République Algérienne Démocratique et Populaire Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A. Mira de Béjaïa

Faculté des Sciences Exactes
Département de Mathématiques

Mémoire de Fin de Cycle

en vue d'obtention du diplôme de Master en Mathématiques

Option : Probabilités Statistique et Applications

Théme

Introduction au LASSO et aux méthodes de régularisation

Réalisé par :

Brahmi Amira

Soutenu le 02/Juillet/2024, Devant le jury composé de :

Présidente	Mme. TIMERIDJINE	Pr - Université de Béjaïa.
Examinatrice	Mme. TABTI	M.C.B - Université de Béjaïa.
Promoteur	M. MAOUCHE	M.C.A - Université de Béjaïa.

Promotion : 2023/2024

Dédicace

Avec une immense gratitude, je dédie humblement ce modeste travail à ceux que nulle expression ne saurait suffire à témoigner pleinement de mon affection profonde.

À la femme, qui a souffert sans me laisser souffrir, qui m'a soutenu et encouragé durant ces années d'étude : mon adorable mère.

À l'homme, mon précieux offre du Dieu, ma réussite et tout mon respect : mon cher père.

À mes frères Ala-eddine, Mohamed et Oussama, qui ont partagé avec moi tous les moments d'émotion lors de la réalisation de ce mémoire. Ils m'ont chaleureusement supporté et encouragé tout au long de mon parcours.

À mes grand-mères, dont les prières et les bénédictions m'ont accompagné à chaque étape.

À mes chères amies Lysa, Cylina, Thizri et Lynda.

À tous mes oncles, tantes, cousins et cousines.

À tous mes enseignants tout au long de mes études.

À tous mes camarades de promotion, avec qui j'ai partagé les défis stimulants et les succès gratifiants de cette riche expérience universitaire.

B.Amira

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude envers Dieu le Tout-Puissant qui m'a accordé la santé, la persévérance, l'intelligence, la force et l'énergie nécessaires pour mener à bien ce travail et l'ensemble de mon cursus scolaire.

Mes sincères remerciements vont à mon encadrant **Mr. MAOUCHE**, pour son accompagnement précieux, pour avoir accepté de diriger ce mémoire et pour l'aide inestimable qu'il m'a prodiguée tout au long de cette expérience.

J'adresse également mes vifs remerciements aux membres du jury **Mme .TABTI** et **Mme .TIMERIDJINE** qui me font l'honneur d'évaluer mon travail.

Mes remerciements s'étendent également à tous mes enseignants durant les années d'études.

Enfin, j'exprime ma gratitude à toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce modeste travail.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Dans de nombreux domaines tels que l'économie, la finance, la biologie et les sciences sociales, les chercheurs et les analystes sont souvent confrontés à des ensembles de données de grande dimension comportant un grand nombre de variables explicatives. L'objectif est alors de modéliser la relation entre une variable réponse et cet ensemble de variables explicatives. La régression linéaire multiple est un outil statistique puissant utilisé pour atteindre cet objectif. Cependant, lorsque le nombre de prédicteurs est élevé par rapport à la taille de l'échantillon, les modèles de régression linéaire classiques peuvent souffrir de plusieurs problèmes, notamment le sur-apprentissage et la multicolinéarité. Pour faire face à ces défis, les méthodes de régularisation ont été développées. Ces techniques consistent à ajouter une pénalité aux coefficients du modèle, afin de réduire leur complexité et d'améliorer leur capacité de généralisation. L'une des premières méthodes de régularisation, la régression Ridge, a été introduite par **Hoerl** et **Kennard en 1970**[25]. Elle ajoute une pénalité basée sur la norme L2 des coefficients, ce qui a pour effet de réduire leur amplitude globale. Bien que la régression Ridge soit utile, elle garde toutes les variables dans le modèle. Parfois, on veut choisir seulement les variables les plus importantes, surtout quand il y en a beaucoup. Pour répondre à ce besoin de sélection de variables, **Robert Tibshirani** a introduit en **1996** le Lasso (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator)[49]. Cette méthode de régularisation L1, connue sous le nom de Lasso, permet à la fois la sélection de variables et l'estimation des coefficients. Elle ajoute une pénalité

L1, basée sur la valeur absolue des coefficients, à la fonction de perte des moindres carrés ordinaires. Le Lasso produit ainsi des modèles parcimonieux, mettant automatiquement à zéro certains coefficients, ce qui simplifie l'interprétation et réduit le surajustement. Cette méthode statistique a évolué progressivement au fil des années. Tout a commencé en **1995** quand **Breiman** a proposé une façon de simplifier les modèles statistiques "garrot"[4]. Un an plus tard, en 1**996**, **Tibshirani** a amélioré cette idée en créant une technique qui choisit automatiquement les informations les plus importantes. Cette méthode est devenue très populaire chez les chercheurs. En **2005**, **Zou et Hastie** ont créé une version plus flexible[56], et en **2006**, **Yuan et Lin** ont développé une variante qui peut traiter des groupes d'informations[55]. Depuis lors, cette méthode a continué à évoluer et s'est imposée comme un outil essentiel dans divers domaines de la recherche et de l'analyse de données.

En apprentissage statistique, le Lasso et les autres méthodes de régularisation sont considérés comme des approches avancées d'apprentissage supervisé, capables de gérer efficacement la complexité des modèles et d'améliorer leur généralisation. Ces méthodes offrent un compromis entre le biais et la variance, permettant d'obtenir des modèles plus performants .

Dans ce contexte, notre étude cherche à répondre à la question suivante : Dans quelle mesure le Lasso et les méthodes de régularisation L1 améliorent-ils la performance prédictive et l'interprétabilité des modèles de régression en haute dimension, et comment leur utilisation influence-t-elle la sélection de variables et la stabilité des estimations dans divers domaines d'application ?

L'objectif de ce travail est d'introduire et d'analyser le Lasso ainsi que les méthodes de régularisation L1 dans le cadre des techniques de régression statistique.

Ce mémoire est constitué d'une introduction générale, de trois chapitres et d'une conclusion.

Le premier chapitre est consacré à la présentation du cadre général des régressions

linéaires multiples et Ridge ainsi que les éléments théoriques utilisés pour l'élaboration de ces méthodes.

Le deuxième chapitre vise à fournir une compréhension approfondie de la régression Lasso, ses fondements théoriques, ses propriétés et ses applications pratiques dans le contexte de la régression régularisée et de la sélection de variables.

Le dernier chapitre est réservé à une étude comparative entre la régression linéaire multiple, Ridge et Lasso, que nous avons traitées théoriquement dans les chapitres 1 et 2. Nous cherchons à analyser les résultats et à évaluer les performances de ces méthodes à travers deux exemples distincts : un ensemble de données sur la consommation d'essence des véhicules et les données immobilières de Boston.

On mentionne que tous les travaux, présentés dans ce mémoire, sont traités à l'aide du logiciel R, version 4.2.3 (voir Ihaka, R. et Gentleman, R[34]), qui est présenté dans l'annexe A.

Chapitre 1

Généralités sur les régressions linéaires multiple, Ridge

1.1 Introduction

La régression linéaire constitue l'une des méthodes de modélisation statistique, largement utilisée dans divers domaines [2, 21]. Elle vise à établir une relation linéaire entre une variable dépendante et une ou plusieurs variables indépendantes, permettant ainsi de prédire ou d'expliquer le comportement de la variable dépendante en fonction des variables indépendantes[30].

Dans ce chapitre, nous explorerons les généralités de la régression linéaire, en mettant particulièrement l'accent sur deux techniques importantes : la régression linéaire multiple et la régression linéaire ridge. Cette dernière étend le cadre de la régression linéaire classique en introduisant des ajustements qui améliorent la stabilité des estimations et la performance prédictive du modèle.

1.2 Régression linéaire multiple

La régression linéaire multiple permet de modéliser la relation entre une variable à expliquer (dépendante) et plusieurs variables explicatives (indépendantes). Est un modèle explicatif qui quantifie l'influence de chaque variable indépendante sur la variable dépendante[46]. La variable dépendante doit être continue, tandis que les variables indépendantes peuvent être continues, discrètes ou catégorielles[19]. L'objectif est d'estimer les coefficients de régression associés à chaque variable explicative, afin de prédire la valeur de la variable dépendante en fonction des valeurs des variables indépendantes .

1.2.1 Modélisation

Le modèle de régression linéaire multiple (noté par RLM) est une généralisation du modèle de régression simple lorsque les variables explicatives sont en nombre quelconque. Nous supposons donc que les données collectées suivent le modèle suivant [7, 13] :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \qquad \forall i \in \{1.....n\}, \tag{1.1}$$

où $\beta_0, \beta_1, \beta_2..., \beta_p$ sont les paramètres inconnus du modèle (à estimer) .

On remarque dans ce modèle de RLM que :

- $i = 1...n$ correspond à la i-ème observation.

- $(x_{i,1}, ..., x_{ip})$ est une réalisation du vecteur aléatoire réel $(X_1, ..., X_p)$.

- y_i est une réalisation de Y.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon_i \tag{1.2}$$

- ε_i est l'erreur aléatoire due à la régression .

Notation matricielle

Ce modèle s'écrit sous la forme matricielle[6, 9] :

$$Y = X\beta + \varepsilon. \tag{1.3}$$

où :

$$Y=\begin{pmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ . \\ . \\ . \\ \mathrm{y}_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & ... & \mathrm{x}_{1p} \\ 1 & x_{21} & ... & \mathrm{x}_{2p} \\ . & . & ... & . \\ . & . & & . \\ . & . & ... & . \\ . & . & & . \\ 1 & x_{n1} & ... & \mathrm{x}_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Y : est le vecteur à expliquer de taille $n \times 1$.

X : est la matrice, de taille $n \times (p+1)$, qui contient l'ensemble des observations des variables explicatives , avec une première colonne formée par la valeur 1 indiquant que l'on intègre la constante β_0 dans l'équation où p étant le nombre de variables explicatives réelles.

β : vecteur des paramètres inconnus à estimer.

ε : vecteur d'erreurs de longueur n.

1.2.2 Estimation des paramètres

Une étape cruciale pour l'explication de données avec un modèle de régression paramétrique est l'estimation de ses paramètres. En générale, l'estimation des paramètres d'un modèle statistique se fait par la méthode de vraisemblance, des moindres carrés, des moments ou encore des techniques bayésiennes. Dans le cas de régression linéaire,

on utilise la méthode des moindres carrés. L'idée consiste à minimiser une distance entre le modèle supposé et les observations [37].

Définition 1.2.1 *(Estimateur des MCO)[18] :*

L'estimateur des moindres carrés ordinaires (MCO) $\hat{\beta}$ est défini comme étant l'estimateur qui minimise la fonction objective :

$$S(\beta) = (Y - X\beta)^t(Y - X\beta).$$

L'estimateur peut également s'écrire sous la forme suivante :

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta \in R^P} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \beta_i x_{ij})^2. \\ &= \arg\min_{\beta \in R^p} (Y - X\beta)^t(Y - X\beta). \\ &= \arg\min_{\beta \in R^p} ||Y - X\beta||^2.\end{aligned}$$

Proposition 1.2.1 *(Expression de l'estimateur des MCO)[8].*

L'estimateur par la méthode des moindres carrés ordinaires (MCO) $\hat{\beta}$ de β dans le modèle de régression linéaire multiple est :

$$\hat{\beta}^{(MCO)} = (X^tX)^{-1}X^tY. \tag{1.4}$$

Démonstration

Il suffit de développer la fonction objective et ensuite trouver le minimum de la fonction :

$$\begin{aligned} S(\beta) &= (Y - X\beta)^t(Y - X\beta) \\ &= (Y^t - \beta^t X^t)(Y - X\beta) \\ &= Y^tY - Y^tX\beta - \beta^tX^tY + \beta^tX^tX\beta \\ &= Y^tY - 2\beta^tX^tY + X^t\beta^tX\beta \ . \end{aligned}$$

car $Y^tX\beta$ est un scalaire, il est égal à sa transposée.

Pour déterminer le minimum de $S(\beta)$, on réalise la dérivation matricielle :

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^tY + 2X^tX\beta$$

$$\begin{aligned} \frac{\partial S(\beta)}{\partial \beta} = 0 &\Leftrightarrow -2X^tY + 2X^tX\beta = 0 \\ &\Leftrightarrow X^tX\beta = X^tY \end{aligned}$$

Donc,

$$\hat{\beta}^{(MCO)} = (X^tX)^{-1}X^tY.$$

1.2.3 Hypothèses relatives au modèle de la RLM

L'inférence statistique relative à la régression repose principalement sur les hypothèses liées au terme d'erreur ε qui résume les informations absentes du modèle.

Ces hypothèses, connues sous le nom d'hypothèses de Gauss-Markov dans le contexte de la régression linéaire, sont :

– H_1 : Les erreurs sont centrées .

$$E(\varepsilon_i) = 0, \qquad \forall i = \{1, ..., n\}.$$

– H_2 : La variance des erreurs est constante, on parle d'homoscédasticité [5] :

$$V(\varepsilon_i) = \sigma^2 < \infty \ , \qquad \forall i = \{1, ..., n\}.$$

– H_3 : Les erreurs relative à deux terme aléatoires ne sont pas corrélés, on dit n'y a pas de corrélation sérielle :

$$COV(\varepsilon_i, \varepsilon_j) = 0, \quad \forall \ i \neq j.$$

1.2.4 Propriétés des estimateurs MCO :

Proposition 1.2.2 *(Biais et matrice de covariance)[42, 48].*

L'estimateur $\hat{\beta}$ des moindres carrés est sans biais de β, c'est dire

$$E(\hat{\beta}) = \beta$$

et sa matrice de variance covariance est donnée par

$$VARCOV(\hat{\beta}) = \sigma^2 (X^t X)^{-1}.$$

Remarque 1.2.1 *La matrice de variance covariance des coeffcients de dimension $(p+1; p+1)$, est donneé par :*

$$VARCOV(\hat{\beta}) = \begin{pmatrix} V(\hat{\beta}_0) & COV(\hat{\beta}_0, \hat{\beta}_1) & ... & COV(\hat{\beta}_0, \hat{\beta}_p) \\ . & V(\hat{\beta}_1) & & COV(\hat{\beta}_1, \hat{\beta}_p) \\ . & . & & . \\ . & . & ... & . \\ . & . & & . \\ . & . & & V(\hat{\beta}_p) \end{pmatrix}$$

Cette matrice est symétrique.

Démonstration

Montrons que $E(\hat{\beta}) = \beta$,on a :

$$\begin{aligned} E(\hat{\beta}) &= E((X^tX)^{-1}X^tY) \\ &= (X^tX)^{-1}X^tE(Y) \\ &= (X^tX)^{-1}X^tE(X\beta + \varepsilon) \end{aligned}$$

et puisque $E(\varepsilon) = 0$, il vient

$$E(\hat{\beta}) = (X^tX)^{-1}X^tX\beta = \beta.$$

L'estimateur $\hat{\beta}$ est donc sans biais.

La variance est :

$$\begin{aligned} V(\hat{\beta}) &= V((X^tX)^{-1}X^tY) \\ &= (X^tX)^{-1}X^tV(Y)X(X^tX)^{-1} \end{aligned}$$

où $V(Y) = V(X\beta + \varepsilon)$ et $V(\varepsilon) = \sigma^2 I$, donc

$$\begin{aligned} V(\hat{\beta}) &= \sigma^2 (X^t X)^{-1} X^t X (X^t X)^{-1} \\ &= \sigma^2 (X^t X)^{-1}. \end{aligned}$$

L'intérêt de l'estimateur **MCO** est que parmi tous les estimateurs linéaires et sans-biais, il est celui qui a la plus petite variance. De plus, cet estimateur est simple à calculer. Pour ces raisons, c'est l'estimateur le plus utilisé en régression linéaire multiple. Par contre, bien qu'on puisse calculer $\hat{\beta}$, on n'a pas directement accès à sa variance $V(\hat{\beta})$ puisqu'elle dépend de σ^2. Il est alors important de comprendre comment estimer σ^2 afin de pouvoir estimer la variance $V(\hat{\beta})$.

1.2.5 Les résidus et la variance de l'estimateur MCO

Pour estimer la variance de $\hat{\beta}$, il est nécessaire de définir le concept de résidu. Le résidu d'une observation donnée correspond à la différence entre sa valeur observée et sa valeur prédite par le modèle. En utilisant les résidus, nous pouvons construire un estimateur sans biais pour la variance σ^2 .

Définition 1.2.2 *(Résidus)[14].*

Soit $\hat{\varepsilon} = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)^t$ le vecteur de dimension n des résidus défnie par :

$$\hat{\varepsilon} = Y - \hat{Y}$$

où $\hat{Y} = X\hat{\beta}$ est le vecteur des valeurs prédites étant donné les variables explicatives observées. En développant l'équation, on obtient

$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - X(X^t X)^{-1} X^t Y = Y - HY = (I_n - H)Y,$$

où $H = X(X^tX)^{-1}X^t$ *est la matrice de projection (la matrice chapeau) .*

I_n *est la matrice unité de dimension* $(n; n)$.

Les résidus permettent de quantifier la distance entre la réponse observée Y_i et la réponse prédite $\hat{Y}_i$ pour $i \in \{1, ..., n\}$. Ainsi, il est important que les valeurs $\hat{Y}_i$ se rapprochent le plus possible des vraies valeurs Y_i , ce qui témoigne d'un modèle de régression linéaire bien ajusté pour les données.

Proposition 1.2.3 *(Espérance et variance des résidus)[50].*

Le vecteur de résidus est tel que $E(\hat{\varepsilon}) = 0$ *et la matrice de variance covariance des résidus satisfait* $V(\hat{\varepsilon}) = \sigma^2(I_n - H)$. *Les résidus sont centrés et corrélés .*

L'espérance et la variance sont calculées par rapport à la distribution de la variable aléatoire Y :

$$E(\hat{\varepsilon}) = E((I_n - H)Y) = (I_n - H)E(Y) = (I_n - H)X\beta = X\beta - X(X^tX)^{-1}X^tX\beta = 0.$$

Sachant que $(I_n - H)$ est une matrice symétrique et idempotente, la variance satisfait

$$\begin{cases} Symetrique : H^t = H \\ Idempotente \ : H^2 = H \end{cases}$$

$$V(\hat{\varepsilon}) = V((I_n - H)Y) = (I_n - H)V(Y)(I_n - H)^t = \sigma^2(I_n - H).$$

l'espérance de $\hat{\varepsilon}$ est nulle et sa variance est $\sigma^2(I_n - H)$. Puisque la trace de la matrice identité est n et que celle de la matrice de projection est p, alors selon les propriétés de la trace,

$$Tr(V(\hat{\varepsilon})) = Tr(\sigma^2(I_n - H)) = \sigma^2(n - p).$$

Proposition 1.2.4 *[53]*

Soit S^2, défini comme suit

$$S^2 = \frac{||\hat{\varepsilon}||^2}{n-p} = \frac{SRC}{n-p}.$$

alors, S^2 est sans biais pour σ^2.

Démonstration

Sachant que $E(||\hat{\varepsilon}||^2) = \sigma^2(n-p)$,alors

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

Ayant identifié un estimateur pour σ^2, il est maintenant possible d'estimer la variance de $\hat{\beta}$.

Proposition 1.2.5 *[53]*

Soit la variance estimée de l'estimateur MCO $\hat{\beta}$, définie comme

$$\hat{V}(\hat{\beta}) = S^2(X^tX)^{-1}, \tag{1.5}$$

alors, $\hat{V}(\hat{\beta})$ est sans biais pour $V(\hat{\beta})$.

Démonstration

Sachant que $E(S^2) = \sigma^2$, alors

$$E(\hat{V}(\hat{\beta})) = E(S^2)(X^tX)^{-1} = \sigma^2(X^tX)^{-1}.$$

1.2.6 Somme des carrés

– La somme des carrés totaux (**SCT**) est définie comme suit[45] :

$$SCT = ||Y - \bar{Y}||^2,$$

– La somme résiduelle des carrés(**SRC**) est définie comme suit :

$$SRC = ||\hat{\varepsilon}||^2 = ||Y - \hat{Y}||^2,$$

où $\hat{Y} = X\hat{\beta}$ est le vecteur de valeurs prédites.

– La somme des carrés expliqués (par la régression)(**SCE**) est :

$$SCE = ||\hat{Y} - \bar{Y}||^2.$$

1.3 Sélection de modèle

La sélection de modèle dans la régression linéaire multiple est un processus essentiel pour garantir que notre modèle soit adapté aux données et puisse être utilisé de manière fiable pour des prévisions et des analyses futures.

1.3.1 Méthodes de sélection de variables traditionnelles

Les méthodes de sélection de variables traditionnelles en régression linéaire sont très intéressantes et souvent performantes. Ces méthodes nous permettent de choisir un sous-modèle offrant des prédictions de bonne qualité et une interprétation adéquate du modèle. Dans le but de bien comprendre ces méthodes, nous introduisons maintenant quelques mesures d'ajustement de modèles telles que les coefficients R^2et R^2 ajusté (R_a^2), ainsi que le C_p de Mallows.

Cœfficient de détermination $\mathbf{R}^2$

Le rapport entre SCE et SCT représente la proportion de variance expliquée et porte le nom de Cœfficient de détermination , noté par R^2[15].

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

Ce Cœfficient R^2 est compris entre 0 et 1 : plus il est proche de 1 , signifie que la variabilité des différentes valeurs prises par la variable dépendante Y autrement dit meilleure est la régression. Inversement, un Cœfficient R^2 proche de0 indique qu'aucune variabilité n'est expliquée. .

Coeffcient de détermination ajusté $\mathbf{R}_a^2$

Le coeffcient R^2 est un indicateur de la qualité de l'ajustement des valeurs observées par le modèle mais il a le défaut de ne pas tenir compte du nombre de variables explicatives utilisés dans le modèle. On ne peut pas l'utiliser pour comparer plusieurs modèles entre eux car, si on ajoute une variable explicative à un modèle, la part des erreurs diminue forcément et donc le coeffcient R^2 augmente : cela signifie que plus il y a de variables explicatives et plus le R^2 est élevé. Or un modèle n'est pas nécessairement meilleur parce qu'il a plus de variables explicatives.

On défnit donc un coeffcient R^2 ajusté qui tient compte des degrés de liberté. Ce coeffcient, noté par R_a^2 , est défni comme suit[18] :

$$R_a^2 = 1 - \frac{SRC/(n-p)}{SCT/(n-1)}.$$

Remarque 1.3.1 *On a R_a^2 est toujours inférieur à R^2, et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs (variables explicatives).*

C_p de Mallows

Le critère C_p de Mallows, proposé par Mallows en 1973 [5] est une mesure d'ajustement de modèle largement utilisée dans la sélection de modèle pour la régression linéaire multiple. Cette mesure est conçue pour évaluer la qualité de prédiction du modèle tout en prenant en compte sa complexité.

$$C_p = \frac{SRC_p}{S^2} - n + 2(p+1).$$

où

SRC_p : la somme des carrés des résidus pour un modèle avec p variables prédictrices.

S^2 : la moyenne des carrés des résidus pour le modèle (estimée par l'EMQ) .

n : la taille de l'échantillon.

p : le nombre de variables prédictrices.

En pratique, il s'agit de conserver le modèle ayant le plus petit C_p.

AIC, BIC

En régression linéaire, le critère AIC est équivalent au C_p de Mallows selon **Akaike (1998)[36]**. Les critères AIC et BIC (respectivement Critère d'Information d'Akaike et Critère d'Information Bayésienne). Sont très similaires, mais se distinguent dans la pénalité qu'ils appliquent relativement à la taille du modèle. Le critère AIC pénalise chaque paramètre additionnel par un facteur de 2[1] , alors que le BIC utilise plutôt un facteur de $log(n)$. Ainsi, BIC a tendance à choisir des modèles plus parcimonieux que AIC.

Les formules des deux critères sont les suivantes :

$$AIC = n \ln \frac{SRC}{n} 2p.$$

$$BIC = n \ln \frac{SRC}{n} \ln(n)p$$

On remarque que lorsque $n > e^2 \approx 7$, on constate que le critère BIC pénalise plus fortement les modèles complexes.

1.4 Régression pénalisée

L'estimateur des moindres carrés est à la fois sans biais et possède la variance la plus faible parmi tous les estimateurs linéaires sans biais de β. Ainsi, pour obtenir une prédiction plus précise que celle de l'estimateur des moindres carrés, il est nécessaire de réduire la variance, même au détriment du biais. On espère que la réduction de la variance compensera la perte de l'absence de biais, et que le gain net en termes de précision de la prédiction sera positif. **Hoerl et Kennard** [27]Ont proposé des méthodes de régressions pénalisés, qui forcent les éléments de β à avoir une certaine forme en vu de reduire l'erreur de prédiction. La régression pénalisée ajoute à la fonction objective à minimiser, $(Y - X\beta)^t(Y - X\beta)$, une pénalité $p(\beta)$ qui est une fonction de β.

Définition 1.4.1 *Le coefficient de régression pénalisée est défini comme l'estimateur* $\hat{\beta}^{(P)}$ *qui minimise*

$$\hat{\beta}^{(P)} = \arg\min_{\beta \in R^p} \{||Y - X\beta||^2\}, \tag{1.6}$$

Sous la contrainte

$$||\beta||_q^r \leq t,$$

où la constante t *est appelée « borne absolue » de la contrainte*

– pour $q = r = 2$: Régression Ridge.

– pour $q = r = 1$: Régression Lasso.

– Régression Elastic Net : Combine les propriétés de Ridge et Lasso.

1.4.1 Régression ridge

La régression Ridge est une technique de régression linéaire régularisée introduite par Hoerl et Kennard (1970)[27]. Elle permet de traiter les problèmes de surajustement et de multicolinéarité en ajoutant une pénalisation sur les coefficients du modèle.

La pénalité utilisée dans cette méthode est définie par :

$$p(\beta) = ||\beta||_2^2 \ ,$$

avec

$$||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2 \ .$$

1.4.2 Estimateur ridge

L'estimateur Ridge $\hat{\beta}^{(R)}$ est défini par[31] :

$$\hat{\beta}^{(R)} = \arg\min_{\beta \in R^p} \{ \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \},$$

Sous la contrainte que, pour un certain $t > 0$,

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

où $\lambda \geq 0$ est un paramètre de régularisation.

La fonction à minimiser peut s'écrire sous forme matricielle :

$$\hat{\beta}^{(R)} = \arg\min_{\beta} \{ ||Y - X\beta||^2 + \lambda ||\beta||_2^2 \tag{1.7}$$

ou encore

$$\hat{\beta}^{(R)} = (Y - X\beta)^t(Y - X\beta) + \lambda\beta^t\beta,$$
$$\hat{\beta}^{(R)} = Y^tY - 2\beta^tX^tY + \beta^tX^tX\beta + \lambda\beta^t\beta.$$

Si on dérive par rapport à β et qu'on pose les dérivées égales à zéro, on trouve :

$$X^tY = X^tX\beta + \lambda\beta,$$
$$= (X^tX + \lambda I_p)\beta.$$

La matrice symétrique X^tX étant définie semi-positive, toutes ses valeurs propres sont non négatives et donc $X^tX + \lambda I_p$ est symétrique et définie positive ($\lambda > 0$), donc inversible. Alors, l'estimateur Ridge est donné par la formule explicite[8] :

$$\hat{\beta}^{(R)} = (X^tX + \lambda I_p)^{-1}X^tY \ . \tag{1.8}$$

Dans le cas où X est orthogonale,

$$\hat{\beta}^{(R)} = (1 + \lambda)^{-1}X^tY = (1 + \lambda)^{-1}\hat{\beta}^{MCO}.$$

et

$$\hat{\beta}_j^{(R)} = \frac{\hat{\beta}_j^{MCO}}{(1 + \lambda)} \ .$$

avec $\hat{\beta}_j^{MCO}$ est donné par l'équation (1.4).

Remarques

La constante β_0 n'intervient pas dans la pénalité, sinon, le choix de l'origine pour Y aurait une influence sur l'estimation de l'ensembles des paramètres. On obtient $\hat{\beta}_0 =$

$\bar{Y}$, ajouter une constante à Y ne modifie pas les $\hat{\beta}_j^{(R)}$ pour $j \geq 1$.

L'estimateur ridge n'est pas invariant par renormalisation des vecteurs X_j , il est préférable de normaliser les vecteurs avant de minimiser le critère [37].

On montre que l'estimateur ridge revient encore à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur β des paramètres ne soit pas trop grande :

$$\hat{\beta}^{(R)} = \arg\min_{\beta}\{||Y - X\beta||^2 ; \ ||\beta|| \leq t\}.$$

Propriétés de l'estimateur ridge

Espérance Revenons au définition des estimateurs ridge et des moindres carrée MCO [28] :

$$\hat{\beta}^{(R)} = (X^t X + \lambda I_p)^{-1} X^t Y \ . \tag{1.9}$$

et

$$\hat{\beta}^{(MCO)} = (X^t X)^{-1} X^t Y. \tag{1.10}$$

En multipliant (1.8) par $X^t X$, on obtient $X^t X \hat{\beta}^{(MCO)} = X^t Y$.

D'aprés la définition de l'estimateur MCO vient :

$$\hat{\beta}^{(R)} = (X^t X + \lambda I_p)^{-1} X^t X \hat{\beta}^{(MCO)}$$

Cette écriture permet de calculer facilement le biais et la variance de l'estimateur ridge. En utilisant les propriétés de l'estimateur MCO, l'espérance de l'estimateur ridge est

$$\begin{aligned} E(\hat{\beta}^{(R)}) &= E(X^tX + \lambda I_p)^{-1}X^tXE(\hat{\beta}) \\ &= (X^tX + \lambda I_p)^{-1}X^tX\beta \\ &= (X^tX + \lambda I_p)^{-1}(X^tX + \lambda I_p - \lambda I_p)\beta \\ &= \beta - \lambda(X^tX + \lambda I_p)^{-1}\beta. \end{aligned}$$

La variance

Proposition 1.4.1 *[24](Biais et variance des estimateurs Ridge). Si les conditions de Gauss Markov sont satisfaites, alors l'estimateur $\hat{\beta}^{(R)}$ est un estimateur possédant un biais égal à :*

$$Biais(\hat{\beta}^{(R)}) = -\lambda(X^tX + \lambda I_p)^{-1}\beta \ .$$

et une variance satisfaisant

$$V(\hat{\beta}^{(R)}) = \sigma^2(X^tX + \lambda I_p)^{-1}X^tX(X^tX + \lambda I_p)^{-1}.$$

Démonstration

Le biais de cet estimateur vaut

$$Biais(\hat{\beta}^{(R)}) = E(\hat{\beta}^{(R)} - \beta).$$

$$Biais(\hat{\beta}^{(R)}) = -\lambda(X^tX + \lambda I_p)^{-1}\beta.$$

ce qui montre que l'estimateur Ridge $\hat{\beta}^{(R)}$ a un biais non nul (un estimateur biaisé)[25].Contrairement à l'estimateur MCO.

La variance est

$$V\ (\hat{\beta}^{(R)}) = V\ (X^tX + \lambda I_p)^{-1}X^tY$$
$$= \ (X^tX + \lambda I_p)^{-1}X^tV(Y)X(X^tX + \lambda I_p)^{-1}$$

où cette hypothèse

$$V(\varepsilon) = \sigma^2 I_n$$

Et on obtient finalement la variance de l'estimateur ridge :

$$V\ (\hat{\beta}^{(R)}) = \sigma^2(X^tX + \lambda I_p)^{-1}X^tX(X^tX + \lambda I_p)^{-1}.$$

remarques

1. Si $\lambda \to 0$, alors $\hat{\beta}^{(R)} \to \hat{\beta}$, $B(\hat{\beta}^{(R)}) \to 0$, $V(\hat{\beta}^{(R)}) \to \infty$.
2. Si $\lambda \to \infty$, alors $\hat{\beta}^{(R)} \to 0, B(\hat{\beta}^{(R)}) \to \infty$, $V(\hat{\beta}^{(R)}) \to 0$.

1.4.3 Interprétation géométrique de la régression ridge

Dans cette figure 1.1, les ellipses illustrent les contours des sommes des carrés des résidus (SCR). Plus l'ellipse est proche du centre, plus la SCR est faible, et son minimum correspond aux estimations obtenues par les moindres carrés ordinaires (MCO). En régression ridge, la contrainte est représentée par un cercle de rayon , $||\beta^2|| \leq t$.

Nous cherchons à réduire simultanément la taille de l'ellipse et du cercle en régression ridge. L'estimation ridge se trouve au point de contact entre l'ellipse et le cercle. Il y a un compromis entre le terme de pénalité et la SRC . Une grande valeur de λ pourrait donner une meilleure somme des carrés des résidus, mais augmenterait le terme de pénalité. C'est pourquoi on peut préférer des λ plus petits, même si cela

donne une RSS moins bonne. D'un point de vue optimisation, le terme de pénalité équivaut à une contrainte sur les β. La fonction reste la somme des carrés des résidus, mais on contraint maintenant la norme des β à être inférieure à une constante t. Il existe une correspondance entre λ et t. Plus λ est grand, plus on préfère que les β soient proches de zéro. Dans le cas extrême où $\lambda = 0$, on effectue une régression linéaire normale. À l'autre extrême, lorsque λ tend vers l'infini, on fixe tous les β à zéro.

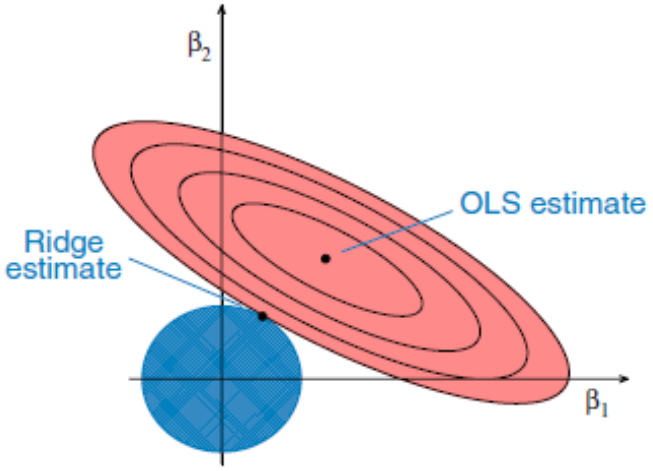

FIG. 1.1 – Illustration de la regression Ridge.

La régression Ridge est une méthode intuitive introduisant un paramètre d'ajustement λ dans l'équation de $\hat{\beta}$. Ceci nous permet, au prix d'un biais, de rétrécir les coefficients ($\hat{\beta}^{(R)} \rightarrow 0$) et de diminuer la variance afin de régler le problème de multicolinéarité. Malgré ses vertus, la régression ridge ne fait pas de sélection de modèle, ce qui nous oblige à utiliser le modèle complet.

1.5 Conclusion

En conclusion, la régression linéaire et la régression pénalisée jouent un rôle fondamental dans l'analyse statistique contemporaine, en fournissant des solutions efficaces pour modéliser des phénomènes complexes et prendre des décisions éclairées basées sur les données. Une utilisation judicieuse de ces méthodes permet d'obtenir des modèles plus précis, plus facilement interprétables et plus généralisables, ce qui contribue significativement à l'avancement de la recherche scientifique, de l'analyse des données et de la prise de décision dans divers domaines d'application.

Chapitre 2

La regression en L1

2.1 Introduction

La régression en L1, connue sous le nom de régression Lasso (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) ce qui signifie "Réduction et Sélection des Coefficients par la Méthode de la Moindre Valeur Absolue " . introduit par **Tibshirani (1996)[49]** , est une technique populaire en statistiques et en apprentissage automatique[36]. Cette méthode est reconnue pour améliorer la généralisation et prévenir le surajustement[43].

Dans ce chapitre, nous allons explorer en détail l'estimateur Lasso, le paramètre de régularisation, la validation croisée, l'interprétation géométrique et bayésienne de la régression Lasso, quelques généralisations, puis l'Elastic Net.

2.2 Estimateur Lasso

Les problèmes de régression en grande dimension, où le nombre de variables explicatives est très élevé, posent des défis importants pour obtenir des estimations fiables et facilement interprétables.

2.2.1 Définition de l'estimateur Lasso

L'estimateur LASSO introduit pour la première fois par Tibshirani [49], cet estimateur est défini comme l'estimateur des moindres carrés sous une contrainte de type L1[23] :

$$\begin{cases} \hat{\beta}^{(L)} = \arg\min_{\beta \in R^p} ||Y - X\beta||_2^2 \\ \text{Sous la contrainte. } ||\beta||_1 \leq t \end{cases} \tag{2.1}$$

où $\beta \in R^p$, Le paramètre s contrôle le niveau de régularisation des coefficients estimés, $t \geq 0$ et $||.||_1$ étant la norme L1[16].

De manière équivalente :

$$\hat{\beta}^{(L)} = \arg\min_{\beta \in R^p} ||Y - X\beta||_2^2 + \lambda||\beta||_1. \tag{2.2}$$

où le paramètre de régularisation $\lambda \geq 0$ a une correspondance biunivoque avec le paramètre s de l'équation **??**. Ainsi, l'estimateur du lasso remplace la pénalité L2 de l'estimateur de ridge (**Hoerl & Kennard, 1970)** par une pénalité L1. Ce problème d'optimisation est convexe mais, en raison de la pénalité L1 , non différentiable en zéro. Pour un ensemble de données donné avec une taille d'échantillon finie, il est toujours possible d'obtenir une erreur attendue plus faible avec une estimation biaisée (régularisée) [29].

Définition 2.2.1 *[14]*

Le coefficient de régression Lasso est défini comme l'estimateur $\hat{\beta}^{(L)}$ *qui minimise.*

$$\hat{\beta}^{(L)} = \arg\min_{\beta \in R^p} \{\frac{1}{2n}(Y - X\beta)^t(Y - X\beta) + \lambda||\beta||_1\},$$

La norme L1 dans la pénalité Lasso nous empêche d'avoir accès à une solution analytique pour l'estimateur $\hat{\beta}^{(L)}$. De ce fait, il faut procéder avec des algorithmes itéra-

tifs pour trouver l'estimateur Lasso. Il est alors recommandé d'ajouter la constante $1/(2n)$ devant la fonction objective, ce qui permet d'accélérer les calculs pour les différents algorithmes.

2.2.2 Propriétés

- Si λ =0, le Lasso correspond à une régression linéaire classique. on retrouve l'estimateur des Moindres Carrés. *ie* $\hat{\beta}^{(L)} = \hat{\beta}^{(MCO)}$. La méthode du Lasso sélectionne les variables sans exception.
- Si $\lambda \to \infty$, tous les coefficients de $\hat{\beta}$ sont nuls.$\hat{\beta}_j^{(L)} = 0 \Rightarrow \hat{\beta}^{(L)} = 0$.Dans ce cas, le Lasso ne sélectionne aucune variable explicative.
- Si $\lambda \in]0, +\infty[$, le nombre de variable sélectionnées par le Lasso diminue lorsque devient grand. c'est-à-dire, si λ est grand, la contrainte exercé sur le vecteur β l'est également. En d'autres termes, l'augmentation du paramètre λ induit la diminution de certains coefficients de $\hat{\beta}^{(L)}$ vers jusqu'à ce qu'ils soient exactement nuls.

2.2.3 Calcul analytique de la solution de la méthode Lasso

L'estimateur Lasso n'a pas de formule analytique dans le cadre général. Cependant, dans le cas où $X^tX = I_p$, le Lasso a une solution explicite. L'estimateur Lasso correspond alors à un seuillage doux de $\hat{\beta}$ la solution des moindres carrées .

Cas orthonormal

Proposition 2.2.1 *[53]*

Soit des variables explicatives orthonormées, c'est à dire satisfaisant $X^tX = I_p$ *. Il*

est alors possible d'obtenir une solution fermée pour l'estimateur Lasso, soit

$$\hat{\beta}_j^{(L)} = sign\ (\hat{\beta}_j^{(MCO)})(|\hat{\beta}_j^{(MCO)}| - n\lambda)^+, \tag{2.3}$$

pour $j = 1, ..., p$, *où* $(\hat{\beta}_j^{MCO})$ *sont les coefficients des moindres carrés ordinaires,* $(x)^+ = max(0, x)$ *est l'opérateur qui prend uniquement des valeurs positives et la fonction sign(·) est définie comme suit*

$$sign\ (\hat{\beta}_j^{MCO}) = \begin{cases} -1 & si\ \ \hat{\beta}_j^{(MCO)} < 0, \\ 0 & si\ \ \hat{\beta}_j^{(MCO)} = 0, \\ 1 & si\ \ \hat{\beta}_j^{(MCO)} > 0. \end{cases}$$

Démonstration

Sachant que les variables explicatives sont orthonormées, la solution de l'estimateur **MCO** devient $\hat{\beta}^{(MCO)} = XY^t$. Il suffit maitenant de développer la fonction objective et de trouver son minimum. Cette fonction satisfait

$$\begin{aligned} S(\beta) &= \frac{1}{2n}(Y - X\beta)^t(Y - X\beta) + \lambda||\beta||_1^2 \\ &= \frac{1}{2n}Y^tY - \frac{1}{n}Y^tX\beta + \frac{1}{2n}X^t\beta^tX\beta + \lambda||\beta||_1 \end{aligned}$$

Exprimée à l'aide de sommes, celle-ci devient

$$S(\beta) = \frac{1}{2n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\sum_{j=1}^{p}\ \hat{\beta}_j^{MCO}\beta_j + \frac{1}{2n}\sum_{j=1}^{p}\beta_j + \lambda\sum_{j=1}^{p}|\beta_j|,$$

pour $j = 1, ..., p$. Maintenant, en examinant cette équation, il est évident que les $\hat{\beta}_j$ qui la minimiseront seront de mêmes signes que les $\hat{\beta}_j^{(MCO)}$ correspondants. En effet, pour minimiser cette équation, il est capital de faire en sorte que chacun des termes dans la somme $-\frac{1}{n}\sum_{j=1}^{p}\ \hat{\beta}_j^{(MCO)}\beta_j$ soit négatif.

En dérivant par rapport à β_j, nous trouvons donc

$$\begin{aligned}\frac{\partial S(\beta_j)}{\beta_j} &= -\frac{1}{n}\hat{\beta}_j^{(MCO)} + \frac{1}{n}\beta_j + \lambda\ sign(\beta_j) \\ &= -\frac{1}{n}\hat{\beta}_j^{(MCO)} + \frac{1}{n}\beta_j + \lambda\ sign\ (\hat{\beta}_j^{(MCO)}).\end{aligned}$$

L'estimateur $\hat{\beta}_j^{(L)}$ satisfait alors l'équation

$$-\frac{1}{n}\hat{\beta}_j^{(MCO)} + \frac{1}{n}\hat{\beta}_j^{(L)} + \lambda\ sign\ (\hat{\beta}_j^{(MCO)}) = 0.$$

En isolant $\hat{\beta}_j^{(L)}$, nous trouvons

$$\begin{aligned}\hat{\beta}_j^{(L)} &= \hat{\beta}_j^{(MCO)} - n\lambda\ sign\ (\hat{\beta}_j^{(MCO)}) \\ &= \ sign\ (\hat{\beta}_j^{(MCO)})(|\hat{\beta}_j^{(MCO)}| - n\lambda).\end{aligned}$$

Sachant que $\hat{\beta}_j^{(L)}$ et $\hat{\beta}_j^{(MCO)}$ doivent être de même signe, il est important que $(|\hat{\beta}_j^{(MCO)}| - n\lambda)$ soit positif. De ce fait, on utilise l'opérateur $(\cdot)^+$, qui retourne uniquement des valeurs non-négatives, soit

$$\hat{\beta}_j^{(L)} = sign\ (\hat{\beta}_j^{(MCO)})(|\hat{\beta}_j^{(MCO)}| - n\lambda)^+. \tag{2.4}$$

En travaillant avec des variables explicatives orthonormées, il est également possible d'obtenir une solution réduite pour l'estimateur Ridge, soit

$$\hat{\beta}^{(R)} = (1+\lambda)^{-1}X^tY = (1+\lambda)^{-1}\hat{\beta}^{(MCO)} \tag{2.5}$$

Remarques

– Le Lasso met tous les coefficients à 0 lorsque le $max_j|\hat{\beta}_j^{(MCO)}| \leq n\lambda$, ce qui témoigne de sa supériorité par rapport à l'estimateur Ridge.

– Cette formule explicite n'est valable que dans le cas orthogonal. Dans le cas général, il n'existe pas de formule analytique pour l'estimateur Lasso et des algorithmes itératifs doivent être utilisés, comme la descente de gradient proximal (ISTA, FISTA).

2.2.4 Cas particulier simple pour Lasso

Considérons un cas spécial simple où $n = p$ et où X est une matrice diagonale avec des 1 sur la diagonale et des 0 sur tous les éléments hors diagonale. Pour rendre le problème plus simple, ajoutons l'hypothèse que nous effectuons la régression sans interception. Dans ce contexte, le problème des moindres carrés devient la recherche des coefficients $\beta_1, ..., \beta_p$ qui minimisent :

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 \tag{2.6}$$

Dans ce cas, la solution des moindres carrés est donnée par

$$\hat{\beta}_j = y_i.$$

Et dans ce cadre, Le lasso revient à trouver les coefficients tels que

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

les estimations du Lasso prennent la forme suivante :

$$\hat{\beta}_j^{(L)} = \begin{cases} y_j - \lambda/2 & \text{si } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{si } y_j < \lambda/2; \\ 0 & \text{si } y_j < \lambda/2. \end{cases} \tag{2.7}$$

Ce cas spécial permet de visualiser de manière plus claire comment le Lasso agit dans un scénario simplifié, offrant ainsi une compréhension plus approfondie de cette technique de régression régularisée.

2.3 Paramètre de régularisation

La régression pénalisée possède un paramètre d'ajustement λ qui doit être sélectionné par l'utilisateur[47]. Ce paramètre doit être choisi judicieusement, car différents choix mèneront à différentes estimations des coefficients Tel que mentionné dans les remarques, si le paramètre λ est trop grand, les coefficients de l'estimateur Ridge seront rétrécis et la variance sera diminuée, mais le biais sera augmenté. Pour le Lasso, si le paramètre λ est trop petit, le modèle risque de comporter trop de variables alors que s'il est trop grand, le modèle sera trop simple. Il faut donc trouver un juste milieu pour le choix de ce paramètre ; pour ce faire, nous utilisons la technique de validation croisée .

2.3.1 La validation croisée

Dans le processus de construction d'un modèle de régression Lasso, le choix du paramètre de régularisation λ joue un rôle crucial pour garantir à la fois la précision du modèle[47]. Pour trouver la valeur optimale de ce paramètre, nous utilisons une méthode éprouvée appelée validation croisée.

Pour déterminer la valeur optimale du paramètre λ parmi les candidats $\lambda \in \{\lambda_1, \ldots, \lambda_m\}$, nous procédons comme suit :

Regroupement

Nous divisons l'échantillon sous l'étude en k groupes de même taille, habituellement entre 5 et 10. En effet, nous laissons un groupe pour la validation et les autres $k-1$ groupes pour l'apprentissage. Par exemple, on prend le groupe 1 pour la validation, on ajuste le modèle pour les autres $k-1$ groupes et on calcule l'erreur MSE_1 . Ensuite, on refait le même processus pour les

autres groupes en calculant les erreurs $MSE_2, MSE_3, ..., MSE_k$.

Calcul de l'Erreur

Nous calculons la moyenne des K estimations de l'erreur de prédiction pour chaque valeur de λ, selon la formule[36] :

$$CV = K^{-1} \sum_{k=1}^{K} MSE_i,$$

Sélection du Meilleur λ

Nous identifions le paramètre λ qui minimise l'erreur de prédiction moyenne sur l'ensemble des données, noté λ_{min} [29] :

$$\lambda_{\min} = \arg \min_{\lambda \in \{\lambda_1, ..., \lambda_n\}} CV.$$

Nouvel Ajustement du Modèle

Un nouveau modèle est ajusté en utilisant ce λ_{min} pour obtenir un modèle final plus précis.

Cette approche de validation croisée nous permet d'évaluer efficacement la performance du modèle pour différentes valeurs de λ et de sélectionner celle qui minimise l'erreur de prédiction.

2.4 Interprétation géométrique

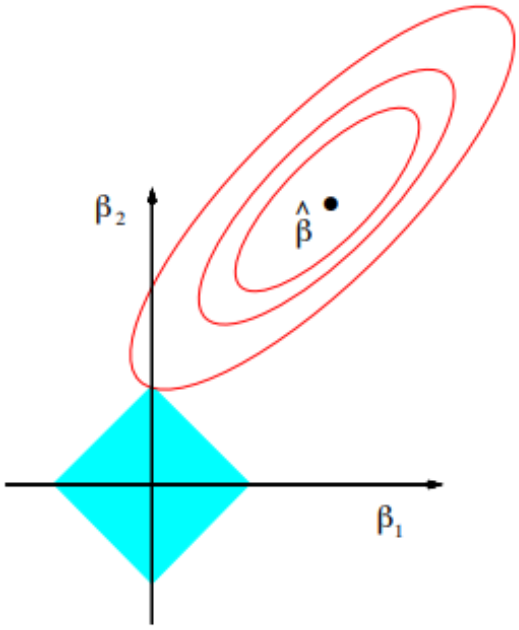

FIG. 2.1 – Interprétation géométrique de la régression Lasso.

La figure 2.1 [40], illustre les estimateurs de la régression LASSO en utilisant des contours elliptiques pour la somme des carrés des erreurs, centrés sur l'estimateur des moindres carrés ordinaires (MCO). Les zones ombrées représentent l'ensemble de contraintes pour le Lasso. La géométrie L1 crée un diamant convexe avec des bords. La solution LASSO est le point où les contours elliptiques touchent l'ensemble de contraintes. Contrairement à la régression Ridge, l'estimation LASSO peut être fixée à zéro lorsque les contours elliptiques touchent un sommet du diamant. Cela signifie que certains coefficients peuvent être exactement nuls, ce qui facilite la sélection de variables pertinentes. Dans un modèle simplifié, par exemple, si β_2 est le seul paramètre pertinent, l'intersection des contours elliptiques avec le sommet du diamant conduira à $\beta_1 = 0$.

Pour des dimensions plus élevées, le diamant se transforme en polyèdre, augmentant ainsi le nombre de possibilités pour que certaines estimations soient tronquées à zéro. Un exemple en trois dimensions est décrit par Tibshirani (1996)dans la figure 2.2[49].

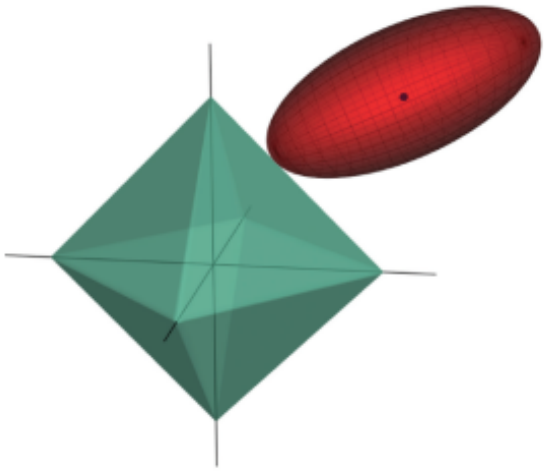

FIG. 2.2 – Interprétation géométrique de la régression Lasso en 3D.

2.5 Interprétation bayésienne de la régression Lasso

On peut voir la régression lasso d'un point de vue bayésien. Un point de vue bayésien pour la régression suppose que le vecteur coefficients β a une certaine distribution a priori, disons $P(\beta)$, où $\beta = (\beta_0, \beta_1, ..., \beta_p)$. La probabilité des données peut s'écrire $f\ (Y\ \backslash X\ ,\beta)$ où $X = (X_1, ..., X_p)$. En multipliant la distribution a priori par la vraisemblance, on obtient la distribution a posteriori, qui prend la forme[49] :

$$P(\beta\ \backslash X, Y) \propto\ f(Y\ \backslash X, \beta)P(\beta\ \backslash X) =\ f(Y\ \backslash X, \beta)P(\beta\ ),$$

où la proportionnalité ci-dessus découle du théorème de Bayes, et l'égalité ci-dessus

découle de l'hypothèse que X est fixe [28]. Nous supposons que le modèle linéaire habituel,

$Y = \beta_0 + X_1\beta_1 + ... + X_p\beta_p + \varepsilon$, et supposons que les erreurs soient indépendantes et tirées d'une distribution normale. De plus, supposons que $P(\beta\ ) = \prod_{j=1}^{p} g(\beta_j)$, pour une certaine fonction de densité g. Il s'avère que la régression lasso et ridge découlent naturellement de deux cas particuliers de g :

- Si g est une distribution gaussienne avec une moyenne nulle et un écart type une fonction de λ , il s'ensuit que le mode a posteriori de β est la solution de régression de ridge. En fait, la solution de régression de crête est également la moyenne a posteriori.
- Si g est une distribution exponentielle double avec une moyenne nulle et un paramètre d'échelle une fonction de λ , il s'ensuit que le mode postérieur pour β est la solution LASSO. Cependant, la solution lasso n'est pas la moyenne a posteriori, et en fait, la moyenne a posteriori ne donne pas un vecteur de coefficient clairsemé[33, 37].

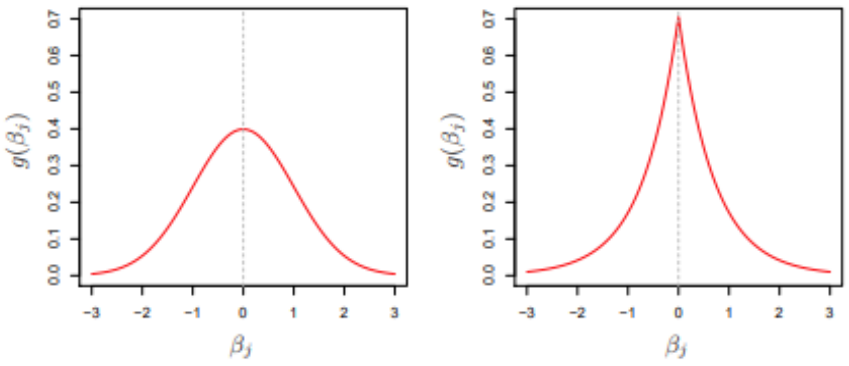

FIG. 2.3 – Densités a priori implicites pour la régression Lasso et Ridge.

Le diagramme2.3 de gauche illustre la régression de ridge en tant que mode a poste-

riori pour β sous une a priori gaussienne. Le diagramme de droite représente le lasso en tant que mode a posteriori pour β sous une a priori double-exponentielle. Ainsi, d'un point de vue bayésien, la régression de crête et le lasso découlent directement de l'hypothèse du modèle linéaire habituel avec des erreurs normales, ainsi que d'une distribution a priori simple pour β.

La Figure [28] indique que le lasso utilise une distribution a priori double-exponentielle tandis que la régression de crête utilise une distribution a priori gaussienne. Cela montre que la régression de crête et le lasso découlent directement de l'hypothèse du modèle linéaire standard avec des erreurs normales, ainsi que d'une distribution a priori simple. La distribution a priori du lasso présente un pic abrupt autour de zéro, tandis que la distribution gaussienne est plus plate et s'étale progressivement vers zéro. Par conséquent, le lasso suppose a priori que de nombreux coefficients sont exactement égaux à zéro, tandis que la régression de crête suppose que les coefficients sont répartis de manière aléatoire autour de zéro.

2.6 Quelques généralisations des régressions Ridge et Lasso

	Ridge	Lasso
Qualités	- Stabilité face à la multicolinéarité. - Efficacité en haute dimension.	- Efficacité en haute dimension. - Sélection automatique des prédicteurs.
Défauts	- ne sélectionne aucune variable, mais rétrécit les coefficients.	- sensible à la multicolinéarité. - biais vers 0 des coefficients importants.

Les modèles de régressions Lasso et Ridge ne donnent pas toujours les meilleures performances en termes de prédiction. Cependant, ils ont l'avantage de fournir un modèle parcimonieux, ce qui est souhaitable pour faciliter l'interprétation des résul-

tats. Pour combiner les avantages des pénalités L1 et L2, **Zou et Hastie (2005)** [56]ont proposé la méthode de l'Elastic Net. Cette approche combine les propriétés souhaitables des deux méthodes : la sélection de variables du Lasso et la gestion de la multicolinéarité de la régression Ridge.

2.7 Elastic Net

La méthode EN est une technique de régularisation qui combine les deux normes[56], la norme L_1 et la norme L_2 . Ainsi, cette méthode est un compromis entre la méthode Lasso et la méthode Ridge. Sa pénalité est donnée par :

$$p(\beta) = ||\beta||_1 + ||\beta||_2^2 \ .$$

Le modèle Elastic Net minimise une fonction de coût qui inclut une combinaison pondérée des termes L1 et L2 :

$$\hat{\beta}^{(EN)} = \arg\min_{\beta \in R^p} \{\sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2\}.$$

où

- $\lambda_1$et λ_2 sont les hyperparamètres de régularisation respectifs pour les termes L1 et L2.
- $\alpha = 1$, on retrouve la méthode LASSO.
- α= 0, on retrouve la régression ridge.

2.8 Conclusion

En conclusion, la régression Lasso est un outil puissant pour analyser des données complexes. Elle choisit automatiquement les variables importantes et simplifie le

modèle, ce qui est très utile quand on a beaucoup de variables. Le Lasso crée des modèles plus simples à comprendre et souvent plus précis. Choisir le bon niveau de simplification peut être difficile et demande des tests, mais le Lasso aide à gérer les problèmes de variables trop liées entre elles . C'est pourquoi le Lasso est très apprécié dans de nombreux domaines pour analyser des données.

Chapitre 3

Comparaison entre la régression linéaire multiple, Ridge et Lasso

3.1 Introduction

Dans ce chapitre, nous allons explorer deux méthodes de régression régularisée, à savoir le Lasso et le Ridge, en les comparant à la régression linéaire multiple traditionnelle. Ces techniques, indispensables pour la modélisation de données complexes , offrent des avantages distincts en termes de performance et de robustesse. Nous verrons comment chacune de ces approches peut être appliquée efficacement dans le contexte de l'analyse statistique et de la modélisation prédictive . Pour cette étude, nous utiliserons le logiciel R pour implémenter ces méthodes, analyser les résultats et évaluer les performances des modèles.

3.2 La consommation d'essence pour les différents véhicules

La consommation d'essence est un sujet préoccupant à la fois pour les automobilistes et les constructeurs automobiles. Elle dépend de nombreux facteurs tels que le poids, la puissance, la cylindrée et la vitesse du véhicule. Comprendre les principaux déterminants de la consommation d'essence est essentiel pour concevoir des véhicules plus économes en carburant . Le tableau de données fourni contient des informations sur la consommation d'essence mesurée pour différents modèles de véhicules[44]. Chaque ligne représente un véhicule avec ses caractéristiques techniques (poids, puissance, cylindrée,prix) ainsi que sa consommation d'essence moyenne observée lors de tests. Notre objectif est d'utiliser ces données pour développer un modèle prédictif de la consommation d'essence en fonction des caractéristiques des véhicules. Cela permettra non seulement de mieux comprendre les facteurs clés, mais aussi de pouvoir estimer la consommation d'un nouveau modèle sans avoir à réaliser des tests coûteux.

3.2.1 Présentation de la base de données :

Cette base de données comprend des informations sur la consommation de carburant (en litres par 100 km) de 31 véhicules différents, décrits par quatre variables explicatives. Les données incluent également le type de voiture (marque et modèle), ce qui peut aider à contextualiser les observations.

Description des variables

Les variables de la base de données sont les suivantes :

- **Type de voiture** : Marque et modèle du véhicule.
- **Prix (x1)** : Prix du véhicule en euros.
- **Cylindrée (x2)** : Volume du moteur en centimètres cubes (cm^3).

– **Puissance (x3)** : Puissance du moteur en chevaux (CV).

– **Poids (x4)** : Poids du véhicule en kilogrammes (kg).

– **Cons (y)** : Consommation de carburant en litres par 100 kilomètres (L/100km).

Notons $\mathbf{X}$ la matrice des données, elle est $n \times p$ avec $n = 31$ *et* $p = 5$.

Les données utilisées dans cette analyse sont issues de l'étude réalisée par Rakotomalala intitulée "Pratique de la régression linéaire multiple "[44]. Dans cette étude, Rakotomalala a appliqué la régression linéaire multiple sur cet ensemble de données décrivant la consommation de carburant de différents véhicules.

Ces données ont également été reprises et analysées dans un mémoire de master de Dris, L., & Hachemi, W. "Régression linéaire multiple et modèle linéaire général"[10]. Leur étude a porté sur plusieurs aspects importants de la régression linéaire multiple :(estimation des paramètres, tests sur les paramètres, test d'ajustement ,analyse de la variance (ANOVA) et la prédiction).

Dans la présente étude, nous allons appliquer et comparer différentes techniques de régression statistique : la régression LASSO, la régression ridge, et la régression linéaire multiple . pour modéliser et prédire la consommation d'essence de ces véhicules.

Tableau des données

TAB. 3.1 – Tableau de données CONSO- consommation d'essence.

Avant de commencer toute étude ou modélisation statistique d'un ensemble relativement important de données, il est préférable de faire quelques statistiques descriptives. Ces dernières servent à organiser et à résumer des observations.

	Prix	Cylindrées	Puissance	Poids	Cons
Minimum	10450	658	29.0	650	5.700
Médiane	28750	1984	101.0	1155	9.200
Moyenne	60341	2069	103.8	1258	9.884
Maximum	560900	5987	325.0	2250	21.300

TAB. 3.2 – Tableau Statistiques descriptives des variables.

Après avoir examiné les statistiques descriptives pour chaque variable, nous pouvons passer à l'étude des relations entre les différentes variables. La matrice de corrélation nous permet d'évaluer les corrélations entre les variables explicatives (prix, cylindrée, puissance et poids) et la variable cible (consommation de carburant).

La matrice de corrélation

$$\begin{pmatrix} 1 & 0.51 & 0.51 & 0.45 & 0.58 \\ 0.51 & 1 & 0,96 & 0.83 & 0.94 \\ 0.51 & 0,96 & 1 & 0.82 & 0.95 \\ 0.41 & 0.83 & 0.82 & 1 & 0.86 \\ 0.58 & 0.94 & 0.95 & 0.86 & 1 \end{pmatrix}$$

La matrice de corrélation montre une forte multicolinéarité entre la plupart des variables explicatives du modèle. Cette multicolinéarité peut poser des problèmes dans les modèles de régression linéaire classiques, entraînant une instabilité des estimations des coefficients et une variance élevée.

3.2.2 Pourquoi utiliser la régression lasso ?

L'avantage de la régression lasso est sa capacité à gérer la multicolinéarité et à améliorer les performances prédictives en trouvant un compromis entre biais et variance. En ajustant λ, nous pouvons obtenir un modèle plus robuste et plus performant sur un ensembles de données non vus. Cela signifie que le modèle ajusté par la régression

lasso produira des erreurs de test plus petites que le modèle ajusté par la régression par les moindres carrés.

$$\hat{\beta}^{(L)} = \arg\min(\sum_{i=1}^{31}(y_i-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}))^2+\lambda(|\beta_1|+|\beta_2|+|\beta_3|+|\beta_4|).$$

Dans cette étude, nous allons voir étape par étape comment effectuer une régression lasso en **R** sur les données de consommation d'essence des véhicules.

choix de λ

La première étape pour construire un modèle lasso est de trouver la valeur optimale du paramètre de régularisation λ (lambda).nous réaliserons une validation croisée sur l'ensemble d'entraînement. Cela nous permettra de sélectionner la valeur de λ qui minimise l'erreur de prédiction.

Sélection du meilleur lambda Pour sélectionner la meilleure valeur du paramètre de régularisation lambda (λ). Nous utiliserons la fonction **cv.glmnet()** de la librairie **glmnet** pour réaliser une validation croisée sur l'ensemble d'entraînement. Voici le code ci-dessous pour trouver la valeur de λ.

```
> # Ajustement du modèle Lasso avec validation croisée
> set.seed(42)
> lasso_model <- cv.glmnet(X, y, alpha = 1)
> # Meilleur lambda
> best_lambda <- lasso_model$lambda.min
> best_lambda
[1] 0.6269591
```

– La fonction **cv.glmnet()** ajuste une grille de modèles lasso avec différentes valeurs

de λ et évalue leur performance à l'aide d'une validation croisée à 10 fois (**k-fold = 10**).

- Le paramètre **alpha = 1** spécifie que nous voulons une régression lasso (**alpha = 0** correspondrait à la régression ridge).
- Lambda_optimal contient la valeur de λ qui minimise l'erreur de prédiction sur les échantillons de validation croisée .

La valeur optimale du paramètre de régularisation λ pour le modèle de régression lasso sur les données de consommation d'essence est de 0.62, trouvée par validation croisée. Cette valeur équilibre la parcimonie du modèle en éliminant les prédicteurs non pertinents tout en conservant les prédicteurs clés, assurant ainsi une bonne qualité prédictive.

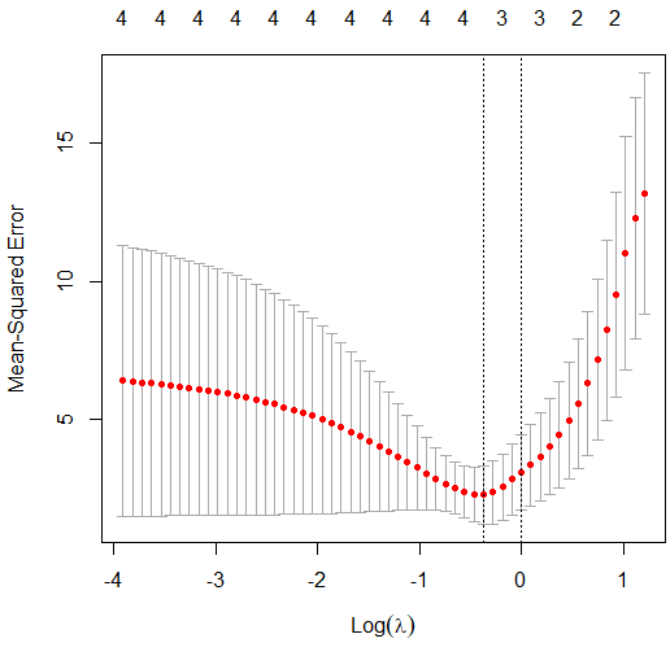

FIG. 3.1 – Lasso - Taux d'erreur en validation croisée vs. log(λ).

Ce graphique représente le tracé de l'erreur de validation croisée (Mean Squared Error ou MSE) en fonction du logarithme de la valeur du paramètre de régularisation lambda (Log(λ)) pour le modèle de régression lasso appliqué aux données de consommation d'essence.

Interprétation des résultats

Une fois que nous avons la valeur optimale de lambda, Le code suivant montre comment trouver les nouveaux coefficients du modèle. .

```
> # Coefficients du modèle avec le meilleur lambda
> lasso_coefficients <- coef(lasso_model, s = best_lambda)
> print(lasso_coefficients)
```

Ces résultats représentent les coefficients du modèle de régression lasso final avec le meilleur lambda sélectionné.

(Intercept)	9.8838710
Prix	.
Cylindrees	0.9734754
Puissance	1.5823385
Poids	0.3049755

L'interprétation de ces coefficients

- (Intercept) =9.8838710 : ceci est la valeur du terme constant (ordonnée à l'origine) dans le modèle. Lorsque toutes les variables explicatives sont nulles, la consommation d'essence prédite est de 9,88 L/100km.
- Prix = . :un point représente un coefficient nul. Cela signifie que la variable Prix a été complètement supprimée du modèle par la pénalité lasso. Elle n'a pas d'effet significatif sur la consommation d'essence selon ce modèle.
- Cylindrées = 0.9734754 : le coefficient positif indique qu'une augmentation de la cylindrée est associée à une augmentation de la consommation d'essence. Plus

précisément, une augmentation de 1 unité de la cylindrée entraîne une hausse de 0,97 L/100km de la consommation, toutes autres variables égales par ailleurs.

- Puissance =1.5823385 : de même, une augmentation de 1 unité de la puissance du moteur est associée à une hausse de 1,58 L/100km de la consommation d'essence.
- Poids=0.3049755 : le coefficient positif du poids indique qu'une augmentation du poids du véhicule entraîne aussi une hausse de la consommation, à hauteur de 0,31 L/100km par unité de poids supplémentaire.

Ainsi, selon ce modèle lasso, les variables Cylindrée, Puissance et Poids ont un effet positif significatif sur la consommation d'essence, tandis que le Prix n'a pas été retenu comme prédicteur pertinent.

Évaluation des performances du modèle

Une fois le modèle final de régression Lasso entraîné sur l'ensemble d'apprentissage avec le lambda optimal, il est essentiel d'évaluer ses capacités prédictives sur un jeu de données indépendant. Cette étape est essentielle pour s'assurer que le modèle est capable de généraliser correctement à de nouvelles données, et non seulement de bien s'ajuster aux données . En effet, un modèle pourrait très bien avoir un excellent coefficient de détermination R^2 sur l'ensemble d'apprentissage, mais se révéler beaucoup moins performant sur de nouvelles observations .

Voici le code pour calculer le coefficient de détermination R^2 sur l'ensemble de test :

```
> r2 <- 1 - sum((y - predictions)^2) / sum((y - mean(y))^2)
> print(paste("R^2 Score :", r2))
```

Ce code calcule le R^2 en comparant les valeurs réelles (y) et les prédictions du modèle sur l'ensemble de test. Le résultat obtenu ici est : "R^2 Score : 0.9679", indiquant que le modèle explique 96,79% de la variance de la variable réponse sur les données de test.

3.2.3 Résultats de la comparaison

Nous allons utiliser le paquet **glmnet** qui est disponible sur **R** pour estimer le modèle. Notre but est de comparer les techniques suivantes :

- la méthode Lasso : $\alpha = 1$;
- la méthode Ridge : $\alpha = 0$;
- la régression multiple (RLM).

Technique	Lasso	Ridge	RLM
Intercept	9.883	9.883	3.137
Prix	.	0.440	3.475
Cylindrees	0.973	0.977	9.121
Puissance	1.582	1.028	2.450
Poids	0.304	0.766	1.676

TAB. 3.3 – Les coefficients pour les différentes techniques : Lasso, Ridge , et RLM.

Remarques

– Lasso tend à produire des modèles plus parcimonieux, ce qui est évident ici avec le coefficient de Prix réduit à zéro. Cela peut aider à simplifier le modèle en éliminant les variables non significatives.

– Ridge régularise tous les coefficients, ce qui les réduit tous en magnitude mais ne les rend pas nuls. Cela aide à éviter le surajustement sans éliminer complètement aucune variable.

– RLM (régression linéaire multiple) donne les coefficients les plus élevés, ce qui peut indiquer un surajustement aux données.

3.2.4 Comparaison les performances des modèles

Technique	R^2	R^2_a
Lasso	0.96	0.94
Ridge	0.97	0.96
RLM	0.94	0.93

TAB. 3.4 – Comparaison des performances des techniques de modélisation.

Ces résultats suggèrent que le choix de la méthode dépend des priorités : pour des prédictions légèrement meilleures, Ridge est préférable ; pour un modèle plus simple et plus facile à interpréter, Lasso est recommandé ; et pour une compréhension complète des effets des variables, la régression linéaire multiple est adéquate.

3.3 Étude comparative sur les données de logement à Boston

Après avoir comparé différentes méthodes de régression sur des données de consommation d'essence, nous allons maintenant appliquer ces techniques sur un jeu de données plus volumineux lié au marché immobilier de Boston[12]. Ce jeu de données contient 506 observations de quartiers de Boston avec 13 variables explicatives, l'objectif étant de prédire le prix médian des maisons. En travaillant sur ces données de plus grande dimension, nous pourrons mieux évaluer les forces de la régression Lasso, notamment sa capacité à gérer un grand nombre de variables potentielles et à sélectionner les prédicteurs les plus pertinents. Cette analyse permettra de compléter nos conclusions précédentes et d'identifier les cas d'utilisation les plus appropriés pour le Lasso, la régression Ridge et la régression linéaire multiple.

3.3.1 Description des Données

Le jeu de données comprend 506 observations, chacune représentant un quartier ou une zone de recensement de Boston[22]. Pour chaque observation, nous disposons de 13 variables explicatives .

Variables explicatives :

- CRIM : taux de criminalité par habitant.
- ZN : proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés.
- INDUS : proportion d'acres commerciales non-retail par ville.
- CHAS : variable binaire pour la rivière Charles (1 si le tract borde la rivière, 0 sinon).
- NOX : concentration d'oxydes d'azote (en parties par 10 millions).
- RM : nombre moyen de pièces par logement.
- AGE : proportion de logements occupés par leur propriétaire construits avant 1940.
- DIS : distances pondérées à cinq centres d'emploi de Boston.
- RAD : indice d'accessibilité aux autoroutes radiales.
- TAX : taux d'imposition foncière pour 10 000$.
- PTRATIO : ratio élèves-enseignants par ville.
- B : 1000(Bk - 0.63)$^{\hat{}}2$ où Bk est la proportion de personnes de couleur par ville.
- LSTAT : pourcentage de la population considérée comme de statut inférieur.

Variable cible :

- MEDV : Valeur médiane des logements occupés par leur propriétaire en milliers de dollars .

Notons $\mathbf{X}$ la matrice des données, elle est $n \times p$ avec $n = 506$ et $p = 14$.

Pour notre étude du marché immobilier de Boston, nous utilisons le jeu de données 'Boston Housing', disponible dans la bibliothèque **MASS** de R. Ce jeu de données,

largement utilisé dans l'analyse statistique, contient des informations sur diverses caractéristiques des quartiers de Boston et les prix médians des maisons. Nous avons importé ces données en utilisant le code suivant :

```
># Charger la bibliothèque nécessaire
>library(MASS)
> data("Boston")
```

3.3.2 Modélisation

Dans cette étude, nous utilisons des techniques de modélisation statistique pour prédire MEDV .

1. Régression Linéaire Multiple

Présentation du modèle. Le modèle mathématique s'écrit comme suit :

$$MEDV = \beta_0 + \beta_1 CRIM + \beta_2 ZN + ... + \beta_{13} LSTAT + \varepsilon$$

Où :

β_0 : est l'intercept (la valeur de MEDV lorsque toutes les variables explicatives sont nulles).

β_1 à β_{13} : sont les coefficients de régression pour chaque variable explicative.

ε : est le terme d'erreur.

Ajustement du modèle Après avoir importé les données de la bibliothèque **MASS**, nous avons procédé à l'ajustement de notre modèle de régression linéaire multiple. Pour ce faire, nous avons utilisé la méthode des moindres carrés ordinaires (MCO), qui vise à minimiser la somme des carrés des résidus, c'est-à-dire la différence entre les valeurs observées et les valeurs prédites par le modèle. L'ajustement a été réalisé

à l'aide du logiciel statistique R, en utilisant la fonction **lm()** de la manière suivante :

```
>lm_model <- lm(medv ~., data = Boston)
```

L'ajustement du modèle nous a fourni des estimations pour les coefficients β_1 à β_{13}, représentant l'effet de chaque variable explicative sur le prix médian des maisons.

2. Régression Ridge

Pour améliorer notre modèle, nous utilisons la régression Ridge, une technique de régularisation qui peut nous aider à gérer la complexité du modèle et la multicolinéarité. Le principe fondamental de la régression Ridge est d'ajouter un terme de pénalité à la fonction de coût de la régression linéaire classique. Cette pénalité est basée sur la somme des carrés des coefficients de régression, pondérée par un paramètre de régularisation, généralement noté λ.

$$\hat{\beta}^{(R)} = \arg\min(\sum_{i=1}^{506}(MEDV_i - (\beta_0 + \beta_1 CRIM_i + \beta_2 ZN_i + ... + \beta_{13} LSTAT_i))^2 + \lambda(\beta_1^2 + \beta_2^2 + ... + \beta_{13}^2)))$$

Ajustement du modèle Pour ajuster notre modèle de régression Ridge et sélectionner la valeur optimale du paramètre λ, nous avons utilisé une approche de validation croisée. Cette méthode nous permet de trouver un équilibre entre la complexité du modèle et sa capacité. Voici le code avec R

```
# Effectuer la validation croisée pour trouver le meilleur lambda
> cv_ridge <- cv.glmnet(X, y, alpha = 0)
> best_lambda <- cv_ridge$lambda.min
> print(paste("Le meilleur lambda est :", best_lambda))
```

Notre analyse de validation croisée a révélé que la valeur optimale du paramètre de régularisation λ pour notre modèle de régression Ridge est de 0,678. nous indique que notre modèle de régression Ridge a trouvé un bon équilibre entre ajustement

aux données et généralisation.

3. Régression Lasso

Maintenant, nous passons à une autre technique de régularisation : la régression Lasso. Elle vise à améliorer la performance du modèle et à gérer la multicolinéarité, mais avec une approche différente de la pénalisation. L'équation pour la régression Lasso est :

$\hat{\beta}^{(R)} = \arg\min(\sum_{i=1}^{506}(MEDV_i - (\beta_0 + \beta_1 CRIM_i + \beta_2 ZN_i + ... + \beta_{13} LSTAT_i))^2 + \lambda(|\beta_1| + |\beta_2| + ... + |\beta_{13}|)))$

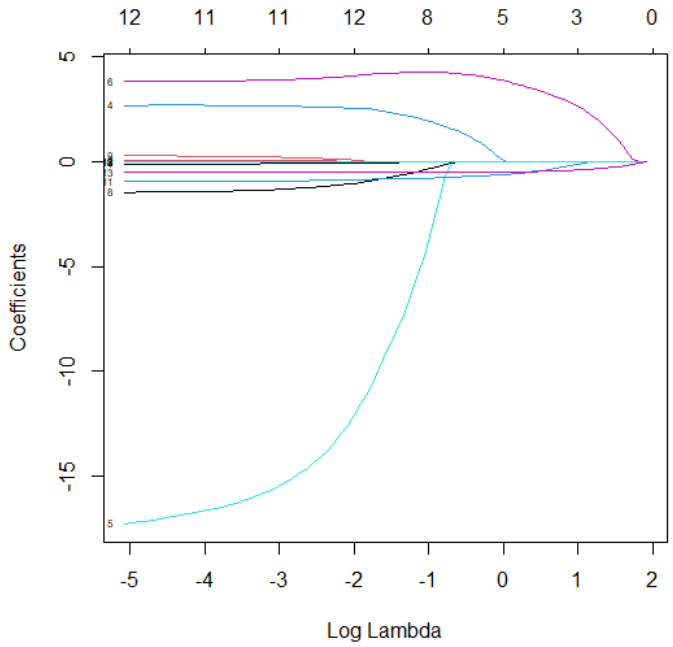

FIG. 3.2 – Lsso path coefficients.

Le graphe 3.2 illustre l'évolution des coefficients d'un modèle Lasso en fonction du paramètre de régularisation λ (en échelle logarithmique). On voit sur le graphe que

les courbes, qui représentent les coefficients, commencent à des valeurs différentes sur la gauche. Lorsque λ augmente (en se déplaçant vers la droite du graphe), ces courbes se rapprochent toutes de zéro. montrant que le Lasso élimine progressivement certaines variables du modèle.

Ajustement du modèle

Pour calculer les coefficients du modèle, nous avons fait une validation croisée pour choisir le coefficient de régularisation λ.qui nous donne l'erreur minimale en utilisant la fonction **cv.glmnet** disponible dans le paquet **glmnet** de **R**. On trouve λ_{min} (c.-à.d. la valeur de λ qui correspond à la plus petite valeur de l'erreur de prédictions).

```
> # Trouver la valeur optimale de lambda qui minimise l'erreur quadratique moyenne
> best_lambda <- cv_model$lambda.min
> print(paste("Meilleur lambda pour Lasso :", best_lambda))
[1] "Meilleur lambda pour Lasso : 0.025517430892623"
```

Ensuite, on utilise la fonction **coef** pour calculer les coefficients qui convient à $\lambda_{\min}$. Les résultats de cette étude sont présentés sous forme de tableau de comparaison.

3.3.3 Résultats de la Modélisation

Le tableau suivant présente les coefficients estimés pour chaque variable explicative selon ces trois méthodes. Cette comparaison nous permettra d'évaluer l'importance relative des différents facteurs influençant les prix des maisons et d'observer comment chaque technique de régression traite ces variables.

Variable	Linéaire Multiple	Ridge	Lasso
Intercept	37.73	28.05	34.64
crim	-0.09	-0.08	-0.09
zn	0.03	0.03	0.04
indus	-0.01	-0.03	.
chas	2.29	2.90	2.6
nox	-17.13	-12.0	-16.41
rm	3.49	4.01	3.85
age	0.009	-0.003	.
dis	-1.39	-1.12	-1.40
rad	0.33	0.15	0.25
tax	-0.012	-0.005	-0.01
ptrati	-0.96	-0.85	-0.93
black	0.009	0.009	0.009
stat	-0.56	-0.47	-0.52

TAB. 3.5 – Tableau Comparatif des Coefficients des Modèles.

Nous remarquons que les coefficients sont relativement proches pour les différentes techniques, ce qui suggère une certaine cohérence dans l'importance relative des variables à travers les modèles. Cependant, il y a des choses importantes à remarquer :

– La technique Lasso a effectivement pénalisé certains coefficients en les réduisant à zéro. Spécifiquement, les variables 'indus' et 'age' ont été exclues du modèle Lasso (représentées par des points dans le tableau). Cela suggère que ces variables sont considérées comme moins importantes.
– Certaines variables comme 'crim', 'nox', 'rm', 'dis', et 'ptratio' ont des coefficients non nuls et relativement importants dans tous les modèles, indiquant leur importance constante pour la prédiction.

– nous observons que les coefficients de Ridge et Lasso sont légèrement réduits par rapport à la régression linéaire multiple, ce qui est l'effet attendu de la régularisation.

Comparaison des Performances

Technique	R^2	R^2ajusté
RLM	0.73	0.72
Ridge	0.73	0.72
LASSO	0.74	0.73

TAB. 3.6 – Comparaison des coefficients de détermination R^2.

Ce tableau 3.6 présente deux mesures de la qualité d'ajustement pour trois modèles différents : la régression linéaire multiple (RLM), la régression Ridge et la régression Lasso. les résultats montrent des performances très similaires en termes de R^2 et R^2 ajusté, avec un léger avantage pour le Lasso.

3.4 Discussion et Conclusion

Notre étude comparative des méthodes de régression a révélé les atouts spécifiques de chaque approche. Le Lasso s'est distingué par sa capacité à améliorer l'interprétabilité des modèles tout en maintenant une bonne performance prédictive, particulièrement efficace pour la sélection de variables dans les contextes de haute dimension. Il a démontré sa stabilité face à la multicolinéarité et sa polyvalence dans divers domaines d'application.

La régression Ridge excelle dans la gestion de la multicolinéarité, tandis que la régression linéaire multiple reste essentielle pour comprendre les relations fondamentales entre variables. Le choix de la méthode dépend du contexte spécifique de l'étude,

des caractéristiques des données et des objectifs d'analyse, soulignant l'importance d'une approche adaptée en modélisation statistique pour obtenir des résultats fiables et pertinents.

Conclusion générale

En conclusion, cette introduction au Lasso et aux méthodes de régularisation L1 a mis en lumière leur importance cruciale dans l'analyse statistique moderne. Le Lasso, en combinant régression et sélection de variables, offre une solution élégante aux défis posés par les ensembles de données de haute dimensionnalité.

Ce travail nous a permis d'acquérir de nouvelles connaissances et d'approfondir d'autres, tant sur le plan théorique que pratique. Nous avons pu explorer les concepts fondamentaux du Lasso dans un contexte théorique tout en appliquant ces concepts à des problèmes pratiques, enrichissant ainsi notre compréhension et notre expertise dans le domaine des méthodes de régularisation.

L'étude comparative entre la régression linéaire multiple, Ridge et Lasso a mis en évidence les avantages spécifiques de chaque approche, soulignant la capacité du Lasso à produire des modèles parcimonieux et interprétables. Cependant, nous avons également reconnu ses limites, notamment sa sensibilité au choix du paramètre de régularisation et son instabilité potentielle face à des prédicteurs fortement corrélés.

Le Lasso et les méthodes de régularisation L1 ont profondément influencé notre approche de l'analyse de données et de la modélisation prédictive. Leur évolution continue promet d'ouvrir de nouvelles perspectives dans le domaine de l'analyse prédictive et de la compréhension des systèmes complexes, renforçant leur importance fondamentale dans l'ensemble des méthodes statistiques avancées.

Bibliographie

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.),Second International Symposium on Information Theory, (pp. 267–281). Academiai Kiado : Budapest.

[2] Andersen, R. (2009). Nonparametric methods for modeling nonlinearity in regression analysis. Annual Review of Sociology, 35, 67-85.

[3] Berkani, F. (2016). Application de la régression linéaire multiple sur la balance commerciale algérienne , Université Kasdi Merbah Ouargla.

[4] Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics, 37(4), 373-384.

[5] Chatterjee, S., & Hadi, A. S. (2015). Regression analysis by example. John Wiley & Sons.

[6] Confais, J., & Le Guen, M. (2007). Premiers pas en régression linéaire . Documents de travail du Centre d'Économie de la Sorbonne.

[7] Cornillon, P. A., & Matzner-Lober, E. (2011). Régression avec R (pp. 157-168). France : Springer.

[8] Cornillon, P. A., & Matzner-Løber, E. (2007). La régression linéaire multiple. Régression : Théorie et applications, 33-50.

[9] De Micheaux, P. L., Drouilhet, R., & Liquet, B. (2014). Régression linéaire simple et multiple. In Le logiciel R (pp. 489-540). Springer, Paris.

[10] Dris, L., & Hachemi, W. (2016). Régression linéaire multiple et modèle linéaire général , Université Abderrahmane Mira - Béjaia.

[11] Dodge, Y., & Rousson, V. (2004). Analyse de régression appliquée (2ème éd.). Dunod.

[12] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA : University of California, School of Information and Computer Science.

[13] Dubois, J. F. (2003). Quelques pièges cachés des méthodes de sélection de variables en régression linéaire multiple. Bibliothèque nationale du Canada, Ottawa.

[14] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., & et al. (2004). Least angle regression. The Annals of Statistics, 32(2), 407-499.

[15] Elftouh, N., & Froda, S. (2008). Étude de tests de permutation en régression multiple. Université du Québec à Montréal.

[16] Frank, I., & Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35, 109-148.

[17] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.

[18] Gruber, M. H. (2010). Regression estimators : A comparative study. JHU Press.

[19] Guyader, A. (2011). Régression linéaire. Université Rennes, 2, 60-61.

[20] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning : data mining, inference, and prediction (Vol. 2, pp. 1-758). New York : Springer.

[21] Hastie, T. J., & Pregibon, D. (2017). Generalized linear models. In Statistical models in S (pp. 195-247). Routledge.

[22] Harrison, D., & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81-102.

[23] Hebiri, M. (2009). Quelques questions de sélection de variables autour de l'estimateur LASSO (Doctoral dissertation, Université Paris-Diderot-Paris VII).

[24] Horel, A. E. (1964). Ridge analysis. Chemical Engineering Progress Symposium Series 60, 67-77.

[25] Hoerl, A. E. and Kennard, R. W. (1968). On regression analysis and biased estimation. Technometrics 10, 422-423. Abstract.

[26] Hoerl, A., Kennard, R., & Baldwin, K. (1975). Ridge regression : Some simulations. Communication in Statistics - Theory and Methods, 4(1), 105-123.

[27] Hoerl, A., & Kennard, R. (1970). Ridge regression : Biased estimates for nonorthogonal problems. Technometrics, 12(1), 55-67.

[28] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York : Springer.

[29] James, W., & Stein, C. (1961). Estimation with quadratic loss. In 4th Berkeley Symposium, Vol. 1 (pp. 361-379).

[30] Jobson, J. D., & Jobson, J. D. (1991). Multiple linear regression. In Applied multivariate data analysis : regression and experimental design (pp. 219-398).

[31] Kharoubi, R. (2016). Une nouvelle approche pour la sélection des variables dans le cas de modèles de discrimination en grandes dimensions (mémoire de maîtrise). Université de Québec à Montréal.

[32] Kerroum, K. (2016). Régression multiple et ridge (mémoire de master) , Université Dr Tahar Moulay - Saïda.

[33] Kumar, D. (2019). Ridge regression and Lasso estimators for data analysis.

[34] Ihaka, R. and Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 : 299-314.

[35] Mallows C.L., 1964. Choosing variables in a linear regression : A graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

[36] Mallows, C. L. (1973). Some comments on Cp. Technometrics, 15(4), 661-675.

[37] Manski, C. F. (1991). Regression. Journal of Economic Literature, 29(1), 34-50.

[38] Montgomery, D. C., & Peck, E. A. (1982). Introduction to linear regression analysis. New York : John Wiley and Sons.

[39] Mood, A. M. (1950). Introduction to the Theory of Statistics.

[40] O'Brien, C. M. (2016). Statistical learning with sparsity : the Lasso and generalizations.

[41] Ounaissi, D. (2016). Méthodes quasi-Monte Carlo et Monte Carlo : Application aux calculs des estimateurs Lasso et Lasso bayésien (Doctoral dissertation, Thesis).

[42] Palm, R., & Iemma, A. F. (1995). Quelques alternatives à la régression classique dans le cas de la colinéarité. Revue de statistique appliquée, 43(2), 5-33.

[43] Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348.

[44] Rakotomalala, R. (2011). Pratique de la régression linéaire multiple. Diagnostic et sélection de variables.

[45] Rousson, V. (2013). Régression linéaire multiple. In Statistique appliquée aux sciences de la vie (pp. 219-258). Springer, Paris.

[46] Sabatier, R., Reynes, C., & Vivien, M. (2016). Grain 7 : Régression linéaire. Chemoocs, Session, 1.

[47] Salifou, M., & Houessou, M. (2020). Apprentissage statistique du modèle Cox-logistique : application à la survie des enfants de moins de 5 ans au Bénin. Journal of Mathematics and Statistics Studies, 1(2), 1-14.

[48] Seber, G. A., & Lee, A. J. (2012). Linear regression analysis. John Wiley and Sons.

[49] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society : Series B (Methodological), 58(1), 267-288.

[50] Tillé, Y. (2011). Résumé du cours de modèles de régression. Institut de statistique, Université de Neuchâtel, Suisse.

[51] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5.

[52] Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, 234-240.

[53] Watts, Y. (2023). Le Lasso linéaire : une méthode pour des données de petites et grandes dimensions en régression linéaire.Université de Montréal.

[54] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley and Sons.

[55] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 68(1), 49-67.

[56] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 67(2), 301-320.

Annexe A : Logiciel R

3.5 Qu'est-ce-que le langage R ?

• Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$E(.)$	:	Espérence mathématique.
$V(.)$	:	Variance mathématique.
$Cov(X,Y)$	:	Covariance mathématique du couple (X,Y).
$B(.)$	:	Biais.
$EQM(.)$	:	Erreur quadratique moyenne.
$Tr(.)$	:	La trace de la matrice.
SCT	:	La somme des carrés totaux.
SRC	:	La somme résiduelle des carrés.
SCE	:	La somme des carrés expliqués.
MCO	:	moindres carrés ordinaires.
RLM	:	Régression linéaire multiple.
$LASSO$	:	**L**east absolute shrinkage and selection operator.
$L1$	:	La norme $\|\|.\|\|$.
$L2$	:	la norme $\|\|.\|\|^2$.
$BLUE$	:	Best Linear Unbiased Estimator.

Résumé

Ce mémoire explore l'utilisation de la régularisation L1, notamment le Lasso (Least Absolute Shrinkage and Selection Operator), dans le cadre des modèles de régression. L'objectif principal est de présenter et d'analyser les techniques de régularisation L1 pour améliorer la performance prédictive et l'interpértabilité des modèles de régression, en particulier dans des contextes de haute dimension. Nous comparons la méthode de régression du Lasso aux méthodes de régression linéaire multiple et Ridge à travers des exemples pratiques.

Mots-clés : Lasso, sélection de variables, surajustement, régularisation L1, Modèles Statistiques.

Abstract

This work explores the use of L1 regularization, specifically the Lasso (Least Absolute Shrinkage and Selection Operator), in regression models. The main objective is to present and analyze L1 regularization techniques to improve the predictive performance and interpretability of regression models, especially in high-dimensional contexts. We compare the LASSO Method to a general linear regression and Ridge methods through practical examples.

Keywords: Lasso, variable selection, overfitting, L1 regularization, statistical models.