Sélection de modèle en régression linéaire

Résumé

*Le modèle linéaire gaussien ou régression multiple est considéré avec pour objectif la prévision d'une variable quantitative par un ensemble de variables quantitatives ou un mélande de quantitatives et qualitatives (analyse de covariance). Recherche d'un modèle parcimonieux assurant un bon équilibre entre la qualité de l'ajustement et la variance des paramètres afin de minimiser le risque empirique. Algorithmes (*backward, forward, stepwise...*) de sélection de modèle par sélection de variables et minimisation de critères pénalisés (C_p, AIC, BIC). Algorithmes de sélection de modèle par pénalisation* ridge, *Lasso,* elastic net.

Retour à l'introduction.

Tous les tutoriels sont disponibles sur le dépôt :
`github.com/wikistat`

1 Régression multiple

Les modèles classiques de régession (linéaire, logistique) sont anciens et moins l'occasion de battage médiatique que ceux récents issus de l'apprentissage machine. Néanmoins, compte tenu de leur robustesse, de leur stabilité face à des fluctuations des échantillons, de leur capacité à passer à l'échelle des données massives... tout ceci fait qu'ils restent toujours très utilisés en production notamment lorsque la fonction à modéliser est bien linéaire et qu'il serait contre productif de chercher plus compliqué.

1.1 Modèle

Une variable quantitative $\mathbf{Y}$ dite *à expliquer* (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p$ dites *explicatives* (ou encore de contrôle, endogènes, indépendantes, régresseurs, prédicteurs).

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n ($n > p+1$) de $\mathbb{R}^{(p+1)}$:

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i) \quad i = 1, \ldots, n.$$

L'écriture du *modèle linéaire* dans cette situation conduit à supposer que l'espérance de $\mathbf{Y}$ appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^p\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de 1s. C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

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

avec les hypothèses suivantes :

1. Les ε_i sont des termes d'erreur indépendants et identiquement distribués ; $E(\varepsilon_i) = 0, Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.
2. Les termes $\boldsymbol{X}^j$ sont supposés déterministes (facteurs contrôlés) **ou bien** l'erreur $\boldsymbol{\varepsilon}$ est indépendante de la distribution conjointe de $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p$. On écrit dans ce dernier cas que :
   $E(\mathbf{Y}|\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p) = \beta_0 + \beta_1\boldsymbol{X}^1 + \beta_2\boldsymbol{X}^2 + \cdots + \beta_p\boldsymbol{X}^p$ et $Var(\mathbf{Y}|\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p) = \sigma^2$.
3. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.
4. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur $\boldsymbol{\varepsilon}$ ($\mathcal{N}(0, \sigma^2\mathbf{I})$). Les ε_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice $\mathbf{X}(n \times (p+1))$ de terme général X_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($X_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général Y_i. En notant les vecteurs $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_p]'$ et $\boldsymbol{\beta} = [\beta_0\beta_1 \cdots \beta_p]'$, le modèle s'écrit matriciellement :

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

1.2 Estimation

Conditionnellement à la connaissance des valeurs des $\boldsymbol{X}^j$, les paramètres inconnus du modèle : le vecteur $\boldsymbol{\beta}$ et σ^2 (paramètre de nuisance), sont estimés par minimisation des carrés des écarts (M.C.) ou encore, en supposant (4.), par maximisation de la vraisemblance (M.V.). Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

1.3 Estimation par moindres carrés

L'expression à minimiser sur $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^1 - \cdots - \beta_p X_i^p)^2 &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

Par dérivation matricielle de la dernière équation on obtient les *équations normales* :

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

dont la solution correspond bien à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. Si cette hypothèse n'est pas vérifiée, il suffit en principe de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Une approche de réduction de dimension (régression *ridge*, Lasso, PLS ...) est à mettre en œuvre.

Alors, l'estimation des paramètres β_j est donnée par :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{Y}$ ont pour expression :

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (*hat matrix*). Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace Vect($\mathbf{X}$) engendré par les vecteurs colonnes de $\mathbf{X}$.

On note

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

le vecteur des résidus ; c'est la projection de $\mathbf{Y}$ sur le sous-espace orthogonal de Vect($\mathbf{X}$) dans $\mathbb{R}^n$.

1.4 Propriétés

Les estimateurs des M.C. $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ sont des estimateurs sans biais : $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, et, parmi les estimateurs sans biais fonctions linéaires des Y_i, ils sont de variance minimum (théorème de Gauss-Markov) ; ils sont donc BLUE : *best linear unbiaised estimators*. Sous hypothèse de normalité, les estimateurs du M.V. sont uniformément meilleurs (efficaces) et coïncident avec ceux des moindres carrés.

On montre que la matrice de covariance des estimateurs se met sous la forme

$$E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

celle des prédicteurs est

$$E[(\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})(\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})'] = \sigma^2\mathbf{H}$$

et celle des estimateurs des résidus est

$$E[\mathbf{e}\mathbf{e}'] = \sigma^2(\mathbf{I} - \mathbf{H})$$

tandis qu'un estimateur sans biais de σ^2 est fourni par :

$$\widehat{\sigma}^2 = \frac{\|\mathbf{e}\|^2}{n-p-1} = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{n-p-1} = \frac{\text{SSE}}{n-p-1}.$$

Ainsi, les termes $\widehat{\sigma}^2 h_i^i$ sont des estimations des variances des prédicteurs $\widehat{Y}_i$.

Conséquence importante : si la matrice $\mathbf{X}'\mathbf{X}$ est mal conditionnée (déterminant proche de 0), son inversion fait apparaître des termes très élevés sur la diagonale et conduit donc à des variances très importantes des estimations des paramètres.

1.5 Sommes des carrés

SSE est la somme des carrés des résidus (*sum of squared errors*),

$$\text{SSE} = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|^2 = \|\mathbf{e}\|^2.$$

On définit également la somme totale des carrés (*total sum of squares*) par

$$\text{SST} = \left\|\mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1}\right\|^2 = \mathbf{Y}'\mathbf{Y} - n\overline{\mathbf{Y}}^2$$

et la somme des carrés de la régression (*regression sum of squares*) par

$$\text{SSR} = \left\|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\mathbf{1}\right\|^2 = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} - n\overline{\mathbf{Y}}^2 = \mathbf{Y}'\mathbf{H}\mathbf{Y} - n\overline{\mathbf{Y}}^2 = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\overline{\mathbf{Y}}^2.$$

On vérifie alors : SST = SSR + SSE.

1.6 Coefficient de détermination

On appelle *coefficient de détermination* le rapport

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

qui est donc la part de variation de $\mathbf{Y}$ expliquée par le modèle de régression. Géométriquement, c'est un rapport de carrés de longueur de deux vecteurs. C'est donc le cosinus carré de l'angle entre ces vecteurs : $\mathbf{Y}$ et sa projection $\widehat{\mathbf{Y}}$ sur Vect($\mathbf{X}$).

La quantité R est appelée *coefficient de corrélation multiple* entre $\mathbf{Y}$ et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{Y}$ et sa prévision $\widehat{\mathbf{Y}}$.

Par construction, le coefficient de détermination croît avec le nombre p de vartiables.

1.7 Inférence dans le cas gaussien

En principe, l'hypothèse optionnelle (4.) de normalité des erreurs est nécessaire pour cette section. En pratique, des résultats asymptotiques, donc valides pour de grands échantillons, ainsi que des études de simulation, montrent que cette hypothèse n'est pas celle dont la violation est la plus pénalisante pour la fiabilité des modèles.

Inférence sur les coefficients

Pour chaque coefficient β_j on note $\widehat{\sigma}_j^2$ l'estimateur de la variance de $\widehat{\beta}_j$ obtenu en prenant j-ème terme diagonal de la matrice $\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. On montre que la statistique

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_j}$$

suit une loi de Student à $(n-p-1)$ degrés de liberté. Cette statistique est donc utilisée pour tester une hypothèse $H_0 : \beta_j = a$ ou pour construire un intervalle de confiance de niveau $100(1-\alpha)\%$:

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

Attention, cette statistique concerne un coefficient et ne permet pas d'inférer conjointement sur d'autres coefficients car leurs estimateurs sont corrélés. De plus elle dépend des absences ou présences des autres variables $\boldsymbol{X}^k$ dans le modèle. Par exemple, dans le cas particulier de deux variables $\boldsymbol{X}^1$ et $\boldsymbol{X}^2$ très corrélées, chaque variable, en l'absence de l'autre, peut apparaître avec un coefficient significativement différent de 0 ; mais, si les deux sont présentes dans le modèle, l'une peut apparaître avec un coefficient insignifiant.

De façon plus générale, si $\mathbf{c}$ désigne un vecteur non nul de $(p+1)$ constantes réelles, il est possible de tester la valeur d'une combinaison linéaire $\mathbf{c}'\beta$ des paramètres en considérant l'hypothèse nulle $H_0 : \mathbf{c}'\beta = a$; a connu. Sous H_0, la statistique

$$\frac{\mathbf{c}'\widehat{\boldsymbol{\beta}} - a}{(\widehat{\sigma}^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})^{1/2}}$$

suit une loi de Student à $(n-p-1)$ degrés de liberté.

Inférence sur le modèle

Le modèle peut être testé globalement. Sous l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, la statistique

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

suit une loi de Fisher avec p et $(n-p-1)$ degrés de liberté. Les résultats sont habituellement présentés dans un tableau *d'analyse de la variance"* sous la forme suivante :

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

Inférence sur un modèle réduit

Le test précédent amène à rejeter H_0 dès que l'une des variables $\boldsymbol{X}^j$ est liée à $\mathbf{Y}$. Il est donc d'un intérêt limité. Il est souvent plus utile de tester un modèle réduit c'est-à-dire dans lequel certains coefficients, à l'exception de la constante, sont nuls contre le modèle complet avec toutes les variables. En ayant éventuellement réordonné les variables, on considère l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0, q < p$.

Notons respectivement SSR_q, SSE_q, R_q^2 les sommes de carrés et le coefficient de détermination du modèle réduit à $(p-q)$ variables. Sous H_0, la statistique

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

suit une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

Dans le cas particulier où $q = 1$ ($\beta_j = 0$), la F-statistique est alors le carré de la t-statistique de l'inférence sur un paramètre et conduit donc au même test.

1.8 Prévision

Connaissant les valeurs des variables $\boldsymbol{X}^j$ pour une nouvelle observation : $\mathbf{x}_0' = [x_0^1, x_0^2, \ldots, x_0^p]$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y}_0$ de $\mathbf{Y}$ ou $E(\mathbf{Y})$ est donnée par :

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0^1 + \cdots + \widehat{\beta}_p x_0^p.$$

Les intervalles de confiance des prévisions de $\mathbf{Y}$ et $E(\mathbf{Y})$, pour *une* valeur $\mathbf{x}_0 \in \mathbb{R}^p$ et en posant $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$, sont respectivement

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

Les variances de ces prévisions, comme celles des estimations des paramètres, dépendent directement du conditionnement de la matrice $\mathbf{X}'\mathbf{X}$.

1.9 Diagnostics

La validité d'un modèle de régression multiple et donc la fiabilité des prévisions, dépendent de la bonne vérification des hypothèses :

- homoscédasticité : variance σ^2 des résidus constante,
- linéarité du modèle : paramètres β_j constant,
- absence de points influents par la distance de Cook :

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

- éventuellement la normalité des résidus,
- le conditionnement de la matrice $\mathbf{X}'\mathbf{X}$.

Tracer le graphe des résidus standardisés en fonction des valeurs ajustés montre leur plus ou moins bonne répartition autour de l'axe $y = 0$. La forme de ce nuage est susceptible de dénoncer une absence de linéarité ou une hétéroscédasticité.

Le conditionnement de la matrice $\mathbf{X}'\mathbf{X}$ est indiqué par le rapport $\kappa = \lambda_1/\lambda_p$ où $\lambda_1, \ldots, \lambda_p$ sont les valeurs propres de la matrice des corrélations $\mathbf{R}$ rangées par ordre décroissant. Ainsi, des problèmes de variances excessives voire même de précision numérique apparaissent dès que les dernières valeurs propres sont relativement trop petites.

1.10 Exemple

Les données sont extraites de Jobson (1991)[3] et décrivent les résultats comptables de 40 entreprises du Royaume Uni.

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

La procédure SAS/REG fournit les résultats classiques de la régression multiple.

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

```
Parameter Estimates
                  Parameter       Standard     T for H0:                      Variance
Variable  DF       Estimate          Error    Parameter=0 Prob>|T|   Tolerance    Inflation
                        (1)            (2)            (3)       (4)           (5)          (6)
INTERCEP   1       0.188072    0.13391661         1.404     0.1716   .             0.00000000
WCFTCL     1       0.215130    0.19788455         1.087     0.2866    0.03734409  26.77799793
WCFTDT     1       0.305557    0.29736579         1.028     0.3133    0.02187972  45.70441500
GEARRAT    1      -0.040436    0.07677092        -0.527     0.6027    0.45778579   2.18442778
LOGSALE    1       0.118440    0.03611612         3.279     0.0029    0.10629382   9.40788501
LOGASST    1      -0.076960    0.04517414        -1.704     0.0999    0.21200778   4.71680805
...
```

(1)	estimations des paramètres $(\widehat{\beta}_j)$
(2)	écarts-types de ces estimations $\widehat{\sigma}_j$
(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=$1/(1 - R^2_{(j)})$

Ces résultats soulignent les problèmes de colinéarités. De grands VIF (facteurs d'inflation de la variance) sont associés à de grands écart-types des estimations des paramètres. D'autre part les nombreux tests de Student non significatifs montrent que trop de variables sont présentes dans le modèle. Cette idée est renforcée par le calcul de l'indice de conditionnement : 8.76623/0.00125.

2 Analyse de covariance (AnCoVa

L'*analyse de covariance* se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Les cas les plus complexes associent plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynomiale. Le principe général, dans un but explicatif ou décisionnel, est toujours d'estimer des modèles *intra-groupes* et de faire apparaître (tester) des effets différentiels *inter-groupes* des paramètres des régressions. Ainsi, dans le cas plus simple où seulement une variable parmi les explicatives est quantitative, des tests interrogent l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

Ce type de modèle permet également, avec un objectif prédictif, de s'intéresser à la modélisation d'une variable quantitative par un ensemble de variables explicatives à la fois quantitatives et qualitatives.

La possible prise en compte d'*interactions* entre les variables complique la procédure de sélection de variables.

2.1 Modèle

Le modèle est explicité dans le cas élémentaire où une variable quantitative $\mathbf{Y}$ est expliquée par une variable qualitative $\boldsymbol{T}$ à J niveaux et une variable quantitative, appelée encore covariable, $\boldsymbol{X}$. 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 $\mathbf{Y}$; $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon.

En pratique, avant de lancer une procédure de modélisation et tests, une démarche exploratoire s'appuyant sur une représentation en couleur (une par modalité j de T) du nuage de points croisant $\mathbf{Y}$ et $\boldsymbol{X}$ et associant les droites de régression permet de se faire une idée sur les effets respectifs des variables : parallélisme des droites, étirement, imbrication des sous-nuages.

On suppose que les moyennes conditionnelles $E[\mathbf{Y}|\boldsymbol{T}]$, c'est-à-dire calculées à l'intérieur de chaque cellule, sont dans le sous-espace vectoriel engendré par les variables explicatives quantitatives, ici $\boldsymbol{X}$. Ceci s'écrit :

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \ ; \quad j = 1, \ldots, J; \quad i = 1, \cdots, n_j$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0,\sigma^2)$ pour la construction des tests.

Notons $\mathbf{Y}$ le vecteur des observations $[Y_{ij}|i=1,n_j;j=1,J]'$ mis en colonne, $\mathbf{x}$ le vecteur $[X_{ij}|i=1,n_j;j=1,J]'$, $\boldsymbol{\varepsilon}=[\varepsilon_{ij}|i=1,n_j;j=1,J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. On note encore $\mathbf{x}.\mathbf{1}_j$ le produit terme à terme des deux vecteurs, c'est-à-dire le vecteur contenant les observations de $\mathbf{x}$ sur les individus prenant le niveau j de $\boldsymbol{T}$ et des zéros ailleurs.

La résolution simultanée des J modèles de régression est simplement obtenue en considérant globalement le modèle :

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

dans lequel $\mathbf{X}$ est la matrice $n\times 2J$ constituée des blocs $[\mathbf{1}_j|\mathbf{X}.\mathbf{1}_j]\;;j=1,\ldots,J$. L'estimation de ce modèle global conduit, par bloc, à estimer les modèles de régression dans chacune des cellules.

Comme pour l'analyse de variance (AnOVa), les logiciels opèrent une reparamétrisation faisant apparaître des effets différentiels par rapport au dernier niveau ou par rapport à un effet moyen, afin d'obtenir directement les bonnes hypothèses dans les tests. Ainsi, dans le premier cas, on considère la matrice de même rang (sans la Jème indicatrice)

$$\mathbf{X}=[\mathbf{1}|\mathbf{X}|\mathbf{1}_1|\cdots|\mathbf{1}_{J-1}|\mathbf{x}.\mathbf{1}_1|\cdots|\mathbf{x}.\mathbf{1}_{J-1}]$$

associée aux modèles :

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

2.2 Tests

Différentes hypothèses sont alors testées en comparant le modèle complet

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

à chacun des modèles réduits :

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

par un test de Fisher. Ceci revient à considérer les hypothèses suivantes :

- 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 $\mathbf{Y}$.

Commencer par évaluer *i*; si le test n'est pas significatif, regarder *ii* qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si *i* n'est pas significatif, s'intéresser à *iii* pour juger de l'effet du facteur $\boldsymbol{T}$.

3 Choix de modèle par sélection de variables

3.1 Introduction

De façon schématique, la pratique de la modélisation statistique vise trois objectifs éventuellement complémentaires.

Descriptif : rechercher de façon exploratoire les liaisons entre $\mathbf{Y}$ et d'autres variables, potentiellement explicatives, $\boldsymbol{X}^j$ qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. À cette stratégie, à laquelle peuvent contribuer des Analyses en Composantes Principales, correspond des algorithmes de recherche (pas à pas) moins performants mais économiques en temps de calcul si p est grand.

Attention, si n est petit, et la recherche suffisamment longue avec beaucoup de variables explicatives, il sera toujours possible de trouver un modèle expliquant y ; c'est l'effet *data mining* dans les modèles économétriques appelé maintenant *data snooping*.

Explicatif : Le deuxième objectif est sous-tendu par une connaissance *a priori* du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels permettent de construire *le* bon test conduisant à la prise de décision recherchée. Utilisées hors de ce contexte, les statistiques de test n'ont qu'une valeur indicative au même titre que d'autres critères plus empiriques.

Prédictif : Dans le troisième cas, l'accent est mis sur la qualité des prévisions. C'est la situation rencontrée en *apprentissage*. Ceci conduit à rechercher des modèles *parcimonieux* c'est-à-dire avec un nombre volontairement restreint de variables explicatives pour réduire la variance. Le modèle ainsi obtenu peut favoriser des estimateurs biaisés au profit d'une variance plus faible même si le théorème de Gauss-Markov indique que, parmi les estimateurs sans biais, celui des moindres carrés est de variance minimum. Un bon modèle n'est donc plus celui qui explique le mieux les données au sens d'un R^2 maximum mais celui conduisant aux prévisions les plus fiables.

Ceci est illustré ceci par un exemple simple (mais pédagogique) en régression polynomiale : Les figures 1 et 2) représentent un jeu de données simulées : $Y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$ et $x_i \in [0, 1]$ sur lesquelles des polynômes de degrés croissants sont ajustés. L'ajustement du modèle mesuré par le R^2 croît logiquement avec le nombre de paramètres et atteint la valeur 1 lorque le polynôme interpole les observations.

Le R^2 ne peut-être un bon critère de sélection de modèles ; il ne peut servir qu'à comparer des modèles de même dimension car sinon conduit à sélectionner le modèle le plus complexe, c'est-à-dire celui correspond au plus grand espace de projection, et conduit donc au sur-ajustement.

Il y a principalement deux façons de biaiser un modèle linéaire dans le but de restreindre la variance :

- en réduisant le nombre de variables explicatives et donc en simplifiant le modèle (sélection ou pénalisation Lasso l_1),
- en contraignant les paramètres du modèle, en les rétrécissant (*schrinkage*), par une régression *ridge* qui opère une régularisation par pénalisation l_2.

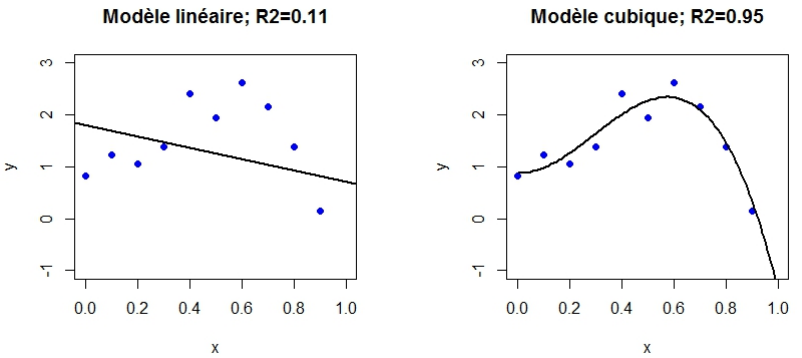

FIGURE 1 – Régression polynomiale : ajustement par, à gauche, $y = \beta_0 + \beta_1 x + \epsilon$, et à droite, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$

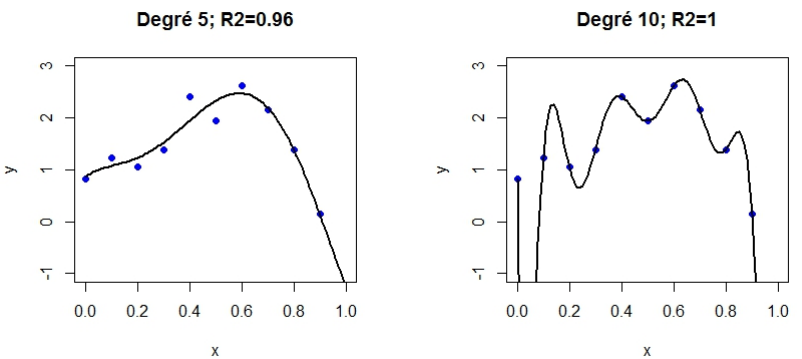

FIGURE 2 – Régression polynomiale : ajustement par, à gauche : $y = \beta_0 + \beta_1 x + \ldots + \beta_5 x^5 + \epsilon$, et à droite, $y = \beta_0 + \beta_1 x + \ldots + \beta_{10} x^{10} + \epsilon$.

3.2 Critères de sélection de variables

De nombreux critères de choix de modèle sont présentés dans la littérature sur la régression linéaire multiple.

Une tradition ancienne, encore présente dans certains livres ou fonction de logiciels, propose d'utiliser la statistique du test de Fisher de comparaiso d'un modèle avec un sous-modèle comme critère de sélection de variable *Attention*, la significativité de la présence d'une variable basée sur la p-valeu du test de nullité de son coefficient n'est du tout une indication sur l'importanc de cette variable pour la qualité de la prévision. Explicatif ou prédictif sont de objectifs différents de la modélisation ; ne pas les confondre. D'autre part, pou éviter le caractère croissant du coefficient de détermination R^2 en fonctio du nombre de variables, une version pénalisée a été proposée : le R^2 ajust mais qui conduit très généralement à des modèles trop complexes. Ces deu approches : statistique du test de Fisher et R^2 ajusté sont à oublier.

D'autres critères sont eux basés sur une qualité de prévision. Le C_p de Mallows, le critère d'information d'Akaïke (AIC), celui bayésien de Sawa (BIC)... Ils sont équivalents, également avec le R^2, lorsque le nombre de variables à sélectionner, ou complexité du modèle, est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de complexitée différentes. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression *pénalisée* de la fonction de vraisemblance afin de favoriser des modèles parcimonieux. En pratique, les plus utilisés ou ceux généralement fournis par les logiciels sont les suivants.

C_p de Mallows

L'indicateur proposé par Mallows (1973)[5] est une estimation de l'erreur quadratique moyenne de prévision qui s'écrit aussi comme la somme d'une variance et du carré d'un biais. L'erreur quadratique moyenne de prévision s'écrit ainsi :

$$\text{MSE}(\widehat{Y}_i) = \text{Var}(\widehat{Y}_i) + [\text{Biais}(\widehat{Y}_i)]^2$$

puis après sommation et réduction :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\text{MSE}(\widehat{Y}_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Var}(\widehat{Y}_i) + \frac{1}{\sigma^2}\sum_{i=1}^{n}[\text{Biais}(\widehat{Y}_i)]^2.$$

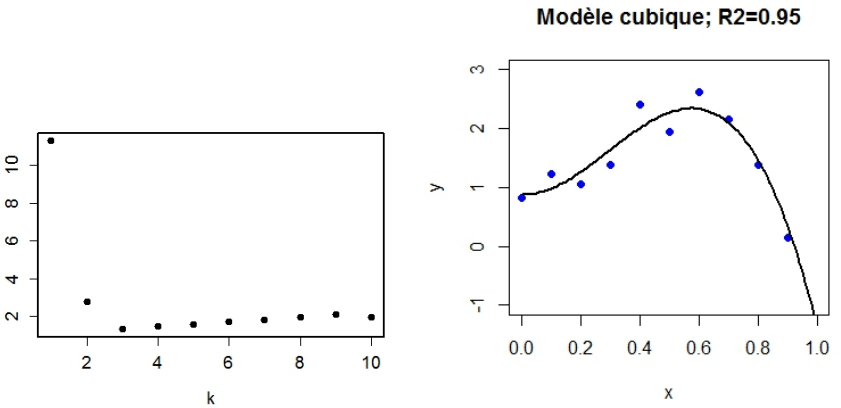

FIGURE 3 – C_p de Mallows en fonction du degré du polynôme et modèle sélectionné de degré 3.

En supposant que les *estimations du modèle complet sont sans biais* et en utilisant des estimateurs de $\text{Var}(\widehat{Y}_i)$ et σ^2, l'expression de l'erreur quadratique moyenne totale standardisée (ou réduite) pour un modèle à j variables explicatives s'écrit :

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

et définit la valeur du C_p de Mallows pour les q variables considérées. Il est alors d'usage de rechercher un modèle qui minimise le C_p généralement proche de $(q+1)$. Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise.

La figure 3 montre le comportement du C_p dans l'exemple de la régression polynomial. Ce critère décroît avec le biais jusqu'à un choix optimal de dimension 3 avant de ré-augmenter avec la variance.

AIC, BIC et PRESS

Dans le cas du modèle linéaire, et si la variance des observations est supposée connue, le critère AIC (*Akaïke's Information criterion*) est équivalent au

critère C_p de Mallows.

Le PRESS de Allen (1974)[**?**] est l'introduction historique de la validation croisée ou *leave one out (loo)*. On désigne par $\widehat{Y}_{(i)}$ la prévision de Y_i calculée sans tenir compte de la $i$ème observation $(Y_i, X_i^1, \ldots, X_i^p)$, la somme des erreurs quadratiques de prévision (PRESS) est définie par

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{f}^{(-i)}(\boldsymbol{x}_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1-h_{ii}}\right]^2.$$

et permet de comparer les capacités prédictives de deux modèles.

La vignette sur Qualité de prévision et risque donne plus de détails sur ces derniers critères.

3.3 Algorithmes de sélection de variables

Dans le cas général et évidemment le plus courant en pratique, les variables ne sont pas pré-ordonnées par importance. Lorsque p est grand, il n'est pas raisonnable de penser explorer les 2^p modèles possibles afin de sélectionner le meilleur au sens de l'un des critères ci-dessus. Différentes stratégies sont donc proposées qui doivent être choisies en fonction de l'objectif recherché, de la valeur de p et des moyens de calcul disponibles. Deux types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire c'est-à-dire par nombre croissant de modèles considérés parmi les 2^p et donc par capacité croissante d'optimalité.

Pas à pas

Stratégie correspondant à la fonction `StepAIC` de R. Comme écrit ci-dessus, oublier les sélections basées sur la statistique de Fisher.

Sélection (`forward`) À chaque pas, une variable est ajoutée au modèle. C'est celle qui permet de réduire au mieux le critère AIC du modèle obtenu. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque AIC ne décroît plus.

Élimination (`backward`) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable dont l'élimination conduit à l'AIC le plus faible est supprimée. La procédure s'arrête lorsque AIC ne décroît plus.

Mixte (`stepwise`) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Global

L'algorithme de Furnival et Wilson (1974)[2] (librairie `leaps` de R) est utilisé pour comparer tous les modèles possibles en cherchant à optimiser l'un des critères : C_p, AIC, BIC choisi par l'utilisateur. Par souci d'économie, cet algorithme évite de considérer des modèles de certaines sous-branches de l'arborescence dont on peut savoir *a priori* qu'ils ne sont pas compétitifs. Cet algorithme affichent le ou les meilleurs modèles de chaque niveau q. Rappel : à q fixé tous les critères sont équivalents mais les choix de q optimal peut différer d'un critère à l'autre. Il n'est pas raisonnable de considérer plus d'une quinzaine de variables avec cet algorithme.

3.4 Sélection en analyse de covariance

Un modèle d'analyse de covariance pose des problèmes spécifiques de sélection notamment par la prise en compte possible d'interactions entre variables dans la définition du modèle. La recherche d'un modèle efficace, donc parcimonieux, peut conduire à négliger des interactions ou effets principaux lorsqu'une faible amélioration du R^2 le justifie et même si le test correspondant apparaît comme significatif. L'utilisation du C_p est théoriquement possible mais en général ce critère n'est pas calculé car d'utilisation délicate. En effet, il nécessite la considération d'un modèle de référence sans biais ou tout du moins d'un modèle de faible biais pour obtenir une estimation raisonnable de la variance de l'erreur. En régression multiple (toutes les variables explicatives quantitatives), le modèle complet est considéré comme étant celui de faible biais mais en analyse de covariance quels niveaux de complexité des interactions faut-il considérer pour construire le modèle complet jugé de faible biais ? Il est alors plus simple et plus efficace d'utiliser le critère AIC, choix par défaut dans plusieurs logiciels comme R.

L'algorithme de recherche descendant est le plus couramment utilisé avec la contrainte suivante :

un effet principal n'est supprimé qu'à la condition qu'il n'apparaisse plus

dans une interaction.

Voici, à titre d'exemple, une étape intermédiaire d'une sélection de variables pas à pas *stepwize* avec l'option *b*oth de la fonction `StepAIC` de R. A chaque étape, le critère AIC est évalué par suppression ou rajout de chacune des variables. L'option minimisant le critère AIC est retenue avant de passer à l'étape suivante. Le modèle ne comprend pas d'interactions.

```
            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
```

En effet, supprimer un effet principal qualitatif alors que la variable est présente dans une interaction ne change en rien le modèle car l'espace engendré par l'ensemble des indicatrices sélectionnées reste le même ; la matrice $\mathbf{X}$ est construite sous contrainte de rang et retirer une colonne (effet principal) fait automatiquement entrer une indicatrice d'interaction supplémentaire. Le modèle est inchangé mais l'interprétation plus compliquée car le modèle ne se décompose plus en un effet principal puis ses interactions.

3.5 Exemple de sélection

Tous les modèles (parmi les plus intéressants selon l'algorithme de Furnival et Wilson) sont considérés. Seul le meilleur pour chaque niveau, c'est-à-dire pour chaque valeur p du nombre de variables explicatives sont donnés. Il est alors facile de choisir celui minimisant l'un des critères globaux (C_p ou BIC). Cet exemple calculé avec SAS est choisi pour comparer différents critères.

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp2 ;
model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       / selection=rsquare cp rsquare bic best=1;
run;
```

```
   N = 40     Regression Models for Dependent Variable: RETCAP
  R-sq. Adjust. C(p)   BIC     Variables in Model
In       R-sq
 1 0.105 0.081 78.393 -163.2 WCFTCL
 2 0.340 0.305 50.323 -173.7 WCFTDT QUIKRAT
 3 0.615 0.583 17.181 -191.1 WCFTCL NFATAST CURRAT
 4 0.720 0.688  5.714 -199.2 WCFTDT LOGSALE NFATAST CURRAT
 5 0.731 0.692  6.304 -198.0 WCFTDT LOGSALE NFATAST QUIKRAT CURRAT
 6 0.748 0.702  6.187 -197.2 WCFTDT LOGSALE NFATAST INVTAST QUIKRAT CURRAT
 7 0.760 0.707  6.691 -195.7 WCFTDT LOGSALE LOGASST NFATAST FATTOT QUIKRAT CURRAT
 8 0.769 0.709  7.507 -193.8 WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT CURRAT
 9 0.776 0.708  8.641 -191.5 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT
                             CURRAT
10 0.783 0.708  9.744 -189.1 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST PAYOUT
                             QUIKRAT CURRAT
11 0.786 0.702 11.277 -186.4 WCFTCL WCFTDT LOGSALE LOGASST NFATAST CAPINT FATTOT INVTAST
                             PAYOUT QUIKRAT CURRAT
12 0.788 0.695 13.000 -183.5 WCFTCL WCFTDT GEARRAT LOGSALE LOGASST NFATAST CAPINT FATTOT
                             INVTAST PAYOUT QUIKRAT CURRAT
```

Dans cet exemple, C_p et BIC se comportent de la même façon. Avec peu de variables, le modèle est trop biaisé. Ils atteignent un minimum pour un modèle à 4 variables explicatives puis croissent de nouveau selon la première bissectrice. La maximisation du R^2 ajusté conduirait à une solution beaucoup moins parcimonieuse. On note par ailleurs que l'algorithme remplace WCFTCL par WCFTDT. Un algorithme par sélection ne peut pas aboutir à la solution optimale retenue.

4 Régression régularisée ou pénalisée

4.1 Régression *ridge*

Modèle et estimation

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables explicatives pour des raisons d'interprétation, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur biaisé des paramètres par une procédure de régularisation.

Soit le modèle linéaire :

$$\mathbf{Y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon},$$

où

$$\widetilde{\mathbf{X}} = \begin{pmatrix} 1 & X_1^1 & X_1^2 & . & X_1^p \\ 1 & X_2^1 & X_2^2 & . & X_2^p \\ . & . & . & . & . \\ 1 & X_n^1 & X_n^2 & . & X_n^p \end{pmatrix},$$

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

où $\boldsymbol{X}^0 = (1, 1, \ldots, 1)'$, et $\mathbf{X}$ désigne la matrice $\widetilde{\mathbf{X}}$ privée de sa première colonne. L'estimateur *ridge* est défini par un critère des moindres carrés, avec une pénalité de type $\mathbb{L}^2$:

DÉFINITION 1. — *L'estimateur* ridge *de* $\widetilde{\boldsymbol{\beta}}$ *dans le modèle*

$$\mathbf{Y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon},$$

est défini par :

$$\widehat{\boldsymbol{\beta}}_{ridge} = argmin_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right),$$

où λ *est un paramètre positif, à choisir.*

À noter que le paramètre β_0 n'est pas pénalisé.

PROPOSITION 2. — *L'estimateur* ridge *s'exprime aussi sous la forme :*

$$\widehat{\beta_0}_{ridge} = \bar{Y}, \quad \begin{pmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \\ . \\ . \\ \widehat{\beta_p} \end{pmatrix}_{ridge} = argmin_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\mathbf{Y}^{(c)} - \mathbf{X}^{(c)}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right).$$

où $\mathbf{X}^{(c)}$ *désigne la matrice* $\boldsymbol{X}$ *recentrée (par colonnes) et* $\mathbf{Y}^{(c)}$ *désigne le vecteur* $\boldsymbol{Y}$ *recentré.*

Supposant désormais que $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont centrés, l'estimateur *ridge* est obtenue en résolvant les équations normales qui s'expriment sous la forme :

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)\beta.$$

Conduisant à :

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}.$$

La solution est donc explicite et linéaire en $\mathbf{Y}$.

Remarques :

1. $\mathbf{X}'\mathbf{X}$ est une matrice symétrique positive (pour tout vecteur $\boldsymbol{u}$ de $\mathbb{R}^p$, $\boldsymbol{u}'(\mathbf{X}'\mathbf{X})\boldsymbol{u} = \|\mathbf{X}\boldsymbol{u}\|^2 \geq 0$. Il en résulte que pour tout $\lambda > 0$, $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p$ est nécessaire inversible.
2. La constante β_0 n'intervient pas dans la pénalité, sinon, le choix de l'origine pour $\mathbf{Y}$ aurait une influence sur l'estimation de l'ensemble des paramètres. Alors : $\widehat{\beta}_0 = \overline{\mathbf{Y}}$; ajouter une constante à $\mathbf{Y}$ ne modifie pas les $\widehat{\beta}_j$ pour $j \geq 1$.
3. L'estimateur *ridge* n'est pas invariant par renormalisation des vecteurs $X^{(j)}$, il est préférable de normaliser (réduire les variables) les vecteurs avant de minimiser le critère.
4. La régression *ridge* revient encore à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur $\boldsymbol{\beta}$ des paramètres ne soit pas trop grande :

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \ ; \ \|\boldsymbol{\beta}\|^2 < c \right\}.$$

   La régression ridge conserve toutes les variables mais, contraignant la norme des paramètres β_j, elle les empêche de prendre de trop grandes valeurs et limite ainsi la variance des prévisions.

Optimisation de la pénalisation

La figure 4 montre quelques résultats obtenus par la méthode *ridge* en fonction de la valeur de la pénalité $\lambda = l$ sur l'exemple de la régression polynomiale. Plus la pénalité augmente et plus la solution obtenue est régulière ou encore, plus le biais augmente et la variance diminue. Il y a sur-ajustement

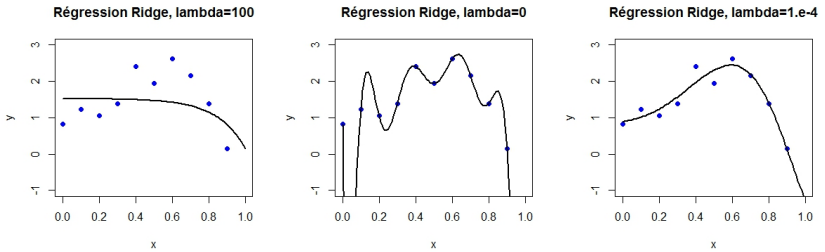

FIGURE 4 – Pénalisation *ridge* du modèle polynomial

avec une pénalité nulle : le modèle passe par tous les points mais oscille dangereusement ; il y a sous-ajustement avec une pénalité trop grande.

Comme dans tout problème de régularisation, le choix de la valeur du paramètre λ est crucial est déterminera le choix de modèle. La validation croisée est généralement utilisée pour optimiser le choix car la lecture du graphique (cf. figure 5) montrant l'évolution des paramètres en fonction du coefficient ou *chemins de régularisation ridge* n'est pas suffisante pour déterminer une valeur optimale.

Le principe de la validation croisée qui permet d'estimer sans biais une erreur de prévision est détaillé par ailleurs.

4.2 Régression LASSO

La régression *ridge* permet donc de contourner les problèmes de colinéarité même en présence d'un nombre important de variables explicatives ou prédicteurs ($p > n$). La principale faiblesse de cette méthode est liée aux difficultés d'interprétation car, sans sélection, toutes les variables sont concernées dans le modèle. D'autres approches par pénalisation permettent également une sélection, c'est le cas de la régression Lasso.

Modèle et estimation

La méthode Lasso (Tibshirani, 1996)[8] correspond à la minimisation d'un critère des moindres carrés avec une pénalité de type l_1 (et non plus l_2 comme

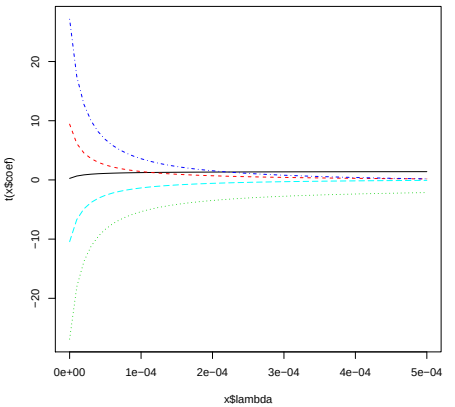

FIGURE 5 – *Modèle polynomial : Chemin de régularisation en régression* ridge *en fonction du paramètre de la pénalisation.*

dans la régression *ridge*). Soit $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.

DÉFINITION 3. — *L'estimateur Lasso de $\boldsymbol{\beta}$ dans le modèle*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

est défini par :

$$\widehat{\boldsymbol{\beta}}_{Lasso} = argmin_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

où λ est un paramètre positif, à choisir.

On peut montrer que ceci équivaut au problème de minimisation suivant :

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}} = \text{argmin}_{\boldsymbol{\beta}, \|\boldsymbol{\beta}\|_1 \leq t} (\|\mathbf{Y} - \mathbf{X}\beta\|^2),$$

pour un t convenablement choisi.

Comme dans le cas de la régression *ridge*, le paramètre λ est un paramètre de régularisation :

- Si $\lambda = 0$, on retrouve l'estimateur des moindres carrés.
- Si λ tend vers l'infini, on annule tous les $\hat{\beta}_j$, $j = 1, \ldots, p$.

La solution obtenue est dite parcimonieuse (*sparse* en anglais), car elle comporte des coefficients nuls.

Autre pénalisation

La méthode Lasso équivaut à minimiser le critère

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

sous la contrainte $\sum_{j=1}^{p} |\beta_j| \leq t$, pour un $t > 0$.

Le logiciel R introduit une contrainte sous forme d'une borne relative pour $\sum_{j=1}^{p} |\beta_j|$: la contrainte s'exprime sous la forme

$$\sum_{j=1}^{p} |\beta_j| \leq \kappa \sum_{j=1}^{p} |\hat{\beta}_j^{(0)}|,$$

où $\hat{\beta}^{(0)}$ est l'estimateur des moindres carrés et $\kappa \in [0, 1]$.

Avec $\kappa = 1$ c'est l'estimateur des moindres carrés (pas de contrainte) et pour $\kappa = 0$, tous les $\hat{\beta}_j$, $j \geq 1$, sont nuls (contrainte maximale).

Utilisation de la régression Lasso

La pénalisation est optimisée comme en régression *ridge* par validation croisée.

Grâce à ses solutions parcimonieuses, cette méthode est surtout utilisée pour sélectionner des variables dans des modèles de grande dimension ; on peut l'utiliser si $p > n$ c'est-à-dire s'il y a plus de variables que d'observations. Bien entendu, dans ce cas, les colonnes de la matrice $\mathbf{X}$ ne sont pas linéairement indépendantes. Il n'y a donc pas de solution explicite, on utilise des procédures d'optimisation pour trouver la solution. Il faut néanmoins utiliser la méthode avec précaution lorsque les variables explicatives sont corrélées. Pour que la méthode fonctionne, il faut que le nombre de variables influentes (correspondant à des β_j différents de 0) ne dépasse pas n et que les variables non influentes ne soient pas trop corrélées avec celles qui le sont.

4.3 Elastic Net

La méthode *Elastic Net* permet de combiner la régression *ridge* et la régression Lasso, en introduisant les deux types de pénalités simultanément.

Le critère à minimiser est :

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^{(1)} - \beta_2 X_i^{(2)} - \ldots - \beta_p X_i^{(p)})^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

- Pour $\alpha = 1$, on retrouve la méthode LASSO.
- Pour $\alpha = 0$, on retrouve la régression *ridge*.

Il y a dans ce dernier cas deux paramètres à optimiser par validation croisée.

4.4 Sélection par réduction de dimension

Le principe de ces approches consiste à calculer la régression sur un ensemble de variables orthogonales deux à deux. Celles-ci peuvent être obtenues à la suite d'une analyse en composantes principales ou par décomposition en valeur singulière de la matrice $\mathbf{X}$: c'est la régression sur les composantes principales associées aux plus grandes valeurs propres.

L'autre approche ou régression PLS (*partial least square* consiste à rechercher itérativement une composante linéaire des variables de plus forte covariance avec la variable à expliquer sous une contrainte d'orthogonalité avec les composantes précédentes.

Ce deux méthodes sont développées dans une vignette spécifique.

5 Exemples

5.1 Prévision de la concentration d'ozone

Les données

Les données proviennent des services de Météo-France et s'intéresse à la prévision de la concentration en Ozone dans 5 stations de mesure ; ces sites ont été retenus pour le nombre important de pics de pollution qui ont été détectés dans les périodes considérées (étés 2002, 2003, 2005). Un pic de pollution est défini ici par une concentration dépassant le seuil de $150\mu g/m^3$. Météo-France dispose déjà d'une prévision (MOCAGE), à partir d'um modèle physique basé sur les équations du comportement dynamique de l'atmosphère (Navier et Stockes). Cette prévision fait partie du dispositif d'alerte des pouvoirs publics et prévoit donc une concentration de pollution à 17h locale pour le lendemain. L'objet du travail est d'en faire une évaluation statistique puis de l'améliorer en tenant compte d'autres variables ou plutôt d'autres prévisions faites par Météo-France. Il s'agit donc d'intégrer ces informations dans un modèle statistique global.

Les variables

Certaines variables de concentration ont été transformées afin de rendre symétrique (plus gaussienne) leur distribution.

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.

Modèle physique

Les graphiques de la figure 6 représente la première prévision de la concentration d'ozone observée, ainsi que ses résidus, c'est-à-dire celle obtenue par

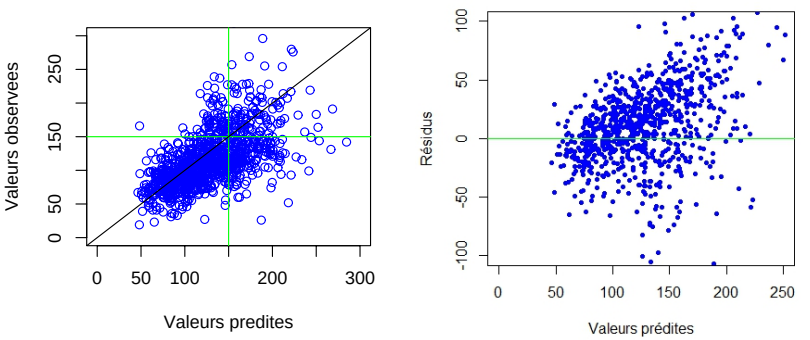

FIGURE 6 – Ozone : prévision et résidus du modèle MOCAGE de Météo-France pour 5 stations.

le modèle physique MOCAGE. Ces graphes témoignent de la mauvaise qualité de ce modèle : les résidus ne sont pas répartis de façon symétrique et les deux nuages présentent une légère forme de "banane" signifiant que des composantes non linéaires du modèle n'ont pas été prises en compte. D'autre part, la forme d'entonnoir des résidus montrent une forte hétéroscédasticité. Cela signifie que la variance des résidus et donc des prévisions croît avec la valeur. En d'autre terme, la qualité de la prévision se dégrade pour les concentrations élevées justement dans la zone sensible.

Modèle sans interaction

Un premier modèle est estimé avec R :

```
fit.lm=lm(O3-o~O3-pr+vmodule+lno2+lno+s-rmh2o+
          jour+station+TEMPE,data=donne)
```

Il introduit l'ensemble des variables explicatives mais sans interaction. Les résultats numériques sont fournis ci-dessous.

```
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
```

A l'exception de la variable indiquant la nature du jour, l'ensemble des coefficients sont jugés significativement différent de zéro mais la qualité de l'ajustement est faible (R^2).

Modèle avec interaction

La qualité d'ajustement du modèle précédent n'étant pas très bonne, un autre modèle est considéré en prenant en compte les interactions d'ordre 2 entre les variables. Compte tenu de la complexité du modèle qui un découle, un choix automatique est lancé par élimination successive des termes non significatifs (algorithme backward). Le critère optimisé est celui (AIC) d'Akaïke. Plusieurs interactions ont été éliminées au cours de la procédure mais beaucoup subsistent dans le modèle. Attention, les effets principaux `lno2, vmodule` ne peuvent être retirés car ces variables apparaissent dans une interaction. En revanche on peut s'interroger sur l'opportunité de conserver celle entre la force du vent et la concentration de dioxyde d'azote.

```
                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 *
```

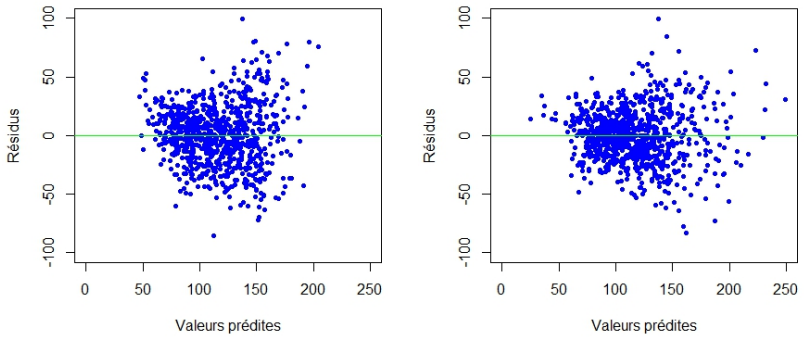

FIGURE 7 – Ozone : Résidus des modèles linéaire et quadratique.

```
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 ***
```

Ce sont surtout les graphes de la figure 7 qui renseignent sur l'adéquation des modèles. Le modèle quadratique fournit une forme plus "linéaire" des résidus et un meilleur ajustement avec un R^2 de 0,64 mais l'hétéroscédasticité reste présente, d'autres approches s'avèrent nécessaires afin de réduire la variance liée à la prévision des concentrations élevées.

Sélection Lasso

Les résulats précédents ont été obtenus avec R qui propose des algorithmes de sélection de variables classique. Ce n'est pas le cas de la librairie `scikit-learn` qui se limite à des sélections par pénalisation Lasso mais sans pouvoir intégrer facilement les interactions alors que celle-ci sont justement importantes pour ces données. L'optimisation du paramètre de pénalisation et les chemins de régularisation sont obtenus par validation croisée (figur 8).

Encore un peu de travail est nécessaire pour obtenir les coefficients finale-

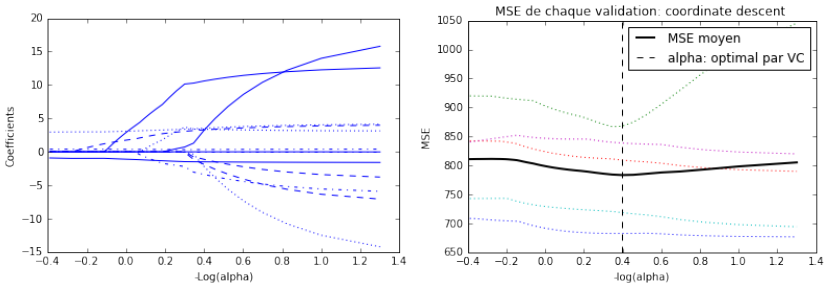

FIGURE 8 – Ozone : Régression lasso ; chemin de régularisation des paramètres et optimisation de la pénalisation.

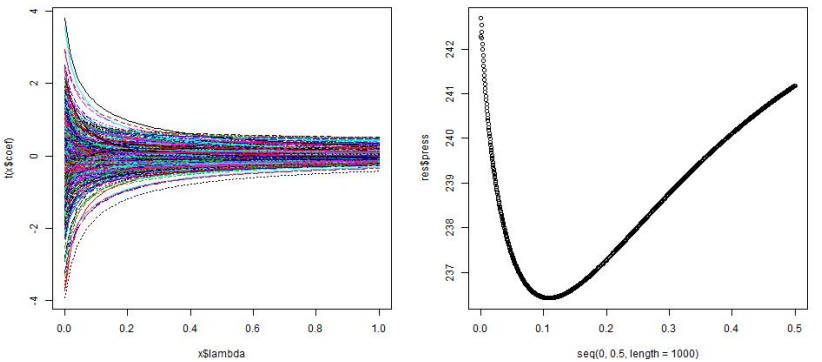

FIGURE 9 – Cookies : Régression *ridge* ; chemin de régularisation des paramètres et optimisation de la pénalisation avec `scikit-learn` en Python.

ment sélectionnés par la procédure.

5.2 Données de spectrométrie NIR

Objectif

Ce type de problème se rencontre en contrôle de qualité sur une chaîne de fabrication agroalimentaire, ici des biscuits (*cookies*). Il est nécessaire de contrôler le mélange des ingrédients avant cuisson afin de s'assurer que les proportions en lipides, sucre, farine, eau, sont bien respectées. Il s'agit de savoir s'il est possible de dépister au plus tôt une dérive afin d'intervenir sur les équipements concernés. Les mesures et analyses, faites dans un laboratoire classique de chimie, sont relativement longues et coûteuses ; elles ne peuvent être entreprises pour un suivi régulier ou même en continue de la production. Dans ce contexte, un spectromètre en proche infrarouge (NIR) mesure l'absorbance c'est-à-dire les spectres dans les longueurs d'ondes afin de construire un modèle de prévision de la concentration en sucre.

Les données

Les données originales sont dues à Osbone et al. (1984) [6] et ont été souvent utilisées pour la comparaison de méthodes (Stone et al. 1990 [7], Brown et al. 2001 [1], Krämer et al. 2008 [4]). Elles sont accessibles dans R au sein de la librairie *ppls*. Les mesures ont été faites sur deux échantillons, l'un de taille 40 prévu pour l'apprentissage, l'autre de taille 32 pour les tests. Pour chacun de ces 72 biscuits, les compositions en lipides, sucre, farine, eau, sont mesurées par une approche classique tandis que le spectre est observé sur toutes les longueurs d'ondes entre 1100 et 2498 nanomètres, régulièrement espacés de 2 nanomètres. Nous avons donc 700 valeurs observées, ou variables potentiellement explicatives, par échantillon de pâte à biscuit.

Résultats par régression pénalisée

Typiquement, cette étude se déroule dans un contexte de très grande dimension avec $p >> n$. L'étude détaillée de ces données fait l'objet d'un scénario avec le logiciel R.

Voici quelques résultats partiels concernant les méthodes de régression par régression *ridge* et régression LASSO. La comparaison globale des résultats des différentes approches de modélisation est reportée en conclusion.

Références

[1] P.J. Brown, T. Fearn et M. Vannucci, *Bayesian Wavelet Regression on*

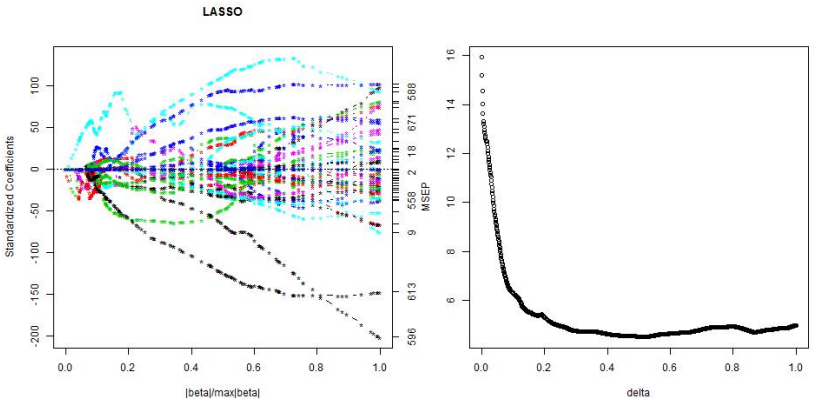

FIGURE 10 – Cookies : Régression lasso ; chemin de régularisation des paramètres et optimisation de la pénalisation.

Curves with Applications to a Spectroscopic Calibration Problem, Journal of the American Statistical Society **96** (2001), 398–408.

[2] G. M. Furnival et R. W. Wilson, *Regression by leaps and bounds*, Technometrics **16** (1974), 499–511.

[3] J.D. Jobson, *Applied Multivariate Data Analysis*, t. I : Regression and experimental design, Springer-Verlag, 1991.

[4] Nicole Krämer, Anne Laure Boulesteix et Gerhard Tutz, *Penalized Partial Least Squares with applications to B-spline transformations and functional data*, Chemometrics and Intelligent Laboratory Systems **94** (2008), 60–69.

[5] C.L. Mallows, *Some Comments on Cp*, Technometrics **15** (1973), 661–675.

[6] B. G. Osborne, T. Fearn, A. R. Miller et S. Douglas, *Application of Near Infrared Reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs*, J. Sci. Food Agric. **35** (1984), 99–105.

[7] M. Stone et R. J. Brooks, *Continuum regression : cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression*, Journal of The Royal Statistical Society B **52** (1990), 237–269.

[8] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288.