Régression en grande dimension

Mise en pratique

Angelina Roche

M1 MA

8 mars 2021

Présentation des données (I)

Les données sont mesurées sur $n = 442$ patients atteints de diabète.

Variable d'intérêt Y_i $(i = 1, ..., n)$: mesure quantitative évaluant la progression de la maladie.

Variables explicatives (covariables) x_i^j, $i = 1, ..., n$, $j = 1, ..., d$ $(d = 64)$:

- âge ;
- sexe ;
- indice de masse corporelle ;
- pression sanguine moyenne ;
- + diverses mesures sérologiques (6 variables) ;
- + 54 variables d'interaction.

Présentation des données (II)

i	y	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	...
1	151.00	0.04	0.05	0.06	0.02	-0.04	-0.03	-0.04	-0.00	0.02	-0.02	...
2	75.00	-0.00	-0.04	-0.05	-0.03	-0.01	-0.02	0.07	-0.04	-0.07	-0.09	...
3	141.00	0.09	0.05	0.04	-0.01	-0.05	-0.03	-0.03	-0.00	0.00	-0.03	...
4	206.00	-0.09	-0.04	-0.01	-0.04	0.01	0.02	-0.04	0.03	0.02	-0.01	...
5	135.00	0.01	-0.04	-0.04	0.02	0.00	0.02	0.01	-0.00	-0.03	-0.05	...
6	97.00	-0.09	-0.04	-0.04	-0.02	-0.07	-0.08	0.04	-0.08	-0.04	-0.10	...
7	138.00	-0.05	0.05	-0.05	-0.02	-0.04	-0.02	0.00	-0.04	-0.06	-0.04	...
8	63.00	0.06	0.05	-0.00	0.07	0.09	0.11	0.02	0.02	-0.04	0.00	...
9	110.00	0.04	0.05	0.06	-0.04	-0.01	0.01	-0.03	-0.00	-0.01	0.01	...
10	310.00	-0.07	-0.04	0.04	-0.03	-0.01	-0.03	-0.02	-0.00	0.07	-0.01	...
...												

Les données sont centrées et normalisées (i.e. standardisées) de façon à ce que

$$\sum_{i=1}^{n} x_i^j = 0 \text{ et } \sum_{i=1}^{n} \left(x_i^j\right)^2 = 1 \text{ pour tout } j = 1, ..., d.$$

Source : Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2003). Least Angle Regression, *The Annals of Statistics*.

Présentation des données (II)

i	y	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	...
1	151.00	0.04	0.05	0.06	0.02	-0.04	-0.03	-0.04	-0.00	0.02	-0.02	...
2	75.00	-0.00	-0.04	-0.05	-0.03	-0.01	-0.02	0.07	-0.04	-0.07	-0.09	...
3	141.00	0.09	0.05	0.04	-0.01	-0.05	-0.03	-0.03	-0.00	0.00	-0.03	...
4	206.00	-0.09	-0.04	-0.01	-0.04	0.01	0.02	-0.04	0.03	0.02	-0.01	...
5	135.00	0.01	-0.04	-0.04	0.02	0.00	0.02	0.01	-0.00	-0.03	-0.05	...
6	97.00	-0.09	-0.04	-0.04	-0.02	-0.07	-0.08	0.04	-0.08	-0.04	-0.10	...
7	138.00	-0.05	0.05	-0.05	-0.02	-0.04	-0.02	0.00	-0.04	-0.06	-0.04	...
8	63.00	0.06	0.05	-0.00	0.07	0.09	0.11	0.02	0.02	-0.04	0.00	...
9	110.00	0.04	0.05	0.06	-0.04	-0.01	0.01	-0.03	-0.00	-0.01	0.01	...
10	310.00	-0.07	-0.04	0.04	-0.03	-0.01	-0.03	-0.02	-0.00	0.07	-0.01	...
...												

variables explicatives variables d'intérêt

Les données sont centrées et normalisées (i.e. standardisées) de façon à ce que

$$\sum_{i=1}^{n} x_i^j = 0 \text{ et } \sum_{i=1}^{n} \left(x_i^j\right)^2 = 1 \text{ pour tout } j = 1, \ldots, d.$$

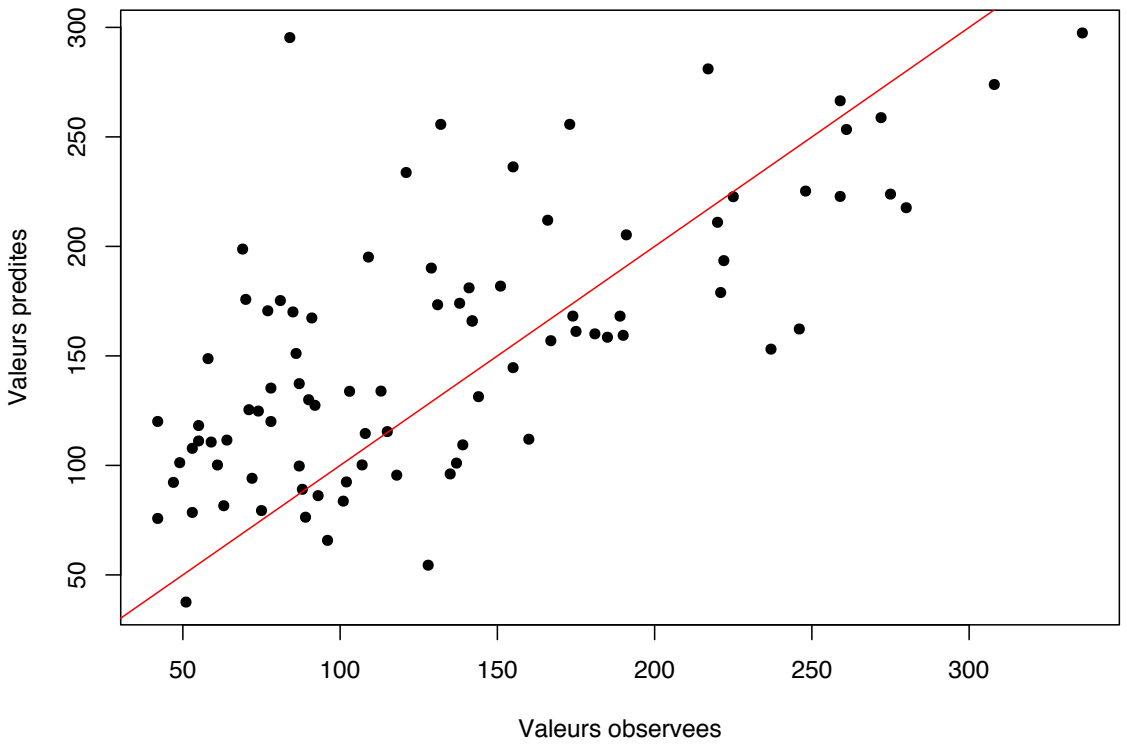
Extraction de l'échantillon de test

- Pour pouvoir comparer les différentes méthodes (moindres carrés, Ridge, Lasso), nous allons extraire aléatoirement un échantillon de test (environ 10% des données soit 44 individus) qui ne servira pas au calcul des estimateurs.

Estimateur des moindres carrés

j	Variable j	$\left[\widehat{\beta}^{(MCO)}\right]_j$
1	(Intercept)	156.23
2	age	81.66
3	sex	-288.70
4	bmi	473.06
5	map	312.14
6	tc	-36569.39
7	ldl	32189.39
...		
12	age^2	63.46
13	bmi^2	74.65
...		
21	age×sex	118.17
...		

Valeurs prédites par $\widehat{\beta}^{(MCO)}$ sur l'échantillon de test



Erreur moyenne :

$$\frac{1}{n_{test}} \sum_{i \in I_{test}} \left(Y_i - \hat{Y}_i^{(MCO)} \right)^2 \approx 3115.$$

Étude de l'estimateur Ridge

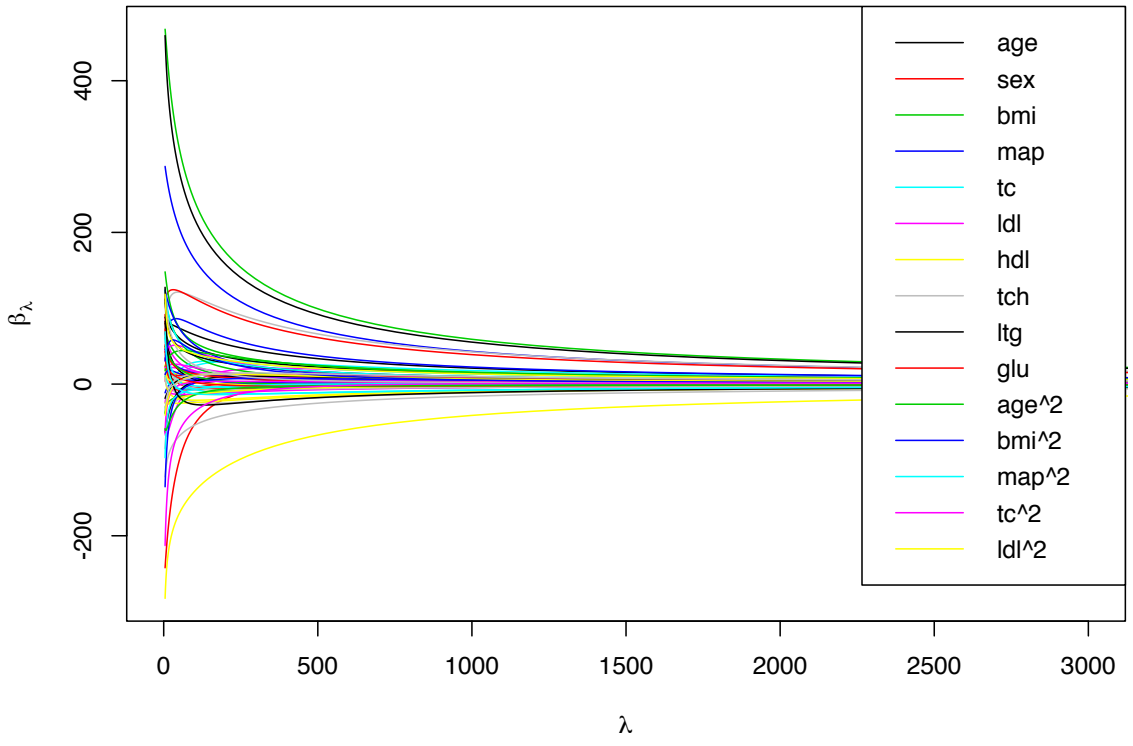


Figure – Évolution des coefficients $\left[\widehat{\beta}_\lambda^{(R)}\right]_j$ en fonction de λ

Sélection du paramètre λ

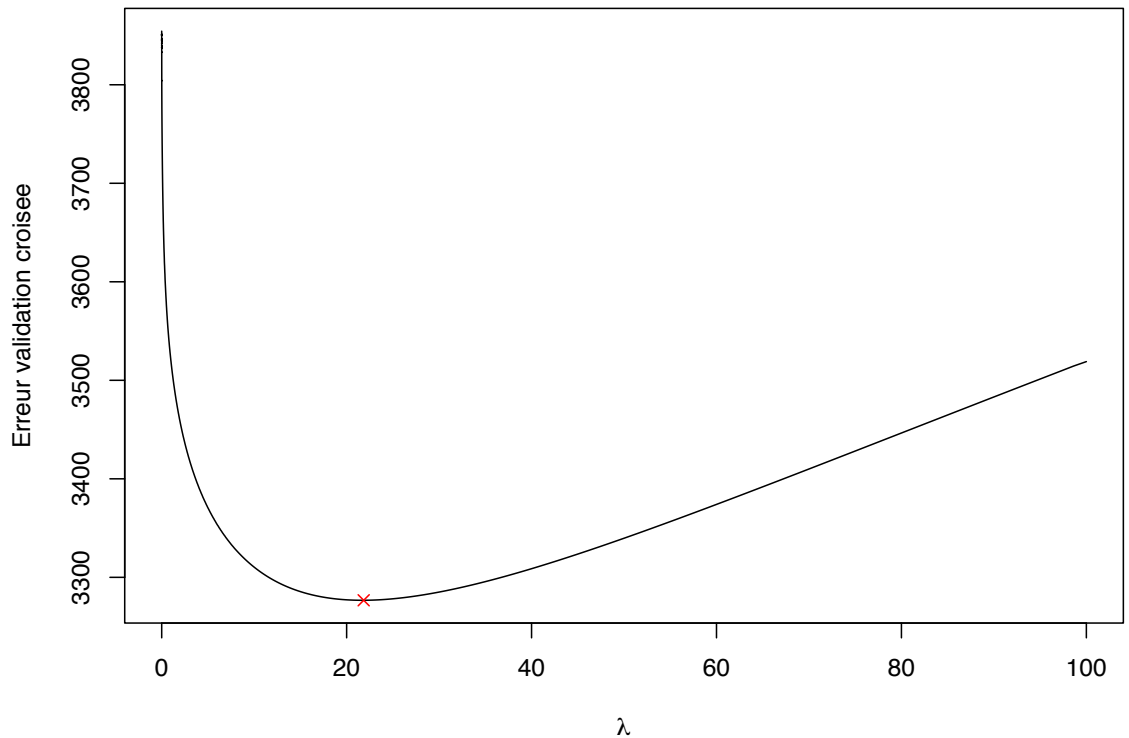


Figure – Évolution de l'erreur de validation croisée $\frac{1}{n_{app}} \sum_{i \in I_{app}} \left(Y_i - \hat{Y}_i^{(R,-i)} \right)^2$ en fonction de λ.

$\widehat{\lambda}^{(R)} \approx 21{,}8.$

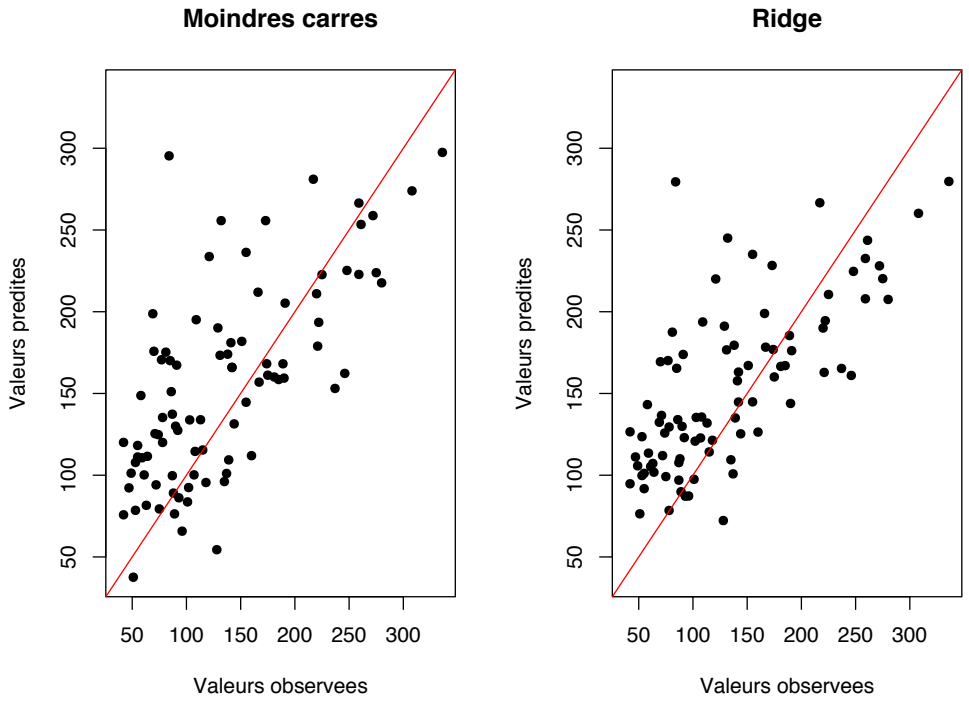
Estimateur $\widehat{\beta}^{(R)}_{\widehat{\lambda}}$

j	Variable j	$\left[\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}\right]_j$
2	age	79.01
3	sex	-163.69
4	bmi	387.58
5	map	245.95
6	tc	-8.90
7	ldl	-52.76
...		
12	age^2	50.80
13	bmi^2	83.01
...		
21	age×sex	98.10
...		

Comparaison avec $\widehat{\beta}^{(MC)}$

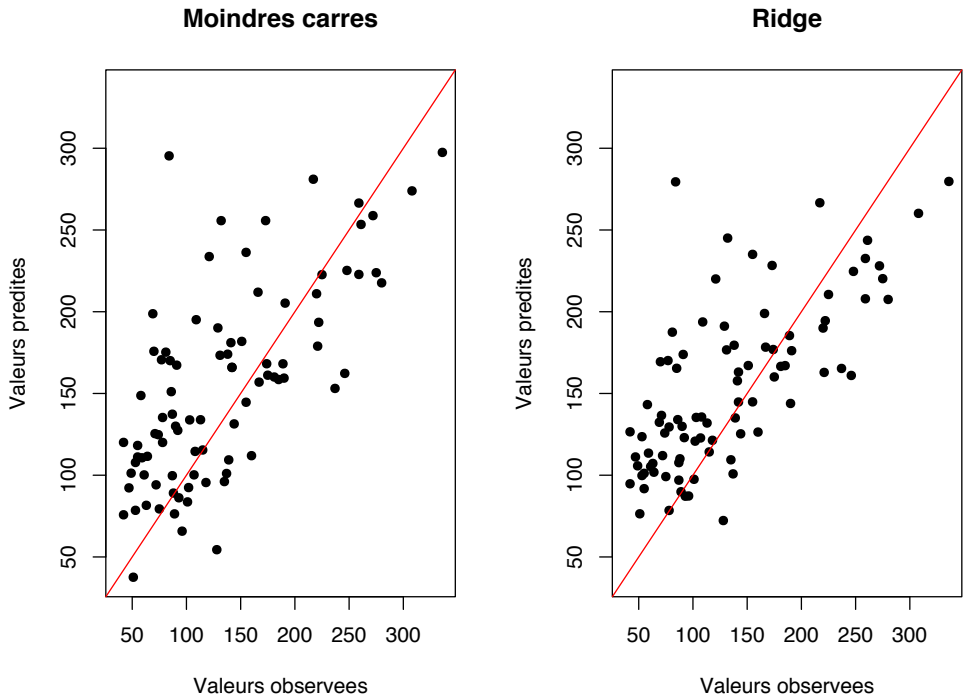
Nom variable	$\left[\widehat{\beta}^{(MCO)}\right]_j$	$\left[\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}\right]_j$
age	81.66	79.01
sex	-288.70	-163.69
bmi	473.06	387.58
map	312.14	245.95
tc	-36569.39	-8.90
ldl	32189.39	-52.76
...	...	...
age^2	63.46	50.80
bmi^2	74.65	83.01
...	...	...
age×sex	118.17	98.10
...	...	...

Valeurs prédites par $\widehat{\beta}^{(MC0)}$ et $\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}$ sur l'échantillon de test



Estimateur	$\widehat{\beta}^{(MCO)}$	$\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}$
Erreur sur I_{test}	3115	2829

Valeurs prédites par $\widehat{\beta}^{(MC0)}$ et $\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}$ sur l'échantillon de test



Estimateur	$\widehat{\beta}^{(MCO)}$	$\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}$
Erreur sur I_{test}	3115	2829

⚠ Ne pas confondre erreur sur I_{test} : $\frac{1}{n_{test}} \sum_{i \in I_{test}} (Y_i - x_i^t \widehat{\beta})^2$

... et erreur de validation croisée : $\frac{1}{n_{app}} \sum_{i \in I_{app}} \left(Y_i - x_i^t \widehat{\beta}^{(-i)}\right)^2$!

Étude de l'estimateur Lasso

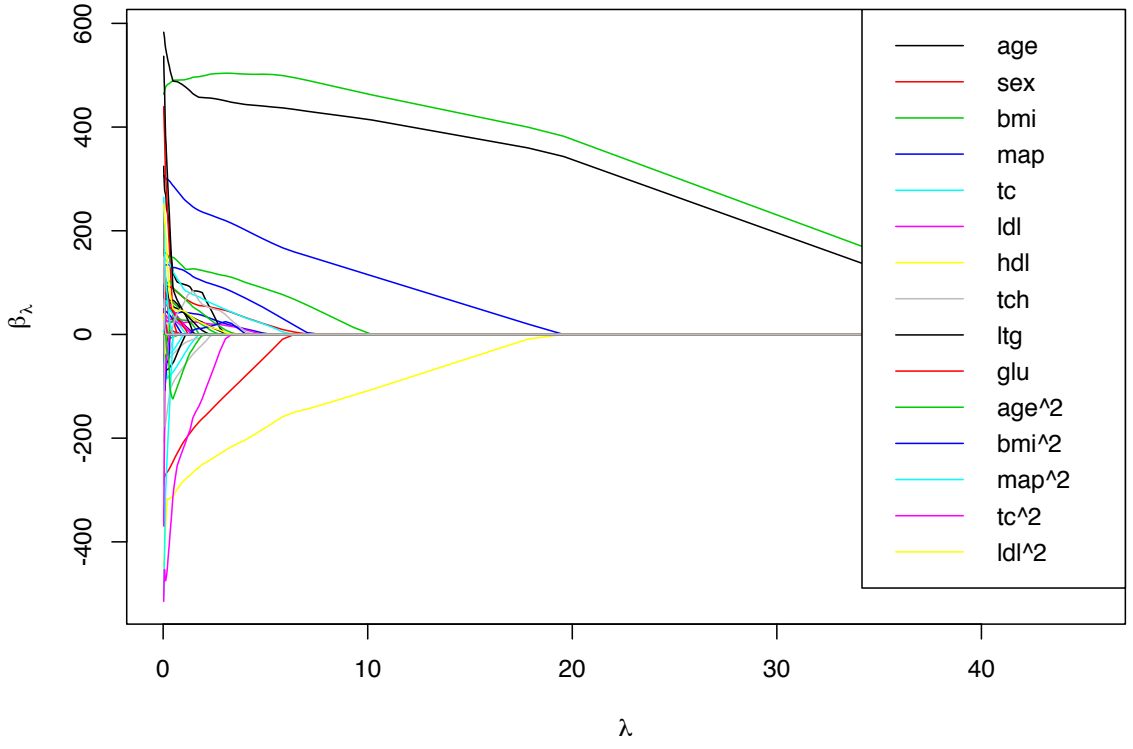


Figure – Évolution des coefficients $\left[\widehat{\beta}_\lambda^{(L)}\right]_j$ en fonction de λ

Sélection du paramètre λ (I)

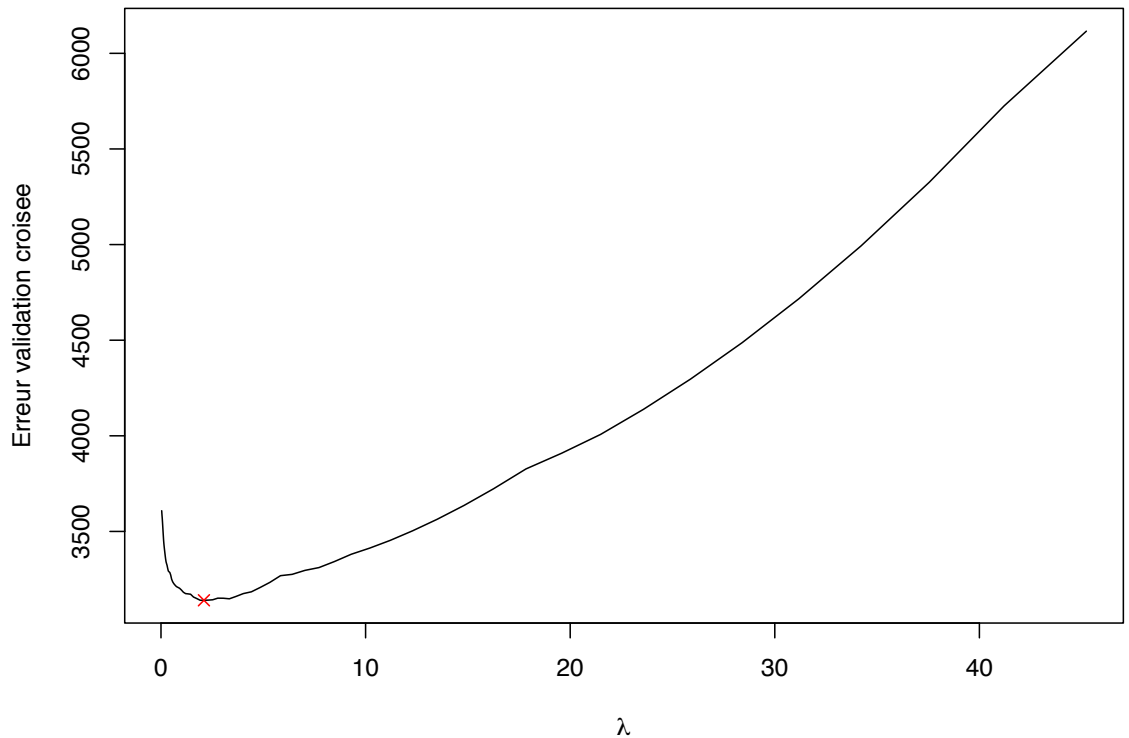


Figure – Évolution de l'erreur de validation croisée $\frac{1}{n_{app}} \sum_{i \in I_{app}} \left(Y_i - \widehat{Y}_i^{(L,-i)} \right)^2$ en fonction de λ.

$\widehat{\lambda}^{(L,CV)} \approx 2,10.$

Sélection du paramètre λ (II)

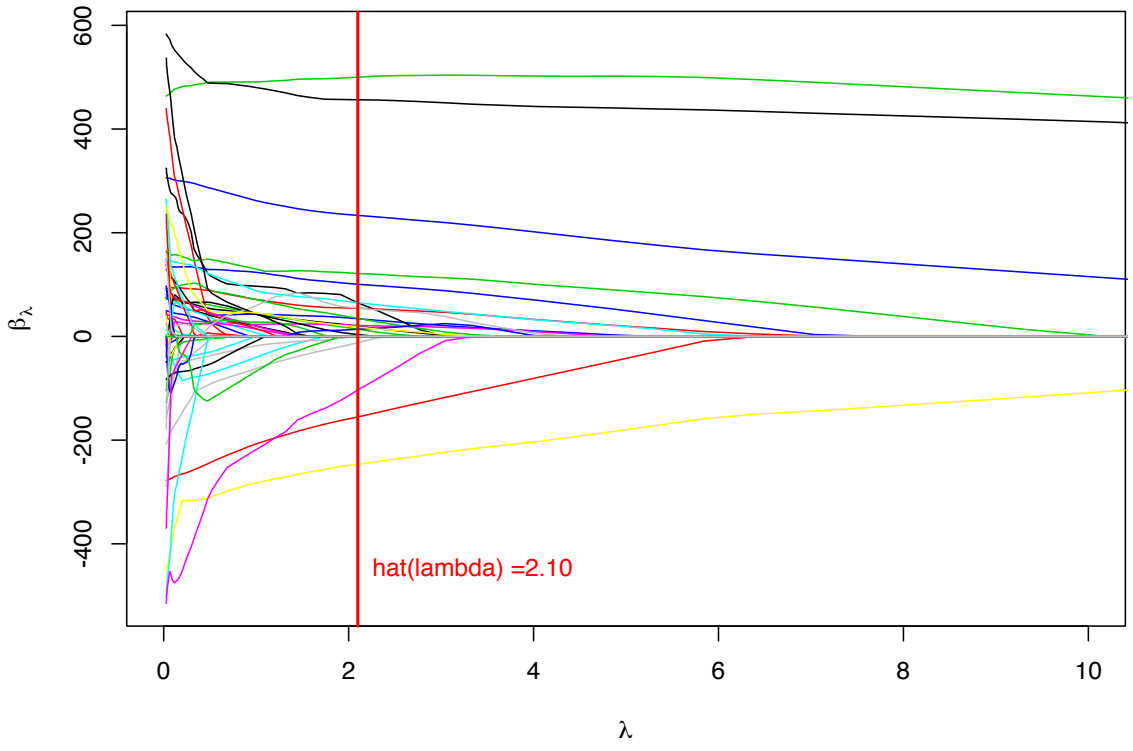


Figure – Évolution des coefficients $\left[\widehat{\beta}_\lambda^{(L)}\right]_j$ en fonction de λ.

Estimateur $\widehat{\beta}^{(L)}_{\widehat{\lambda}^{(L,CV)}}$ (I)

j	Variable j	x
2	age	19.76
3	sex	-155.37
4	bmi	499.94
5	map	233.46
6	tc	0.00
7	ldl	0.00
...		
12	age^2	14.05
13	bmi^2	34.14
...		
21	age×sex	100.54
...		

Estimateur $\widehat{\beta}^{(L)}_{\widehat{\lambda}^{(L,CV)}}$ (II)

21 variables sélectionnées

	x
age	19.76
sex	-155.37
bmi	499.94
map	233.46
hdl	-246.74
ltg	456.43
glu	54.33
age^2	14.05
bmi^2	34.14
tch^2	64.91
glu^2	121.37
age :sex	100.54
age :map	21.16
age :ldl	-12.17
age :ltg	35.07
sex :hdl	3.48
bmi :map	65.18
ldl :tch	54.77
hdl :ltg	13.86
tch :ltg	-103.28
tch :glu	18.32

Estimateur $\widehat{\beta}^{(L)}_{\widehat{\lambda}^{(L,CV)}}$ (II)

21 variables sélectionnées

age	19.76
sex	-155.37
bmi	499.94
map	233.46
hdl	-246.74
ltg	456.43
glu	54.33
age^2	14.05
bmi^2	34.14
tch^2	64.91
glu^2	121.37
age :sex	100.54
age :map	21.16
age :ldl	-12.17
age :ltg	35.07
sex :hdl	3.48
bmi :map	65.18
ldl :tch	54.77
hdl :ltg	13.86
tch :ltg	-103.28
tch :glu	18.32

Nom	Signification	Effet
bmi	IMC	aggravant
ltg	lamotrigine	aggravant
hdl	"bon" cholestérol	bénéfique
map	pression sanguine	aggravant
glu	glucose	aggravant
age	âge	aggravant
tch	cholestérol total	aggravant

Estimateur $\widehat{\beta}^{(L)}_{\widehat{\lambda}^{(L,CV)}}$ (II)

21 variables sélectionnées

age	19.76
sex	-155.37
bmi	499.94
map	233.46
hdl	-246.74
ltg	456.43
glu	54.33
age^2	14.05
bmi^2	34.14
tch^2	64.91
glu^2	121.37
age :sex	100.54
age :map	21.16
age :ldl	-12.17
age :ltg	35.07
sex :hdl	3.48
bmi :map	65.18
ldl :tch	54.77
hdl :ltg	13.86
tch :ltg	-103.28
tch :glu	18.32

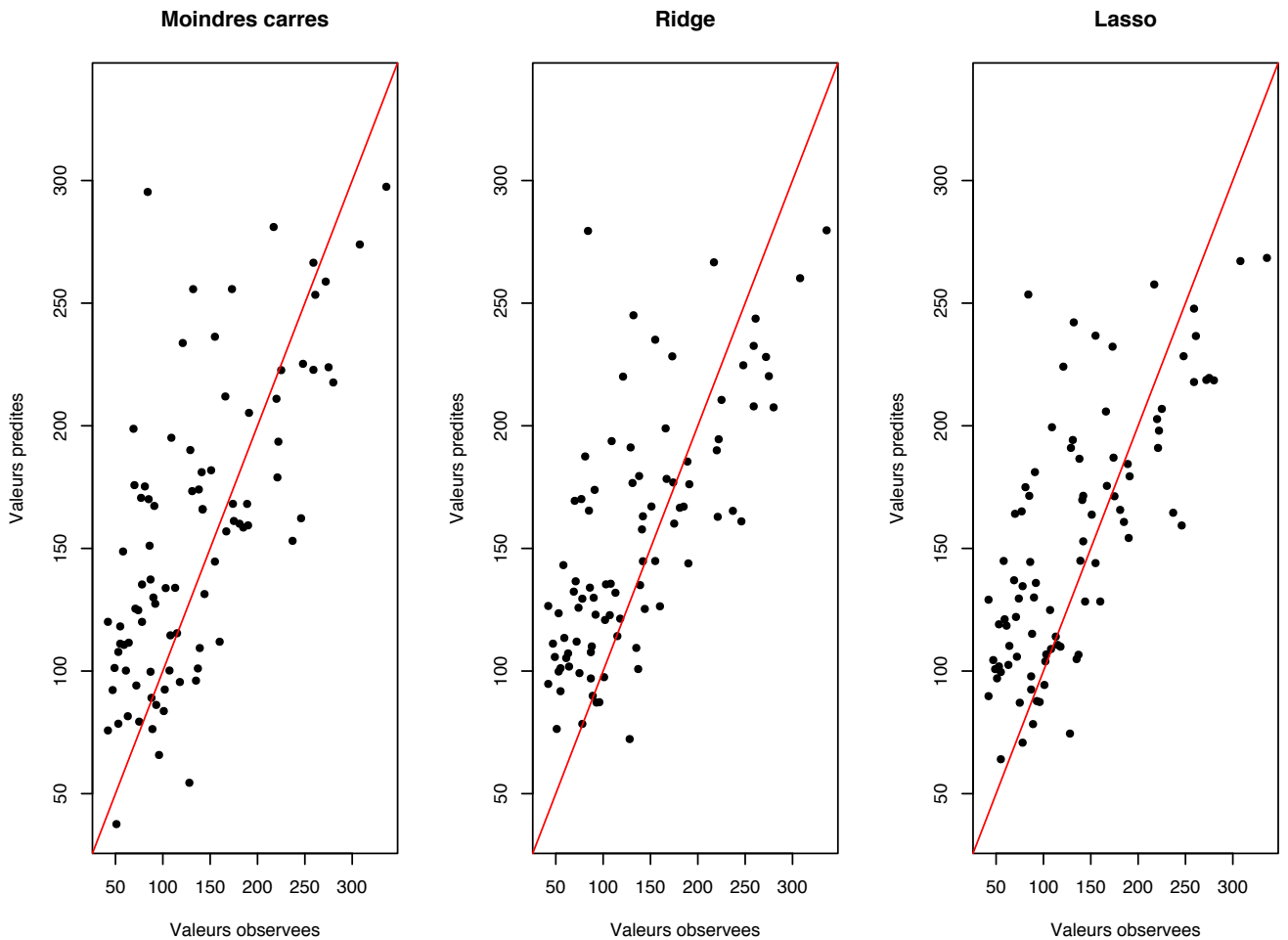
Nom	Signification	Effet
bmi	IMC	aggravant
ltg	?	aggravant
hdl	"bon" cholestérol	bénéfique
map	pression sanguine	aggravant
glu	glucose	aggravant
age	âge	aggravant
tch	cholestérol total	aggravant
ldl	"mauvais" cholestérol	

- Interactions jouant un rôle négatif : IMC et pression sanguine, "mauvais" cholestérol et cholestérol total, âge et ltg, âge et pression sanguine, cholestérol total et glucose, "bon" cholestérol et ltg.
- Interactions jouant un rôle bénéfique : cholestérol total et ltg, âge et "mauvais" cholestérol.

Comparaison avec $\widehat{\beta}^{(MCO)}$ et $\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}$

Nom variable	$\left[\widehat{\beta}^{(MCO)}\right]_j$	$\left[\widehat{\beta}^{(R)}_{\widehat{\lambda}^{(R)}}\right]_j$	$\left[\widehat{\beta}^{(L)}_{\widehat{\lambda}^{(L,CV)}}\right]_j$
age	81.66	79.01	19.76
sex	-288.70	-163.69	-155.37
bmi	473.06	387.58	499.94
map	312.14	245.95	233.46
tc	-36569.39	-8.90	0.00
ldl	32189.39	-52.76	0.00
...	...	...	
age^2	63.46	50.80	14.05
bmi^2	74.65	83.01	34.14
...	...	...	
age×sex	118.17	98.10	100.54
...	...	...	
Erreur sur I_{test}	3115	2829	2694

Valeurs prédites par $\widehat{\beta}^{(MC)}$, $\widehat{\beta}^{(R)}_{\widehat{\lambda}}$ et $\widehat{\beta}^{(L)}_{\widehat{\lambda}}$ sur l'échantillon de test



Remarques finales

- Contrairement à l'estimateur Ridge, le Lasso sélectionne les variables ⇒ avantage dans l'interprétation des résultats.
- Toutefois...
  - il peut ajouter un biais trop important (les coefficients ayant tendance à être trop petit),
  - l'estimateur est plus difficile à calculer que l'estimateur Ridge.
- D'autres méthodes existent (autres pénalisations) : Elastic Net (cf TD), Gauss-Lasso, Lasso adaptatif,...