Cours n°3 – Régression avancée

L. Carraro

Option MAFQ

Plan

- Bootstrap et bagging
- Retour sur les prédicteurs corrélés
- Régression ridge
- Les degrés de liberté
- LASSO
- Vers les méthodes L^1 avancées

Le bootstrap

- Méthode générale en statistiques
- Exemple élémentaire : intervalle de confiance
  - Soit (x_1 ,...,x_n) un n-échantillon de loi normale N(m,1)
  - Problème : intervalle de confiance pour m basé sur l'échantillon
- Ici, calcul théorique :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ est de loi } N(m,1/n)$$

Cas général

- Un intervalle de confiance de niveau α doit vérifier :

$$P_{\theta_0}\left(\theta_0 \in [l(X_1,...,X_n), u(X_1,...,X_n)]\right) = 1-\alpha$$

  - Pour l'exemple qui précède, on trouve un intervalle valable pour tout m=θ et non seulement pour le « bon » θ_0

- Bootstrap :

  Remplacer P_{θ_0} par une estimation notée P_n^*

  - choix le plus courant : loi empirique

$$P_n^* = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

Algorithme

- Pour b=1 à B, faire :
  - simuler un échantillon de taille n parmi $\{x_1,...,x_n\}$, noté X^{*b}
  - stocker l'estimation de θ obtenue, notée θ^{*b}
- Fin faire
- Evaluer les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de l'échantillon des θ^{*b}, b=1,...,B

$$IC_{bootstrap} = \left[q_{\alpha/2}, q_{1-\alpha/2}\right]$$

- Résultats : voir simulations R

Le bagging

- Bagging = Boostraping and Aggregating

- Idée :
  - rééchantillonner parmi les (x_i , y_i)
  - construire un modèle sur cet échantillon
  - moyenner les modèles obtenus

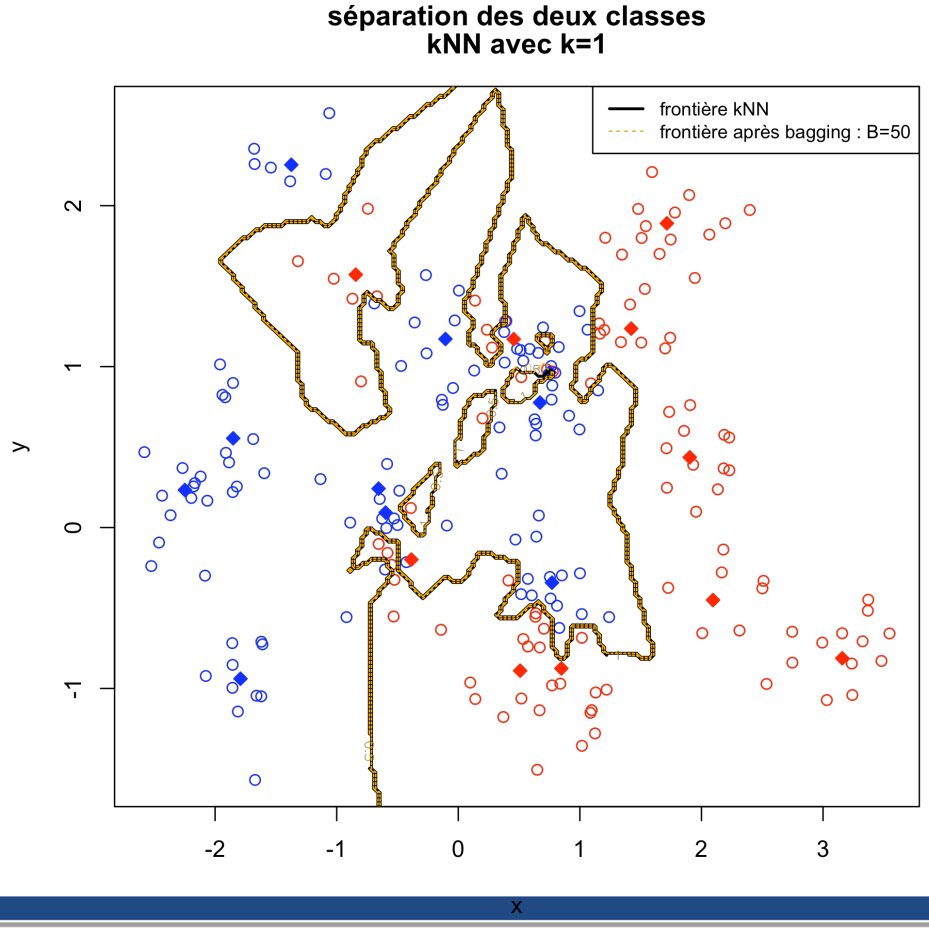
Algorithme pour le bagging

- Soit Z = (X,Y) ou (X,G) l'ensemble d'apprentissage
  - Pour b=1 à B, faire :
    - simuler un échantillon de taille n parmi $\{z_1,...,z_n\}$, noté Z^{*b}
    - estimer un modèle à partir de Z^{*b}
    - stocker le modèle obtenu, noté $\hat{Y}_b^*$
  - Fin faire
  - Poser :

$$\hat{Y}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_b^*(x)$$

Warning

- Inutile si régression linéaire
- Inutile si modèle linéaire en les réponses



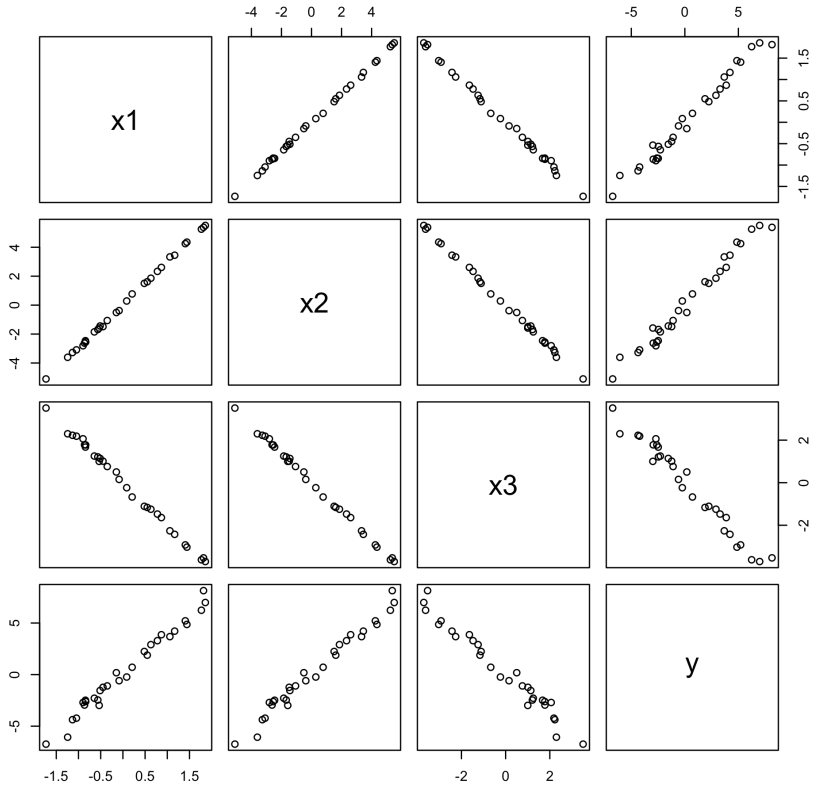
Retour sur prédicteurs corrélés

- Exemple synthétique
- On suppose que y dépend linéairement de la moyenne de 3 prédicteurs x_1, x_2, x_3 :

$$y = 2 \times (x_1 + x_2 + x_3) + \varepsilon$$

  - ε de loi N(0,0.1)
- Plan d'expériences à 30 points :
  - x_1 de loi N(0,1), x_2/x_1 de loi N($3x_1$,0.1)
  - $x_3/x_1,x_2$ de loi N(x_1-x_2,0.1)

pairplot



Modèle linéaire

```
> mod1<-lm(y~x1+x2+x3)
> summary(mod1)

Call:
lm(formula = y ~ x1 + x2 + x3)
```

```
Residual standard error: 0.4217 on
   26 degrees of freedom
Multiple R-squared: 0.9897,
        Adjusted R-squared: 0.9885
F-statistic: 835.2 on 3 and 26 DF,
   p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02610    0.07898   0.330 0.743746
x1           8.00830    3.10282   2.581 0.015847 *
x2           1.25169    1.35297   0.925 0.363401
x3           3.89028    0.92633   4.200 0.000277 ***
```

Analyse de modèles emboités

```
> mod1<-lm(y~x1+x2+x3)
> mod2<-lm(y~x3-1)
> anova(mod2,mod1)
```

```
Analysis of Variance Table

Model 1: y ~ x3-1
Model 2: y ~ x1 + x2 + x3
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	14.9925				
2	26	4.6244	3	10.368	19.431	8.084e-07 ***

Remarque : si conserve le modèle 2 :

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x3 -1.86923    0.06418  -29.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.719 on 29 degrees of freedom
Multiple R-squared: 0.9669,   Adjusted R-squared: 0.9658
F-statistic: 848.1 on 1 and 29 DF,  p-value: < 2.2e-16
```

Seconde simulation

```
> mod1<-lm(y~x1+x2+x3)
> summary(mod1)

Call:
lm(formula = y ~ x1 + x2 + x3)
```

```
Residual standard error: 0.6267 on
   26 degrees of freedom
Multiple R-squared: 0.9745,
        Adjusted R-squared: 0.9716
F-statistic: 331.1 on 3 and 26 DF,
   p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05785    0.11948  -0.484    0.632
x1           2.92216    3.91420   0.747    0.462
x2           1.64002    1.55865   1.052    0.302
x3           1.95736    1.02395   1.912    0.067 .
```

Rien de significatif avec $R^2 = 97\%$!!

Troisième simulation

```
> mod1<-lm(y~x1+x2+x3)
> summary(mod1)

Call:
lm(formula = y ~ x1 + x2 + x3)
```

```
Residual standard error: 0.5013 on
   26 degrees of freedom
Multiple R-squared: 0.9859,
        Adjusted R-squared: 0.9843
F-statistic: 606.6 on 3 and 26 DF,
   p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11513    0.09604   1.199  0.24141
x1           0.89016    2.84315   0.313  0.75671
x2           2.96368    1.25139   2.368  0.02559 *
x3           2.88135    1.03635   2.780  0.00996 **
```

Troisième simulation modèles emboités


```
> mod1<-lm(y~x1+x2+x3)
> mod2<-lm(y~x2+x3-1)
> anova(mod2,mod1)
```

```
Analysis of Variance Table

Model 1: y ~ x2+x3-1
Model 2: y ~ x1 + x2 + x3
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	6.8955				
2	26	6.5343	2	0.3612	0.7186	0.4969

Ici, on conserve le modèle 2 :

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x2   3.1844     0.6725   4.735 5.72e-05 ***
x3   2.7640     1.0057   2.748   0.0104 *

Residual standard error: 0.4963 on 28 degrees of freedom
Multiple R-squared: 0.9852,   Adjusted R-squared: 0.9842
F-statistic: 933.8 on 2 and 28 DF,  p-value: < 2.2e-16
```

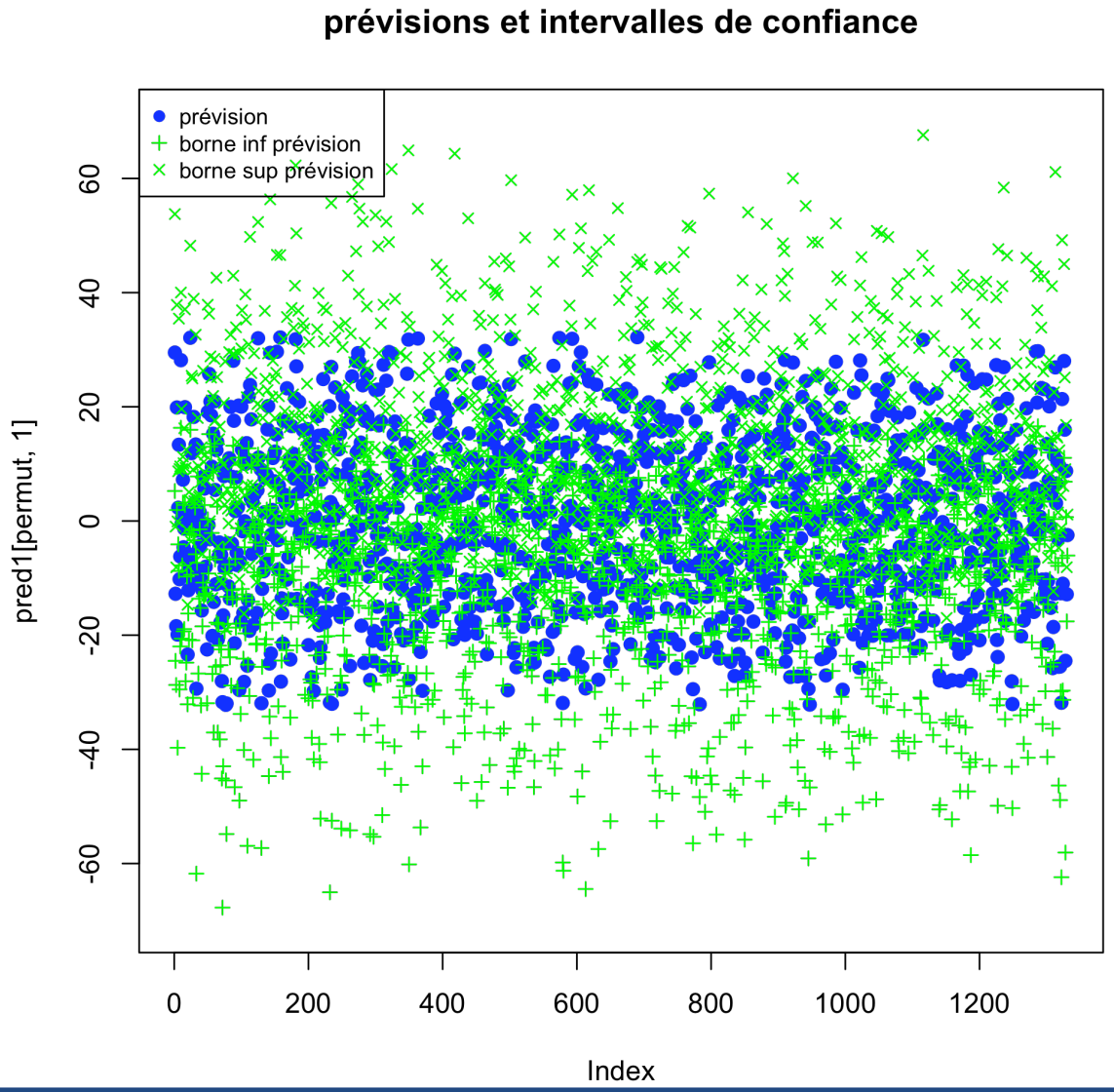
Qualité de prédiction dernier modèle

- Plan test factoriel à 1331 points
  - x_1 entre -2 et 2
  - x_2 entre -6 et 6
  - x_3 entre -4 et 4
- Ecart type des écarts prédit – simulé :
  - 13.96 (au lieu de 0.5 !)
- Pourcentage de réponses dans l'intervalle de prévision à 95%
  - 11,8 %

Qualité de prédiction

- Troisième simulation
- modèle complet
- Ecart type des écarts prédit – simulé :
  - 13.96 (au lieu de 0.5 !)
- Pourcentage de réponses dans l'intervalle de prévision à 95%
  - 53 %
  - La colinéarité des prédicteurs augmente la taille de l'intervalle de prévision

Illustration

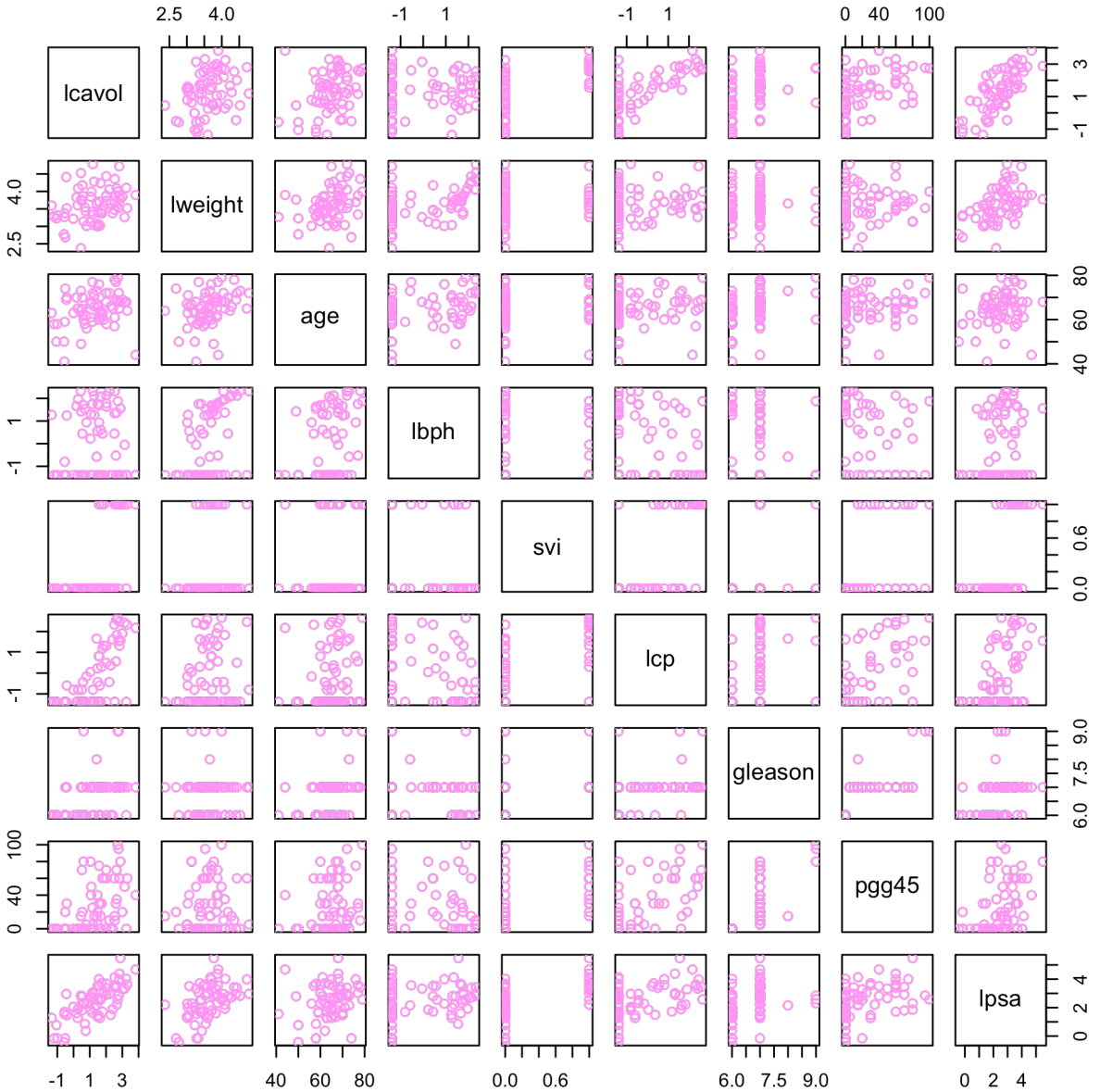


Conclusion

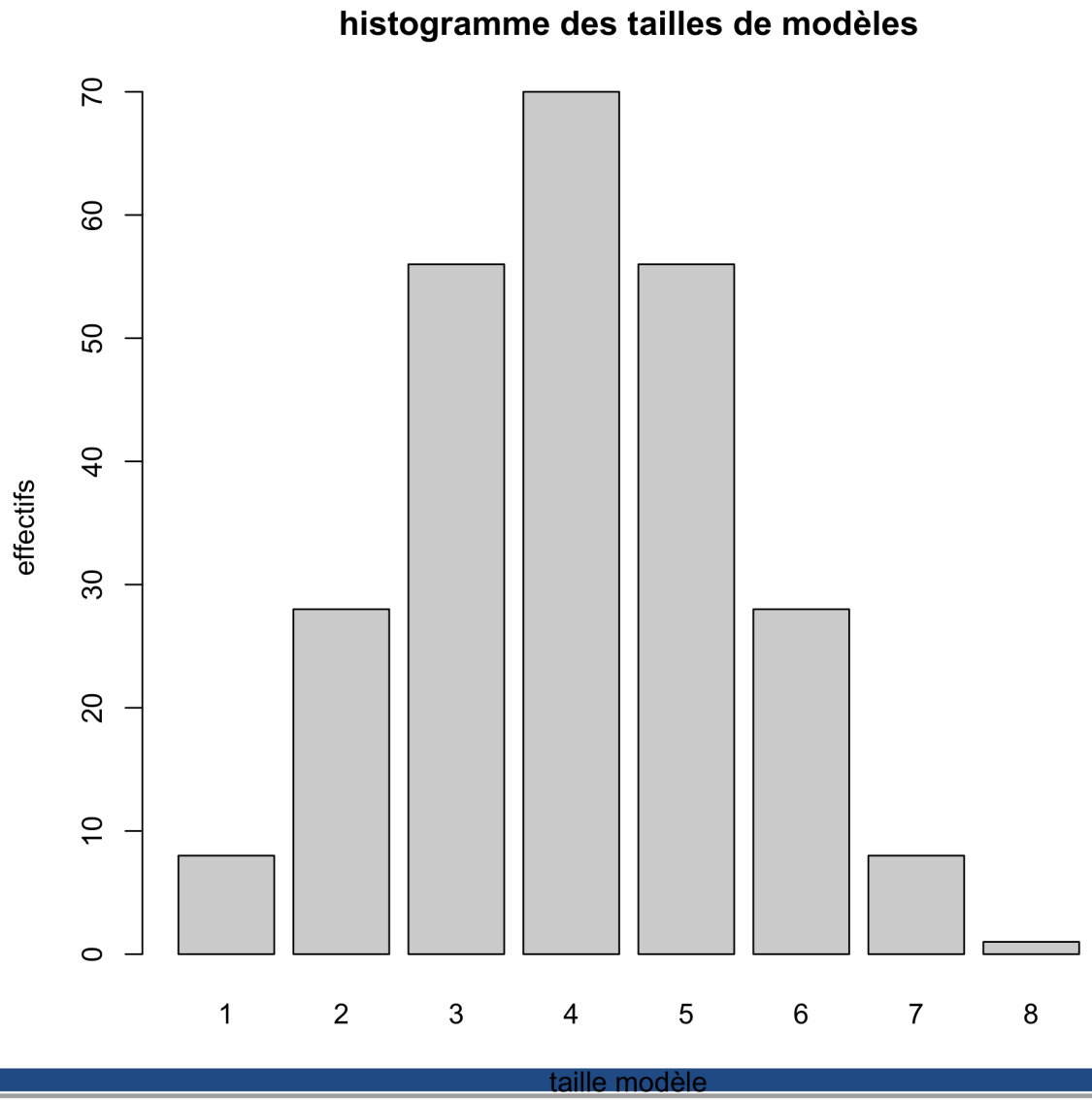
- Corrélation des prédicteurs entraîne :
  - des choix de modèles erronés
  - des modèles souvent structurellement aberrants (cf. signe des coefficients)
  - des modèles instables
  - des modèles strictement valables dans le domaine d'apprentissage

Exemple « prostate »

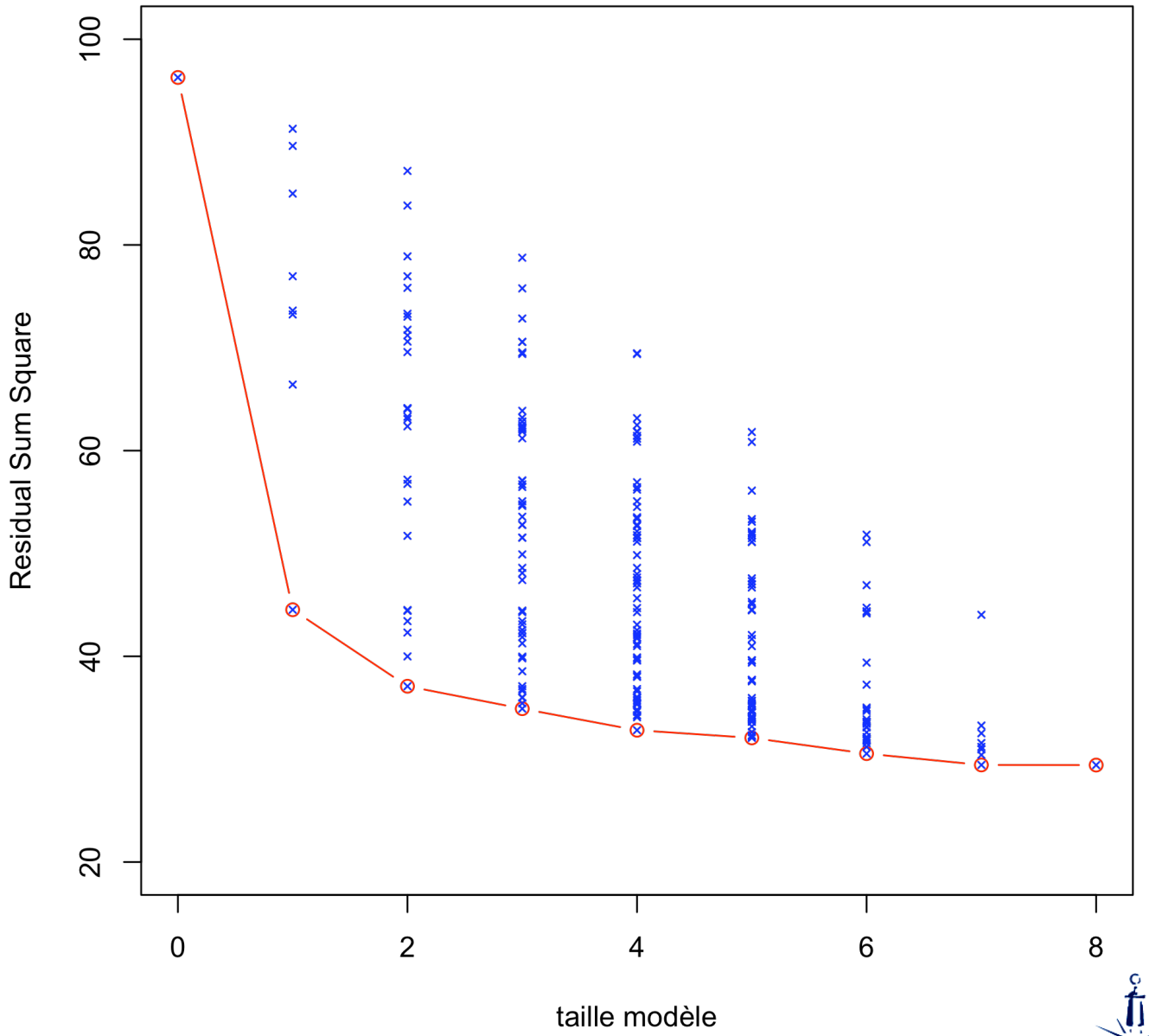
- Réponse : taux de PSA (prostate-specific antigen) : lpsa
- Prédicteurs :
  - lcavol (log volume cancer)
  - lweight (log poids prostate)
  - age (en années)
  - lbph (log benign prostatic hyperplasia)
  - svi (invasion vésicule séminale)
  - lcp (log pénétration capsulaire)
  - gleason (indice : 6, 7, 8, 9)
  - pgg45(pourcentage de Gleason 4 ou 5)


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
lpsa

On essaye tous les sous-modèles



régression sur tous les sous-modèles



Pourquoi un sous-modèle ?

- Prévision
  - intervalles de prévision corrects
  - intervalles de prévision de taille raisonnable
- Interprétation
  - on veut la « big picture »
  - détails sans importance

Régression robuste « ridge regression »

- On part d'un modèle linéaire :

  $Y = X\beta + \varepsilon$

  on suppose les colonnes de X normées

- Estimation de β

$$Arg\min_{\beta}\left\{\|Y - X\beta\|^2 + \lambda\|\beta\|^2\right\}$$

ou

$$Arg\min_{\|\beta\|\leq c}\left\{\|Y - X\beta\|^2\right\}$$

Ridge : estimation de β

$$\hat{\beta}_{ridge} = \left(X'X + \lambda Id\right)^{-1} X'Y$$

- Augmentation du conditionnement matriciel
- Remarque : si les colonnes de X sont orthogonales

$$\hat{\beta}_{ridge} = \frac{1}{1+\lambda} X'Y = \frac{1}{1+\lambda}\hat{\beta}$$

Vision bayésienne

- On suppose B de loi a priori $N(0,\eta^2 Id)$
- Le modèle de régression donne :
  - $Y/B=\beta$ de loi $N(X\beta,\sigma^2 Id)$
- Loi (a posteriori) de B :
  - $f_{B/Y=y}(\beta) = f_{Y/B=\beta}(y)\, f_B(\beta)/f_Y(y)$
  - Espérance et mode (a posteriori) de Y sont égaux à l'estimateur ridge, avec $\lambda=\sigma^2/\eta^2$

Evaluations matricielles

- Décomposition en valeurs singulières de X :
  - X = U D V'
    - U n*p orthogonale : U'U=Id_p
    - D p*p diagonale
    - V p*p orthogonale : V'V=Id_p

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

NB : U'Y donne les coordonnées de Y dans la base des colonnes de U

Cas de l'estimateur ridge

$$\hat{Y}_{ridge} = X\hat{\beta}_{ridge} = X(X'X + \lambda Id)^{-1}X'Y$$

$$= UD(D^2 + \lambda Id)^{-1}DU'Y$$

$$= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j' Y$$

Décomposition de Y sur la base des colonnes de U, puis diminution des coefficients gouvernée par λ

Les degrés de liberté

- En régression linéaire :
  - p = dim(ev{colonnes de X})
    = dim(Im(H))

  où H est la matrice chapeau : $H=X(X'X)^{-1}X'$

  H est un projecteur : dim(Im(H))=Tr(H)
- En régression ridge :

  $H_\lambda=X(X'X+\lambda\ Id)^{-1}X'$ n'est pas un projecteur
- Définition très générale (ex. splines)

Degrés de liberté en régression ridge

- On pose :

  $df(\lambda) = Tr(H_\lambda)$

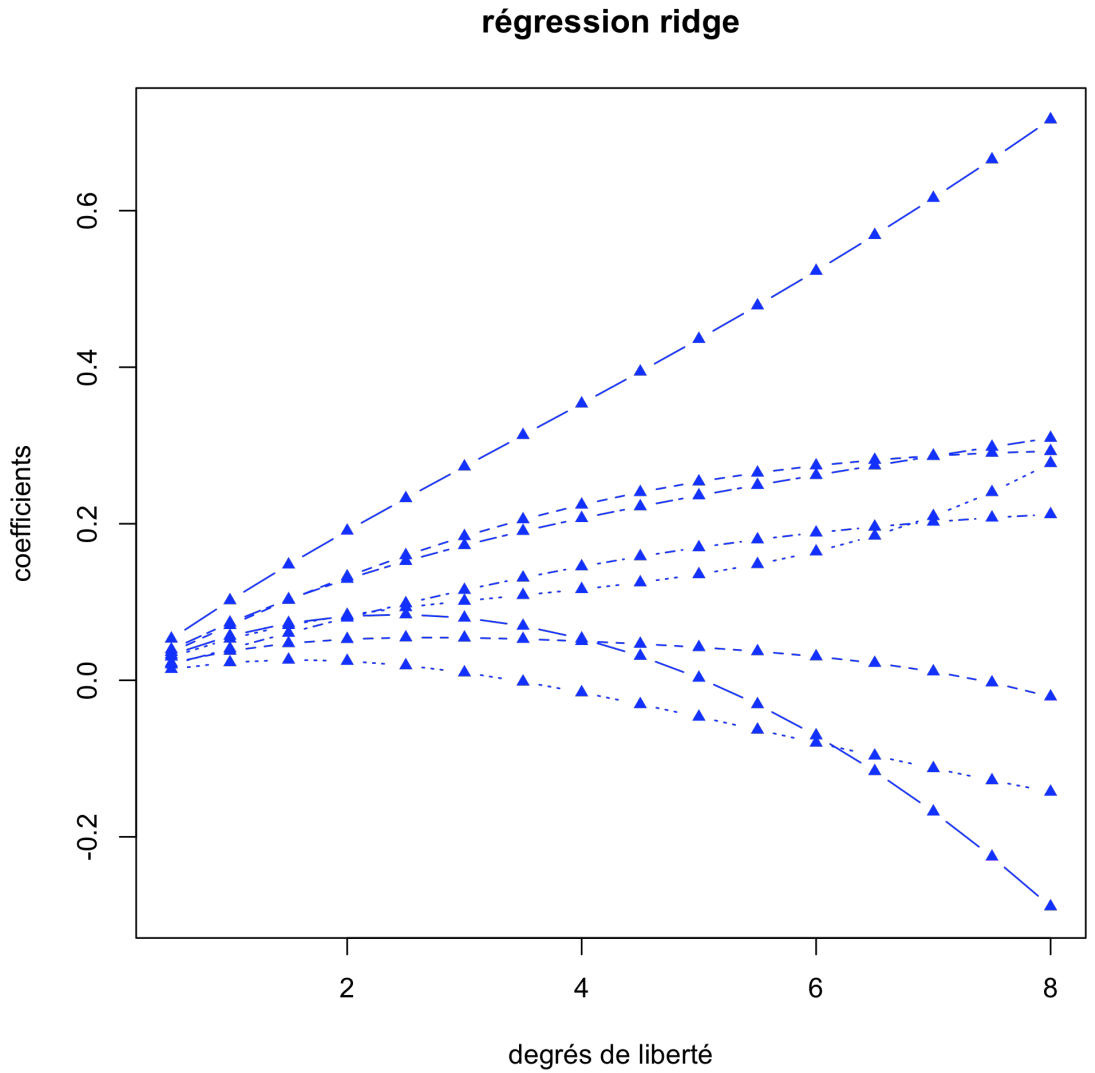
- Propriétés :

$$Tr(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

  – df(0)=p, df(.) tend vers 0 à l'infini

  – df(.) est décroissante

Evolution des coefficients



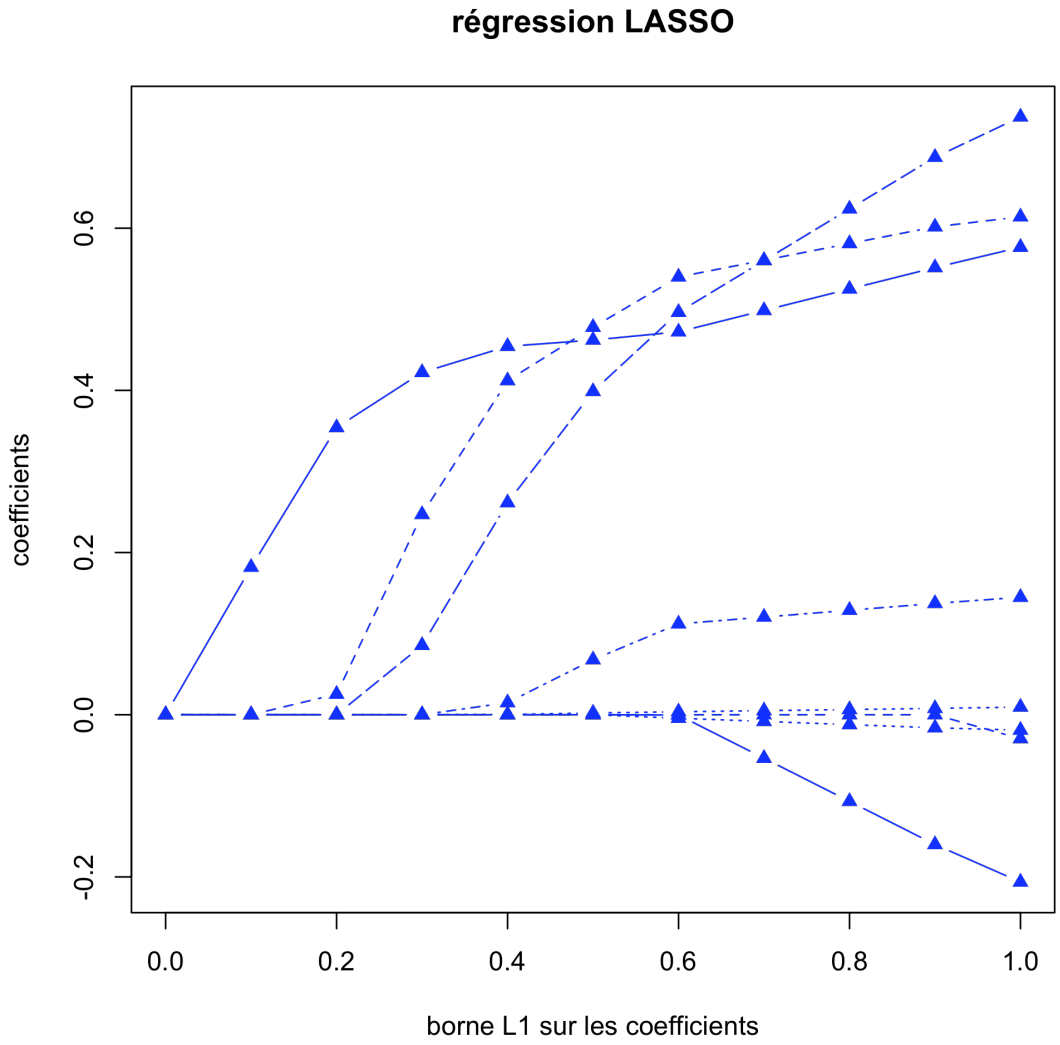
Le LASSO

- Même idée que ridge :

$$\underset{\|\beta\|_1 \leq c}{Arg\min}\left\{\|Y - X\beta\|^2\right\}$$

- Mais :
  - on remplace $\|\beta\|_2$ par $\|\beta\|_1$

Evolution des coefficients



Différences fondamentales

- Estimation de β par LASSO via un algorithme de programmation quadratique
- Estimation de β par ridge par inversion matricielle
- LASSO donne des coefficients nuls