrice $(X'X)^{-1}$ possède une petite trace, plus l'estimateur sera précis. C'est

donc vers ce type de design qu'il faut s'orienter.

Une deuxième manière de contrôler la pertinence d'un estimateur est de s'intéresser à sa variance, en l'occurrence ici à la matrice de covariance $V(\hat{\theta}_{MC})$. Cette dernière est donnée par

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

Il est alors courant, afin de contrôler cette dernière, de demander un 'grand' déterminant pour la matrice $(X'X)$.

En résumé, nous retiendrons deux critères d'optimalité :

- critère du plus grand déterminant : le déterminant de la matrice $X'X$ doit être le plus grand possible,
- critère de la plus petite trace : la trace de la matrice $(X'X)^{-1}$ doit être la plus petite possible.

Ces critères sont assez rapides à calculer et permettent donc une prise de décision rapide.

Annexe A

Rappels et compléments

A.1 Théorème de Cochran.

Théorème A.1 (*admis*) Soit X_1, \dots, X_n i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. On note X le vecteur $(X_1, \dots, X_n) \in \mathbb{R}^n$. Soit $E_1 \oplus E_2 \oplus \dots \oplus E_p$ une décomposition de \mathbb{R}^n en p sous-espaces orthogonaux de dimensions respectives r_1, \dots, r_p . On note X_{E_i} la projection orthogonale de X sur E_i . Alors les vecteurs $X_{E_1}, X_{E_2}, \dots, X_{E_p}$ sont indépendants, de plus, pour tout i , la variable $\|X_{E_i}\|^2$ a pour loi $\sigma^2 \chi^2(r_i)$.

PREUVE. Par soucis de simplicité, on se contentera de la preuve dans le cas où $p = 2$. Soient E_1 et E_2 tels que $\mathbb{R}^n = E_1 \oplus E_2$. En particulier, $E_2 = E_1^\perp$ et $r_2 = n - r_1$. Soient (e_1, \dots, e_{r_1}) et (e_{r_1+1}, \dots, e_n) des bases orthonormées de respectivement E_1 et E_2 . Puisque $\mathbb{R}^n = E_1 \oplus E_2$, (e_1, \dots, e_n) est une base orthonormée de \mathbb{R}^n . Par la suite, on note U la matrice de passage de la base canonique à la base (e_1, \dots, e_n) , ce qui permet d'exprimer les projections orthogonales sur E_1 et E_2 sous la forme

$$P_{E_1} = UI_{r_1}^1 U', \quad P_{E_2} = UI_{n-r_1}^2 U',$$

où $I_{r_1}^1$ désigne la matrice diagonale avec que des 1 sur les r_1 premiers coefficients, et des 0 ensuite, le contraire pour $I_{n-r_1}^2$.

Avec toutes notations, on peut ré-écrire \mathbf{X}_{E_1} et \mathbf{X}_{E_2} sous la forme

$$\mathbf{X}_{E_1} = UI_{r_1}^1 U' \mathbf{X} \quad \text{et} \quad \mathbf{X}_{E_2} = UI_{n-r_1}^2 U' \mathbf{X}.$$

Pour commencer, on s'intéresse à la loi de $U' \mathbf{X}$. Le vecteur \mathbf{X} étant un vecteur gaussien centré réduit, $\mathbf{Y} := U' \mathbf{X}$ possède les mêmes propriétés. En effet, on a en particulier $\text{Cov}(U' \mathbf{X}) = U' U = Id = I_n$. On remarque alors que

$$I_{r_1}^1 U' \mathbf{X} = I_{r_1}^1 \mathbf{Y} = (Y_1, \dots, Y_{r_1}, 0, \dots, 0) \quad \text{et} \quad I_{n-r_1}^2 U' \mathbf{X} = I_{n-r_1}^2 \mathbf{Y} = (0, \dots, 0, Y_{r_1+1}, \dots, Y_n),$$

ce qui, au final, nous donne bien l'indépendance de \mathbf{X}_{E_1} et \mathbf{X}_{E_2} . Pour finir, on obtient

$$\|\mathbf{X}_{E_1}\|^2 = \|UI_{r_1}^1 U' \mathbf{X}\|^2 = \|I_{r_1}^1 U' \mathbf{X}\|^2 = \|(Y_1, \dots, Y_{r_1}, 0, \dots, 0)\|^2 \sim \chi_{r_1}^2,$$

dans la mesure où la norme d'un vecteur n'est pas affecté par transformations orthogonale, et $\mathbf{Y} \sim \mathcal{N}(0, I_n)$. Avec le même raisonnement, on obtient $\|\mathbf{X}_{E_2}\|^2 \sim \chi_{r_2}^2$, ce qui permet de conclure.

□

A.2 La méthode de Newton-Raphson

Soit $t : \mathbb{R} \rightarrow \mathbb{R}$ une fonction \mathcal{C}^1 donnée. La problématique consiste à trouver x^* tel que $t(x^*) = 0$. Par définition de la dérivée, on a

$$t'(x^*) = \lim_{h \rightarrow +\infty} \frac{t(x^* + h) - t(x^*)}{h}.$$

La méthode de Newton est basée sur l'heuristique suivante. Si x est suffisamment 'proche' de x^* , alors moralement

$$t'(x) \simeq \frac{t(x) - t(x^*)}{x - x^*} \Leftrightarrow x - x^* \simeq \frac{t(x) - t(x^*)}{t'(x)},$$

par définition de x^* . On va utiliser cette méthode de manière itérative en initialisant un x_0 puis en posant, pour tout $n \in \mathbb{N}$,

$$x_n = x_{n-1} - \frac{t(x_{n-1}) - t(x^*)}{t'(x_{n-1})}.$$

Sous des hypothèses assez souples (fonction t deux fois différentiable au voisinage de x^* par exemple), on peut démontrer que $x_n \rightarrow x^*$ quand $n \rightarrow +\infty$.

A.3 Théorème central limite : condition de Lindeberg

Le théorème suivant généralise le Théorème central limite à des suites de variables indépendantes mais non identiquement distribuées. Ce type de résultat est particulièrement intéressant pour le modèle linéaire généralisé.

Théorème A.2 Soient X_1, \dots, X_n des variables aléatoires indépendantes d'espérances et de variances respectives m_i et σ_i^2 . Soit $S_n^2 = \sum_{i=1}^n \sigma_i^2$ et pour tout $i \in \{1, \dots, n\}$, F_i la fonction de répartition des variables $X_i - m_i$. Si

$$\lim_{n \rightarrow +\infty} \left[\frac{1}{S_n^2} \sum_{i=1}^n \int_{|x| > S_n} x^2 dF_i(x) \right] = 0, \quad (\text{A.1})$$

alors,

$$\frac{\sum_{i=1}^n (X_i - m_i)}{\sqrt{S_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } n \rightarrow +\infty.$$

Annexe B

Exams given in previous sessions

Partial exam
30 minutes

Documents, calculators and cellular phones are not allowed.

Exercise

We consider the classical linear regression model

$$Y_i = aZ_i + b + \epsilon_i, \quad i = \{1, \dots, n\}.$$

For all $i \in \{1, \dots, n\}$, we assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, the variance σ^2 being supposed to be **known**.

1. Write the model in a matrix form.
2. Let \hat{a} and \hat{b} the least square estimators. Provide (the answer should be justified) the law of the random variable

$$\sum_{i=1}^n (\hat{a}Z_i + \hat{b} - Y_i)^2.$$

3. We are interested in the assumption $H_0 : a = 0$. Let \tilde{b} be the maximum likelihood estimator of b under H_0 . Provide an explicit expression for \tilde{b} and its law under H_0 .
4. Now, we want to test

$$H_0 : a = 0 \quad \text{against} \quad H_1 : a \neq 0.$$

To this end, we consider the likelihood ration defined as

$$\Lambda = 2 \ln \frac{\max_{\theta \in \mathbb{R}^2} L(\mathbf{Y}, \theta)}{\max_{\theta \in \Theta_0} L(\mathbf{Y}, \theta)},$$

where

$$\Theta_0 = \{\theta = (a, b)' \in \mathbb{R}^2, \quad a = 0\}.$$

Prove that

$$\Lambda_n = \frac{1}{\sigma^2} \left[\sum_{i=1}^n (\tilde{b} - Y_i)^2 - \sum_{i=1}^n (\hat{a}Z_i + \hat{b} - Y_i)^2 \right].$$

5. Prove that the random variable Λ_n follows a χ^2 distribution (the degree of freedom should be made precise).

Partial exam
30 minutes

Documents, calculators and cellular phones are not allowed.

Exercise

We deal with a sample Y_1, \dots, Y_N where $Y_i \sim \mathcal{B}(n_i, p_i)$ for all $i \in \{1, \dots, N\}$, the variables Y_i being supposed independent. Our aim is to explain these variables (i.e. the parameter p_i) thanks to the deterministic explanatory variables Z_1, \dots, Z_N according to the relationship

$$p_i = \frac{e^{aZ_i+b}}{1 + e^{aZ_i+b}} \quad \forall i \in \{1, \dots, n\}.$$

In particular, we want an estimator of the parameter $\theta = (a \ b)'$.

1. Write the log-likelihood $l(\mathbf{Y}, \theta)$ associated to this model.
2. Check that the score satisfies

$$\begin{aligned} S(\mathbf{Y}, \theta) &= \left(\frac{\partial}{\partial a} l(\mathbf{Y}, \theta) \quad \frac{\partial}{\partial b} l(\mathbf{Y}, \theta) \right)', \\ &= \left(\sum_{i=1}^N (Y_i - n_i p_i) Z_i \quad \sum_{i=1}^N (Y_i - n_i p_i) \right)'. \end{aligned}$$

3. Compute the Fisher information matrix.
4. Check that this matrix corresponds to the covariance matrix of the score.

Final exam
2 hours

Calculators and cellular phones are not allowed.

Exercise 1.

We are interested here in the life expectation of 17 patients suffering from leukaemia. The following table links the life expectation (Y_i variable, in weeks since the detection of the disease) and the X_i variable corresponding to the \log_{10} of the initial number of white blood cells.

y_i	65	156	100	134	16	108	121	4	39
x_i	3.36	2.88	3.63	3.41	3.78	4.02	4	4.23	3.73

y_i	143	56	26	22	1	1	5	65
x_i	3.36	2.88	3.63	3.41	3.78	4.02	4	4.23

1. The relationship between X_i and Y_i is modeled through the expression $\mathbb{E}[Y_i] = e^{-\beta_1 - \beta_2 X_i}$. The parameter of interest is $\theta = (\beta_1 \ \beta_2)'$. What is the associated link function ?
2. We assume that the Y_i are exponentially distributed. Provide an expression of the score and the Fisher information matrix.
3. Propose a strategy allowing to test the nullity of β_2 .

Indication : A random variable Y_i is exponentially distributed with an associated paramter λ_i if it admits the density

$$f_{Y_i}(y) = \begin{cases} \lambda_i e^{-\lambda_i y} & \text{si } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we have $\mathbb{E}[Y_i] = 1/\lambda_i$ and $\text{Var}(Y_i) = 1/\lambda_i^2$.

Exercise 2.

We are interested in the links between the presence of a fish specie (here the roach - *gardon* in French) in a given river and some environmental factors. The dataset **Gardons.txt**¹ gathers 5 columns associated to the following informations

- the place where the measure have been conducted (variable S),
- a local hydraulic index characterizing the flow speed, the slope and the width of the river (variable V, normalized),
- the position of the station in the gradient upstream/downstream (variable G, normalised),
- the elevation (variable A, transformed and normalized),
- the presence (or not) of roach (variable P binary, 0 for presence, 1 for absence).

Our aim is to explain the variable P .

Provide a complete analysis of these data according to the following R outputs. In particular, you should

- detail the construction of the considered model(s),
- briefly explain the considered tests (test statistic, decision rule, ...),
- Provide an interpretation of the results.

1. Data coming from the Programme National "Indice Poisson". GIP Hydrosystèmes, CSP, Agences de Bassin. Mise au point d'un indice Poisson applicable sur le territoire national : Convention n° 1302 Conseil Supérieur de la pêche / Agence de l'eau Adour-Garonne. Août 1996 - Décembre 2000. Responsable scientifique : T. Oberdorff.

```
#-----#
#  Partie I      #
#-----#
```

```
> Gardon <- read.table('Gardons.txt',skip=2, col.names=c("S","V","G","A","P"))
> Gardon <- Gardon[order(Gardon[, "V"],decreasing=FALSE),]
> reg <- lm(P~V,data=Gardon)
> summary(reg)
```

Call:

```
lm(formula = P ~ V, data = Gardon)
```

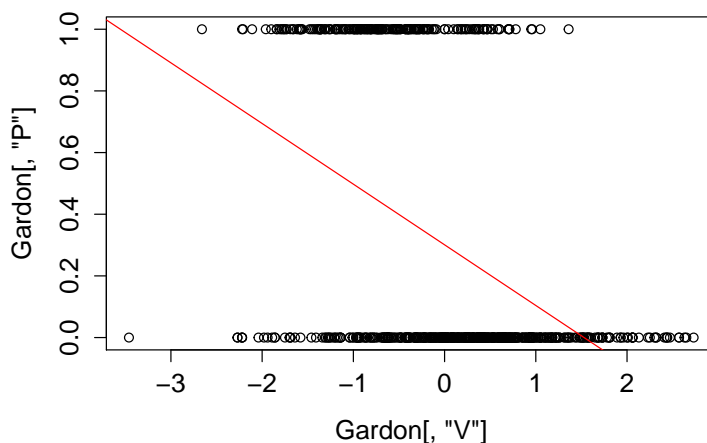
Residuals:

	Min	1Q	Median	3Q	Max
	-0.9816	-0.3008	-0.1572	0.4277	0.9667

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.30082	0.01634	18.41	<2e-16 ***
V	-0.19674	0.01635	-12.03	<2e-16 ***

```
> plot(Gardon[, "V"],Gardon[, "P"])
> abline(0.30082,-0.19674,col="red")
```



```
#-----#
#  Partie II     #
#-----#
```

```
> glm1 <- glm(P~V,family=binomial,data=Gardon)
> summary(glm1)
```

Call:

```
glm(formula = P ~ V, family = binomial, data = Gardon)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.4738	-0.7668	-0.5193	0.9522	2.3434

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.0735	0.1038	-10.340	<2e-16 ***
V	-1.1808	0.1194	-9.885	<2e-16 ***

AIC: 659.87

Number of Fisher Scoring iterations: 4

```
> plot(Gardon[, "V"], Gardon[, "P"])
> lines(Gardon[, "V"], glm1$fitted.values, lwd="2")
> Gardon[, "V2"] <- Gardon[, "V"] * Gardon[, "V"]
> glm2 <- glm(P ~ V + V2, family=binomial, data=Gardon)
> summary(glm2)
```

Call:

glm(formula = P ~ V + V2, family = binomial, data = Gardon)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4022	-0.8283	-0.4372	0.9969	2.8004

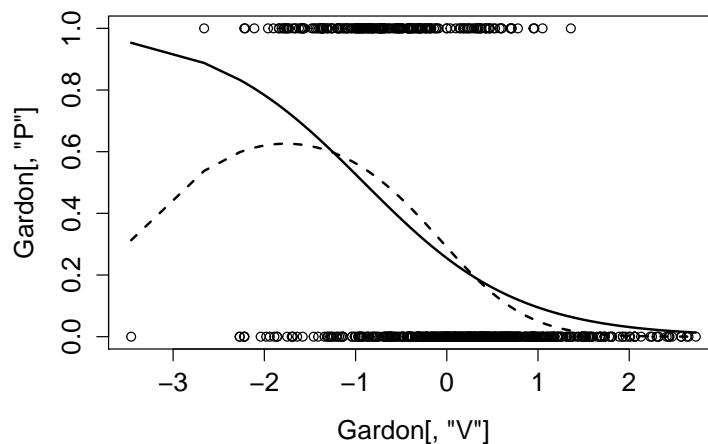
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8934	0.1163	-7.683	1.55e-14 ***
V	-1.5963	0.1864	-8.564	< 2e-16 ***
V2	-0.4524	0.1198	-3.777	0.000158 ***

AIC: 645.8

Number of Fisher Scoring iterations: 6

```
> lines(Gardon[, "V"], glm2$fitted.values, lwd="2", lty=2)
```



```
#-----#
# Partie III #
#-----#
```

```
> glm3 <- glm(P ~ V + G + A, family=binomial, data=Gardon)
> summary(glm3)
```



```
Call:
glm(formula = P ~ V + G + A, family = binomial, data = Gardon)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4755  -0.5949  -0.2865   0.4539   2.8604
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4085     0.1348  -10.448  < 2e-16 ***
V             -0.7205     0.1484   -4.856  1.19e-06 ***
G              1.1832     0.1447    8.176  2.92e-16 ***
A             -0.7057     0.1447   -4.879  1.07e-06 ***
---
AIC: 504.2
```

```
Number of Fisher Scoring iterations: 5
```

```
> glm4 <- glm(P~V+G+A+V*G+V*A+A*G,family=binomial,data=Gardon)
> summary(glm4)
```

```
Call:
glm(formula = P ~ V + G + A + V * G + V * A + A * G, family = binomial,
    data = Gardon)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3955  -0.5968  -0.2980   0.4073   2.8879
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3915     0.1352  -10.289  < 2e-16 ***
V             -0.7056     0.1503   -4.695  2.66e-06 ***
G              1.0441     0.1691    6.173  6.69e-10 ***
A             -0.7288     0.1525   -4.778  1.77e-06 ***
V:G           -0.2338     0.1909   -1.225    0.221
V:A           -0.2392     0.1750   -1.367    0.172
G:A           -0.2047     0.1748   -1.171    0.242
---
AIC: 506.36
```

```
Number of Fisher Scoring iterations: 6
```

Final exam
2 hours

Calculators and cellular phones are not allowed. All paper documents are authorized.

Exercise 1.

The goal of this exercise is to obtain a theoretical validation of the C_p criterium performances in a simplified framework. To this end, we consider the observations :

$$y_k = \theta_k + \sigma \xi_k, \quad k = 1, \dots, n, \quad (\text{B.1})$$

where $\theta = (\theta_1, \dots, \theta_n)'$ is the parameter of interest that has to be estimated, σ the noise level (supposed to be known) and the ξ_k i.i.d. standard Gaussian random variables.

1. Write the model (B.1) in a matrix form.
2. We consider the non-decreasing family of models $\mathcal{M} = (\{1, \dots, m\})_{m=1 \dots p}$. In the following, each model is identified to its size m . Prove that

$$\mathbb{E} \|X\hat{\theta}^{(m)} - X\theta\|^2 = \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|^2 = \sum_{k=m+1}^n \theta_k^2 + \sigma^2 m = - \sum_{k=1}^m \theta_k^2 + \sigma^2 m + \|\theta\|^2.$$

3. Prove that for all $k \in \{1, \dots, n\}$, $y_k^2 - \sigma^2$ is an unbiased estimator of θ_k^2 . Deduce from this result that the $C_p(m)$ criterion can be written as

$$C_p(m) = - \sum_{k=1}^m y_k^2 + 2\sigma^2 m \quad \forall m \in \{1, \dots, n\}.$$

In particular, one can prove that

$$\mathbb{E}_\theta C_p(m) = \mathbb{E} \|\hat{\theta}^{(m)} - \theta\|^2 - \|\theta\|^2 \quad \forall m \in \{1, \dots, m\}.$$

4. In the following, we set

$$\hat{m} := \arg \min_{m \in \{1, \dots, n\}} C_p(m)$$

and we denote by $\hat{\theta}_{(\hat{m})}$ the associated estimator. Prove that

$$\mathbb{E} \|\hat{\theta}_{(\hat{m})} - \theta\|^2 = \mathbb{E}_\theta \left[\sum_{k=\hat{m}+1}^n \theta_k^2 + \sigma^2 \hat{m} \right] + \mathbb{E} \left[\sigma^2 \sum_{k=1}^{\hat{m}} (\xi_k^2 - 1) \right].$$

5. Establish that

$$\mathbb{E}_\theta \left[\sum_{k=\hat{m}+1}^n \theta_k^2 + \sigma^2 \hat{m} \right] = \|\theta\|^2 + \mathbb{E}_\theta [C_p(\hat{m})] + \mathbb{E} \left[\sigma^2 \sum_{k=1}^{\hat{m}} (\xi_k^2 - 1) \right] + 2\sigma \mathbb{E} \left[\sum_{k=1}^{\hat{m}} \theta_k \xi_k \right].$$

6. In the following, we assume that there exists a constant C_1 such that

$$\mathbb{E} \left[\sigma^2 \sum_{k=1}^{\hat{m}} (\xi_k^2 - 1) \right] \leq C_1 \sigma^2 (\mathbb{E}[\hat{m}] + \ln(1/\sigma))$$

and

$$\sigma \mathbb{E} \left[\sum_{k=1}^{\hat{m}} \theta_k \xi_k \right] \leq C_1 (\|\theta\|^2 + \sigma^2 \mathbb{E}[\hat{m}] + \ln(1/\sigma)).$$

Deduce from all the previous results that there exists a constant C_2 such that, for all $m \in \{1, \dots, n\}$,

$$\mathbb{E} \|\hat{\theta}^{(\hat{m})} - \theta\|^2 \leq C_2 [\|\theta\|^2 + \mathbb{E}_\theta[C_p(m)] + \sigma^2 \ln(1/\sigma)],$$

and hat

$$\mathbb{E} \|\hat{\theta}^{(\hat{m})} - \theta\|^2 \leq C_2 \inf_{m=1 \dots n} \mathbb{E}_\theta \|\hat{\theta}^{(m)} - \theta\|^2 + C_2 \sigma^2 \ln(1/\sigma).$$

Conclude.

Exercise 2.

We are interested here in the link between mortality rate and smoking. The data are gathered bellow

	age	smoke	pop	dead
1	40-44	no	656	18
2	45-59	no	359	22
3	50-54	no	249	19
4	55-59	no	632	55
5	60-64	no	1067	117
6	65-69	no	897	170
7	70-74	no	668	179
8	75-79	no	361	120
9	80+	no	274	120
10	40-44	cigarPipeOnly	145	2
11	45-59	cigarPipeOnly	104	4
12	50-54	cigarPipeOnly	98	3
13	55-59	cigarPipeOnly	372	38
14	60-64	cigarPipeOnly	846	113
15	65-69	cigarPipeOnly	949	173
16	70-74	cigarPipeOnly	824	212
17	75-79	cigarPipeOnly	667	243
18	80+	cigarPipeOnly	537	253
19	40-44	cigarrettePlus	4531	149
20	45-59	cigarrettePlus	3030	169
21	50-54	cigarrettePlus	2267	193
22	55-59	cigarrettePlus	4682	576
23	60-64	cigarrettePlus	6052	1001
24	65-69	cigarrettePlus	3880	901
25	70-74	cigarrettePlus	2033	613
26	75-79	cigarrettePlus	871	337
27	80+	cigarrettePlus	345	189
28	40-44	cigarretteOnly	3410	124
29	45-59	cigarretteOnly	2239	140
30	50-54	cigarretteOnly	1851	187
31	55-59	cigarretteOnly	3270	514
32	60-64	cigarretteOnly	3791	778
33	65-69	cigarretteOnly	2421	689
34	70-74	cigarretteOnly	1195	432
35	75-79	cigarretteOnly	436	214
36	80+	cigarretteOnly	113	63

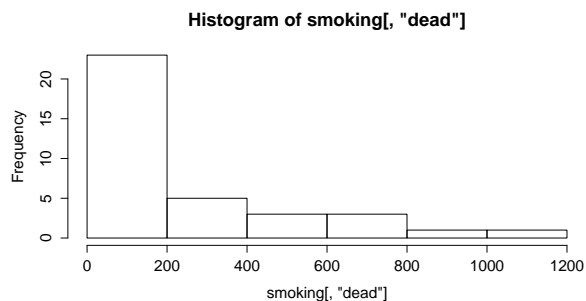
Construct a complete analysis of this dataset according to the R outputs proposed bellow. In particular, you should

- detail the construction of the considered model(s),
- briefly explain the considered testing procedures (test statistics, decision rule, and so on...)
- provide an interpretation of the results.

```
> hist(smoking[, "dead"])
> glm1 <- glm(dead ~ smoke, family=poisson, data=smoking)
> summary(glm1)
```

Call:

```
glm(formula = dead ~ smoke, family = poisson, data = smoking)
```



Deviance Residuals:

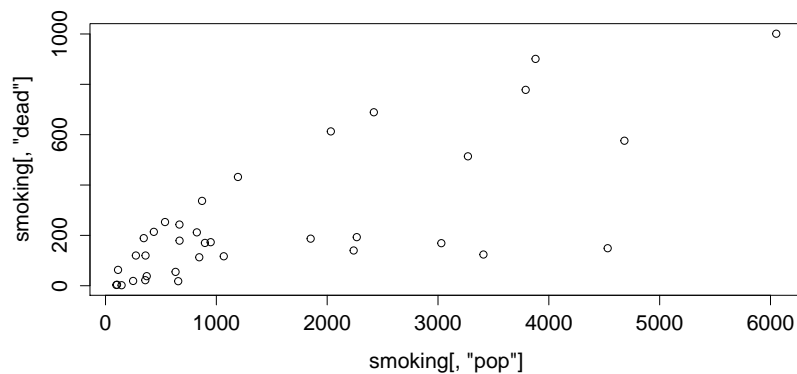
Min	1Q	Median	3Q	Max
-18.876	-13.030	-2.168	7.529	21.858

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	4.75071	0.03099	153.280	< 2e-16 ***
smokecigarretteOnly	1.10436	0.03576	30.880	< 2e-16 ***
smokecigarrettePlus	1.37761	0.03468	39.721	< 2e-16 ***
smokeno	-0.23863	0.04669	-5.111	3.21e-07 ***

Signif. codes: 0

```
> plot(smoking[, "pop"], smoking[, "dead"])
> summary(lm(dead ~ pop, data = smoking))
```



Call:

```
lm(formula = dead ~ pop, data = smoking)
```

Residuals:

Min	1Q	Median	3Q	Max
-470.79	-78.99	-8.96	98.88	361.41

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	61.53578	42.97638	1.432	0.161

```
pop          0.12321    0.01961    6.283    3.7e-07 ***
```

```
---
```

```
Multiple R-squared:  0.5373,          Adjusted R-squared:  0.5237
```

```
> smoking[,5] <- smoking[,4]/smoking[,3]
```

```
> glm3 <- glm(smoking[,5]~smoking[,2],family=binomial,weights=smoking[,3])
```

```
> summary(glm3)
```

```
Call:
```

```
glm(formula = smoking[, 5] ~ smoking[, 2], family = binomial,  
     weights = smoking[, 3])
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-25.910	-6.530	-2.579	9.772	17.405

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.21287	0.03530	-34.357	<2e-16 ***
smoking[, 2]cigaretteOnly	-0.38890	0.04036	-9.636	<2e-16 ***
smoking[, 2]cigarettePlus	-0.52902	0.03913	-13.521	<2e-16 ***
smoking[, 2]no	-0.45415	0.05192	-8.747	<2e-16 ***

```
---
```


Références

- [1] J.M. Azais, J.M. Bardet. Le modèle linéaire par l'exemple. Dunod, 2005.
- [2] L. Birgé and P. Massart. Gaussian model selection. J. Eur. Math. Soc. (3), pp 203-268, (2001).
- [3] J-J. Daudin. Le modèle linéaire et ses extensions. Ellipses, 2015.
- [4] G. Der and B/ Everitt. A handbook of statistical analyses using SAS, 3rd edition. CRC Press, 2008.
- [5] A.J. Dobson and A.G. Barnett. An introduction to generalized linear models. CRC Press, Taylor and Francis (2008).
- [6] S. Mallat. A wavelet tour of signal processing : the sparse way. Academic press, 2008.
- [7] P. McCullagh and J.A. Nelder. Generalized Linear Models, 2nd edition. Chapman et Hall (1989).
- [8] G. Saporta. Probabilités, analyse des données et statistiques 2nde édition. Edition Technip, 2006.
- [9] D.C. Weber and J.H. Skillings. A first course in the design of experiments. CRC Press, 2000.