Projet de statistique
Sujet : Étude de la régression en grande dimension via l'estimateur LASSO

PERILLAT PIRATOINE Henri

MONTMEAT Florian

2A ENSAE

Décembre 2018

Question 1 :

Nous disposons d'un échantillon de données $(x_i, Y_i)_{i=1,...,n}$, chaque couple étant à valeurs dans $\mathbb{R}^p \times \mathbb{R}$. Nous considérons un modèle de régression avec erreurs gaussiennes, ce qui s'écrit matriciellement : Y = Xβ + ϵ où ϵ $\sim \mathcal{N}(0, \sigma^2 I_n)$ avec $\epsilon = (\epsilon_1, ...,\epsilon_n)'$,Y = $(Y_1, ..., Y_n)'$ et $X = (x'_1, ..., x'_p) \in \mathrm{R}^{n*p}$. Les x_i sont déterministes et la variance du bruit est connue.

L'estimateur des moindres carrés vise à minimiser l'erreur quadratique, c'est à dire : $\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{argmin} ||Y - X\beta||_2^2$

On a que $||Y - X\beta||_2^2 = (Y - X\beta)'(Y - X\beta)$
Donc

$$\frac{d||Y - X\beta||_2^2}{d\beta} = 0 \Leftrightarrow -2X'(Y - X\beta) = 0 \Leftrightarrow X'Y = X'X\beta \quad (1)$$

On se pose la question suivante : la matrice (X'X) est-elle inversible ?
On sait que rang(X'X) min(n,p), et que la matrice X'X est de dimension p*p, donc si p<n, alors la matrice est inversible.
On a donc lorsque p>n, $\hat{\beta} = (X'X)^{-1}X'Y$.

Si p>n, alors X'X est non inversible et il y a donc plusieurs solutions au problème de minimisation. L'estimateur des moindres carrés n'est donc plus unique.

Question 2 :

Même lorsque n>p, si le nombre de variables est élevé mais que l'estimateur des moindres carrés reste unique, notre modèle de prédiction est assez flou. Il est clairement plus facile d'arriver à interpréter les coefficients de β lorsque p=2 (il s'agit simplement d'une droite de régression) par exemple, que lorsque p=15 où on ne peut plus interpréter en disant comment les X_i impactent les Y_i.
Lorsque p>n, comme l'estimateur des moindres carrés n'est pas unique, on ne peut plus interpréter.

Question 3 :

La pénalisation en statistique a pour but de contracter la valeurs des coefficients en les contenant dans une boule, ou en en limitant le nombre d'éléments non nuls par exemple. Ici, par le biais de l'élément $\lambda||\beta||_1$ on va pénaliser le modèle quand il y a trop de coefficients de β non nul.

Question 4 :

On va montrer que résoudre le problème LASSO est en fait équivalent à résoudre un problème de minimisation sous contrainte. L'estimateur LASSO vérifie le problème : $\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{argmin} \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$

Soit le nouveau problème : $\hat{\beta} \in \underset{\beta \in \mathbb{R}^p, ||\beta||_1 \leqslant t}{argmin} ||Y - X\beta||_2^2$

Par le théorème de Kuhn et Tucker, résoudre ce problème de minimisation sous contrainte revient à annuler la dérivée du Lagrangien du problème, où le Lagrangien s'écrit : $L(\beta) = ||Y - X\beta||_2^2 + \lambda(||\beta||_1 - t)$.

Or le Lagrangien du problème $\underset{\beta \in \mathbb{R}^p}{min} \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$ s'écrit $L(\beta) = \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$ où la condition du premier ordre est la même que celle du problème précédent (seul le coefficient $\frac{1}{2n}$ change, et il s'élimine donc ne change pas la résolution du problème).

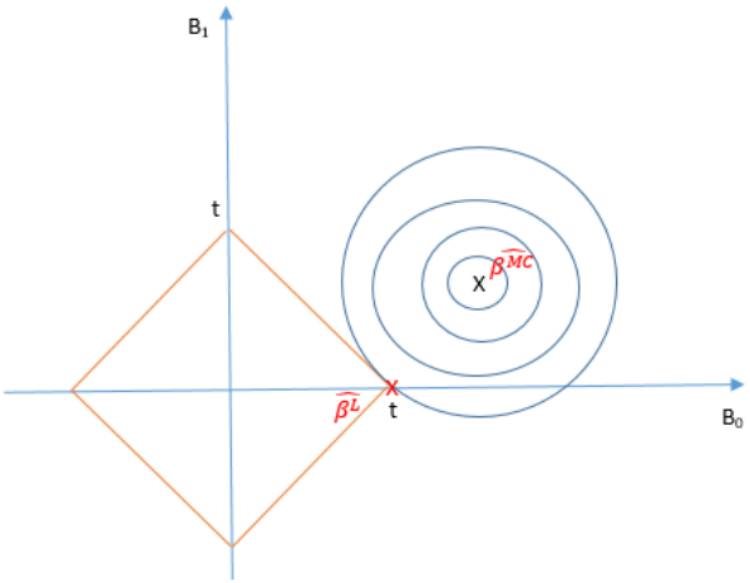
Donc ces 2 problèmes sont équivalents pour une valeur de λ donnée. La variable t dépend de λ.

Question 5 :

On aurait pu se poser la question, au lieu de mettre comme contrainte $||\beta||_1 \leqslant t$, de mettre plutôt $||\beta||_0 \leqslant s$ où la norme 0 correspond au nombre de β_j non nuls. Mais la minimisation de ce problème n'a pas de solution

explicite car ce problème n'est pas convexe. Et de plus il n'y a pas d'algorithme simple pour minimiser ce problème en un temps raisonnable.

Question 6 :



On observe géométriquement en dimension 2 que l'estimateur LASSO est fixé sur une des coordonnées lorsqu'on force une autre à être nulle.

Question 7 :

On cherche $\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{argmin} \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$

Tout d'abord, on introduit la définition du sous-différentiel :

Définition : $\partial \mathcal{N}_1(x_0) = \{x \in \mathbb{R}^p, x_j = sign(x_{0,j}) \ si \ x_{0,j} \neq 0, x_j \in [-1,1] \ si \ x_{0,j} = 0\}$
alors $x_0 \in argmin(f(x) + \lambda||x||_1)$ ssi $\delta \in \partial \mathcal{N}_1(x_0)$ tel que $f'(x_0) + \lambda\delta = 0$

Donc pour résoudre le problème LASSO, $\exists \delta \in \partial \mathcal{N}_1(\hat{\beta_\lambda}^{(L)})$ tel que $\frac{X'X}{n}\hat{\beta_\lambda}^{(L)} - \frac{1}{n}X'Y + \lambda\delta = 0$
Donc $\hat{\beta_\lambda}^{(L)} = \frac{1}{n}X'Y - \lambda\delta$ sous $\frac{X'X}{n} = I_p$, et alors $X'Y = n\hat{\beta}^{(MC)}$.

Donc la forme finale de l'estimateur LASSO est : On note $\hat{\beta_\lambda}^{(L)} = \begin{pmatrix} \hat{\beta}_{\lambda,1}^{(L)} \\ \vdots \\ \hat{\beta}_{\lambda,p}^{(L)} \end{pmatrix}$

Donc si $\hat{\beta}_{\lambda,j}^{(L)} \neq 0$ on en déduit $\hat{\beta}_{\lambda,j}^{(L)} = (\frac{1}{n}X'Y)_j - \lambda \ sign(\hat{\beta}_{\lambda,j}^{(L)})$
Et donc cela implique $sign((\frac{1}{n}X'Y)_j) = sign(\hat{\beta}_{\lambda,j}^{(L)}) \ si \ |(\frac{1}{n}X'Y)_j| > \lambda$
Si on a pas $|(\frac{1}{n}X'Y)_j| > \lambda$ alors $sign((\frac{1}{n}X'Y)_j) \neq sign(\hat{\beta}_{\lambda,j}^{(L)})$

Dans le cas où $|(\frac{1}{n}X'Y)_j| \leqslant \lambda$ alors on ne peut pas supposer $\hat{\beta}_{\lambda,j}^{(L)} \neq 0$ et donc on aura $\hat{\beta}_{\lambda,j}^{(L)} = 0$
Donc

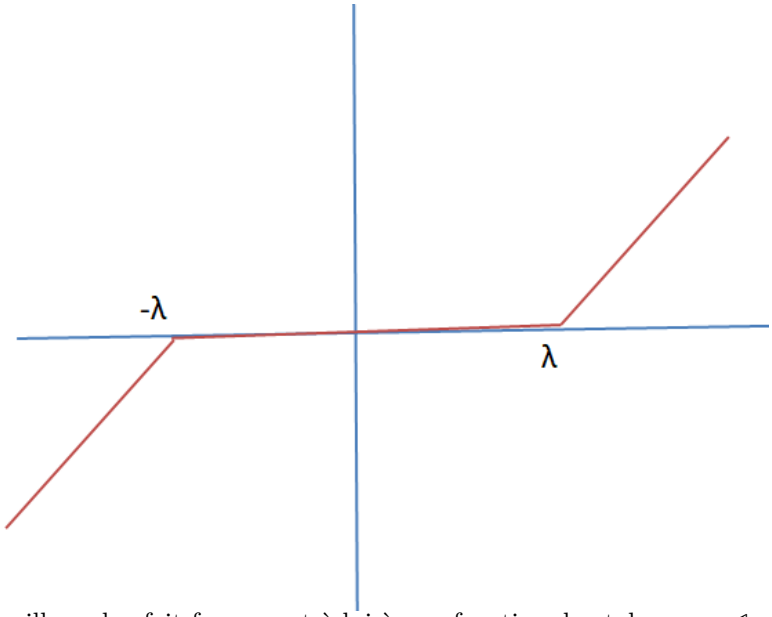
$$\hat{\beta}_{\lambda,j}^{(L)} = \begin{cases} (\frac{1}{n}X'Y)_j - \lambda \ sign((\frac{1}{n}X'Y)_j) \ si \ |(\frac{1}{n}X'Y)_j| > \lambda, \\ 0 \quad sinon, \end{cases} \tag{2}$$

On déduit $\hat{\beta}_{\lambda,j}^{(L)} = (\frac{1}{n}X'Y)_j(1 - \frac{n\lambda}{|(X'Y)_j|})_+$

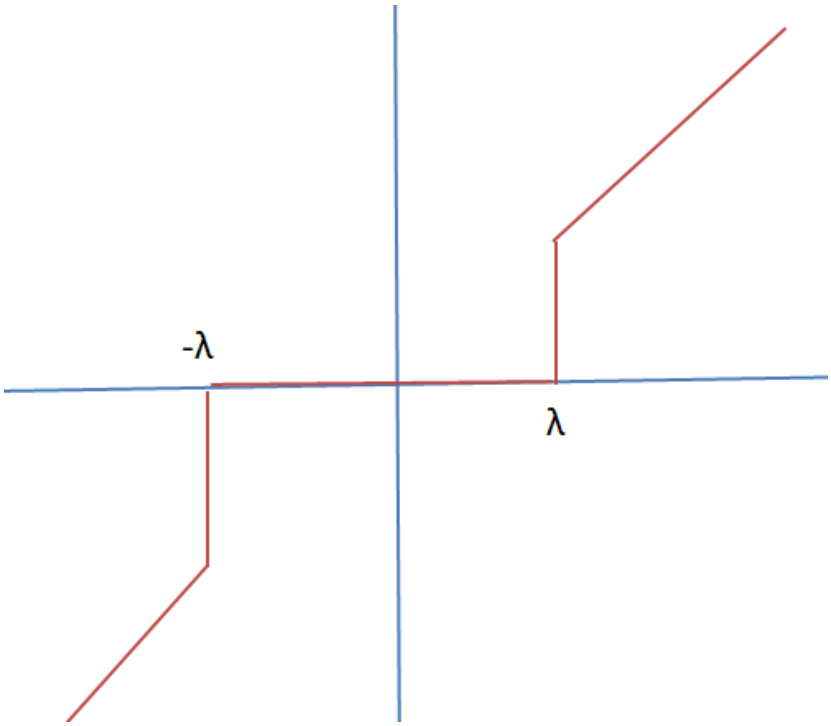
Or on sait que si $\frac{X'X}{n} = I_p$ alors $X'Y = n\hat{\beta}^{(MC)}$ donc la forme finale de l'estimateur LASSO est :

$$\hat{\beta}_{\lambda,j}^{(L)} = \hat{\beta}_j^{(MC)}(1 - \frac{\lambda}{|\hat{\beta}_j^{(MC)}|})_+ \tag{3}$$

On peut parler de seuillage doux et de seuillage dur. Ici nous faisons face à un seuillage doux du fait que $\hat{\beta}_{\lambda,j}^{(L)}$ est continue. C'est à dire que $\hat{\beta}_{\lambda,j}^{(L)}$ n'arrivera pas brusquement à 0, mais de manière douce contrairement au seuillage dur où il y a une marche pour arriver à 0. Nous allons illustré ceci par les graphes du seuillage doux et du seuillage dur.
Un seuillage doux fait face à une fonction du style : $y = sign(x)(|x| - \lambda)$



Un seuillage dur fait face quant à lui à une fonction du style : $y = x\mathbb{1}_{(|x|>\lambda)}$



Question 8 :

On considère que $\beta \sim f$ a priori où $f(\beta) = exp(-\frac{\lambda}{2\sigma^2}||\beta||_1)$

Par les formules de Bayes on sait que β suit à posteriori $f_{Y|\beta}(y)f_\beta(\beta)$ où Y $\sim \mathcal{N}(X\beta,\, \sigma^2 I_n)$

$f_{\beta|Y}(\beta) \propto exp(-\frac{(Y-X\beta)'(Y_X\beta)}{2\sigma^2})exp(-\frac{\lambda}{2\sigma^2}||\beta||_1) = exp(-\frac{1}{2\sigma^2}(||Y-X\beta||_2^2 + \lambda||\beta||_1))$

Il est donc clair que minimiser $||Y-X\beta||_2^2+\lambda||\beta||_1$ en β et donc trouver l'estimateur LASSO revient à maximiser la loi a posteriori de β.

Question 9 :

Précision relative au code : Dans le code, la première partie sera composé de la partie implémentée sans les fonctions R. Les résultats trouvés sont relativement similaire dans l'ensemble à ceux trouvé par les fonctions R. Par ailleurs cette partie du code ayant été faite après avoir d'abord résolu les questions avec les fonctions R, les graphiques du Latex sont obtenus avec les fonctions R.

L'algorithme dit de coordinate descent sert pour l'optimisation de fonction convexe multivariée f. Il s'agit d'un algorithme de descente mais qui à l'itération k+1 pour le calcul de la coordonnée j se sert des coordonnées $i \in [1, j-1]$ calculées précédemment à l'étape k+1 et des coordonnées $i \in [j+1, p]$ calculées précédemment à l'étape k.

$$x_j^{k+1} \in \underset{y\in\mathbb{R}}{argmin} f(x_1^{k+1}, ..., x_{j-1}^{k+1}, y, x_{j+1}^k, x_p^k) \tag{4}$$

L'algorithme de descente de l'estimateur LASSO pourra donc s'écrire :
Début
Initialiser $\beta = \beta_{init}$

Répéter jusqu'à convergence
Pour j = 0,...,p, on pose $\beta_j^{k+1} = R_j(1 - \frac{\lambda}{|R_j|})_+$ avec $\quad R_j = \frac{1}{n}\sum_{i=1}^n x_{ij}(y_i - \sum_{h<j} x_{ih}\beta_h^{k+1} - \sum_{h>j} x_{ih}\beta_h^k)$

Fin Pour
Condition d'arrêt : on s'arrête en k+1 si $||\hat{\beta}^{k+1} - \hat{\beta}^{k+1}|| < \epsilon$ ou si le nombre d'itérations est supérieur à une limite préalablement fixée
Retourner β
Fin

Question 11 :

La méthode par validation croisée répète n fois une méthode de choix de λ pour une méthode classique où l'on a divisé notre échantillon en X_{train} et X_{test}, et où l'on a choisi le λ qui minimisait l'erreur de prédiction sur l'échantillon de test.

Nos deux échantillons sont $I_{train} = \{j \neq i, j = 1, ..., n\}$ et $I_{test} = \{i\}$. Donc pour une validation croisée, on réalise pour i = 1,...,n :

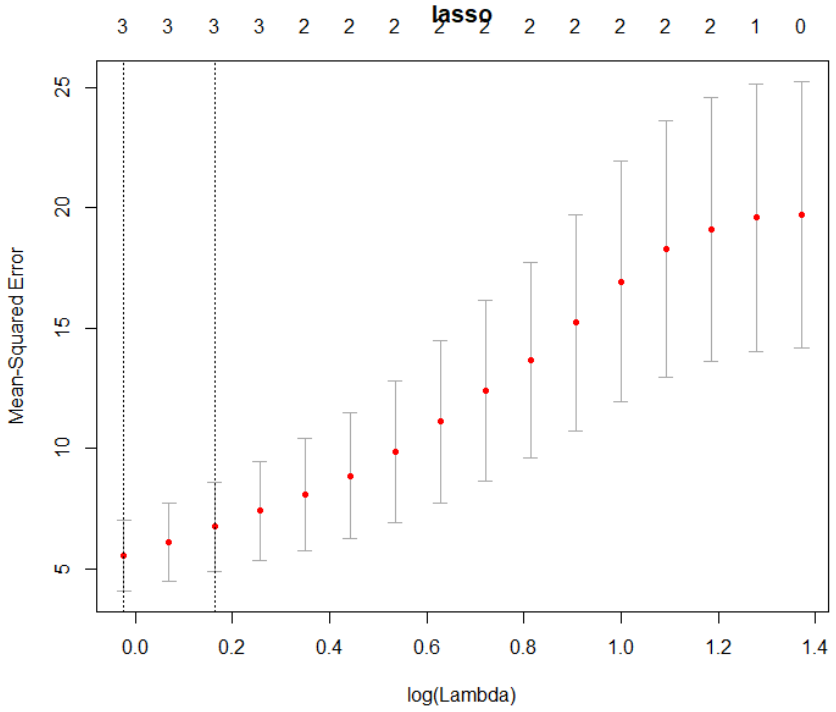
- Pour tout $\lambda \in \Lambda$, on calcule $\hat{\beta}_\lambda^{(L_j - i)}$ à partir de $\{(x_j, y_j), j \neq i\}$
- On calcule $\hat{Y}_{i,\lambda}^{(L_j-i)} = x_i'\hat{\beta}_\lambda^{(L_j-i)}$

Puis on prend :

$$\hat{\lambda}^{(L)} \in \underset{\lambda\in\Lambda}{argmin} \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_{i,\lambda}^{(L_j-i)})^2 \tag{5}$$

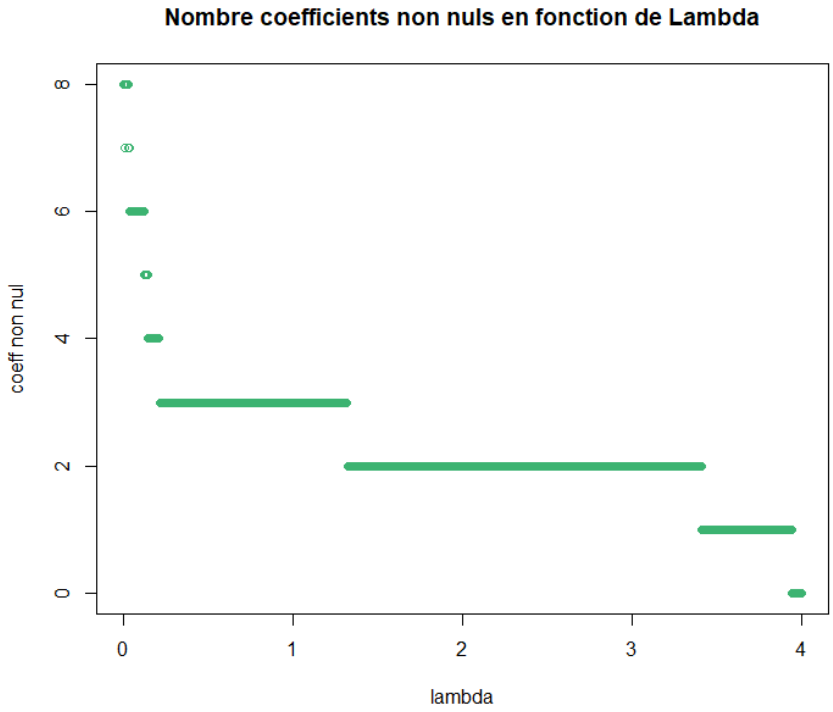
L'estimateur final est donc $\hat{\beta}^{(L)}_{\hat{\lambda}^{(L)}}$.

Nous avons grâce à la fonction cv.glmnet de R tracé la valeur des erreurs quadratiques en fonction de $log(\lambda)$. Cette fonction nous donne directement la valeur de λ qui minimise le problème, et qui est donc le $\hat{\lambda}$ souhaité. Ici nous avons trouvé $\hat{\lambda} = 0.97$.
Puis en simulant la régression sur l'échantillon avec ce λ trouvé par validation croisée, nous obtenons les coefficients suivants : $\hat{\beta} = (2.18, 0.4, 0, 0, 1.01, 0, 0, 0)$

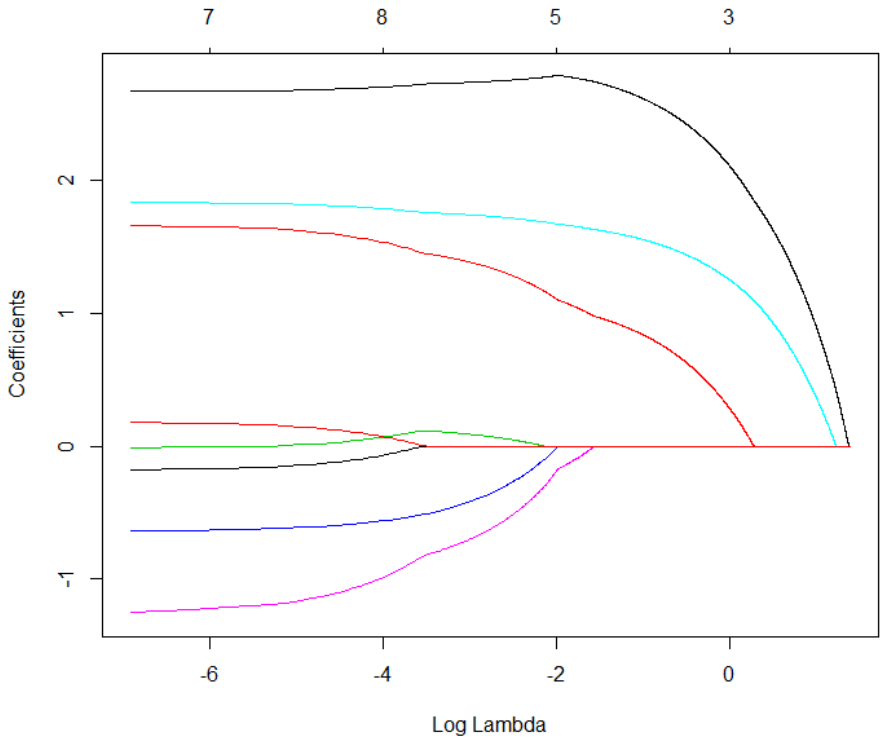


Question 12 :

On remarque graphiquement que plus le λ est élevé, plus le nombre de coefficients non nuls est faible. Ceci est logique dans le sens où plus le λ est élevé et plus on pénalise, donc plus on force des coefficients à être nuls. On observe que à partir de $\lambda = 4$ environ, le modèle ne sélectionne plus aucune variable. De plus pour le λ trouvé par validation croisée, le modèle sélectionne 3 variables. on observe bien que pour $\hat{\lambda}^{(L)} = 0.97$, alors 3 variables sont sélectionnées.



Sur ce second graphique, nous observons les valeurs des coefficients en fonction de la valeur de $log(\lambda)$. On observe bien que plus λ augmente et plus on fait une sélection restrictive des variables.

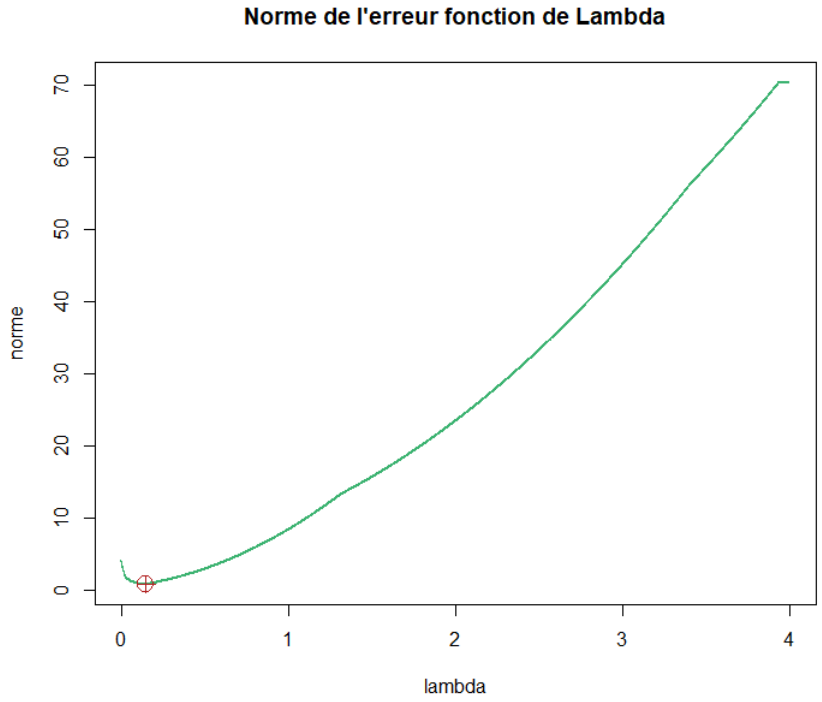


pour la validation croisée et les résultats qui suivent, il est important de faire remarquer que en simulant sous R, le nombre d'individus étant tellement faible, le résultat sorti diffère suivant les x.train et x.test choisi. Donc en réalisant d'autre simulation, il est possible de trouver un résultat différent.

Question 13 :

On suppose que les données réelles ont été simulées suivant $\beta^0 = (3, 1.5, 0, 0, 2, 0, 0, 0)$. Or lors de la validation croisée, le $\hat{\beta}$ obtenu ne comportait que 3 éléments non nuls et était égal à (2.13, 0.3, 0, 0, 1.26, 0, 0, 0).

Ici, en ayant tracé $||X(\hat{\beta}_\lambda - \beta^0)||$ en fonction de λ on remarque que le minimum de cette fonction est différent de celui trouvé par validation croisée question 11. Sur cette courbe on remarque bien que la courbe de l'erreur fait une cloche, comme on s'y attendait. Pour un $\lambda = 0$, cela correspond bien sûr à une régression classique et on observe que la minimum est observé en $\lambda = 0.199$. Ensuite, plus la valeur de λ augmente et plus l'erreur croit, ce qui est cohérent car on a observé que lorsque λ est trop fort, il réduit toutes les valeurs des coefficients à 0.



Sur le graphique on observe clairement que pour $\lambda = 0$, il s'agit d'un sur-apprentissage et donc que l'erreur n'est pas optimale. Par ailleurs, lorsque λ est grand, il s'agit de sous-apprentissage, tous les coefficients de β étant fixés à 0 (on prédit donc tout le temps $\hat{Y} = 0$. Donc, la fonction tracée est égale à $||X\beta^0|| = 70$ ici.

Enfin, lorsque l'on choisit finalement de faire la régression avec le λ optimal trouvé pour cette question, nous obtenons $\beta = (2.79, 1.09, 0, 0, 1.67, -0.15, 0, 0)$ qui sélectionne donc 4 variables ici contrairement à celui de la validation croisée qui sélectionnait 3 variable comme β^0. Par contre les variables 1,2 et 5 sélectionnées ont un coefficient β_i qui se rapproche plus de celui avec lequel les variables ont réellement été simulées.

Question bonus :

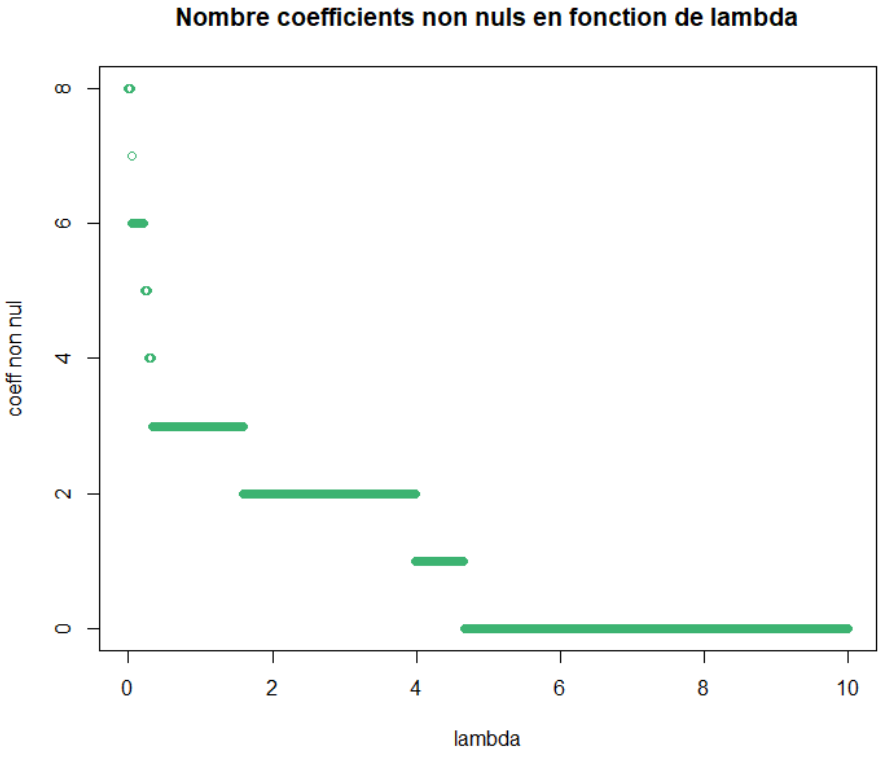
Le principe Elastic Net est une combinaison des algorithmes LASSO et Ridge. Ridge fonctionne sous le même principe que LASSO, mais en pénalisant avec la norme 2 des β.
L'estimateur Elastic Net prend la forme suivante :

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{argmin} ||Y - X\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2 \tag{6}$$

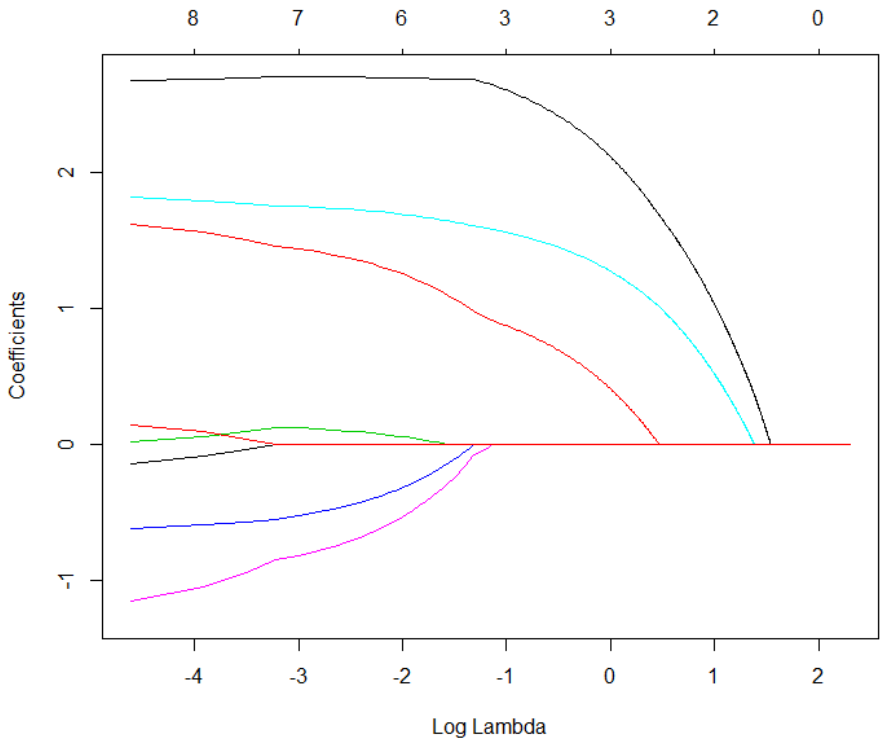
Un problème alternatif est : $\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{argmin} ||Y - X\beta||_2^2 + \lambda((1-\alpha)||\beta||_1 + \alpha||\beta||_2^2)$
Dans le deuxième problème alternatif, nous voyons bien que le cas $\alpha = 0$ correspond à l'estimation Lasso et que le cas $\alpha = 1$ correspond à l'estimation Ridge.

Le Lasso est connu pour que le nombre de variables sélectionnées par le Lasso soit limité par la taille de l'échantillon, donc ce nombre ne peut pas dépasser n. De plus, le Lasso a tendance à sélectionner une seule variable parmi un groupe de variables très corrélées. Enfin, les prédictions du Lasso sont nettement moins bonnes que celle de l'estimateur Ridge si les variables sont très corrélées, donc avec l'Elastic Net on enlève ce risque en faisant une combinaison des 2 pénalisations.

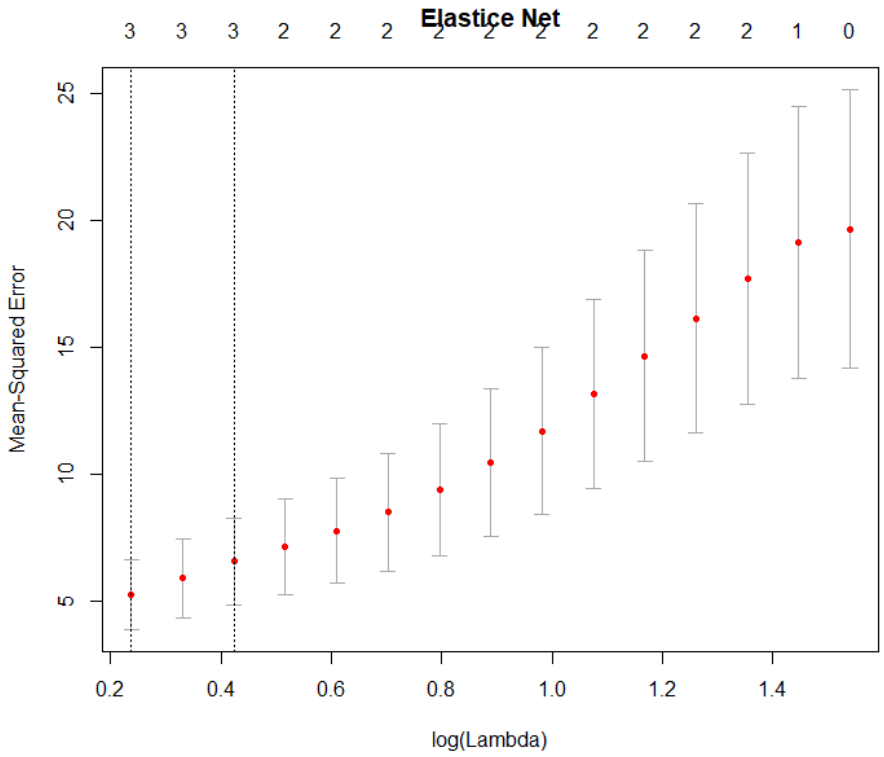
Ici, nous ferons Elastic Net avec $\alpha = 0.75$ (choix arbitraire). En comparaison par rapport au Lasso, on observe qu'il faut aller jusqu'à $\lambda = 4.3$ pour remarquer que la régression Elastic Net ne sélectionne plus aucune variable (avec Lasso $\lambda = 4$



Sur ce second graphique, nous observons la valeurs des coefficients en fonction de la valeur de $log(\lambda)$. On observe bien que plus λ augmente et plus on fait une sélection restrictive des variables.



En faisant une validation croisée de la même manière que pour le Lasso mais à coefficient α fixé, nous pouvons obtenir le λ optimal de l'estimation Elastic Net de ce problème. Le λ optimal trouvé ici est égale à 0.76. Par ailleurs en faisant la régression Elastic Net ensuite avec ce λ, nous trouvons comme coefficients $\beta = (2.25, 0.61, 0, 0, 1.10, 0, 0, 0)$ ce qui est comme le Lasso une sélection des variables 1,2 et 5. Par contre lorsque l'on regarde l'erreur à laquelle on s'intéresse à la question 13, pour Elastic Net on a $||X(\hat{\beta}_\lambda - \beta^0)|| = 16.63$, alors que pour le Lasso on avait $||X(\hat{\beta}_\lambda - \beta^0)|| = 20.37$ avec les λ trouvés par validation croisée. Finalement, l'estimation par Elastic Net correspond mieux aux variables simulées.



Les avantages de l'Elastic Net par rapport à l'estimateur Lasso sont donc qu'il va fixer moins rapidement les coefficients que l'ont souhaite estimer à 0. On a observé aussi sur ce jeu de données que l'estimation par Elastic Net fournit des estimations qui correspondent mieux à la réalité. De plus en grande dimension (p nettement supérieur à n), ce qui n'est pas notre cas dans ce jeu de données, l'Elastic Net apporte aussi plus de robustesse du fait qu'il ne limite pas le nombre de coefficients non nuls à n. Si p est nettement supérieur à n, il est clair que ne laisser que n coefficients pour expliquer le modèle semble assez restrictif.