République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abou Bekr Belkaid Tlemcen

Faculté des Sciences
Département de Mathématiques
MÉMOIRE DE MASTER

En vue de l'obtention du
Diplôme de master en mathématiques.
Option : Probabilités et statistiques

QUELQUES METHODES D'APPRENTISSAGE SUPERVISE

Présenté par : AISSAOUI Fatiha

Mémoire soutenu le date devant le jury composé de :

A. Allam	MCA	UABB	Tlemcen	**Président**
F. Boukhari	Professeur	UABB	Tlemcen	**Examinateur**
M. Kada Kloucha	MCB	UABB	Tlemcen	**Directrice de thèse**

Année universitaire : 2020-2021

Dédicace

A mes très chers parents qui ont toujours été là pour moi, et qui m'ont donné un magnifique modèle de labeur et de persévérance. J'espère qu'ils trouveront dans ce travail toute ma reconnaissance et tout mon amour.

A mes chers frères et soeurs.

A mes meilleurs amies.

Remerciements

Avant tout, je remercie ALLAH de m'avoir donné le courage et la patience d'entamer et de finir ce mémoire. Je remercie vivement Madame Kada kloucha maître de conférences à l'Université Abou bekr Belkaid-Tlemcen, pour son aide précieux et conseils éclairés durant la réalisation de ce travail, ainsi que le temps qu'elle a bien voulu me consacrer.

J'exprime également ma gratitude à Monsieur A. ALLAM maître de conférence et chef de département à l'Université Abou Bekr Belkaid-Tlemcen, pour sa compétence, sa disponibilité et son soutient durant ces deux années de Master, et de m'avoir fait l'honneur de présider le jury de ce mémoire.

Mes vifs remerciement s'adresse aussi à Monsieur F. Boukhari professeur à l'Université Abou Bekr Belkai-Tlemcen, d'avoir accepter d'examiner ce mémoire et faire partie du jury.

Je tiens enfin à remercier tous ceux qui ont contribué de prés ou de loin à la réalisation de ce travail, qu'ils trouvent ici la traduction de mon gratitude et ma reconnaissance.

Table des matières

Introduction

L'apprentissage automatique (Machine Learning) constitue une grande avancée lorsqu'on veut créer une intelligence artificielle ou tenter d'obtenir un aperçu de toutes les données qui ont été collectée.
On distingue trois techniques de " Machine Learning " :
Apprentissage supervisé : où on dispose d'un ensemble d'objets et pour chaque objet une valeur cible associée.
Apprentissage non supervisé : où on dispose d'un ensemble d'objets sans aucune valeur cible associée.
Apprentissage semi-supervisé : où on dispose d'un petit ensemble d'objets avec pour chacun une valeur cible associée et d'un plus grand ensemble d'objets sans valeur cible.
Il existe une large variété d'algorithmes de l'apprentissage automatique, certains sont toute fois plus couramment utilisés que d'autres.
Dans ce mémoire nous allons nous intéresser à l'une des étapes initiales du Machine Learning : "**l'apprentissage supervisé**", en introduisant 2 algorithmes principaux :
1. La régression linéaire :
L'un des algorithmes d'apprentissage supervisé les plus populaires, lorsque la valeur cible à prédire est continue, est de la régression linéaire , il est utilisé pour expliquer la relation entre une variable dépendante et une ou plusieurs variables indépendantes, et donner une prédiction des valeurs continues (températures, Consommation de l'énergie électrique,...)et elle est utilisée dans d'autres domaines (sciences et techniques), où on peut expliquer les variations d'une variable donnée Y en fonction des variations d'autres variables $X_1, X_2, ..., X_p$.
La solution classique à ce problème est le modéle linéaire suivant

$$Y = b_0 + b_1X_1 + ... + b_pX_p + \varepsilon. \tag{1}$$

Avec :

- $b = (b_0, b_1, ..., b_p)$ est le vecteur des paramètres inconnus (coefficients de régression) ;
- ε est l'erreur de la régression ;
- Y est la variable à expliquée (endogène) ;
- $X_1, ..., X_p$ les variables explicatives.

pour $p = 1$ on parle d'une régression linéaire simple

$$Y = b_0 + b_1X_1 + \varepsilon$$

Plusieur méthodes de régression peuvent être utilisées, pour estimer le vecteur des paramètres b à partir des observations.

Le modèle de régression le plus connu, est l'estimateur linéaire des moindre carrés (MCO) qui consiste à minimiser la somme des carrés résiduels, c'est à dire :

$$\hat{\beta} = argmin_{\beta \in \mathbb{R}^p} \|Y - X\beta\| \text{ avec } \|.\| \text{ est la norme } l_2(\forall x \in \mathbb{R}, \quad \|x\| = x'x). \tag{2}$$

Nous imposons les hypothèses suivantes :

$$\textbf{H1} : \quad \text{La matrice} \quad X \quad \text{est de plein rang.} \tag{3}$$

$$\textbf{H2 :}\ E(\varepsilon_i) = 0 \quad \forall i, \quad cov(\varepsilon_i, \varepsilon_j) = 0 \quad (\forall i \neq j), \quad Var(\varepsilon_i) = \sigma^2. \tag{4}$$

Sous l'hypothèse H1, l'estimateur des MC $\hat{b}$ de b se présente comme suit :

$$\hat{b} = (X'X)^{-1}X'Y \tag{5}$$

Pour la méthode des moindre carrées, on suppose que la matrice X^tX est inversible, c'est à dire, que X est de plein rang, d'après le théorème de Gauss-Markov, parmi les estimateurs sans biais, celui des moindres carrés est de variance minimale.
Mais en pratique, ce sont les variables colinéaires qui rendent la matrice X^tX singulière, par ailleurs, les MCO supposent également que $n > p$ (nombre d'observations > nombre de variables),
Par conséquent, pour résoudre ce problème, on peut avoir intérêt à sacrifier le biais et utiliser l'une des méthodes suivantes :

- **Méthodes de sélection :** c'est de choisir le sous-modèle dont l'estimation du risque de prévision est minimale.
- **Réduction de dimension :** c'est de projeter les p variables, évoluant en toute généralité dans un espace vectoriel de dimension p, dans un sous-espace vectoriel de dimension beaucoup plus petit.
  On peut alors régresser Y sur ce sous-espace afin d'éviter les écueils de la grande dimension, comme titre d'exemple les méthodes PCR et PLS.
- **Les estimations contraintes :** l'idée est d'utiliser une méthode d'estimation qui contrainte les paramètres à ne pas exploser (contrairement aux MCO(5) en grande dimension).
  Ainsi l'estimation est moins variable et les prévisions plus fiables, parmi ces méthodes, les techniques de type "Lasso" conduisent à estimer certains coefficients par 0, auquel cas une sélection de variables s'opère dans le même temps.

2. La régression logistique

les prévisions de la régression logistique sont des valeurs discrètes, c'est à dire, un ensemble fini de valeurs (Vrai ou faux par exemple). La régression logistique convient mieux à la classification binaire (dichotomique) où on peut considérer un ensemble de données ($y = 0$ ou 1) , où 1 représente la classe par défaut.
Au contraire de la régression linéaire, la régression logistique, propose le résultat sous forme de probabilités de la classe par défaut. Le résultat, donc, appartient à l'intervalle [0, 1].
Pour combiner entre les différentes caractéristiques, on utilise une fonction linéaire (exactement comme la régression linéaire) :

$$C(X) = a_0 + a_1X_1 + ... + a_pX_p,$$

Cette valeur est transformée à une probabilité en utilisant la fonction logistique. Donc, la probabilité qu'un échantillon avec les caractéristiques $x_1, \ldots, x_p$ appartienne à une classe y est calculée comme suit :

$$F(x) = P(Y = 1/X = x) = \frac{1}{1 + e^{-C(x)}}.$$

C'est la fonction de répartition de la loi logistique.

Pour prédire si un échantillon x appartient à une classe donnée (classe positive) $Y = 1$, on calcule sa probabilité en utilisant l'équation précédante. Un seuil est ensuite appliqué pour forcer cette probabilité dans une classification binaire.

On peut utiliser le seuil 0.5, dans ce cas

— Si $p(Y = 1|X) \geq 0.5$ donc la classe est positive

— Sinon la classe est négative

Ce mémoire est organisé en trois chapitres :
Dans le premier nous rappelons les résultats essentiels obtenus sur le modèle linéaire, en présentant la méthode des moindres carrés, qui consiste à déterminer l'estimateur du paramètre de régression.

Dans le deuxième chapitre, nous développons des résultats sur les régressions PCR, Ridge et Lasso en s'appuiyant principalement sur le livre de Pierre-André Cornillon et Eric Matzner-Lober, "Régression.Théorie et application" [13].

Le troisième chapitre est consacré à la "Régression Logistique", nous nous intéressons au model dichotomique, dans lequel la variable expliquée ne peut prendre que deux modalités, puis nous présentons le modèle Logit proposé par Verhulst (1804 − 1849) ([14], [16]et [17]), puis développé par Berkson (1944, 1951) ([4], [4]). Dans une seconde section, nous nous intéressons au problème de l'estimation des paramètres de ces modèles.

Dans le dernier chapitre, nous présentons des résultats de simulations numériques en utilisant le logiciel *R* sur différents exemples.

Chapitre 1

Régression linéaire multiple

La régression linéaire multiple est une méthode statistique fréquemment utilisée pour analyser les données lorsqu'elles existent plusieurs variables independantes, elle est utilisée en finance, et aussi dans d'autres disciplines scientifiques, tandis que dans certaines phénomènes, on peut expliquer les variations d'une variable donnée Y en fonction des variations d'autres variables $X_1, X_2, ..., X_p$.
le modèle de la régression linéaire multiple est donné par la formule suivante

$$Y = b_0 + b_1 X_1 + ... + b_p X_p + \varepsilon. \tag{1.1}$$

Avec :

- $b_0, b_1, ..., b_p$ sont des parèmètres inconnus (coefficients de régression) ;
- ε est l'erreur de la régression ;
- Y est la variable à expliquée (endogène) ;
- $X_1, ..., X_p$ les variables explicatives.
  pour $p = 1$ on parle d'une régression simple définie par :

$$Y = b_0 + b_1 X_1 + \varepsilon$$

1.1 Modélisation

La forme matricielle du model (1.1) est donnée par

$$Y = Xb + \varepsilon.$$

Ou encore

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ 1 & x_{1,2} & \dots & x_{p,2} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{1.2}$$

Avec :

- Y est un vecteur aleatoire de dimenssion n ;
- X est une matrice déterministe de taille $(n \times (p + 1))$ appelée une matrice du plan d'éxpérience ;
- b le vecteur de dimension $(p + 1)$ des paramètres inconnues du modèle ;

— ε est le vecteur des erreurs de dimension n. Pour la ième observation on a :

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_p x_{pi} + \varepsilon_i$$

On note par $\mathbf{H_1}$, $\mathbf{H_2}$ les hypothèses suivantes :

$$\mathbf{H_1} : \quad \text{La matrice} \quad X \quad \text{est de plein rang.} \tag{1.3}$$

$$\mathbf{H_2} : E(\varepsilon_i) = 0 \quad \forall i, \quad cov(\varepsilon_i, \varepsilon_j) = 0 \quad (\forall i \neq j), \quad Var(\varepsilon_i) = \sigma^2 \tag{1.4}$$

1.2 Estimation des paramètres du modèle

1.2.1 Estimateur des moindres carrés ordinaires

Définition 1. *L'estimateur des moindres carrés ordinaires $\hat{b}$ est définie comme suit :*

$$\begin{aligned} \hat{b} &= \arg\min_{b \in \mathbb{R}^{p+1}} \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2 \\ &= \arg\min_{b \in \mathbb{R}^{p+1}} \|Y - Xb\|^2. \end{aligned}$$

Avec : $x_{01} = \ldots = x_{0p} = 1$.

Proposition 1.2.1. *[8] Sous l'hypothèse* (**H1**) *(1.3), l'estimateur $\hat{b}$ des moindres carées ordinaires (**MCO**) de b est donné par :*

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

Remarque
L'hypothèse ($\mathbf{H_1}$) (1.3) assure que la matrice $(X'X)$ est bien inversible.
En effet, supposons qu'il existe un vecteur $b \in \mathbb{R}^{p+1}$ tel que

$$(X'X)b = 0.$$

Alors

$$\|Xb\|^2 = b'(X'X)b = 0.$$

Donc $Xb = 0$, par suite $b = 0$ car $rg(X) = p + 1$, c'est à dire la matrice $(X'X)$ est définie positive.
Preuve

[1] Posons $S(b) = \|Y - Xb\|^2$.
On cherche à déterminer le vecteur $b \in \mathbb{R}^{p+1}$ qui minimise S.

$$\begin{aligned} S(b) &= \|Y - Xb\|^2 \\ &= (Y - Xb)'(Y - Xb) \\ &= (Y' - (Xb)')(Y - Xb) \\ &= Y'Y - Y'Xb - b'X'Y + b'X'Xb \\ &= Y'Y - (X'Y)'b - b'X'Y + b'X'Xb. \end{aligned}$$

S est de type quadratique en b , avec la matrice $X'X$ symetrique definie positive donc S admet une unique solution.

$$\begin{aligned}\frac{\partial S(b)}{\partial b} &= -X'Y - X'Y + ((X'X)' + (X'X))b \\ &= -2X'Y + 2X'Xb \\ &= -2X'Y + 2(X'X)'b.\end{aligned}$$

$$\begin{aligned}\frac{\partial S(b)}{\partial b} = 0 &\iff X'Y = (X'X)'b \\ &\implies \hat{b} = (X'X)^{-1}X'Y.\end{aligned}$$

Vérifiant maintenant que $\hat{b}$ définit bien un minimum :

$$\frac{\partial^2 S}{\partial b^2} = 2(X'X)$$

et

$$\forall x \in \mathbb{R}^{p+1}, x'(2X'X)x \geq 0$$

donc

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

est un minimum. ■

1.2.2 Interprétation géometrique de MCO

Rappelons le principe : Y est le vecteur des variables à expliquer , la matrice X est formée de $(p+1)$ vecteurs colonne (la première colonne est étant généralement constituée de 1) .
Le sous espace de $\mathbb{R}^{p+1}$ engendré par les $(p+1)$ vecteurs de X est appelé espace image ou espace des solutions noté $M(X)$, est de dimension $(p+1)$ par hypothèse $(\mathbf{H_1})$ (1.3).
Tout vecteur de cet espace est de la forme $X\alpha$ où α est un vecteur de $\mathbb{R}^{p+1}$:

$$X\alpha = \alpha_0 + \alpha_1 X_1 + ... + \alpha_p X_p$$

Selon le modèle (1.1) le vecteur Y est la somme d'un élément de $M(X)$ et d'un bruit élément de $\mathbb{R}^n$, lequel n'a aucun raison d'appartenir à $M(X)$.
Minimiser $\|Y - X\alpha\|^2$ revient à rechercher un élément de $M(X)$ qui soit le plus proche de Y au sens de la norme euclidienne classique.
Cet unique élément est par définition le projeté orthogonal sur $M(X)$, il sera noté

$$\hat{Y} = P_X Y.$$

Où P_X est la projection orthogonale sur $M(X)$, il peut s'écrire aussi sous la forme

$$\hat{Y} = X\hat{b}.$$

Où $\hat{b}$ est l'éstimateur des MCO.
$M^{\perp}(X)$ est l'espace orthogonal de $M(X)$ ou l'espace des résidus de plus

$$dim M^{\perp}(X) = dim\mathbb{R}^n - dim M(X)$$

.

Proposition 1.2.2. *[7] $X\hat{b}$ est la projection orthogonale de Y sur* M(X).

Preuve
On veut montrer que $X\hat{b}$ est la projection orthogonale de Y sur $M(X)$, c'est à dire :

$$\forall Z \in M(X), \langle Y - X\hat{b}, Z\rangle = 0$$

$$Z \in M(X) \Rightarrow \exists \lambda \in \mathbb{R}^{p+1}, \quad Z = X\lambda.$$

$$\begin{aligned}
\langle Y - X\hat{b}, Z\rangle &= \langle Y - X\hat{b}, X\lambda\rangle \\
&= (X\lambda)'(Y - X\hat{b}) \\
&= \lambda' X'Y - \lambda' X'X\hat{b} \\
&= \lambda' X'Y - \lambda'(X'X)(X'X)^{-1}X'Y \\
&= \lambda' X'Y - \lambda' X'Y \\
&= 0
\end{aligned}$$

donc $X\hat{b}$ est la projection orthogonale de Y sur $M(X)$. ■

Définition 2. *La matrice de projection orthogonale sur* M(X)*, P_X (noté aussi par H), est définit par*

$$P_X = (X'X)^{-1}X'$$

Propriétés

— P_X est symetrique ($P'_X = P_X$).
— P_X^2=P_X.
— rg(P_X)=Tr(P_X)=p+1.

Remarques
Par définition on a :

— $\hat{Y} = P_X Y = X\hat{b} = X(X'X)^{-1}X'Y$.
— $P_{X^\perp} = (I - P_X)$ est la matrice de projection orthogonale sur $M^\perp(X)$.
— P_X vérifier bien que $P_X^2 = P_X$ et que P_X est symétrique.
— $Tr(P_X) = p + 1$ et $Tr(P_{X^\perp}) = n - p - 1$.
— On a la décomposition suivante :

$$Y = \hat{Y} + (Y - \hat{Y}) = P_X Y + (I - P_X)Y = P_X Y + P_{X^\perp} Y.$$

Par projection on peut tirer l'expréssion de $\hat{b}$, en effet : le projeté orthogonale $\hat{Y} = X\hat{b}$ est défiué comme l'unique vecteur tel que $(Y - \hat{Y})$ soit orthogonale à $M(X)$, puisque $M(X)$ est engendré par les vecteurs $X_1, X_2, ..., X_p$ ceci revient à dire que $(Y - \hat{Y})$ est orthogonal à chacun des X_i, ie :

$$\langle X_i, Y - X\hat{b}\rangle = 0, \quad i = 1 \dots p.$$

Alors :

$$\begin{aligned}
X'(Y - X\hat{b}) = 0 &\Rightarrow X'Y = X'X\hat{b} \\
&\Rightarrow \hat{b} = (X'X)X'Y.
\end{aligned}$$

Ce qui donne le même résultat de la proposition (1.2.1).

1.2.3 Quelques propriétés

On rappelle que la matrice de covariance du vecteur aléatoire $\hat{b}$ est définie par :

$$C_{\hat{b}} = E[(\hat{b} - E(\hat{b}))(\hat{b} - E(\hat{b}))'] = E(\hat{b}\hat{b}') - E(\hat{b})E(\hat{b})'$$

elle est de dimension $(p+1) \times (p+1)$.
De plus pour toute matrice A de taille $m \times (p+1)$ et tout vecteur u de dimension m déterministes on a :

$$E(A\hat{b} + u) = AE(\hat{b}) + u,$$

et

$$C_{A\hat{b}+u} = AC_{\hat{b}}A'.$$

Ces propriétés élémentaires seront appliquées dans la suite.

Proposition 1.2.3. *L'estimateur $\hat{b}$ des MCO (1.2.1) est un estimateur sans biais et sa matrice de covariance est donnée par :*

$$C_{\hat{b}} = \sigma^2(X'X)^{-1}$$

Preuve
pour le biais il suffit d'écrire :

$$\begin{aligned} E(\hat{b}) &= E((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'E(Xb + \varepsilon) \\ &= (X'X)^{-1}X'Xb \\ &= b, \end{aligned}$$

donc l'estimateur $\hat{b}$ des MCO est un estimateur sans biais.
Et pour la matrice de covariance on a :

$$\begin{aligned} C_{\hat{b}} &= (X'X)^{-1}X'C_Y X(X'X)^{-1} \\ &= (X'X)^{-1}X'C_{Xb+\varepsilon}X(X'X)^{-1} \\ &= (X'X)^{-1}X'C_{\varepsilon}X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

■

Théorème 1. *[3]**Théorème de Gauss :** L'estimateur $\hat{b}$ des MCO(1.2.1) est optimal (de variance minimale) parmi les estimateurs linéaires sans biais de b.*

Remarque :
Soient S_1 et S_2 deux matrices, on dit que $S_1 \geq S_2$ si et seulement si $S = (S_1 - S_2)$ est une matrice symétrique définie positive.
Preuve
On va montrer que pour tout autre estimateur $\tilde{b}$ de b linéaire et sans biais on a :

$$C_{\tilde{b}} \geq C_{\hat{b}}.$$

On sait que la matrice de covariance de la somme de deux vecteurs aléatoire u et v

est comme suit :

$$C_{u+v} = C_u + C_v + C_{u,v} + C_{v,u},$$

on décompose la matrice de covariance de $\tilde{b}$,

$$C_{\tilde{b}} = C_{\tilde{b}-\hat{b}+\hat{b}} = C_{\tilde{b}-\hat{b}} + C_{\hat{b}} + C_{\tilde{b}-\hat{b},\hat{b}} + C_{\hat{b},\tilde{b}-\hat{b}}.$$

Le but est de montrer que

$$C_{\tilde{b}-\hat{b},\hat{b}} = 0.$$

Posons pour tout matrice B de taille $(p+1) \times (n)$

$$\tilde{b} = BY$$

comme $\tilde{b}$ est un estimateur sans biais, alors pour tout $b \in \mathbb{R}^{p+1}$, on a :

$$\begin{aligned} E(\tilde{b}) = b &\Rightarrow E(BY) = b \\ &\Rightarrow E(B(Xb+\varepsilon)) = b \\ &\Rightarrow E(BXb) + E(B\varepsilon) = b \\ &\Rightarrow E(BXb) = b \\ &\Rightarrow BXb = b \\ &\Rightarrow BX = I. \end{aligned}$$

On a :
$C_{\tilde{b}-\hat{b},\hat{b}} = C_{\tilde{b},\hat{b}} - C_{\hat{b}} = C_{\tilde{b},\hat{b}} - \sigma^2(X'X)^{-1}$.
Comme $\hat{b} = (X'X)^{-1}X'Y = AY$, on a :

$$\begin{aligned} C_{\tilde{b},\hat{b}} &= C_{BY,AY} \\ &= AC_Y B' \\ &= (X'X)^{-1}X'(\sigma^2 I_n)B' \\ &= \sigma^2(X'X)^{-1}X'B' \\ &= \sigma^2(X'X)^{-1}(BX)' \\ &= \sigma^2(X'X)^{-1}, \end{aligned}$$

par suite :

$$C_{\tilde{b}-\hat{b},\hat{b}} = 0$$

de mème

$$C_{\hat{b},\tilde{b}-\hat{b}} = 0.$$

Donc :

$$C_{\tilde{b}} = C_{\tilde{b}-\hat{b}} + C_{\hat{b}}.$$

Comme $C_{\tilde{b}-\hat{b}} = C_{\tilde{b}} - C_{\hat{b}}$ est définie positive,
alors :

$$C_{\tilde{b}} > C_{\hat{b}}$$

d'où le resultat. ■

1.3 Résidu et variance résiduelle

Le résidu est définie par :

$$\hat{\varepsilon} = Y - \hat{Y} = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\varepsilon,$$

car $Y = Xb + \varepsilon$ et $Xb \in M(X)$.
on peut alors énoncer le résultat suivant :
propriétés
Sous les hyothèses (**H1**) (1.3),(**H2**) (1.4).
On a :

1) $E(\hat{\varepsilon}) = 0$.
2) $C_{\hat{\varepsilon}} = \sigma^2 P_{X^\perp}$.
3) $E(\hat{Y}) = Xb$.
4) $C_{\hat{Y}} = \sigma^2 P_X$.
5) $cov(\hat{\varepsilon}, \hat{Y}) = 0$.

Preuve
1)

$$\begin{aligned} E(\hat{\varepsilon}) &= E(P_{X^\perp}\varepsilon) \\ &= P_{X^\perp}E(\varepsilon) \\ &= 0. \end{aligned}$$

2)

$$\begin{aligned} C_{\hat{\varepsilon}} &= P_{X^\perp}C_\varepsilon P'_{X^\perp} \\ &= P_{X^\perp}C_\varepsilon P_{X^\perp} \\ &= \sigma^2 P_{X^\perp}P_{X^\perp} \\ &= \sigma^2 P_{X^\perp}. \end{aligned}$$

3)

$$\begin{aligned} E(\hat{Y}) &= E(X\hat{b}) \\ &= Xb. \end{aligned}$$

4)

$$\begin{aligned} C_{\hat{Y}} &= C_{X\hat{b}} \\ &= XC_{\hat{b}}X' \\ &= \sigma^2 X(X'X)^{-1}X' \\ &= \sigma^2 P_X. \end{aligned}$$

5) La covariance entre deux vecteurs aléatoires est bilinéaire :

$$\begin{aligned} C_{\hat{\varepsilon},\hat{Y}} &= C_{\hat{\varepsilon},Y-\hat{\varepsilon}} \\ &= C_{\hat{\varepsilon},Y} - C_{\hat{\varepsilon},\hat{\varepsilon}} \\ &= C_{\hat{\varepsilon},Y} - C_{\hat{\varepsilon}} \\ &= C_{P_{X^\perp}Y,Y} - \sigma^2 P_{X^\perp}. \end{aligned}$$

Puisque $C_Y = \sigma^2 I_n$ on a :

$$\begin{aligned} C_{\hat{\varepsilon},\hat{Y}} &= P_{X^\perp}C_Y - \sigma^2 P_{X^\perp} \\ &= 0. \end{aligned}$$

■

Un estimateur naturel de la variance résiduelle est donné par :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=0}^{n}\hat{\varepsilon}_i^{\,2} = \frac{1}{n}\|\hat{\varepsilon}\|^2$$

Cet estimateur est biaisé mais il est facile de le corriger comme le montre le résultat suivant :

Proposition 1.3.1. *[1] La statistique* $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p-1}$ *est un estimateur sans biais de* σ^2.

Preuve
On calcule $E(\|\hat{\varepsilon}\|^2)$ puisque c'est un scalaire donc il est égale à sa trace :

$$E(\|\hat{\varepsilon}\|^2) = E(Tr(\|\hat{\varepsilon}\|^2)) = E(Tr(\hat{\varepsilon}'\hat{\varepsilon})$$

et puisque pour tout matrice A, on a $Tr(A'A) = Tr(AA') = \sum_{i,j} a_{ij}^2$, alors :

$$E(\|\hat{\varepsilon}\|^2) = E(Tr(\hat{\varepsilon}'\hat{\varepsilon}) = Tr(E(\hat{\varepsilon}\hat{\varepsilon}')) = Tr(C_{\hat{\varepsilon}}) = Tr(\sigma^2 P_{X^\perp}) = \sigma^2 Tr(P_{X^\perp})$$

et comme $P_{X^\perp}$ est la matrice de projection orthogonale sur un espace de dimenssion $(n-p-1)$.
On a bien :

$$E(\|\hat{\varepsilon}\|^2) = (n-p-1)\sigma^2.$$

Proposition 1.3.2. *[1] Sous l'hypothèse* $\mathbf{H_2}$ *(1.4) on a :*

1) $\hat{b} \rightsquigarrow N(b, \sigma^2(X'X)^{-1})$.

2) $(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \rightsquigarrow \chi^2_{n-p-1}$.

3) $\hat{b}$ *et* $\hat{\sigma}^2$ *sont independants.*

Preuve
1)Soit $z \in \mathbb{R}^{p+1}$, $A = (X'X)^{-1}X'$
$\phi_{\hat{b}}$ est la fonction caractéristique du vecteur $\hat{b}$:

$$\begin{aligned}\phi_{\hat{b}}(z) &= E(\exp(i\langle z, \hat{b}\rangle)) \\ &= E(\exp(i\langle z, AY\rangle)) \\ &= E(\exp(i\langle A'z, Y\rangle)) \\ &= \phi_Y(A'z)\end{aligned}$$

d'autre part on a $Y \rightsquigarrow N(Xb, \sigma^2 I_n)$ car $Y = Xb + \varepsilon$, d'où.

$$\phi_{\hat{b}}(z) = \exp(i\langle A'z, Xb\rangle) - \frac{1}{2}q_{\sigma^2 I_n}(A'z).$$

Or :
$q_{\sigma^2 I_n}(A'z)$ est la forme quadratique associer à la matrice $\sigma^2 I_n$

$$\begin{aligned}q_{\sigma^2 I_n}(A'z) &= (A'z)'(\sigma^2 I_n)(A'z) \\ &= z'A\sigma^2 I_n A'z \\ &= z'(\sigma^2 AA')z \\ &= z'(\sigma^2(X'X)^{-1}X'X(X'X)^{-1})z \\ &= z'(\sigma^2(X'X)^{-1})z \\ &= q_{\sigma^2(X'X)^{-1}}(z).\end{aligned}$$

De plus

$$\begin{aligned}\langle A'z, Xb\rangle &= \langle z, AXb\rangle \\ &= \langle z, (X'X)^{-1}X'Xb\rangle \\ &= \langle z, b\rangle,\end{aligned}$$

d'où

$$\phi_{\hat{b}}(z) = \exp(i\langle z, b\rangle) - \frac{1}{2}q_{\sigma^2(X'X)^{-1}}(z),$$

ainsi

$$\hat{b} \rightsquigarrow N(b, \sigma^2(X'X)^{-1}).$$

2) On a :

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-p-1}\|Y - \hat{Y}\|^2 \\ &= \frac{1}{n-p-1}\|\hat{\varepsilon}\|^2.\end{aligned}$$

Avec $\hat{Y} = X\hat{b} = P_X Y$ est la projection orthogonale de Y sur $M(X)$ telle que $P_X = X(X'X)^{-1}X'$ et $P_{X^\perp} = (I - P_X)$.
et

$$\begin{aligned}\hat{\varepsilon} &= Y - \hat{Y} \\ &= (I - P_X)Y \\ &= P_{X^\perp}Y \\ &= P_{X^\perp}\varepsilon \\ &= (I - P_X)\varepsilon.\end{aligned}$$

D'aprés le théorème de Cochron on a :

$$\frac{\|\hat{\varepsilon}\|^2}{\sigma^2} = \frac{\|P_{X^\perp}(\varepsilon) - P_{X^\perp}(E(\varepsilon))\|^2}{\sigma^2} \rightsquigarrow \chi^2_{n-p-1}.$$

Or

$$\begin{aligned}\hat{\sigma}^2 = \frac{1}{n-p-1}\|\hat{\varepsilon}\|^2 &\Rightarrow (n-p-1)\hat{\sigma}^2 = \|\hat{\varepsilon}\|^2 \\ &\Rightarrow (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \rightsquigarrow \chi^2_{n-p-1}.\end{aligned}$$

Donc

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \rightsquigarrow \chi^2_{n-p-1}.$$

3) $\hat{b}$ et $\hat{\sigma}^2$ sont indépendants, en effet :

$$\hat{\varepsilon} = (I - P_X)\varepsilon = (I - X(X'X)^{-1}X')\varepsilon$$

$$\begin{aligned}
C_{\hat{b},\hat{\varepsilon}} &= E(\hat{b}\hat{\varepsilon}') - E(\hat{b})E(\hat{\varepsilon}') \\
&= E(\hat{b}\hat{\varepsilon}') \\
&= E((X'X)^{-1}X'Y\varepsilon'(I - P_X)') \\
&= E((X'X)^{-1}X'Y\varepsilon' - (X'X)^{-1}X'Y\varepsilon'P_X') \\
&= E((X'X)^{-1}X'(Xb + \varepsilon)\varepsilon' - (X'X)^{-1}X'(Xb + \varepsilon)\varepsilon'P_X') \\
&= E(b\varepsilon' + (X'X)^{-1}X'\varepsilon'\varepsilon - b\varepsilon P_X' - (X'X)^{-1}\varepsilon\varepsilon'P_X') \\
&= E((X'X)^{-1}X'\varepsilon\varepsilon') - E((X'X)^{-1}X'\varepsilon\varepsilon'P_X') \\
&= (X'X)^{-1}X'E(\varepsilon\varepsilon') - (X'X)^{-1}X'E(\varepsilon\varepsilon'P_X') \\
&= (X'X)^{-1}X'\sigma^2 I_n - (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1}X' \\
&= 0
\end{aligned}$$

On déduit que $\hat{b}$ et $\hat{\varepsilon}$ sont non corrélé donc indépendant et comme
$\hat{\sigma^2} = \frac{1}{n-p-1}\|\hat{\varepsilon}\|^2$
alors $\hat{b}$ et $\hat{\sigma^2}$ sont indépendants aussi. ■

1.4 Evaluation globale de la régréssion

1.4.1 Coefficient de détermination

$$\begin{aligned}
\sum_{i=0}^{n}(y_i - \bar{y_n})^2 &= \sum_{i=0}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y_n})^2 \\
&= \sum_{i=0}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=0}^{n}(\hat{y}_i - \bar{y_n})^2 + 2\sum_{i=0}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y_n}).
\end{aligned}$$

puisque $\hat{Y}$ est la projection orthogonale de Y sur $M(X)$ alors :

$$\sum_{i=0}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y_n}) = 0.$$

d'où

$$\sum_{i=0}^{n}(y_i - \bar{y_n})^2 = \sum_{i=0}^{n}(\hat{y}_i - \bar{y_n})^2 + \sum_{i=0}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=0}^{n}(\hat{y}_i - \bar{y_n})^2 + \|\hat{\varepsilon}\|^2.$$

où $\hat{\varepsilon} = Y - \hat{Y}$.

On en déduit l'équation fondamentale de l'analyse de variance :

$$SCT = SCE + SCR$$

avec
SCT : est la variabilité totale ;
SCE : la variabilité expliqué ;
SCR : la variabilité résiduelle.
Cette équation permet de juger la qualité d'ajustement d'un modèle.
Ce pendant, ces valeurs dependent des unités de mesure , c'est pourquoi il'est préférable d'utiliser le nombre sans dimension.

Définition 3. *Le coefficient de détermination(ou coefficient de corrélation multiple) qui mesure la proportion de la variance de Y expliquée par la régréssion de Y sur X,noté R^2, est définie par :*

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Propriété :

— $0 < R^2 < 1$, plus la valeur de R^2 est proche de 1, plus le modèle est plus significatif.

— Si $n < p$ alors on calcule le coefficient de détermination corrigé(ajusté) :

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{n-1}{n-p-1}\frac{SCR}{SCT}$$

.

Test du critère significatif d'un des coefficients de régression :
test de student :
Tester l'influence directe de la variable explicative sur la variable endogène, revient à tester son coefficient de régression s'il est égale ou différent de 0, pour un seuil choisi, en général $\alpha = 0,05$.
Le test d'hypothèse est le suivant : est appelé (test bilatéral)

$$\mathcal{H}_0 : b_j = 0$$

contre :

$$\mathcal{H}_1 : b_j \neq 0$$

La statistique de student est la suivante :

$$T_c^{\alpha} = \left| \frac{\hat{b_j} - b_j}{\hat{\sigma}_{\hat{b_j}}} \right| \rightsquigarrow T_{n-p-1,\frac{\alpha}{2}}$$

où :

— T_c^{α} : désigne la valeur critique de la statistique T (dite calculée).

— $\hat{b_j}$: désigne la valeur estimée du paramètre b_j.

— $\hat{\sigma}_{\hat{b_j}}$: désigne la valeur de l'écart-type du $\hat{b_j}$ qui vaut

$$\hat{\sigma}_{\hat{b_j}} = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$$

c'est à dire la racine carrée du j-ème terme de la diagonale de $C_{\hat{b}}$.

— $n - p - 1$ est le degré de liberté.

Règle de décision :
si $|T_c| \leq T_t^{\alpha=0,05}$ on accepte l'hypothèse $\mathcal{H}_0$, la variable X_j n'est pas contributive à l'explication de Y.
Test unilatéral : ce test est utilisé lorsque : $\mathcal{H}_1 : b_j > 0$ ou $b_j < 0$.

$$T_c^{\alpha} = \left| \frac{\hat{b_j} - b_j}{\hat{\sigma}_{\hat{b_j}}} \right| \rightsquigarrow T_{n-p-1,\alpha}$$

Chapitre 2

Régréssion avec biais

On se place dans le cadre d'un modèle de régréssion linéaire (1.1), comme on l'a vu dans le premier chapitre ce modèle repose sur les hypothèses ($\mathbf{H_1}$) (1.3), et ($\mathbf{H_2}$) (1.4).
Mais il y a des cas où ces hypothèses ne seront pas vérifier :

— Si $p \gg n$ (la présence de nombreuses variables explicatives)

— Si $n \geq p$ mais les variables explicatives $X_1, X_2, ...X_{p+1}$ sont liées autrement-dit une(ou plusieurs) variable(s) sont linéairement redondante(s) , ie :

$$\exists j, X_j = \sum_{j \neq k}^{n} (\alpha_k X_k)$$

Alors $rg(X) < p$ (c'est à dire $X'X$ n'est pas inversible) qui rend l'estimateur des moindres carrées peu fiable car sa variance explose.
Ainsi en présence de ses cas fréquents on ramène à un modèle mal estimé (parcimonieu), donc il est nécéssaire d'introduire des méthodes alternatives aux MCO(1.2.1) afin d'estimer au mieux le modèle lineaire (1.1).

— Méthodes de sélection : c'est de choisir le sous-modèle tel que l'estimation du risque de prévision est d'une valeur minimale.

— Réduction de dimension : c'est de projeter les p variables, dans un espace vectoriel de dimension p, dans un sous-espace vectoriel de dimension beaucoup plus petit.
On peut alors régresser Y sur ce sous-espace afin d'éviter les obstacle de la grande dimension, comme titre d'exemple les méthodes PCR et PLS.

— Les estimations contraintes, l'idée est de trouver une méthode d'estimation qui met les contraintes sur les paramètres à ne pas exploser (contrairement aux MCO(1.2.1) en grande dimension).
Alors l'estimation est moins variable et par conséquent les prévisions plus fiables, parmi ces techniques , les méthodes de "Lasso" amènent à estimer certains coefficients par 0, dans ce cas une sélection de variables se produire dans le même temps.

Dans tous la suite on suppose que les variables explicatives $X_1, X_2, ...X_{p+1}$ sont centées et réduites.

2.1 Régression sur composantes principales (PCR)

Le principe de l'analyse en composantes principales (ACP ou PCA en anglais) est de trouver une base orthogonale de l'espace vectoriel $M(X)$ dont les vecteurs

$Z_1, ..., Z_{p+1}$, appelés composantes principales , sont construits de telle sorte à garder le plus "d'information" possible contenue dans les vecteurs initiaux $X_1, ..., X_{p+1}$.
L'information est caractérisé par la variance des coordonnées des n individus sur chaque nouvel axe :
Plus la variance est élevée et plus le nuage de points se répartit bien sur l'axe, gardant ainsi le maximum de la diversité initiale du nuage , par contre au cas extrême inverse où tous les points seraient projetés au même endroit qui conduit à une variance nulle.

2.1.1 Modélisation

Soit X la matrice des variables centrées réduites.

— Le premier axe Z_1 est la combinaison linéaire de $X_1, ..., X_{p+1}$ de variance maximale :

$$Z_1 = X\alpha_1$$

avec

— $\alpha_1 \in \mathbb{R}^{p+1}$ représente la direction du premier axe principal et $\|\alpha_1\| = 1$.

— $X\alpha_1 \in \mathbb{R}^n$ est l'ensemble des coordonnées du nuage de points sur cet axe.

— $Var(X\alpha_1)$ maximale parmi tous les vecteurs de la forme $X\alpha$.

— De la même manière construisant le second axe, avec la contrainte d'être orthogonal à Z_1, qu'on peut l'écrit à l'aide du produit scalaire

$$\langle X\alpha_1, X\alpha_2 \rangle = 0.$$

— D'une façon générale, pour $j = 1, ..., p + 1$ le j-ème axe est

$$Z_j = X\alpha_j$$

où :

$$\alpha_j = \underset{\alpha \in \mathbb{R}^{p+1}}{\arg\max} Var(X\alpha)$$

Sous les contraintes :

— $\|\alpha\| = 1$

— $\alpha' X' X\alpha_l = 0$ pour tout $l = 1, ..., j - 1$.

— Puisque les variables sont supposées centrées, on a

$$Var(X\alpha) = \alpha' X' X\alpha.$$

Finalement tous les axes sont orthogonaux et ils sont ordonnés du plus "informatif " Z_1, au moins informatif Z_{p+1}, au sens où la variance des coordonnées des individus sur les axes décroit.
Pour résoudre le problème d'optimisation précédent on utilise l'algorithme suivant : les directions α_j représentent les vecteurs propres (normalisés) de la matrice $X'X$, ordonnés par ordre décroissant de leur valeur propre associée.
La régression sur composantes principales (PCR) est basée sur l'idée suivante :
Plutôt que de régresser Y sur $X_1, ..., X_{p+1}$, on régresse Y sur les premiers axes principaux $Z_1, ..., Z_m$, où le nombre m de composantes obtenues est à choisir.

Ce principe permet de réduire la dimension du problème de $p + 1$ à m, sans perdre l'orthogonalité des variables explicatives retenues.
Le modèle s'écrit ainsi :

$$Y = \gamma_0 + \sum_{j=1}^{m} \gamma_j Z_j + \varepsilon$$

les estimateurs par moindres carrés associés sont

$$\hat{\gamma}_0 = \bar{Y}, \hat{\gamma}_j = \frac{Y'Z_j}{Z_j'Z_j}.$$

L'interprétation des m composantes principales retenues et des coefficients estimés $\hat{\gamma}_j$ de la PCR n'est pas toujours aisée.
Mais, il est possible de revenir aux variables initiales puisque $Z_j = X\alpha_j$.
On obtient

$$\hat{Y} = \bar{Y} + X\hat{b}$$

avec :

$$\hat{b} = \sum_{j=1}^{m} \hat{\gamma}_j \alpha_j$$

2.2 Régression Ridge

En présence de nombreuses variables explicatives, la matrice $X'X$ n'est pas de plein rang et n'est donc plus inversible, ce problème rend l'estimateur des moindres carrés incalculable.
Aussi que , si les variables explicatives sont fortement corrélées entre elles (présence du problème de multicolinéarité), la matrice $X'X$ peut être inversible mais son inverse est instable , dans le sens où une légère modification des données peut conduire à une matrice inverse radicalement différente, et par suite l'instabilité de l'estimateur.
L'influence de ce phénomène est claire sur la variance de l'estimateur par MCO qui vaut

$$\sigma^2(X'X)^{-1}.$$

Lorsque $X'X$ est non inversible ou d'inverse instable, cela s'influe sur ses valeurs propres , certaines sont nulles ou quasiment nulles.
Une astuce pour régler le problème d'inversion consiste à ajouter une petite valeur $k > 0$ aux valeurs propres.
On remplace donc $X'X$ par $X'X + kI_p$ où I_p est la matrice identité de taille p, il s'agit de la régularisation de Tikhonov, connue en statistique sous le nom de régression ridge.

Définition 4. *L'estimateur Ridge de b est défini comme solution du problème de minimisation suivant :*

$$\hat{b}_{Ridge} = \underset{b \in R^p}{\arg\min} \|Y - Xb\|^2$$

sous la contrainte

$$\sum_{j=1}^{p} b_j^2 \leq t$$

ou bien

$$\hat{b}_{Ridge} = \underset{b \in R^p}{\arg\min} \|Y - Xb\|^2 + k(\|b\|^2 - t)$$

où k est le paramètre de régularisation (lié à t) qu'il convient de choisir en pratique.

Proposition 2.2.1. *L'estimateur Ridge de b dans un modèle de régression est défini par l'expréssion suivante :*

$$\hat{b}_{Ridge} = (X'X + kI)^{-1}X'Y$$

Preuve
Posons

$$RSS(k) = \|Y - Xb\|^2 + k(\|b\|^2 - t)$$

qui peut s'écrire sous la forme matricielle

$$RSS(k) = (Y - Xb)'(Y - Xb) + kb'b - ktI_p$$

ou encore

$$RSS(k) = Y'Y + 2b'X'Y + b'X'Xb + kb'b - ktI_p$$

en dérivant par rapport à b et en posant cette dérivé égal à zéro, on obtient :

$$X'Y = X'Xb + kb = (X'X + kI_p)b$$

Comme la matrice symetrique $X'X$ est définie semi-positive , alors $(X'X + kI_p)$ est aussi, et donc inversible.
D'où la formule explicite de l'estimateur Ridge :

$$\hat{b}_{\text{Ridge}} = (X'X + kI)^{-1}X'Y.$$

■

Remarque :
Le but d'introduire le paramètre k est de biaiser l'estimation d'une part, et de diminuer sa variance d'autre part (la stabilisation), remarquons que au sens d'erreur quadratique, l'estimateur ridge est préférable à l'estimateur par MCO pour des valeurs petites de k.

Proposition 2.2.2. *(l'espérance et la variance de l'estimateur Ridge)*

1. $\mathbb{E}(\hat{b}_{Ridge}) = b - k(X'X + KI)^{-1}b.$
2. $C_{\hat{b}_{Ridge}} = \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1}.$

Preuve
1. Pour le calcule de l'esperance, rappellons que l'estimateur de b de MCO est défini par :

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

d'ou

$$(X'X)\hat{b} = X'Y$$

alors

$$\begin{aligned}
\mathbb{E}(\hat{b}_{\text{Ridge}}) &= \mathbb{E}((X'X + kI)^{-1}X'Y) \\
&= \mathbb{E}((X'X + kI)^{-1}(X'X)\hat{b}) \\
&= (X'X + kI)^{-1}X'X\mathbb{E}(\hat{b}) \\
&= (X'X + kI)^{-1}X'Xb
\end{aligned}$$

2.Pour la variance on a :

$$\begin{aligned}
C_{\hat{b}_{\text{Ridge}}} &= (X'X + kI)^{-1}X'C_YX(X'X + kI)^{-1} \\
&= \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1}.
\end{aligned}$$

■

Avantages et inconvénients de la méthode Ridge :

— **Avantages :**
  — La méthode Ridge améliore l'erreur de prédiction, en réduisant la variance des estimateurs.
  — robuste à la multicolinéarité.
  — robuste à la grande dimension.
— **Inconvénients :**
  — La méthode Ridge ne pénalise pas les variables nuisibles par des coefficients exactement nuls, donc elle ne produit pas de parcimonie dans le modèle.
  — ne sélectionne aucune variable mais rétrécit les coefficients.

2.3 Régression Lasso

Modélisation
Le principe de la régression Lasso (Least Absolute Shrinkage and Selection Operator) est le même que la régression ridge, mais au lieu de prendre la contrainte sur la norme l^2 de b à ne pas exploser, elle contraint sa norme l^1 :

$$\hat{b}_{Lasso} = \underset{b\in\mathbb{R}^p}{\arg\min} \|Y - Xb\|^2$$

sous la contrainte

$$\sum_{j=1}^{p} |b_j| \leq \kappa$$

Le problème Lasso peut s'écrire en un problème équivalent avec le lagrangien

$$\hat{b}_{Lasso} = \arg\min_{b\in R^p} \left(\|Y - Xb\|^2 + 2\lambda \sum_{j=1}^{p} |b_j| \right) \tag{2.1}$$

avec 2λ est le multiplicateur de Lagrange lié à κ par la contrainte $\sum_{j=1}^{p} |b_j| \leq \kappa$,
on a utiliser 2λ et non plus λ pour simplifier quelques calculs dans la résolution du problème .

Calcul analytique de la solution de la méthode Lasso :
Cherchons la forme explicite de la solution du problème (2.1) pour chaque b_j à condition de garder tous les autres paramètres fixes.
Le problème (2.1) est equivalent à l'ecriture matricielle suivante :

$$\begin{aligned} \hat{b}_{Lasso} &= \arg\min_{b\in R^p} \left((Y - Xb)'(Y - Xb) + 2\lambda \sum_{j=1}^{p} |b_j| \right) \\ &= \arg\min_{b\in R^p} \left(Y'Y - Y'Xb - (Xb)'Y + (Xb)'Xb + 2\lambda \sum_{j=1}^{p} |b_j| \right) \end{aligned}$$

En gardant que les quantitées qui dépendent de b_j, posons $L(b_j)$ la fonction qui minimise le problème en b_j tel que :

$$L(b_j) = -2X_j'Yb + \sum_{i=1}^{n} x_{ij}^2 b_j^2 + 2 \sum_{i=1}^{n} \sum_{k\neq j} x_{ik}x_{ij}b_k b_j + 2\lambda|b_j|.$$

c'est une equation quadratique en b_j dérivable pour $b_j \neq 0$:

$$\frac{\partial L(b_j)}{\partial b_j} = -2X_j'Y + 2\sum_{i=1}^{n} x_{ij}^2 b_j + 2\sum_{k\neq j} x_{ik}x_{ij}b_k + 2\lambda signe\{b_j\}$$

D'ou

$$-2X_j'Y + 2\sum_{i=1}^{n} x_{ij}^2 b_j + 2\sum_{k\neq j} x_{ik}x_{ij}b_k + 2\lambda signe\{b_j\} = 0$$

On a $\sum_{i=1}^{n} x_{ij}^2 = 1$, ainsi :

$$\begin{aligned} \hat{b}_j &= X_j'Y - \sum_{k\neq j} x_{ik}x_{ij}b_k - \lambda signe\{b_j\} \\ &= \sum_{i=1}^{n}(y_i - r_i^{(j)})x_{ij} - \lambda signe\{b_j\} \end{aligned}$$

avec $\sum_{k\neq j} x_{ik}b_k = r_i^{(j)}$ D'où

$$\hat{b}_j = signe\{\sum_{i=1}^{n}(y_i - r_i^{(j)})x_{ij}\}(|\sum_{i=1}^{n}(y_i - r_i^{(j)})x_{ij}| - \lambda)_+.$$

avec $x_+ = max(x, 0)$

Avantages et inconvénients de la méthode LASSO :

- **Avantages :**
  - La méthode LASSO élimine les variables nuisibles dans le modèle en estimant leur coefficients par des zéros.
  - Elle choisit les variables qui contribuent le plus dans le modèle .
  - Elle rétrécit les coefficients vers zéro.
- **Inconvénients :**
  - La méthode Lasso est une méthode non appropriée pour la sélection des groupes des prédicteurs. En effet, si des prédicteurs sont fortement corrélés entre eux, elle choisit un prédicteur et pénalise les autres avec des coefficients nuls ;
  - Si p > n, l'approche Lasso choisie au maximum n variables.

2.4 Régression Elastic Net

La régression Elastic net est une méthode qui combine les deux normes, la norme L1 et la norme L2, ainsi est un compromis entre Ridge et Lasso, elle est introduite la première fois par[15].
Le principe est de garantir une meilleure robustesse en cas de multicolinéarité, propriété de l'estimateur Ridge,tout en tirant partie des qualités de sélection de l'estimateur Lasso.
L'estimateur elastic-net est la solution du problème :

$$\hat{b}_{E-N} = \underset{b\in R^p}{\arg\min} \|Y - Xb\|^2$$

sous la contrainte

$$\sum_{j=1}^{p}((1-\alpha)b_j^2 + \alpha|b_j|) \leq t$$

ou bien

$$\hat{b}_{E-N} = \underset{b\in R^p}{\arg\min} \|Y - Xb\|^2 + \lambda \sum_{j=1}^{p}((1-\alpha)b_j^2 + \alpha|b_j|)$$

où λ est le paramètre de régularisation (lié à t) qu'il convient de choisir en pratique.
Avantages de la méthode Elastic Net :

— La méthode Elastic Net tient compte de la corrélation entre les prédicteurs.

— Elle favorise la parcimonie .

— La méthode peut sélectionner $p > n$ variables dans le modèle, contrairement à la méthode Lasso .

Chapitre 3

Régression logistique Binaire

La régression logistique est l'une des méthodes d'apprentissage automatique, consiste à construire un modèle permettant d'expliquer les valeurs prises par une variable cible, le plus souvent, variable dichotomique telle que "présence ou absence", "oui ou non" et "A ou B"..., on parle dans ce cas de la régression logistique binaire (si elle possède plus de 2 modalités, on parle de régression logistique polytomique) à partir d'un ensemble de variables explicatives quantitatives ou qualitatives.

3.1 Notations et hypothèses :

L'objectif de la régréssion logistique est de prédire une variable dichotomique Y.
On dispose d'un échantillon Ω de taille n , La valeur prise pour un individu ω est notée $Y(\omega)$,
et à partir d'une collection de descripteurs, $X = (X_1, X_2, ..., X_p)$, le vecteur d'observation pour un individu ω s'écrit
$X(\omega) = (X_1(\omega), X_2(\omega), ..., X_p(\omega))$.

Il s'agit en quelque sorte, de mêttre en évidence l'existence d'une liaison fonctionnelle sous-jacente entre ces variables de la forme

$$Y = f(X, \alpha)$$

avec

— $f(.)$ est le modèle de prédiction (appelé aussi classifieur),

— α est le vecteur des paramètres de la fonction f .

Dans le cadre de la discrimination binaire, on considère une variable dépendante Y à deux modalités (positif " + " ou négatif " − ") , cette question est connue sous le nom d'apprentissage "supervisé", on cherche à prédire correctement les valeurs de Y , et on peut quantifier la probabilité d'un individu à être positif (ou négatif).
La probabilité a priori d'être positif s'écrit $p'(\omega) = P(Y(\omega) = +)$, lorsqu'il ne peut y avoir d'ambiguïtés, nous la noterons simplement p'.
Le classifieur bayesien est celui qui répond d'une manière optimale aux spécifications ci-dessus, il s'agit de calculer les probabilités conditionnelles (probabilité a posteriori)

$$pi(\omega) = P(Y(\omega) = y_k/X(\omega))$$

pour chaque modalité y_k de Y , de même , lorsqu'il ne peut y avoir de confusions, nous écrirons pi.
On affecte à l'individu la modalité la plus probable $\hat{Y}$ c'est à dire :

$$\hat{Y} = \arg\max_k P(Y = y_k/X)$$

3.2 Estimation des paramètres :

le but est d'estimer la probabilité conditionnelle $P(Y/X)$.

Théorème 2. *Théorème de Bays :*
Soit X ,Y deux variables aléatoires, la probabilité conditionnelle $P(Y/X)$ est définie par :

$$P(Y(\omega) = y_k/X(\omega)) = \frac{P(Y = y_k) \times P(X/Y = y_k)}{P(X)} \tag{3.1}$$

$$= \frac{P(Y = y_k) \times P(X/Y = y_k)}{\sum_{l=1}^{k} P(Y = y_l)P(X/Y = y_l)} \tag{3.2}$$

Dans le cas de deux classes nous devons simplement comparer $P(Y = +/X)$ et $P(Y = -/X)$. On obtient

$$\frac{P(Y = +/X)}{P(Y = -/X)} = \frac{P(Y = +)}{P(Y = -)} \times \frac{P(X/Y = +)}{P(X/Y = -)} \tag{3.3}$$

La règle de decision devient

$$Si \quad \frac{P(Y = +/X)}{P(Y = -/X)} > 1 \quad alors \quad Y = +$$

Le rapport $\frac{P(Y=+)}{P(Y=-)}$ est facile à estimer à partir des données issu d'un tirage aléatoire dans la population, indepandamment des classes d'appartenance des individus, il suffit de prendre le rapport entre le nombre d'observations positive et négative $\frac{n_+}{n_-}$.
Donc, le problème est de donner une estimation du rapport des probabilités

$$\frac{P(X/Y = +)}{P(X/Y = -)}.$$

La régression logistique repose sur l'hypothèse fondamentale suivante :

$$\ln\left(\frac{P(X/Y = +)}{P(X/Y = -)}\right) = b_0 + b_1X_1 + ... + b_pX_p. \tag{3.4}$$

Cette hypothèse couvre une large classe de lois distributions des données [2]
- La loi normale ;
- Les lois exponentielles ;
- Les lois discrètes ;
- Les lois Beta, les lois Gamma et les lois de Poisson ;
- Un mélange de variables explicatives binaires (0/1) et numeriques, cette propriété est très importante car elle rend opérationnelle la régression logistique dans de très nombreuses configurations.

Contrairement à l'Analyse Discriminante Linéaire, que l'on qualifie de méthode paramétrique car on émet une hypothèse sur les distributions respectives de $P(X/Y = +)$ et $P(X/Y = -)$ (loi normale).
La régression logistique est une méthode semi-paramétrique car l'hypothèse porte uniquement sur le rapport de ces probabilités, elle est moins restrictive, et son champ d'action est donc théoriquement plus large.

3.3 Le modèle LOGIT :

On appelle transformation **LOGIT** de $pi(\omega) = P(Y(\omega) = +/X(\omega))$ l'expression

$$\ln(\frac{pi}{1-pi}) = a_0 + a_1X_1 + ... + a_pX_p \tag{3.5}$$

comme on est dans le cadre binaire

$$P(Y = -/X) = 1 - pi$$

la quantité

$$\frac{pi}{1-pi}$$

exprime un odds (ou bien un rapport de chance).
Avec la notation

$$C(X) = a_0 + a_1X_1 + ... + a_pX_p,$$

nous obtenons

$$\begin{aligned} pi &= \frac{\exp(C(X))}{1+\exp(C(X))} \\ &= \frac{1}{1+\exp(-C(X))} = F(Xa), \end{aligned} \tag{3.6}$$

c'est la fonction de répartition de la loi logistique.
Comparaison entre les deux approches :
Les deux approches correspondent à deux facettes d'un même problème.
En effet :

$$\begin{aligned} \ln\frac{pi}{1-pi} &= a_0 + a_1X_1 + ... + a_pX_p \\ &= \ln\frac{P(Y=+)}{P(Y=-)} \times \frac{P(X/Y=+)}{P(X/Y=-)} \\ &= \ln\frac{P(Y=+)}{P(Y=-)} + \ln\frac{P(X/Y=+)}{P(X/Y=-)} \\ &= \ln\frac{p'}{1-p'} + b_0 + b_1X_1 + ... + b_pX_p(\text{ avec } p' = P(Y=+)). \end{aligned}$$

Par suite les deux formules (3.4 et 3.5) sont identiques à une constante près

$$a_0 = \ln[\frac{p'}{1-p'}] + b_0$$

Remarques

- Le LOGIT $=C(X)$ est théoriquement définie entre $-\infty$ et $+\infty$.
- $0 < pi < 1$ représente une probabilité.
- La règle d'affectation peut être basée sur p comme suit
  - Si $\frac{pi}{1-pi} > 1$ alors $Y = +$.
  - Si $pi > 0.5$ alors $Y = +$.
- Et elle peut être basée sur $C(X)$
  - Si $C(X) > 0$ alors $Y = +$.

3.4 Estimation des paramètres par la méthode du maximum de vraisemblance

Pour estimer les paramètres de la régression logistique par la méthode du maximum de vraisemblance, il faut d'abord déterminer la loi de distribution de $P(Y/X)$.
Puisque Y est une variable binaire définie dans $\{+,-\}$ (ou $\{0,1\}$ pour simplifier) , pour un individu $\omega \in \Omega$ on modélise la probabilité à l'aide de la loi binomiale $\mathbf{B}(1, pi)$, avec

$$P(Y(\omega) = y(\omega)/X(\omega)) = (pi(\omega))^{y(\omega)} \times (1 - pi(\omega))^{(1-y(\omega))}$$

Cette modélisation est cohérente avec ce qui a été dit précédemment, en effet

— Si $y(\omega) = 1$, alors $P(Y(\omega) = 1/X(\omega)) = pi$

— Si $y(\omega) = 0$, alors $P(Y(\omega) = 0/X(\omega)) = 1 - pi$

Remarque
Si Y était d'une autre nature, on utiliserait d'autres modèles comme Poisson, Multinomial, ...

Définition 5. *La vraisemblance associée à un échantillon Ω s'écrit sous la forme*

$$L(Y, a) = \prod_{i=0}^{n} pi(\omega)^{Y(\omega)} \times (1 - pi(\omega))^{(1-Y(\omega))}$$

La méthode du maximum de vraisemblance consiste à produire les paramètres $a = (a_0, a_1, ..., a_p)$ de la régression logistique pour que les probabilités des réalisations observées soient aussi maximum.
Pour simplifier les calculs, on préfère travailler sur la log-vraisemblance, qui vaut

$$\begin{aligned} LogL(Y, a) &= \ln(\prod_{\omega}(pi(\omega))^{Y(\omega)} \times (1 - pi(\omega))^{(1-Y(\omega))}) \\ &= \sum_{\omega} Y(\omega)\ln(pi(\omega)) + (1 - Y(\omega))\ln(1 - pi(\omega)). \\ &= \sum_{\omega} Y(\omega)\ln(F(x_i a)) + (1 - Y(\omega))\ln(1 - F(x_i a)). \end{aligned} \tag{3.7}$$

Pour alléger l'écriture, nous écrirons pour la suite

$$LogL(Y, a) = \sum_{\omega} y_i \ln(F(x_i a)) + (1 - y_i)\ln(1 - F(x_i a)). \tag{3.8}$$

L'estimateur du maximum de vraisemblance des paramètres a est obtenu en maximisant soit la fonction de vraisemblance L soit la fonction de log-vraisemblance $LogL$.
En dérivant La log-vraisemblance par rapport aux éléments du vecteur a de dimension p, on obtient un vecteur noté $G(a)$ appelé vecteur du gradient.

$$G(a) = \frac{\partial LogL(Y, a)}{\partial a} \tag{3.9}$$

$$= \sum_{i=1}^{n} y_i \frac{f(x_i a)}{F(x_i a)} x_i^t + (y_i - 1)\frac{f(x_i a)}{1 - F(x_i a)} x_i^t. \tag{3.10}$$

où f est la fonction de densité associée à F et x_i^t est la transposée du vecteur x_i.
On obtient

$$G(a) = \sum_{i=1}^{n} \frac{(y_i - F(x_i a))f(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i^t. \tag{3.11}$$

Proposition 3.4.1. *[10] L'estimateur $\hat{a}$ du maximum de vraisemblance du vecteur de paramètres a dans un modèle dichotomique est défini par la résolution du système de p équations non linéaires en a*

$$\begin{aligned} \hat{a} &= \arg\max_{a} LogL(Y,a) \\ &\Leftrightarrow \frac{\partial LogL(Y,\hat{a})}{\partial \hat{a}} = \sum_{i=1}^{n} \frac{(y_i - F(x_i\hat{a}))f(x_i\hat{a})}{F(x_i\hat{a})(1-F(x_i\hat{a}))} x_i^t = G(\hat{a}) = 0. \end{aligned} \tag{3.12}$$

Proposition 3.4.2. *[10] Puisque $\hat{a}$ est un estimateur du maximum de vraisemblance, il en possède les propriétés suivantes*

- *Il est asymptotiquement sans biais.*
- *Il est de variance minimale.*
- *Il est asymptotiquement gaussien.*
  Ces éléments seront très importants pour l'inférence statistique (intervalle de confiance, test de signifcativité, etc.).

Remarque
le système défini par l'équation 3.12 est non linéaire, par suite, l'estimateur $\hat{a}$ ne peut être obtenu directement.
Plusieurs algorithmes d'optimisation numérique de la vraisemblance existent, les logiciels s'appuient souvent sur l'algorithme de Newton-Raphson [10].

3.5 Méthode du Maximum de Vraisemblance

3.5.1 Matrices Hessiennes et Matrices d'information de Fischer associées à la log-vraisemblance

Définition 6. *la matrice hessienne $H(a)$, associée à la log-vraisemblance d'un échantillon de taille n est définie par*

$$H(a) = \frac{\partial^2 LogL(Y,a)}{\partial a.\partial a^t}$$

Proposition 3.5.1. *La matrice hessienne $H(a)$ est de dimension $(p \times p)$ d'expression générale*

$$\begin{aligned} H_{(p,p)}(a) &= -\sum_{i=1}^{n} \left(\frac{y_i}{F(x_i a)^2} + \frac{1-y_i}{(1-F(x_i a))^2} \right) f(x_i a)^2 x_i^t x_i \\ &+ \sum_{i=1}^{n} \left(\frac{y_i - F(x_i a)}{F(x_i a)(1-F(x_i a))} \right) f'(x_i a) x_i^t x_i. \end{aligned} \tag{3.13}$$

avec $f'(.)$ désigne la dérivée de la fonction de densité $f(.)$ associée à $F(.)$.

Preuve

$$
\begin{aligned}
H(a) &= \frac{\partial^2 \log L(Y,a)}{\partial a \partial a^t} \\
&= \frac{\partial}{\partial a}\left(\frac{\partial \log L(Y,a)}{\partial a}\right) \\
&= \frac{\partial}{\partial a} G(a)^t \\
&= \frac{\partial}{\partial a} \sum_{i=1}^{n} \frac{(y_i - F(x_i a)) f(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i^t \\
&= \sum_{i=1}^{n} \frac{(F(x_i a))(1 - F(x_i a))}{(F(x_i a))^2 (1 - F(x_i a))^2} \frac{\partial (y_i - F(x_i a)) f(x_i a)}{\partial a} x_i \\
&- \sum_{i=1}^{n} \frac{(y_i - F(x_i a)) f(x_i a)}{(F(x_i a))^2 (1 - F(x_i a))^2} \frac{\partial (F(x_i a))(1 - F(x_i a))}{\partial a} \\
&= \sum_{i=1}^{n} \frac{-f^2(x_i a) + (y_i - F(x_i a)) f'(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i^t x_i \\
&- \sum_{i=1}^{n} \frac{(y_i - F(x_i a)) f(x_i a)}{(F(x_i a))^2 (1 - F(x_i a))^2} \times \left[(1 - F(x_i a)) f(x_i a) - F(x_i a) f(x_i a)\right] x_i^t x_i \\
&= -\sum_{i=1}^{n} \left[\frac{y_i}{F(x_i a)^2} + \frac{1 - y_i}{(1 - F(x_i a))^2}\right] f^2(x_i a) x_i^t x_i \\
&+ \sum_{i=1}^{n} \left[\frac{y_i - F(x_i a)}{F(x_i a)(1 - F(x_i a))^2}\right] f'(x_i a) x_i^t x_i .
\end{aligned}
$$

■

Remarque

Dans le cas des modèles logit, il n'y a pas d'expression simplifiée de la matrice hessienne, par contre, l'espérance de la matrice hessienne est plus simple à traiter.

Définition 7. *Pour un modèle dichotomique univarié, la matrice d'information de Fisher est définie par*

$$
I(a) = -\mathbb{E}\left(\frac{\partial^2 LogL}{\partial a . \partial a^t}\right) = \sum_{i=1}^{n} \frac{f_i^2(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i^t x_i
$$

Proposition 3.5.2. *L'information de fisher pour le modèle Logit 3.6, s'écrit sous la forme suivante*

$$
I(a) = \sum_{i=1}^{n} \frac{\exp(x_i a)}{(1 + \exp(x_i a))^2} x_i^t x_i . \tag{3.14}
$$

Preuve

d'aprés la définition on a

$$
\begin{aligned}
I(a) &= -\mathbb{E}\left[\frac{\partial^2 LogL}{\partial a . \partial a^t}\right] \\
&= \sum_{i=1}^{n} \frac{f_i^2(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i^t x_i
\end{aligned}
$$

Or

$$
f(x_i a) = F(x_i a)(1 - F(x_i a))
$$

$$\begin{aligned} F(x_i a)(1 - F(x_i a)) &= \left(\frac{\exp(x_i a)}{1 + \exp(x_i a)}\right)\left(1 - \frac{\exp(x_i a)}{1 + \exp(x_i a)}\right) \\ &= \left(\frac{\exp(x_i a)}{(1 + \exp(x_i a))^2}\right). \end{aligned}$$

et

$$\begin{aligned} \frac{f^2(x_i a)}{F(x_i a)(1 - F(x_i a))} &= \left(\frac{(1 + \exp(x_i a))^2}{\exp(x_i a)}\right)\left(\frac{(\exp(x_i a))^2}{((1 + \exp(x_i a))^2)^2}\right) \\ &= \frac{\exp(x_i a)}{(1 + \exp(x_i a))^2} \end{aligned}$$

d'où le resultat. ■

Proposition 3.5.3. *[10] Dans un modèle dichotomique univarié, la fonction de log-vraisemblance $logL(y, a)$ est strictement concave, ce qui garantit l'unicité du maximum de cette fonction.*

Preuve
L'equation (3.8) nous montre que si $logL$ est concave alors $logF$ est concave , donc :

$$\begin{aligned} \frac{\partial \log F(X)}{\partial X} &= \frac{1}{F(X)}\frac{\partial F(X)}{\partial X} \\ &= \frac{1 + \exp(X)}{\exp(X)}\frac{\exp(X)}{(1 + \exp(X))^2} \\ &= \frac{1}{1 + \exp(X)} \end{aligned}$$

et

$$\frac{\partial^2 \log F(X)}{\partial X^2} = \frac{\partial}{\partial X}\left(\frac{1}{1 + \exp(X)}\right) = \frac{-\exp(X)}{(1 + \exp(X))^2} < 0$$

$$\begin{aligned} \frac{\partial \log(1 - F(X))}{\partial X} &= -\frac{1}{1 - F(X)}\frac{\partial F(X)}{\partial X} \\ &= -\frac{1 + \exp(X)}{1}\frac{\exp(X)}{(1 + \exp(X))^2} \\ &= -\frac{\exp(X)}{(1 + \exp(X))} \\ &= -F(X) \end{aligned}$$

$$\frac{\partial^2 \log(1 - F(X))}{\partial X^2} = -\frac{\partial F(X)}{\partial X} = \frac{-\exp(X)}{(1 + \exp(X))^2} < 0$$

■

3.5.2 L'algorithme de Newton-Raphson

[10] L'algorithme de Newton-Raphson est l'une des méthodes numériques les plus utilisées pour optimiser la log-vraisemblance , il démarre avec une initialisation quelconque du vecteur de paramètre a , pour passer de l'étape (i) à l'étape $(i + 1)$, il se rapproche de la solution finale $\hat{a}$ en utilisant la formule suivante

$$\hat{a}_{i+1} = \hat{a}_i - \left(\frac{\partial^2 LogL(y, a)}{\partial a . \partial a^t}\right)^{-1}\Big|_{a=\hat{a}_i} \times \frac{\partial LogL(y, a)}{\partial a}\Big|_{a=\hat{a}_i} \tag{3.15}$$

ou encore

$$\hat{a}_{i+1} = \hat{a}_i - H^{-1}(\hat{a}_i) \times G(\hat{a}_i) \tag{3.16}$$

Les itérations sont arrêtés, lorsque la différence entre a_{i+1} et a_i est inférieure à un certain seuil fixé dans le programme.

Proposition 3.5.4. *La suite $(a_i)_i$ définie par la relation récurrence (3.16), converge vers l'estimateur du maximum de vraisemblance $\hat{a}$.*

Preuve
Posons $\tilde{a} = \lim_{i\to\infty} \hat{a}_i$ en utilisons (3.16), on trouve :

$$\tilde{a} = \tilde{a} - H^{-1}(\tilde{a}) \times G(\tilde{a}) \Leftrightarrow H^{-1}(\tilde{a}) \times G(\tilde{a}) = 0$$

comme $H(\tilde{a})$ est définie positive, alors $G(\tilde{a}) = 0$
Donc la suite $\hat{a}_i$ des estimateurs obtenus par l'algorithme de Newton Raphson, converge vers $\tilde{a}$ solution des équations des vraisemblance.

■

3.5.3 Loi asymptotique de l'estimateur du Maximum de Vraisemblance

On suppose que les $(X_i)_i$ sont des variables aléatoires continues ou des variables déterministes, pour garantir la convergence et la normalité asymptotique des estimateurs, deux approches sont retenues :

1. Si les variables explicatives sont aléatoires continues on suppose qu'elles sont indépendantes identiquement distribuées de même loi iid, en admettant l'existence de moments d'ordre suffisant.
2. Si les variables explicatives sont déterministes, on impose la condition suivante :

$$\exists m > 0, \exists M < \infty \quad \text{tq} \quad m < |X_i^k| < M, \forall k \in \mathbb{R}, \forall i = 1 \cdots n.$$

Proposition 3.5.5. *[10] Sous ces conditions (1) et (2) sur les variables explicatives, l'estimateur $\hat{a}$ du maximum de vraisemblance converge et suit asymptotiquement une loi normale de moyenne égale à la vraie valeur des paramètres a_0, et de matrice de variance covariance égale à l'inverse de la matrice d'information de Fisher $I(a_0)$ évaluée au point a_0,*

$$\sqrt{n}\,(\hat{a} - a_0) \xrightarrow[n\to\infty]{L} N(0, I(a_0)^{-1})$$

Preuve
D'aprés la proposition précédente l'estimateur du maximum de vraisemblance satisfait à la condition $G(\hat{a}) = 0$.
On considère le développement limité d'ordre 1 autour de la vraie valeur des paramètres a_0, on obtient :

$$G(\hat{a}) = G(a_0) + H(a_0)(\hat{a} - a_0) = 0$$

En multipliant cette quantité par $H^{-1}(a_0)$ on trouve :

$$(\hat{a} - a_0) = -H^{-1}(a_0)G(a_0)$$

Ou encore

$$\sqrt{n}(\hat{a} - a_0) = -\left[\frac{1}{n}H(a_0)\right]^{-1} [\sqrt{n}\bar{g}(a_0)]$$

avec $\bar{g}$ est le vecteur de taille $(p, 1)$ défini comme suit :

$$\bar{g}(a_0) = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} \frac{\partial \log L(y_i,a)}{\partial a_1} \\ \sum_{i=1}^{n} \frac{\partial \log L(y_i,a)}{\partial a_2} \\ \vdots \\ \sum_{i=1}^{n} \frac{\partial \log L(y_i,a)}{\partial a_p} \end{pmatrix}$$

Posons que les variables $Z_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log L(y_i,a)}{\partial a_j}$ sont iid, on a

$$E(Z_j) = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\partial \log L(y_i, a)}{\partial a_j}\right] = 0$$

car

$$\begin{aligned} E\left[\frac{\partial \log L(y_i, a)}{\partial a_j}\right] &= E\left[\sum_{i=1}^{n} \frac{(y_i - F(x_i a))f(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i\right] \\ &= \sum_{i=1}^{n} E\left[\frac{(y_i - F(x_i a))f(x_i a)}{F(x_i a)(1 - F(x_i a))} x_i\right] \\ &= 0 \end{aligned}$$

puisque

$$E(y_i) = p_i = F(x_i a)$$

D'ou

$$E(\bar{g}(a_0)) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

par le théorème central limite(4.2)

$$Z_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log L(y_i, a)}{\partial a_j} \hookrightarrow N(0, 1)$$

D'où

$$\bar{g}(a_0) \hookrightarrow N(0, I_d)$$

De plus par la loi des grands nombres(4.2)

$$\frac{H(a_0)}{n} \underset{n\to\infty}{\longrightarrow} E(H(a_0)) = I_f(a_0)$$

Où $I_f(a_0)$ est l'information de Fischer évaluée au point a_0.
On applique les deux théorèmes(4.2, 4.2) à la formule suivante :

$$\sqrt{n}(\hat{a} - a_0) = -\left[\frac{1}{n}H(a_0)\right]^{-1} [\sqrt{n}\bar{g}(a_0)]$$

$\sqrt{n}(\hat{a} - a_0)$ a une distribution normale de moyenne 0 et de matrice de covariance $-E(H(a_0))$. ■

Chapitre 4

Simulation

4.1 Consommation de l'énergie électrique de la Société des Ciments de Benisaf

Nous considérons les données de mesure de la consommation de l'énergie électrique de la Société des Ciments de BeniSaf (S.CI.BS) Wilaya de AIN TEMOUCHENT (c'est un client alimenté en électricité haute tension 60000 Volts par la société Gestionnaire Réseau de Transport d'électricité (G.R.T.E) filiale de SONELGAZ, sa source d'alimentation est depuis un Poste 220000/60000 Volts qui se trouve à AMIR ABELKADER sur la route National N35A).
Des données de puissance en mégua watt sont observés par un compteur d'électricité tout les 10 minute pendant la période de 11/02/2018 (00h00) jusqu'au 30/06/2018 (23H50), representé dans le graphe suivant (Fig. 4.1).
Notre but est de prédire la consommation d'éléctricité du dernier jour noté Y en fonction des variations de la consommation des 69 jours precédents , tel que chaque jour renvoie 144 valeurs , donc on a la matrice X des variables explicatives de dimension (144×69) , et le vecteur prédicteur Y de taille (144×1).

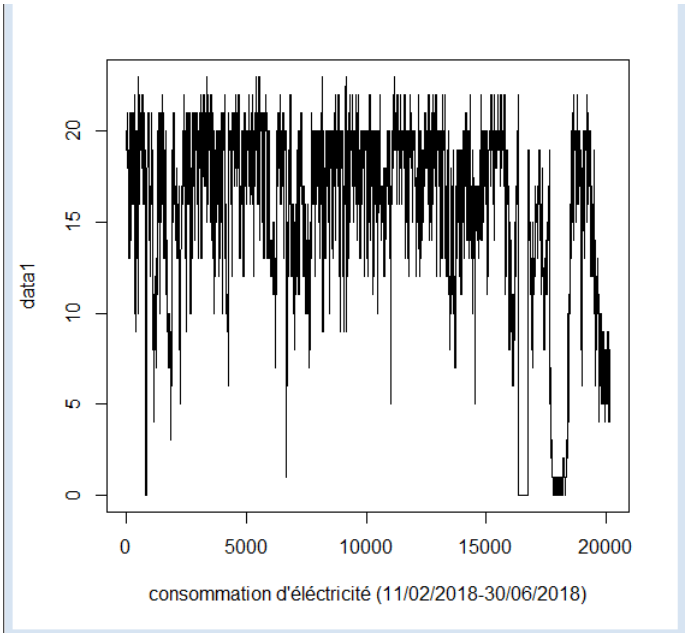

FIGURE 4.1 – consommation d'electricité (11.02.2018/30.06.2018)

4.1.1 Régression lineaire multiple

On pose le modèle lineaire correspond des données de notre exemple :

$$y_i = b_0 + \sum_{j=1}^{69} b_j x_{ij} + \varepsilon_i \qquad i = 1, \ldots, 144$$

En utilisant le logiciel R pour montrer quelque resultats importantes :
->RLM<-lm(Y ~ X)
-> summary(RLM)
Le tableau suivant montre un résumé sur le résidu .

Residuals :

Min	1Q	Median	3Q	Max
-1.09115	-0.27292	0.05197	0.29445	0.97324

Un deuxième tableau donne l'estimation des coeficients b_i dans la première colonne , et l'ecart type estimé dans la deuxième colonne , et dans les dernière colonne on trouve le test de nullité du coeficient et la p-valeur.

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.4321151	8.3312976	1.012	0.314788	
x[, 1]	-0.0706081	0.1033426	-0.683	0.496587	
x[, 2]	0.0710117	0.0976291	0.727	0.469299	
x[, 3]	-0.0988608	0.0871246	-1.135	0.260161	
x[, 4]	0.0755429	0.0613695	1.231	0.222239	
x[, 5]	-0.2002148	0.1053635	-1.900	0.061299	.
x[, 6]	-0.0255965	0.0397320	-0.644	0.521419	
x[, 7]	0.2163738	0.0964785	2.243	0.027910	*
x[, 8]	0.0572383	0.0650018	0.881	0.381405	
x[, 9]	0.1286130	0.0759971	1.692	0.094788	.
x[, 10]	0.0448309	0.0800325	0.560	0.577063	
x[, 11]	-0.0122981	0.0714733	-0.172	0.863856	
x[, 12]	0.0175979	0.0783123	0.225	0.822821	
x[, 13]	-0.0176917	0.0893810	-0.198	0.843638	
x[, 14]	0.0077284	0.0665257	0.116	0.907831	
x[, 15]	-0.1024394	0.0994695	-1.030	0.306431	
x[, 16]	-0.1124070	0.0698593	-1.609	0.111864	
x[, 17]	-0.1047947	0.0972647	-1.077	0.284793	
x[, 18]	-0.0335226	0.0808751	-0.414	0.679708	
x[, 19]	0.0862086	0.0709971	1.214	0.228511	
x[, 20]	0.0360582	0.0780901	0.462	0.645614	
x[, 21]	0.1033224	0.0881817	1.172	0.245077	
x[, 22]	0.1693490	0.0806883	2.099	0.039246	*
x[, 23]	-0.0366942	0.0966408	-0.380	0.705258	
x[, 24]	-0.0809948	0.0782270	-1.035	0.303863	
x[, 25]	0.0008533	0.0631533	0.014	0.989256	

x[, 26]	-0.0324564	0.0824934	-0.393	0.695124	
x[, 27]	0.2794792	0.0821192	3.403	0.001078	**
x[, 28]	0.0252413	0.0861346	0.293	0.770308	
x[, 29]	-0.0331397	0.0876560	-0.378	0.706464	
x[, 30]	-0.1114202	0.0914830	-1.218	0.227118	
x[, 31]	0.2221593	0.0821610	2.704	0.008496	**
x[, 32]	-0.0953911	0.0870025	-1.096	0.276451	
x[, 33]	-0.2279203	0.0954577	-2.388	0.019510	*
x[, 34]	0.0812383	0.0897848	0.905	0.368502	
x[, 35]	-0.0618791	0.1019094	-0.607	0.545578	
x[, 36]	-0.0542617	0.1035985	-0.524	0.602004	
x[, 37]	0.1076837	0.0724697	1.486	0.141550	
x[, 38]	-0.0651864	0.1034150	-0.630	0.530415	
x[, 39]	-0.1189453	0.0912823	-1.303	0.196598	
x[, 40]	0.0217087	0.0632510	0.343	0.732409	
x[, 41]	0.1805513	0.1196139	1.509	0.135442	
x[, 42]	0.1754639	0.1178837	1.488	0.140882	
x[, 43]	0.1313155	0.1298491	1.011	0.315172	
x[, 44]	0.0111560	0.0819913	0.136	0.892141	
x[, 45]	0.0599999	0.0855827	0.701	0.485455	
x[, 46]	-0.2763387	0.0817664	-3.380	0.001161	**
x[, 47]	-0.0022570	0.0401023	-0.056	0.955270	
x[, 48]	-0.0082349	0.0852094	-0.097	0.923271	
x[, 49]	-0.0364544	0.0842968	-0.432	0.666670	
x[, 50]	-0.1011499	0.0842635	-1.200	0.233813	
x[, 51]	-0.2131857	0.1106820	-1.926	0.057932	.
x[, 52]	0.2114610	0.1132587	1.867	0.065854	.
x[, 53]	0.1164271	0.0793917	1.466	0.146752	
x[, 54]	0.1795419	0.0705382	2.545	0.013001	*
x[, 55]	0.0913672	0.0660295	1.384	0.170600	
x[, 56]	-0.0187545	0.0758548	-0.247	0.805406	
x[, 57]	0.0062695	0.0611763	0.102	0.918651	
x[, 58]	-0.0189798	0.0747332	-0.254	0.800225	
x[, 59]	-0.0062871	0.0675997	-0.093	0.926151	
x[, 60]	0.0312791	0.0818045	0.382	0.703288	
x[, 61]	-0.1376184	0.0862836	-1.595	0.114985	
x[, 62]	-0.1455093	0.0788226	-1.846	0.068887	.
x[, 63]	0.0600944	0.0961488	0.625	0.533884	
x[, 64]	-0.0787030	0.0785447	-1.002	0.319601	
x[, 65]	0.1895216	0.1019930	1.858	0.067121	.
x[, 66]	0.0374558	0.0825656	0.454	0.651409	
x[, 67]	0.0947009	0.0758624	1.248	0.215848	
x[, 68]	-0.1434628	0.0748590	-1.916	0.059171	.
x[, 69]	0.2523113	0.0734578	3.435	0.000975	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error : 0.5633 on 74 degrees of freedom
Multiple R-squared : 0.9758, Adjusted R-squared : 0.9533
F-statistic : 43.32 on 69 and 74 DF, p-value : 2.2e-16

L’estimation de l’écart type nous donne $\hat{\sigma}$ =0.5633 avec 74 degré de liberté , et la qualité de l’ajustement du modèle caractérisé par R^2=0.9758 , aussi que la statistique ajusté par R_a^2=0.9533.

Ainsi que la racine de l'erreur quadratique moyenne RMSE=0.4038366

4.1.2 Ridge

On applique la méthode de la regréssion ridge sur les même données, en utilisant le package glmnet qui a définie pour α=0 (coefficient de ponderation), λ>0 (coefficient de pénalité) comme suit :
->library(Matrix)
->library(nlme)
->library(glmnet)
->ridge1<- glmnet(X,Y,alpha=0)
->plot(ridge1,xvar="lambda")
Le graphe suivant permèttant de juger de la dispertion des coefficients au regard de λ (4.2) , plus λ augmente, plus la norme des coefficients b_i diminue , tous les coefficients sont nul lorsque λ=λ_{max}.
On a les logarithmes des λ_i en abscisse de notre graphique , il varient de $\log(0,3678794) = -1$ à $\log(2980,958) = 8$.

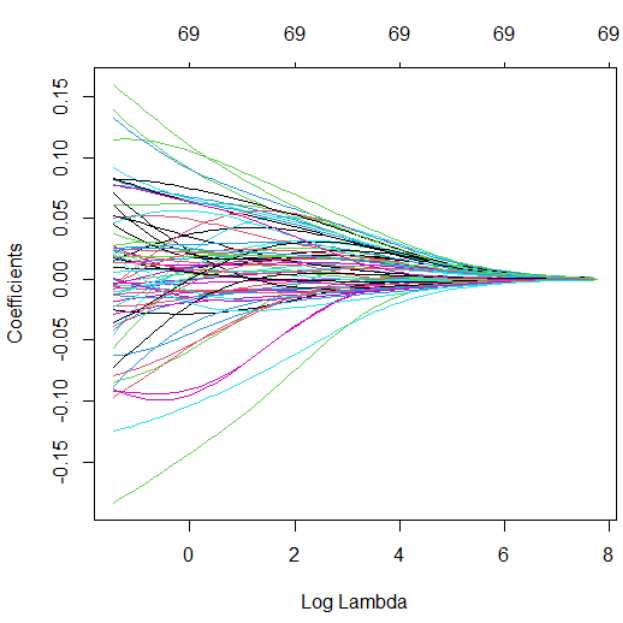

FIGURE 4.2 – Coeffcient Ridge

Le choix de λ optimal vient de la valeur de λ à partir de laquelle les coefficients commencent à se stabiliser, mais cette démarche ne s'assure pas de trouver la solution qui optimise les qualités prédictives du modèle, la solution est de passer par la validation croisé qui fait l'appel à la fonction cv.glmnet.
– >ridge2=cv.glmnet(X,Y,lambda=10^ seq(4,-1,-.1),alpha=0)
– >plot(ridge2)

Le graphe suivant permet de mêttre en relation les valeurs log(λ) avec le taux d'erreur moyen en validation croisée(les points rouges).
Un intervale de confiance est proposé, défini par ± ecart-type de l'erreur en validation croisée.

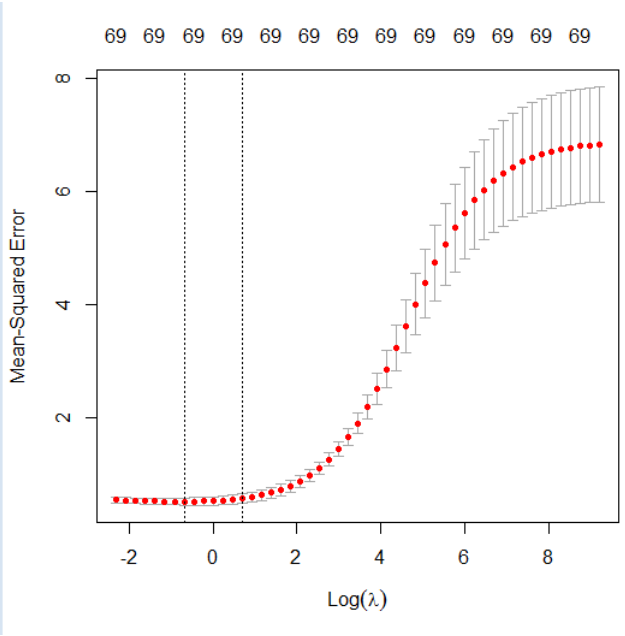

FIGURE 4.3 – Taux d'erreur en validation croisée par log(λ)

On peut visualiser quelques éléments importants :
l'erreur minimal par la commande :
– >print(min(ridge2$cvm))
– >0.524456.
Puis cherchons le λ optimal correspond à cette erreur :
– >print(ridge2$ lambda.min)
– >0.50111872
et son logarithme :
– >print(log(ridge2$ lambda.min))
– > −0.697755
Cette coordonnée est matérialisée par le premier trait pointillé (à gauche) dans le graphique (figure 2)(4.3) de la validation croisée.
On aperçoit un deuxième trait dans le graphique(figure 2)(4.3) correspond à la plus grand valeur de λ (noté $\lambda.1se$) tel que son erreur moyenne en validation croisée (points rouges) est inferieure à la borne haute de l'intervale de confiance de l'erreur optimale, on peut la visualiser par la commande :
– >print(ridge2$ lambda.1se)
– >print(log(ridge2$ lambda.1se))
– >0.6907755
finalement la valeur du racine de l'erreur quadratique moyenne en utilisant le package "nlme", et " Metrics" :
– >print(rmse(Y,Ychap))
– >0.6432905.

4.1.3 LASSO

Même démarche pour la regréssion lasso avec λ >0 et α=1 .
En uilisant la commande :
– >lasso1<- glmnet(X,Y,alpha=0)
– >plot(lasso1,xvar="lambda")
Le graphe suivant montre la dispertion des coefficients au regard de λ (4.2) , plus λ augmente, plus la norme des coefficients b_i diminue , tous les coefficients sont nul lorsque λ=λ_{max}.

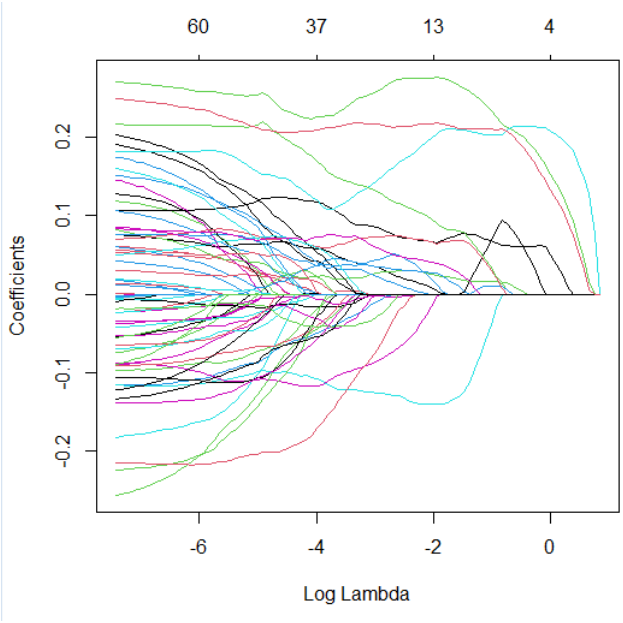

FIGURE 4.4 – LASSO,log(λ)

Pour obtenir le lambda optimale il suffit de passer à la validation croisée :
– >lam<-10^ seq(4,-1,-.1)
– >lasso2 <-cv.glmnet(X,Y,lambda=lam,alpha=1)
– >plot(lasso2)

Le graphe suivant exprime la relation entre les valeurs log(λ) et le taux d'erreur moyen en validation croisée(les points rouges).

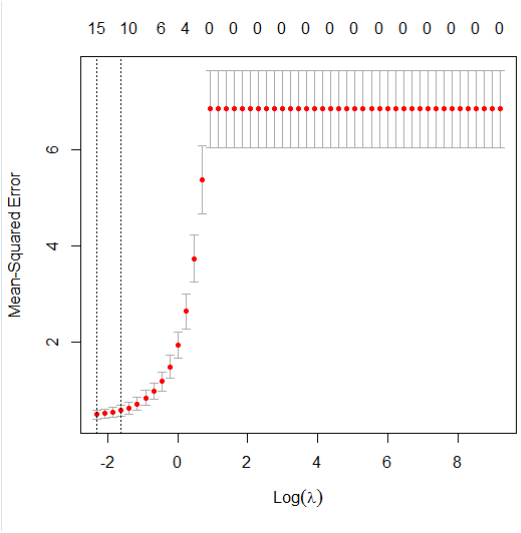

FIGURE 4.5 – Taux d'erreur en validation croisée par log(λ)

On peut visualiser quelque résultat importante :
L'erreur minimale :
− >print(min(lasso2$cvm))
0.4990347
λ optimale correspond à cette erreur :
print(lasso2$lambda.min)
0.1

Logarithme de λ optimale :
− >print(log(lasso2$lambda.min))
-2.302585

lambda .1se :
− >print(lasso2$lambda.1se)
0.1995262
Logarithme de lambda .1se :
print(log(lasso2$lambda.1se))
−1.61181
Ainsi que le RMSE= 0.697841.

4.1.4 Elastic-net

Pour la régression elastic-net, on va paramétrer par $\lambda > 0$ et $0<\alpha<1$.
Avec glmnet(), fixons la valeur de α .
– >elastic1<- glmnet(X,Y,alpha= 0.8)
– >plot(elastic1,xvar="lambda")
On dispose d'un graphique qui permet de caractériser la relation entre les coefficients b_i avec les λ_i.

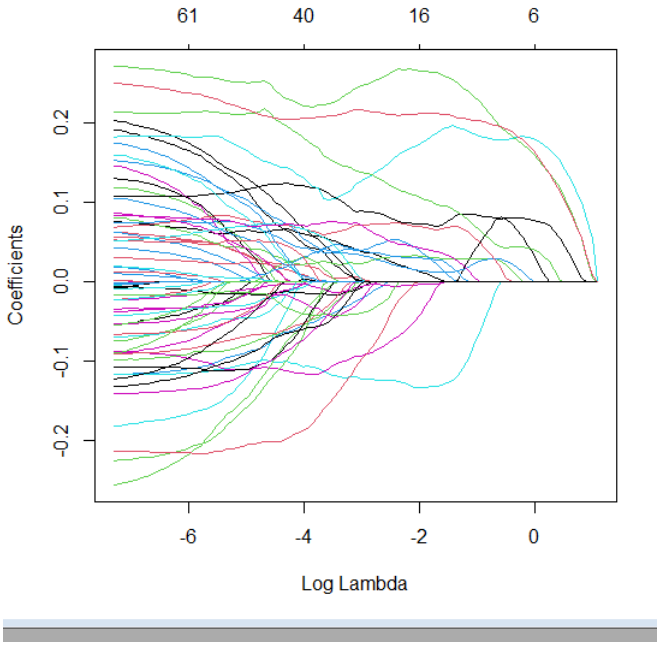

FIGURE 4.6 – Elastic-net,log(λ)

Passons par la validation croisée pour détecter la valeur optimale de λ pour $\alpha = 0.8$.
– >lam<-10^ seq(4,-1,-.1)
– >elastic2 <-cv.glmnet(X,Y,lambda=lam,alpha=0.8)
– >plot(elastic2)

On pourrais confirmer grace au graphe suivant qui met en relation log(λ) avec le taux d'erreur moyen en validation croisée :

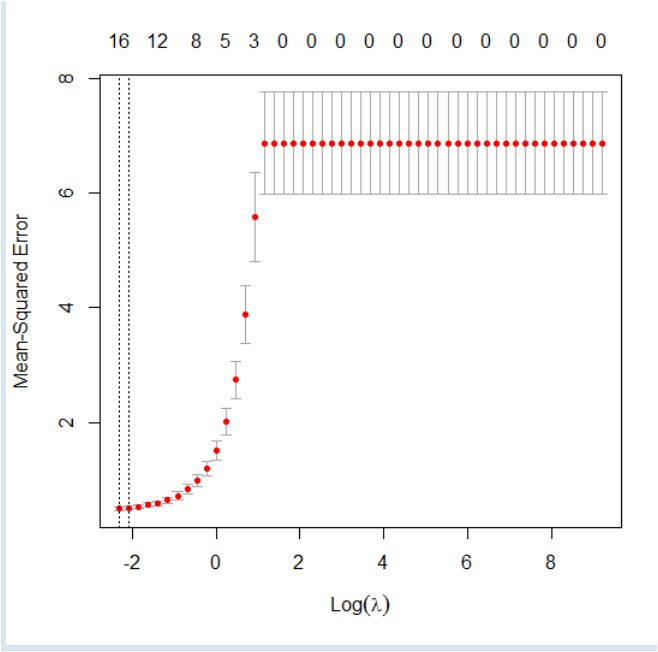

FIGURE 4.7 – Taux d'erreur en validation croisée par log(λ)

.

On peut montrer quelques résultats importantes :

— L'erreur minimale :
− > print(min(elastic2$cvm))
0.4962776

— λ optimale correspond à l'erreur minimale :
print(elastic2$lambda.min)
0.1

— Logarithme de λ optimale :
print(log(elastic2$lambda.min))
−2.302585

— lambda .1se :
− >print(elastic$lambda.1se)
0.1258925

— Logarithme de lambda .1se :
print(log(elastic2$lambda.1se))
−2.072327

— RMSE= 0.6308387.

Finalement on peut visualiser les coeficients estimés par les trois méthodes ridge, lasso et elastic-net.
− > r=coef(ridge2,s="lambda.min")
− > l=coef(lasso2,s="lambda.min")
− > e=coef(elastic2,s="lambda.min")
− > print(cbind(r,l,e))

	ridge	lasso	elastic-net
(Intercept)	9.8670327561	4.133108e+00	4.170266668
V1	0.0312491308	.	.
V2	0.0515686478	.	.
V3	0.0028710365	.	.
V4	0.0160357985	.	.
V5	0.0037256874	.	.
V6	-0.0042708318	.	.
V7	0.0285712973	.	.
V8	0.0160858816	.	.
V9	0.0084906153	.	.
V10	-0.0076617459	.	.
V11	-0.0035708127	.	.
V12	-0.0014731001	.	.
V13	0.0083934716	.	.
V14	-0.0086347460	.	.
V15	-0.0693270101	.	.
V16	-0.0221847683	-1.073305e-02	-0.010504857
V17	-0.1115629357	-1.249720e-01	-0.123003475
V18	-0.0127490601	.	.
V19	0.0408818077	3.517227e-02	0.034391236
V20	0.0234959740	.	.
V21	0.0617077571	2.767369e-02	0.028966510
V22	0.0690913564	2.387238e-02	0.028571535
V23	0.0007462237	.	.
V24	-0.0937469603	-8.632089e-02	-0.086475898
V25	-0.0280795144	.	.
V26	-0.0183454572	.	.
V27	0.1034646215	2.568261e-01	0.253175726
V28	-0.0149517469	.	.
V29	-0.0024691172	.	.
V30	-0.0213834785	-2.366410e-03	-0.003773376
V31	0.0365082257	.	.
V32	-0.0641474858	.	.
V33	-0.0132077059	.	.
V34	0.0709344666	4.476626e-02	0.046423534
V35	-0.0003782638	.	.
V36	-0.0991458786	-4.663993e-05	.
V37	0.0692867119	8.444861e-02	0.085748697
V38	0.0284558593	.	.
V39	0.0298887694	.	.
V40	0.0276944810	.	.

V41	0.0724321717	1.401509e-01	0.136876360
V42	0.0120760126	.	.
V43	0.0239043144	.	.
V44	-0.0102246380	.	.
V45	0.1114460723	7.305892e-02	0.073093032
V46	-0.0498090109	.	.
V47	0.0076624287	.	.
V48	0.0060491965	.	.
V49	-0.0122034502	.	.
V50	-0.0141842825	.	.
V51	-0.1571959584	-7.498274e-02	-0.075772818
V52	0.1017396818	1.323341e-01	0.131830909
V53	-0.0017829570	.	.
V54	0.0113388475	.	.
V55	0.0086920479	.	.
V56	0.0111996950	.	.
V57	0.0190638219	.	.
V58	-0.0534283901	-3.049751e-02	-0.030922859
V59	0.0559577791	3.606698e-02	0.036158349
V60	-0.0114599493	.	.
V61	-0.0355591591	.	.
V62	0.0161984997	.	.
V63	0.0265621428	.	.
V64	0.0259426534	.	.
V65	-0.0115553477	.	.
V66	0.0689121081	.	.
V67	0.0787633136	6.083991e-02	0.060492016
V68	-0.0687979221	.	.
V69	0.1251670432	2.162828e-01	0.213901964

Remarque : On remarque Instantanément du tableau qu'il y a des coefficients qui sont rétrecit vers zéro.
Or le principe du lasso et elastic-net est d'illiminé les variables nuisibles dans le modèle en estimant leurs coefficients par zéro.

4.2 Régression logistique

Conciderons l'exemple des données d'une maladie cardiaque comportant 20 observations et 3 variables prédictives pour illustrer la régression logistique binaire. Le but est de prédire la présence ou l'absence d'une maladie cardiaque à partir de :

— X_1 : l'Age d'individu (quantitative).

— X_2 : le Taux-max de la pression sanguine(quantitative).

— X_3 : l'occurence d'une Angine de poitrine(qualitative).

avec :

$$Y(\omega) = \begin{cases} 1, & \text{si la présence de la maladie.} \\ 0, & \text{si le cas contraire.} \end{cases}$$

Age	Taux-max	Angine	Coeur
50	126	1	presence
49	126	0	presence
46	144	0	presence
49	139	0	presence
62	154	1	presence
35	156	1	presence
67	160	0	absence
65	140	0	absence
47	143	0	absence
58	165	0	absence
57	163	1	absence
59	145	0	absence
44	175	0	absence
41	153	0	absence
54	152	0	absence
52	169	0	absence
57	168	1	absence
50	158	0	absence
44	170	0	absence
49	171	0	absence

TABLE 4.1 – Données de Coeur

La modélisation sous R nous donne les résulats suivants :

```
− >getwd()
− > maladie <- read.table("maladie.txt",header=TRUE)
− >modele<-glm(Y ~ Age+Taux.max+Angine,data=maladie,family="binomial")
− > print(summary(modele))
```

Call :

glm(formula = $Y \sim Age + Taux.max + Angine$, family = "binomial", data = maladie)

Deviance Residuals :

Min	1Q	Median	3Q	Max
-1.9773	-0.5437	-0.3876	0.5093	1.7577

Coefficients :

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	14.49379	7.95464	1.822	0.0684	.
Age	-0.12563	0.09380	-1.339	0.1805	
Taux.max	-0.06356	0.04045	-1.572	0.1161	
Angine	1.77901	1.50449	1.182	0.2370	

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance : 24.435 on 19 degrees of freedom
Residual deviance : 16.618 on 16 degrees of freedom
AIC : 24.618
Number of Fisher Scoring iterations : 5

On peut également visualiser les coefficients estimés sous forme d’un vecteur $a = (a_0, a_1, a_2, a_3)$.
− >a=coef(modele)
− >a
(Intercept) Age Taux.max Angine
14.4937905 -0.1256341 -0.0635603 1.7790129
D’où le modèle logistique :

$$C(X) = 14.4937905 - 0.1256341X_1 - 0.0635603X_2 + 1.7790129X_3$$

Calculons les quantités $C(X)$ et p_i pour chaque individu.

```
> X<-matrix(nrow=20,ncol=4)
> for(iin1:20){
  +X[i,1]<-1
+ X[i,2]<-maladie$Age[i]
+ X[i,3]<-maladie$Taux.max[i]
+ X[i,4]<-maladie$Angine[i]
+ }
> C<-a%*%t(X)
> pi<-matrix(nrow=20,ncol=1)
> for(iin1:20)
+ pi[i,]<-exp(C[,i])/(1+exp(C[,i]))
> pi
> s=cbind(t(C),pi)
> s.
```

Le tableau suivant résume les résultats obtenus sur $C(X)$, p_i et la prédiction $\hat{Y}$ de Y en passant par la règle de décision suivante :

$$\text{si } C(X) > 0 \text{ alors } Y = + \text{ (présence).}$$

ou bien

$$\text{si } p_i > 0.5 \text{ alors } Y = +.$$

Coeur	C(X)	pi	prédiction
présence	1.9824993	0.87894733	présence
présence	0.3291205	0.58154537	présence
présence	-0.4380624	0.39220275	absence
présence	-0.4971633	0.37820752	absence
présence	-1.3047987	0.21335852	absence
présence	1.9602024	0.87655486	présence
absence	-4.0933440	0.01640958	absence
absence	-2.5708698	0.07103688	absence
absence	-0.5001363	0.37750865	absence
absence	-3.2804383	0.03624840	absence
absence	1.8022236	0.85841939	présence
absence	-2.1348665	0.10575388	absence
absence	-2.1571633	0.10366373	absence
absence	-0.3819344	0.40566043	absence
absence	-1.9516179	0.12437705	absence
absence	-2.7808746	0.05836647	absence
absence	-1.5664721	0.17271990	absence
absence	-1.8304431	0.13818549	absence
absence	-1.8393618	0.13712678	absence
absence	-2.5310928	0.07370700	absence

Il y a plusieurs indicateurs qui nous permet d'evaluer l'éfficacité de notre modèle éstimé comme le Taux d'erreur et le Taux de succés.
Pour ce la on introduit la matrice de confusion définie comme suit :

$Y \times \hat{Y}$	$\hat{+}$	$\hat{-}$	total
+	a	b	$a + b$
-	c	d	$c + d$
total	$a + c$	$b + d$	$n = a + b + c + d$

avec

— a :les individus qui ont été classés positifs et qui le sont réellement.

— b : classées négatifs alors qu'ils sont positifs.

— c : les observations qui sont negatifs et classés positfs.

— d : les individus qui ont été classés négatifs et qui le sont réellement.

— Taux d'erreur : nombre de mauvais classement rapporté à l'effectif total, qui est défini par le rapport ;

$$\tau = \frac{b + c}{n}$$

— Taux de succés : la probabilité de bon classement du modèle ;

$$\vartheta = 1 - \tau = \frac{a + d}{n}$$

dans notre exemple ona :
$a = 3$, $b = 3$, $c = 1$, $d = 13$, par suite :

$$\tau = \frac{3 + 1}{20} = 0.2$$

et

$$\vartheta = \frac{3 + 13}{20} = 0.8$$

le taux d'erreur a eu une petite valeur donc on peut dire que notre modèle estimé est un bon modèle .

Annexe

1)**Theoreme central limite :** [12]
Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes ayant la même distribution avec $\mathbb{E}(X_i) = \mu_i$, et $V(X_i) = \hat{\sigma}_i^2$ pour $i = 1, \ldots, n$.
Alors la variable aléatoire

$$Z = \frac{\sum_{i=1}^{n} (X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n} \hat{\sigma}_i^2}}$$

suit approximativement une loi normale $N(0, 1)$ si n est grand.
2)**Theoreme (Loi Forte des grands nombres) :** [11]
Soit $(X_n)_{n>0}$ une suite de variables alétaoires indépendentes, identiquement distribuées. Posons

$$S_n = X_1 + \ldots + X_n$$

et supposons que $\mathbb{E}[|X_1|] < \infty$. Alors $\frac{S_n}{n}$ converge presque surement vers $\mathbb{E}[X_1]$.

Conclusion

L'apprentissage automatique (machine Learning), est un élément principal quand il s'agit de l'intelligence artificielle. L'apprentissage automatique constitue une grande avancée si vous voulez créer une intelligence artificielle ou tentez simplement d'obtenir un aperçu de toutes les données que vous avez collectées.
Dans ce mémoire nous avons présenté quelques algorithmes d'apprentissage supervisé (où on connait déjà les réponses qu'on attend d'elle) :

— nous avons rappelé l'essentiel des différentes méthodes de régression : linéaire multiple, RIDGE, LASSO, PCR et Elastic Net : où l'on utilise ces estimations de régression pour expliquer la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes.

— Régression logistique : les prévisions de la régression logistique sont des valeurs discrètes (vrai ou faux par exemple).
Nous avons résumé l'essentiel des résultats et travaux sur la régression logistique binaire, en particulier, les travaux de Hurlin [10],
nous avons étudié le modèle dichotomique simple "logit", nous avons présenté le modèle principal, puis nous nous sommes intéressés à l'estimation des paramètres de ce modèle par la méthode du maximum de vraisemblance, en utilisant la méthode d'optimisation de NEWTON RAPHSON pour optimiser la log-vraisemblance et nous avons étudié la loi asymptotique de l'estimateur du M V, sous certaines conditions, l'estimateur du M.V. est convergent et suit asymptotiquement une loi normale.

des applications réelles ont été considérées " La prévision de la consommation de l'énergie éléctrique de la Société des Ciments de BeniSaf" pour le cas d'une variable dépendante continue et "un exemple des données d'une maladie cardiaque" pour le cas d'une variable dépendante discrète, avec des erreurs relatives de prévision très faibles.
Nous souhaitons étudier d'autres méthodes similaire d'apprentissage supervisé, notamment aux "arbres de classification et de régression" et "Réseau de neurones artificiels".

Bibliographie

[1] A. Allam. Cours de régression lineaire. M1.

[2] M. Bardos, Analyse discriminante - Application au risque et scoring financier, Chapitre 3, "Discrimination logistique", pages 61-79, Dunod, 2001.

[3] B. Bercu and D. Chafaï. Modélisation stochastique et simulation. Dunod, Paris, 2007.

[4] J. BERKSON, 1950,“Are There Two Regressions ?”, Journal of the American Statistical Association, 45 (250) : 164–180.

[5] J. BERKSON, 1944, “ Application of the Logistic Function to Bio–Assay”, Journal of the American Statistical Association , 39 (227 : 357)–65.

[6] Cornillon, "Régression avec R", P-A. Cornillon, E. Matzner-Løber.

[7] ESL, "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.

[8] Giraud, "Introduction to high-dimensional statistics", C. Giraud.

[9] Hastie, "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.

[10] C. Hurlin, Modèles Dichotomiques Univariés Modèles Probit, Logit et Semi-Paramétriques, Polycopié de Cours ,Université d’Orléans 2003.

[11] Oscar Sheyin, "On the Law of Large Numbers" , page 5.

[12] Pierre-Simon Laplace, "Mémoire sur les approximations des formules qui sont fonctions de trés-grands nombres, et sur leur application aux probabilités" , page 353-415.

[13] Pierre-André Cornillon, Eric Matzner-Lober.*"Régression. Théorie et application" (Springer 2007).*

[14] P.F. Verhulst, 1838, "Notice sur la Loi que la Population Poursuit dans son Accroissement", Correspondance mathématique et physique, 10, 113–121.

[15] Wang, 1., Zhu, J. et Zou, H. (2006). The doubly regularized support vector machine. Statistica Sinica, 589-615.

[16] P.F. Verhulst, 1845, "Recherches Mathématiques sur la Loi d’Accroissement de la Population", Nouveaux Mémoires de l’Académie Royale des Sciences et Belles-Lettres de Bruxelles, 18, 1–42.

[17] P.F. Verhulst,1847, "Deuxième Mémoire sur la loi d’Accroissement de la Population" , Mémoires de l’Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique, 20, 1-32.

[18] Tibshirani, R. (1996) . Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267-288.

[19] Tseng, P. et Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming, 117(1-2), 387-423.

[20] Talon, C., Dautreme, E., Remy, E., Dirat, Y.,Dinse, C.ANALYSE DE DIFFÉRENTS ALGORITHMES DE CLASSIFICATION PAR APPRENTISSAGE AUTOMATIQUE SUR UN CAS D'USAGE DU DOMAINE NUCLÉAIRE, 21e Congrès de Maitrise des Risques et Sûreté de Fonctionnement $\lambda\mu$21 Reims 16-18 octobre 2018.

[21] Cédric A.(2016) . Le modèle Logit Théorie et application, Institut National de la Statistique et des Etudes Economiques.