

Le modèle linéaire pour l'apprentissage automatique

Laurent Risser
Institut de Mathématiques de Toulouse

ISAE-SUPAERO - 2020/21

Ce cours de Laurent Risser (IR CNRS, IMT) se base initialement sur une sélection des notes des cours de Philippe Besse (Pr INSA Toulouse, IMT) accessibles sur wikistat (<http://wikistat.fr/>).

Contents

1	Introduction	3
1.1	Modèle linéaire en Sciences de la Décision ?	3
1.2	Petit rappels en Probabilités/Statistique	5
1.2.1	Notions de variable aléatoire et de densité de probabilité	5
1.2.2	Théorème central limite	5
1.2.3	Estimation empirique des paramètres d'un modèle	6
2	Regression Linéaire	10
2.1	Régression Linéaire simple	10
2.1.1	Modèle	10
2.1.2	Estimation	11
2.1.3	Prédiction	12
2.1.4	Inférence	12
2.1.5	Qualité d'ajustement	13
2.1.6	Détection d'outliers	14
2.2	Régression Linéaire Multiple	16
2.2.1	Modèle	16
2.2.2	Estimation	17
2.2.3	Prévision	18
2.2.4	Qualité d'ajustement	18
3	Sélection de modèle en régression linéaire multiple	20
3.1	Introduction	20
3.1.1	Intérêt de modèles parcimonieux	20
3.1.2	Fléau de la dimension	22
3.1.3	Compromis biais-variance	23
3.2	Sélection de modèle par sélection de variables et minimisation de critères pénalisés	24
3.3	Sélection de modèle par régularisation	27
3.3.1	Régression ridge	27
3.3.2	Régression LASSO	29
3.3.3	Régression Elastic Net	31
3.3.4	Sélection par réduction de dimension	32
3.4	Validation croisée	32
3.4.1	Subdivision des observations en deux ensembles de données	32
3.4.2	K-folds	33
3.4.3	Leave-one-out	33

4 Analyse de variance	34
4.1 Introduction	34
4.2 Modèle ANOVA à un facteur	34
4.2.1 Modèle	35
4.3 Test sur la moyenne	37
4.4 Recherche de moyennes significativement différentes	38
4.5 Extension à deux facteurs	39
4.6 Analyse de covariance	42
5 Modèle linéaire mixte	45
5.1 Écriture du modèle	45
5.2 Estimation des β	47
5.3 Estimation de V	47
5.4 Tests de significativité des facteurs	48
6 Ouvertures	49
6.1 Régression logistique	49
6.2 Méthode Partial Least Squares	50
A Quelques densités de probabilités	54

Chapter 1

Introduction

1.1 Modèle linéaire en Sciences de la Décision ?

Motivation

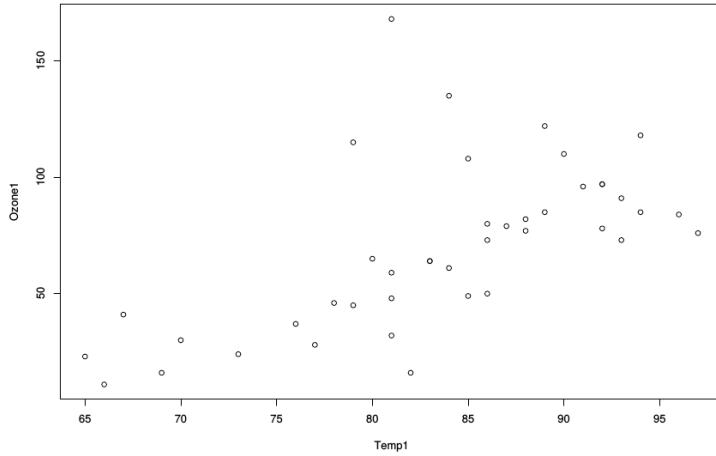
Pour étudier un phénomène à $p \geq 1$ variables d'entrée $X^{(1)}, X^{(2)}, \dots, X^{(p)}$ et une variable de sortie Y , un modèle vise à expliquer par une relation mathématique les valeurs observées de Y en fonction des variables d'entrée : $Y = g(X^{(1)}, X^{(2)}, \dots, X^{(p)})$. Le modèle essaie de refléter le plus fidèlement possible la réalité à partir de n observations. Il a pour but de mieux comprendre le phénomène étudié, et aussi de permettre de prédire Y sans devoir nécessairement réaliser des expériences.

On distingue deux types de modèles :

1. Modèle déterministe : équation qui émane souvent de lois physiques, chimiques, ou économiques, et représente le comportement attendu du phénomène.
2. Modèle statistique : Souvent, il est difficile de développer un modèle théorique car le phénomène étudié est trop complexe. On a alors recours à un modèle statistique basé non pas sur une théorie mais sur des données observées.

Exemple introductif

On étudie la pollution de l'air à New-York. On a mesuré pendant 111 jours la concentration en ozone, noté O_i (en ppm), et la température de l'air, notée T_i (en degrés Fahrenheit). Le tableau ci-dessous représente une partie des observations (celles pour lesquelles la vitesse du vent et le rayonnement solaire sont dans une certaine plage).



On constate que la concentration en ozone croît avec la température. La relation est approximativement linéaire dans la zone représentée ici. En considérant les T_i variables d'entrées (ou observations) et les O_i comme variables de sorties (ou scores), on introduit alors le modèle :

$$O_i = a + bT_i + \varepsilon_i, i = 1, \dots, n. \quad (1.1)$$

Ce modèle est appelé modèle de régression linéaire simple.

Questions posées dans ce cours

La résolution et l'étude du problème introduit ci-dessus sont discutés au début de ce cours (Chapitre 2). Beaucoup d'autres questions permettent de bien comprendre les bases de l'apprentissage statistique, qui est une composante importante de l'Intelligence Artificielle :

- Peut-on s'assurer qu'il y a une relation entre les entrées et les sorties ?
- Quel est le niveau d'incertitude sur cette relation ?
- Peut-on détecter des valeurs abérantes ?
- Que faire si la dimension des entrées (p) est plus grande que le nombre d'observations (n) ?
- Que faire si le niveau de bruit n'est pas le même pour différents groupes de variables ou si différents groupes de variables ont un *bruit* de moyenne non nulle.
- ...

Ces questions seront typiquement abordées dans le cadre de ce cours.

1.2 Petit rappels en Probabilités/Statistique

1.2.1 Notions de variable aléatoire et de densité de probabilité

Variable aléatoire Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans \mathbb{R} ou un sous-ensemble de \mathbb{R} . On distinguera en particulier le *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple $X \in \{0, 1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face.

Loi de probabilité La *loi de probabilité* d'une v.a. décrit la probabilité d'obtenir les différents résultats de cette variable.

Loi de probabilité discrète Par exemple si l'on joue à pile ou face avec une pièce parfaitement équilibrée, on a $\mathbb{P}(X = 0) = 1 - p = 0.5$ et $\mathbb{P}(X = 1) = p = 0.5$. On remarquera que la somme des probabilités de tous les résultats possibles dans le cas discret est toujours 1.

Loi de probabilité continue Dans le cas continu, écrire $\mathbb{P}(X = x)$ n'a aucun sens puisque la probabilité d'une valeur exacte est infinitésimale. On pourra par contre utiliser la *fonction de répartition* $F_X(x) = \mathbb{P}(X \leq x)$ pour représenter comment se répartissent les probabilités des différents résultats de X . Il sera alors possible de quantifier les chances que X soit sur une certaine gamme de valeurs $\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Naturellement, on aura toujours $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$. De manière purement équivalente à la fonction de répartition $p_X(x)$, la *densité de probabilité* pourra de même représenter la loi de probabilité d'une v.a. X suivant :

$$p_X(x) = \frac{\partial F_X}{\partial x}(x)$$

En utilisant les densités de probabilités, les chances que X tombe sur une gamme de valeurs $[x_1, x_2]$ sera alors

$$\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x) dx.$$

1.2.2 Théorème central limite

Afin de montrer l'importance de la loi Normale en probabilités/statistique, ainsi que de manipuler les concepts énoncés ci-dessus, il est intéressant de présenter maintenant le Théorème Central Limite (TCL).

Supposons que n variables aléatoires X_1, X_2, \dots, X_n indépendantes mais suivant une même loi de probabilité soient tirées. L'espérance (ou moyenne) m et l'écart type s de leur loi est connue. Le nombre d'observations n est

aussi supposé grand (typiquement $n > 30$). Alors, la somme des X_i peut être approchée par une loi normal de moyenne nm et d'écart type $s\sqrt{n}$, i.e. :

$$\sum_{i=1}^n X_i \sim \mathcal{N}(nm, s^2 n),$$

où la *densité de probabilité* de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est (voir aussi appendice A) :

$$f_{\theta=\{\mu, \sigma\}}(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

On peut de même montrer que la loi de $\sum_{i=1}^n X_i$ tend de même vers $\mathcal{N}(nm, s^2 n)$ lorsque n tend vers l'infini. Nous ne le montrerons pas ici, mais il est aisément de trouver la preuve de ce théorème.

Afin de nous familiariser avec les notions énoncées ci-dessus, nous proposons de vérifier empiriquement le TCL dans le cas d'une pièce tirée à pile ou face. Le protocole expérimental sera le suivant :

- Chaque étudiant de la classe tire $n = 10$ fois une pièce à pile ou face avec et compte le nombre de fois que la pièce est tombée sur pile. Pile correspond alors à $X_i = 1$ et face à $X_i = 0$.
- On suppose que $\mathbb{P}(X = 1) = 0.5$ et $\mathbb{P}(X = 0) = 0.5$, ce qui est sans doute très proche de la réalité. Ainsi l'espérance (moyenne) de X est $m = 0.5$ et son écart type est $s = 0.5$.
- On va dessiner un graphique dans lequel l'abscisse représente le nombre de 'piles' potentiellement obtenus par un étudiant (entre 0 et 10) et l'ordonnée représente le nombre d'étudiant qui ont obtenus ce nombre de 'piles' divisé par le nombre total d'étudiants.
- On constatera que cette courbe approche la densité de la loi normale de moyenne $10m$ et d'écart type $s\sqrt{10}$ (voir appendice A).

Au-delà de la connaissance du TCL lui-même et de l'illustration des notions de la section 1.2.1, cet exemple nous amène un enseignement qui est (à mes yeux) l'essence de la modélisation statistique. En assemblant plusieurs variables aléatoires, nous avons créé un modèle aléatoire dont on peut étudier les propriétés statistiques telles que la moyenne mais aussi d'une certaine manière la précision/étendue/sensibilité. Ce type de modélisation se distingue alors de la modélisation déterministe qui ne s'intéresse qu'à l'équivalent de la moyenne ici.

1.2.3 Estimation empirique des paramètres d'un modèle

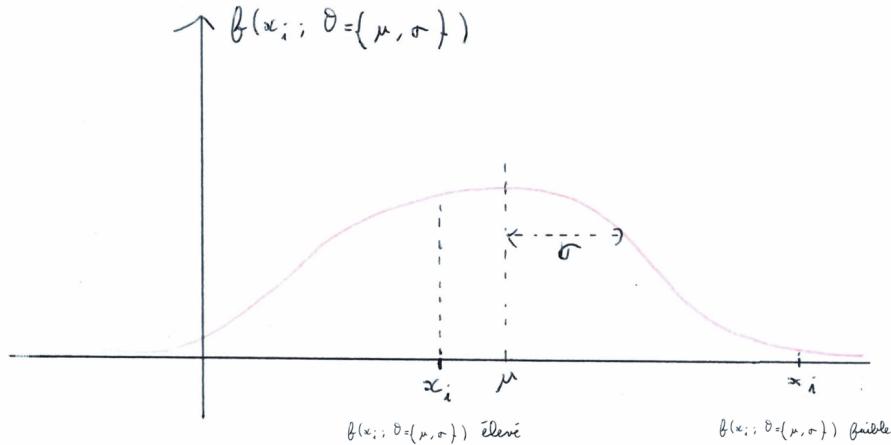
Un des composantes importantes de ce cours est de donner des méthodes pour l'estimation des paramètres de lois à partir d'observations, ou plus spécifiquement de paramètres de modèles contenant des variables aléatoires (c'est à dire avec des sources d'aléa). Cette estimation est classiquement effectuée en suivant le principe du maximum de vraisemblance ou plus simplement une estimation au sens des moindres carrés.

Maximum de vraisemblance

On dénote X une variable aléatoire (v.a.) supposée suivre une loi discrète (e.g. Bernoulli) ou continue (e.g. Normale) de paramètres θ . On note aussi $x_1, \dots, x_i, \dots, x_n$ les observations de X .

Pour une observation x_i donnée, on modélise alors la loi de X avec la fonction $f(x_i; \theta)$. Cette fonction vaut $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$ si X est une v.a. discrète et $f(x_i; \theta) = f_\theta(x_i)$ si X est continue, où $f_\theta(x_i)$ est la densité de la loi en fonction de ses paramètres θ .

Pour des paramètres θ donnés (ex : moyenne et écart type d'une loi normale), $f(x_i; \theta)$ sera alors d'autant plus élevée que x_i a des chances d'être tirée en fonction des θ .



La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

Dans l'exemple de pile ou face, supposons que l'on souhaite vérifier empiriquement si une pièce est équilibrée ou non. On modélisera $\mathbb{P}(X = 1) = f(X_i = 1; \theta = \{p\}) = p$ et $\mathbb{P}(X = 0) = f(X_i = 0; \theta = \{p\}) = 1 - p$, puis on réalisera n observations de X en tirant à pile ou face. La vraisemblance sera alors $L(\theta = \{p\}) = \prod_{i=1}^n (1_{X_i=1}p + 1_{X_i=0}(1-p))$. Supposons que sur $n = 10$ tirages, on observe 4 'piles' et 6 'faces'. En simplifiant légèrement les notations, la vraisemblance du paramètre p par rapport à notre modèle et nos observations empiriques sera alors $L(p) = p^4(1-p)^6$. Calculons alors la vraisemblance pour plusieurs valeurs de p : $L(0.2) = 0.00042$, $L(0.5) = 0.00098$, $L(0.8) = 0.00002$. De ces trois valeurs, $p = 0.5$ semble le plus vraisemblable.

De manière générale, on calculera le maximum de vraisemblance :

$$\hat{\theta} = \arg \max_{\theta} L(\theta),$$

qui renverra les paramètres les plus vraisemblables en fonction des observations et de la loi choisie.

Dans l'exemple de pile ou face, la meilleur vraisemblance sera obtenue pour $p = 0.4$ avec $L(0.4) = 0.00119$. Si la pièce est bien équilibré, le nombre de 'piles' et de 'faces' obtenus sera de plus en plus proche quand $n \rightarrow +\infty$ et $p = 0.5$ aura ainsi la meilleure vraisemblance.

Pour des raisons numériques, il est aussi bien pratique de maximiser la log-vraisemblance au lieu de la vraisemblance brute :

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \log(L(\theta)) \\ &= \arg \max_{\theta} \log \left(\prod_{i=1}^n f(x_i; \theta) \right) \\ &= \arg \max_{\theta} \sum_{i=1}^n \log f(x_i; \theta)\end{aligned}$$

Vu que la fonction \log est strictement croissante les paramètres optimum $\hat{\theta}$ seront les mêmes avec la log-vraisemblance ou la vraisemblance.

Estimation au sens des moindres carrés

On suppose disposer d'observations $\{y_i\}_{i=\{1, \dots, n\}}$ que l'on souhaite prédire/deviner à partir de observations correspondantes $\{x_i\}_{i=\{1, \dots, n\}}$, où chaque y_i correspond à x_i (voir l'exemple introductif par exemple). Dans ce cours, et très souvent en apprentissage automatique, on va alors optimiser les paramètres θ d'un modèle f_θ pour prédire au mieux les y_i avec $\hat{y}_i = f_\theta(x_i)$.

Faisons l'hypothèse que les erreurs d'approximation du modèle $e_i = y_i - f_\theta(x_i)$ suivent une loi normale centrée, i.e. $e_i \sim \mathcal{N}(0, \sigma)$. Ce choix par défaut est commun et semble raisonnable quand f_θ est bien calibré. Nous pouvons alors utiliser le principe de maximum de vraisemblance pour estimer les paramètres θ du modèle f_θ .

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{e_i^2}{2\sigma^2} \right) \\ &= \arg \max_{\theta} \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n e_i^2 \right) \\ &= \arg \min_{\theta} \sum_{i=1}^n e_i^2 \\ &= \arg \min_{\theta} \sum_{i=1}^n (y_i - f_\theta(x_i))^2\end{aligned}$$

Cette technique d'estimation est celle dite au sens des moindres carrés. Nous la retrouvons très couramment en apprentissage automatique et son interprétation est particulièrement intuitive. Elle doit notamment sa popularité au fait qu'il

est ais  de calculer son gradient par rapport aux param tres θ si on sais calculer le gradient de f_θ par rapport ´a θ :

$$\nabla_\theta e_i^2 = 2(y_i - f_\theta(x_i))\nabla_\theta f_\theta(x_i)$$

Cela ouvre la porte aux techniques d'optimisation par descente de gradient qui sont quasi syst matiques en apprentissage automatique.

Pour un public avis , il faudra se souvenir du fait que la pertinence de l'estimation de param tres d'un mod le au sens des moindres carr s repose sur une hypoth se de normalit  de l'erreur.

Chapter 2

Regression Linéaire

2.1 Régression Linéaire simple

2.1.1 Modèle

On note Y la variable aléatoire réelle à expliquer (ou encore de réponse, dépendante) et X la variable explicative (ou encore déterministe, de contrôle) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, l'estimation $\mathbb{E}(Y)$, est une fonction affine de X .

$$\mathbb{E}(Y) = f(X) = \beta_0 + \beta_1 X.$$

Pour une séquence d'observations aléatoires identiquement distribuées $\{(y_i, x_i), i = 1, \dots, n\}$, avec $n > 2$ et les x_i non tous égaux, le modèle s'écrit à partir des observations :

$$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, \dots, n$$

ou bien sous forme matricielle :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

où le vecteur \mathbf{u} contient les erreurs.

Les hypothèses relatives à ce modèle sont les suivantes :

- la distribution de l'erreur \mathbf{u} est indépendante de X ou bien X est fixe.
- l'erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, \dots, n : E(u_i) = 0, \text{Var}(u_i) = \sigma_u^2.$$

- β_0 et β_1 sont constants, il n'y a pas de rupture du modèle.
- Hypothèse complémentaire pour les inférences : $u \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Ce point important est développé dans l'appendice A.

Remarque On a fait une hypothèse de linéarité ici mais en pratique cette hypothèse n'est pas toujours valide. Quand ce n'est pas le cas, il existe aussi des méthodes de régression non-paramétriques qui ne sont pas abordées dans le cours mais peuvent être très utiles. Il est aussi possible d'effectuer des transformations élémentaires sur les données, comme par exemple $y_i = \beta_0 + \beta_1 \ln x_i$ ou bien $y_i = \beta_0 + \beta_1(x_i)^\alpha$.

2.1.2 Estimation

L'estimation des paramètres β_0 , β_1 , σ_u^2 peut être obtenue en minimisant la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données $\{(y_i, x_i), i = 1, \dots, n\}$, le critère des moindres carrés s'écrit :

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Pour minimiser ce critère, on pose :

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ r &= \frac{s_{xy}}{s_x s_y}\end{aligned}$$

On peut alors montrer que les estimateurs de β_0 et β_1 au sens des moindres carrés sont :

$$\begin{aligned}b_1 &= \frac{s_{xy}}{s_x^2}, \\ b_0 &= \bar{y} - b_1 \bar{x}.\end{aligned}$$

On montre que ce sont des estimateurs sans biais et de variance minimum parmi les estimateurs fonctions linéaires des y_i . Cela signifie que pour $i \in \{0, 1\}$ alors $Biais(b_i) = \mathbb{E}(b_i) - \beta_i = 0$ et ainsi que $Var(b_i) = \mathbb{E}(b_i - \mathbb{E}(b_i))^2 = \mathbb{E}(b_i - \beta_i)^2$ est minimum. À chaque valeur x_i de X correspond la valeur estimée (ou prédite, ajustée) de Y :

$$\hat{y}_i = b_0 + b_1 x_i$$

les résidus calculés ou estimés sont :

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

La variance σ_u^2 est enfin estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n e_i^2.$$

2.1.3 Prédiction

Une fois les paramètres β_0 et β_1 estimés par b_0 et b_1 , il est immédiat de prédire la valeur \hat{y}_0 qui a le plus de chance d'être associée à une observation x_0 avec :

$$\hat{y}_0 = b_0 + b_1 x_0.$$

Il est important de remarquer que le principe d'**estimation** des paramètres d'un modèle à partir de données d'apprentissage (les x_i et y_i) puis de **prédition** de *scores/labels/variables de sortie* (ici y_0) à partir de nouvelles observations (ici x_0) est au coeur de l'apprentissage machine.

2.1.4 Inférence

Niveau d'incertitude lié à l'estimation de b_0 et b_1

On rappel qu'une hypothèse a été faite sur les résidus $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$ dans la sous-section 2.1.1 (où e est noté u). Les estimateurs b_0 et b_1 sont alors des variables aléatoires réelles. Ils ne font qu'approcher les valeurs β_0 et β_1 que l'on connaît à coup sûr si on disposait d'une infinité d'observations (ou si l'on contrôle le modèle). Ceci est intuitivement évident, si on compare les b_0 et b_1 obtenus sur disons 4 observations pour lesquelles e est faible avec ceux obtenus sur 3 observations avec e faible et une dernière où e est grand, ce qui peut arriver puisque $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Les valeurs de b_0 et b_1 seront différentes alors que le modèle est ses paramètres sont les mêmes.

Sous l'hypothèse de Gaussianité des résidus, on montre que

$$\frac{(n-2)s^2}{\sigma_u^2} \sim \chi_{(n-2)}^2$$

où la loi du χ^2 suit une densité de probabilité donnée appendice A. Alors, les statistiques

$$(b_0 - \beta_0) \sqrt{s} \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2}$$

et

$$(b_1 - \beta_1) \sqrt{s} \left(\frac{1}{(n-1)s_x^2} \right)^{1/2}$$

suivent des lois de Student à $(n-2)$ degrés de liberté. Ceci permet de tester l'hypothèse de nullité d'un de ces paramètres à partir de tests d'hypothèses. On va par exemple tester si le b_1 obtenu est significativement différent de 0, en fonction d'un coefficient α qui représente la probabilité avec laquelle on accepte de se tromper. Typiquement α correspond à 5% de chances de se tromper, ci qui est raisonnablement faible (voir le cours de Statistique pour aller plus loin). Notons, que si b_1 est significativement différent de 0, on peut considérer qu'il existe une relation de dépendance entre les x_i et les y_i .

Intervalles de confiance

Il est de même possible de construire des intervalles de confiance pour les valeurs de b_0 et b_1 , toujours en fonction d'un niveau de confiance dépendant de α :

$$b_0 \pm s \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$

$$b_1 \pm s \left(\frac{1}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$

où $t_\nu(\alpha)$ est la distribution de Student à ν degrés de liberté (voir appendice A). En observant bien ces intervalles de confiance ainsi que les distributions de Student, il est intéressant de noter que plus on a d'observations n , plus les intervalles de confiances sont resserrés autour des b_0 et b_1 estimés. Plus on dispose d'information, moins le risque d'erreur est en effet grand par rapport aux valeurs réelles.

Attention : une inférence conjointe sur β_0 et β_1 ne peut être obtenue en considérant séparément les intervalles de confiance. La région de confiance est en effet une ellipse d'équation :

$$n(b_0 - \beta_0)^2 + 2(b_0 - \beta_0)(b_1 - \beta_1) \sum_{i=1}^n x_i + (b_1 - \beta_1)^2 \sum_{i=1}^n x_i^2 = 2s^2 \mathcal{F}_{\alpha;2,(n-2)}$$

où $\mathcal{F}_{\alpha;d_1,d_2}$ est la distribution de Fisher-Snedecor avec les paramètres d_1 et d_2 (voir appendice A).

Niveau d'incertitude lié à l'estimation d'un y_0 à partir d'un x_0

Enfin, connaissant une valeur x_0 , on définit deux intervalles de confiance de prédiction à partir de la valeur prédictive $\hat{y}_0 = b_0 + b_1 x_0$. Le premier encadre $E(Y)$ sachant $X = x_0$; le deuxième, encadre y_0 et est plus grand car il tient compte de la variance totale $\sigma_u^2 + Var(\hat{y}_0)$:

$$\hat{y}_0 \pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2},$$

$$\hat{y}_0 \pm t_{\alpha/2;(n-2)} s \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$

2.1.5 Qualité d'ajustement

On rappelle que la variance σ_u^2 est estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n e_i^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - b_0 + b_1 x_i)^2.$$

et que :

$$\begin{aligned}s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n \left(x_i - \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \right)^2 \\ s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n \left(y_i - \left(\frac{1}{n} \sum_{i=1}^n y_i \right) \right)^2 \\ s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n \left(x_i - \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \right) \left(y_i - \left(\frac{1}{n} \sum_{i=1}^n y_i \right) \right)\end{aligned}$$

Dans l'optique de mesurer la qualité d'ajustement du modèle, il est d'usage de décomposer les sommes de carrés des écarts à la moyenne sous la forme ci-dessous :

- Sum of Squares Total : $SST = (n-1)s_y^2$
- Sum of Squares Regression : $SSR = (n-1) \frac{s_{xy}^2}{s_x^2}$
- Sum of Squares Errors : $SSE = (n-1)s^2$

On appelle alors *coefficient de détermination* la quantité :

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{s^2}{s_y^2} = \frac{SSR}{SST}$$

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale. En pratique, si R^2 vaut par exemple 0.79, cela signifie que 79% de la variabilité de Y a été capturée par le modèle linéaire et que seulement 21% restent à expliquer.

2.1.6 Détection d'outliers

Le critère des moindres carrés est très sensible à des observations atypiques hors "norme" (outliers) c'est-à-dire qui présentent des valeurs trop singulières. L'étude descriptive initiale permet sans doute déjà d'en repérer mais c'est insuffisant. Un diagnostic doit être établi dans le cadre spécifique du modèle recherché afin d'identifier les observations influentes c'est-à-dire celles dont une faible variation du couple (x_i, y_i) induisent une modification importante des caractéristiques du modèle.

Ces observations repérées, il n'y a pas de remède universel : supprimer une valeur aberrante, corriger une erreur de mesure, construire une estimation robuste (en norme L_1), ne rien faire... , cela dépend du contexte et doit être négocié avec le commanditaire de l'étude.

Effet levier

Une première indication est donnée par l'éloignement de x_i par rapport à la moyenne \bar{x} . En effet, écrivons les prédicteurs y_i comme combinaisons linéaires

des observations :

$$\hat{y}_i = b_0 + b_1 x_i = \sum_{j=1}^n h_{ij} y_j$$

avec

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

en notant \mathbf{H} la matrice (hat matrix) des h_{ij} ceci s'exprime encore matriciellement :

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation de \hat{y}_i .

Résidus

Différents types de résidus sont définis afin d'affiner leurs propriétés :

- Résidus : $e_i = y_i - \hat{y}_i$
- Résidus_i : $e_{(i)i} = y_i - \widehat{y}_{(i)i} = \frac{e_i}{1-h_{ii}}$

où $\widehat{y}_{(i)i}$ est la prévision de y_i calculée sans la i ème observation (x_i, y_i) .

Afin de supprimer l'influence de la variance dans les résidus, on remarque d'abord que $Var(e_i) = \sigma_u^2(1 - h_{ii})$. En supposant que $E(e_i) = 0$, les résidus peuvent alors être standardisés de deux manières. Les *résidus standardisés* r_i sont calculés avec :

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}.$$

La standardisation ci-dessus dépend cependant de e_i dans le calcul de s (qui estime $Var(e_i)$). Une estimation non biaisée de cette variance est basée sur

$$s_{(i)}^2 = \left((n-1)s^2 - \frac{e_i^2}{1-h_{ii}} \right) / (n-3)$$

qui ne tient pas compte de la i ème observation. On définit alors les *résidus studentisés* par :

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

Sous hypothèse de normalité, on montre que ces résidus suivent une loi de Student à $(n-3)$ degrés de liberté.

Il est ainsi possible de construire un test d'hypothèse pour tester la présence d'observations atypique. Plusieurs observations peuvent de même être simultanément considérées en utilisant l'inégalité de Bonferroni. En pratique, les résidus studentisés sont souvent comparés aux bornes ± 2 . Si un résidu studentisé n'est pas dans cet intervalle de valeurs, il est considéré comme atypique.

Diagnostics

Un dernier indicateur couramment utilisé est la distance de Cook :

$$D_i = \frac{\sum_{j=1}^n (\widehat{y}_{(i)j} - \widehat{y}_j)^2}{2s^2} = \frac{h_{ii}}{2(1-h_{ii})} r_i^2, \forall i$$

qui mesure l'influence de chaque observation i sur l'ensemble des prévisions en prenant en compte effet levier et importance des résidus.

2.2 Régression Linéaire Multiple

Les modèles classiques de régression (linéaire, logistique) sont anciens et moins l'occasion de battage médiatique que ceux récents issus de l'apprentissage machine. Néanmoins, ils présentent un grand intérêt compte tenu de leur robustesse, de leur stabilité face à des fluctuations d'échantillons et de leur capacité à passer à l'échelle pour des données massives. Ils restent ainsi toujours très utilisés en production notamment lorsque la fonction à modéliser est bien linéaire et qu'il serait contre productif de chercher plus compliqué.

2.2.1 Modèle

Une variable quantitative \mathbf{Y} dite à expliquer (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $\mathbf{X}^1, \dots, \mathbf{X}^p$ dites explicatives (ou encore de contrôle, endogènes, indépendantes, régresseurs, prédicteurs).

Les données sont supposées provenir d'un échantillon statistique de n observations, chacune étant dans $\mathbb{R}^{(p+1)}$ (avec $n > p + 1$):

$$(x_i^1, \dots, x_i^j, \dots, x_i^p, y_i), i = 1, \dots, n$$

L'écriture du modèle linéaire dans cette situation conduit à supposer que l'espérance de \mathbf{Y} appartient au sous-espace de \mathbb{R}^n engendré par $\{\mathbf{1}, \mathbf{X}^1, \dots, \mathbf{X}^p\}$ où $\mathbf{1}$ désigne le vecteur de \mathbb{R}^n constitué de 1s. C'est-à-dire que les $(p + 1)$ variables aléatoires vérifient :

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \dots + \beta_p X_i^p + \varepsilon_i, i = 1, 2, \dots, n$$

avec les hypothèses suivantes :

- Les ε_i sont des termes d'erreur indépendants et identiquement distribués, *i.e.* $E(\varepsilon_i) = 0$, $Var(\varepsilon) = \sigma^2 \mathbf{I}$.
- Les termes de \mathbf{X}^j , *i.e.* du vecteur qui contient les observations de la j^{eme} variable, sont supposés déterministes (facteurs contrôlés). Dans certains contextes, on suppose alternativement que l'erreur ε est indépendante de la distribution conjointe de $\mathbf{X}^1, \dots, \mathbf{X}^p$. On écrit dans ce cas que $E(\mathbf{Y}|\mathbf{X}^1, \dots, \mathbf{X}^p) = \beta_0 + \beta_1 \mathbf{X}^1 + \beta_2 \mathbf{X}^2 + \dots + \beta_p \mathbf{X}^p$ et que $Var(Y|\mathbf{X}^1, \dots, \mathbf{X}^p) = \sigma^2$.
- Les paramètres inconnus β_0, \dots, β_p sont supposés constants.

- En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur ε (*i.e.* $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$). Les ε_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice \mathbf{X} de taille $(n \times (p + 1))$ de terme général X_i^j , dont la première colonne contient le vecteur $\mathbf{1}$ (c'est à dire $X_0^i = 1$), et dans un vecteur \mathbf{Y} de terme général Y_i . En notant les vecteurs $\varepsilon = [\varepsilon_1 \dots \varepsilon_n]'$ et $\beta = [\beta_0 \beta_1 \dots \beta_p]'$, le modèle s'écrit matriciellement :

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

Ce modèle est détaillé ci-dessous :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^p \\ 1 & x_2^1 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

2.2.2 Estimation

Conditionnellement à la connaissance des valeurs des \mathbf{X}^j , les paramètres inconnus du modèle, le vecteur β et le paramètre de nuisance σ^2 , sont estimés par minimisation des carrés des écarts (M.C.) ou encore par maximisation de la vraisemblance (M.V.) en considérant l'hypothèse de la normalité de la variable d'erreur. Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Etudions l'estimation par moindres carrés. L'expression à minimiser sur $\beta \in \mathbb{R}^{p+1}$ s'écrit :

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i^1 - \dots - \beta_p X_i^p)^2 = \|\mathbf{Y} - \mathbf{X}\beta\|^2 \quad (2.1)$$

$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \quad (2.2)$$

Par dérivation matricielle de la dernière équation on obtient les équations normales :

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0$$

dont la solution correspond à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice \mathbf{X} est de rang $(p + 1)$ et donc qu'il n'existe pas de

colinéarité entre ses colonnes. Si cette hypothèse n'est pas vérifiée, il suffit en principe de supprimer des colonnes de \mathbf{X} et donc des variables du modèle. Une approche de réduction de dimension (régression ridge, Lasso, PLS...) est en pratique à mettre en oeuvre (voir Section 3.3). Alors, l'estimation des paramètres β_j est donnée par :

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

et les valeurs ajustées (ou estimées, prédictes) de \mathbf{Y} ont pour expression :

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

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est connue sous le nom *hat matrix*. Géométriquement, c'est la matrice de projection orthogonale dans \mathbb{R}^n sur le sous-espace $Vect(\mathbf{X})$ engendré par les vecteurs colonnes de \mathbf{X} . On note alors :

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

le vecteur des résidus.

Notons finalement qu'il est possible d'inférer sur l'estimation des paramètres β_j comme dans la régression linéaire simple mais nous nous intéresserons dans ce cours à d'autres aspects de la régression multiple, notamment la sélection de modèle.

2.2.3 Prévision

Connaissant les valeurs des variables \mathbf{X}^j pour une nouvelle observation : $x_0 = [x_0^1, x_0^2, \dots, x_0^p]'$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée \hat{y}_0 de \mathbf{Y} ou $E(\mathbf{Y})$ est donnée par :

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0^1 + \cdots + \hat{\beta}_p x_0^p.$$

Les intervalles de confiance des prévisions de \mathbf{Y} et $E(\mathbf{Y})$, pour une valeur $\mathbf{x}_0 \in \mathbb{R}^p$ et en posant $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$, sont respectivement

$$\begin{aligned}\hat{y}_0 &\pm t_{\alpha/2;(n-p-1)}\hat{\sigma}(1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}, \\ \hat{y}_0 &\pm t_{\alpha/2;(n-p-1)}\hat{\sigma}(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}.\end{aligned}$$

Il est intéressant de remarquer que ces intervalles de confiance dépendent d'une loi de Student à $n-p-1$ degrés de liberté (voir appendice A). A dimension des observations fixée p , plus n est grand, plus les valeurs de la loi de Student seront faible, et ainsi les marges seront réduites. Cependant, plus p est proche de n , plus les marges sont élevées. En parallèle, les variances de ces prévisions, comme celles des estimations des paramètres, dépendent aussi directement du conditionnement de la matrice $\mathbf{X}'\mathbf{X}$ de par le terme $\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0$.

2.2.4 Qualité d'ajustement

Tout comme dans le modèle linéaire simple (Sous-section 2.1.5), la qualité d'ajustement du modèle peut être mesurée avec $p > 1$ variables par coefficient

de détermination R^2 . On note SSE la somme des carrés des résidus (sum of squared errors) :

$$SSE = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|e\|^2.$$

On définit également la somme totale des carrés (total sum of squares) par

$$SST = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2.$$

et la somme des carrés de la régression (regression sum of squares) par

$$SSR = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n\bar{\mathbf{Y}}^2 = \mathbf{Y}'\mathbf{H}\mathbf{Y} - n\bar{\mathbf{Y}}^2 = \hat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2.$$

où $\bar{\mathbf{Y}}\mathbf{1}$ est le vecteur de même taille que \mathbf{Y} dont tous les termes sont égaux à la moyenne des valeurs observées de \mathbf{Y} . Le *coefficient de détermination* est alors le rapport

$$R^2 = \frac{SSR}{SST} = \frac{\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2}$$

qui est donc la part de variation de \mathbf{Y} expliquée par le modèle de régression. La quantité R est appelée coefficient de corrélation multiple entre \mathbf{Y} et les variables explicatives, c'est le coefficient de corrélation usuel entre \mathbf{Y} et sa prévision $\hat{\mathbf{Y}}$.

Notons que le coefficient de détermination croît avec le nombre p de variables par construction. D'une manière générale, plus un modèle est complexe plus il va pouvoir coller aux données, mais moins il sera explicable et sera généralisable.

Chapter 3

Sélection de modèle en régression linéaire multiple

3.1 Introduction

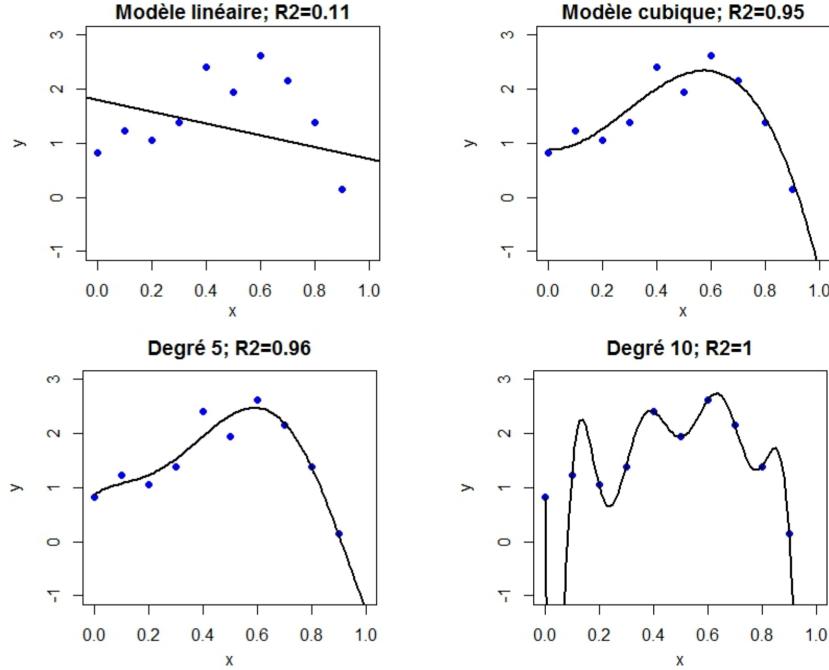
La pratique de la modélisation statistique vise trois objectifs éventuellement complémentaires.

1. Descriptif : Cherche de façon exploratoire les liaisons entre \mathbf{Y} et d'autres variables, potentiellement explicatives, \mathbf{X}^j qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. L'Analyses en Composantes Principales peut contribuer à cette recherche (voir cour de Statistique). Si p est grand, des algorithmes de recherche moins performants mais économiques en temps de calcul sont aussi à considérer.
2. Explicatif : Le deuxième objectif est sous-tendu par une connaissance a priori du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels permettent de construire le bon test conduisant à la prise de décision recherchée (voir cour de Statistique).
3. Prédicatif : Dans le troisième cas, l'accent est mis sur la qualité des prévisions. C'est la situation rencontrée en apprentissage. Ceci conduit à rechercher des modèles parcimonieux (sparse) c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Dans ce contexte, un bon modèle n'est pas celui qui explique le mieux les données au sens d'un Coefficient de détermination R^2 maximum, mais celui conduisant aux prévisions les plus fiables.

3.1.1 Intérêt de modèles parcimonieux

Ceci est illustré ceci par un exemple simple en régression polynomiale sur un jeu de données simulées. Notons qu'on sort ici du cadre linéaire pour illustrer sur un graphique 2D l'intérêt d'un modèle parcimonieux. On approche les (x_i, y_i) à l'aide de polynomes de degrés K : $y_i = \beta_0 + \sum_{k=1}^K (\beta_k x_i^k) + \varepsilon$. Les résultats après

estimation des β_k et le coefficient de détermination R^2 sont donnés ci-dessous pour $K = 1, 3, 5, 10$.



L'ajustement du modèle mesuré par le R^2 croît naturellement avec le nombre de paramètres K et atteint la valeur 1 lorsque le polynôme interpolate les observations (quand $K = n$). Dans un but de prédire de nouvelles valeurs de y , le meilleur modèle n'est cependant clairement pas celui ayant le polynôme le plus élevé. Il vaudra mieux utiliser un modèle plus contraint qui ne capturent pas le bruit inhérent aux données. Ils sont ainsi plus génériques et moins spécifique aux données observées.

Ce phénomène est connu sous le nom de sur-apprentissage en apprentissage machine. Sélectionner les variables/dimensions les plus pertinentes peut aussi être particulièrement souhaitable pour expliquer au mieux les données. Ce principe est connu sous le nom de parcimonie (sparseness) en apprentissage machine et conduit à avoir peu de $\beta_k > 0$.

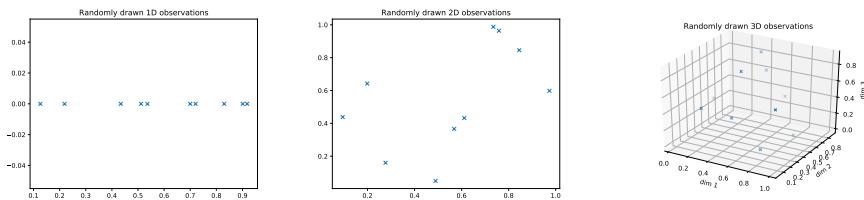
Nous nous focaliserons ici sur des critères de qualité de prévision. Le C_p de Mallows, le critère d'information d'Akaike (AIC), celui bayésien de Sawa (BIC) sont les plus classiques. Ils sont équivalents avec le R^2 lorsque le nombre de variables à sélectionner (ou complexité du modèle) est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de complexité différentes. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression pénalisée de la fonction de vraisemblance afin de favoriser des modèles parcimonieux.

3.1.2 Fléau de la dimension

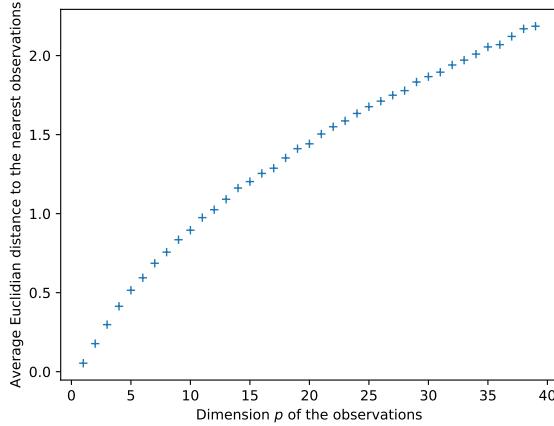
Un autre angle d'attaque pour motiver le besoin de parcimonie en apprentissage automatique est celui du *fléau de la dimension*. L'idée générale de cette notion est que lorsque la dimension p des données augmente, le volume de l'espace dans lequel les données vivent croît rapidement, si bien que les données se retrouvent isolées et deviennent éparses. Imaginons par exemple que l'on souhaite mettre en lien les profils $x_i = (x_i^1, \dots, x_i^p)$ de n étudiants $i = 1, \dots, n$ à leur note à un concours $y_i \in \mathbb{R}$. On peut représenter dans x_i la moyenne des étudiants au cours de l'année passée en Français et en Mathématiques, ce qui donne un profil général avec seulement $p = 2$ variables. On peut aussi utiliser les moyennes dans toutes les matières pour être plus fin, ce qui permet d'avoir environ $p = 10$ variables. On peut aussi utiliser des variables variées, comme le revenu moyen dans le quartier des parents, la taille, des indicateurs issus des sites internet visités, pour arriver à $p \sim 50$ variables.

Chaque fois qu'une dimension sera ajoutée, on aura plus d'information sur les élèves mais chaque élève sera aussi de plus en plus unique. Si on veut deviner la note y_{test} d'un nouvel élève, il sera alors de plus en plus difficile de trouver plusieurs élèves qui lui ressemblent dans la base d'apprentissage afin d'interpoler leurs notes avec le modèle de prédiction. Il semble alors totalement intuitif de sélectionner quelles variables sont les plus pertinentes dans le jeu d'apprentissage pour prédire les y_{test} .

Voilà une petite expérience pour illustrer l'influence de la dimension sur le niveau d'isolement d'observations. On tire suivant une loi uniforme $n = 10$ observations dans un domaine $[0, 1]^p$, avec différentes valeurs de p . Les figures ci-dessous représentent des tirages dans $[0, 1]$, dans $[0, 1]^2$ et dans $[0, 1]^3$.



On voit bien que le $[0, 1]^p$ est de moins en moins densément rempli quand p augmente. Mesurons alors empiriquement la distance euclidienne moyenne d'une nouvelle observation dans $[0, 1]^p$ au point le plus proche, parmi dix points tirés suivant cette loi uniforme. Nous obtenons les distances moyennes à un point en fonction de p ci-dessous :



La distance sera 3.8 fois plus grande si $p = 2$ que $p = 1$, elle sera plus de 10 fois plus grande si $p = 5$ que $p = 1$, plus de 40 fois plus grande si $p = 30$ que $p = 1$, ... bref il sera rapidement difficile de trouver un modèle pouvant interpoler efficacement des observations à n fixé.

Ceci est encore pire en posant le problème à l'envers. Pour avoir la même distance moyenne à un des 10 points dans $[0, 1]$, il faudra environ 140 observations dans $[0, 1]^2$, environ 2800 observations dans $[0, 1]^3$, En particulier, quand le nombre d'observations ne peut pas être trop grand et que les observations sont en grande dimension, il faudra définir un modèle efficace pour interpoler les observations (voir les réseaux de neurones convolutionnels), soit réduire leur dimension préalablement (voir l'ACP, la PLS ou encore les espaces latents de réseaux de neurones), soit sélectionner les variables les plus pertinentes. Dans le cadre du cours de modèle linéaire, nous allons voir cette troisième option qui est extrêmement classique et la plus interprétable.

3.1.3 Compromis biais-variance

Avant de rentrer dans les méthodes de sélection de modèle, discutons de la formalisation du problème de compromis biais-variance. Considérons les données d'apprentissage (x_i, y_i) , $i = 1, \dots, n$. De manière générale, on suppose que les x_i peuvent expliquer partiellement les y_i , et que d'autres paramètres indépendants des x_i entrent aussi en jeu. Généralement, on modélisera alors le problème sous cette forme :

$$y_i = f(x_i) + \epsilon_i,$$

où f est une fonction inconnue et ϵ_i suit une loi Normale de moyenne nulle et d'écart type σ . Le but de la regression est alors de trouver une fonction \hat{f} qui approxime au mieux f . Ceci se fait en fixant d'abord un modèle (linéaire, polynôme, arbre de décision, réseau de neurones, ...) puis en apprenant ses q paramètres à partir de ce que l'on connaît, c'est à dire les (x_i, y_i) . Le problème qui émerge naturellement est le suivant : Comment simultanément estimer f au mieux et tenir le moins possible compte du bruit ϵ sachant que les deux sont inconnus ? C'est la question clé du compromis biais-variance.

Plus formellement, on minimise l'espérance empirique de $(y - \hat{f}(x))^2$ sur les (x_i, y_i) , c'est à dire l'erreur au carré moyenne (Mean Squared Error – MSE). Elle peut être décomposée sous cette forme :

$$\mathbb{E}[(y - \hat{f}(x))^2] = \underbrace{\mathbb{E}[\hat{f}(x) - f(x)]^2}_{bias[\hat{f}(x)]} + \underbrace{\mathbb{E}[\hat{f}(x)^2] - \mathbb{E}[\hat{f}(x)]^2}_{variance[\hat{f}(x)]} + \sigma^2 \quad (3.1)$$

Cette représentation de la MSE peut être démontré en utilisant les relations suivantes :

- $\mathbb{E}[f(x)] = f(x)$ car $f(x)$ est déterministe
- $\mathbb{E}[y] = \mathbb{E}[f(x) + \epsilon] = \mathbb{E}[f(x)] + \mathbb{E}[\epsilon] = \mathbb{E}[f(x)] = f(x)$
- $Var[\epsilon] = \mathbb{E}[\epsilon^2] + (\mathbb{E}[\epsilon])^2 = \mathbb{E}[\epsilon^2] = \sigma^2$
- $Var[y] = \mathbb{E}[(y - \mathbb{E}[y])^2] = \mathbb{E}[(y - f(x))^2] = \mathbb{E}[(f(x) + \epsilon - f(x))^2] = \sigma^2$

Plus intéressant ici, les différents termes d'Eq. (3.1) peuvent être interprétés comme suit :

- Le terme de biais $\mathbb{E}[\hat{f}(x) - f(x)]^2$ représente à quel point le modèle \hat{f} approxime la fonction inconnue f .
- Le terme de variance $\mathbb{E}[\hat{f}(x)^2] - \mathbb{E}[\hat{f}(x)]^2 = Var[\hat{f}(x)]$ représente le niveau de variabilité de \hat{f} , sans tenir compte de f .
- Le terme σ^2 représente enfin le niveau de bruit dans les données (x_i, y_i) , qui tout comme f est inconnu.

Pour une MSE (i.e. $\mathbb{E}[(y - \hat{f}(x))^2]$) donnée, un \hat{f} représentera alors un compromis entre qualité d'approximation de f au niveau des observations $\{x_i\}_{i=1,\dots,n}$ et sa stabilité. Une trop grande qualité d'approximation au niveau des observations impliquera alors des fonctions \hat{f} instables et ainsi moins généralisables en dehors des $\{x_i\}_{i=1,\dots,n}$ (sur-apprentissage). A contrario, des fonctions \hat{f} trop stables captureront mal les relations entre les x_i et les y_i et auront de même un faible pouvoir prédictif.

Trouver un bon compromis entre biais et variance pourra se faire en réduisant explicitement la dimension d'un modèle (Section 3.2) ou en régularisant l'estimation des paramètres d'un modèle (Section 3.3). Dans tous les cas, il sera plus que recommandé d'estimer à quel point le modèle appris est généralisable à l'aide d'une technique de validation croisée (Section 3.4).

3.2 Sélection de modèle par sélection de variables et minimisation de critères pénalisés

Considérons un modèle linéaire \mathcal{M} à q variables $\mathbf{X}^{(j)}$, $j = 1, \dots, q$. Dans ce modèle $q < p$ et chaque $\mathbf{X}^{(j)}$ correspond à une des p variables observées \mathbf{X}^k , $k = 1, \dots, p$. Ce modèle s'écrit :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(q)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(q)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m^{(1)} & x_m^{(2)} & \dots & x_m^{(q)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

La sélection de modèle consiste à la fois à choisir les meilleures variables explicatives des y_i et à estimer les paramètres β_i optimaux. Nous développons dans cette section plusieurs critères de sélection de modèle.

Critère C_p de Mallows

On rappelle que la somme des carrés des résidus $SSE = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|e\|^2$. On dénote alors la *mean square error*:

$$MSE = \frac{SSE}{n - p - 1},$$

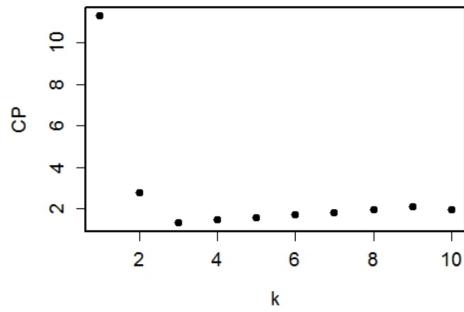
où $n - p - 1$ est le nombre de degrés de liberté du modèle compétant à p variables et n observations.

L'indicateur proposé par Mallows en 1973 pour évaluer la qualité d'un modèle donné \mathcal{M} à q variables est alors

$$C_p = (n - (q + 1)) \frac{MSE_{\mathcal{M}}}{MSE} - (n - 2(q + 1))$$

où $MSE_{\mathcal{M}}$ est la MSE calculée pour le modèle \mathcal{M} .

Il est alors d'usage de rechercher un modèle qui minimise le C_p . Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise. A qualité de modèle constante $MSE_{\mathcal{M}}/MSE$, plus q est faible, plus C_p est faible. Par contre si l'erreur du modèle \mathcal{M} augmente à q fixé, C_p augmente. Voici ci-dessous l'évolution de C_p en fonction de K dans l'exemple introductif du chapitre. Ici, le meilleur modèle contient $q = 3$ variables.



Critères AIC, BIC et PRESS

Dans le cas du modèle linéaire, et si la variance des observations est supposée connue, le critère AIC (Akaike's Information criterium) est équivalent au critère C_p de Mallows. Le critère BIC (Bayesian Information Criterion) est une extension d'AIC dans lequel le terme de pénalité est plus important. Le PRESS (somme des erreurs quadratiques) de Allen (1974) est l'introduction historique de la validation croisée ou leave-one-out (loo). Ces critères peuvent être résumés par :

- AIC : $AIC(\mathcal{M}) = n \log MSE_{\mathcal{M}} + 2(q + 1)$
- BIC : $AIC(\mathcal{M}) = n \log (MSE_{\mathcal{M}}) + \log(n)(q + 1)$
- PRESS : On désigne par $\widehat{y}_{(-i)j}$ la prévision de y_j calculée sans tenir compte de la i ème observation lors de l'estimation des paramètres alors :

$$PRESS = \sum_{i=1}^n (y_i - \widehat{y}_{(-i)i})^2$$

et permettent de comparer les capacités prédictives de différents modèles.

Algorithmes de sélection de variables

Dans le cas général les variables ne sont pas pré-ordonnées par importance. C'est d'ailleurs le cas le plus courant en pratique ! Lorsque p est grand, il n'est pas raisonnable d'explorer les 2^p modèles possibles afin de sélectionner le meilleur au sens de l'un des critères ci-dessus. Différentes stratégies existent pour explorer efficacement les modèles possibles. Elles doivent être choisies en fonction de l'objectif recherché, de la valeur de p et des moyens de calcul disponibles. Deux types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire, c'est-à-dire par nombre croissant de modèles considérés explorés parmi les 2^p et ainsi par capacité croissante d'optimalité.

Sélection (forward) A l'état initial $q = 1$ et toutes les p variables sont testées. La variable qui permet de réduire au mieux le critère du modèle obtenu est sélectionnée, on la dénote (1). On teste alors si une des $p - 1$ variables restantes améliore la qualité du modèle avec $q = 2$ et (1) déjà sélectionné... et ainsi de suite. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque le critère ne décroît plus.

Elimination (backward) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable dont l'élimination conduit la valeur du critère la plus faible est supprimée. La procédure s'arrête lorsque la valeur du critère ne décroît plus.

Mixte (stepwise) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

3.3 Sélection de modèle par régularisation

Les méthodes de régression régularisée sont à utiliser quand le problème est mal conditionné, et typiquement quand le nombre d'observations n est plus petit que la dimension des observations p . Ce cas est très courant en pratique, par exemple quand chaque observation coûte cher à obtenir mais est en très grande dimension, comme c'est le cas en génomique ou dans de nombreuses applications industrielles.

3.3.1 Régression ridge

Modèle et estimation

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables explicatives pour des raisons d'interprétation, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur biaisé des paramètres par une procédure de régularisation. Soit le modèle linéaire :

$$\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \epsilon$$

où :

$$\begin{aligned}\tilde{\mathbf{X}} &= \begin{pmatrix} 1 & X_1^1 & X_1^2 & \dots & X_1^p \\ 1 & X_2^1 & X_2^2 & \dots & X_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n^1 & X_n^2 & \dots & X_n^p \end{pmatrix}, \\ \tilde{\boldsymbol{\beta}} &= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}\end{aligned}$$

où $\mathbf{X}^0 = (1, 1, \dots, 1)'$ et \mathbf{X} désigne la matrice $\tilde{\mathbf{X}}$ privée de sa première colonne. L'estimateur ridge est défini par un critère des moindres carrés, avec une pénalité de type \mathbb{L}^2 par :

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \left(\sum_{i=1}^n (Y_i - \sum_{j=0}^p X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

où λ est un paramètre positif. Notez que le paramètre β_0 n'est pas pénalisé.

En supposant \mathbf{X} et \mathbf{Y} centrés, l'estimateur ridge est obtenu en résolvant les équations normales qui s'expriment sous la forme :

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)\boldsymbol{\beta}$$

Conduisant à :

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y}$$

La solution est donc explicite et linéaire en \mathbf{Y} . Remarquons alors que :

- $\mathbf{X}'\mathbf{X}$ est une matrice symétrique positive, *i.e.* pour tout vecteur \mathbf{u} de \mathbb{R}^p : $\mathbf{u}'(\mathbf{X}'\mathbf{X})\mathbf{u} \geq 0$. Il en résulte que pour tout $\lambda > 0$, $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p$ est inversible.

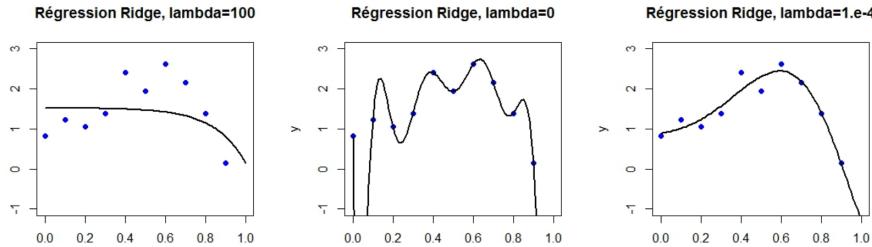
- La constante β_0 n'intervient pas dans la pénalité, sinon, le choix de l'origine pour \mathbf{Y} aurait une influence sur l'estimation de l'ensemble des paramètres. Alors : $\widehat{\beta}_0 = \bar{\mathbf{Y}}$; ajouter une constante à \mathbf{Y} ne modifie pas les $\widehat{\beta}_j$ pour $j \geq 1$.
- L'estimateur ridge n'est pas invariant par renormalisation des vecteurs $X^{(j)}$, il est préférable de normaliser (réduire les variables) des vecteurs avant de minimiser le critère.
- La régression ridge est aussi équivalent à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur β des paramètres ne soit pas trop grande :

$$\widehat{\beta}_{ridge} = \arg \min_{\beta \in \mathbb{R}^p, \|\beta\|^2 < c} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

La régression ridge conserve toutes les variables mais, contraignant la norme des paramètres β_j , elle les empêche de prendre de trop grandes valeurs et limite ainsi la variance des prévisions.

Optimisation de la pénalisation

La figure ci-dessous montre quelques résultats obtenus par la méthode ridge en fonction de la valeur de la pénalité λ sur l'exemple de la régression polynomiale (toujours pour pouvoir représenter les résultats dans un graphique 2D mais le principe est le même dans le cas linéaire multiple).



On peut remarquer que plus la pénalité λ augmente et plus la solution obtenue est régulière ou encore, plus le biais augmente (on s'éloigne des données) et la variance diminue (les estimation varient moins) :

- Il y a sur-ajustement avec une pénalité nulle : le modèle passe par tous les points mais oscille dangereusement.
- Il y a par contre sous-ajustement avec une pénalité trop grande.

Comme dans tout problème de régularisation, le choix de la valeur du paramètre λ est alors crucial et déterminera le choix de modèle. La validation croisée est généralement utilisée pour optimiser le choix (voir Section 3.4). La lecture du graphique montrant l'évolution des paramètres en fonction du coefficient ou chemins de régularisation ridge est suffisant pour définir un bon choix mais n'est pas suffisante pour déterminer une valeur optimale et est de plus laborieuse.

3.3.2 Régression LASSO

La régression ridge permet de contourner les problèmes de colinéarité même en présence d'un nombre important de variables explicatives ou prédicteurs ($p > n$). La principale faiblesse de cette méthode est cependant liée à la difficulté d'interprétation. Sans sélection, toutes les variables sont concernées dans le modèle : elles ont une valeur non-nulle et on ne peut pas se ramener au problème posé au début de Section 3.2.

Pour comprendre l'équivalence entre sélectionner explicitement des variables, comme dans Section 3.2, et sélectionner des variables en ne considérant que les $|\beta_i| > 0$, imaginons que l'on ai 4 variables $\{1, 2, 3, 4\}$ et que les deux variables sélectionnées soient $(1) = 1$ et $(2) = 3$. Alors on a :

$$\begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} \\ 1 & x_2^{(1)} & x_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & x_m^{(1)} & x_m^{(2)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{(1)} \\ \beta_{(2)} \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2^1 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m^1 & x_m^2 & x_m^3 & x_m^4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ 0 \\ \beta_3 \\ 0 \end{pmatrix}$$

avec $|\beta_1| > 0$ et $|\beta_3| > 0$

D'autres approches par pénalisation permettent une sélection, c'est le cas de la régression Lasso.

Modèle et estimation

La méthode Lasso (Tibshirani, 1996) correspond à la minimisation d'un critère des moindres carrés avec une pénalité de type L_1 (et non L_2 comme dans la régression ridge). Soit $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$.

L'estimateur Lasso de β dans le modèle $\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\beta} + \epsilon$ est alors défini par :

$$\hat{\beta}_{LASSO} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \left(\sum_{i=1}^n (Y_i - \sum_{j=0}^p X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$

où λ est un paramètre positif. On peut montrer que ceci équivaut au problème de minimisation suivant

$$\hat{\beta}_{LASSO} = \arg \min_{\beta \in \mathbb{R}^p, \|\beta\|_1 \leq t} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

pour un t convenablement choisi. Comme dans le cas de la régression ridge, le paramètre λ est un paramètre de régularisation :

- Si $\lambda = 0$, on retrouve l'estimateur des moindres carrés.
- Si λ tend vers l'infini, on annule tous les $\hat{\beta}_j$, $j = 1, \dots, p$.

La solution obtenue est dite parcimonieuse (sparse en anglais), car elle comporte des coefficients nuls.

Pourquoi la pénalisation L_1 sélectionne-t-elle les variables ?

On se place dans un cadre général dans lequel on minimise une fonction d'erreur sur l'attache aux données $f(\beta_1, \dots, \beta_p)$. Cette fonction est continue et deux fois dérivable et les β_i ont une pénalité soit L_1 ou soit L_2 :

$$\widehat{\beta}_{L_1} = \arg \min_{\beta \in \mathbb{R}^p} f(\beta_1, \dots, \beta_p) + \lambda \sum_{j=1}^p |\beta_j|$$

$$\widehat{\beta}_{L_2} = \arg \min_{\beta \in \mathbb{R}^p} f(\beta_1, \dots, \beta_p) + \lambda \sum_{j=1}^p (\beta_j)^2$$

A l'état optimal, *i.e.* pour $\beta = \widehat{\beta}_{L_1}$ ou $\beta = \widehat{\beta}_{L_2}$, les gradients des fonctions optimisées sont nulles. Il existe alors un équilibre entre les gradients de f qui tendent minimiser l'erreur sur l'attache aux données, et les gradients du terme de régularisation qui tendent à ramener β vers $\mathbf{0}$. Pour un β_j donné, on a alors :

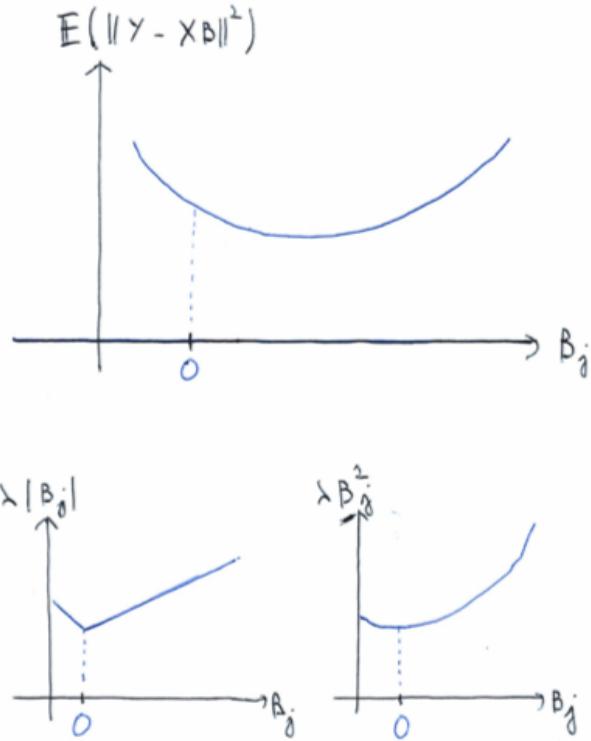
$$\text{Cas } L_1: \frac{\partial f(\beta_1, \dots, \beta_p)}{\partial \beta_j} = \lambda \text{sign}(\beta_j)$$

$$\text{Cas } L_2: \frac{\partial f(\beta_1, \dots, \beta_p)}{\partial \beta_j} = 2\lambda \beta_j$$

où $\text{sign}(\beta_j)$ vaut 1 si $\beta_j > 0$ et -1 si $\beta_j < 0$. Le cas $\beta_j = 0$ n'est pas bien défini puisque $|\beta_j|$ n'est pas dérivable en 0. En pratique sa dérivée peut être approchée par la fonction définie partout $\beta_j/(|\beta_j| + \epsilon)$ avec $\epsilon > 0$, où l'on considère que les valeurs de $\beta_j < \epsilon$ sont négligeables.

Dans le cas L_1 , β_j est nul si $|\partial f(\dots)/\partial \beta_j| < \lambda$ ce qui permet de ne sélectionner que les β_j ayant réellement une influence sur f . Dans le cas L_2 , $2\lambda \beta_j$ est à l'équilibre avec $f(\dots)/\partial \beta_j$, c'est à dire que plus β_j est faible, moins il pénalise l'attache de f aux données. Il a ainsi très peu de chances d'être nul.

Pour avoir une meilleure intuition de ces principes, la figure ci-dessous permet d'illustrer sur une dimension β_j la différence d'impact entre les pénalités L_1 et L_2 au regard d'un terme quadratique d'attache aux données :



Dans cette figure et ne s'intéressant qu'à la dimension j , le minimum de $\mathbb{E}(Y - X\beta) + \lambda(\beta_j)^2$ ne sera jamais en $\beta_j = 0$ pour un λ fini. Le minimum de $\mathbb{E}(Y - X\beta) + \lambda|\beta_j|$ sera par contre $\beta_j = 0$ si λ est suffisamment grand par rapport à la dérivée de $\mathbb{E}(Y - X\beta)$ en ce point.

Utilisation de la régression Lasso

La pénalisation est optimisée comme en régression ridge par validation croisée (voir Section 3.4).

Grâce à ses solutions parcimonieuses, cette méthode est surtout utilisée pour sélectionner des variables dans des modèles de grande dimension ; on peut l'utiliser si $p > n$ c'est-à-dire s'il y a plus de variables que d'observations. Bien entendu, dans ce cas, les colonnes de la matrice X ne sont pas linéairement indépendantes. Il n'y a donc pas de solution explicite, on utilise des procédures d'optimisation pour trouver la solution. Il faut néanmoins utiliser la méthode avec précaution lorsque les variables explicatives sont corrélées. Pour que la méthode fonctionne, il faut que le nombre de variables influentes (correspondant à des β_j différents de 0) ne dépasse pas n et que les variables non influentes ne soient pas trop corrélées avec celles qui le sont.

3.3.3 Régression Elastic Net

La méthode Elastic Net permet de combiner la régression ridge et la régression Lasso, en introduisant les deux types de pénalités simultanément.

Le critère à minimiser est:

$$\hat{\beta}_{E.N.} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \left(\sum_{i=1}^n (Y_i - \sum_{j=0}^p X_i^{(j)} \beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + (1-\alpha) \sum_{j=1}^p \beta_j^2 \right) \right)$$

- Pour $\alpha = 1$, on retrouve la méthode LASSO.
- Pour $\alpha = 0$, on retrouve la régression ridge

Il y a dans ce dernier cas deux paramètres à optimiser par validation croisée.

3.3.4 Sélection par réduction de dimension

Le principe de ces approches consiste à calculer la régression sur un ensemble de variables orthogonales deux à deux. Celles-ci peuvent être obtenues à la suite d'une analyse en composantes principales ou par décomposition en valeur singulière de la matrice \mathbf{X} : c'est la régression sur les composantes principales associées aux plus grandes valeurs propres.

L'autre approche ou régression PLS (Partial Least Squares, Section 6.2) consiste à rechercher itérativement une composante linéaire des variables de plus forte covariance avec la variable à expliquer sous une contrainte d'orthogonalité avec les composantes précédentes.

3.4 Validation croisée

Considérons la formule générique optimisée dans la section précédente :

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \left(\sum_{i=1}^n (Y_i - \sum_{j=0}^p X_i^{(j)} \beta_j)^2 + \lambda R(\beta_1, \dots, \beta_p) \right)$$

où R est la fonction de pénalisation de β qui regularise le problème d'optimisation.

Trois méthodes de validation croisée (cross-validation) pour valider le choix du paramètre λ et éventuellement de α sont largement utilisées en apprentissage machine (pas seulement en regression linéaire).

3.4.1 Subdivision des observations en deux ensembles de données

La méthode élémentaire est de subdiviser les n observations en deux sous ensembles d'observations :

- Les données d'apprentissage.
- les données de validation.

Les données seront idéalement séparées de manière aléatoire, par exemple $i = 1, \dots, n_1$ pour les données d'apprentissage et $i = n_1 + 1, \dots, n$ pour les données de validation.

Après avoir estimé $\hat{\beta}$ sur les données d'apprentissage, l'erreur d'approximation moyenne peut être estimée sur les données de validation

$$e_{split} = \frac{1}{n - n_1} \sum_{i=n_1+1}^n |Y_i - \hat{Y}_i|$$

Si le problème est trop régularisé, les tendances des données seront mal capturées par le modèle et les \hat{Y}_i auront de grandes chances de mal estimer les Y_i . L'erreur e_{split} sera alors élevé. A contrario, si le problème n'est pas assez régularisé, le modèle va trop coller aux données d'apprentissage (sur-apprentissage, overfitting) sans fort pouvoir de prédiction pour d'autres données. L'erreur e_{split} sera alors de même élevé.

Les paramètres optimaux λ et éventuellement α sont alors ceux qui minimisent e_{split} . L'optimisation des paramètres peut typiquement se faire par une *grid search*, une descente de gradient ou un algorithme stochastique (ex : recuit simulé).

3.4.2 K-folds

Afin de quantifier la stabilité de l'estimation des β_j en fonction des données il est intéressant de reproduire plusieurs fois le test de séparation de données en jeu d'apprentissage et jeu d'estimation.

La méthode la plus simple est celle dite des K-folds. Elle consiste à subdiviser les n observations (Y_i, \mathbf{X}_i) en K jeux de données de taille similaires δ_k , i.e. avec δ_k proche de n/K . Pour simplifier les notations, on suppose ici que $\delta_k = n/K$ est entier.

La méthode d'apprentissage-validation décrite dans la sous-section précédente est alors effectuée K fois, avec pour l'itération k :

- Les données d'apprentissage (Y_i, \mathbf{X}_i) , $i = 1, \dots, (k-1)\delta_k, k\delta_k + 1, \dots, n$ sont utilisées pour estimer les β_j^k .
- Les données de validation (Y_i, \mathbf{X}_i) , $i = (k-1)\delta_k + 1, \dots, k\delta_k$. sont utilisées pour calculer e_{split}^k .

$K > 1$ estimation de l'erreur e_{split}^k et des paramètres β_j^k sont alors effectués. Ceci permet d'en mesurer l'erreur de manière plus robuste qu'avec $K = 1$. De plus cela permet de quantifier la variabilité sur l'estimation des β_j : On peut simplement en calculer leur moyenne et écart type. Si une stratégie de sélection de modèle a été effectuée, on peut aussi étudier quels sont les β_j systématiquement sélectionnés et quels sont ceux qui le sont moins.

3.4.3 Leave-one-out

La méthode de validation croisée dite *Leave-one-out* est extrêmement populaire en apprentissage machine et est équivalente aux K-folds avec $K = n$. À chaque itération, l'apprentissage est effectué en enlevant une observation du jeu de données et la validation est faite sur cette observation. Cette méthode est plus lente que les K-folds en particulier quand n est grand, mais est la plus robuste et recommandée quand n est petit.

Chapter 4

Analyse de variance

4.1 Introduction

Les techniques dites d'analyse de variance sont des outils entrant dans le cadre général du modèle linéaire, où une *variable quantitative* est expliquée par une ou plusieurs *variables qualitatives*. Ici une variable qualitative va modéliser par exemple l'appartenance à un groupe, par exemple $T = 0$ signifie qu'un patient est sain, $T = 1$ signifie qu'il a une pathologie donnée, et $T = 2$ signifie qu'il a une autre pathologie.

L'objectif essentiel est alors de comparer les moyennes empiriques de la variable quantitative observées pour différentes catégories d'unités statistiques. Ces catégories sont définies par l'observation des variables qualitatives ou facteurs prenant différentes modalités ou encore de variables quantitatives découpées en classes ou niveaux.

Il s'agit donc de savoir si un facteur ou une combinaison de facteurs (interaction) a un effet sur la variable quantitative en vue, par exemple, de déterminer des conditions optimales de production ou de fabrication, une dose optimale de médicaments. Ces techniques apparaissent aussi comme des cas particuliers de la régression linéaire multiple en associant à chaque modalité une variable indicatrice (dummy variable) et en cherchant à expliquer une variable quantitative par ces variables indicatrices. L'appellation *analyse de variance* (ANOVA pour ANalysis Of Variance) vient de ce que les tests statistiques sont bâtis sur des comparaisons de sommes de carrés de variations.

Notons que l'analyse de variance avancée conduit à l'étude de plans d'expérience. Ces derniers ne seront pas abordé dans ce cours mais sont un champs important de l'analyse statistique.

4.2 Modèle ANOVA à un facteur

Cette situation est un cas particulier d'étude de relations entre deux variables statistiques : une quantitative Y admettant une densité et une qualitative T ou facteur qui engendre une partition ou classification de l'échantillon en J groupes (ou cellules, classes, ...) indiquées par j . L'objectif est de comparer les distributions de Y pour chacune des classes en particulier les valeurs des moyennes et variances.

Notons qu'avant de lancer une analyse avec les outils présentés ci-dessous, il est recommandé de réaliser un graphique constitué de boites à moustaches parallèles, une pour chaque modalité. Cette représentation donne une première appréciation de la comparaison des distributions (moyenne, variance) internes à chaque groupe.

4.2.1 Modèle

Présentation

On dispose de n observations comme dans les sections précédentes mais on ne considère que les y_i . Chaque niveau j de T avec $j = 1, \dots, J$ et $J < n$ correspond à un groupe d'observations : Pour chaque j , on observe n_j valeurs y_{1j}, \dots, y_{1n_j} de la variable Y où $n = \sum_{j=1}^J n_j$. On suppose qu'à l'intérieur de chaque groupe, les observations sont indépendantes équidistribuées de moyenne μ_j et de variance homogène $\sigma_j^2 = \sigma^2$. Ceci s'écrit :

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

où les ε_{ij} sont i.i.d. suivent une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests. Les espérances μ_j ainsi que le paramètre de nuisance σ^2 sont les paramètres inconnus à estimer.

Estimation des paramètres

On note respectivement :

$$\begin{aligned}\bar{y}_{.j} &= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \\ s_j^2 &= \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2 \\ \bar{y}_{..} &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} y_{ij}\end{aligned}$$

les moyennes et variances empiriques de chaque groupe, et la moyenne générale de l'échantillon. Alors :

- Les paramètres μ_j sont estimés sans biais par les moyennes $\bar{y}_{.j}$.
- Comme $y_{ij} = \bar{y}_{.j} + (y_{ij} - \bar{y}_{.j})$, l'estimation des erreurs e_{ij} dans chaque groupe j est naturellement :

$$e_{ij} = (y_{ij} - \bar{y}_{.j}).$$

- Les valeurs prédites dans chaque groupe j sont $\hat{y}_{ij} = \bar{y}_{.j}$.

- Sous l'hypothèse d'homogénéité des variances, la meilleure estimation sans biais de σ^2 s'écrit donc comme une moyenne pondérée des variances empiriques de chaque groupe :

$$\begin{aligned}s^2 &= \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2}{n - J} \\ &= \frac{1}{n - J} ((n_1 - 1)s_1^2 + \dots + (n_J - 1)s_J^2).\end{aligned}$$

Ecriture sous forme vectorielle

On note :

- \mathbf{y} le vecteur colonne de taille n des observations y_{ij} pour $i = 1, \dots, n_j$ et $j = 1, \dots, J$.
- \mathbf{u} le vecteur colonne de taille n des erreurs ε_{ij} pour $i = 1, \dots, n_j$ et $j = 1, \dots, J$.
- $\mathbf{1}_j$ le vecteur colonne de taille n des indicatrices du fait d'être dans la classe j . Son i ème élément vaut 1 si la i ème observation est dans la classe j , et 0 sinon.

Comme dans le cas de la régression linéaire multiple, le modèle consiste à écrire que l'espérance de la variable Y appartient au sous-espace linéaire engendré par les variables explicatives, ici les variables indicatrices :

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{1}_1 + \dots + \beta_J \mathbf{1}_J + \mathbf{u}$$

où $\mathbf{1}$ vaut 1 partout. Cette équation est détaillée ci-dessous :

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{J1} \\ \vdots \\ y_{Jn_J} \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \dots + \beta_J \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{J1} \\ \vdots \\ \varepsilon_{Jn_J} \end{pmatrix}$$

La matrice \mathbf{X} , équivalente à celle du modèle linéaire multiple, peut être construite en agrégeant horizontalement les $\mathbf{1}_j$, $j = 1, \dots, J$. La matrice $\mathbf{X}'\mathbf{X}$ n'est cependant pas inversible en général et le modèle admet une infinité de solutions. Nous disons alors que les paramètres β_j ne sont pas estimables ou identifiables. En revanche, certaines combinaisons linéaires de ces paramètres sont estimables et appelées contrastes. On estimera alors les paramètres comme dans la sous-section précédente.

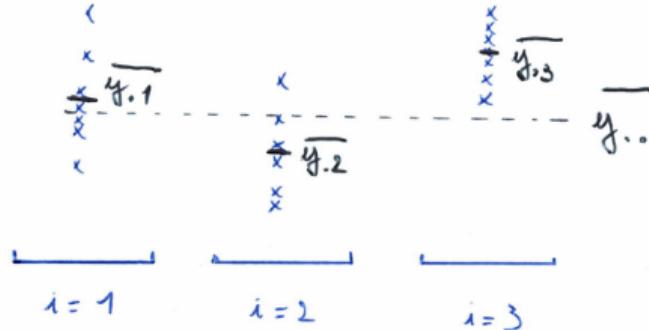
4.3 Test sur la moyenne

Test standard

On désigne les différentes sommes des carrés des variations par :

$$\begin{aligned} \text{SST} &= \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} y_{ij}^2 - n\bar{y}_{..}^2, \\ \text{SSW} &= \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^J n_j \bar{y}_{.j}^2, \\ \text{SSB} &= \sum_{j=1}^J n_j (\bar{y}_{.j} - \bar{y}_{..})^2 = \sum_{j=1}^J n_j \bar{y}_{.j}^2 - n\bar{y}_{..}^2, \end{aligned}$$

où T signifie totale, W (within) intra ou résiduelle et B (between) inter ou expliquée par la partition. La figure ci-dessous permet d'illustrer ces notations :



On considère alors l'hypothèse à tester

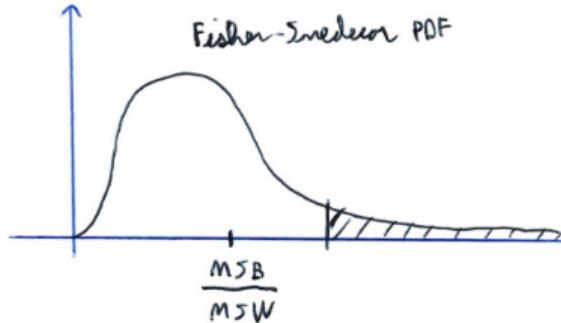
$$H_0 : \mu_1 = \dots = \mu_J,$$

qui revient à dire que la moyenne est indépendante du groupe ou encore que le facteur n'a pas d'effet, contre l'hypothèse

$$H_1 : \exists (j, k) \text{ tel que } \mu_j \neq \mu_k,$$

qui revient à reconnaître un effet ou une influence du facteur sur la variable \mathbf{Y} .

L'étude de cette hypothèse revient à comparer par un test de Fisher d'égalité des variances un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres β_j . La statistique de test est alors MSB/MSW , où $MSB = SSB/(J-1)$ et $MSW = SSW/(n-J)$. Cette statistique suit une distribution de Fischer à $(J-1)$ et $(n-J)$ degrés de libertés (voir appendice A).



Notons qu'une hypothèse a été faite ici sur l'homogénéité de la variance dans toutes les classes. Le test de Barlett et celui de Levene permettent de tester cette hypothèse.

Tests non-paramétriques

On rappelle que l'on a supposé dans cette section que les ε_{ij} suivent une loi $\mathcal{N}(0, \sigma^2)$. Lorsque l'hypothèse de normalité n'est pas satisfaite et que la taille de l'échantillon est trop petite, et ne permet ainsi pas de supposer des propriétés asymptotiques, une procédure non-paramétrique peut encore être mise en œuvre. Elle est une alternative intéressante au test de Fisher pour tester l'égalité des moyennes.

La procédure la plus utilisée est la construction du test de Kruskal-Wallis basée sur les rangs. Toutes les observations sont ordonnées selon les valeurs y_{ij} qui sont remplacées par leur rang r_{ij} , les ex-equo sont remplacés par leur rang moyen. On montre que la statistique de ce test, construite sur la somme des rangs à l'intérieur de chaque groupe, suit asymptotiquement une loi du χ^2 à $(J - 1)$ degrés de liberté.

4.4 Recherche de moyennes significativement différentes

Si l'hypothèse nulle est rejetée, il est légitime de se demander quel sont les groupes qui possèdent des moyennes significativement différentes. De nombreux tests et procédures ont été proposés dans la littérature pour répondre à cette question.

Procédure naïve

Une procédure naïve consiste à exprimer, pour chaque paire j et l de groupes, un intervalle de confiance au niveau $100(1 - \alpha)\%$ de la différence $(\mu_j - \mu_l)$, avec α typiquement égal à 0.05 :

$$(\bar{\mu}_{j,l} - \bar{\mu}_{l,j}) \pm t_{n-J}(\alpha/2) s \left(\frac{1}{n_j} + \frac{1}{n_l} \right)^{1/2} .$$

où $t_{n-J}(\cdot)$ est donné appendice A. Si 0 est inclus dans cet intervalle pour un couple de groupes (j, l) , leurs moyennes ne sont pas jugées significativement différentes au niveau α .

L'orthogonalité des facteurs rendent les tests indépendants mais elle ne peut être systématisée si J est grand. Dans ce cas, il y a en effet un total de $J(J - 1)/2$ comparaisons à tester ce qui peut être long. De manière plus critique encore, on peut s'attendre à ce que sur le simple fait du hasard $\alpha \times J(J - 1)/2$ paires de groupes soient jugés de moyennes significativement différentes même si le test global accepte l'égalité des moyennes. Par définition même du niveau d'incertitude, le test se trompe en effet environ $\alpha \times 100\%$ des fois.

Procédure de Bonferroni

D'autres procédures visent à corriger cette démarche afin de contrôler globalement le niveau des comparaisons. La plus standard est la procédure de Bonferroni qui propose des intervalles plus conservatifs (plus grands) en ajustant le niveau $\alpha' < \alpha$ définissant les valeurs critiques $t_{n-J}(\alpha'/2)$ dans la loi de Student avec:

$$\alpha' = \frac{2\alpha}{J(J - 1)}$$

Cette procédure est tout de même plus conservative que la procédure naïve et a ainsi la propriété d'augmenter le nombre de faux positifs *i.e.* de moyennes significativement différentes non détectées. D'autres méthodes comme celle de Scheffe (encore plus conservatrice) ou bien d'autres basées sur des intervalles studentisés avec des valeurs critiques spécifiques existent. La recherche est encore active dans ce domaine.

4.5 Extension à deux facteurs

Introduction

La considération de deux (ou plus) facteurs explicatifs, dans un modèle d'analyse de variance, engendre plusieurs complications. Nous aborderons ici celles qui concernent l'interaction entre variables explicatives. Cette section décrit alors le cas de deux facteurs explicatifs croisés c'est-à-dire dont les niveaux d'un facteur ne sont pas conditionnés par ceux de l'autre. On note :

- Les niveaux du 1er facteur sont notés par l'indice j variant de 1 à J .
- Les niveaux du 2eme facteur sont notés par l'indice k variant de 1 à K .
- Pour chaque combinaison (j, k) , on dispose de n_{jk} observations y_{ijk} , $i = 1, \dots, n_{jk}$.

Un plan d'expérience peut être soit *équilibré* (ou équirépété) soit *déséquilibré*. Un plan sera dit équilibré si pour chaque combinaison (j, k) , on dispose du même nombre d'observations : $n_{jk} = c, \forall(j, k)$. Ce cas introduit des simplifications importantes dans l'estimations des paramètres ainsi que dans la décomposition des variances. On se placera dans le cas équilibré dans la suite de cette sous-section.

Modèle complet

On écrit un modèle de variance à un facteur présentant $J \times K$ niveaux (j, k) :

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$

avec $j = 1, \dots, J$, $k = 1, \dots, K$ et $i = 1, \dots, n_{jk}$. On se place dans le cas équilibré avec $n_{jk} = c$, $\forall(j, k)$. En supposant que les termes d'erreur ε_{ijk} sont mutuellement indépendants et de même loi. Chacun des paramètres μ_{jk} est estimé sans biais par la moyenne

$$\bar{y}_{.jk} = \frac{1}{c} \sum_{i=1}^c y_{ijk}.$$

On définit de même les moyennes suivantes :

$$\begin{aligned}\bar{y}_{.j.} &= \frac{1}{K} \sum_{i=1}^K \bar{y}_{.jk} \\ \bar{y}_{..k} &= \frac{1}{J} \sum_{i=1}^J \bar{y}_{.jk} \\ \bar{y}_{...} &= \frac{1}{J} \sum_{i=1}^J \bar{y}_{.j.} = \frac{1}{K} \sum_{i=1}^K \bar{y}_{..k}\end{aligned}$$

qui n'ont de sens que dans le cas équilibré. La même convention du point en indice est également utilisée pour exprimer les moyennes des paramètres μ_{ijk} .

On estime alors différents termes avec :

- L'effet général $\mu_{..}$ est estimée avec $\bar{y}_{...}$.
- L'effet différentiel du niveau j du 1er facteur $\alpha_j = \mu_{j.} - \mu_{..}$ est estimé avec $\bar{y}_{.j.} - \bar{y}_{...}$.
- L'effet différentiel du niveau k du 2eme facteur $\beta_k = \mu_{..k} - \mu_{..}$ est estimé avec $\bar{y}_{..k} - \bar{y}_{...}$.
- L'effet de l'interaction des niveaux j et k $\gamma_{jk} = \mu_{jk} - \mu_{j.} - \mu_{..k} + \mu_{..}$ est estimé avec $\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}$.

Un modèle d'analyse de variance à deux facteurs s'écrit alors :

$$y_{ijk} = \mu_{..} + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

avec les contraintes :

$$\begin{aligned}\sum_{j=1}^J \alpha_j &= 0 \\ \sum_{k=1}^K \beta_k &= 0 \\ \sum_{j=1}^J \gamma_{jk} &= 0, \forall k \\ \sum_{k=1}^K \gamma_{jk} &= 0, \forall j\end{aligned}$$

qui découlent de la définition des effets et assurent l'unicité de la solution.

Modèles de régression

Comme dans le cas du modèle à un facteur, l'analyse d'un plan à deux facteurs se ramène à l'estimation et l'étude de modèles de régression sur variables indicatrices. En plus de celles des niveaux des deux facteurs $\mathbf{1}_1^1, \dots, \mathbf{1}_J^1$, et $\mathbf{1}_1^2, \dots, \mathbf{1}_K^2$, la prise en compte de l'interaction nécessite de considérer les indicatrices de chaque cellule ou traitement obtenues par produit des indicatrices des niveaux associés $\mathbf{1}_{jk}^{1 \times 2} = \mathbf{1}_j^1 \times \mathbf{1}_k^2$. On peut alors écrire un modèle de régression comme dans le cas à un facteur (sous-section 4.2.1) en considérant toutes les combinaisons de (j, k) possibles. Il conduit aussi dans le cas général à inverser une matrice $\mathbf{X}'\mathbf{X}$ non inversible. Il est alors usuel de supprimer des indicatrices.

Stratégie de test

On considère les sommes de carrés spécifiques au cas équirépété :

$$\begin{aligned}\text{SST} &= \sum_{i=1}^c \sum_{j=1}^J \sum_{k=1}^K (y_{ijk} - \bar{y}_{...})^2 &= \sum_{i=1}^c \sum_{j=1}^J \sum_{k=1}^K y_{ijk}^2 - cJK\bar{y}_{...}^2, \\ \text{SS1} &= cK \sum_{j=1}^J (\bar{y}_{.j.} - \bar{y}_{...})^2 &= cK \sum_{j=1}^J y_{.j.}^2 - cJK\bar{y}_{...}^2, \\ \text{SS2} &= cJ \sum_{k=1}^K (\bar{y}_{..k} - \bar{y}_{...})^2 &= cJ \sum_{k=1}^K y_{..k}^2 - cJK\bar{y}_{...}^2, \\ \text{SSI} &= c \sum_{j=1}^J \sum_{k=1}^K (\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...})^2 &= c \sum_{j=1}^J \sum_{k=1}^K \bar{y}_{.jk}^2 - cK \sum_{j=1}^J y_{.j.}^2 - \\ &\quad - cJ \sum_{k=1}^K y_{..k}^2 + cJK\bar{y}_{...}^2, \\ \text{SSE} &= \sum_{i=1}^c \sum_{j=1}^J \sum_{k=1}^K (y_{ijk} - \bar{y}_{.jk})^2 &= \sum_{i=1}^c \sum_{j=1}^J \sum_{k=1}^K y_{ijk}^2 - c \sum_{j=1}^J \sum_{k=1}^K \bar{y}_{.jk}^2.\end{aligned}$$

Dans le cas équirépétré, il est facile de montrer que tous les doubles produits des décompositions s'annulent (théorème de Pythagore) et que

$$SST = SS1 + SS2 + SSI + SSE.$$

On parle alors de plans orthogonaux et les trois hypothèses suivantes peuvent être considérées de façon indépendante :

- $H_{03} : \gamma_{11} = \dots = \gamma_{JK}$, i.e. pas d'effet d'interaction. Hypothèse testée par un test de Fisher avec la statistique MSI/MSE où $MSI = SSI/((J-1)(K-1))$ et $MSE = SSE/(JK(c-1))$.
- $H_{02} : \beta_1 = \dots = \beta_K$ et H_{03} i.e. pas d'effet du 2ème facteur. Hypothèse testée par un test de Fisher avec la statistique $MS2/MSE$ où $MS2 = SS2/(K-1)$.
- $H_{01} : \alpha_1 = \dots = \alpha_J$ et H_{03} i.e. pas d'effet du 1er facteur. Hypothèse testée par un test de Fisher avec la statistique $MS1/MSE$ où $MS1 = SS1/(J-1)$.

En pratique, des questions supplémentaires se posent pour des plans déséquilibrés ou incomplets. Il existe en fait toute une littérature liée au plans d'expérience. De même l'analyse de covariance peut être effectuée pour analyser l'interaction entre les variables mais déborde du cadre de ce cours.

4.6 Analyse de covariance

L'analyse de covariance se situe dans un contexte où une variable quantitative Y est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Le principe général est alors d'estimer des modèles *intra-groupes* et de tester des effets différentiels *inter-groupes* des paramètres des régressions.

Nous nous intéressons ici au cas le plus simple où seulement une variable X parmi les explicatives est quantitative. Nous sommes alors amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

Modèle

On considère une variable quantitative Y expliquée par une variable qualitative T à J niveaux et une variable quantitative (appelée encore covariable) X . Pour chaque niveau j de T , on observe n_j valeurs $x_{1j}, \dots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \dots, y_{n_j j}$ de Y . La taille de l'échantillon est alors $n = \sum_{j=1}^J n_j$.

Pour chaque $j = 1, \dots, J$ et $i = 1, \dots, n_j$, on suppose alors que :

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}$$

où mes ε_{ij} sont i.i.d. de loi centrée et de variance σ^2 . Pour la construction de tests, on suppose que $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. La résolution simultanée des J modèles de régression est alors obtenue en considérant globalement le modèle :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

où les matrices et vecteurs sont construits comme ci-dessous :

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{m_11} \\ y_{12} \\ \vdots \\ y_{m_JJ} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 & \cdots & 0 & 0 \\ 1 & x_{21} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{m_11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & x_{12} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{m_JJ} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{02} \\ \beta_{12} \\ \vdots \\ \beta_{0J} \\ \beta_{1J} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{m_11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{m_JJ} \end{pmatrix}$$

On dénote alors $\mathbf{1}_j$ et $\mathbf{x} \cdot \mathbf{1}_j$ les colonnes $2(j-1)+1$ et $2(j-1)+2$ de la matrice \mathbf{X} qui contiennent respectivement des 1 et les valeurs de x_{ij} pour le j ème groupe de T . L'estimation de ce modèle global conduit à estimer les modèles de régression par bloc, dans chacune des cellules.

Tests

Afin d'obtenir directement les bonnes hypothèses dans les tests, il est standard de reparamétriser les β_{ij} de manière à ce qu'ils expriment une différence avec un autre paramètre. Typiquement des β_{0J} et β_{1J} sont estimés sur tout \mathbf{x} et les effets différentiels consistent à estimer des $\beta'_{ij} = (\beta_{ij} - \beta_{iJ})$, $i \in \{0, 1\}$ et $j = 1, \dots, J-1$. Le modèle complet est alors :

$$\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} (\beta'_{0j}\mathbf{1}_j + \beta'_{1j}\mathbf{x} \cdot \mathbf{1}_j) + \mathbf{u}$$

Différentes hypothèses peuvent alors être testées en comparant ce modèle à un des modèles réduits suivants :

$$\begin{aligned} (i) \quad & \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} (\beta'_{0j}\mathbf{1}_j) + \mathbf{u} \\ (ii) \quad & \mathbf{y} = \beta_{0J}\mathbf{1} + \sum_{j=1}^{J-1} (\beta'_{0j}\mathbf{1}_j + \beta'_{1j}\mathbf{x} \cdot \mathbf{1}_j) + \mathbf{u} \\ (iii) \quad & \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + \sum_{j=1}^{J-1} (\beta'_{1j}\mathbf{x} \cdot \mathbf{1}_j) + \mathbf{u} \end{aligned}$$

On n'a pas de $\beta'_{1j}\mathbf{x} \cdot \mathbf{1}_j$ avec l'hypothèse (i), pas de $\beta_{1J}\mathbf{x}$ avec l'hypothèse (ii) et pas de $\beta'_{0j}\mathbf{1}_j$ avec l'hypothèse (iii). Un test de Fisher teste alors l'égalité des variances des résidus entre les modèles comparées après estimations des β_{ij} optimums. En comparant le modèle complet à (i), (ii) ou (iii) on teste alors respectivement :

- $H_0^{(i)}$: pas d'interaction, i.e. $\beta_{11} = \dots = \beta_{1J}$. Les droites partagent la même pente que β_{1J} .

- $H_0^{(ii)} : \beta_{1,J} = 0$
- $H_0^{(iii)} : \beta_{01} = \dots = \beta_{0J}$. Les droites partagent la même origine que $\beta_{0,J}$.

La démarche à suivre pour analyser un jeu de données avec ce modèle est alors : On commence donc par évaluer (i). Si le test n'est pas significatif, on regarde (ii) qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X . De même, toujours si (i) n'est pas significatif, on s'intéresse à (iii) pour juger de l'effet éventuel du facteur T .

Chapter 5

Modèle linéaire mixte

On appelle modèle mixte un modèle statistique dans lequel on considère à la fois des facteurs à effets fixes (qui interviennent sur la moyenne dans différents groupes du modèle) et des facteurs à effets aléatoires (qui interviennent sur la variance du modèle). Un modèle est dit mixte lorsqu'il y a au moins un facteur de chaque nature. Dans le cadre de ce cours, nous ne considérons que des modèles linéaires gaussiens mixtes, mais la notion de modèle mixte se rencontre également dans d'autres contextes, notamment dans le modèle linéaire généralisé.

5.1 Écriture du modèle

Modèle

Un modèle linéaire gaussien mixte à n observations s'écrit sous la forme matricielle suivante :

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\beta + \sum_{k=1}^K \mathbf{Z}_k \mathbf{A}_k + U \\ &= \mathbf{X}\beta + \mathbf{Z}\mathbf{A} + U\end{aligned}$$

où :

- \mathbf{Y} est le vecteur aléatoire réponse de \mathbb{R}^n .
- \mathbf{X} est la matrice $n \times p$ relative aux effets fixes du modèle, où p est le nombre total d'effets fixes pris en compte dans le modèle.
- β est le vecteur des p effets fixes β_j , $j = 1, \dots, p$ à estimer.
- Z_k est la matrice des indicatrices (disposées en colonnes) des niveaux du k ème facteur à effets aléatoires ($k = 1, \dots, K$). On note q_k le nombre de niveaux de ce facteur. \mathbf{Z}_k est alors de dimension $n \times q_k$.
- On note A_{kl} la v.a.r. associée au l ème niveau du k ème facteur à effets aléatoires avec $l = 1, \dots, q_k$. Pour tout l lié au facteur k , on suppose $A_{kl} \sim \mathcal{N}(0, \sigma_k^2)$.

- Pour un facteur k donnée, on note $A_k = (A_{k1}, \dots, A_{kq_k})'$ le vecteur colonne des A_{kl} . On suppose que $A_k \sim \mathcal{N}(0, \sigma_k^2 \mathbf{I}_{q_k})$.
- Enfin, U est le vecteur aléatoire des erreurs du modèle qui vérifie $U \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$.

Le modèle peut alors être écrit sous la forme :

$$\begin{array}{c}
 \text{m} \\
 \downarrow \\
 \left(\begin{array}{c} Y \\ \vdots \\ Y \end{array} \right) = \left(\begin{array}{c} X \\ \vdots \\ X \end{array} \right) \times \left(\begin{array}{c} \beta \\ \vdots \\ \beta \end{array} \right) + \left(\begin{array}{c} Z_1 \\ \vdots \\ Z_K \\ \vdots \\ Z_K \end{array} \right) + \left(\begin{array}{c} A_1 \\ \vdots \\ \vdots \\ A_K \\ \vdots \\ A_K \end{array} \right) + \left(\begin{array}{c} U \\ \vdots \\ U \end{array} \right)
 \end{array}$$

réponse Effets fixes Vecteur des effets fixes
 Vecteur des qk niveaux du k^e facteur à effets aléatoires Bruit

Moments du modèle

Il est évident de montrer que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta$ d'après les hypothèses sur le modèle. On note aussi \mathbf{V} la variance de \mathbf{Y} qui se calcule par :

$$\begin{aligned}
 \mathbf{V} &= \text{Var}(\mathbf{Y}) \\
 &= \text{Var}(\mathbf{ZA}) + \text{Var}(\mathbf{U}) \\
 &= \sum_{k=1}^K \left(\sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k' \right) + \sigma^2 \mathbf{I}_n \\
 &= \mathbf{Z}_k \mathbf{G} \mathbf{Z}_k' + \sigma^2 \mathbf{I}_n
 \end{aligned}$$

où $\mathbf{G} = \text{diag}(\sigma_1^2 \mathbf{I}_{q_1}, \dots, \sigma_K^2 \mathbf{I}_{q_K})$. On obtient alors $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\beta, \mathbf{V})$.

Les composantes de \mathbf{Y} ne sont ainsi pas indépendantes au sein de chaque niveau l d'un facteur aléatoire k donné. Ceci est évident si on observe un exemple de ce à quoi peut ressembler \mathbf{Z} :

$$\mathbf{Z} = \left(\begin{array}{cccccc}
 1 & 0 & 1 & 0 & 0 & \cdots \\
 1 & 0 & 1 & 0 & 0 & \cdots \\
 1 & 0 & 0 & 1 & 0 & \cdots \\
 0 & 1 & 0 & 1 & 0 & \cdots \\
 0 & 1 & 0 & 1 & 0 & \cdots \\
 0 & 1 & 0 & 0 & 1 & \cdots \\
 0 & 1 & 0 & 0 & 1 & \cdots
 \end{array} \right) \quad \begin{array}{c} n = 7 \\ \\ \\ \\ \\ \\ \\ \end{array}$$

Z_1
 avec $q_1 = 2$ Z_2
 avec $q_2 = 3$

Dans cet exemple, le vecteur \mathbf{ZA} aura la forme $(\xi_{11} + \xi_{21}, \xi_{11} + \xi_{21}, \xi_{11} + \xi_{22}, \xi_{12} + \xi_{22}, \xi_{12} + \xi_{22}, \xi_{12} + \xi_{23}, \xi_{12} + \xi_{23})'$ où les $\xi_{kl} \sim \mathcal{N}(0, \sigma_k^2)$ ce qui induit les dépendances intra-niveau des Y .

5.2 Estimation des β

L'expression que l'on obtient dans le cas général pour $\hat{\beta}$ fait intervenir l'estimation de la matrice des variances-covariances \mathbf{V} de \mathbf{Y} . Cette expression obtenue est fournie par la méthode des moindres carrés généralisés notée $GLSE(\beta)$ (pour Generalized Least Squares Estimator) :

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

où $\hat{\mathbf{V}} = \sum_{k=1}^K \hat{\sigma}_k^2 \mathbf{Z}_k \mathbf{Z}_k'$ et les $\hat{\sigma}_k^2$ et $\hat{\sigma}^2$ sont les composantes de variances. On a alors :

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

et il est nécessaire d'estimer les composantes de covariance, ce qui se fait typiquement par maximum de vraisemblance.

On remarquera que dans le cas équilibré où tous les q_k ont la même valeur, on a plus simplement :

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

où *OLSE* signifie *Ordinary Least Squares Estimator*.

5.3 Estimation de \mathbf{V}

Pour estimer \mathbf{V} par maximum de vraisemblance, on note d'abord $\Psi = (\hat{\sigma}_1^2, \dots, \hat{\sigma}_K^2, \hat{\sigma}^2)$ les paramètres à estimer dont dépend \mathbf{V} . La log-vraisemblance du modèle mixte gaussien s'écrit :

$$l(y, \beta, \mathbf{V}(\Psi)) = -\frac{1}{2} \log (\det (\mathbf{V}(\Psi))) - \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{V}(\Psi))^{-1} (\mathbf{Y} - \mathbf{X}\beta)$$

On en déduit le système de p équations :

$$\frac{\partial l}{\partial \beta} = \mathbf{X}' \mathbf{V}^{-1} - \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} \beta$$

dont découlent les équations normales pour $\hat{\beta}$.

On remarque ensuite que :

$$\frac{\partial \mathbf{V}}{\partial \sigma_k^2} = \mathbf{Z}_k \mathbf{Z}_k'$$

On déduit alors que pour chaque σ_k^2 :

$$\frac{\partial l}{\partial \sigma_k^2} = -\frac{1}{2} \text{tr}(\mathbf{V}(\Psi) \mathbf{Z}_k \mathbf{Z}_k') + \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{V}(\Psi))^{-1} \mathbf{Z}_k \mathbf{Z}_k' (\mathbf{V}(\Psi))^{-1} (\mathbf{Y} - \mathbf{X}\beta)$$

On obtient ainsi un système de $K + 1 + p$ équations non linéaires à $K + 1 + p$ inconnues que l'on résoud par une méthode numérique itérative. Ces procédures numériques fournissent en plus, à la convergence, la matrice des variances-covariances asymptotiques des estimateurs.

5.4 Tests de significativité des facteurs

Ces tests sont standards dans le cas équilibré (tous les q_k sont égaux), mais deviennent assez problématiques dans le cas déséquilibré. Dans le cas équilibré le test de Fisher sur les variances est en effet valable (comme dans ANOVA sous-section 4.3 ou ANCOVA sous-section 4.6). Il n'y a cependant pas de test exact, ni même de test asymptotique, qui permette de tester les effets, que ce soient les effets fixes ou les effets aléatoires, dans un modèle mixte avec un plan déséquilibré. Il existe seulement des tests approchés (dont on ne contrôle pas réellement le niveau, et encore moins la puissance)

Chapter 6

Ouvertures

6.1 Régression logistique

Modèle

On se pose maintenant dans le cas où une variable qualitative Y a 2 modalités : 1 ou bien 0. Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\beta$ ne prend pas des valeurs simplement binaires. Si l'on ne connaît que Y , on pourra estimer le paramètre Π de la loi de Bernoulli, $\mathbb{P}(Y = 1) = \Pi$ et $\mathbb{P}(Y = 0) = 1 - \Pi$, en calculant la moyenne empirique de $\mathbb{P}(Y = 1)$. On va cependant s'intéresser ici au cas où Y est lié à n observations en dimension p .

On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X . On suppose alors que :

$$\ln \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} = \beta_0 + \sum_{j=1}^p \beta_j X_j \quad (6.1)$$

où les X_j sont les p composantes de X . Il est intéressant de remarquer que $\ln(\mathbb{P}/(1 - \mathbb{P}))$ est une fonction de $\mathbb{P}(\cdot)$ strictement croissante qui :

- tend vers $-\infty$ quand $\mathbb{P}(\cdot)$ se rapproche de 0,
- vaut 0 pour $\mathbb{P}(\cdot) = 0.5$,
- tend vers $+\infty$ quand $\mathbb{P}(\cdot)$ se rapproche de 1.

On en conclue que Y à plutôt des chances de valoir 0 si $\beta_0 + \sum_{j=1}^p \beta_j X_j$ est négatif, et que Y à plutôt des chances de valoir 1 $\beta_0 + \sum_{j=1}^p \beta_j X_j$ est positif. Notons aussi que le modèle de régression est dit logistique car la loi de probabilité est modélisée à partir d'une loi logistique. Ce modèle est extrêmement populaire en apprentissage machine car il se montre performant quand on n'a que deux classes à distinguer, et il passe facilement à l'échelle.

Apprentissage des β_j

Après transformation de l'équation, on obtient :

$$\mathbb{P}(Y = 1|X) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}$$

que l'on notera pour une observation i , $i = 1, \dots, n$:

$$p(y_i = 1|x_i^1, \dots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_i^j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_i^j}}.$$

Pour résoudre estimer les β_j à l'aide d'un jeu de n observations y_i, x_i^1, \dots, x_i^p , $i = 1, \dots, n$ et de la méthode du maximum de vraisemblance, on note la contribution à la vraisemblance de l'observation i :

$$(p(y_i = 1|x_i^1, \dots, x_i^p))^{y_i} \cdot (1 - p(y_i = 1|x_i^1, \dots, x_i^p))^{1-y_i}$$

qui vaut $p(y_i = 1|x_i^1, \dots, x_i^p)$ si $y_i = 1$ et qui vaut $p(y_i = 0|x_i^1, \dots, x_i^p)$ si $y_i = 0$. La vraisemblance des observations s'écrit alors :

$$L(\beta) = \prod_{i=1}^n \left[(p(y_i = 1|x_i^1, \dots, x_i^p))^{y_i} \cdot (1 - p(y_i = 1|x_i^1, \dots, x_i^p))^{1-y_i} \right]$$

Les paramètres β_j qui maximisent cette quantité sont les estimateurs du maximum de vraisemblance de la régression logistique. Ils seront estimés typiquement en utilisant une méthode itérative. Pour des raisons numériques la log-vraisemblance, *i.e.* $n^{-1} \log(L)$, sera aussi maximisé plutôt que L .

Prédiction

Une fois les β_j appris, on se réfère à Eq. (6.1) et son interprétation pour prédire le label d'un y_0 en fonction de x_0^j observés. On calculera simplement $\beta_0 + \sum_{j=1}^p \beta_j x_0^j$. Si le signe est positif alors $\hat{y}_0 = 1$ et si le signe est négatif alors $\hat{y}_0 = 0$.

Sélection de modèle

Notons enfin qu'il est possible et même classique de sélectionner un modèle en régression logistique en pénalisant les β_j lors de la maximisation de la Log-vraisemblance, typiquement avec une méthode de type Lasso. On résoudra alors le problème suivant :

$$\hat{\beta} = \arg \max_{\beta} (n^{-1} \log(L(\beta))) - \lambda |\beta|_1$$

où comme pour la régression linéaire multiple, on sera amené à trouver un λ qui offrira un bon compromis entre pouvoir prédictif et explicabilité du modèle.

6.2 Méthode Partial Least Squares

Intuition

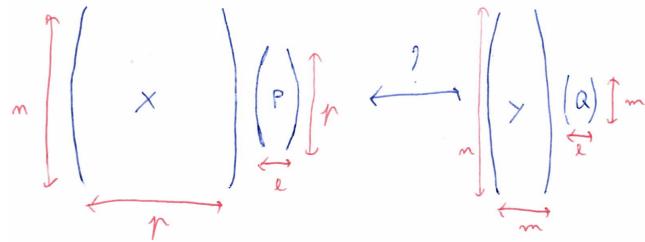
On a vu dans le cours de statistique que l'Analyse en Composantes Principales (ACP) était un outil essentiel pour explorer un ensemble d'observations $X_i = (x_i^1, \dots, x_i^p)$, $i = 1, \dots, n$ regroupées en ligne dans une matrice \mathbf{X} . L'ACP consiste en effet à maximiser la variance des projections des observations X_i , ce qui permet entre autres d'expliquer comment les variables interagissent entre

elles. Plus spécifiquement, le 1er vecteur propre v_1 est celui qui maximise la variance des projections des X_i . En supposant que les X_i sont centrés (et idéalement réduits), cela signifie que

$$v_1 = \arg \max_{v \text{ t.q. } \|v\|_2=1} \sum_{i=1}^n (X_i v)^2$$

Le 2ème vecteur propre v_2 est choisi suivant le même principe, une fois enlevée l'influence de v_1 dans \mathbf{X} ; et ainsi de suite.

L'idée de la méthode *Partial Least Squares* (PLS) est relativement similaire, mais maintenant on s'intéresse au lien entre \mathbf{X} et une matrice $n \times m$ de réponses \mathbf{Y} . Pour chaque observation X_i de \mathbf{X} , la matrice \mathbf{Y} contient une réponse Y_i en dimension m . Si $m = 1$, on a les mêmes données d'entrée que dans le cadre de la régression linéaire multiple (Section 2.2). L'approche d'analyse est cependant totalement différente : On cherche les transformations linéaires \mathbf{P} et \mathbf{Q} de \mathbf{X} et de \mathbf{Y} (si $m > 1$), respectivement, telles que : La 1ère colonne de \mathbf{P} est celle qui projette les X_i de manière à séparer au mieux les y_i projetés par la première colonne de \mathbf{Q} ; et ainsi de suite. Cette idée est schématisée ci-dessous :



Modèle

La méthode PLS (Partial Least Squares) repose toujours sur une hypothèse de modèle linéaire $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, où \mathbf{U} modélise le bruit. L'approche utilisée pour estimer le lien entre \mathbf{Y} et les variables explicatives de \mathbf{X} est cependant différente de celle du modèle linéaire classique. En particulier, le modèle sur le bruit est totalement différent et va dépendre de la covariance entre des combinaisons linéaires de \mathbf{X} et \mathbf{Y} .

Plus spécifiquement, on suppose :

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}' + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}' + \mathbf{F} \end{aligned}$$

où

- \mathbf{X} est la matrice $n \times p$ de prédicteurs. Elle est supposée centrée/réduite,
- \mathbf{Y} est la matrice $n \times m$ de réponses. Elle est supposée centrée/réduite,
- \mathbf{P} et \mathbf{Q} sont respectivement des matrices $p \times l$ et $m \times l$ de projection. Leurs colonnes sont orthonormées.
- \mathbf{T} et \mathbf{U} sont les projections de \mathbf{X} et de \mathbf{Y} respectivement par \mathbf{P} et \mathbf{Q} . Elles sont de taille $n \times l$.

- \mathbf{E} et \mathbf{F} sont des termes d'erreur de même taille que \mathbf{X} et \mathbf{Y} . Ils sont supposés *i.i.d.* et distribués suivant une loi normale.

Les projections de \mathbf{X} et de \mathbf{Y} dans \mathbf{T} et \mathbf{U} sont aussi toutes deux de même taille $n \times l$ avec $l \leq p$. La PLS consiste alors à calculer les projecteurs \mathbf{P} et \mathbf{Q} qui maximisent la covariance entre \mathbf{T} et \mathbf{U} . On dénote $\bar{\mathbf{T}}_j$ et $\bar{\mathbf{U}}_j$ la moyenne des valeurs des colonnes j de \mathbf{T} et \mathbf{U} . On maximise alors $\sum_{j=1}^l \sum_{i=1}^n (\mathbf{T}_{ij} - \bar{\mathbf{T}}_j)(\mathbf{U}_{ij} - \bar{\mathbf{U}}_j)$.

Estimation

Géométriquement, la régression PLS consiste à calculer une projection des \mathbf{X} sur un hyperplan qui est à la fois une bonne estimation de \mathbf{X} et dont les projections sont de bons prédicteurs des \mathbf{Y} . En vue de la définition d'une stratégie d'optimisation, le problème peut être vu sous la forme plus classique $\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}} + \mathbf{B}_0$. Nous donnons Alg. 1 l'algorithme PLS1 qui permet de résoudre le problème pour $m = 1$, c'est à dire \mathbf{Y} est un vecteur colonne. Dans ce cas là, $\hat{\mathbf{B}}$ est un vecteur de taille p dont l'interprétation est similaire aux vecteurs $\hat{\beta}$ du modèle linéaire multiple mais avec un modèle sous-jacent différent.

Alg. 1 Fonction $PLS1(\mathbf{X}, \mathbf{y}, l)$

```

1:  $\mathbf{X}^{(0)} \leftarrow \mathbf{X}$ 
2:  $\mathbf{w}^{(0)} \leftarrow \mathbf{X}'\mathbf{y}/|\mathbf{X}'\mathbf{y}|_2$ .
3: for  $k = 0, \dots, l-1$  do
4:    $\mathbf{t}^{(k)} \leftarrow \mathbf{X}^{(k)}\mathbf{w}^{(k)}$ 
5:    $t_k \leftarrow \mathbf{t}^{(k)} \mathbf{t}^{(k)'} \mathbf{t}^{(k)}$ 
6:    $\mathbf{t}^{(k)} \leftarrow \mathbf{t}^{(k)}/t_k$ 
7:    $\mathbf{p}^{(k)} \leftarrow \mathbf{X}^{(k)} \mathbf{t}^{(k)'} \mathbf{t}^{(k)}$ 
8:    $q_k \leftarrow \mathbf{y}' \mathbf{t}^{(k)}$ 
9:   if  $q_k = 0$  then
10:     $l \leftarrow k$  et sort de la boucle for (toute la variabilité est capturée).
11:   end if
12:   if  $k < (l-1)$  then
13:      $\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} - t_k \mathbf{t}^{(k)} \mathbf{p}^{(k)'} \mathbf{p}^{(k)}$ 
14:      $\mathbf{w}^{(k+1)} \leftarrow \mathbf{X}^{(k+1)'} \mathbf{y}/|\mathbf{X}^{(k+1)'} \mathbf{y}|_2$ 
15:   end if
16: end for
17: W est la matrice composée des colonnes  $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(l-1)}$ .
18: P est la matrice composée des colonnes  $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \dots, \mathbf{p}^{(l-1)}$ .
19: q est le vecteur composé des scalaires  $q_0, q_1, \dots, q_{l-1}$ .
20:  $\mathbf{B} \leftarrow \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \mathbf{q}$ 
21:  $\mathbf{B}_0 \leftarrow q_0 - \mathbf{P}^{(0)'} \mathbf{B}$ 
22: return  $\mathbf{B}, \mathbf{B}_0$ 

```

Sparse PLS

On a vu Chapitre 3 l'intérêt pratique des méthodes de régularisation, telles que LASSO, qui sélectionnent des modèles parcimonieux (sparse). En plus de

bien contraindre le problème de régression, ces modèles sont en effet simples à interpréter, même pour des non-spécialistes de l'analyse de données.

Ce principe peut aussi s'appliquer dans le cas de la PLS, afin de trouver à la fois des changements de bases qui mettent en lien les \mathbf{X} et \mathbf{Y} de manière optimale, et qui permettent de regrouper des blocs de variables ayant une influence similaire lorsque \mathbf{X} et \mathbf{Y} sont mis en lien. La méthode de la *sparse PLS* est alors extrêmement puissante d'un point de vue pratique.

En posant par exemple une pénalisation L_1 sur les éléments de la base \mathbf{P} avec une pondération λ_P , Alg. 1 sera légèrement modifié en rajoutant la ligne suivante entre les lignes 7 et 8 : $\mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \lambda_P sign(\mathbf{p}^{(k)})$, où $sign(\mathbf{p}^{(k)})$ est un vecteur colonne contenant des $\{-1, 0, 1\}$ en fonction de si chaque élément de $\mathbf{p}^{(k)}$ est respectivement négatif, nul ou positif. Notons que tout comme pour la régularisation LASSO en régression, un élément de $\mathbf{p}^{(k)}$ sera mis à zéro si son signe change pendant cette opération. Seuls les éléments de $\mathbf{p}^{(k)}$ ayant une réelle influence (en fonction de la *pression* de λ_P) auront alors une valeur non nulle et seront sélectionnés.

Appendix A

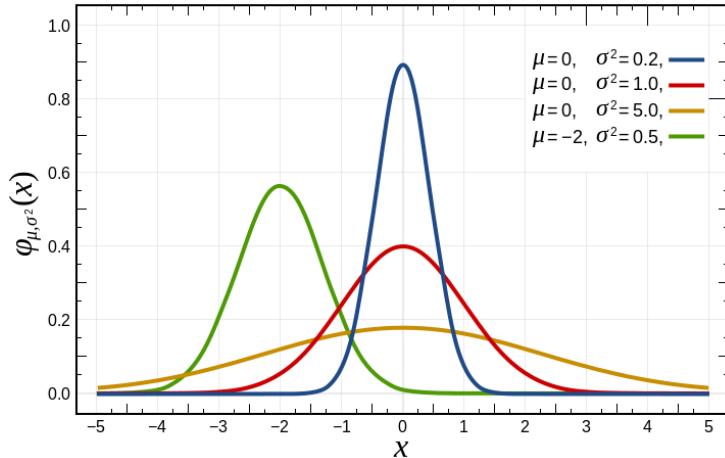
Quelques densités de probabilités

Loi normale

La *densité de probabilité* de la loi normale de moyenne μ et d'écart type σ s'écrit :

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

En voici ci-dessous une illustration pour différentes valeurs de μ et σ issue de wikipedia :



Alors :

- Si la *variable aléatoire* X suit une loi normale de moyenne μ et d'écart type σ , on écrit $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Pour connaître la probabilité qu'un réalisation de X ait une valeur entre x_1 et x_2 , on calcule $\int_{x_1}^{x_2} \varphi_{\mu,\sigma^2}(x)dx$.
- Une probabilité proche de 0 indique que cette réalisation a très peu de chances d'arriver et plus la probabilité est proche de 1 plus la réalisation a des chances d'être observée.

Loi de Student

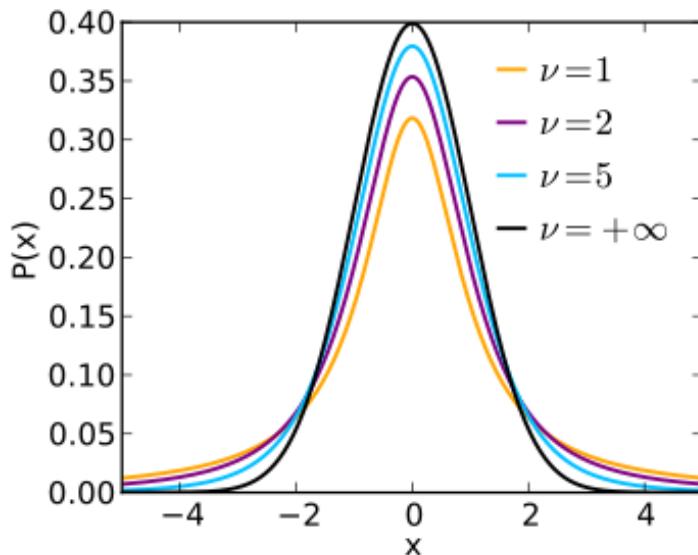
La densité de probabilité de la loi de Student t_k à ν degrés de liberté est :

$$f_T(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \nu > 0$$

où $\Gamma(x)$ est la fonction Gamma d'Euler :

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$

En voici ci-dessous une illustration pour plusieurs valeurs de ν issue de wikipedia :



Loi de Fisher-Snedecor

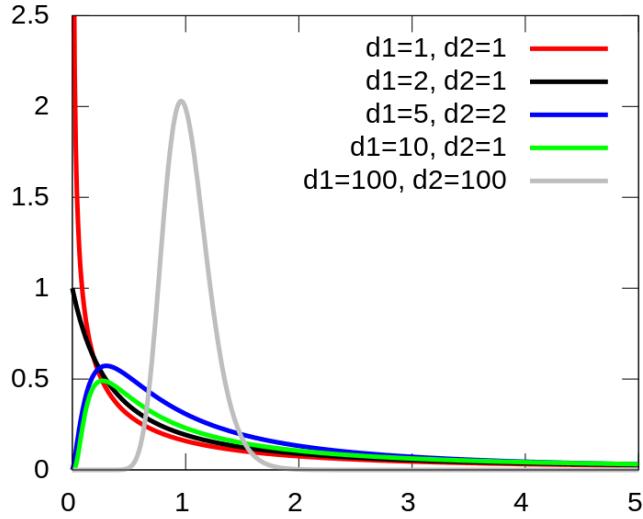
La densité de probabilité d'une loi de Fisher $\mathcal{F}(d_1, d_2)$ de degrés de liberté d_1 et d_2 est :

$$f_F(x) = \frac{\left(\frac{d_1 x}{d_1 x + d_2}\right)^{d_1/2} \left(1 - \frac{d_1 x}{d_1 x + d_2}\right)^{d_2/2}}{x \mathbf{B}(d_1/2, d_2/2)}$$

où $\mathbf{B}(x, y)$ est la fonction Beta :

$$\mathbf{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt$$

En voici ci-dessous une illustration pour plusieurs valeurs de d_1 et d_2 issue de wikipedia :



Loi du χ^2

La densité de probabilité de la loi du χ^2 à k degrés de libertés :

$$f_k(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{x/2-1} e^{-x/2}$$

où $\Gamma(x)$ est la fonction Gamma d'Euler :

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$

En voici ci-dessous une illustration pour plusieurs valeurs de k issue de wikipedia :

