Apprentissage automatique / Statistique

Sélection de modèle dans le cas gaussien

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Objectifs

- Expliquer Y quantitative avec $X^1, \ldots, X^p$
- Modèle gaussien et linéaire général
- Dianostic : multicolinéarité (influence, tests, résidus)
- Choix de modèle par sélection de variables
- Choix de modèle par pénalisation (*ridge*, Lasso)

Hypothèses du Modèle linéaire

- Échantillon taille n : $(x_i^1, \ldots, x_i^p, y_i); i = 1, \ldots, n$

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \varepsilon_i; i = 1, \ldots, n$$

- Hypothèses
  - $E(\varepsilon_i) = 0, Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$
  - X^j déterministes ou bien ε indépendant des X^j
  - $\beta_0, \ldots, \beta_p$ constants
  - Option $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$

Expression matricielle

- $E(\varepsilon_i) = 0, Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$
- $\mathbf{X}(n \times (p+1))$ de terme général x_i^j avec $\boldsymbol{x}^0 = \mathbf{1}$
- $\mathbf{Y}$ de terme général y_i
- $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_p]'$
- $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Estimateur des moindres carrés

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2$$

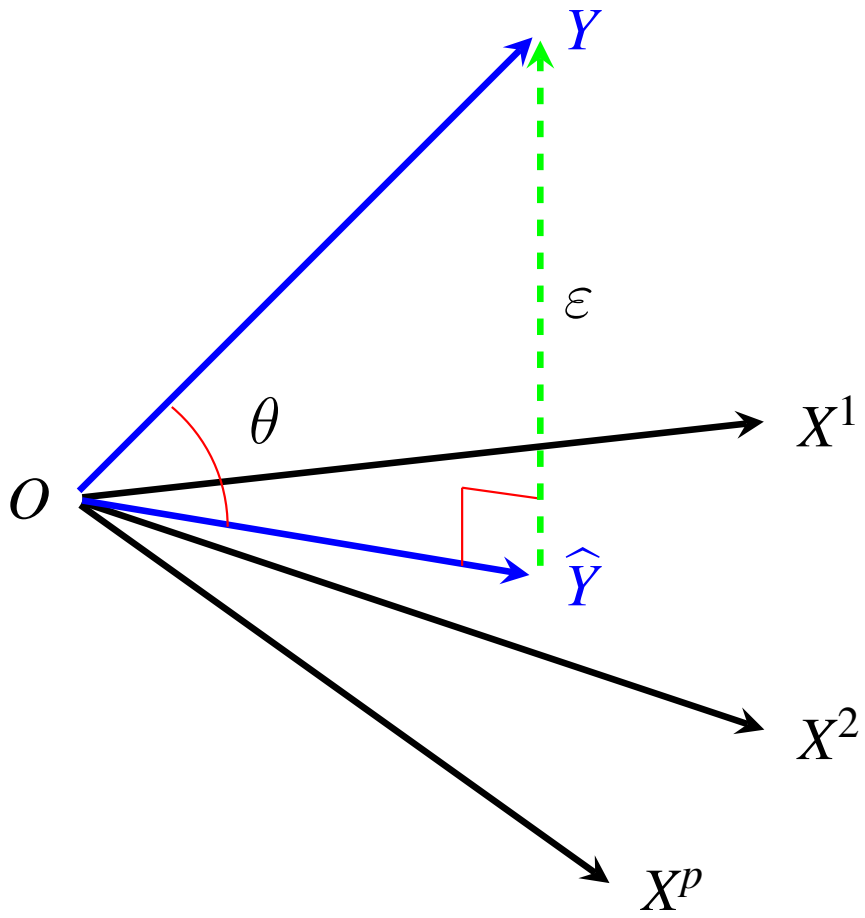
- Equations normales : $\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=0$
- et si $\mathbf{X}'\mathbf{X}$ inversible
- Estimation de β : $\widehat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Prédiction de $\mathbf{Y}$: $\widehat{\mathbf{Y}}=\mathbf{X}\mathbf{b}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=\mathbf{H}\mathbf{Y}$
- $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$: projection orthog. sur Vect($\mathbf{X}$)
- Résidus : $\mathbf{e}=\mathbf{Y}-\widehat{\mathbf{Y}}=\mathbf{Y}-\mathbf{X}\mathbf{b}=(\mathbf{I}-\mathbf{H})\mathbf{Y}$

Covariances des estimateurs

$$\begin{aligned} E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ E[(\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})(\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})'] &= \sigma^2\mathbf{H} \\ E[\mathbf{e}\mathbf{e}'] &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

Estimation de σ^2

$$s^2 = \frac{\left\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|^2}{n-p-1} = \frac{\text{SSE}}{n-p-1}$$



Projection $\widehat{Y}$ de Y sur l'espace vectoriel Vect$\{\mathbf{1}, X^1, \ldots, X^p\}$

Sommes des carrés

$$
\begin{aligned}
\text{SSE} &= \left\|\mathbf{Y} - \overline{\mathbf{Y}}\right\|^2 = \|\mathbf{e}\|^2 \\
\text{SST} &= \left\|\mathbf{y} - \overline{Y}\mathbf{1}\right\|^2 = \mathbf{Y}'\mathbf{Y} - n\overline{Y}^2 \\
\text{SSR} &= \left\|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}\right\|^2 = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2 \\
\text{SST} &= \text{SSR} + \text{SSE} \\
R^2 &= \frac{\text{SSR}}{\text{SST}} \quad \textit{Coefficient de détermination}
\end{aligned}
$$

Cosinus carré de l'angle entre $\mathbf{Y}$ et $\widehat{\mathbf{Y}}$

Inférence sur les coefficients

La statistique $\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_j} \sim$ Student à $(n-p-1)$ ddl

H_0 : $\beta_j = a$ et intervalle de confiance de niveau $100(1-\alpha)\%$:

$$\widehat{\beta}_j \pm t_{\alpha/2;(n-p-1)} \widehat{\sigma}_j$$

Attention les coefficients sont corrélés entre eux

Inférence sur le modèle

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

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

Fisher avec p et $(n-p-1)$ ddl

Tableau d'analyse de la variance

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	F
Régression	p	SSR	MSR=$\frac{\text{SSR}}{p}$	MSR/MSE
Erreur	$n-p-1$	SSE	MSE=$\frac{\text{SSE}}{(n-p-1)}$	
Total	$n-1$	SST		

Inférence sur un modèle réduit

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0, q < p$
SSR_q, SSE_q, R^2_q du modèle réduit à $(p-q)$ variables

$$\frac{(\text{SSR} - \text{SSR}_q)/q}{\text{SSE}/(n-p-1)} = \frac{(R^2 - R^2_q)/q}{(1-R^2)/(n-p-1)}$$

Fisher à q et $(n-p-1)$ ddl

Inférence sur la Prévision

Pour $\mathbf{x}_0$:

$$\widehat{y}_0 = b_0 + b_1 x_0^1 + \cdots + b_p x_0^p.$$

Intervalles de confiance des prévisions de Y et $E(Y)$

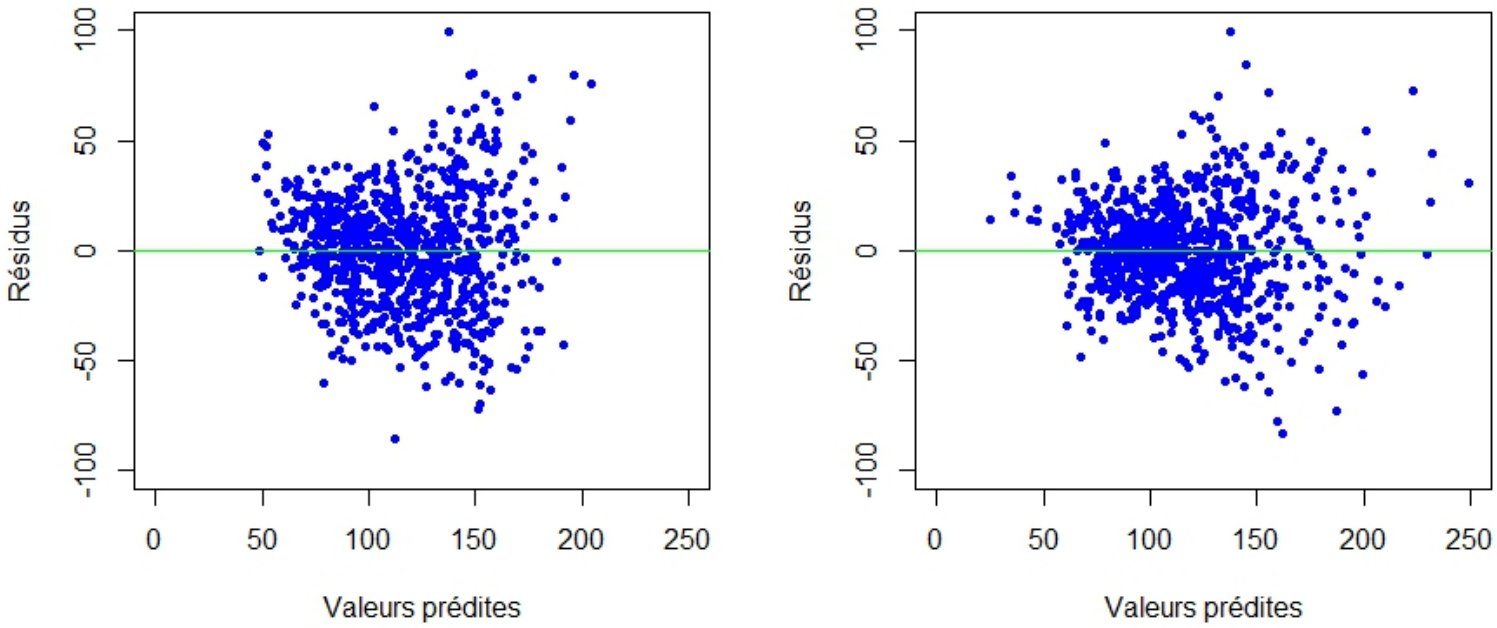
$$\begin{aligned} \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)} s(1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2} \\ \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)} s(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2} \end{aligned}$$

avec $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$

Diagnostics des Résidus

- Homoscédasticité, linéarité, normalité
- Effet levier : $\mathbf{H}_i i$ et résidu *studentisé* grand par
- Distance de *Cook* :

$$D_i = \frac{1}{s^2(p+1)}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})$$



Ozone : Résidus des modèles linéaire et quadratique.

Diagnostics de colinéarité

- conditionnement de $\mathbf{X}'\mathbf{X}$
- Facteurs d'inflation de la variance (VIF) : $V_j = \frac{1}{1-R_j^2}$
- Conditionnement : $\kappa = \lambda_1/\lambda_p$

Retour sur capital

40 entreprises du Royaume Uni décrites par

RETCAP	Return on capital employed
WCFTDT	Ratio of working capital flow to total debt
LOGSALE	Log to base 10 of total sales
LOGASST	Log to base 10 of total assets
CURRAT	Current ratio
QUIKRAT	Quick ratio
NFATAST	Ratio of net fixed assets to total assets
FATTOT	Gross sixed assets to total assets
PAYOUT	Payout ratio
WCFTCL	Ratio of working capital flow to total current liabilities
GEARRAT	Gearing ratio (debt-equity ratio)
CAPINT	Capital intensity (ratio of total sales to total assets)
INVTAST	Ratio of total inventories to total assets

Modèle complet

Analysis of Variance

Source	DF (1)	Sum of Squares	Mean Square	F Value	Prob>F
Model	12	0.55868 (2)	0.04656 (5)	8.408 (7)	0.0001 (8)
Error	27	0.14951 (3)	0.00554 (6)		
C Total	39	0.70820 (4)			

Root MSE	0.07441 (9)	R-square	0.7889 (12)
Dep Mean	0.14275 (10)	Adj R-sq	0.6951 (13)
C.V.	52.12940 (11)		

(1)	d.d.l. de la loi de Fisher du test global	(8)	$P(f_{p;n-p-1} > F)$; H_0 rejetée au niveau α si $P < \alpha$
(2)	SSR	(9)	s =racine de MSE
(3)	SSE ou déviance	(10)	moyenne empirique de la variable à expliquée
(4)	SST=SSE+SSR	(11)	Coefficient de variation $100\times$ (9)/(10)
(5)	SSR/DF	(12)	Coefficient de détermination R^2
(6)	MSE=SSE/DF est l'estimation de σ_u^2	(13)	Coefficient de détermination ajusté R'^2
(7)	Statistique F de Fisher du test global		

Paramètres du modèle

Parameter Estimates

Variable	DF	Parameter Estimate (1)	Standard Error (2)	T for H0: Parameter=0 (3)	Prob>|T| (4)	Tolerance (5)	Variance Inflation (6)
INTERCEP	1	0.188072	0.13391661	1.404	0.1716	.	0.0000000
WCFTCL	1	0.215130	0.19788455	1.087	0.2866	0.03734409	26.777998
WCFTDT	1	0.305557	0.29736579	1.028	0.3133	0.02187972	45.704415
GEARRAT	1	-0.040436	0.07677092	-0.527	0.6027	0.45778579	2.184428
LOGSALE	1	0.118440	0.03611612	3.279	0.0029	0.10629382	9.407885
LOGASST	1	-0.076960	0.04517414	-1.704	0.0999	0.21200778	4.716808

...

(1) estimations des paramètres (β_j)
(2) écarts-types de ces estimations (s_j)
(3) statistique T du test de Student de $H_0 : \beta_j = 0$
(4) $P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$
(5) $1 - R^2_{(j)}$
(6) VIF=$1/(1 - R^2_{(j)})$

AnCoVa élémentaire

- $\mathbf{Y}$ expliquée par
- $\boldsymbol{T}$ à J niveaux et
- $\boldsymbol{X}$ quantitative (covariable)
- Pour chaque niveau j de $\boldsymbol{T}$, on observe
- n_j valeurs $X_{1j}, \ldots, X_{n_j j}$ de $\boldsymbol{X}$ et
- n_j valeurs $Y_{1j}, \ldots, Y_{n_j j}$ de $\mathbf{Y}$;
- $n = \sum_{j=1}^{J} n_j$ taille de l'échantillon
- $E[\mathbf{Y}|\boldsymbol{T}]$ est fonction affine des variables explcatives
- $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$; $\quad j = 1, \ldots, J; \quad i = 1, \cdots, n_j$
- ε_{ij} supposés i.i.d éventuellement $\mathcal{N}(0, \sigma^2)$

Notations de l'AnCoVa

- $\boldsymbol{Y}$ observations $[Y_{ij}|i = 1, n_j; j = 1, J]'$
- $\mathbf{x}$ vecteur $[X_{ij}|i = 1, n_j; j = 1, J]'$
- $\boldsymbol{\varepsilon} = [\varepsilon_{ij}|i = 1, n_j; j = 1, J]'$ vecteur des erreurs
- $\mathbf{1}_j$ variables indicatrices des niveaux
- $\mathbf{x}.\mathbf{1}_j$ valeurs pour le niveau j, 0 ailleurs
- $\mathbf{X}$ matrice $n \times 2J$ $[\mathbf{1}_j|\mathbf{x}.\mathbf{1}_j]$ $; j = 1, \ldots, J$

Modèle et paramètres

- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- $\mathbf{X}$ est la matrice $n \times 2J$: $[\mathbf{1}_j | \mathbf{X}.\mathbf{1}_j]$; $j = 1, \ldots, J$
- Reparamétrisation :
- $\mathbf{X} = [\mathbf{1} | \mathbf{X} | \mathbf{1}_1 | \cdots | \mathbf{1}_{J-1} | \mathbf{x}.\mathbf{1}_1 | \cdots | \mathbf{x}.\mathbf{1}_{J-1}]$

$$Y_{ij} = \beta_{0J} + (\beta_{0j} - \beta_{0J}) + \beta_{1J} X_{ij} + (\beta_{1j} - \beta_{1J}) X_{ij} + \varepsilon_{ij} \; ;$$
$$j = 1, \ldots, J-1; i = 1, \ldots, n_j.$$

Tests

Comparer le modèle complet :

$$\begin{aligned}\mathbf{Y} = \beta_{0J}\mathbf{1} \quad &+ \quad (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \\ &+ \quad (\beta_{11} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \boldsymbol{\varepsilon}\end{aligned}$$

A chacun des modèles réduits :

$$(i) \quad \mathbf{Y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \boldsymbol{\varepsilon}$$

$$(ii) \quad \mathbf{Y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \boldsymbol{\varepsilon}$$

$$(iii) \quad \mathbf{Y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + (\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} +$$

$$(iv) \quad \mathbf{Y} = \beta_{0J}\mathbf{1} + \boldsymbol{\varepsilon}$$

Hypothèses testées

- H_0^i : pas d'interaction entre variables $\mathbf{X}$ et $\mathbf{T}$, $\beta_{11} = \cdots = \beta_{1J}$, les droites partagent la même pente β_{1J}.
- H_0^{ii} : $\beta_{11} = \cdots = \beta_{1J}$=0 (pas d'effet de $\mathbf{x}$)
- H_0^{iii} :$\beta_{01} = \cdots = \beta_{0J}$, les droites partagent la même constante à l'origine β_{0J}.
- H_0^{iv} les variables $\mathbf{X}$ et $\mathbf{T}$ n'ont aucun effet sur $\mathbf{Y}$.

Données marketing

Observations des
Consommation de lait après deux mois de
6 familles de taille 1 à 6 dans
4 villes ou campagnes de pub de
5 régions
Modéliser la consommation en fonction de la taille de la famille conditionnellement au type de campagne publicitaire

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PUB	3	227.1807	75.7269	0.57	0.6377 (1)
TAILLE	1	40926.0157	40926.0157	306.57	0.0001 (2)
TAILLE*PUB	3	309.8451	103.2817	0.77	0.5111 (3)

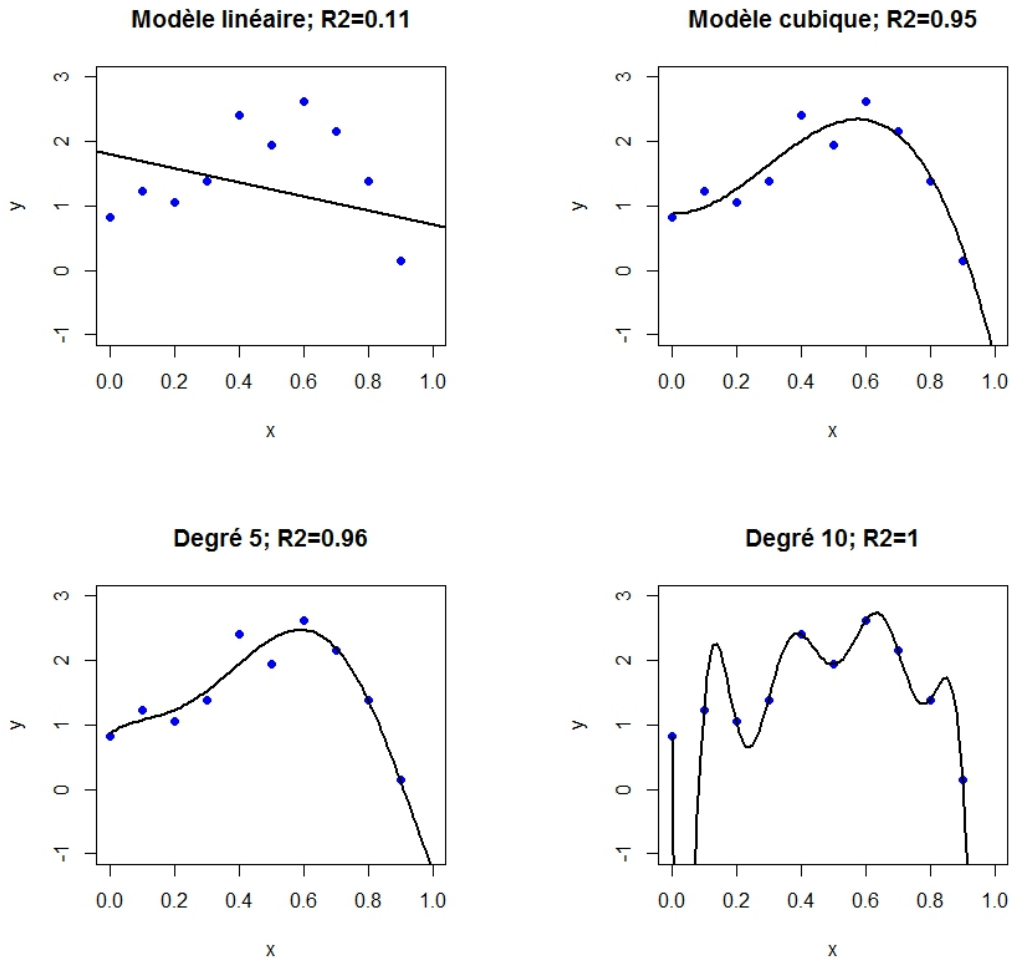
Tests

Attention aux interactions

Région	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	PUB	3	72.02974	24.00991	4.62	0.0164
1	TAILLE	1	7178.32142	7178.32142	1380.25	0.0001
	TAILLE*PUB	3	217.37048	72.45683	13.93	0.0001
	PUB	3	231.73422	77.24474	30.36	0.0001
2	TAILLE	1	8655.25201	8655.25201	3402.34	0.0001
	TAILLE*PUB	3	50.15069	16.71690	6.57	0.0042
	PUB	3	79.54688	26.51563	6.01	0.0061
3	TAILLE	1	6993.30160	6993.30160	1585.35	0.0001
	TAILLE*PUB	3	173.19305	57.73102	13.09	0.0001
	PUB	3	415.66664	138.55555	15.23	0.0001
4	TAILLE	1	9743.37830	9743.37830	1071.32	0.0001
	TAILLE*PUB	3	361.39556	120.46519	13.25	0.0001
	PUB	3	15.35494	5.11831	0.79	0.5168
5	TAILLE	1	8513.28516	8513.28516	1314.71	0.0001
	TAILLE*PUB	3	52.75119	17.58373	2.72	0.0793

Objectif de parcimonie en prévision

- Modèle
  - Explicatif (tests)
  - Prédictif
  - Le R^2 n'est pas un bon critère
  - Biaiser le modèle pour réduire la variance
    - Réduire le nombre de variables
    - Contraindre les paramètres



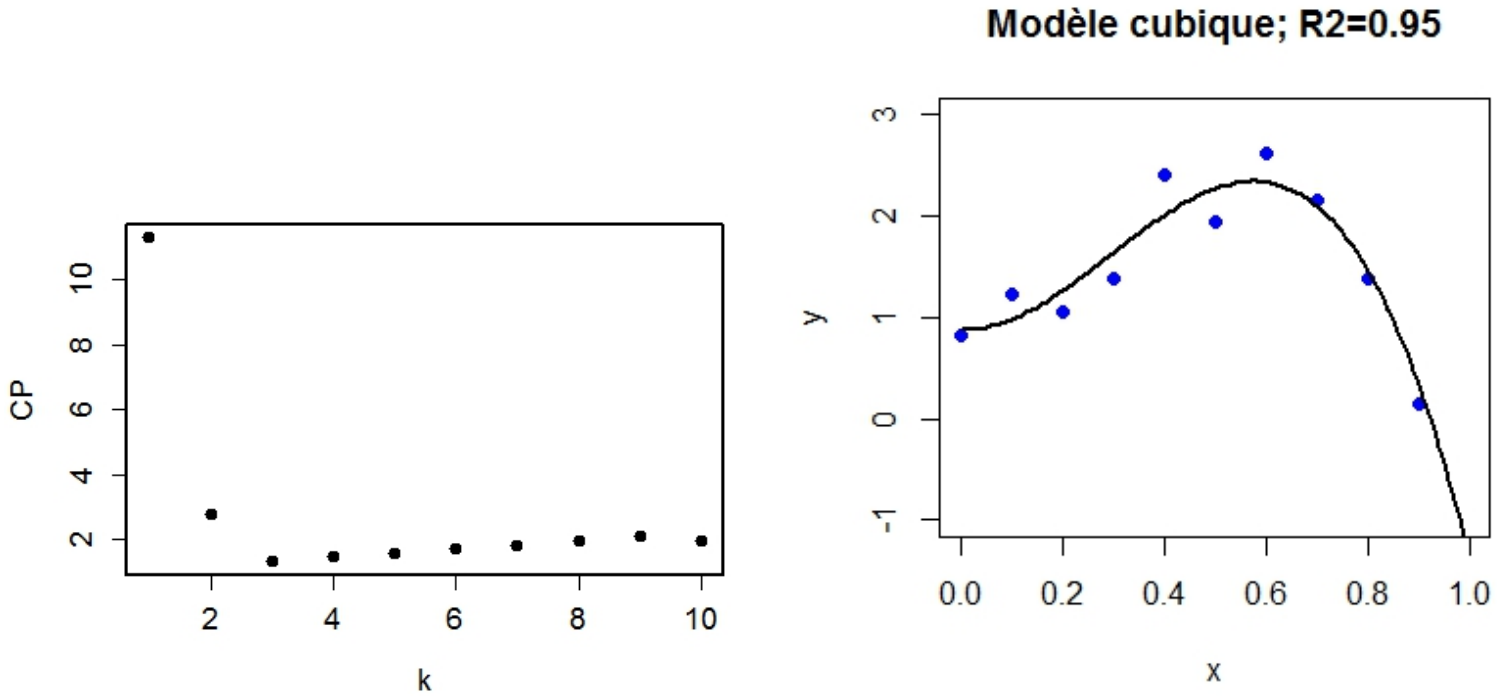
Régression polynomiale de degré 1, 2, 5 et 10

Critères de prévision

- Tous les critères sont équivalents avec q fixé
- Problème : optimisé le choix de q
- C_p de Mallow $\text{MSE}(\widehat{y}_i) = \text{Var}(\widehat{y}_i) + [\text{Biais}(\widehat{y}_i)]^2$
  On suppose le modèle complet sans biais

$$C_j = (n-j-1)\frac{\text{MSE}_j}{\text{MSE}} - [n-2(j+1)]$$

- $C_p = \widehat{R_n}(\widehat{f}(\boldsymbol{d}^n), \boldsymbol{d}^n) + 2\frac{d}{n}\widehat{\sigma}^2$
- $\text{AIC} = -2\mathcal{L} + 2\frac{d}{n}$
- $\text{BIC} = -2\mathcal{L} + \log(n)\frac{d}{n}$
- $\text{PRESS} = \sum_{i=1}^{n}(y_i - \widehat{y}_{(i)})^2$ (*leave one out cross validation*)



Régression polynomiale : minimisation du C_p de Mallows.

Algorithmes de sélection

Rechercher dans le graphe des 2^p modèles possibles

- Sélection ascendante (*forward*)
- Élimination descendante (*backward*)
- Mixte (pas à pas ou *step wise*)
- Globale (Furnival & Wilson, 1974), (`leaps` de R)
- Analyse de covariance :
  - AIC mais pas le C_p
  - Interactions et effets principaux

Stepwise et AIC avec R

```
            Step: AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

           Df Sum of Sq  RSS      AIC
- pgg45     1  0.6590    45.526  -61.374
<none>                   44.867  -60.788
+ lcp       1  0.6623    44.204  -60.231
- age       1  1.2649    46.132  -60.092
- lbph      1  1.6465    46.513  -59.293
+ gleason   3  1.2918    43.575  -57.622
- lweight   1  3.5646    48.431  -55.373
- svi       1  4.2503    49.117  -54.009
- lcavol    1 25.4190    70.286  -19.248

            Step: AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi
```

Retour sur capital avec SAS

```
N = 40     Regression Models for Dependent Variable: RETCAP
R-square Adjusted C(p)   BIC    Variables in Model
In       R-square
1 0.1055 0.0819 78.3930 -163.3 WCFTCL
2 0.3406 0.3050 50.3232 -173.7 WCFTDT QUIKRAT
3 0.6154 0.5833 17.1815 -191.1 WCFTCL NFATAST CURRAT
4 0.7207 0.6888 5.7146 -199.20 WCFTDT LOGSALE NFATAST CURRAT
5 0.7317 0.6923 6.3047 -198.05 WCFTDT LOGSALE NFATAST QUIKRAT CURRAT
6 0.7483 0.7025 6.1878 -197.25 WCFTDT LOGSALE NFATAST INVTAST QUIKRAT CURRAT
7 0.7600 0.7075 6.6916 -195.77 WCFTDT LOGSALE LOGASST NFATAST FATTOT QUIKRAT CURRAT
8 0.7692 0.7097 7.5072 -193.87 WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT CURRAT
9 0.7760 0.7088 8.6415 -191.59 WCFTCL WCFTDT  LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT
                                CURRAT
10 0.7830 0.7082 9.7448 -189.2 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST PAYOUT
                                QUIKRAT CURRAT
11 0.7867 0.7029 11.277 -186.4 WCFTCL WCFTDT LOGSALE LOGASST NFATAST CAPINT FATTOT INVTAST
                                PAYOUT QUIKRAT CURRAT
12 0.7888 0.695  13.000 -183.5 WCFTCL WCFTDT GEARRAT LOGSALE LOGASST NFATAST CAPINT FATTOT
                                INVTAST PAYOUT QUIKRAT CURRAT
```

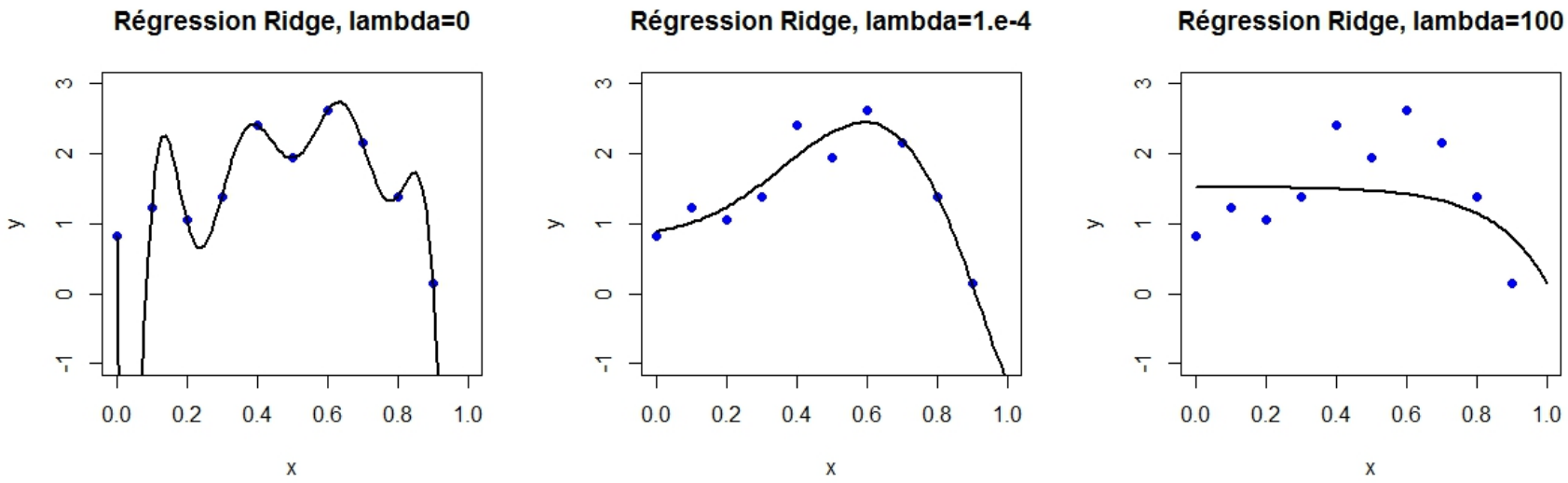
Définition de la régression *ridge*

$$\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}$$

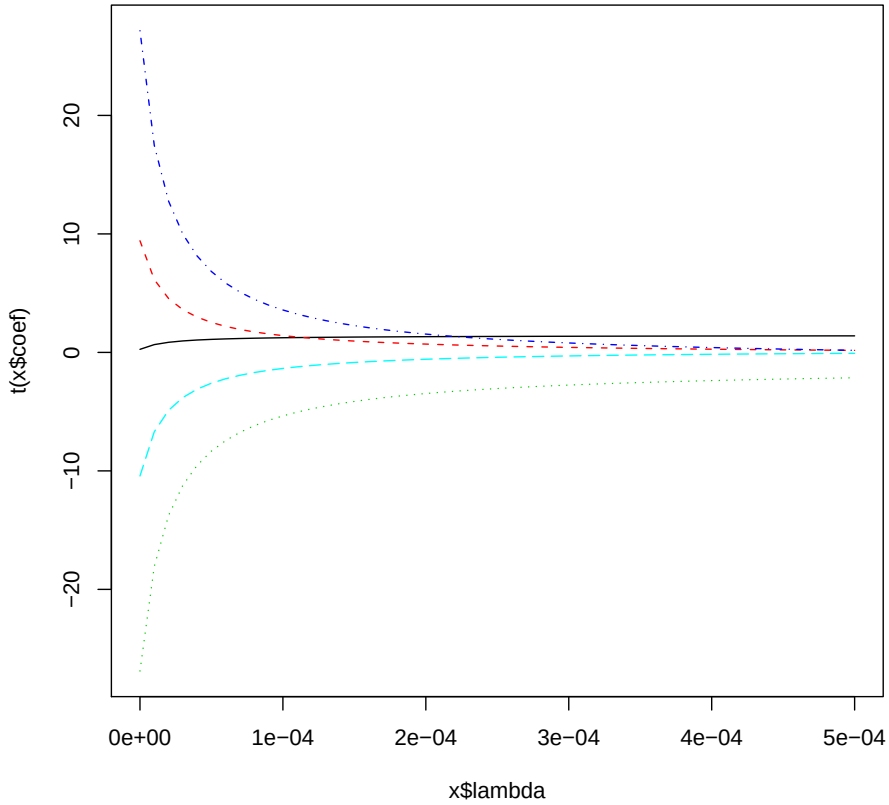
- $\boldsymbol{X}^0 = (1, 1, \ldots, 1)'$, et $\mathbf{X}$ la matrice $\widetilde{\mathbf{X}}$ privée de $\boldsymbol{X}^0$
- $\mathbf{Y}$ et bmX sont **centrés**
- $\mathbf{Y} = \mathbf{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$
- $\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$
- λ paramètre positif à optimiser
- $\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y}$

Propriétés de la régression ridge

1. $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p$ est inversible, mieux conditionnée
2. β_0 n'intervient pas : centrer $\mathbf{X}$
3. Dépend des unités : réduire $\mathbf{X}$
4. Forme équivalente :
   $\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \ ; \ \|\boldsymbol{\beta}\|^2 < c \right\}$
5. Chemin de régularisation
6. Optimisation de λ par V-fold validation croisée



Régression polynomiale : pénalisation ridge



Régression polynomiale : chemin de régularisation de la régression ridge

Régression LASSO ou *sparse* (1996)

- Ridge toujours calculable mais problème d'interprétation
- Objectif : associer pénalisation et sélection
- $\widehat{\beta}_{\text{Lasso}} =$
  $\text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (Y_i - \sum_{j=0}^p X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$
- $\widehat{\beta}_{\text{Lasso}} = \text{argmin}_{\beta, \|\beta\|_1 \leq t} (\|\mathbf{Y} - \mathbf{X}\beta\|^2)$
- λ est le paramètre de pénalisation
  - $\lambda = 0$: estimateur des moindres carrés.
  - λ tend vers l'infini, $\hat{\beta}_j = 0, j = 1, \ldots, p$.
- $\beta_j = \text{signe}(\widehat{\beta}_j)(|\widehat{\beta}_j| - \lambda)\mathbf{1}_{|\widehat{\beta}_j| \geq \lambda}$

Utilisation de la régression Lasso

- Utilisable si $p > n$
- Procédures de programmation linéaire ou algorithme LARS
- Nombre de variables influentes $q < n$
- Pas ou peu corrélées avec les autres

Régression *elastic net*

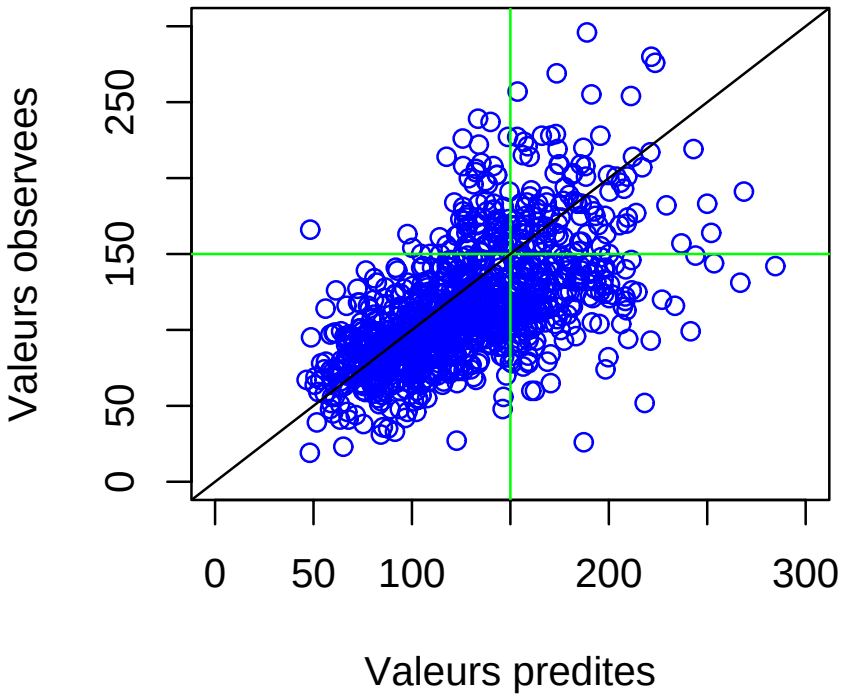
$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^{(1)} - \beta_2 X_i^{(2)} - \ldots - \beta_p X_i^{(p)})^2$$

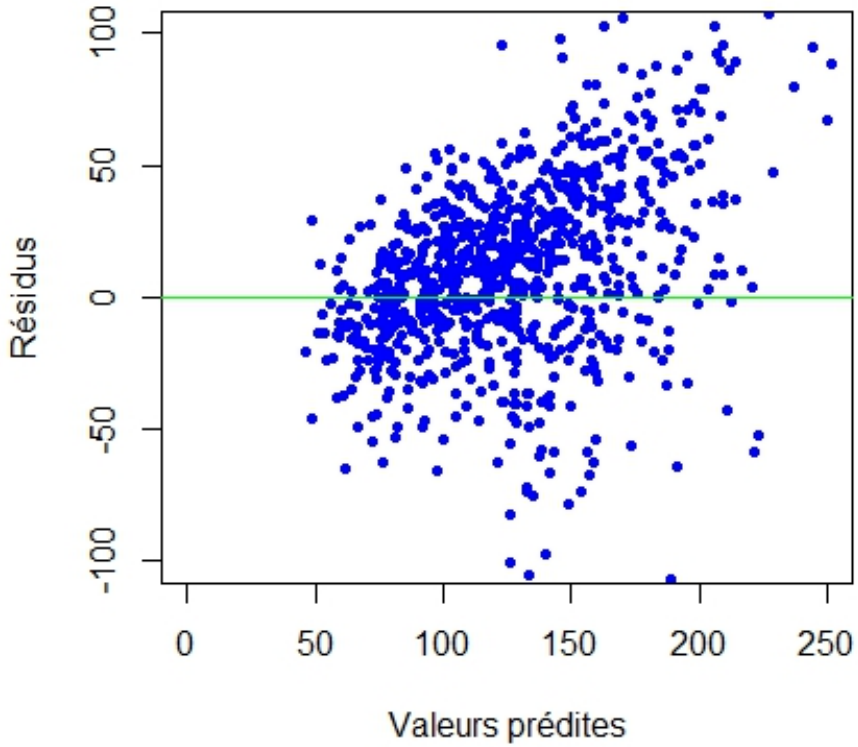
$$+\lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

- Pour $\alpha = 1$, régression Lasso
- Pour $\alpha = 0$, régression ridge

Concentration d'ozone

O3-o	Concentration d'ozone effectivement observée ou variable à prédire,
03-pr	prévision "mocage" qui sert de variable explicative ;
Tempe	Température prévue pour le lendemain,
vmodule	Force du vent prévue pour le lendemain,
lno	Logarithme de la concentration observée en monoxyde d'azote,
lno2	Logarithme de la concentration observée en dioxyde d'azote,
rmh20	Racine de la concentration en vapeur d'eau,
Jour	Variable à deux modalités pour distinguer les jours "ouvrables" (0) des jours "fériés-WE" (1).
Station	Une variable qualitative indique la station concernée : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache, et Plan de Cuques.





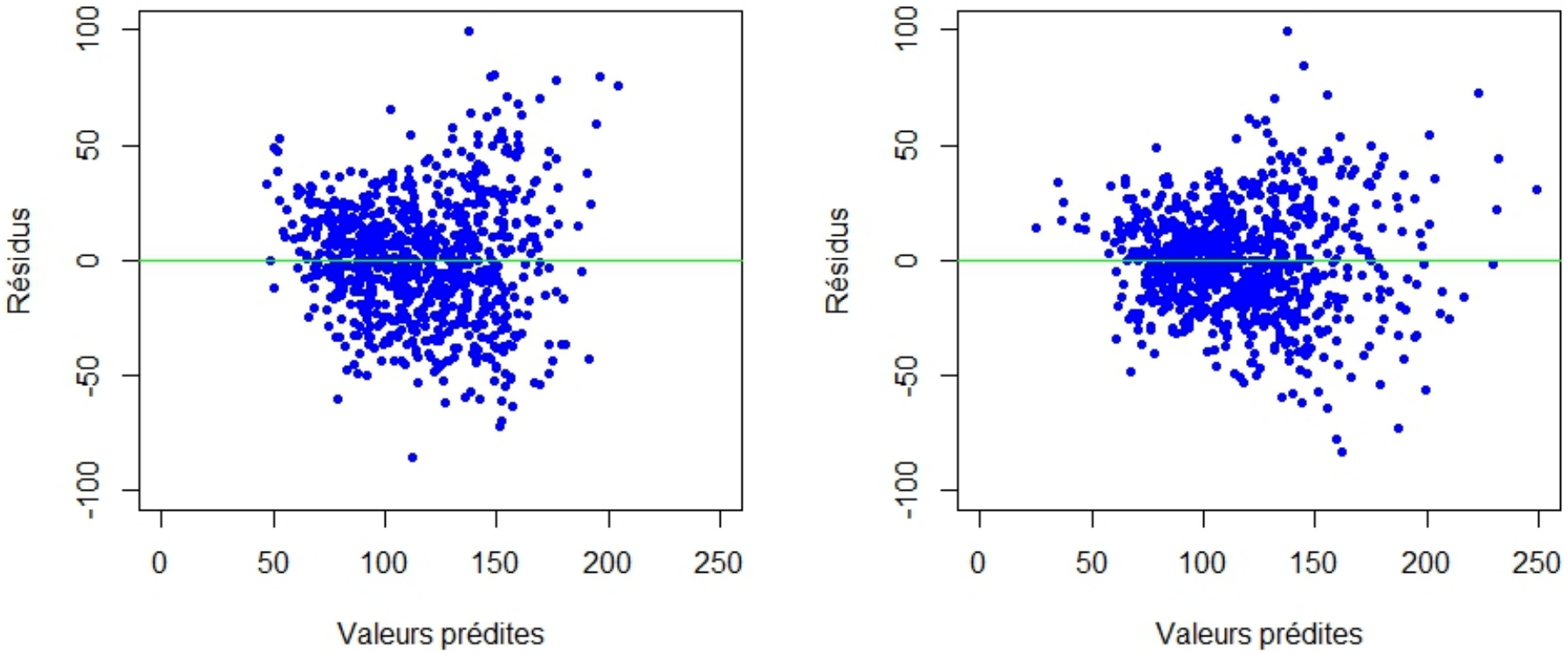
Ozone : Estimation et résidus de MOCAGE.

Modèle linéaire

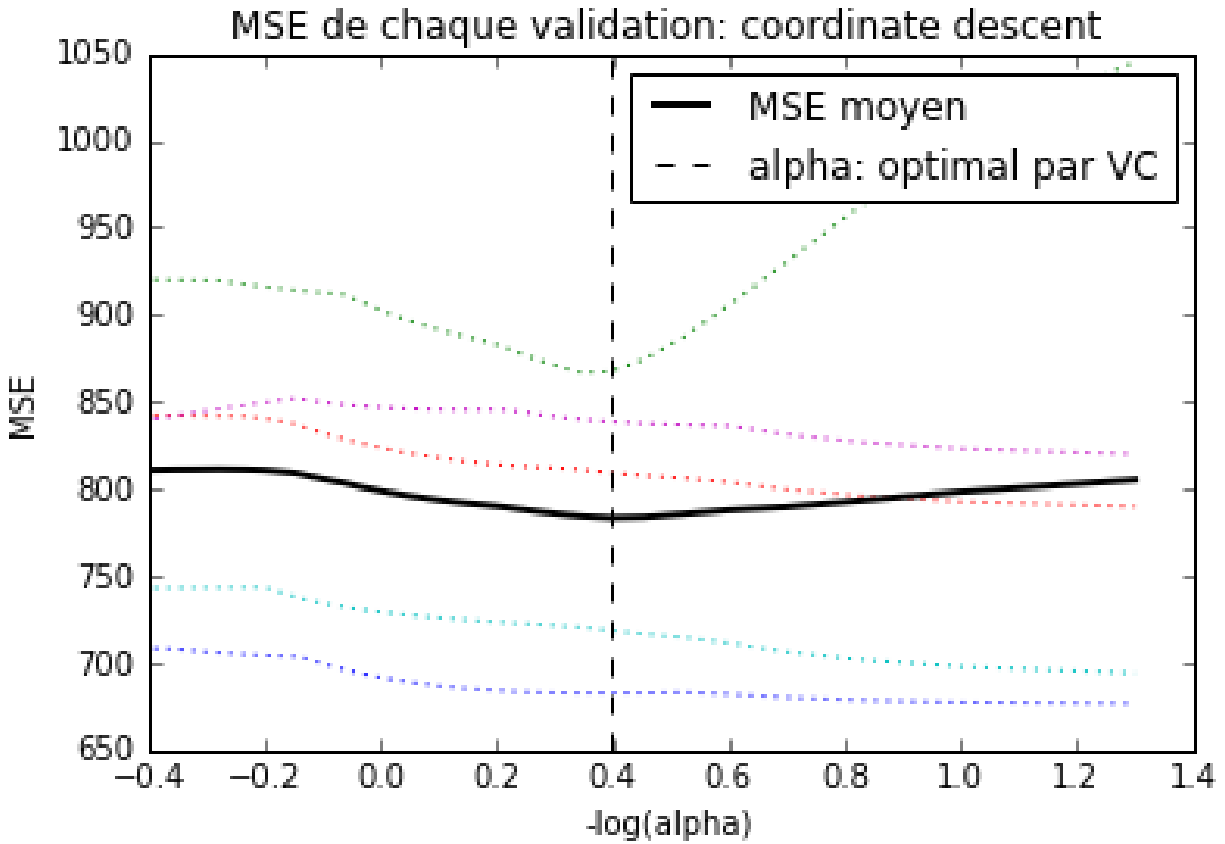
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.99738    7.87028  -0.635  0.52559
O3_pr         0.62039    0.05255  11.805  < 2e-16 ***
vmodule      -1.73179    0.35411  -4.891 1.17e-06 ***
lno2        -48.17248    6.19632  -7.774 1.83e-14 ***
lno          50.95171    5.98541   8.513  < 2e-16 ***
s_rmh2o     135.88280   50.69567   2.680  0.00747 **
jour1        -0.34561    1.85389  -0.186  0.85215
stationAls    9.06874    3.37517   2.687  0.00733 **
stationCad   14.31603    3.07893   4.650 3.76e-06 ***
stationPla   21.54765    3.74155   5.759 1.12e-08 ***
stationRam    6.86130    3.05338   2.247  0.02484 *
TEMPE         4.65120    0.23170  20.074  < 2e-16 ***

Residual standard error: 27.29 on 1028 degrees of freedom
Multiple R-Squared: 0.5616,     Adjusted R-squared: 0.5569
F-statistic: 119.7 on 11 and 1028 DF,  p-value: < 2.2e-16
```

Modèle quadratique

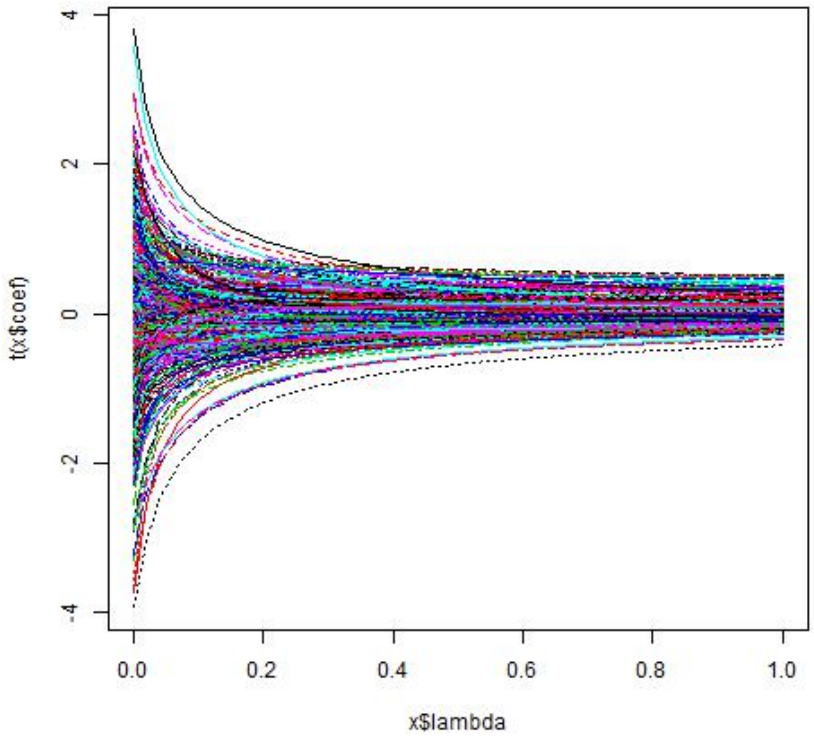
	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)	
NULL			1039	1745605			
O3_pr	1	611680	1038	1133925	969.9171	< 2.2e-16	***
station	4	39250	1034	1094674	15.5594	2.339e-12	***
vmodule	1	1151	1033	1093523	1.8252	0.1769957	
lno2	1	945	1032	1092578	1.4992	0.2210886	
s_rmh2o	1	24248	1031	1068330	38.4485	8.200e-10	***
TEMPE	1	248891	1030	819439	394.6568	< 2.2e-16	***
O3_pr:station	4	16911	1026	802528	6.7038	2.520e-05	***
O3_pr:vmodule	1	8554	1025	793974	13.5642	0.0002428	***
O3_pr:TEMPE	1	41129	1024	752845	65.2160	1.912e-15	***
station:vmodule	4	7693	1020	745152	3.0497	0.0163595	*
station:lno2	4	12780	1016	732372	5.0660	0.0004811	***
station:s_rmh2o	4	19865	1012	712508	7.8746	2.997e-06	***
station:TEMPE	4	27612	1008	684896	10.9458	1.086e-08	***
vmodule:lno2	1	1615	1007	683280	2.5616	0.1098033	
vmodule:s_rmh2o	1	2407	1006	680873	3.8163	0.0510351	.
lno2:TEMPE	1	4717	1005	676156	7.4794	0.0063507	**
s_rmh2o:TEMPE	1	42982	1004	633175	68.1543	4.725e-16	***

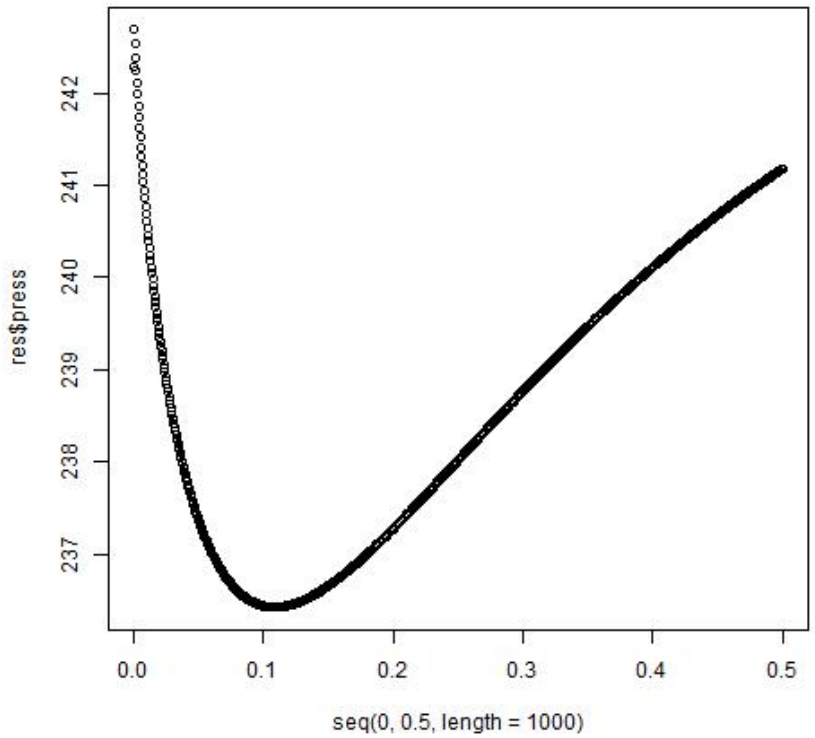


Ozone : Résidus des modèles linéaire et quadratique

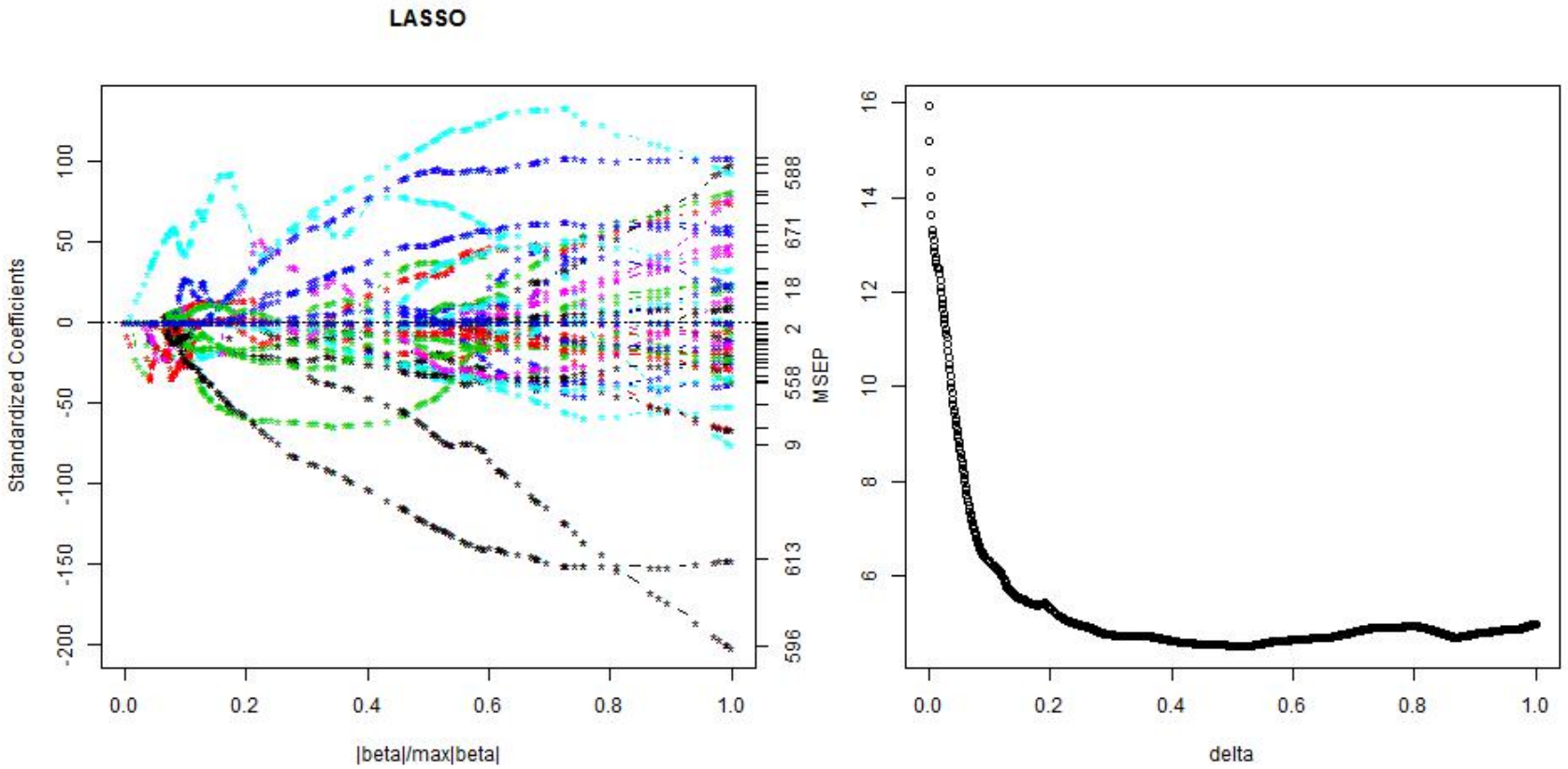


Ozone : optimisation de régularisation lasso par validation croisée.





Cookies : Régression ridge de données NIR.



Cookies : Régression Lasso de données NIR.