SMDA - TP2: Modèles de regression linéaire Ridge et Lasso

Maha ELBAYAD

15 Novembre 2015

Application I: Modèle de régression linéeaire

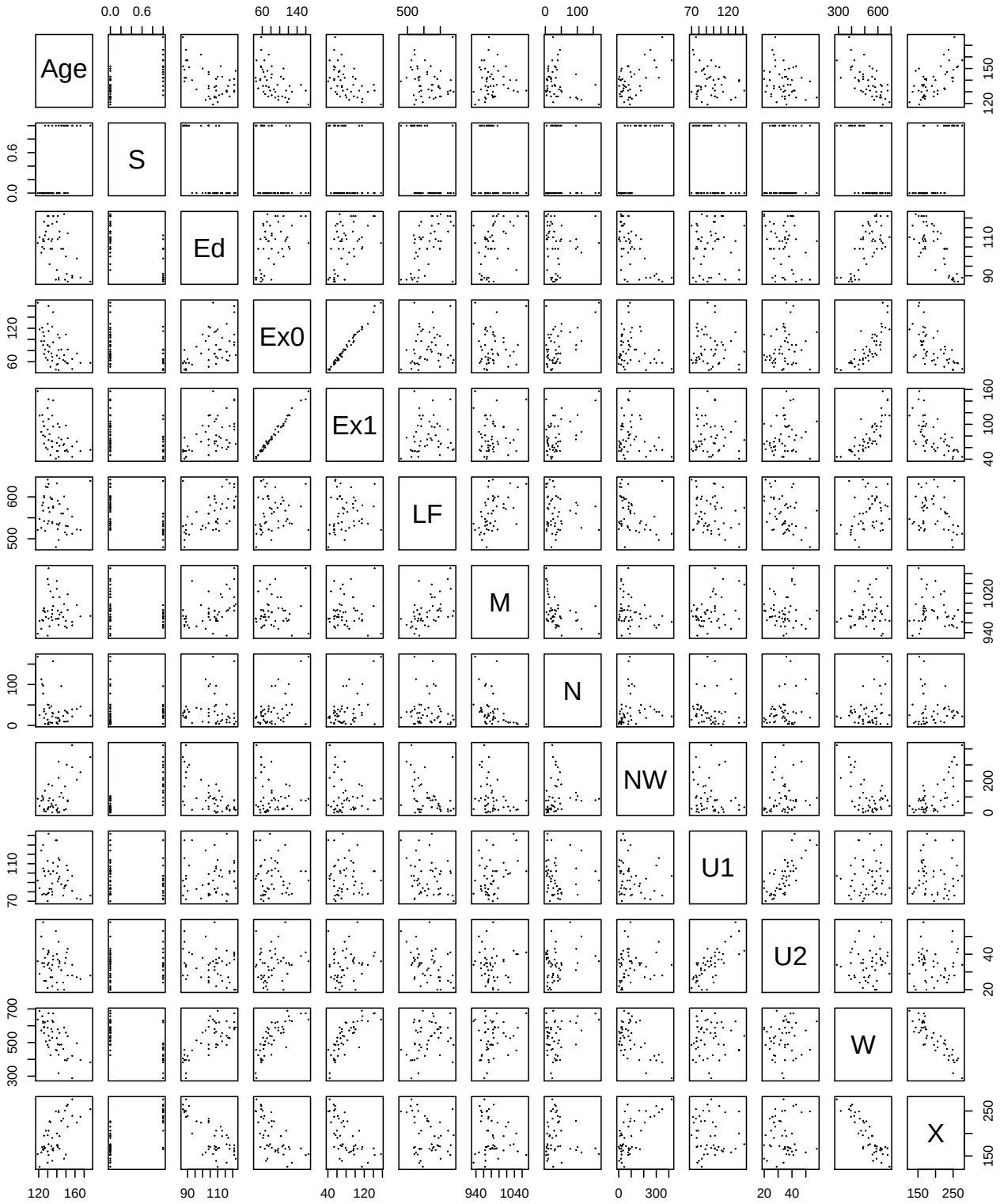
Préliminaires

```
## 'data.frame':    47 obs. of  14 variables:
##  $ R  : num  79.1 163.5 57.8 196.9 123.4 ...
##  $ Age: int  151 143 142 136 141 121 127 131 157 140 ...
##  $ S  : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed : int  91 113 89 121 121 110 111 109 90 118 ...
##  $ Ex0: int  58 103 45 149 109 118 82 115 65 71 ...
##  $ Ex1: int  56 95 44 141 101 115 79 109 62 68 ...
##  $ LF : int  510 583 533 577 591 547 519 542 553 632 ...
##  $ M  : int  950 1012 969 994 985 964 982 969 955 1029 ...
##  $ N  : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW : int  301 102 219 80 30 44 139 179 286 15 ...
##  $ U1 : int  108 96 94 102 91 84 97 79 81 100 ...
##  $ U2 : int  41 36 33 39 20 29 38 35 28 24 ...
##  $ W  : int  394 557 318 673 578 689 620 472 421 526 ...
##  $ X  : int  261 194 250 167 174 126 168 206 239 174 ...
```

Dans le fichier 'UsCrime.txt' on dispose de 47 observations de 14 variables.

La matrice de corrélation entre les différentes variables:

	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
Age	1												
S	0.58	1											
Ed	-0.53	-0.7	1										
Ex0	-0.51	-0.37	0.48	1									
Ex1	-0.51	-0.38	0.5	**0.99**	1								
LF	-0.16	-0.51	0.56	0.12	0.11	1							
M	-0.03	-0.31	0.44	0.03	0.02	0.51	1						
N	-0.28	-0.05	-0.02	0.53	0.51	-0.12	-0.41	1					
NW	0.59	**0.77**	-0.66	-0.21	-0.22	-0.34	-0.33	0.1	1				
U1	-0.22	-0.17	0.02	-0.04	-0.05	-0.23	0.35	-0.04	-0.16	1			
U2	-0.24	0.07	-0.22	0.19	0.17	-0.42	-0.02	0.27	0.08	**0.75**	1		
W	-0.67	-0.64	**0.74**	**0.79**	**0.79**	0.29	0.18	0.31	-0.59	0.04	0.09	1	
X	0.64	**0.74**	**-0.77**	-0.63	-0.65	-0.27	-0.17	-0.13	0.68	-0.06	0.02	**-0.88**	1



On remarque que certaines covariables sont fortement corrélés, notamment (‘Ex0’, ‘Ex1’) et (‘X’,‘W’)

Le modèle:

On considère le modèle linéaire:

$$Y = X\beta + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

1.Résultats *lm()*:

	Estimate	Pr(>|t|)	signif
(Intercept)	-691.838	0	***
Age	1.04	0.019	*
S	-8.308	0.581	
Ed	1.802	0.009	**
Ex0	1.608	0.138	
Ex1	-0.667	0.565	
LF	-0.041	0.791	
M	0.165	0.438	
N	-0.041	0.752	
NW	0.007	0.911	
U1	-0.602	0.178	
U2	1.792	0.044	*
W	0.137	0.203	
X	0.793	0.002	**

Table 1: Coefficients estimés β

2. Le modèle global:

le coefficient de détermination R^2 mesure de la qualité du modèle linéaire

$$R^2 = \frac{var(\hat{Y})}{var(Y)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},\ o\grave{u}\ \hat{Y} = X\hat{\beta}$$

Dans notre cas $R^2 = 0.769236$, plus R^2 est proche de 1, plus les covariables expliquent parfaitement la cible.

Pour évaluer la signifiance du modèle on considère la statistique de Fisher:

$$F = \frac{R^2(n-p-1)}{p(1-R^2)} \sim F_{p,n-p-1}(1-\alpha)$$

Le test a pour hypothèses.[1]:

$$\begin{cases} H_0 :\ \beta_1 = \beta_2 = ... = \beta_{13} = 0 \\ H_1 :\ \exists j \in \{1...13\},\ \beta_j \neq 0 \end{cases}$$

Au seuil critique α on rejète H_0 i.e on considère que la regression est significative si $F > q_\alpha$ soit une $p - value < \alpha$. Dans notre modèle $F = 8.46179$ avec une $p - value = 3.686 \times 10^{-7}$. Comme la p-value est très petite on conclut que le modèle linéaire est retenu.

3. Les coefficients du modèle:

Pour chaque covariable on considère le test de significativité de β_j avec la statistique de student.

$$T = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 S_{j,j}}},\ S_{j,j}\ \text{jème terme diagonal de}(X^T X)^{-1}$$

[1] On exclut β_0 la constante de régression

Le test a pour hypothèses:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Au seuil critique α on rejète H_0 si $T > t_{n-p}(1 - \frac{\alpha}{2})$ ou $p - value < \alpha$.

Dans les $p-values$ du tableau 4, on voit que les covaribales avec les plus petites p-values sont plus significatives ($***$: p-value <.001, $**$: p-value <.01, $*$: p-value <.05, $\bullet$: p-value <.1). Pour ce modèle particulier, les covariables {Ed,Age,U2,X} s'avèrenet plus significatives, mais comme les covariables ont de fortes corrélation on ne peut virer les covariables qui paraissent à priori sans utilité.

	2.5%	97.5%
(Intercept)	-1008.99	-374.68
Age	0.18	1.90
S	-38.65	22.03
Ed	0.48	3.12
Ex0	-0.55	3.76
Ex1	-3.00	1.67
LF	-0.35	0.27
M	-0.26	0.59
N	-0.30	0.22
NW	-0.12	0.14
U1	-1.49	0.29
U2	0.05	3.53
W	-0.08	0.35
X	0.31	1.27

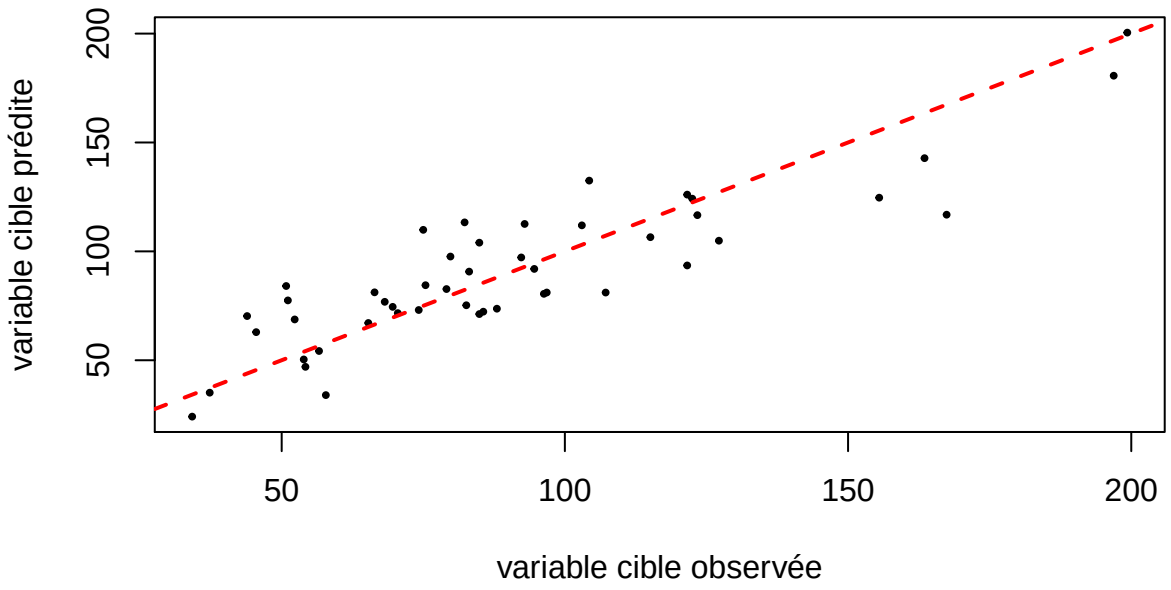
Table 2: Intervalles de confiance au risque de 5%

	0.5%	99.5%
(Intercept)	-1117.92	-265.75
Age	-0.12	2.20
S	-49.07	32.45
Ed	0.03	3.58
Ex0	-1.29	4.50
Ex1	-3.81	2.47
LF	-0.46	0.38
M	-0.41	0.74
N	-0.40	0.31
NW	-0.17	0.18
U1	-1.80	0.59
U2	-0.55	4.13
W	-0.15	0.43
X	0.15	1.44

Table 3: Intervalles de confiance au risque de 1%

La p-value de la question précédente présente la probabilité de se retrouver en dehors de l'intervalle de confiance. La constante de régression (intercept), Ed et X étant les seules avec p-value<.01, leurs intervalles de confiance au risque de 1% ne comprennet pas la valeur 0. Pour Age et U2 les intervalles de confiance(5%) ne comprennent plus 0 comme leurs p-value est <.05.

4. Etude des valeurs prédites:



On remarque que les points sont tous proches de la première bissectrice (y=x) le modèle est donc valide.

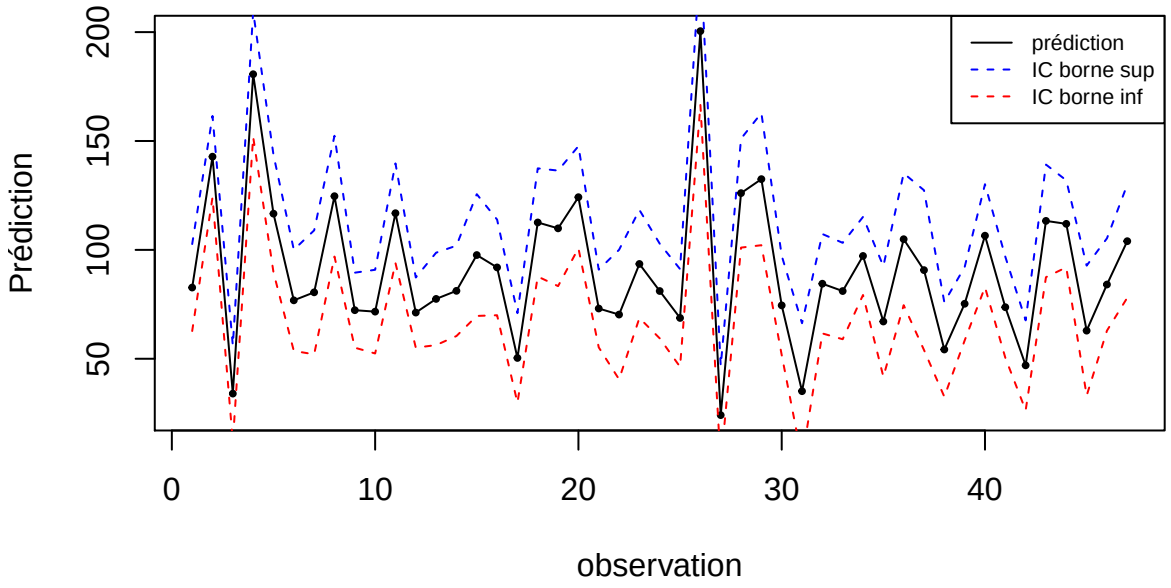
	1	2	3	4	5	6	7	8	9	10	11
fit	82.70	142.80	33.99	180.67	116.63	76.83	80.50	124.65	72.30	71.63	116.84
lwr	62.73	124.09	11.90	151.89	89.66	53.60	52.03	96.92	55.08	52.42	93.98
upr	102.67	161.50	56.07	209.44	143.61	100.06	108.97	152.39	89.53	90.85	139.70

	12	13	14	15	16	17	18	19	20	21	22
fit	71.20	77.48	81.18	97.62	91.94	50.37	112.59	109.88	124.19	73.07	70.29
lwr	55.10	56.26	60.43	69.66	69.98	29.81	87.75	83.36	100.75	55.22	40.55
upr	87.31	98.70	101.94	125.58	113.91	70.93	137.44	136.41	147.63	90.91	100.03

	23	24	25	26	27	28	29	30	31	32	33
fit	93.54	81.10	68.74	200.45	24.11	126.06	132.49	74.52	35.11	84.46	81.10
lwr	68.72	59.37	46.29	166.63	1.59	101.05	102.15	51.72	3.91	61.64	58.96
upr	118.35	102.84	91.19	234.28	46.63	151.08	162.83	97.31	66.32	107.29	103.23

	34	35	36	37	38	39	40	41	42	43	44
fit	97.21	67.11	104.89	90.70	54.24	75.26	106.52	73.67	46.97	113.31	111.96
lwr	79.27	41.74	74.60	54.03	32.55	58.40	82.88	50.48	26.31	87.46	91.98
upr	115.14	92.48	135.19	127.38	75.94	92.11	130.16	96.86	67.63	139.17	131.94

	45	46	47
fit	62.90	84.10	103.99
lwr	33.04	62.89	78.02
upr	92.76	105.30	129.96

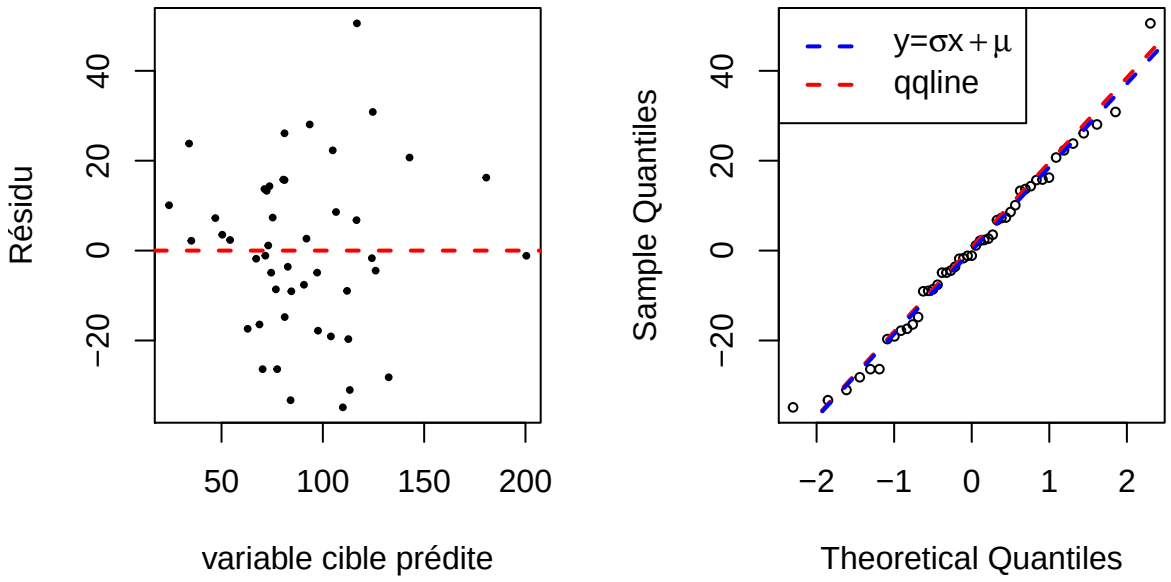


5. Etude des résidus:

L'erreur quadratique des résidus $RSE = \|\hat{\epsilon}\|_2^2 = 1.5878701 \times 10^4$.

On pose $\hat{\sigma}^2 = \dfrac{\|\hat{\epsilon}\|_2^2}{n-p}$. Avec le théorème de Cochran on peut écrire: $\|\hat{\epsilon}\|_2^2 \sim \sigma^2 \chi^2(n-p)$. Ainsi notre estimateur de σ^2 est sans biais ($E(\hat{\sigma}^2) = \sigma^2$).

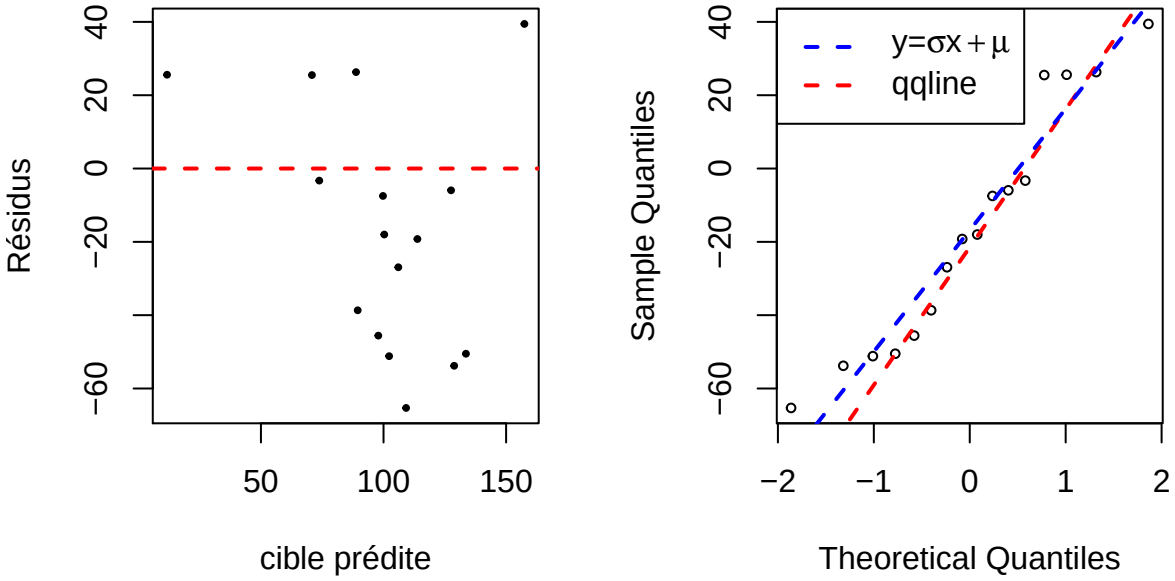
$$\hat{\sigma}^2 = 467.02061$$



On effectue le test de Shapiro-Wilk de normalité, H_0 : $\hat{\epsilon}$ suit une loi normale, donc si p-value < 0.01, l'échantillon ne suit pas une loi normale. Pour ce modèle p-value=0.8216155. Graphiquement, on remarque que les résidus sont distribués de façon aléatoire ($\sim \mathcal{N}(0, \sigma^2)$), ils ne semblent plus contenir d'information. Sur le graphe Q.Q aucun point aberrant n'a été soulevé et tous les résidus sont alignés sur la droite $y = \sigma x + \mu \sim qqline = D(Q1, Q3)$. On constate aussi que les résidus sont plutôt hétéroscédastiques.

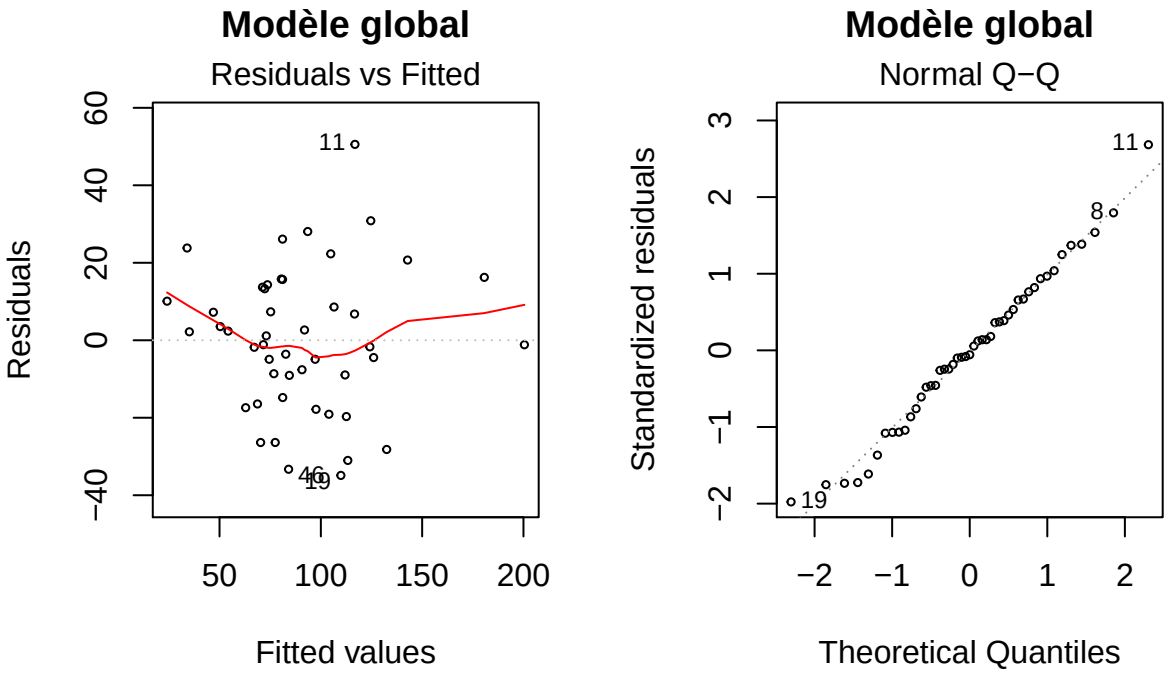
6. Performances du moèle sur de nouvelles données:

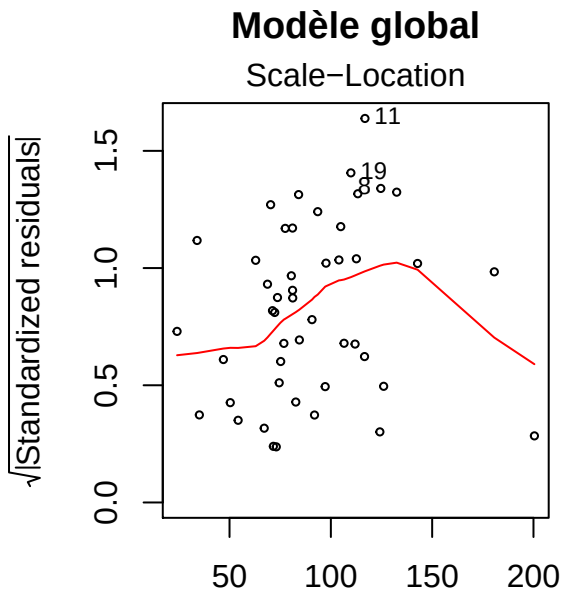
- La moyenne des erreurs quadratiques= 2.097739×10^4.
- La moyenee des erreurs = -16.818
- L'écart type des erreurs quadratiques= 33.118.

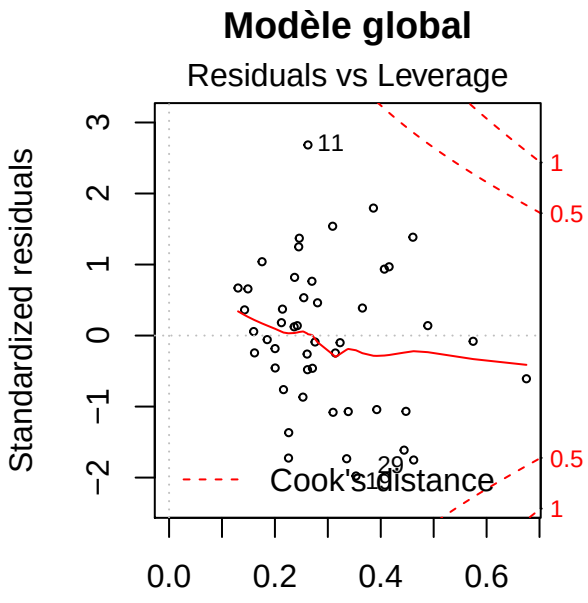


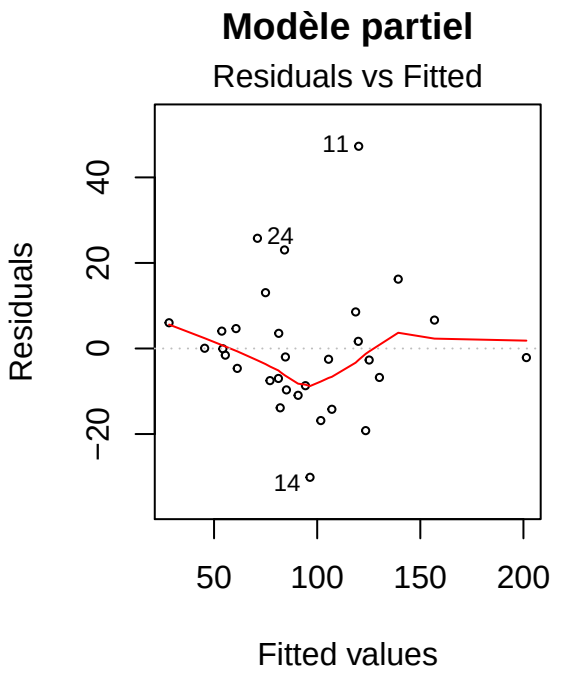
Sur les données du Test set, les résidus ont une moyenne $\neq 0$ et sur le qq-plot on voit que la droite du 1er quartile- 3ème quartile est écratée de $y = \mu + \sigma x$.

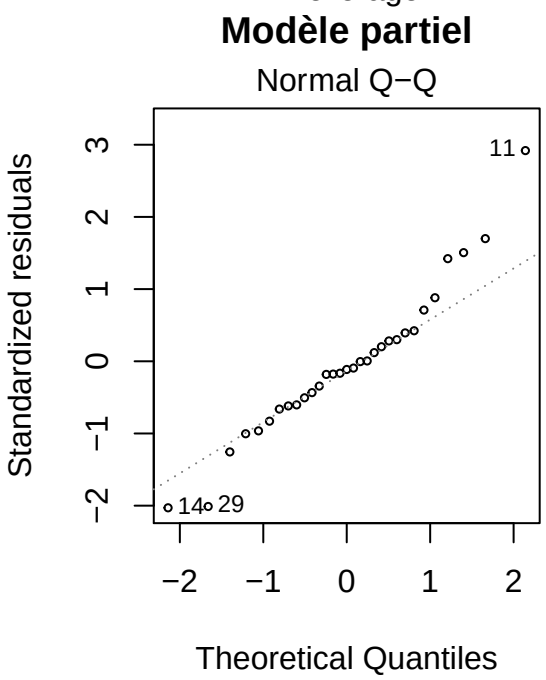
Les graphiques *lm*:

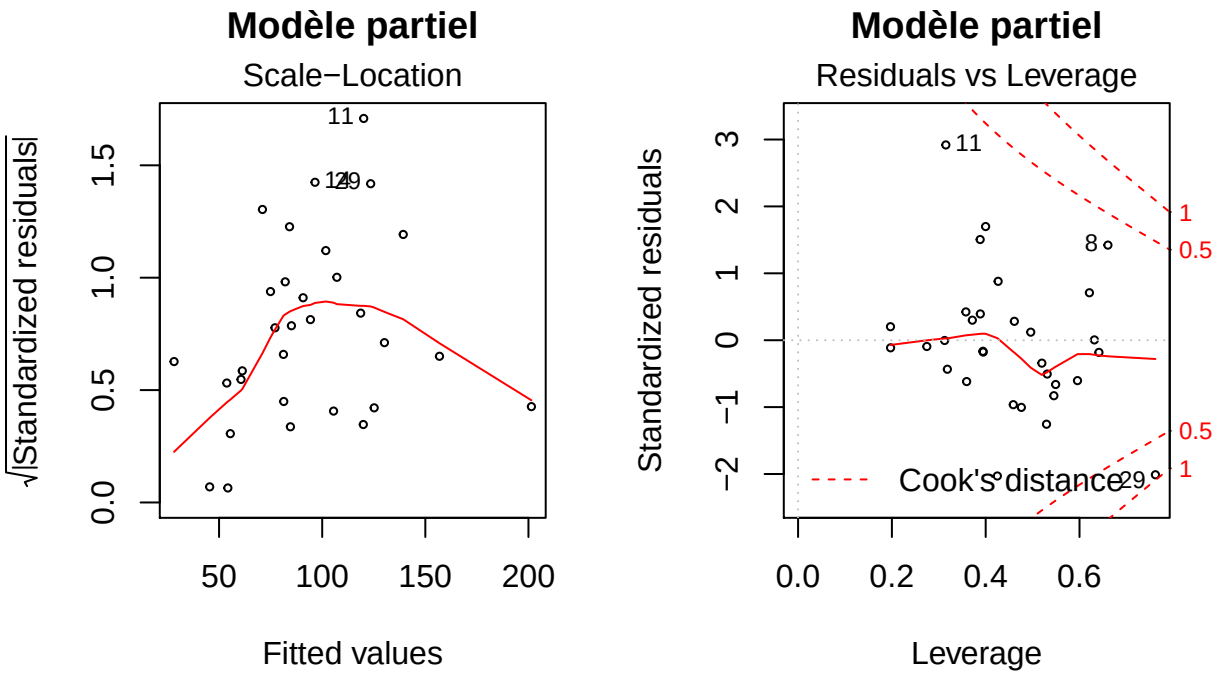



Modèle global
Scale−Location
√|Standardized residuals|
Fitted values
11
19
0.0
0.5
1.0
1.5
50
100
150
200


Modèle global
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
11
29
1
0.5
0.5
1
−2
−1
0
1
2
3
0.0
0.2
0.4
0.6


Modèle partiel
Residuals vs Fitted
Residuals
Fitted values
11
24
14
−20
0
20
40
50
100
150
200


Modèle partiel
Normal Q−Q
Standardized residuals
Theoretical Quantiles
11
14
29
−2
−1
0
1
2
3



Sur le graphe 'Residuals vs fitted' on vérifie l'hypothèse d'homoscédasticité: variance constante des résidus pour chaque prédiction $\hat{y}$. Le graphe scale-location nous permet de repérer les outliers: points qui ont des résidus studentisés >3. Le QQ-plot teste l'hypothèse de normalité des résidus. Le dernier graphe Residuals vs Leverage montre l'impact de chaque observation sur le modèle et permet de repérer les points leviers avec une large distance de Cook $\propto$ différence entre les valeurs prédites avec et sans l'observation en question.

Sélection de modèles:

1. Régression *Backward*:

Dans la sélection *Backward*, on retire du modèle le régresseur non significatif qui porte le score AIC le plus faible. Dans le modèe linéaire des *UScrime* on retire dans l'ordre: NW, LF, N,S,Ex1, M,U1. le modèle final est donc:

$$R \sim Age + Ed + Ex0 + U2 + W + X$$

	Estimate	Pr(>|t|)	signif
(Intercept)	-618.503	0	***
Age	1.125	0.003	**
Ed	1.818	0.001	**
Ex0	1.051	0	***
U2	0.828	0.06	•
W	0.16	0.097	•
X	0.824	0	***

Table 4: Coefficients estimés β - sélection backward

Comparaison avec le modèle initial:

On constate que le modèle *backward* est plus performant.[2] et avec des régresseurs tous significatifs.

[2] R^2_{adj} prenant en considération le nombre de covariables est donc la plus appropriée pour comparer les deux modèles

stat	Modèle complet	Modèle backward
R^2	0.769	0.748
R^2_{adj}	0.678	0.71
F-stat p-value	3.686e-07	1.441e-10

2. Régression *Forward*:

Similaire à la sélection *Backward*, cette fois on ajoute au modèle le régresseur avec le plus petit AIC. Dans le modèe linéaire des *UScrime* on ajoute dans l'ordre: Ex0,X,Ed,Age,U2,W. le modèle final est donc:

$$R \sim Ex0 + X + Ed + Age + U2 + W$$

On se retrouve alors avec le même modèle qu'en sélection *Backward*. L'approche *Backward* ou *Forward* ne prend pas en considération la corrélation entre les régresseurs, l'ajout ou la supression dans une étape altèrent la signifiance des regresseurs sélectionnés.

3. Régression *Stepwise*:

Dans l'apprche *Stepwise* on part du modèle gloabl (ou du modèle null) et à chaque étape k on ajoute ou on retire le regresseur avec le moins d'impact sur le AIC global du modèle. On s'arrête lorsque le modèle se stabilise.

On effectue les ajouts/supression suivantes: - NW - LF -N -S -EX1 - M -U1 Pour finir avec le même modèle que dans les sélections précédentes:

$$R \sim Age + Ed + Ex0 + U2 + W + X$$

5. Sélection avec BIC

Pour substituer AIC par BIC , on choisit $k = log(n)$. Les régresseurs sélectionnés en *Stepwise* sont dans l'ordre: +Ex0 + X + Ed + Age + U2 Le modèle final :

$$R \sim Ex0 + X + Ed + Age + U2$$

On retrouve ce même modèle en sélection *Backward* ou *Forward*.

stat	Modèle complet	Modèle AIC	modèle BIC
R^2	0.769	0.748	0.73
R^2_{adj}	0.678	0.71	0.697
F-stat p-value	3.686e-07	1.441e-10	1.105e-10

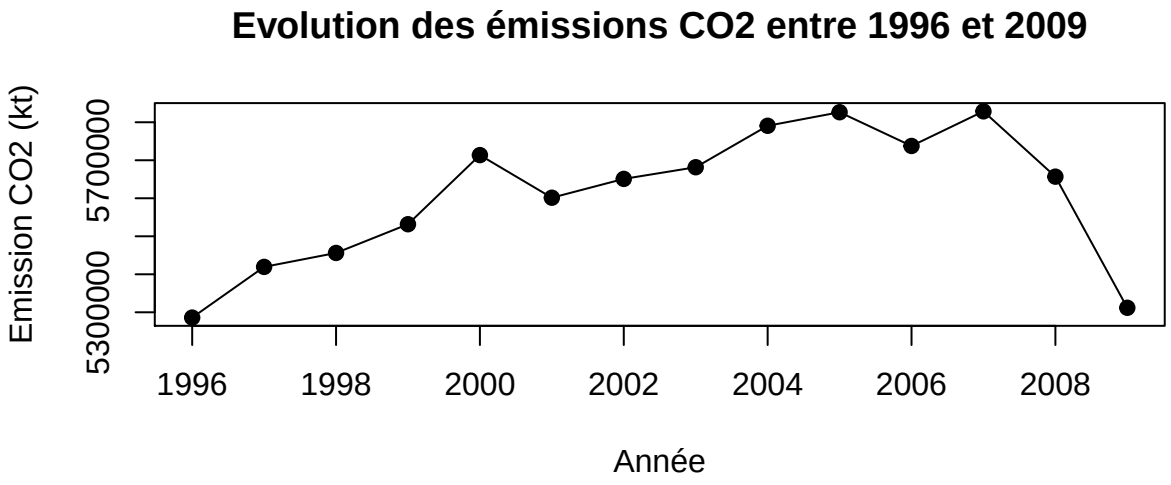
Application II: Régression Ridge et Lasso:

Préliminaires

1,2,3.

Dans le fichier *usa_indicators.txt*, on dispose de 14 observations de 110 variables. Comme $n \ll p$, un modèle de régression linéaire simple n'est pas pertinent ($X^T X$ singulière).

La variable cible est: EN.ATM.CO2E.KT (CO2 emissions (kt)) dont une partie est la variable EN.CO2.BLDG.MT (CO2 emissions from residential buildings and commercial and public services (million metric tons))



4.

Avoir des données de nature très diverse en entrée du modèle, ralentit la convergence vers une solution si on utilise la méthode du gradient. lm() utilisant une décomposition QR pour trouver les paramètres du modèle, la normalisation des régresseurs premet tout juste d'avoir des poids β interprétables.

Regression Ridge:

1.

La régression Ridge consiste à minimiser la fonction cost:

$$E(\beta) = \|Y - X\beta\|^2, \text{ sous la contrainte } \|\beta\|_2^2 \leq c$$

2.

On effectue la ridge régression avec $\lambda = 0$ puis $\lambda = 100$, on liste dessous les 5 régresseurs les plus influents pour les deux modèles. On note que model$coef ne sont pas dans la bonne échelle, on utilise coeff(model) pour extraire les coefficients du modèle.

$\lambda = 0$	-	$\lambda = 100$	-
AG.LND.TOTL.K2	2.93	AG.LND.TOTL.K2	1.61
Intercept	-2.26	Intercept	-0.463
EG.USE.COMM.FO.ZS	0.51	SP.RUR.TOTL	-0.135
AG.LND.AGRI.K2	0.284	EG.USE.COMM.FO.ZS	0.12
SP.RUR.TOTL	-0.246	AG.SRF.TOTL.K2	-0.0956

Où:

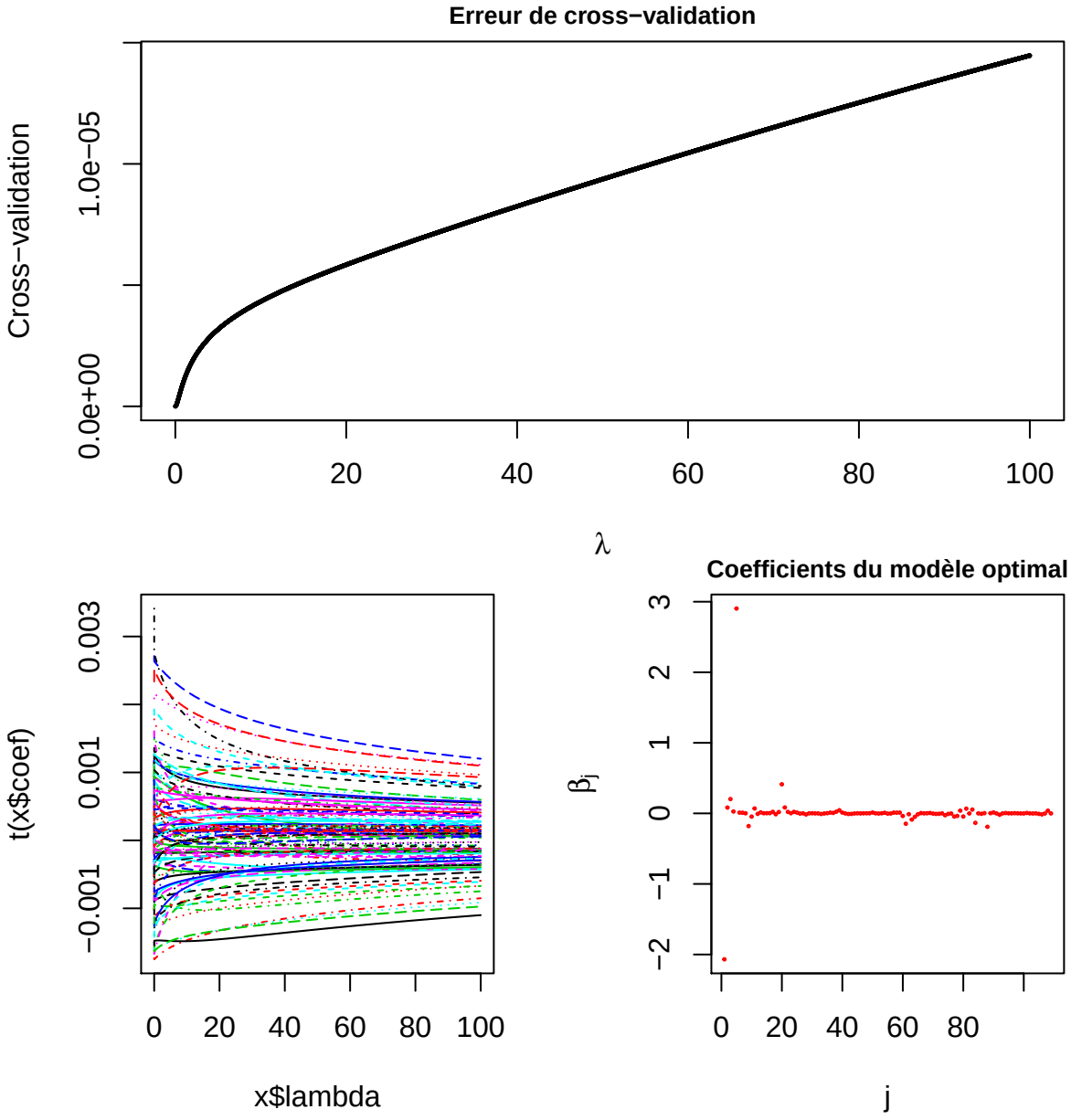
- AG.LND.TOTL.K2: Land area (sq. km)
- EG.USE.COMM.FO.ZS: Fossil fuel energy consumption (% of total)
- SP.RUR.TOTL: Rural population
- AG.LND.AGRI.K2: Agricultural land (sq. km)

- AG.SRF.TOTL.K2: Surface area (sq. km)

Ces régresseurs sont vraisemblables, si on prend en considération que, par exemple, le coefficient pour la poulation rurale (<0) a absorbé ceux de la population totale et de la population urbane du fait de leur forte corrélation.

3.

On effectue une régression ridge pour λ allant de 0 à 100 par pas de 0.01 et on trace la croube de l'erreur Cv pour chaque λ ainsi que les différents coefficients du modèle.



On voit bien que les coefficients tendent à s'annuler quand on augmente la valeur de λ. Cependant, le modèle optimal, ayant la plus petite erreur CV correspond à $\lambda = 0.01$ (ou plûtot la plus petite valeur non nulle de λ qu'on donne au modèle). Les coefficients pour $\lambda = 0.01$ sont affichés dans le graphe ci-dessus.

4. Erreur quadratique du modèle optimal ($\lambda = 0.01$):

On obtient une erreur quadratique moyenne de 3.591×10^{-11}.

Regression Lasso:

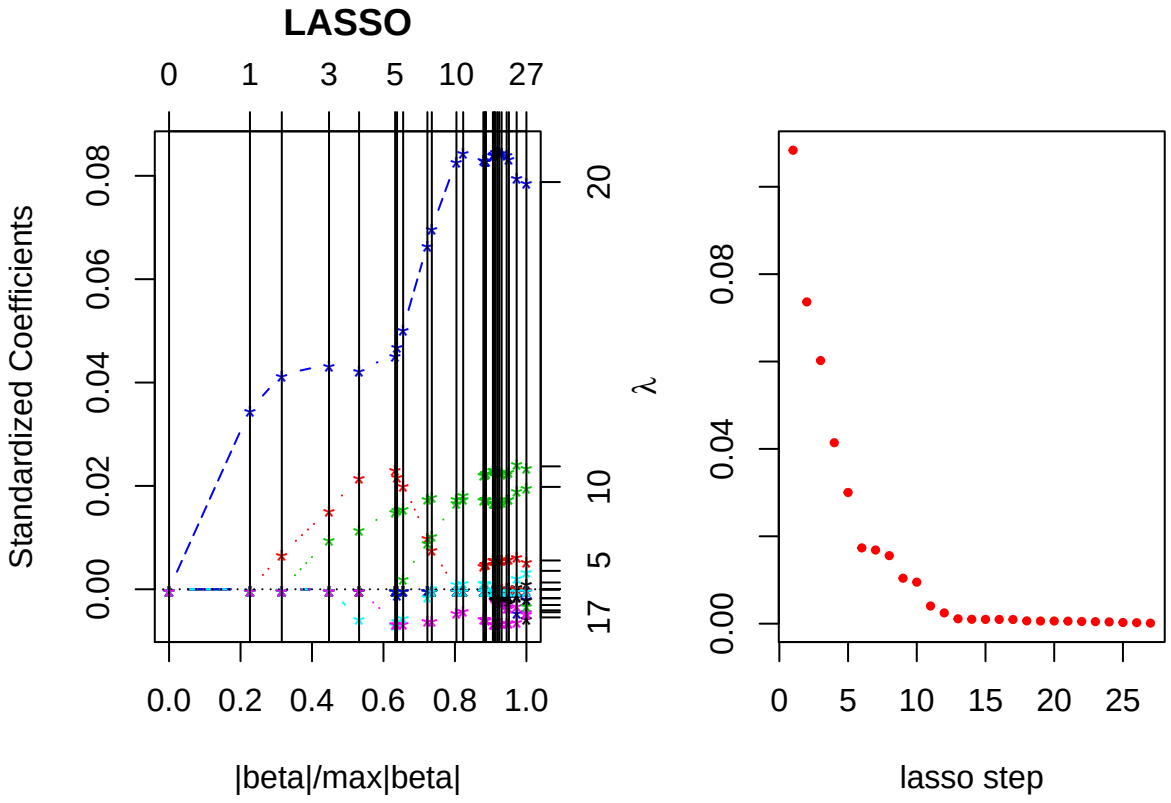
1.

La régression Lasso consiste à minimiser la fonction cost:

$$E(\beta) = \|Y - X\beta\|^2, \text{ sous la contrainte } \|\beta\|_1 \leq c$$

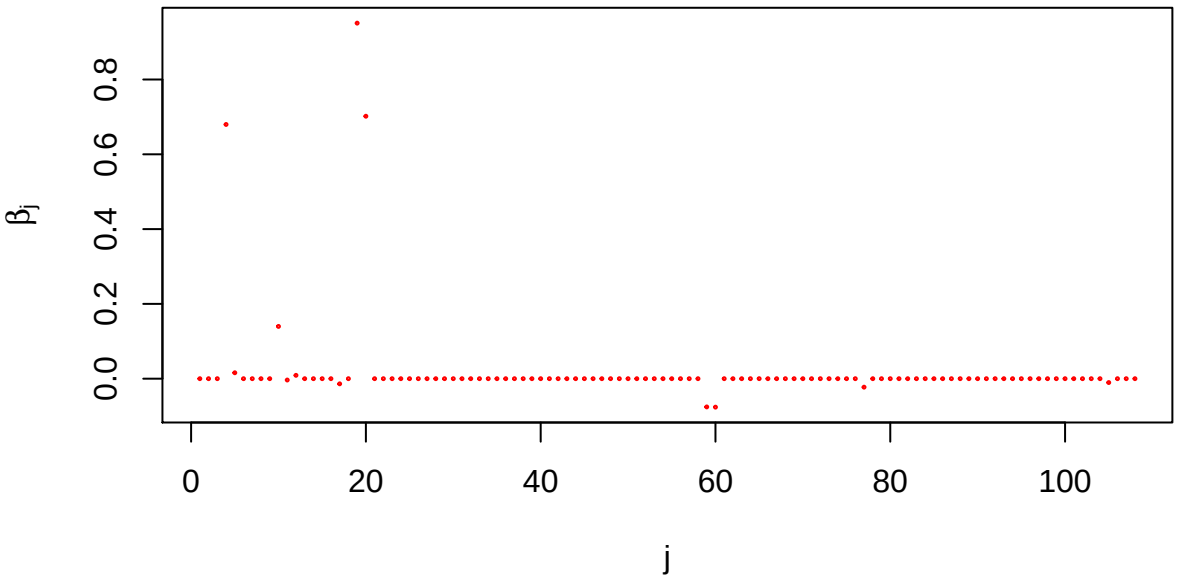
2.

Le graphe ci-dessous compare les chemins de régularisation en traçant l'évolution des coefficients pour différentes valeurs de λ. Les λ de chaque étape lasso sont visualisés dans la figure de droite.



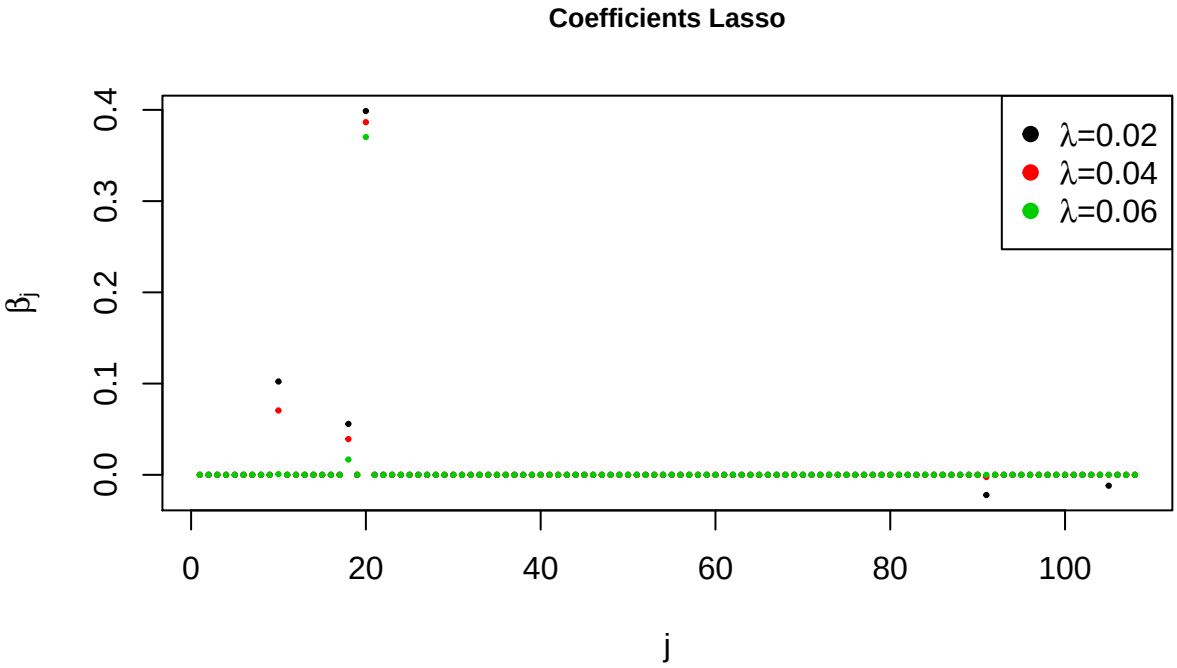
3.

Pour $\lambda = 0$ on estime les coefficients du modèle LASSO. Il s'agit d'un modèle sans régularisation sur l'ensemble des régresseurs choisis par LASSO.



4.

On estime de nouveau les coefficients pour $\lambda \in \{.02, .04, .06\}$



En augmentant λ on force les coefficients de plus en plus à s'annuler pour minimiser $\|\beta\|_1$. avec $\lambda.02$ on se trouve déjà avec seulement 5 régresseurs qui semblent plus vraisemblables que les variables Ridge en substituant par exemple des variables de surface par des variables économiques de l'industrie agroalimentaire.

- EG.ELC.COAL.KH: Electricity production from coal sources (kWh)
- EG.IMP.CONS.ZS: Energy imports, net (% of energy use)
- EG.USE.COMM.KT.OE: Energy use (kt of oil equivalent)

- TM.VAL.FOOD.ZS.UN: Food imports (% of merchandise imports)
- TX.VAL.FOOD.ZS.UN: Food exports (% of merchandise exports)

5. Erreur quadratique moyenne:

On obtient une erreur quadratique moyenne de 3.082×10^{-4}, ce qui est plus important que l'erreur de la régression Ridge, cependant le modèle LASSO est plus sparse comparé aux coefficients de Ridge qui sont certes très petits mais $\neq 0$. Il faut aussi rappeler qu'on a $p \gg n$ est donc considérer toutes les variables n'est pas consistant.