République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieure et de la recherche scientifique

N° Attribué par la bibliothèque

Année Univ.: 2015/2016

Régression multiple et ridge

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université Dr Tahar Moulay - Saïda
Discipline : MATHEMATIQUES
Spécialité :Analyse Stochastique Statistique des Processus et Applications

par

Kerroum **K**arima [1]

Sous la direction de

Dr. F.Mokhtari

Soutenue le 1 Juin 2016 devant le jury composé de

A.Kandouci	Maître de conférences	Université Dr Tahar Moulay - Saïda	Président
F.Mokhtari	Maître de conférences	Université Dr Tahar Moulay - Saïda	Rapporteur
I.Ouassou	Professeur	Université Kadi Ayad - Marrakech	Co-encadreur
F.Benziadi	Maître de conférences	Université Dr Tahar Moulay - Saïda	Examinatrice
F.Maref	Maître de conférences	Université Dr Tahar Moulay - Saïda	Examinatrice

1. e-mail :kerroumkarima67@gmail.com

Remerciements

Grace à "Dieu" le tout puissant, j'ai pu achever ce mémoire.
Mes vifs remerciements accompagnés de ma gratitude vont tout d'abord à mon encadreur madame F.Mokhtari, sans oublier mon co-encadreur monsieur Idir Ouassou, docteur en mathématiques à l'université de Marrakech, pour avoir proposé ce sujet et dirigé mon travail, et qui a mis à ma disposition son précieux temps, ses connaissances scientifiques, et sa marque de confiance durant mon séjour au laboratoire du mathématiques à Marrakech.

Mes remerciements vont également à monsieur A.Kandouci pour son aide précieuse, et tous les enseignants qui ont contribué de prés ou de loin à ma formation à la faculté des sciences à l'université Docteur Moulay Tahar Saida.

Table des matières

Résumé

La littérature statistique a plusieurs méthodes pour faire à la multicolinéarité. Ce memoire présente un nouvel estimateur appelé estimateur modifié du ridge sans biais (MUR), cet estimateur est obtenu à partir de l'estimateur ridge sans biais (URR), De la même manière que l'estimateur ridge ordinaire (ORR) est obtenu à partir des moindres carrés ordinaire (MCO), on compare ensuite l'estimateur ridge modifie (MUR) par l'estimateur ridge ordinaire (ORR) et l'estimateur (URR) en terme d'erreur quadratique moyenne MMSE.

Abstract

Statistical litérature has several methods for coping with multicollinearity. We introduce in this memoir a new shrinkage estimator, called modified unbiased ridge (MUR). This estimator is obtained from unbiased ridge regression (URR) in the same way that ordinary ridge regression (ORR) is obtained from ordinary least squares (OLS). Properties of MUR are derived. Results on its matrix mean squared error (MMSE) are obtained. MUR is compared with ORR and URR in terms of MMSE.

Title

Least squares régression and ridge régression.

Mots clés.

Estimateur des moindres carrés ordinaires, estimateur ridge ordinaire, estimateur ridge sans biais, estimateur ridge modifié.

Key Words.

Ordinary least squares estimator, ordinary ridge estimator, unbiased ridge estimator, modified unbiased ridge estimator.

Introduction

Le terme "régression" a une origine curieuse. Il remonte à l'étude du physiologiste et anthropologue Francis Galton (vers la fin du XIX-éme siècle) sur la relation entre la taille des parents et celle des enfants. Galton a introduit le mot régression pour désigner, dans des problèmes d'héréditè, la diminution progressive (régression) des écarts par rapport à la moyenne, d'une génération à la suivante. Le terme de régression a ensuite été utilisé d'une manière tout à fait générale, sans aucune relation avec sa signification initiale.

La régression linéaire est l'un des modèles statistiques les plus employés, son champ d'application s'étend de la description et de l'analyse des données expérimentales jusqu'à la prévision, et il est aussi utilisé pour l'interpolation. Une situation courante en sciences biologiques est d'avoir à disposition deux ensembles de données de taille n, $y_1, ..., y_n$ et $x_1, ..., x_n$, obtenus exprémentalement et mesurés sur une population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme y=f(x). Lorsque la relation recherchée est affine c'est à dire de la forme y=ax+b, on parle de régression linéaire.

Ce memoire comporte trois chapitres. Dans le premier chapitre on introduit les modèles linéaires simples, en commençant par donner un exemple sur la pollution d'air. On donne l'estimateur des moindres carrés de cette régression qui permet de chercher l'éventuelle relation fonctionnelle linéaire qui existerait entre une valeur explicative (ou indépendante) x et une variable aléatoire à expliquer (ou dépendante) y, aprés avoir expliciter les hypothèses nécessaires du modèle, on donnes quelques notions d'estimation des paramètres, une prévision par intervalle de confiance.

Le deuxième chapitre est consacré à la régression linéaire multiple, l'outil statistique le plus habituellement mis en ouvre par l'étude des données multidimensionnelles, cas particulier du modèle linéaire, il constitue la généralization naturelle de la régression simple.

Dans le dernier chapitre, on commence par étudier l'estimateur ridge ordinaire, sa représentation dans le modèle de base et sa représentation dans le modèle canonique. Des propriétés conçernant ce processus sont établies, telles que l'espérance et la variance. En fin une comparaison entre cet estimateur et l'estimateur MCO. On examine ensuite un nouvel estimateur appelé (MUR), estimateur modifié du ridge sans biais obtenu à partir d'une modification faite sur l'estimateur ridge sans biais (URR). De la même façon que le premier modèle le, on étudie ces propréités et on le compare avec d'autres estimateurs au sens d'erreur quadratique.

Chapitre 1

La régression linéaire simple

Dans ce chapitre, nous allons analyser la régression linéaire simple : nous pouvons la voir comme une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y). Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

1.1 Exemple : la pollution de l'air

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études epidemiologiques ont permis de mettre en evidence l'influence sur la santé de certains composes chimiques comme le dioxyde de souffre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3) ou des particules sous forme de poussières contenues dans l'air. L'influence de cette pollution est notable sur les personnes sensibles (nouveau-nés, asmatiques, personnes ageés ...). La prevision des pics de concentration de ces composes est donc importante.
Nous allons nous interesser plus particulièrement à la concentration en ozone. Nous possédons quelques connaissances à priori sur la manière dont se forme l'ozone, grace aux lois régissant les équilibres chimiques. La concentration de l'ozone sera fonction de la température ; plus la température sera élevée, plus la concentration en ozone va augmenter. Cette relation très vague doit être améliorée afin de pouvoir prédire les pics d'ozone.

Afin de mieux comprendre ce phénomène, l'association Air Breizh (surveillance de la qualité de l'air en Bretagne) mesure depuis 1994 la concentration en O_3 (en $\mu g/ml$) toute les 10 minutes et obtient donc le maximum journalier de cette concentration. Air Breizh collecte également à certaines heures de la journée des données mèteorologiques comme la température, la nebulosite, le vent...
Le tableau suivant donne les 10 données journalières de température et d'Ozone.

Individu	1	2	3	4	5	6	7	8	9	10
T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

Nous allons donc chercher, en particulier, à savoir si on peut expliquer le taux maximal d'ozone de la journée par la température T_{12} à $12h$. D'un point de vue pratique le but de cette régression est double
– **Ajuster un modèle pour expliquer O_3 en fonction de $T12$;**
– **Prédire les valeurs d'O_3 pour de nouvelles valeurs de $T12$.**
Pour analyser la relation entre les x_i (température) et les y_i (ozone), nous allons chercher une fonction f telle que :

$$y_i \approx f(x_i).$$

Pour préciser le sens de $\approx$ il va falloir se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. Il faudra aussi se donner une classe de fonctions F dans laquelle nous supposerons que se trouve la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\arg\min_{f \in F} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

ou n représente le nombre de données à analyser et $l(\cdot)$ est appelée fonction de coût ou fonction de perte.

1.2 Modélisation mathématique

Etant donnés les points (x_i, y_i), le but est de trouver une fonction affine f telle que la quantité

$$\sum_{i=1}^{n} l(y_i - f(x_i))$$

soit minimale. Pour pouvoir déterminer f, il faut préciser la fonction du coût l. Deux fonctions sont classiquement utilisées :

– le coût absolu $f(u) = |u|$;

– le coût quadratique $f(u) = u^2$.

Ces fonctions sont positives, symétriques, elles donnent donc la même valeur lorsque l'erreur est positive ou négative et s'annulent lorsque u vaut zéro. La fonction du coût l peut aussi être vue comme la distance entre une observation (x_i, y_i) et son point correspondant sur la droite $(x_i, f(x_i))$.

Le coût quadratique est le coût le plus souvent utilisé, ceci pour plusieurs raisons : historique, calculabilité, et propriétés mathématiques. En 1800, il n'existait pas d'ordinateur et l'utilisation du coût quadratique permettait de calculer explicitement les estimateurs à partir des données. A propos de l'utilisation d'autres fonctions de coût, voici ce que disait Gauss (1809) : " Mais de tous ces principes, celui des moindres carrés est le plus simple par contre avec les autres, nous serions conduits aux calculs les plus complexes".

En conclusion, seul le coût quadratique sera automatiquement utilisé dans la suite.

1.3 Modélisation statistique

Dans de nombreuses situations, une idée naturelle est de supposer que la variable à expliquer Y est une fonction affine de la variable explicative X, c'est-à-dire de chercher f dans l'ensemble F des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}$. Lorsque nous ajustons par une droite les données, nous supposons implicitement qu'elles étaient de la forme

$$Y = \beta_1 + \beta_2 X.$$

Dans l'exemple de l'ozone, nous supposons donc un modèle où la concentration d'ozone dépend linéairement de la température. Nous savons pertinemment que toutes

les observations mesurées ne sont pas sur la droite. D'une part, il est irréaliste de croire que la concentration de l'ozone dépend linéairement de la température et de la température seulement. D'autre part, les mesures effectuées dépendent de la précision de l'appareil de mesure, de l'opérateur et il arrive souvent que, pour des valeurs identiques de la variable X, nous observions des valeurs différentes pour Y.
Nous supposons alors que la concentration d'ozone dépend linéairement de la température mais cette liaison est perturbée par un "bruit". C'est le principe de la régression linéaire simple. On suppose disposer dans la suite de n points (x_i, y_i) dans le plan.

Définition 1.3.1. *(Modèle de régression linéaire simple) Un modèle de régression linéaire simple est défini par une équation de la forme :*

$$\forall i \in \{1, \ldots, n\} \quad y_i = \beta_1 + \beta_2 x_i + \varepsilon_i. \tag{1.1}$$

Remarque 1.3.1. *Les quantités ε_i viennent du fait que les points ne sont jamais parfaitement alignés sur une droite. On les appelle erreurs (ou bruits) et elles sont supposées aléatoires.*

Pour pouvoir dire des choses pertinentes sur ce modèle, il faut néanmoins imposer deux hypothèses les concernants :

$$\mathcal{H} \begin{cases} \mathcal{H}_1 : & E[\varepsilon_i] = 0 & \text{pour tout indice } i \\ \mathcal{H}_2 : & Cov(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2 & \text{pour tout couple } (i,j). \end{cases}$$

Les erreurs sont donc supposées centrées, de même variance (homoscedasticite) et non corrélées entre elles.
Notons que le modèle de régression linéaire simple de la définition 1.3.1 peut encore s'écrire de la façon vectorielle suivante :

$$Y = \beta_1 I + \beta_2 X + \varepsilon, \tag{1.2}$$

où :
– le vecteur Y est aléatoire de dimension n,
– le vecteur I est le vecteur de $\mathbb{R}^n$ dont les composantes valent 1,
– le vecteur X est un vecteur de dimension n donné (non aléatoire),
– les coefficients β_1 et β_2 sont les paramètres inconnus du modèle,

– le vecteur ε est aléatoire de dimension n.
Cette notation vectorielle sera commode notamment pour la représentation et l'interprétation géometrique du problème en régression linéaire multiple.

Remarque 1.3.2. *Afin d'estimer les paramètres inconnus du modèle, nous mesurons dans le cadre de la régression simple une seule variable explicative ou variable* **exogène** X *et une variable à expliquer où variable* **endogène** Y*, La variable* X *est souvent considérée comme non aléatoire au contraire de* Y*.*

1.4 Estimateur des Moindres Carrés Ordinaire

Définition 1.4.1 (Estimateurs des Moindres Carrés Ordinaires-MCO)**.** *On appelle estimateurs des Moindres Carrés Ordinaires (en abrégé MCO)* $\hat{\beta}_1$ *et* $\hat{\beta}_2$ *les valeurs minimisant la quantité :*

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Remarque 1.4.1. *La fonction de deux variables* S *est une fonction quadratique et sa minimisation ne pose aucun problème.*

1.5 Calcul des estimateurs de β_1 et β_2

Proposition 1.5.1 (Estimateurs des MCO $\hat{\beta}_1$ et $\hat{\beta}_2$)**.** *Les estimateurs des MCO ont pour expressions :*

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.3}$$

avec

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1.4}$$

Preuve :
La fonction $S(\beta_1, \beta_2)$ est strictement convexe, elle admet donc un minimum unique au point $(\hat{\beta}_1, \hat{\beta}_2)$, lequel est déterminé en annulant les dérivées partielles de S. On

obtient les équations suivantes :

$$\begin{cases} \frac{\partial S}{\partial \beta_1} = & -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) & = & 0 \\ \frac{\partial S}{\partial \beta_2} = & -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) & = & 0, \end{cases}$$

la première équation donne :

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

d'ou l'on déduit immédiatement :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.5}$$

ou $\bar{x}$ et $\bar{y}$ sont comme d'habitude les moyennes empiriques des x_i et des y_i.
La seconde équation donne :

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

et en remplaçant $\hat{\beta}_1$ par son expression (1.5), nous avons :

$$(\bar{y} - \hat{\beta}_2 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

donc

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x}} = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}. \tag{1.6}$$

Cette dernière équation suppose que le dénominateur $\sum_{i=1}^{n}(x_i - \bar{x})^2$ est non nul. Or ceci ne peut arriver que si tous les x_i sont égaux, situation sans interêt pour notre problème et que nous excluons donc à priori pour toute la suite.

Remarque 1.5.1. *La relation $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$, montre que la droite des MCO passe par le centre de gravite du nuage $(\bar{x}, \bar{y})$.*

1.6 Quelques propriétés des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$

Sous les seules hypothèses ($\mathcal{H}1$) et ($\mathcal{H}2$) de centrages, décorrelations et homoscedasticités des erreurs ε_i du modèle, on peut déjà donner certaines propriétés statistiques des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ des moindres carrés.

Théorème 1.6.1 (Estimateurs sans biais de β_1 et β_2)**.** *$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de β_1 et β_2.*

Preuve :
Une autre façon d'écrire $\hat{\beta}_2$ est :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

il suffit de remplacer y_i par $\beta_1 + \beta_2 x_i + \varepsilon_i$ et $\bar{y}$ par $\beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}$ dans l'equation (1.4). Dans cette expression, seuls les bruits ε_i sont aléatoires, et puisqu'ils sont centrés, on en déduit bien que

$$E[\hat{\beta}_2] = \beta_2$$

Pour $\hat{\beta}_1$, on part de l'expression (1.3) :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \quad \bar{y} = \beta_1 + \beta_2 \bar{x} + \bar{\varepsilon} \quad et \quad E[\bar{\varepsilon}] = 0$$

on obtient alors :

$$E[\hat{\beta}_1] = E[\bar{y}] - \bar{x}E[\hat{\beta}_2] = \beta_1 + \bar{x}\beta_2 - \bar{x}\beta_2 = \beta_1.$$

On peut également exprimer les variances et les covariances de nos estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$.

Théorème 1.6.2 (Variances et covariance)**.** *Les variances des estimateurs sont :*

$$Var(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \quad et \quad Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

tandis que leur covariance vaut :

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

.

Preuve :
On part à nouveau de l'expression de $\hat{\beta}_2$ utilisée dans la preuve du non-biais :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où les erreurs ε_i sont décorrélées et de même variance σ^2 donc la variance de la somme est la somme des variances :

$$Var\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Par ailleurs, la covariance entre $\bar{y}$ et $\hat{\beta}_2$ s'écrit :

$$Cov\left(\bar{y}, \hat{\beta}_2\right) = Cov\left(\frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

d'où il vient pour la variance de $\hat{\beta}_1$:

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\bar{y} - \hat{\beta}_2\bar{x}\right) \\
&= Var\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) + \bar{x}^2 Var\left(\hat{\beta}_2\right) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_2)) \\
&= Var\left(\frac{\sum_{i=1}^{n} \beta_1 + \beta_2 x_i + \varepsilon_i}{n}\right) + \bar{x}^2 Var\left(\hat{\beta}_2\right) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

c'est-à-dire :

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Enfin, pour la covariance des deux estimateurs :

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(\bar{y} - \hat{\beta}_2\bar{x}, \hat{\beta}_2) = Cov(\bar{y}, \hat{\beta}_2) - \bar{x}Var(\hat{\beta}_2)) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Remarque 1.6.1. *On a vu que la droite des MCO passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$. Supposons celui-ci fixe et $\bar{x}$ positif, alors il est clair que si on augmente la pente, l'ordonnée à l'origine va baisser et vice versa, on retrouve donc bien le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.*

Les estimateurs des moindres carrés sont en fait optimaux en un certain sens, c'est ce que precise le théorème suivant :

Théorème 1.6.3 (Gauss-Markov)**.** *Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ pour $j = 1, 2$ sont de variance minimale.*

Preuve :
L'estimateur des MCO s'écrit $\hat{\beta}_2 = \sum_{i=1}^n p_i y_i$, avec $p_i = (x_i - \bar{x})/\sum_{i=1}^n (x_i - \bar{x})^2$. Considérons un autre estimateur δ_2 linéaire en y_i et sans biais, c'est-a-dire :

$$\delta_2 = \sum_{i=1}^n \lambda_i y_i.$$

Montrons que $\sum_{i=1}^n \lambda_i = 0$ et $\sum_{i=1}^n \lambda_i x_i = 1$, L'égalité

$$\begin{aligned} E[\delta_2] &= \beta_1 \sum_{i=1}^n \lambda_i + \beta_2 \sum_{i=1}^n \lambda_i x_i + \sum_{i=1}^n \lambda_i E[\varepsilon_i] \\ &= \beta_1 \sum_{i=1}^n \lambda_i + \beta_2 \sum_{i=1}^n \lambda_i x_i \\ &= \beta_2 \end{aligned}$$

est vraie pour tout β_1 et β_2 car l'estimateur δ_2 est sans biais, donc $E[\delta_2] = \beta_2$ pour tout β_2, c'est-à-dire que $\sum_{i=1}^n \lambda_i = 0$ et $\sum \lambda_i x_i = 1$.

Montrons que $Var(\delta_2) \geq Var(\hat{\beta}_2)$.

$$Var(\delta_2) = Var\left(\delta_2 - \hat{\beta}_2 + \hat{\beta}_2\right) = Var\left(\delta_2 - \hat{\beta}_2\right) + Var\left(\hat{\beta}_2\right) + 2Cov\left(\delta_2 - \hat{\beta}_2, \hat{\beta}_2\right).$$

$$Cov\left(\delta_2 - \hat{\beta}_2, \hat{\beta}_2\right) = Cov\left(\delta_2, \hat{\beta}_2\right) - Var\left(\hat{\beta}_2\right) = \frac{\sigma^2 \sum_{i=1}^n \lambda_i (xi - \bar{x})}{\sum_{i=1}^n (xi - \bar{x})^2} - \frac{\sigma^2}{\sum_{i=1}^n (xi - \bar{x})^2} = 0,$$

d'où :

$$Var(\delta_2) = Var(\hat{\beta}_2 - \hat{\beta}_2) + Var(\hat{\beta}_2)$$

comme la variance est toujours positive, donc :

$$Var(\delta_2) \geq Var(\hat{\beta}_2)$$

D'où le résultat. On obtiendrait la même chose pour $\hat{\beta}_1$.

1.7 Résidus et variance résiduelle

Dans $\mathbb{R}^2$ (espace des variables x_i et y_i), $\hat{\beta}_1$ est l'ordonnée à l'origine et $\hat{\beta}_2$ la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée. Les résidus sont définis par :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i = y_i - \bar{y} - \hat{\beta}_2 (x_i - \bar{x}). \tag{1.7}$$

Par construction, la somme des résidus est nulle :

$$\sum_{i=1}^n \hat{\varepsilon}_i = \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i) = \sum_{i=1}^n (y_i - \bar{y}) - \hat{\beta}_2 \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Remarque 1.7.1. *Les variances et la covariance des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ établies dans la section précédente ne sont pas pratiques, car elles font intervenir la variance σ^2 des erreurs, laquelle est en général inconnue. On peut en exprimer un estimateur sans biais grace aux residus.*

Théorème 1.7.1 (Estimateur non biaisé de σ^2)**.** *La statistique $\sigma^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.*

Preuve :
En écrivant les résidus nous constatons que : $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$ et $\beta_1 = \bar{y} - \beta_2\bar{x} - \bar{\varepsilon}$, ce qui donne :

$$\begin{aligned}\hat{\varepsilon}_i &= \beta_1 + \beta_2 x_i + \varepsilon_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \\ &= \bar{y} - \beta_2\bar{x} - \bar{\varepsilon} + \beta_2 x_i + \varepsilon_i - \bar{y} + \hat{\beta}_2\bar{x} - \hat{\beta}_2 x_i \\ &= (\beta_2 - \hat{\beta}_2)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}).\end{aligned}$$

En développant et en nous servant de l'écriture vue plus haut :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

nous avons :

$$\begin{aligned}\sum_{i=1}^n \hat{\varepsilon}_i^2 &= (\beta_2 - \hat{\beta}_2)^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 + 2(\beta_2 - \hat{\beta}_2)\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \\ &= (\beta_2 - \hat{\beta}_2)^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - 2(\beta_2 - \hat{\beta}_2)^2 \sum_{i=1}^n (x_i - \bar{x})^2.\end{aligned}$$

Prenons l'espérance :

$$E\left[\sum_{i=1}^n \hat{\varepsilon}_i^2\right] = E\left[\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2\right] - \sum_{i=1}^n (x_i - \bar{x})^2 Var(\hat{\beta}_2) = (n-2)\sigma^2.$$

Remarque 1.7.2. *Bien sur, lorsque n est grand, cet estimateur diffère très peu de l'estimateur empirique de la variance des résidus.*

1.8 Interprétations géométriques

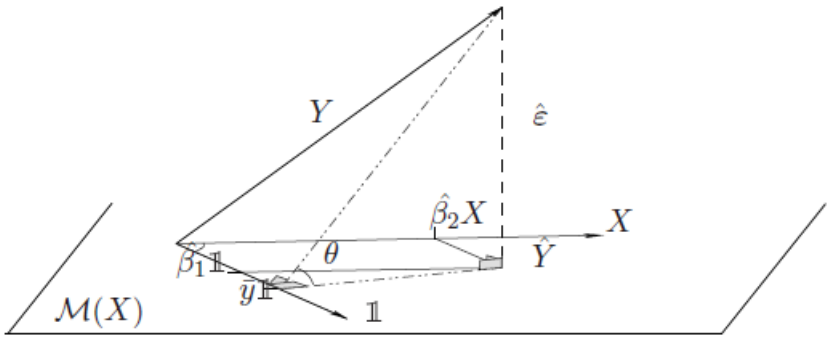


Figure 2.1 - Représentation de la projection dans l'espace des variables.

Si nous abordons le problème d'un point de vue vectoriel, nous avons deux vecteurs à notre disposition : le vecteur $X = [x_1, \ldots, x_n]^t$ des n observations pour la variable explicative et le vecteur $Y = [y_1, \ldots, y_n]^t$ des n observations pour la variable à expliquer. Ces deux vecteurs appartiennent au même espace $\mathbb{R}^n$. Si on ajoute à cela le vecteur $1 = [1, \ldots, 1]^t$, on voit tout d'abord que par l'hypothèse selon laquelle tous les x_i ne sont pas égaux, les vecteurs 1 et X ne sont pas colinéaires : ils engendrent donc un sous-espace de $\mathbb{R}^n$ de dimension 2, noté $\mathcal{M}(X)$. On peut projeter orthogonalement le vecteur Y sur le sous-espace $\mathcal{M}(X)$, notons provisoirement $\tilde{Y}$ cette projection : Puisque $(1, X)$ forme une base de $\mathcal{M}(X)$, il existe une unique décomposition de la forme $\tilde{Y} = \tilde{\beta}_1 1 + \tilde{\beta}_2 X$. Par définition de la projection orthogonale, $\tilde{Y}$ est défini comme l'unique vecteur de $\mathcal{M}(X)$, minimisant la distance euclidienne $||Y - \tilde{Y}||$, ce qui revient au même que de minimiser son carré. Où on a :

$$||Y - \tilde{Y}||^2 = \sum_{i=1}^{n} (y_i - (\tilde{\beta}_1 + \tilde{\beta}_2 x_i))^2,$$

ce qui nous ramène à la methode des moindres carrés ordinaires. On en déduit que $\tilde{Y} = \hat{Y}$, $\hat{\beta}_1 = \tilde{\beta}_1$ et $\hat{\beta}_2 = \tilde{\beta}_2$, avec les expressions de $\hat{Y}$, $\hat{\beta}_1$ et $\hat{\beta}_2$ vues précédemment. Autrement dit, dans $\mathbb{R}^n$, $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme les coordonnées de la projection orthogonale $\hat{y}$ de y sur le sous-espace de $\mathbb{R}^n$ engendré par 1 et x.

Remarque 1.8.1. *Nous avons supposé que* 1 *et* x *est colinéaires. En général, ces vecteurs ne sont pas orthogonaux, ce qui implique que* $\hat{\beta}_1 1$ *n'est pas la projection de* y *sur* 1 *et que* $\hat{\beta}_2 x$ *n'est pas la projection de* y *sur* x*.*

1.8.1 Le coefficient de détermination R^2

Nous conservons les notations du paragraphe précédent, en notant

$$\hat{Y} = [\hat{y}_1, \ldots, \hat{y}_n]^t$$

la projection orthogonale du vecteur Y sur $\mathcal{M}(X)$ et

$$\hat{\varepsilon} = Y - \hat{Y} = [\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n]^t$$

le vecteur des résidus déjà rencontrés. Le théorème de Pythagore donne alors directement :

$$||Y - \bar{y}1||^2 = ||(Y - \hat{Y}) + (\hat{Y} - \bar{y}1)||^2 = ||\hat{Y} - \bar{y}1||^2 + ||\hat{\varepsilon}||^2$$

donc

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

$$SCT = SCE + SCR,$$

où

- $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$ représente la somme des carrés totale ;
- $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ représente la somme des carrés expliquée ;
- $SCR = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$ représente la somme des carrés résiduelle.

Définition 1.8.1. *Le coefficient de determination* R^2 *est defini par :*

$$\mathbb{R}^2 = \frac{SCE}{SCT} = \frac{||\hat{Y} - \bar{y}1||^2}{||Y - \bar{y}1||^2}.$$

On voit que R^2 correspond au cosinus carré de l'angle θ, De façon schématique, on peut différencier les cas suivants :

✠ Si $R^2 = 1$, le modèle explique tout, l'angle θ vaut zero et Y est dans $\mathcal{M}(X)$, c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$ pour tout i.

✠ Si $R^2 = 0$, cela veut dire que $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 0$, donc $\hat{y}_i = \bar{y}$ pour tout i.

Remarque 1.8.2. *On peut aussi voir R^2 comme le carré du coefficient de corrélation empirique entre les x_i et les y_i :*

$$R^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 = \rho^2_{X,Y}.$$

Cas d'erreurs gaussiennes

Mieux que les expressions des estimateurs et celles de leurs variances, on aimerait connaitre leurs lois. Dans cette optique, il faut bien entendu faire une hypothèse plus forte sur notre modèle, à savoir préciser la loi des erreurs. Nous supposerons ici que les erreurs sont gaussiennes. Les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$ deviennent $\mathcal{H}$

$$\begin{cases} (\mathcal{H}_1) & : & \varepsilon_i & \sim & N(0, \sigma^2) & \text{pour} & i = 1, \ldots, n \\ (\mathcal{H}_2) & : & \varepsilon_i & \text{sont} & \text{independants} & \text{pour} & i = 1, \ldots, n. \end{cases}$$

Le modèle de régression simple devient un modèle paramétrique, où les paramètres β_1, β_2, σ^2 sont à valeurs dans $\mathbb{R}$, $\mathbb{R}$ et $\mathbb{R}_+^*$ respectivement. La loi des ε_i étant connue, les lois des y_i s'en déduisent. Nous pouvons donc calculer la vraisemblance de l'échantillon et les estimateurs qui maximisent cette vraisemblance. C'est l'objet de la section suivante.

1.9 Prévision

Un des buts de la régression est de faire de la prévision, c'est-à-dire de prévoir la variable à expliquer y en présence d'une nouvelle valeur de la variable explicative x. Soit donc x_{n+1} une nouvelle valeur de la variable x, nous voulons prédire y_{n+1}.
Le modèle est toujours le même :

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$$

avec $E[\varepsilon_{n+1}] = 0$, $Var(\varepsilon_{n+1}) = \sigma^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$. Nous pouvons prédire la valeur correspondante grace au modèle ajusté :

$$y_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}.$$

Proposition 1.9.1 (Erreur de prévision)**.** *L'erreur de prévision* $\hat{\varepsilon}_{n+1} = (y_{n+1} - \hat{y}_{n+1})$ *satisfait les propriétés suivantes :*

1. $E\left[\hat{\varepsilon}_{n+1}\right] = 0$
2. $Var(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$

Preuve :
Pour l'espérance, il suffit d'utiliser le fait que ε_{n+1} est centrée et que les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont sans biais :

$$E[\hat{\varepsilon}_{n+1}] = E[\beta_1 - \hat{\beta}_1] + E[\beta_2 - \hat{\beta}_2]x_{n+1} + E[\varepsilon_{n+1}] = 0.$$

Nous obtenons la variance de l'erreur de prévision en nous servant du fait que y_{n+1} est fonction de ε_{n+1} seulement, tandis que $\hat{y}_{n+1}$ est fonction des autres erreurs $(\varepsilon_i)_{1 \leq i \leq n}$.

$$Var(\hat{\varepsilon}_{n+1}) = Var(y_{n+1} - \hat{y}_{n+1}) = Var(y_{n+1}) + Var(\hat{y}_{n+1}) = \sigma^2 + Var(\hat{y}_{n+1}).$$

Calculons le second terme

$$\begin{aligned}
Var(\hat{y}_{n+1}) &= Var\left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) \\
&= Var(\hat{\beta}_1) + x_{n+1}^2 Var(\hat{\beta}_2) + 2x_{n+1} Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{\sum_{i=1}^n x_i^2}{n} + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} + \bar{x}^2 + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).
\end{aligned}$$

Au total, on obtient bien :

$$Var(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Pour l'éspérance et la variance, nous avons évidemment les mêmes résultats que ceux obtenus précédamment. De plus, puisque $\hat{y}_{n+1}$ est linéaire en $\hat{\beta}_1$, $\hat{\beta}_2$ et ε_{n+1}, on peut préciser sa loi :

$$y_{n+1} - \hat{y}_{n+1} \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right).$$

A nouveau on ne connait pas σ^2 et on l'estime donc par $\hat{\sigma}^2$.
Comme $(y_{n+1} - \hat{y}_{n+1})$ et $\hat{\sigma}^2(n-2)/\sigma^2$ sont indépendants, on peut énoncer un résultat donnant des intervalles de confiance pour y_{n+1}.

Proposition 1.9.2. *(Loi et intervalle de confiance pour la prédiction) Avec les notations et les hypothèses précédentes, on a :*

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{1/2}} \sim \mathcal{T}_{n-2},$$

d'où l'on déduit l'intervalle de confiance pour y_{n+1} :

$$\left[\hat{y}_{n+1} \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2}\right].$$

Chapitre 2

La régression linéaire multiple

2.1 Introduction

La modélisation de la concentration d'ozone dans l'atmosphère évoquée au Chapitre 1 est relativement simpliste. En effet, d'autres variables peuvent expliquer cette concentration, par exemple le vent qui pousse les masses d'air. Ce phénomène physique est connu sous le nom d'advectance (apport d'ozone) ou de dilution. D'autres variables telles le rayonnement, la précipitation, etc, ont une influence certaine sur la concentration d'ozone. L'association Air Breizh mesure ainsi en même temps que la concentration d'ozone d'autres variables susceptibles d'avoir une influence sur celle-ci. Voici quelques-unes de ces données :

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
V	9.25	-6.15	-4.92	11.57	-6.23	2.76	10.15	13.5	21.27	13.79
N_{12}	5	7	6	5	2	7	4	6	1	4
O_3	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

La variable V est une variable synthétique. En effet, le vent est normalement mesuré en degrés (direction) et mètres par seconde (vitesse). La variable V que nous avons créé est la projection du vent sur l'axe Est-Ouest, elle tient donc compte à la fois de la direction et de la vitesse.
Pour analyser la relation entre la température T, le vent V , la nébulosité à midi N et l'ozone O3, nous allons chercher une fonction f telle que :

$$O_{3i} \approx f(T_i, V_i, N_i)$$

Pour preciser $\approx$ il va falloir définir comme au Chapitre 1 un critère quantifiant la qualité de l'ajustement de la fonction f aux données, ou inversement le cout de non-ajustement.
Minimiser un coût nécessite aussi la connaissance de l'espace sur lequel on minimise, c'est-à-dire la classe de fonctions F dans laquelle nous supposerons que se trouve la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\arg\min_{f \in F} \sum_{i=1}^{n} L(y_i - f(x_i))$$

où n représente le nombre de données à analyser, L(.) est appelée fonction de coût, ou de perte, et x_i est une variable vectorielle pour tout i. La fonction de coût sera la même que celle utilisée précédemment, c'est-à-dire le coût quadratique. En ce qui concerne le choix de la classe F, par analogie avec le chapitre précédent, nous utiliserons la classe suivante :

$$F = \left\{ f : \mathbb{R}^p \to \mathbb{R},\ f(x_1, ..., x_p) = \sum_{j=1}^{n} \beta_j x_j \right\}$$

.

2.2 Modélisation

Le modèle de régression linéaire multiple 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 :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \quad i = 1, ..., n$$

où :

1. les x_{ij} sont des nombres connus, non aléatoires
2. les paramètres β_j du modèle sont inconnus
3. les ϵ_i sont des variables aléatoires inconnues

En utilisant l'écriture matricielle nous obtenons la définition suivante :

Définition 2.2.1. *(Modèle de régression linéaire multiple) Un modèle de régression linéaire est défini par une équation de la forme :*

$$Y = X\beta + \epsilon$$

où :
- *Y est un vecteur aléatoire de dimension n.*
- *X est une matrice de taille $n \times (p+1)$ connue.*
- *β est le vecteur de dimension $p+1$ des paramètres inconnus du modèle.*
- *ϵ est le vecteur de dimension n des erreurs.*

Les hypothèses conçernant le modèle sont

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : rg(X) = p+1 \\ (\mathcal{H}_2) : \mathbb{E}(\epsilon) = 0 \quad Var(\epsilon) = \sigma^2 \end{cases} .$$

L'hypothèse $(\mathcal{H}_2)$ signifie que les erreurs sont centrées, de même variance et non corrélées entre elles

2.3 Estimateur des Moindres Carrés Ordinaire

Conditionnellement à la connaissance des valeurs des X_{ij}, les paramètres inconnus du modèle : le vecteur β et σ^2 (paramètre de nuisance), sont estimés par minimisation du critère des moindres carrés (MCO).

Définition 2.3.1. *L'estimateur des moindres carrés $\hat{\beta}$ est définie comme suit :*

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 = \arg\min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2$$

Dans la suite de cette section, nous allons donner l'expression de l'estimateur $\hat{\beta}$ ainsi que certaines de ses propriétés.

2.4 Calcul de $\hat{\beta}$

Pour déterminer $\hat{\beta}$, une méthode consiste à se placer dans l'espace des variables : $Y = [y_1, ..., y_n]^t$ est le vecteur des variables à expliquer. La matrice X est formée de p vecteurs colonnes (la première colonne étant généralement constituée de 1). Le sous-espace de $\mathbb{R}^n$ engendré par les p vecteurs colonnes de X est appelé espace image, ou espace des solutions, et noté $\mathcal{M}(X)$. Il est de dimension p, par l'hypothèse $\mathcal{H}_1$ et tout vecteur de cet espace est de la forme $X\alpha$, ou α est un vecteur de $\mathbb{R}^p$:

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

Le vecteur Y est la somme d'un élément de $\mathcal{M}(X)$ et d'un élément bruit de $\mathbb{R}^n$, lequel n'a aucune raison d'appartenir à $\mathcal{M}(X)$. Minimiser $\|Y - X\alpha\|$ revient a chercher un élément de $\mathcal{M}(X)$ qui soit le plus proche de Y au sens de la norme euclidienne classique. Cet unique élément est par definition, le projeté orthogonal de Y sur $\mathcal{M}(X)$. Il sera note $\hat{Y} = P_X Y$, où P_X est la matrice de projection orthogonale sur $\mathcal{M}(X)$. Il peut aussi s'écrire sous la forme $\hat{Y} = X\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des MCO de β. L'espace orthogonal à $\mathcal{M}(X)$, note $\mathcal{M}^{\perp}(X)$, est souvent appelé espace des résidus. En tant que supplementaire orthogonal, il est de dimension $n - p = dim(\mathbb{R}^n) - dim(\mathcal{M}(X))$

Proposition 2.4.1. *. L'estimateur $\hat{\beta}$ des moindres carrés ordinaire a pour expression*

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

Preuve
L'expression à minimiser sur $\beta \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - ... - \beta_p x_i^p) &= \|Y - X\beta\|^2 \\
&= (Y - X\beta)^t (Y - X\beta) \\
&= y^t y - 2\beta^t X^t Y + \beta^t X^t X \beta
\end{aligned}$$

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

$$X^tY - X^tX\beta = 0$$

dont la solution correspond bien à un minimum car la matrice X^tX est semi définie-positive.

donc

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

2.5 Quelques propriétés

Comme en régression linéaire simple, l'estimateur obtenu est sans biais. On rappelle que la matrice de covariance du vecteur aléatoire $\hat{\beta}$, est par définition :

$$Var(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])^t]$$

puisque β est de dimension p, elle est de dimension $p \times p$. De plus, pour toute matrice A de taille $n \times p$ et tout vecteur β de dimension n déterministes on a
$\mathbb{E}[A\hat{\beta} + \beta] = A\mathbb{E}[\hat{\beta} + \beta]$ et $Var(A\hat{\beta} + \beta) = AVar(\hat{\beta})A^t$. ces propriétés élémentaires seront constamment appliquées dans la suite.

Proposition 2.5.1. *(Biais et matrice de covariance) L'estimateur $\hat{\beta}$ des moindres carrés est sans biais, i.e $\mathbb{E}[\hat{\beta}] = \beta$, est sa matrice de covariance est*
$Var(\hat{\beta}) = \sigma^2(X^tX)^{-1}$

Preuve
Pour le biais : il suffit d'écrire :

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^tX)^{-1}X^tY] \\ &= (X^tX)^{-1}X^t\mathbb{E}[Y] \\ &= (X^tX)^{-1}X^t\mathbb{E}[X\beta + \epsilon] \end{aligned}$$

et puisque $\mathbb{E}[\epsilon] = 0$, il vient

$$\mathbb{E}[\hat{\beta}] = (X^tX)^{-1}X^tX\beta = \beta$$

Pour la variance, on procède de même :

$$\begin{aligned} Var[\hat{\beta}] &= Var[(X^tX)^{-1}X^tY] \\ &= (X^tX)^{-1}X^tVar[Y]X(X^tX)^{-1} \end{aligned}$$

où $Var[Y] = Var[X\beta + \epsilon] = Var[\epsilon] = \sigma^2\mathbf{I}$, donc

$$\begin{aligned} Var[\hat{\beta}] &= \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} \\ &= \sigma^2(X^tX)^{-1} \end{aligned}$$

Théorème 2.5.1. *(Gauss-Marcov) L'estimateur $\hat{\beta}$ des MCO est de variance minimale parmi les estimateurs linéaires sans biais de β*

Preuve

Nous allons montrer que, pour tout autre estimateur $\tilde{\beta}$ de β linéaire et sans biais, $Var(\tilde{\beta}) \geq Var(\hat{\beta})$. Nous savons que la matrice de covariance de la somme de deux vecteurs aléatoire U et V est :

$$Var(U+V) = Var(U) + Var(V) + cov(U,V) + cov(V,U)$$

Décomposons ainsi la variance de $\tilde{\beta}$:

$$\begin{aligned} Var(\tilde{\beta}) &= Var(\tilde{\beta} - \hat{\beta} + \hat{\beta}) = Var(\tilde{\beta} - \hat{\beta}) + Var(\hat{\beta}) \\ &+ cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) + cov(\hat{\beta}, \tilde{\beta} - \hat{\beta}) \end{aligned}$$

les variances étant semi-définies positives, si nous montrons que $cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$. Puisque $\hat{\beta}$ est linéaire, $\tilde{\beta} = AY$ où A est une matrice (p,n), de plus, nous savons qu'il est sans biais, c'est à dire $\mathbb{E}(\tilde{\beta}) = \beta$ pour tout β, donc $AX = I$. La covariance devient :

$$\begin{aligned} cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) &= cov(AY, (X^tX)^{-1}X^tY) - Var(\hat{\beta}) \\ &= \sigma^2AX(X^tX)^{-1} - \sigma^2(X^tX)^{-1} = 0 \end{aligned}$$

2.6 Résidus et variance résiduelle

Les résidus sont définis par :

$$\hat{\epsilon} = [\hat{\epsilon}_1, ..., \hat{\epsilon}_n]^t = Y - \hat{Y} = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\epsilon$$

car $Y = X\beta$ et $X\beta \in \mathcal{M}(X)$. On peut alors énoncer les résultats suivants.

Propriétés

Sous l'hypothèses $\mathcal{H}$, on a :

1. $\mathbb{E}(\hat{\epsilon}) = 0$.
2. $Var(\hat{\epsilon}) = \sigma^2 P_{X^\perp}Y$.
3. $\mathbb{E}(\hat{Y}) = X\beta$.
4. $Var(\hat{Y}) = \sigma^2 P_X$.
5. $Cov(\hat{\epsilon}, \hat{Y}) = 0$.

Preuve

1. $\mathbb{E}(\hat{\epsilon}) = \mathbb{E}(P_{X^\perp}\epsilon) = P_{X^\perp}\mathbb{E}(\epsilon) = 0$
2. $Var(\hat{\epsilon}) = P_{X^\perp}Var(\epsilon)P_{X^\perp}^t = P_{X^\perp}Var(\epsilon)P_{X^\perp} = \sigma^2 P_{X^\perp}P_{X^\perp} = \sigma^2 P_{X^\perp}$
3. $\mathbb{E}(\hat{Y}) = \mathbb{E}(X\hat{\beta}) = X\mathbb{E}(\hat{\beta}) = X\beta$, car $\hat{\beta}$ est sans biais.
4. $Var(\hat{Y}) = Var(X\hat{\beta}) = XVar(\hat{\beta})X^t = \sigma^2 X(XX^t)^{-1}X^t$
5. Rappelons que la covariance entre deux vecteurs aléatoires est une application bilinéaire et que $Cov(U, U) = Var(U)$, ici, ceci donne :

   $$Cov(\hat{\epsilon}, \hat{Y}) = Cov(\hat{\epsilon}, Y - \hat{\epsilon}) = Cov((\hat{\epsilon}, Y) - Var(\hat{\epsilon}) = Cov(P_{X^\perp}Y, Y) - \sigma^2 P_{X^\perp}$$

   et puisque $Var(Y) = \sigma^2$ nous avons :

   $$Cov(\hat{\epsilon}, \hat{Y}) = P_{X^\perp}Var(Y) - \sigma^2 P_{X^\perp} = 0$$

Proposition 2.6.1. *La statistique*

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n-p} = \frac{SCR}{n-p}$$

est un estimateur sans biais de σ^2.

Preuve
Nous calculons $\mathbb{E}[\|\hat{\epsilon}\|^2]$: puisque c'est un scalaire, il est égal à sa trace, ce qui donne :

$$\mathbb{E}[\|\hat{\epsilon}\|^2] = \mathbb{E}[Tr(\|\hat{\epsilon}\|^2)] = \mathbb{E}[Tr(\hat{\epsilon}^t\hat{\epsilon})]$$

et puisque pour toute matrice A, on a $Tr(AA^t) = Tr(A^tA) = \sum_{i,j} a_{ij}^2$ il vient :

$$\mathbb{E}[\|\hat{\epsilon}\|^2 = \mathbb{E}[Tr(\hat{\epsilon}^t\hat{\epsilon})] = Tr(\mathbb{E}[\hat{\epsilon}\hat{\epsilon}^t]) = Tr(Var[\hat{\epsilon}]) = Tr(\sigma^2 P_{X^\perp})$$

Et comme $P_{X^\perp}$ est la matrice de la projection orthogonale sur un espace de dimension (n-p), on a bien :

$$\mathbb{E}[\|\hat{\epsilon}\|^2] = (n-p)\sigma^2$$

.

2.7 Coefficient de détermination R^2

Définition 2.7.1. *Le coefficient de détermination R^2 est défini par :*

$$R^2 = \cos^2\theta = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = 1 - \frac{\|\hat{\epsilon}\|^2}{\|Y\|^2} = 1 - \frac{SCR}{SCT}$$

ou plus souvent, si la constante fait partie du modèle, par :

$$R^2 = \cos^2\theta = \frac{V.\ expliqu\ par\ le\ modle}{Variation\ totale} = \frac{\|\hat{Y} - \bar{y}\|^2}{\|Y - \bar{y}\|^2} = 1 - \frac{\|\hat{\epsilon}\|^2}{\|\hat{Y} - \bar{y}\|^2} = 1 - \frac{SCR}{SCT}$$

Ce coefficient mesure le cosinus carré de l'angle entre les vecteurs Y et $\hat{Y}$ pris à l'origine ou pris en $\bar{y}$. Néanmoins, on peut lui reprocher de ne pas tenir compte de la dimension de l'espace de projection M(X), d'où la définition du coefficient de détermination ajusté.

Définition 2.7.2. *Le coefficient de détermination ajusté R_a^2 est défini par :*

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

2.8 Estimation par intervalle de confiance

Cherchons un intervalle de confiance pour β. On a

$$\frac{\hat{\beta} - \mathbb{E}(\hat{\beta})}{\sqrt{Var(\hat{\beta})}} \sim \mathcal{N}(0,1)$$

$$\Leftrightarrow \frac{\hat{\beta} - \beta}{\sigma}(X^tX)^{\frac{1}{2}} \sim \mathcal{N}(0,1)$$

En utilisat $\hat{\sigma}$:

$$\frac{\hat{\beta} - \beta}{\sigma}(X^tX)^{\frac{1}{2}} \sim t(N-p-1)$$

L'intervalle de confiance pour un risque α est tel que :

$$1 - \alpha = \mathbb{P}[-c_1\hat{\sigma} \leq (\hat{\beta} - \beta)(X^tX)^{\frac{1}{2}} \leq c_1\hat{\sigma}]$$

avec c_1 le fractile d'ordre $1 - \frac{\alpha}{2}$ de $t(N-p-1)$.
En isolant β dans l'équation ci-dessus, on obtient :

$$1 - \alpha = \mathbb{P}[\hat{\beta} - c_1\hat{\sigma}(X^tX)^{\frac{1}{2}} \leq \beta \leq \hat{\beta} + c_1\hat{\sigma}(X^tX)^{\frac{1}{2}}]$$

L'intervalle obtenu est donc :

$$\beta \in [\hat{\beta} - c_1\hat{\sigma}(X^tX)^{\frac{1}{2}}, \hat{\beta} + c_1\hat{\sigma}(X^tX)^{\frac{1}{2}}]$$

2.9 Prévision

Un des buts de la régression est de proposer des prédictions pour la variable à expliquer y lorsque nous avons une nouvelle valeur de x. Soit donc $x^t_{n+1} = (x_{n+1}, 1, ..., x_{n+1}, p)$ une nouvelle valeur pour laquelle nous voudrions prédire y_{n+1}. Cette variable réponse est définie par $y_{n+1} = x^t_{n+1}\beta + \epsilon_{n+1}$, avec $\mathbb{E}[\epsilon_{n+1}] = 0$, $Var[\epsilon_{n+1}] = \sigma^2$ et
$cov(\epsilon_{n+1}, \epsilon_i) = 0 \quad pour \quad i = 1, ..., n.$

Proposition 2.9.1. *L'erreur de prévision $\hat{\epsilon}_{n+1} = (y_{n+1} - \hat{y}_{n+1})$ satisfait les propriétés suivantes :*

1. $\mathbb{E}[\hat{\epsilon}_{n+1}] = 0$
2. $Var[\hat{\epsilon}_{n+1}] = \sigma^2(1 + x_{n+1}^t(X^tX)^{-1}x_{n+1})$

Preuve.
Comme $\mathbb{E}[\epsilon_{n+1}] = 0$ et puisque $\hat{\beta}$ est un estimateur sans biais de β, il est clair que

$$\begin{aligned}\mathbb{E}[\hat{\epsilon}_{n+1}] &= \mathbb{E}[x_{n+1}^t(\beta - \hat{\beta}) + \epsilon_{n+1}] \\ &= x_{n+1}^t(\beta - \mathbb{E}(\hat{\beta}) + \mathbb{E}[\epsilon_{n+1}]) \\ &= 0\end{aligned}$$

En calculant la variance de l'erreur de prévision, puisque $\hat{\beta}$ dépend uniquement des variables aléatoires $(\epsilon_i)_{1<i<n}$ il vient

$$\begin{aligned}Var(\hat{\epsilon}_{n+1}) &= Var(x_{n+1}^t(\beta - \hat{\beta}) + \epsilon_{n+1}) \\ &= \sigma^2 + x_{n+1}^t Var(\hat{\beta})x_{n+1} \\ &= \sigma^2(1 + x_{n+1}(X^tX)^{-1}x_{n+1})\end{aligned}$$

2.9.1 Inférence sur le modèle

Le modèle peut être testé globalement. Sous l'hypothèse nulle

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

La statistique

$$\frac{SSR/P}{SSE/(n-p-1)} = \frac{MSR}{MSE}$$

suit une loi de Ficher avec p et (n-p-1) degrés de liberté. Les résultats sont habituellement présentés dans un tableau "d'analyse de variance" sous la forme suivante :

Source de variation	d.d.l	Somme des carrés	Variance	F
Régression	p	SSE	MSR=SRE/P	MSR/MSE
Erreur	n-p-1	SSR	MSR=SSR/(n-p-1)	
Total	n-1	SST		

Chapitre 3

La régression ridge

3.1 Introduction

Nous savons que l'estimateur des moindres carrés et sans biais, nous savons même qu'il est l'estimateur ayant la variance la plus petite possible parmi les estimateurs linéaires sans biais de β. Ainsi, si l'on cherche un estimateur dont l'erreur de prédiction soit plus petite que celle de l'estimateur des moindres carrés, nous devons nécessairement gagner sur la variance tout en perdant le fait que l'estimateur soit non biaisé. On espère que le gain dû à la réduction de la variance soit plus grand que la perte due au biais. E. Hoerl et Kennard [8] 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.

3.2 Estimateur ridge ordinaire

L'estimateur ridge ordinaire a été proposée par E. Hoerl et Kennard sans article " Ridge regression, biaised estimation for nonorthogonal problems", ils partent de la constatation suivante : lorsque l'on se trouve en face d'un problème de forte colinéarité des valeurs propres de la matrice X^tX, et l'orsqu'il y a des corrélations entre les variables explicatives, la matrice X^tX à des valeurs propres proches de zéro, dans ce cas l'estimateur des moindres carrés n'est pas stable, car sa variance explose. Hoerl et Kennard proposent d'augmenter faiblement les valeurs propres de cette matrice d'une même constante $k > 0$, ceci dans le but de rendre les estimateurs stables.

3.2.1 Estimateur ridge ordinaire dans le modéle canonique

Soit le modéle défini précédamment

$$Y = X\beta + \epsilon$$

Il existe une matrice orthogonale T de vecteurs propres de la matrice X^tX i.e $(T^tT = I)$ et une matrice diagonale Λ de valeurs propres notée $\lambda_i, i = 1, ..., p$

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & \lambda_p \end{pmatrix}$$

telle que $X^tX = T\Lambda T^t$ et $Y = XTT^t\beta + \epsilon$.
En notant XT=Z et $T^t\beta = \gamma$, le modéle peut s'écrire

$$Y = Z\gamma + \epsilon. \tag{3.1}$$

C'est le modéle sous sa forme canonique.
Remarques

1. On constate que $Z^tZ = T^tX^tXT = \Lambda$, matrice diagonale des valeurs propres, ou des carrés des valeurs singuliéres, d_i, $i = 1...p$

$$Z^tZ = \begin{pmatrix} \lambda_1 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & \lambda_p \end{pmatrix} = \begin{pmatrix} d_1^2 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & d_p^2 \end{pmatrix}$$

2. Dans ces conditions l'estimateur Ridge ordinaire, donne le résultat suivant : Hoerl et kennard proposent donc d'augmenter "un peu" les d_i en ajoutant à tous la même constante k positive. Cela conduit à augmenter tous les termes d'une valeur k, et comme Z^tZ est une matrice diagonale, on obtient

$$Z^tZ + kI = \begin{pmatrix} d_1^2 + k & \cdots & \cdots \\ \cdots & d_i^2 + k & \cdots \\ \cdots & \cdots & d_p^2 + k \end{pmatrix}$$

Pour estimé β, il suffit d'estimer le paramétre γ.

Définition 3.2.1. *L'estimateur ridge ordinaire de γ est donné par*

$$\widehat{\gamma r} = (Z^t Z + kI)^{-1} Z^t Y \tag{3.2}$$

L'estimateur des moindres carrés ordinaire (**MCO**) de γ est donné par :

$$\hat{\gamma}_{MC} \;=\; (Z^t Z)^{-1} Z^t Y$$

où $Z^t Z = \Lambda$
D'aprés la définition (3.2.1), l'estimateur ridge ordinaire (**ORR**) de γ peut être écrit comme :

$$\hat{\gamma}_r = (\Lambda + kI_p)^{-1} Z^t Y. \tag{3.3}$$

3.2.2 Estimateur ridge ordinaire dans le modéle de base

On se place dans le modéle linéaire suivant

$$Y = \beta_0 \mathbf{1}_n + X\beta + \epsilon$$

où

$$X = \begin{pmatrix} x_1^1 & x_1^2 & . & x_1^p \\ x_2^1 & x_2^2 & . & x_2^p \\ . & . & . & . \\ x_n^1 & x_n^2 & . & x_n^p \end{pmatrix}$$

et

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}, \quad \mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}$$

Définition 3.2.2. *L'estimateur ridge de β dans le modèle précédent est défini par :*

$$\begin{aligned} \hat{\beta}(k) &= \arg\min_{\beta \in \mathbb{R}^p} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^j \beta_j)^2 + k \sum_{j=1}^{p} \beta_j^2 \right) \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - \beta_0 \mathbf{1}_n - X\beta\|^2 + k\|\beta\|^2 \right) \end{aligned}$$

le paramétre $k \geq 0$ contrôle le niveau de pénalité : lorsque $k \to 0$, on s'approche de la solution des moindres carrés.

Proposition 3.2.1. *L'estimateur ridge s'exprime aussi sous la forme :*

$$\hat{\beta}_0(k) = \bar{Y}, \quad \hat{\beta}(k) = (X^t X + kI_p)^{-1} X^t Y$$

Preuve
Pour pouvoir résoudre l'équation ci-dessus, il est préférable de centrer les données : remplacer chaque x_{ij} par $x_{ij} - \bar{X}_j$, $\bar{X}_j = \frac{1}{N}\sum_{i=1}^{p} x_{ij}$. En utilisant les variables centrées x_{ij}, et $RSS(\beta) = \|y - \mathbf{1}_n\beta_0 - X\beta\|$ donc la dérivée par rapport à β_0

$$\frac{\partial}{\partial \beta_0} RSS(\beta) + k\|\beta\|^2 = 0 \Leftrightarrow \mathbf{1}_n^T(Y - \beta_0 - X\beta) = 0 \Leftrightarrow \sum_i y_i - n\beta_0 - \mathbf{1}_n^T X\beta = 0.$$

En divisant par n

$$\bar{Y} - \beta_0 - \sum_{j=1}^{p} X_j\beta_j$$

Lorsque X est centrée en colonne, $\hat{\beta}_0 = \bar{Y}$, pour les autres paramètres β l'hypothèse de gradient nul donne :

$$\nabla_\beta = (RSS(\beta) + k\|\beta\|^2) = 0 \Leftrightarrow 2X^t(Y - \beta_0 \mathbf{1}_n - X\beta) + 2k\beta$$

En estime β_0 par $\bar{Y}$ donc

$$X^t Y - X^t X\beta + k\beta = 0 \Leftrightarrow (X^t X + kI_p)\beta = X^t Y$$

d'où

$$\hat{\beta}(k) = (X^t X + kI_p)^{-1} X^t Y \tag{3.4}$$

$\hat{\beta}(k)$ est appelé l'estimateur ridge ordinaire notée ORR dans le modèle de base.

Remarques

1. $X^t X$ est une matrice symétrique positive (pour tout vecteur u de $\mathbb{R}^p$, $u^t(X^t X)u = \|Xu\|^2 \geq 0$. Il en resulte que pour tout $k > 0$, $X^t X + kI_p$ est nécessairement inversible.

2. 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 $\hat{\beta}_j$ pour $j \geq 1$.
3. 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.
4. 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 :

$$\hat{\beta}(k) = \arg\min_{\beta} \left\{ \|Y - X\beta\|^2; \quad \|\beta\| < c \right\}$$

3.2.3 Propriétés de l'estimateur ridge ordinaire (ORR)

Espérance

Revenons au définition de l'estimateur ridge ordinaire ORR et des moindres carrée MCO :

$$\hat{\beta}(k) = (X^t X + kI_p)^{-1} X^t Y \tag{3.5}$$

et

$$\hat{\beta}_{MC} = (X^t X)^{-1} X^t Y \tag{3.6}$$

En multipliant la seconde égalité à gauche par $X^t X$, on obtient $X^t X \hat{\beta}_{MC} = X^t Y$, d'aprés la définition de l'estimateur MCO vient

$$\hat{\beta}(k) = (X^t X + kI_p)^{-1} X^t X \hat{\beta}_{MC}$$

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

$$\begin{aligned} \mathbb{E}(\hat{\beta}(k)) &= (X^t X + kI_p)^{-1} X^t X \mathbb{E}(\hat{\beta}) \\ &= (X^t X + kI_p)^{-1} (X^t X)\beta \\ &= (X^t X + kI_p)^{-1} (X^t X + kI_p - kI_p)\beta \\ &= \beta - k(X^t X + kI_p)^{-1}\beta \end{aligned} \tag{3.7}$$

Le biais de cet estimateur vaut

$$\begin{aligned} Biais(\hat{\beta}(k)) &= \mathbb{E}(\hat{\beta}(k)) - \beta \\ &= -k(X^tX + kI_p)^{-1}\beta \end{aligned}$$

Ce qui montre que l'estimateur ridge est un estimateur biaisé.
variance

$$\begin{aligned} Var(\hat{\beta}(k)) &= Var[(X^tX + kI_p)^{-1}X^tY] \\ &= [(X^tX + kI_p)^{-1}X^t][X(X^tX + kI_p)^{-1}]Var(Y) \end{aligned}$$

Où cette hypothèse

$$Var(Y) = \sigma^2\mathbf{I}$$

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

$$Var(\hat{\beta}(k)) = \sigma^2(X^tX + kI_p)^{-1}X^tX(X^tX + kI_p)^{-1}$$

Conclusion
Nous nous trouvons face à un estimateur biaisé. Est-il meilleur que les MCO ? Mystére car pour l'instant nous ne savons comparer que des estimateurs sans biais : on prend celui de variance minumum. Mais pour comparer des estimateurs sans biais avec des estimateurs biasés il faut introduire un nouveau critére celui du MSE.

3.3 Le critère de l'erreur quadratique moyenne (MSE)

Définition 3.3.1. *Soit $\hat{\beta}$ un estimateur de β, on appelle MSE de $\hat{\beta}$ la formule suivante :*

$$MSE(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta)^t(\hat{\beta} - \beta)]$$

Remarques

1. *Dans la théorie de l'estimation paramétrique, MSE nous permet de comparer deux estimateurs qu'ils soient sans biais ou biaisés. Le travail se fera plutôt avec la trace de cette matrice, on a*

$$traceMSE(\hat{\beta}) = trace\mathbb{E}[(\hat{\beta} - \beta)^t(\hat{\beta} - \beta)]$$

On note trace=tr, puisque la fonction trace est une application linéaire, alors

$$\begin{aligned} trMSE(\hat{\beta}) &= tr\mathbb{E}({}^t(\hat{\beta}-\beta)(\hat{\beta}-\beta)) \\ &= tr\mathbb{E}[(\hat{\beta}-\mathbb{E}(\hat{\beta})^t+\mathbb{E}(\hat{\beta})-\beta)(\hat{\beta}-\mathbb{E}(\hat{\beta})+\mathbb{E}(\hat{\beta})] \\ &= \mathbb{E}[(\hat{\beta}-\mathbb{E}(\hat{\beta}))^t(\hat{\beta}-\mathbb{E}(\hat{\beta}))]+\mathbb{E}[(\mathbb{E}(\hat{\beta})-\beta)^t(\hat{\beta}-\mathbb{E}(\hat{\beta}))] \\ &+ \mathbb{E}[(\hat{\beta}-\mathbb{E}(\hat{\beta}))^t(\mathbb{E}(\hat{\beta})-\beta)]+\mathbb{E}[(\mathbb{E}(\hat{\beta})-\beta)^t(\mathbb{E}(\hat{\beta})-\beta)] \end{aligned}$$

où $\mathbb{E}[(\hat{\beta}-\mathbb{E}(\hat{\beta}))^t(\mathbb{E}(\hat{\beta})-\beta)] = [(\mathbb{E}(\hat{\beta})-\mathbb{E}(\hat{\beta}))^t(\mathbb{E}(\hat{\beta}-\hat{\beta})] = 0 = \mathbb{E}[(\mathbb{E}(\hat{\beta})-\beta)^t(\hat{\beta}-\mathbb{E}(\hat{\beta}))]$
donc

$$trMSE(\hat{\beta}) = tr\mathbb{E}[(\hat{\beta}-\mathbb{E}(\hat{\beta}))^t(\hat{\beta}-\mathbb{E}(\hat{\beta}))]+tr\mathbb{E}[(\mathbb{E}(\hat{\beta})-\beta)^t(\mathbb{E}(\hat{\beta})-\beta)$$

$$trMSE(\hat{\beta}) = \sum_{i=1}^{p} var(\hat{\beta}_i)+\sum_{i=1}^{p}(biais(\hat{\beta}))^2$$

2. *Dans le cas d'un estimateur sans biais* $trMSE(\hat{\beta}) = \sum_{i=1}^{p} var(\hat{\beta}_i)$

3.4 Comparaison de l'estimateur ORR avec l'estimateur MCO

L'estimateur ORR est un estimateur biaisé, la MSE de ORR donnée par (Özkale and Kaçiranlar [10] est

$$MMSE(\hat{\beta}(k)) = \sigma^2 WS^{-1}W^t + k^2 S_k^{-1}\beta\beta^t S_k^{-1}$$

et on a

$$\begin{aligned} trMSE(\hat{\beta}(k)) &= trMSE(\widehat{\gamma r}_i) = \sum_{i=1}^{p} var(\widehat{\gamma r}_i)+\sum_{i=1}^{p}(biais(\widehat{\gamma r}_i))^2 \\ &= \sum_{i=1}^{p} var(\widehat{\gamma r}_i)+\sum_{i=1}^{p}(\gamma r_i-\mathbb{E}(\widehat{\gamma r}_i))^2 \end{aligned}$$

où

$$\mathbb{E}(\widehat{\gamma r}_i) = \gamma - \begin{pmatrix} k/(d_1^2+k) & 0 & \cdots & 0 \\ 0 & k/(d_2^2+k) & \cdots & 0 \\ 0 & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & k/(d_p^2+k) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ .. \\ \gamma_p \end{pmatrix}$$

Puisque

$$\sum_{i=1}^{p} (\gamma_i - \mathbb{E}(\widehat{\gamma r}_i)^2) = \sum_{i=1}^{p} k^2 \gamma_i^2 / (\lambda_i + k)^2$$

La matrice de variance-covariance de $\hat{\gamma}$ s'écrit :

$$\begin{aligned} Var(\widehat{\gamma r}) &= \sigma^2 \begin{pmatrix} 1/(d_1^2+k) & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & 1/(d_p^2+k) \end{pmatrix} \begin{pmatrix} d_1^2 & & 0 \\ 0 & .. & 0 \\ 0 & & d_p^2 \end{pmatrix} \begin{pmatrix} 1/(d_1^2+k) & & 0 \\ 0 & \cdots & 0 \\ 0 & & 1/(d_p^2+k) \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} d_1^2/(d_1^2+k)^2 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & d_p^2/(d_p^2+k)^2 \end{pmatrix} \end{aligned}$$

donc $\sum_{i=1}^{p} var(\widehat{\gamma r}_i) = \sigma^2 \sum_{i=1}^{p} d_i^2/(\lambda_i + k)$ alors

$$trMSE(\widehat{\gamma r}_i) = \sigma^2 \sum_{i=1}^{p} d_i^2/(d_i^2+k) + \sum_{i=1}^{p} k^2 \gamma_i^2/(d_i^2+k)^2$$

La trace MSE de l'estimateur Ridge s'écrit

$$trMSE(\widehat{\gamma r}_i) = \sum_{i=1}^{p} \frac{\sigma^2 d_i^2 + k^2 \gamma_i^2}{(d_i^2+k)^2} \tag{3.8}$$

Comparaison

Nous avons vu que la trMSE de l'estimateur Ridge s'écrit

$$trMSE(\widehat{\gamma r}) = \sum_{i=1}^{p} \frac{\sigma^2 d_i^2 + k^2 \gamma_i^2}{(d_i^2+k)^2}$$

La trMSE de l'estimateur des MCO $\hat{\gamma}$ est (pour k=0)

$$trMSE(\hat{\gamma}) = \sum_{i=1}^{p} \frac{\sigma^2 d_i^2}{(d_i^2)^2} = \sum_{i=1}^{p} \frac{\sigma^2}{\lambda_i}$$

Pour que l'estimateur Ridge soit meilleur que MCO, selon le critére du MSE il faut que :

$$trMSE(\widehat{\gamma r}_i) < trMSE(\hat{\gamma})$$

$$\sum_{i=1}^{p} \frac{\sigma^2 d_i^2 + k^2 \gamma_i^2}{(d_i^2 + k)^2} < \sum_{i=1}^{p} \frac{\sigma^2}{d_i^2}$$

il suffit donc vérifier l'inégalité pour chaque i, ie

$$\frac{\sigma^2 d_i^2 + k^2 \gamma_i^2}{(d_i^2 + k)^2} < \frac{\sigma^2}{\lambda_i}$$

et

$$\sigma^2 d_i^4 + k^2 \gamma_i^2 d_i^2 < \sigma^2 (d_i^4 + k^2 + 2k d_i^2)$$

$$k^2 (\gamma_i d_i^2 - \sigma^2) - 2k\sigma^2 d_i^2 < 0$$

C'est une inégalité de second degré en k.

1. Si $\gamma_i^2 \lambda_i - \sigma^2 > 0$, l'expression est négative entre les racines de l'équation du second degré, soit entre 0 et une $k_i^* = \frac{2\sigma^2 d_i^2}{\gamma_i^2 d_i^2 - \sigma^2} > 0$
2. Si $\gamma_i^2 d_i^2 - \sigma^2 < 0$, les deux expressions sont négatives et la condition est toujours vérifiée.

On en déduit donc que la MSE de l'estimateur Ridge est inférieur a la MSE des MCO si $0 < k < min k_i^*$.

3.5 Estimateur modifié du ridge sans biais (MUR)

Plusieurs autres estimateurs biaisés de β ont été proposées (Swindel [12], Sarkar [11], Batah et Gore [1], Batah et al [2], Batah et al [4]). Swindel [12] a défini un estimateur appelé estimateur ridge modifié (MRR) comme suit

$$\hat{\beta}(k, b) = (X^t X + k I_P)^{-1} (X^t Y + kb), \quad k \geq 0 \tag{3.9}$$

Où b est un estimateur à priori de β. Crouse et al [5] ont defini un estimateur ridge sans biais (URR) par

$$\hat{\beta}(k, J) = (X^t X + kI_P)^{-1}(X^t Y + kJ), \quad k \geq 0 \tag{3.10}$$

où J est une variable aléatoire de loi $N(\beta, \frac{\sigma^2}{k} I_P)$ pour $k > 0$.
Dans cette section on considére l'estimateur modifié du ridge sans biais (MUR).

3.5.1 Définition de l'estimateur MUR

Définition 3.5.1. *L'estimateur modifié du ridge sans biais (MUR) est donné par :*

$$\hat{\beta}_J(k) = [I - k(X^t X + kI_p)^{-1}]\hat{\beta}(k, J) \tag{3.11}$$

Où $\hat{\beta}(k, J)$ est données par l'équation (3.10).
Cet estimateur est appelé estimateur modifié du ridge sans biais (**MUR**), car il est développé à partir de l'estimateur URR.
L'estimateur MUR dans le modéle (3.1) devient

$$\hat{\gamma}_J(k) = [I - k(\Lambda + kI_p)^{-1}]\hat{\gamma}(k, J) \tag{3.12}$$

3.5.2 Propriétés de l'estimateur MUR

Espérance
On a

$$\hat{\beta}_J(k) = [I - k(X^t X + kI_p)^{-1}](X^t X + kI_p)^{-1}(X^t Y + kJ)$$

1. Donc

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_J(k)) &= \mathbb{E}\left[[I - k(X^t X + kI_p)^{-1}](X^t X + kI_p)^{-1}(X^t Y + kJ)\right] \\
&= [I - k(X^t X + kI_p)^{-1}](X^t X + kI_p)^{-1}]\mathbb{E}\left[(X^t Y + kJ)\right] \\
&= [I - k(X^t X + kI_p)^{-1}](X^t X + kI_p)^{-1}](X^t\mathbb{E}(Y) + k\mathbb{E}(J)) \\
&= [I - k(X^t X + kI_p)^{-1}](X^t X + kI_p)^{-1}(X^t X + kI_p)\beta
\end{aligned}$$

alors

$$\mathbb{E}(\hat{\beta}_J(k) = \beta - k(X^t X + kI_p)^{-1}\beta$$

donc l'estimateur MUR est biaisé

$$Biais(\hat{\beta}_J(k)) = -kS_k^{-1}\beta \tag{3.13}$$

où $S = X^tX$, et $S_k = (S + kI)$

2. **Variance**
On note $W = [I - kS_k^{-1}]$, on remplace dans l'équation 3.11 :

$$\hat{\beta}_J(k) = WS_k^{-1}(X^tY + kJ)$$

Donc la variance de MUR est donnée par

$$\begin{aligned} V(\hat{\beta}_J(k)) &= V[WS_k^{-1}(X^tY + kJ)] \\ &= WS_k^{-1}V(X^tY + kJ)W^tS_k^{-1} \\ &= WS_k^{-1}\sigma^2(X^tX + kI_p)W^tS_k^{-1} \\ &= \sigma^2WS_k^{-1}W^t \end{aligned}$$

alors

$$V(\hat{\beta}_J(k)) = \sigma^2WS_k^{-1}W^t \tag{3.14}$$

3. **Matrice d'erreur quadratique moyenne (MMSE)**

$$\begin{aligned} MMSE(\hat{\beta}_J(k)) &= Var(\hat{\beta}_J(k)) + [biais(\hat{\beta}_J(k))][biais(\hat{\beta}_J(k))^t] \\ &= \sigma^2WS_k^{-1}W^t + k^2S_k^{-1}\beta\beta^tS_k^{-1} \end{aligned}$$

donc

$$MMSE(\hat{\beta}_J(k)) = \sigma^2WS_k^{-1}W^t + k^2S_k^{-1}\beta\beta^tS_k^{-1} \tag{3.15}$$

4. **Erreur quadratique scalaire moyenne (SMSE)**

$$\begin{aligned} SMSE(\hat{\beta}_J(k)) &= \mathbb{E}[(\hat{\beta}_J(k) - \beta)^t(\hat{\beta}_J(k) - \beta)] \\ &= tr(MMSE(\hat{\beta}_J(k)) \end{aligned}$$

donc

$$SMSE(\hat{\beta}_J(k)) = \sum_{i=1}^{p} Var(\hat{\beta}_J(k)) + \sum_{i=1}^{p}(biais(\hat{\beta}_J(k)))^2 \tag{3.16}$$

alors

$$SMSE(\hat{\gamma}_J(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i^2}{(\lambda_i + k)^3} + k^2 \sum_{i=1}^{p} \frac{(\lambda_i + k)\gamma_i^2}{\lambda_i + k} \tag{3.17}$$

5. $\hat{\beta}_J(k = 0) = \hat{\beta}_{MC} = (X^t X)^{-1} X^t Y$ c'est l'estimateur des moindres carrés (MCO)
6. $\lim \hat{\beta}_J(k)_{k \to 0} = \hat{\beta}_{MC}$

3.6 Comparaison avec les autres estimateurs

L'estimateur MUR est biaisé et il est donc comparé à d'autres estimateurs en termes d'erreur quadratique mayonne MMSE.

3.6.1 Comparaison avec ORR

L'estimateur ridge ordiaire ORR est

$$\hat{\beta}(k) = [I - k(X^t X + kI_p)^{-1}]\hat{\beta}_{MC}$$

La variance de $\hat{\beta}(k)$ est donnée par

$$\begin{aligned} V(\hat{\beta}(k)) &= V[[I - k(X^t X + kI_p)^{-1}]\hat{\beta}_{MC}] \\ &= [I - k(X^t X + kI_p)^{-1}]V(\hat{\beta}_{MC})[I - k(X^t X + kI_p)^{-1}]^t \\ &= \sigma^2[I - k(X^t X + kI_p)^{-1}](X^t X)^{-1}[I - k(X^t X + kI_p)^{-1}]^t \end{aligned}$$

où $S = X^t X, \quad S_k = (S + kI) \quad et \quad W = [I - kS_k^{-1}]$

$$V(\hat{\beta}(k)) = \sigma^2 W S^{-1} W^t \tag{3.18}$$

d'aprés (3.13) on trouve l'erreur quadratique moyenne comme

$$MMSE(\hat{\beta}(k)) = \sigma^2 W S^{-1} W^t + k^2 S_k^{-1} \beta \beta^t S_k^{-1}$$

et d'aprés la définition de l'estimateur ORR et d'aprés (3.3) on a

$$\begin{aligned} Var(\hat{\gamma}(k)) &= var[(\Lambda + kI_p)^{-1} Z^t Y] \\ &= (\Lambda + kI_p)^{-1} Z^t Z Var(Y) (\Lambda + kI_p)^{-1} \\ &= \sigma^2 \Lambda (\Lambda + kI_p)^{-2} \end{aligned}$$

donc

$$SMSE(\hat{\gamma}(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i + k)^2} \tag{3.19}$$

Considérons

$$\begin{aligned} \Delta &= MMSE(\hat{\beta}(k)) - MMSE(\hat{\beta}_j(k)) \\ &= \sigma^2 W (S^{-1} - S_k^{-1}) W^t \\ &= \sigma^2 H \end{aligned}$$

Puisque $S_k - S = kI_p$ est définie positive (d.p), il est facile de montrer que $S^{-1} - S_k^{-1}$ est définie positive, pour $k > 0$. Par conséquence, nous avons le résultat suivant.
Résultat 1 : MMSE de L'estimateur MUR est plus petite que MMSE de ORR, Pour $k > 0$.

3.6.2 Comparaison avec URR

D'aprés la définition de l'estimateur URR (3.15), ie est un estimateur sans biais, son MMSE est donnée par

$$MMSE(\hat{\beta}(k, J)) = V(\hat{\beta}(k, J))$$

où

$$\begin{aligned} V(\hat{\beta}(k, J)) &= V[(X^t X + kI_p)^{-1} (X^t Y + kJ)] \\ &= (X^t X + kI_p)^{-1} \sigma^2 (X^t X + kI_p) (X^t X + kI_p)^{-1} \\ &= \sigma^2 (X^t X + kI_p)^{-1} \end{aligned}$$

telle que

$$S_k^{-1} = (X^t X + kI_p)^{-1}$$

donc

$$MMSE(\hat{\beta}(k,J)) = \sigma^2 S_k^{-1} \tag{3.20}$$

et on a

$$SMSE(\hat{\beta}(k,J)) = tr(MMSE(\hat{\beta}(k,J)) \tag{3.21}$$

Puisque $\hat{\gamma}(k,J)$ est un estimateur sans biais donc alors

$$MMSE(\hat{\gamma}(k,J)) = V(\hat{\gamma}(k,J))$$

où

$$\begin{aligned} V(\hat{\gamma}(k,J)) &= V[(\Lambda + kI_p)^{-1}(Z^t Y + kI_p)] \\ &= (\Lambda + kI_p)^{-1}\sigma^2(\Lambda + kI_p)^{-1}(\Lambda + kI_p) \\ &= \sigma^2(\Lambda + kI_p)^{-1} \end{aligned}$$

on a aussi $\Lambda = Z^t Z$ alors

$$MMSE(\hat{\gamma}(k,J)) = \sigma^2(\Lambda + kI_p)^{-1}$$

et alors

$$SMSE(\hat{\gamma}(k,J)) = \sigma^2 \sum_{i=1}^{p} \frac{1}{(\lambda_i + k)} \tag{3.22}$$

d'aprés (3.15)

$$\begin{aligned} \Delta &= MMSE(\hat{\beta}(k,J) - MMSE(\hat{\beta}_J(k)) \\ &= \sigma^2[S_k^{-1} - WS_k^{-1}W^t] - k^2 S_k^{-1}\beta\beta^t S_k^{-1} \\ &= S_k^{-1}[k^2\sigma^2(\frac{2}{k}I_p - S_k^{-1}) - k^2\beta\beta^t]S_k^{-1} \end{aligned}$$

où $W = [I - kS_k^{-1}]$, Δ est définie non négative (pour $k > 0$) si et seulement si $\phi = \frac{1}{k^2} S_k \Delta S_k$ est définie non négative de plus

$$\phi = \sigma^2 \left(\frac{2}{k} I_p - S_k^{-1} \right) - \beta\beta^t \tag{3.23}$$

Puisque la matrice $\frac{2}{k} I_p - S_k^{-1}$ est définie positive (d'aprés Farebrother[6]), ϕ est non négative si et seulement si

$$\beta^t [\frac{2}{k} I_p - S_k^{-1}]^{-1} \beta \leq \sigma^2 \tag{3.24}$$

D'où nous avons le résultat suivant
Résultat 2 : MUR a MMSE plus petite que celle de URR si

$$\beta^t [\frac{2}{k} I_p - S_k^{-1}]^{-1} \beta \leq \sigma^2$$

La condition du résultat (2) est verifiée par le test

$$H_0 : \beta^t [\frac{2}{k} I_p - S_k^{-1}]^{-1} \beta \leq \sigma^2$$

contre

$$H_1 = \beta^t [\frac{2}{k} I_p - S_k^{-1}]^{-1} \beta > \sigma^2$$

Puisque $\Lambda - \Lambda^*(k)$ est semi definie positive, la condition du résultat (2) devient $\beta' T \Lambda^*(k)^{-1} T' \beta \leq \sigma^2$ si $\beta^t T \Lambda^{-1} T^t \beta \leq \sigma^2$. Sous la condition de la normalité $\sigma^{-1} \Lambda^*(k)^{-\frac{1}{2}} T^t \hat{\beta}_J(k) \backsim N(\sigma^{-1} \Lambda^*(k)^{-\frac{1}{2}} (I - k\Lambda_k^{-1}) T^t \beta, \Lambda^*(k)^{-1} (I - k\Lambda_k^{-1})^2)$, et le test statistique

$$F = \frac{\hat{\beta}_j(k)^t T \Lambda^{-1} T^t \hat{\beta}_j(k)/p}{\hat{\epsilon}^t \epsilon / n - p} \sim F(p, n-p, \frac{\beta' T \Lambda^{-1} T' \beta}{2\sigma^2})$$

sous H_0. En conclusion l'erreur de l'estimateur MUR et plus petite que URR si H_0 est acceptée.

3.7 Le paramètre ridge optimal

L'erreur quadratique moyenne MMSE de l'estimateur modifié du ridge sans biais MUR dépend du paramètre k, le choix de k est crucial pour la performance de MUR. Par conséquence, on trouve des conditions sur les valeurs de k pour que l'estimateur modifié du ridge sans biais MUR peut être mieux que les autres estimateurs en termes de SMSE.

Résultat 3

Nous avons

1. $SMSE_i(\hat{\gamma}_J(k)) < SMSE_i(\hat{\gamma}(k,J)), \quad pour \quad 0 < k_i < k_{i1}$
2. $SMSE_i(\hat{\gamma}_J(k)) > SMSE_i(\hat{\gamma}(k,j)), \quad pour \quad k_{i1} < k_i < \infty$

où

$$k_{i1} = \frac{(\sigma^2 - \lambda_i\gamma_i^2)}{2\gamma_i^2} + [\frac{(\sigma^2 - \lambda_i\gamma_i^2)^2}{4\gamma_i^4} + \frac{2\sigma^2\lambda_i}{\gamma_i^2}]^{\frac{1}{2}} > 0 \tag{3.25}$$

Preuve

Le résultat (3) peut être prouvé en montrant que

$$(\lambda_i + k)^3[SMSE_i(\hat{\gamma}_J(k)) - SMSE_i(\hat{\gamma}(k,J))] = k_i[\gamma_i^2k_i^2 - (\sigma^2 - \lambda_i\gamma_i^2)k_i - 2\lambda_i\sigma^2]$$

et c'est le cas d'aprés (3.18) et (3.23).

Ensuite, on compare SMSE de $\hat{\gamma}_J(k)$ avec celle MCO. L'estimateur MUR réduit à MCO lorsque $k = 0$. la iéme composante du SMSE de γ de l'estimateur MCO est donnée par

$$SMSE_i(\hat{\gamma}_{MC}) = \frac{\sigma^2}{\lambda_i}, \quad i = 1, 2, ..., p \tag{3.26}$$

on a le résultat suivant

Résultat 4 :

Nous avons

1. Si $\lambda_i\gamma_i^2 - \sigma^2 \leq 0$, alors

$$SMSE_i(\hat{\gamma}_J(k)) < SMSE_i(\hat{\gamma}_{MC}), \quad pour \quad 0 < k_i < \infty$$

2. Si $\lambda_i\gamma_i^2 - \sigma^2 > 0$, alors il existe un k_{i2} positif, telle que

$$SMSE_i(\hat{\gamma}_J(k)) > SMSE(\hat{\gamma}_{MC}), \quad pour \quad 0 < k_i < k_{i2}$$

et

$$SMSE_i(\hat{\gamma}_J(k)), \quad pour \quad k_{i2} < k_i < \infty$$

où

$$k_{i2} = [\frac{(\lambda_i^2\gamma_i^2 - 3\sigma^2\lambda)^2}{4(\lambda_i\gamma_i^2 - \sigma^2)^2} + \frac{3\lambda_i^2\sigma^2}{(\lambda_i\gamma_i^2 - \sigma^2)}]^{\frac{1}{2}} - \frac{(\lambda_i^2\gamma_i^2 - 3\sigma^2\lambda)}{2(\lambda_i\gamma_i^2 - \sigma^2)} > 0 \quad (3.27)$$

Preuve
Le résultat (4) peut être prouvé en montrant que

$$\lambda_i(\lambda_i+k_i)^3[SMSE(\hat{\gamma}_J(k)-SMSE(\hat{\gamma}_{MC})] = k_i[(\lambda_i\gamma_i^2-\sigma^2)k_i^2+(\lambda_i^2\gamma_i^2-3\sigma^2\lambda_i)k_i-3\lambda_i^2\sigma^2]$$

c'est un résultat de (3.18) et (3.27), Par ailleur, la différenciation de $SMSE = (\hat{\gamma}_J(k))$ par rapport à k_i donne :

$$\frac{\partial SMSE(\hat{\gamma}_J(k))}{\partial k} = \frac{2\lambda_i\gamma_i^2k_i^2 + 2\lambda_i^2\gamma_i^2k_i - 3\sigma^2\lambda_i^2}{(\lambda_i + k_i)^4} = 0$$

Ainsi, la valeur optimale du paramétre ridge k est

$$k_{i(FG)} = \frac{\lambda_i}{2}[(1 - (\frac{6\sigma^2}{\gamma_i^2}))^{\frac{1}{2}} - 1] \quad (3.28)$$

d'aprés (3.26),(3.38) et (3.28) , on peut facilement vérifier que $k_{i1} < k_{i(FG)} < k_{i2}$ si $\lambda_i\gamma_i^2 - \sigma^2 > 0$. Dans le cas où $k = k_1 = k_2 = \cdots = k_p$, on obtient k comme la moyenne harmonique dans (3.28). Elle est donnée par

$$k_{FG} = \frac{p\sigma^2}{\sum_{i=1}^{p}[\gamma_i^2/[(\frac{\gamma_i^4\lambda_i^2}{4\sigma^4} + \frac{6\gamma_i^2\lambda_i}{\sigma^2})^{1/2} - \frac{\lambda_i\gamma_i^2}{2\sigma^2}]]} \quad (3.29)$$

En utilisant un argument de Horel et al [8], il est raisonnable d'adopter la moyenne harmonique des coefficients de régression. Le paramètre $k_{(FG)}$ de (3.29) dépend des paramètres inconnus γ et σ^2 à estimer

3.8 L'estimation du paramètre ridge k

Dans cette section, nous proposons de construire l'estimateur modifié du ridge sans biais (MUR) en utilisant le paramètre de ridge proposé par Hoerl et al [8] et Crouse et al [5]. Puisque la moyenne harmonique des valeurs du paramètre ridge optimal, dépend des paramètres inconnus γ et σ^2, nous utilisons dans ce cas l'estimation des moindres carrés MCO.

1. le paramètre ridge opérationelle de (3.29) est

$$\hat{k}_{FG} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}[\hat{\gamma}_i^2/[(\frac{\hat{\gamma}_i^4\lambda_i^2}{4\hat{\sigma}^4} + \frac{6\hat{\gamma}_i^2\lambda_i}{\hat{\sigma}^2})^{1/2} - \frac{\lambda_i\hat{\gamma}_i^2}{2\hat{\sigma}^2}]]} \tag{3.30}$$

   est appelé le paramètre ridge (FG)

2. le paramêtre ridge HKB de (Hoerl et al [8]) est

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\gamma}_{MC}^t\hat{\gamma}_{MC}} \tag{3.31}$$

3. le paramêtre ridge CJH (Crouse et al [5]) :

$$\hat{k}_{CJH} = \begin{cases} \frac{p\hat{\sigma}^2}{(\hat{\beta}_{MC}-j)^t(\hat{\beta}_{MC}-j)-\hat{\sigma}^2tr(X^tX)^{-1}}, & si \quad (\hat{\beta}_{MC}-j)^t(\hat{\beta}_{MC}-j) > \widehat{\sigma}^2tr(X^tX)^{-1} \\ \frac{p\hat{\sigma}^2}{(\hat{\beta}_{MC}-j)^t(\hat{\beta}_{MC}-j)}, & sinon \end{cases}$$

   où $\hat{\sigma}^2 = \frac{(Y-X\hat{\beta}_{MC})^t(Y-X\hat{\beta}_{MC})}{(n-p)}$ est un estimateur sans bias de σ^2. Notons que $\hat{k}_{CJH}$ est une généralisation de $\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}_{MC}^t\hat{\beta}_{MC}}$ de Hoerl et al. [8]

3.9 Simulations et comparaisons

Dans cet exemple on calcule l'estimateur MCO et ORR des données de Boston et on les compare au sens d'erreur quadratique. En utilisant le langage R

Exemple 1 : prix de maisons à Boston (1970)

Description des données

1. Variables de réponse : médiane des prix des maisons(medv).

2. Variables de régression :
   - taux de criminalité (crim) ;% zones terrestres pour les lots (zn)
   - % les heures d'ouverture (indus) ; 1/0 sur la rivière Charles (chas)
   - la concentration en oxyde d'azote (nox) ; nombre moyen de pièces (rm)
   - % construites avant 1940 (age) ; taux d'imposition (tax)
   - distance pondérée à centres d'emploi (dis)
   - population de statut inférieur (lstat) ;% noir (B)
   - l'accessibilité aux autoroutes radiales (rad)
   - ratio élèves / enseignant (ptratio)

Programme

library(MASS)

Boston.lm<- lm(medv∽., data=Boston)

summary(Boston.lm)

Call :

lm(formula = medv∽ ., data = Boston)

Residuals :

Min	$1Q$	$Median$	$3Q$	Max
−15.594	−2.730	−0.518	1.777	26.199

Coefficients :

	Estimate	*Std.Error*	*tvalue*	Pr(>|t|)
$(Intercept)$	$3.646e+01$	$5.103e+00$	7.144	$3.28e-12$
$crim$	$-1.080e-01$	$3.286e-02$	-3.287	0.001087
zn	$4.642e-02$	$1.373e-02$	3.382	0.000778
$indus$	$2.056e-02$	$6.150e-02$	0.334	0.738288
$chas$	$2.687e+00$	$8.616e-01$	3.118	0.001925
nox	$-1.777e+01$	$3.820e+00$	-4.651	$4.25e-06$
rm	$3.810e+00$	$4.179e-01$	9.116	$<2e-16$
age	$6.922e-04$	$1.321e-02$	0.052	0.958229
dis	$-1.476e+00$	$1.995e-01$	-7.398	$6.01e-13$
rad	$3.060e-01$	$6.635e-02$	4.613	$5.07e-06$
tax	$-1.233e-02$	$3.760e-03$	-3.280	0.001112
$ptratio$	$-9.527e-01$	$1.308e-01$	-7.283	$1.31e-12$
$black$	$9.312e-03$	$2.686e-03$	3.467	0.000573
$lstat$	$-5.248e-01$	$5.072e-02$	-10.347	$< 2e-16$

Residual standard error : 4.745 on 492 degrees of freedom Multiple R-Squared : 0.7406, Adjusted R-squared : 0.7338 F-statistic : 108.1 on 13 and 492 DF, p-value : < 2.2e-16

D'aprés les résultats précédents

indus et age ne sont pas significatifs au niveau 0.05

fmBoston=as.formula("medv∽., crim+zn+chas+nox+rm+dis+rad+tax+ptration+black+lstat")

Boston1.lm <- lm(fmBoston∽., data=Boston)

summary(Boston1.lm) Call : lm(formula = fmBoston, data = Boston) Residuals :

Min	$1Q$	*Median*	$3Q$	*Max*
-15.5984	-2.7386	-0.5046	1.7273	26.2373

Coefficients :

	Estimate	*Std.Error*	*tvalue*	Pr(>|t|)
$(Intercept)$	36.341145	5.067492	7.171	$2.73e-12$
$crim$	-0.108413	0.032779	-3.307	0.001010
zn	0.045845	0.013523	3.390	0.000754
$chas$	2.718716	0.854240	3.183	0.001551
nox	3.801579	0.406316	9.356	$<2e-16$
rm	3.801579	0.406316	9.356	$<2e-16$
dis	-1.492711	0.185731	-8.037	$6.84e-15$
rad	0.299608	0.063402	4.72	$63.00e-06$
tax	-0.011778	0.003372	-3.493	0.000521
$ptratio$	-0.946525	0.129066	-7.334	$9.24e-13$
$black$	0.009291	0.002674	3.475	0.000557
$lstat$	-0.522553	0.047424	-11.019	$<2e-16$

Residual standard error : 4.736 on 494 degrees of freedom Multiple R-Squared : 0.7406, Adjusted R-squared : 0.7348 F-statistic : 128.2 on 11 and 494 DF, p-value : $<$ 2.2e-16

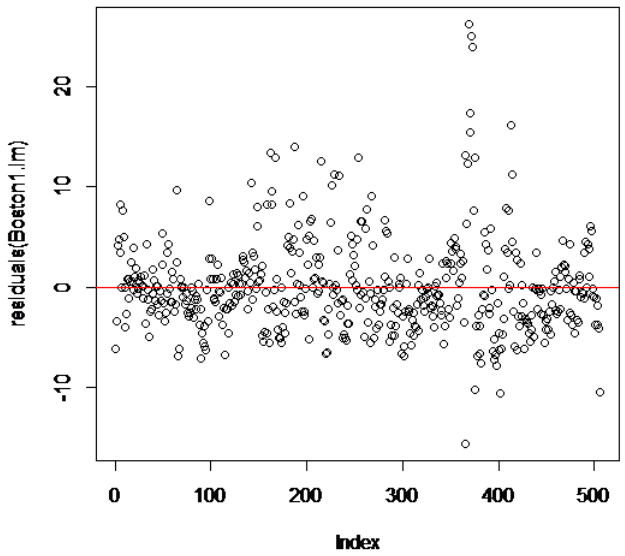
Graphiques de résidus :

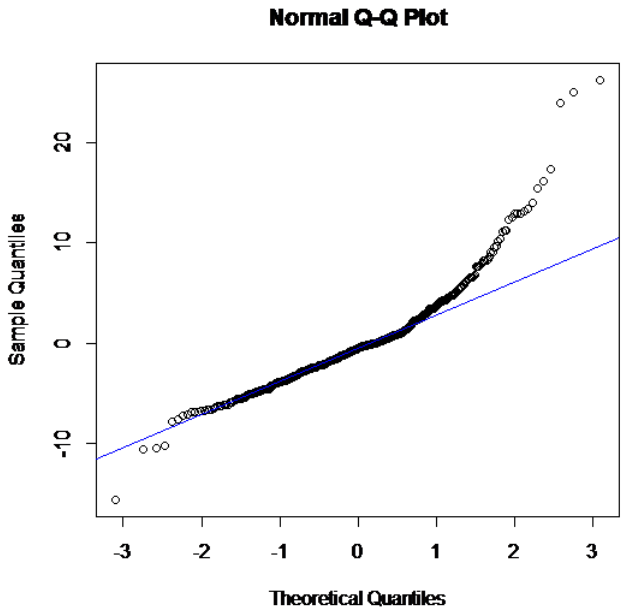
plot(residuals(Boston1.lm))

abline(0,0)

qqnorm(residuals(Boston1.lm))

qqline(residuals(Boston1.lm))





Régression "ridge"sous R

Description de la fonction lm.ridge

Valeurs à la sortie

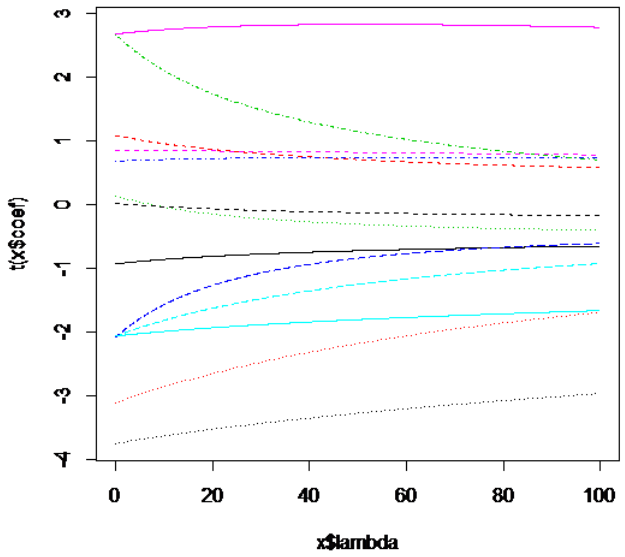
– scales - paramètres d'échelle de la matrice X.

– lambda - vecteur de valeurs de λ.
– ym - moyenne de Y.
– xm - moyenne des colonnes de X.
– GCV - vecteur des valeurs du $V(\lambda)$(GCV).
– kHKB - estimation de la constante du ridge par HKB.
– kLW - estimation de la constante du ridge par L-W.

Boston.ridge<-lm.ridge(medv∽., Boston, lambda=seq(0, 100, 0.1))

plot(Boston.ridge)

Tracé du ridge :



select(Boston.ridge)

modified HKB estimator is 4.594163
modified L-W estimator is 3.961575
smallest value of GCV at 4.3

Boston.ridge.cv<-lm.ridge(medv∽.,Boston,lambda=4.3)

Boston.ridge.cvcoef

crim	*zn*	*indus*	*chas*	*nox*
−0.895001937	1.020966996	0.049465334	0.694878367	−1.943248437
age	*dis*	*rad*	*tax*	*ptratio*
−0.005646034	−2.992453378	2.384190136	−1.819613735	−2.026897293
lstat	*rm*	*black*		
−3.689619529	2.707866705	0.847413719		

L'erreur quadratique de ridge "MSE"

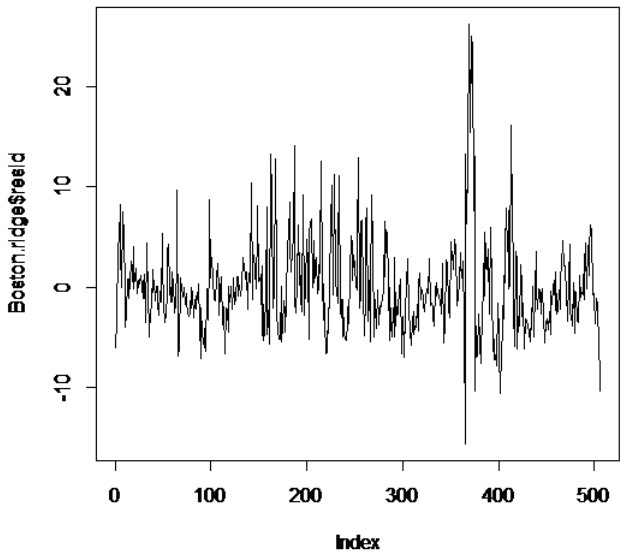
```
lm.ridge1=function (formula, data, subset, na.action, lambda = 0, model = FALSE,
x = FALSE, y = FALSE, contrasts = NULL, ...)
{
m <- match.call(expand.dots = FALSE)
m$model <- m$x <- m$y <- m.contrasts <- m.... <- m$lambda <- NULL
m[[1L]] <- quote(stats : :model.frame)
m <- eval.parent(m)
Terms <- attr(m, "terms")
Y <- model.response(m)
X <- model.matrix(Terms, m, contrasts)
n <- nrow(X)
p <- ncol(X)
offset <- model.offset(m)
if ( !is.null(offset))
Y <- Y - offset
if (Inter <- attr(Terms, "intercept"))
{
Xm <- colMeans(X[, -Inter])
Ym <- mean(Y)
p <- p - 1
X <- X[, -Inter] - rep(Xm, rep(n, p))
Y <- Y - Ym
}
else Ym <- Xm <- NA
```

$Xscale < -drop(rep(1/n, n) * X^2)^{0.5}$

```
X <- X/rep(Xscale, rep(n, p))
Xs <- svd(X)
rhs <- t(Xs$u) %*% Y
d <- Xs$d
lscoef <- Xs$v %*%(rhs/d)
lsfit <- X %*%lscoef
resid <- Y - lsfit
s2 <- −sum(resid^2)/(n − p − Inter)
s3 = sum(resid^2)/n

HKB <- −(p − 2)% ∗ %s2/sum(lscoef^2)
LW <- −(p − 2)% ∗ %s2% ∗ %n/sum(lsfit^2)
k <- length(lambda)
dx <- length(d)
div <- −d^2 + rep(lambda, rep(dx, k))
a <- drop(d * rhs)/div
dim(a) <- c(dx, k)
coef <- Xs$v %* % a

dimnames(coef) <- list(names(Xscale), format(lambda))
GCV <- −colSums((Y − X% ∗ %coef)^2)/(n − colSums(matrix(d^2/div, dx)))^2
res <- list(s3=s3,resid=resid,s=s2,coef = drop(coef), scales = Xscale, Inter = Inter,
lambda = lambda, ym = Ym, xm = Xm, GCV = GCV, kHKB = HKB, kLW = LW)
class(res) <- "ridgelm"
res
}
Boston.ridge<-lm.ridge1(medv∽., Boston, lambda=seq(0, 100, 0.1))
Boston.ridge$resid
plot(Boston.ridge$resid, type="l")
```



Boston.ridge$s3

Boston.ridge$s

Comparaison de MCO avec ridge

Boston

attach(Boston)

Boston.ridge<-lm.ridge(medv∽crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat, lambda=4.3)

Boston.ridge

	crim	*zn*	*chas*	*nox*
34.918471472	−0.104315656	0.043678727	2.747912270	−16.683126171
rm	*dis*	*rad*	*tax*	*ptratio*
3.851764685	−1.426638997	0.272261020	−0.010629921	−0.935344659
black	*lstat*			
0.009278324	−0.516975577			

X.matrix <- cbind(rep(1,length=length(medv)),crim,zn,chas,nox,rm,dis,rad,tax,ptratio,black,lstat)

fitted.vals <- X.matrix %*% c(34.918471472 , -0.104315656, 0.043678727, 2.747912270, -16.683126171, 3.851764685 ,-1.426638997 , 0.272261020 ,-0.010629921 , -0.935344659, 0.009278324 ,-0.516975577)

$sse.ridge < -sum((medv - fitted.vals)^2)$

sse.ridge

11088.5

$sum(resid(bodyfat.reg)^2)$

11081.36

SSE pour régression ridge est pas beaucoup plus élevé que celle du MCO.

Exemple 2

On utilise maintenant les donées bodyfat

$bodyfat.data < -read.table(file = "C : \Users\Desktop\bodyfatdata.txt", header = FALSE, col.names = c("triceps","thigh","midarm","bodyfat"))$

attach(bodyfat.data)

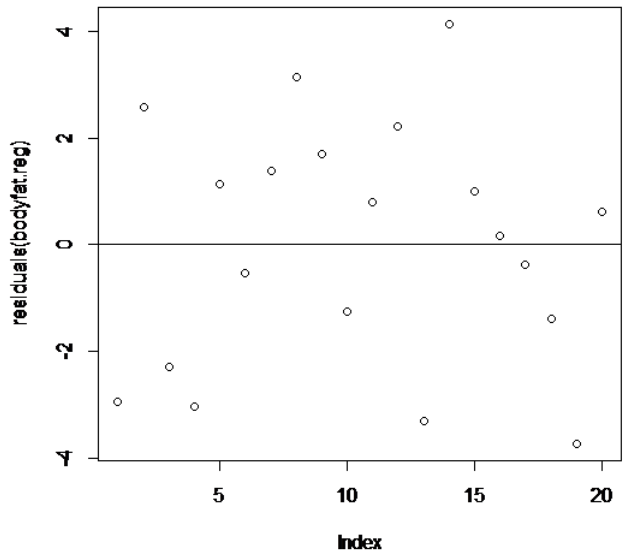
bodyfat.reg <- lm(bodyfat $\sim$ triceps + thigh + midarm)

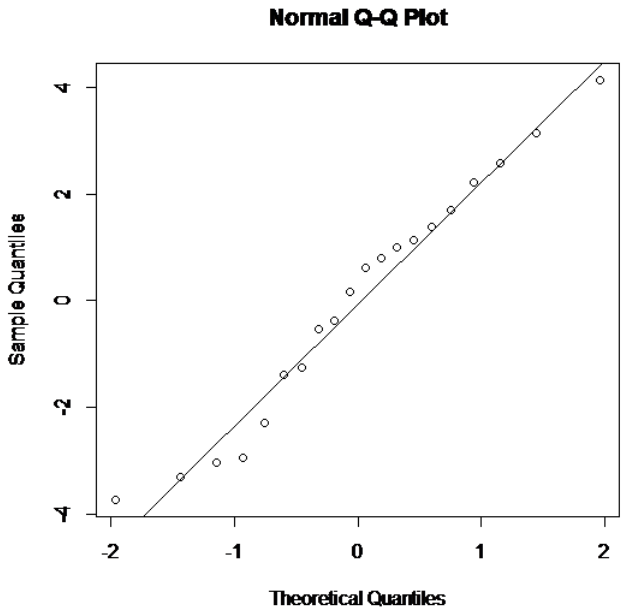
plot(residuals(bodyfat.reg))

abline(0,0)

qqnorm(residuals(bodyfat.reg))

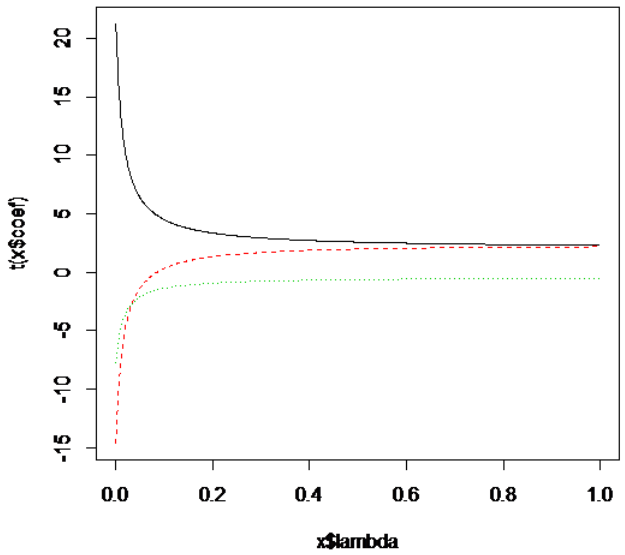
qqline(residuals(bodyfat.reg))





Régression "ridge"sous R

bodyfat.ridge.reg <- lm.ridge(bodyfat triceps + thigh + midarm, lambda =seq(0.1,0.001))
plot(bodyfat.ridge.reg)



select(lm.ridge(bodyfat $\sim$ triceps + thigh + midarm, lambda = seq(0,1,0.001)))
modified HKB estimator is 0.008505093
modified L-W estimator is 0.3098511
smallest value of GCV at 0.019
bodyfat.ridge.reg <- lm.ridge(bodyfat triceps + thigh + midarm, lambda = .019)

	triceps	*thigh*	*midarm*
43.8401126	2.1174933	−0.9597309	−1.0180612

Comparaison la regression ridge avec moindres carrés

X.matrix <- cbind(rep(1,length=length(bodyfat)),triceps, thigh, midarm)

les valeurs ajustées pour regression de ridge :

fitted.vals <- X.matrix %*% c(43.840113, 2.117493, -0.959731, -1.018061)

SSE pour regression ridge :

$sse.ridge < -sum((bodyfat - fitted.vals)^2)$

sse.ridge

101.7287

SSE pour les moindres carrés :

$sum(resid(bodyfat.reg)^2)$

98.40489

SSE pour régression ridge n'est beaucoup plus élevée que celle du MCO.

Bibliographie

[1] F. Batah and S. Gore, Improving Precision for Jackknifed Ridge Type Estimation, Far East Journal of Theoretical Statistics 24 (2008), 157 - 174. MR2474325. Zbl pre05495192.

[2] F. Batah, S. Gore and M. Verma, Effect of Jackknifing on Various Ridge Type Estimators, Model Assisted Statistics and Applications 3 (2008a), 121 - 130.

[3] F. Batah, T. Ramanathan and S. Gore, The Efficiency of Modified Jackknife and Ridge Type Regression Estimators : A Comparison, Surveys in Mathematics and its Applications 3 (2008b), 111 - 122. MR2438671.

[4] F. Batah, S. Gore and M. ¨Ozkale, Combining Unbiased Ridge and Principal Component Regression Estimators, Communication in Statistics - Theory and Methods 38 (2009), 2201 - 2209.

[5] R. Crouse, C. Jin and R. Hanumara, Unbiased Ridge Estimation With Prior Information and Ridge Trace, Communication in Statistics - Theory and Methods 24 (1995), 2341 - 2354. MR1350665(96i :62073). Zbl 0937.62616.

[6] R. Farebrother, Further Results on the Mean Squared Error of the Ridge Regression, Journal of the Royal Statistical Society B. 38 (1976), 248-250.

[7] A. Hoerl and R. Kennard, Ridge Regression : Biased Estimation for Non orthogonal Problems, Technometrics 12 (1970), 55 - 67. Zbl 0202.17205

[8] A. Hoerl, R. Kennard and K. Baldwin, Ridge Regression : Some Simulations, Communication in Statistics - Theory and Methods 4 (1975), 105-123.

[9] A. Hoerl and R. Kennard, Ridge Regression : 1980 Advances, Algorithms, and Applications, Amer.J. Mathemat. Manage. Sci. 1 (1981), 5 - 83. MR0611894(82d :62117). Zbl 0536.62054.

[10] M. ¨Ozkale and S. Ka¸ciranlar, Comparisons of the Unbiased Ridge Estimation to the Other Estimations, Communication in Statistics - Theory and Methods

[11] N. Sarkar, Mean Squared Error Matrix Comparison of Some Estimators in Linear Regression With Multicollinearity, Statistics and Probabilty Letters 21 (1996), 1987 -2000.

[12] B. Swindel, Good Ridge Estimators Based on Prior Information, Communication in Statistics - Theory and Methods 11 (1976), 1065 - 1075. MR0440806(55 :13675). Zbl 0342.62035