Linear models for regression and Model selection

1 Introduction

The linear regression model is the simplest model to study multidimensional data. It assumes that the regression function $\mathbb{E}(\mathbf{Y}/\mathbf{X})$ is linear in the input (or explanatory) variables $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p$. Although very simple, these models are still widely used, because they are very interpretable and often provide an adequate description on the influence of the input variables to the output. For small sample sizes n (with respect to the number of variables p), or when the signal to noise ratio is high, they often outperform more complex models. Furthermore, it is possible to use linear models with nonlinear transformations of the variables, which considerably enlarges the scope of these models. In high dimensional framework, when p is possibly larger than n, model selection for linear models has been this past twenty years and is still a very active field of research in statistics. Some of these methods, such as Ridge or Lasso methods, will be at the core of this course. The main references for this course are the books "Introduction to High-Dimensional Statistics" by C. Giraud [3] and " The elements of Statistical Learning" by T. Hastie et al [4].

2 The Linear model

2.1 The model

We have a quantitative variable $\mathbf{Y}$ *to explain* (or response variable) which is related with p variables $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p$ called *explanatory variables* (or regressors, or input variables).

The data are obtained from the observation of a n sample of $\mathbb{R}^{(p+1)}$ vectors :

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

We assume in a first time that $n > p + 1$. In *the linear model*, the regression function $\mathbb{E}(\mathbf{Y}/\mathbf{X})$ is linear in the input (or explanatory) variables $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p$. We assume for the sake of simplicity that the regressors are deterministic. In this case, this means that $\mathbb{E}(\mathbf{Y})$ is linear in the explanatory variables $\{\mathbf{1}, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^p\}$ where $\mathbf{1}$ denotes the $\mathbb{R}^n$-vector with all components equal to 1. The linear model is defined by :

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

with the following assumptions :

1. The random variables ε_i are independent and identically distributed (i.i.d.) ; $\mathbb{E}(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$.
2. The regressors $\boldsymbol{X}^j$ are assumed to be deterministic **or** the errors $\boldsymbol{\varepsilon}$ are independent of $(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p)$. In this case, we have :

   $E(\mathbf{Y}|\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p) = \beta_0 + \beta_1 \boldsymbol{X}^1 + \beta_2 \boldsymbol{X}^2 + \cdots + \beta_p \boldsymbol{X}^p$ and $Var(\mathbf{Y}|\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p) = \sigma^2$.
3. The unknown parameters $\beta_0, \ldots, \beta_p$ are supposed to be constant.
4. It is sometimes assumed that the errors are Gaussian : $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_n]' \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I_n})$. The variables ε_i are then i.i.d. $\mathcal{N}(0, \sigma^2)$.

The explanatory variables are given in the matrix $\mathbf{X}(n \times (p+1))$ with general term X_i^j, the first column contains the vector $\mathbf{1}$ ($X_0^i = 1$). The regressors $\boldsymbol{X}^j$ can be quantitative variables, nonlinear transformation of quantitative variables (such as log, exp, square ..), interaction between variables $\boldsymbol{X}^j = \boldsymbol{X}^k.\boldsymbol{X}^l$, they can also correspond to qualitative variables : in this case the variables $\boldsymbol{X}^j$ are indicator variables coding the different levels of a factor (we remind that we need identifiability conditions in this case).
The response variable is given in the vector $\mathbf{Y}$ with general term Y_i. We set $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$, which leads to the matricial formulation of the linear model:

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

2.2 Least square estimation

The regressors $\boldsymbol{X}^j$ are observed, the unknown parameters of the model are the vector $\boldsymbol{\beta}$ and σ^2. $\boldsymbol{\beta}$ is estimated by minimizing the residuals sum of square or equivalently, assuming (4.), by maximisation of the likelihood.

We minimise with respect to the parameter $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ the criterion :

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

Derivating the last equation, we obtain the *" normal equations"* :

$$2(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = 0$$

The solution is indeed a minimiser of the criterion since the Hessian $2\mathbf{X}'\mathbf{X}$ is positive semi definite (the criterion is convex) .

We make the additional assumption that the matrix $\mathbf{X}'\mathbf{X}$ is invertible, which is equivalent to the fact that the matrix $\mathbf{X}$ has full rank $(p+1)$ and so that there is no collinearity between the columns of $\mathbf{X}$ (the variables). Under this assumption, the estimation of $\boldsymbol{\beta}$ is give by :

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

and the predicted values of $\mathbf{Y}$ are :

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the "*hat matrix*" ; which puts a "hat" on $\mathbf{Y}$. Geometrically, it corresponds to the matrix of orthogonal projection in $\mathbb{R}^n$ onto the subspace Vect($\mathbf{X}$) generated by the columns of $\mathbf{X}$.

Remark. — We have assumed that $\mathbf{X}'\mathbf{X}$ is invertible, which means that the columns of $\mathbf{X}$ are linearly independent. If it is not the case, this means that the application $\boldsymbol{\beta} \mapsto \mathbf{X}\boldsymbol{\beta}$ is not injective, hence the model is not identifiable and $\boldsymbol{\beta}$ is not uniquely defined. Nevertheless, even in this case, the predicted values $\widehat{\mathbf{Y}}$ are still defined as the projection of $\mathbf{Y}$ onto the space generated by the columns of $\mathbf{X}$, even if there is not a unique $\widehat{\boldsymbol{\beta}}$ such that $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. In practice, if $\mathbf{X}'\mathbf{X}$ is not invertible (which is necessarily the case in high dimension when the number of variables p is larger than the number of observations n - since p vectors of $\mathbb{R}^n$ are necessarily linearly dependent), we have to remove variables from the model or to consider other approches to reduce the dimension (*Ridge*, Lasso, PLS ...) that we will developed in the next chapters.

We define the vector of residuals as :

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

This is the orthogonal projection of $\mathbf{Y}$ onto the subspace Vect$(\mathbf{X})^{\perp}$ in $\mathbb{R}^n$. The variance σ^2 is estimated by

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

2.3 Properties of the least square estimator

THEOREM 1. — *Assuming that*

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

with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\mathbf{I_n})$*, we obtain that* $\widehat{\boldsymbol{\beta}}$ *is a Gaussian vector :*

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}).$$

In particular, the components of $\widehat{\boldsymbol{\beta}}$ *are Gaussian variables :*

$$\widehat{\boldsymbol{\beta}}_j \sim \mathcal{N}(\boldsymbol{\beta}_j, \sigma^2(X'X)^{-1}_{j,j}).$$

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-(p+1)}\chi^2_{(n-(p+1))}$$

and is independent of $\hat{\beta}$.

Exercise. — Prove Theorem 1

$\widehat{\boldsymbol{\beta}}$ is a linear estimator of $\boldsymbol{\beta}$ (it is a linear transformation of the observation $\mathbf{Y}$) and it is unbiased. One can wonder if it has some optimality property. This is indeed the case : the next theorem, called the *Gauss-Markov* theorem, is very famous in statistics. It asserts that the least square estimator $\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimator of $\boldsymbol{\beta}$.

THEOREM 2. — *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *two matrices. We say that* $\mathbf{A} \preceq \mathbf{B}$ *if* $\mathbf{B} - \mathbf{A}$ *is positive semi-definite. Let* $\widetilde{\boldsymbol{\beta}}$ *a linear unbiased estimator of* $\boldsymbol{\beta}$*, with variance-covariance matrix* $\widetilde{\mathbf{V}}$*. Then,* $\sigma^2(\mathbf{X}'\mathbf{X})^{-1} \preceq \widetilde{\mathbf{V}}$.

Exercise. — Prove the Gauss-Markov theorem.

Theorem 2 shows that the estimator $\widehat{\boldsymbol{\beta}}$ is the best among all linear unbiased estimator of $\boldsymbol{\beta}$, nevertheless, in the next section, we will see that it can be preferable to consider biased estimator, if they have a smaller variance than $\widehat{\boldsymbol{\beta}}$, to reduce the quadratic risk. This will be the case for the Ridge, Lasso, PCR, or PLS regression.

2.4 Confidence intervals

One can easily deduce from Theorem 1 that

$$\frac{\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{i,i}}} \sim \mathcal{T}_{(n-(p+1))}$$

follows a Student distribution with $n-(p+1)$ degrees of freedom. This allows to build confidence intervals and tests for the parameters $\boldsymbol{\beta}_j$. The following interval is a 0.95 confidence interval for $\boldsymbol{\beta}_j$:

$$[\widehat{\boldsymbol{\beta}}_j - t_{n-(p+1),0.975}\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{j,j}}, \widehat{\boldsymbol{\beta}}_j + t_{n-(p+1),0.975}\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{j,j}}].$$

In order to test that the variable associated to the parameter $\boldsymbol{\beta}_j$ has no influence in the model, hence $H_0 : \boldsymbol{\beta}_j = 0$ contre $H_1 : \boldsymbol{\beta}_j \neq 0$, we reject the null hypothesis at the level 5% if 0 does not belong to the previous confidence interval.

Exercise. — Recover the construction of the confidence intervals.

2.5 Prediction

As mentioned above, the vector of predicted values is

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

This corresponds to the predicted values at the observation points. Based on the n previous observations, we may be interested with the prediction of the response of the model for a new point : $\mathbf{X_0}' = (1, {X_0}^1, \ldots, {X_0}^p)$:

$$Y_0 = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_0^1 + \boldsymbol{\beta}_2 X_0^2 + \ldots + \boldsymbol{\beta}_p X_0^p + \varepsilon_0,$$

where $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$. The predicted value is

$$\widehat{Y}_0 = \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 {X_0}^1 + \ldots \widehat{\boldsymbol{\beta}}_p {X_0}^p = \mathbf{X_0}'\widehat{\boldsymbol{\beta}}.$$

We derive from Theorem 1 that

$$\mathbb{E}(\widehat{Y}_0) = \mathbf{X_0}'\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_0^1 + \boldsymbol{\beta}_2 X_0^2 + \ldots + \boldsymbol{\beta}_p X_0^p$$

and that $\widehat{Y}_0 \sim \mathcal{N}(\mathbf{X_0}'\boldsymbol{\beta}, \sigma^2 \mathbf{X_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_0})$. We can deduce a confidence interval for the mean response $\mathbf{X_0}'\boldsymbol{\beta}$ at the new observation point $\mathbf{X_0}$:

$$\left[\mathbf{X_0}'\widehat{\boldsymbol{\beta}} - t_{n-(p+1),0.975}\hat{\sigma}\sqrt{\mathbf{X_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_0}},\right.$$

$$\left.\mathbf{X_0}'\widehat{\boldsymbol{\beta}} + t_{n-(p+1),0.975}\hat{\sigma}\sqrt{\mathbf{X_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_0}}\right].$$

A prediction interval for the response Y_0 at the new observation point $\mathbf{X_0}$ is :

$$\left[\mathbf{X_0}'\widehat{\boldsymbol{\beta}} - t_{n-(p+1),0.975}\hat{\sigma}\sqrt{1 + \mathbf{X_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_0}},\right.$$

$$\left.\mathbf{X_0}'\widehat{\boldsymbol{\beta}} + t_{n-(p+1),0.975}\hat{\sigma}\sqrt{1 + \mathbf{X_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_0}}\right].$$

Exercise. — Recover the construction of the prediction intervals. Hint : what is the distribution of $\widehat{Y}_0 - Y_0$?

2.6 Fisher test of a submodel

Suppose that our data obey to a polynomial regression model of degree p and we want to test the null hypothesis that our data obey to a polynomial regression model of degree $k < p$, hence we want to test that the $p - k$ last coefficients of $\boldsymbol{\beta}$ are equal to 0. More generally, assume that our data obey to the model, called Model (1) :

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

where $\beta \in R^p$ and consider another model, called Model (0):

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

where $\theta \in \mathbb{R}^l$ with $l < p$.

DEFINITION 3. — *We define*

$$V = \{\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^p\}$$

and

$$W = \{\tilde{\mathbf{X}}\boldsymbol{\theta}, \boldsymbol{\theta} \in \mathbb{R}^l\}.$$

We say that Model (0) is a submodel of Model (1) if W is a linear subspace of V.

We want to test the hypothesis :
H_0 : "the vector $\mathbf{Y}$ of observations obeys to Model (0)" against the alternative
H_1 : "the vector $\mathbf{Y}$ of observations obeys to Model (1)".
In the Model (0), the least square estimator of $\boldsymbol{\theta}$ is :

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\boldsymbol{\theta}}_0 \\ \widehat{\boldsymbol{\theta}}_1 \\ . \\ . \\ \widehat{\boldsymbol{\theta}}_l \end{pmatrix} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y}.$$

The F-statistics is defined by :

$$F = \frac{\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \tilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|^2/(p-l)}{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p)}.$$

An alternative way to write the F-statistics is :

$$F = \frac{(\text{SSR}_\mathbf{0} - \text{SSR}_\mathbf{1})/(p-l)}{\text{SSR}_\mathbf{1}/(n-p)},$$

where SSR_0 and SSR_1 respectively denote the residuals sum of square under Model (0) and Model (1).

Exercise. — Prove that, under the null hypothesis H_0, the F-statistics is a Fisher distribution with parameters $(p-l, n-p)$.

The numerator of the F-statistics corresponds to $\left\|\widehat{\mathbf{Y}}_\mathbf{0} - \widehat{\mathbf{Y}}_\mathbf{1}\right\|^2$, where $\widehat{\mathbf{Y}}_\mathbf{0}$ and $\widehat{\mathbf{Y}}_\mathbf{1}$ correspond respectively to the predicted values under the sub-model and under the full model. This quantity is small under the null hypothesis, when the sub-model is valid, and becomes larger under the alternative. Hence, the null hypothesis is rejected for large values of F, namely, for a level-α test, when

$$F > f_{p-l,n-p,1-\alpha},$$

where $f_{p,q,1-\alpha}$ is the $(1-\alpha)$ quantile of the Fisher distribution with parameters (p, q). The statistical softwares provide the $p-$ value of the test :

$$P_{H_0}(F > F_{obs})$$

where F_{obs} is the observed value for the F-statistics. The null hypothesis is rejected at level α if the $p-$ value is smaller than α.

2.7 Diagnosis on the residuals

The analysis and visualisation of the residuals allow to verify some hypotheses :

- Homoscedasticity: the variance σ^2 is assumed to be constant,
- The linear model is valid : there is no tendancy in the residuals,
- Detection of possible outliers with the Cook's distance
- Normality of the residuals (if this assumption was used to provide confidence/prediction intervals or tests).

This is rather classical for linear regression, and we focus here on the detection of possible high collinearities between the regressors, since it has an impact on the variance of our estimators. Indeed, we have seen that the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

When the matrix $\mathbf{X}$ is *ill-conditioned*, which means that the determinant of $\mathbf{X}'\mathbf{X}$ is close to 0, we will have high variances for some components of $\widehat{\boldsymbol{\beta}}$. It is therefore important to detect and remedy these situations by removing some variables of the model or introducing some constraints on the parameters to reduce the variance of the estimators.

VIF

Most statistical softwares propose collinearity diagnosis. The most classical il the *Variance Influence Factor* (VIF)

$$V_j = \frac{1}{1 - R_j^2}$$

where R_j^2 corresponds to the determination coefficient of the regression of the variable $\boldsymbol{X}^j$ on the other explanatory variables ; R_j represents also the cosine of the angle in $\mathbb{R}^n$ between $\boldsymbol{X}^j$ and the linear subspace generated by the variables $\{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^{j-1}, \boldsymbol{X}^{j+1}, \ldots, \boldsymbol{X}^p\}$. The more $\boldsymbol{X}^j$ is "linearly" linked with the other variables, the more R_j is close to 1 ; we show that the variance of the estimator of β_j is large in this case. This variance is minimal when $\boldsymbol{X}^j$ is orthogonal to the subspace generated by the other variables.

Condition number

We consider the covariance matrix $\mathbf{R}$ between the regressors. We denote $\lambda_1 \geq \ldots \geq \lambda_p$ the ordered eigenvalues of $\mathbf{R}$. If the smallest eigenvalues are close to 0, the inversion of the matrix $\mathbf{R}$ will be difficult and numerical problems arise. In this case, some components of the least square estimator $\widehat{\boldsymbol{\beta}}$ will have high variances. The condition number of the matrix $\mathbf{R}$ is defined as the ratio

$$\kappa = \lambda_1 / \lambda_p$$

between the largest and the smallest eigenvalues of $\mathbf{R}$. If this ratio is large, then the problem is ill-conditioned.
This condition number is a global indicator of collinearities, while the VIF allows to identify the variables that are problematic.

3 Determination coefficient and Model selection

3.1 R^2 and adjusted R^2

We define respectively the total, explicated and residual sums of squares by

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \left\| \mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1} \right\|^2,$$

$$\text{SSE} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \left\| \widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\mathbf{1} \right\|^2,$$

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 = \left\| \mathbf{Y} - \widehat{\mathbf{Y}} \right\|^2 = \|\mathbf{e}\|^2.$$

Since, by Pythagora's theorem,

$$\left\| \mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1} \right\|^2 = \left\| \mathbf{Y} - \widehat{\mathbf{Y}} \right\|^2 + \left\| \widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\mathbf{1} \right\|^2,$$

we have the following identity :

$$\text{SST} = \text{SSR} + \text{SSE}.$$

We define the determination coefficient R^2 by :

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

Note that $0 \leq R^2 \leq 1$. The model is well adjusted to the n training data if the residuals sum of square SSR is close to 0, or equivalently, if the determination coefficient R^2 is close to 1. Hence, the first hint is that a "good" model is a model for which R^2 is close to 1. This is in fact not true, as shown by the following pedagogical example of polynomial regression. Suppose that we have a training sample $(X_i, Y_i)_{1 \leq i \leq n}$ where $X_i \in [0, 1]$ and $Y_i \in \mathbb{R}$ and we adjust polynomials on these data :

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_i + \boldsymbol{\beta}_2 X_i^2 + \ldots + \boldsymbol{\beta}_k X_i^k + \varepsilon_i.$$

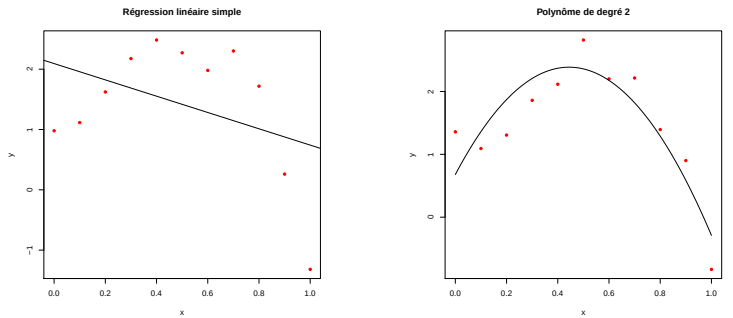

Figure 1: Polynomial regression : adjusted model, on the left : $y = \beta_0 + \beta_1 x + \epsilon$, $R^2 = 0.03$, on the right : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, $R^2 = 0.73$.

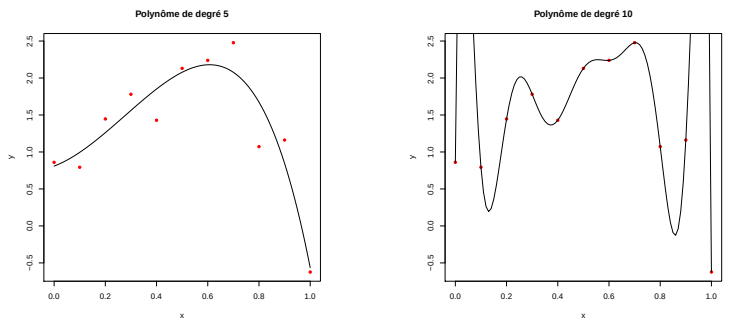

Figure 2: Polynomial regression : adjusted model, on the left : $y = \beta_0 + \beta_1 x + \ldots + \beta_5 x^5 + \epsilon$, $R^2 = 0.874$, , on the right : $y = \beta_0 + \beta_1 x + \ldots + \beta_{10} x^{10} + \epsilon$, $R^2 = 1$.

When k increases, the model is more and more complex, hence $\left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|^2$ decreases, and R^2 increases as shown in Figures 1 and 2.

The determination coefficient is equal to 1 for the polynomial of degree $n-1$ (which has n coefficients) and passes through all the training points. Of course this model is not the best one : it has a very high variance since we estimate as much coefficients as the number of observations. This is a typical case of *overfitting*. When the degree of the polynomial increases, the bias of our estimators decreases, but the variance increases. The best model is the one that realizes the best trade-off between the bias term and the variance term. Hence, we have seen that maximizing the determination coefficient is not a good criterion to compare models with various complexity. It is more interesting to consider the adjusted determination coefficient defined by :

$$R'^2 = 1 - \frac{\text{SSR}/(n-k-1)}{\text{SST}/(n-1)}.$$

The definition of R'^2 takes into account the complexity of the model, represented here by its number of coefficients : $k+1$ for a polynomial of degree k, and penalizes more complex models. One can choose, between several models, the one which maximizes the adjusted R^2. In the previous example, we would choose a polynomial of degree 3 with this criterion.
More generally, we have to define model selection procedures that realize a good compromise between a good adjustment to the data (small bias) and a small variance; and an unbiased estimator is not necessarily the best one in this sense. We will prefer a biased model if this allows to reduce drastically the variance. There are several ways to do that :

- Reducing the number of explanatory variables and by the same way simplifying the model (variable selection or *Lasso* penalization)
- Putting some constraints on the parameters of the model by *shrinking* them (*Ridge* or *Lasso* penalization)

3.2 Example

We consider the **Prostate** cancer data from the R package **lasso 2**.

" These data come from a study that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy."
It is data frame with 97 rows and 9 columns.

source Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989)
Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, *Journal of Urology* 141(5), 1076–1083."

The data frame has the following components:

lcavol	log(cancer volume)
lweight	log(prostate weight)
age	age
lbph	log(benign prostatic hyperplasia amount)
svi	seminal vesicle invasion (0, 1)
lcp	log(capsular penetration)
gleason	Gleason score of the tumor (6, 7, 8, 9)
pgg45	percentage Gleason scores 4 or 5
lpsa	log(prostate specific antigen)

We denote by Y the variable (**lpsa**) to explain. We set $X^1, \dots X^p$ for the explanatory variables (**lcavol** , **lweight** , **gleason** ..). The variables are quantitative (**lcavol** , **lweight** , ...), or qualitative (**gleason** , **svi**). We consider the linear model :

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

For the qualitative variables, we consider indicator functions of the different levels of the factor, and introduce some constraints for identifiability. By default, in R, the smallest value of the factor (0 for **svi** and 6 for **gleason**) are set in the reference. This is an analysis of covariance model (mixing quantitative and qualitative variables).
Exercise. — Write the matrix $\boldsymbol{X}$ of the model with the default constraints of R, and with the constraint that the sum of the coefficients associated to each modality of the qualitative variables is equal to 0 (*contr.sum* in R).

For the least square estimation, with the default constraints of R, we obtain the following results :

```
Coefficients   Estimate    Std.  Error   t value   Pr(>|t|)
(Intercept)    0.913313    0.840836      1.086     0.28043
lcavol         0.569989    0.090100      6.326     1.09e-08 ***
lweight        0.468783    0.169610      2.764     0.00699 **
age           -0.021749    0.011361     -1.914     0.05890 .
lbph           0.099685    0.058984      1.690     0.09464 .
svi1           0.745877    0.247398      3.015     0.00338 **
lcp           -0.125111    0.095591     -1.309     0.19408
gleason7       0.267601    0.219419      1.220     0.22596
gleason8       0.496798    0.769268      0.646     0.52012
gleason9      -0.056230    0.500196     -0.112     0.91076
pgg45          0.004990    0.004672      1.068     0.28847
Residual standard error:  0.7048 on 86 degrees of freedom
 Multiple R-squared:  0.666, Adjusted R-squared:  0.6272
```

Exercise. — Prove that the Student test of $H_0 : \beta_j = 0$ is equivalent to the Fisher test for the same hypothesis.

4 Variable selection

As we have seen, the least square estimator is not satisfactory since it has low bias but generally high variance. In the previous example, several variables seem to be non significant, and we may have better results by removing those variables from the model. Moreover, a model with a small number of variables is more interesting for the interpretation, keeping only the variables that have the strongest effects on the variable to explain. There are several ways to do that.

Assume we want to select a subset of variables among all possible subsets taken from the input variables. Each subset defines a model, and we want to select the "best model". We have seen that maximizing the R^2 is not a good criterion since this will always lead to select the full model. It is more interesting to select the model maximizing the adjusted determination coefficient R'^2. Many other penalized criterion have been introduce for variable selection

such as the Mallow's C_P criterion or the BIC criterion. In both cases, it corresponds to the minimization of the least square criterion plus some penalty term, depending on the number k of parameters in the model m that is considered.

$$\text{Crit}(m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \text{pen(k)}.$$

The Mallow's C_P criterion is

$$\text{Crit}_{C_P}(m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2k\sigma^2,$$

and the BIC criterion penalizes more the dimension of the model with an additional logarithmic term.

$$\text{Crit}_{BIC}(m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \log(n)k\sigma^2.$$

The aim is to select the model (among all possible subsets) that minimizes one of those criterion. On the example of the polynomial models, we obtain the results summarized in Figure 3.

Nevertheless, the number of subsets of a set of p variables is 2^p, and it is impossible (as soon as $p > 30$) to explore all the models to minimize the criterion. Fast algorithms have been developed to find a clever way to explore a subsample of the models. This are the *backward, forward* and *stepwise* algorithms.

Backward/Forward Algorithms :

- **Forward selection :** We start from the constant model (only the intercept, no explanatory variable), and we add sequentially the variable that allows to reduce the more the criterion.
- **Backward selection :** This is the same principle, but starting from the full model and removing one variable at each step in order to reduce the criterion.
- **Stepwise selection:** This is a mixed algorithm, adding or removing one variable at each step in order to reduce the criterion in the best way.

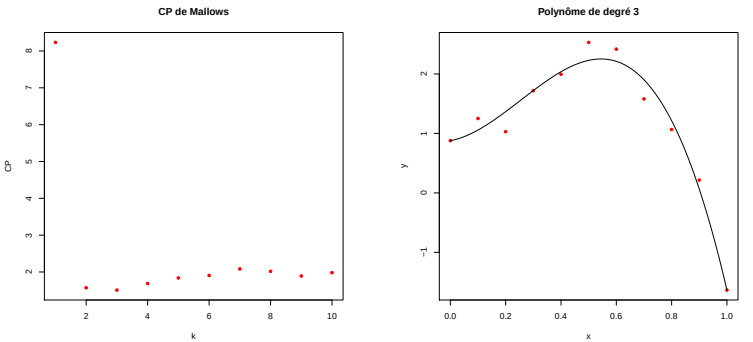

Figure 3: Mallows'C_P in function of the degree of the polynomial. Selected model : polynomial with degree 3.

All those algorithms stop when the criterion can no more be reduced. Let us see some applications of those algorithms on the **Prostate** cancer data.

Stepwise Algorithm

We apply the `StepAIC` algorithm, with the option **both** of the sofware R in order to select a subset of variables, and we present here an intermediate result :

```
                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
```

The algorithm stops when adding or removing a variable does no more allow to reduce the criterion. After 4 iterations, we get the following model :

$$\textbf{lpsa} \sim \textbf{lcavol + lweight + age + lbph + svi}.$$

5 Ridge regression

The principle of the Ridge regression is to consider all the explanatory variables, but to introduce constraints on the parameters in order to avoid overfitting, and by the same way in order to reduce the variance of the estimators. In the case of the Ridge regression, we introduce an l_2 constraint on the parameter $\boldsymbol{\beta}$.

5.1 Model and estimation

If we have an ill-conditionned problem, but we want to keep all the variables, it is possible to improve the numerical properties and to reduce the variance of the estimator by considering a slightly biased estimator of the parameter $\boldsymbol{\beta}$.

We consider the linear model

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

where

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

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

We set $\boldsymbol{X}^0 = (1, 1, \ldots, 1)'$, and $\mathbf{X}$ the matrix $\widetilde{\mathbf{X}}$ where we have removed the first column. The *ridge* estimator is defined by a least square criterion plus a penalty term, with an l_2 type penalty.

DEFINITION 4. — *The* ridge *estimator of $\widetilde{\boldsymbol{\beta}}$ in the model*

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

is defined by

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

where λ is a non negative parameter, that we have to calibrate.

Note that the parameter β_0 is not penalized.

PROPOSITION 5. — *The* ridge *estimator can be expressed as follows :*

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

where $\mathbf{X}^{(c)}$ is the matrix $\boldsymbol{X}$ that has been centered (for each column) and $\mathbf{Y}^{(c)}$ is the vector $\boldsymbol{Y}$, that has been centered.

We now assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered. We can find the *ridge* estimator by resolving the normal equations :

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

We get

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

The solution is therefore explicit and linear with respect to $\mathbf{Y}$.

Remarks :

1. $\mathbf{X}'\mathbf{X}$ is a nonnegative symmetric matrix (for all vector $\boldsymbol{u}$ in $\mathbb{R}^p$, $\boldsymbol{u}'(\mathbf{X}'\mathbf{X})\boldsymbol{u} = \|\mathbf{X}\boldsymbol{u}\|^2 \geq 0$. Hence, for any $\lambda > 0$, $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p$ is invertible.

2. The constant β_0 is not penalized, otherwise, the estimator would depend on the choice of the origin for $\mathbf{Y}$. We obtain $\widehat{\beta}_0 = \overline{\mathbf{Y}}$, adding a constant to $\mathbf{Y}$ does not modify the values of $\widehat{\beta}_j$ for $j \geq 1$.
3. The *ridge* estimator is not invariant by normalization of the vectors $X^{(j)}$, it is therefore important to normalize the vectors before minimizing the criterion.
4. The *ridge* regression is equivalent to the least square estimation under the constraint that the l_2-norm of the vector $\boldsymbol{\beta}$ is not too large:

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

The ridge regression keeps all the parameters, but, introducing constraints on the values of the β_j's avoids too large values for the estimated parameters, which reduces the variance.

Choice of the penalty term

In the Figure 4, we see results obtained by the *ridge* method for several values of the tuning parameter $\lambda = l$ on the polynomial regression example. Increasing the penalty leads to more regular solutions, the bias increases, and the variance decreases. We have overfitting when the penalty is equal to 0 and under-fitting when the penalty is too large.

For each regularization method, the choice of the parameter λ is crucial and determinant for the model selection. We see in Figure 5 the *Regularisation path*, showing the profiles of the estimated parameters when the tuning parameter λ increases.

Choice of the regularization parameter

Most softwares use the cross-validation to select the tuning parameter penalty. The principe is the following :

- We split the data into K sub-samples. For all I from 1 to K :
  - We compute the Ridge estimator associated to a regularization parameter λ from the data of all the subsamples, except the I-th (that will be a "‘test"’ sample).

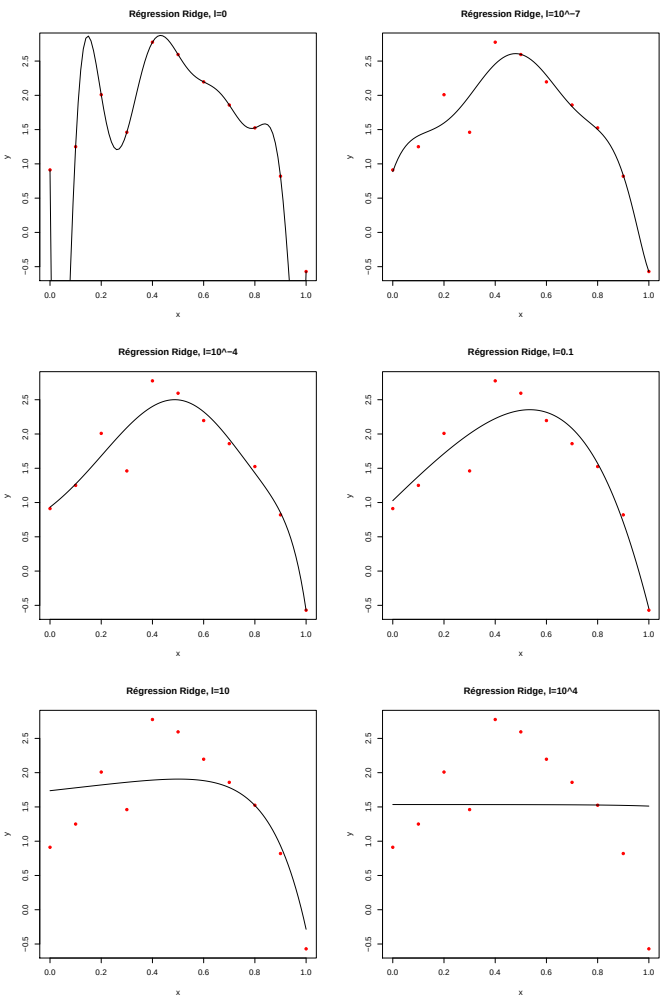

Figure 4: *Ridge* penalisation for the polynomial model

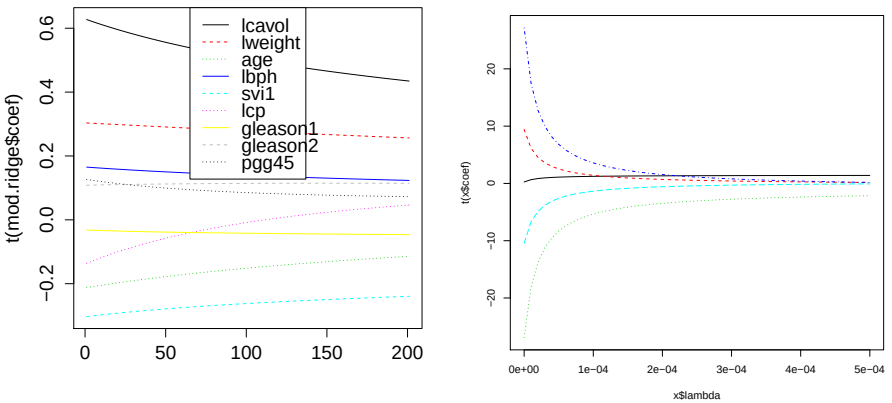

Figure 5: *Regularization paths for the* ridge *regression, as the tuning parameter varies. On the right, the polynomial regression and on the left, the Prostate cancer data).*

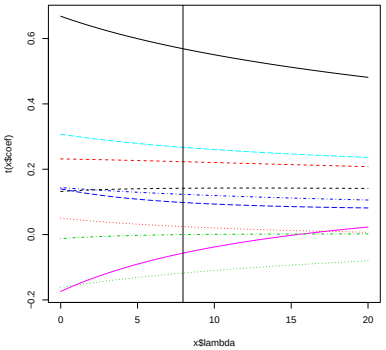

Figure 6: Selection of the regularization parameter by CV

- We denote by $\hat{\boldsymbol{\beta}}_\lambda^{(-I)}$ the obtained estimator.
- We test the performances of this estimator on the data that have not been used to build it, that is the one of the I-th sub-sample.

- We compute the criterion :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_\lambda^{(-\tau(i))})^2.$$

- We choose the value of λ which minimizes $CV(\lambda)$.

Application to the prostate cancer data: The value of λ selected by cross-validation is 7.96. We show the obtained value in Figure 6.

Singular Value Decomposition and Ridge regression

The Singular Value Decomposition (SVD) of the centered matrix $\mathbf{X}$ allows to interpret the *ridge* regression as a shrinkage method. The SVD of the matrix $\boldsymbol{X}$ has the following form :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

where $\mathbf{X}$ is a $n \times p$ matrix, $\mathbf{U}$ is $n \times n$, $\mathbf{D}$ is a $n \times p$ "diagonal" matrix whose all elements are ≥ 0 and ordered by decreasing values, $\mathbf{V}$ is a $p \times p$ matrix. The elements of D are the singular values of the matrix X. $\mathbf{U}$ and $\mathbf{V}$ are orthogonal: $\mathbf{UU}' = \mathbf{U}'\mathbf{U} = \mathbf{I}_n$, $\mathbf{VV}' = \mathbf{V}'\mathbf{V} = \mathbf{I}_p$.
We have

$$\mathbf{X}\widehat{\boldsymbol{\beta}}_R = \mathbf{UD}(\mathbf{D}'\mathbf{D} + \lambda \mathbf{I}_p)^{-1}\mathbf{D}'\mathbf{U}'\mathbf{Y}.$$

Suppose that $n \leq p$. We denote by $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n)}$ the columns of the matrix $\mathbf{U}$. Setting $d_1 \geq \ldots \geq d_p \geq 0$ the diagonal elements of $\mathbf{D}$, $\mathbf{UD}$ is a $n \times p$ matrix whose j-th column is $d_j \boldsymbol{u}^{(j)}$. We therefore have

$$\mathbf{X}\widehat{\boldsymbol{\beta}}_R = \sum_{j=1}^{p} \boldsymbol{u}^{\boldsymbol{j}} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) (\boldsymbol{u}^{\boldsymbol{j}})'\mathbf{Y}.$$

Let us compare this estimator with the least square estimator (which corresponds to $\lambda = 0$):

$$X\widehat{\boldsymbol{\beta}} = \sum_{j=1}^{p} \boldsymbol{u}^{\boldsymbol{j}} (\boldsymbol{u}^{\boldsymbol{j}})'\mathbf{Y}.$$

$(\boldsymbol{u}^{\boldsymbol{j}})'\mathbf{Y}$ corresponds to the j-th component of $\mathbf{Y}$ in the basis $(\boldsymbol{u}^{\boldsymbol{1}}, \ldots, \boldsymbol{u}^{\boldsymbol{n}})$.
In the case of the *ridge* regression, this component is multiplied by the factor $d_j^2 / \left(d_j^2 + \lambda\right) \in]0, 1[$, we can say that this component has been *thresholded*.
Remarks :
1) When the tuning parameter λ increases, the coefficients are more and more thresholded.
2) $x \mapsto x/(x + \lambda)$ is a non decreasing function of x for $x > 0$. The largest coefficients are slightly thresholded : if $d_j^2 >> \lambda$, $d_j^2 / \left(d_j^2 + \lambda\right)$ is close to 1. The threshold decreases when j increases since d_j decreases.

We can give an interpretation in relation with the **Principal Components Analysis** . $\mathbf{X}$ being centered, $\mathbf{X}'\mathbf{X}/n$ is the empirical variance-covariance matrix of the column vectors of the matrix $\boldsymbol{X}$.

$$\mathbf{X}'\mathbf{X} = \mathbf{VD}'\mathbf{DV}',$$

where $\mathbf{D}'\mathbf{D}$ is the diagonal matrix composed by the elements d_i^2. We denote by $\boldsymbol{v_1}, \ldots, \boldsymbol{v_p}$ the column vectors in $\mathbb{R}^p$ of the matrix $\mathbf{V}$.

Let $\boldsymbol{v}$ be an $\mathbb{R}^p$ vector with norm 1.

$$Var(\mathbf{Xv}) = (\mathbf{Xv})'(\mathbf{Xv}) = \boldsymbol{v'}(\mathbf{X}'\mathbf{X})\boldsymbol{v},$$

which is maximal for $\boldsymbol{v} = \boldsymbol{v_1}$ and is equal to d_1^2.
$\boldsymbol{z}_1 = \mathbf{X}\boldsymbol{v_1}$ is the first principal component of the matrix $\mathbf{X}$.
The orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ are the principal directions (or Karhunen Loeve directions) of $\mathbf{X}$. The variables $\boldsymbol{z}_j = X\boldsymbol{v}_j$ are the principal components. We remark that

$$\boldsymbol{z}_j = X\boldsymbol{v}_j = \mathbf{UDV}'\boldsymbol{v}_j = d_j \boldsymbol{u}^{(j)}.$$

We see that the *ridge* regression shrinks slightly the first principal components (for which d_j is large), and more the last principal components.
We can associate to the *ridge* procedure the quantity $df(\lambda)$ which is called the effective number of degrees of freedom in the *ridge* regression and is defined by

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

If $\lambda = 0$, $df(\lambda) = p$ (no shrinkage), if $\lambda \to \infty$, $df(\lambda) \to 0$, at the limit, all the coefficients are equal to 0.

6 The LASSO regression

The *ridge* regression allows to get around the collinearity problems even if the numbers of predictors p is large with possibly $p > n$. The main weakness of this method is related to interpretation difficulties because, without selection, all variables are included in the model. Other regularization approaches also allow selection, as the LASSO regression, which leads to more interpretable solutions.

6.1 Model and estimation

LASSO is the abbreviation of Least Absolute Shrinkage and Selection Operator. The Lasso estimator is introduced in the paper by Tibshirani, R. (1996)[9]: Regression shrinkage and selection via the lasso. J. Royal. Statist.

Soc B., Vol. 58, No. 1, pages 267-288. The Lasso corresponds to the minimization of a least square criterion plus an l_1 penalty term (and no more an l_2 penalization like in the *ridge* regression). We denote $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.

DEFINITION 6. — *The Lasso estimator of $\boldsymbol{\beta}$ in the model*

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

is defined by :

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

where λ is a nonnegative tuning parameter.

We can show that this is equivalent to the minimization problem :

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

where t is suitably chosen, and $\widehat{\beta_0}_{\text{Lasso}} = \bar{Y}$. Like for the Ridge regression, the parameter λ is a regularization parameter:

- If $\lambda = 0$, we recover the least square estimator.
- If λ tends to infinity, all the coefficients $\hat{\beta}_j$ are equal to 0 for $j = 1, \ldots, p$.

The solution to the Lasso is parsimonious (or sparse), since it has many null coefficients.

If the matrix $\mathbf{X}$ is orthogonal : ($\mathbf{X}'\mathbf{X} = Id$), the solution is explicit.

PROPOSITION 7. — *If $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the solution $\boldsymbol{\beta}$ of the minimization of the Lasso criterion*

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1$$

is defined as follows : for all $j = 1, \ldots, p$,

$$\beta_j = sign(\widehat{\beta}_j)(|\widehat{\beta}_j| - \lambda)\mathbf{1}_{|\widehat{\beta}_j| \geq \lambda},$$

where $\widehat{\boldsymbol{\beta}}$ is the least square estimator : $\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$.

The obtained estimator corresponds to a soft thresholding of the least square estimator. The coefficients $\widehat{\beta}_j$ are replaced by $\phi_\lambda(\widehat{\beta}_j)$ where

$$\phi_\lambda : x \mapsto \text{sign}(x)(|x| - \lambda)_+.$$

Exercise. — Prove the proposition 7.

Another formulation

The LASSO is equivalent to the minimization of the criterion

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

under the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$, for some $t > 0$.

The statistical software R introduces a constraint expressed by a relative bound for $\sum_{j=1}^{p} |\beta_j|$: the constraint is expressed by

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

where $\hat{\beta}^{(0)}$ is the least square estimator and $\kappa \in [0, 1]$.

For $\kappa = 1$ we recover the least square estimator (there is no constraint) and for $\kappa = 0$, all the $\hat{\beta}_j$, $j \geq 1$, vanish (maximal constraint).

6.2 Applications

We represent in Figure 7 the values of the coefficients in function of κ for the Prostate cancer data: this are the regularization paths of the LASSO. As for the Ridge regression, the tuning parameter is generally calibrated by cross-validation.

Comparison LASSO/ RIDGE

The Figure 3.12 gives a geometric interpretation of the minimization problems for both the Ridge and Lasso estimators. This explains why the Lasso solution is sparse.

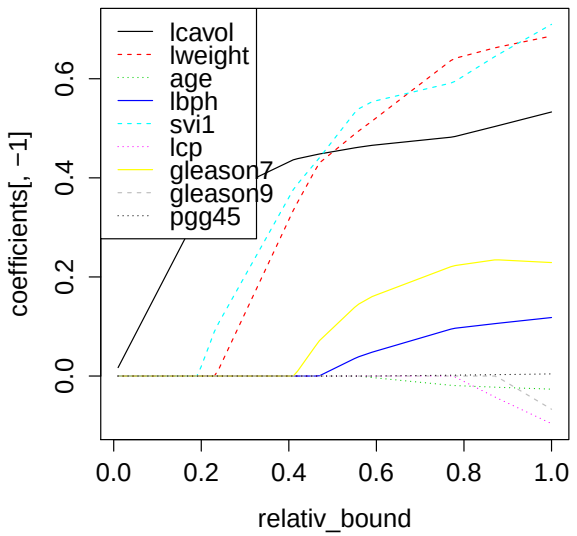

Figure 7: Regularization paths of the LASSO when the penalty decreases

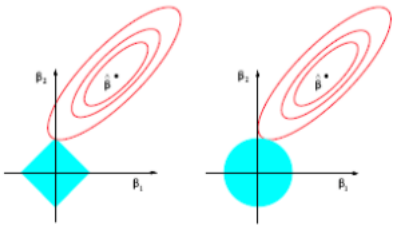

Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

6.3 Optimization algorithms for the LASSO

Convex functions and subgradients

DEFINITION 8. — *A function* $F : \mathbb{R}^n \to \mathbb{R}$ *is convex if* $\forall x, y \in \mathbb{R}^n, \forall \lambda \in [0,1]$,

$$F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y).$$

LEMMA 9. — *When* F *is differentiable in* x*, we have* $F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle \ \forall y \in \mathbb{R}^n$.

When F is non differentiable, we introduce the subdifferential ∂F of F defined by :

DEFINITION 10. — *The subdifferential* ∂F *of* F *is :*

$$\partial F(x) = \{\omega \in \mathbb{R}^n, F(y) \geq F(x) + \langle \omega, y - x \rangle, \forall y \in \mathbb{R}^n\}.$$

A vector $\omega \in \partial F(x)$ *is called a subgradient of* F *in* x.

LEMMA 11. — *F is convex* $\Leftrightarrow \partial F(x) \neq \emptyset \; \forall x \in \mathbb{R}^n$.

Example : subdifferential of the l_1 norm

$$\partial |x|_1 = \{\omega \in \mathbb{R}^n, \omega_j = 1 \text{ for } x_j > 0, \omega_j = -1 \text{ for } x_j < 0,$$

$$\omega_j \in [-1,1] \text{ for } x_j = 0\}.$$

Remark : The subdifferential of a convex function is monotone in the following sense :

$$\langle \omega_x - \omega_y, x - y \rangle \geq 0 \; \forall \omega_x \in \partial F(x), \forall \omega_y \in \partial F(y).$$

Indeed

$$\begin{aligned} F(y) &\geq F(x) + \langle \omega_x, y - x \rangle \\ F(x) &\geq F(y) + \langle \omega_y, x - y \rangle. \end{aligned}$$

By summing, $\langle \omega_x - \omega_y, x - y \rangle \geq 0$.

First optimality condition

PROPOSITION 12. — *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex function.*

$$x_* \in \textit{argmin}_{x \in \mathbb{R}^n} F(x) \Leftrightarrow 0 \in \partial F(x_*).$$

Proof : In both cases,

$$F(y) \geq F(x_*) + \langle 0, y - x_* \rangle.$$

The Lasso estimator

We consider the linear model :

$$Y = X\beta^* + \varepsilon.$$

We assume that the columns of X have norm 1. Let

$$L(\beta) = \|Y - X\beta\|^2 + \lambda |\beta|_1.$$

By definition, the Lasso estimator

$$\hat{\beta}_\lambda \in \text{argmin}_{\beta \in \mathbb{R}^p}(L(\beta)).$$

We deduce from the first order optimality condition that $0 \in \partial L(\hat{\beta}_\lambda)$.
We have that

$$L(\beta) = \|Y\|^2 - \beta^* X^* X \beta - 2\beta^* XY + \lambda |\beta|_1.$$

LEMMA 13. — *Let $h : \beta \mapsto \beta^* A \beta$ where A is a symmetric matrix. Then* $\bigtriangledown h(\beta) = 2A\beta$.

Let $g : \beta \mapsto \beta^* z = z^* \beta = \langle z, \beta \rangle$ where $z \in \mathbb{R}^p$. Then $\bigtriangledown g(\beta) = z$.

Hence we have

$$\partial L(\beta) = 2X^* X\beta - 2X^* Y + \lambda \partial |\beta|_1.$$

$$0 \in \partial L(\hat{\beta}_\lambda) \Leftrightarrow \exists \hat{z} \in \partial |\hat{\beta}_\lambda|_1 \text{ such that :}$$

$$2X^* X \hat{\beta}_\lambda - 2X^* Y + \lambda \hat{z} = 0.$$

This last equality is equivalent to

$$X^* X \hat{\beta}_\lambda = X^* Y - \frac{\lambda}{2} \hat{z} \qquad (E).$$

We have seen that

$$\begin{aligned} \hat{z}_j &= sign((\hat{\beta}_\lambda)_j \text{ if } (\hat{\beta}_\lambda)_j \neq 0 \\ \hat{z}_j &\quad \text{can be any real in } [-1,1] \text{ if } (\hat{\beta}_\lambda)_j = 0. \end{aligned}$$

Orthogonal setting

When $X^* X = I_p$, (E) gives $(\hat{\beta}_\lambda)_j = X_j^* Y - \frac{\lambda}{2} \hat{z}_j$.
Moreover, $\hat{z}_j = sign(\hat{\beta}_\lambda)_j$if $(\hat{\beta}_\lambda)_j \neq 0$. Hence,

$$\begin{cases} (\hat{\beta}_\lambda)_j > 0 \Rightarrow X_j^* Y > \frac{\lambda}{2} \\ (\hat{\beta}_\lambda)_j < 0 \Rightarrow X_j^* Y < -\frac{\lambda}{2}. \end{cases}.$$

$$(\hat{\beta}_\lambda)_j \neq 0 \Rightarrow \begin{cases} |X_j^* Y| > \frac{\lambda}{2} \\ sign((\hat{\beta}_\lambda)_j) = sign(X_j^* Y) \end{cases}.$$

This leads to the **explicit solution of the Lasso in the orthogonal setting**

$$(\hat{\beta}_\lambda)_j = sign(X_j^* Y)\left(|X_j^* Y| - \frac{\lambda}{2}\right)\mathbf{1}_{|X_j^* Y| > \frac{\lambda}{2}}.$$

It corresponds to a **soft thresholding** of the Ordinary Least Square estimator $\hat{\beta}_j = X_j^* Y$.

Non orthogonal setting

In this case, there is no analytic formula for the Lasso estimator $\hat{\beta}_\lambda$.
Let $\hat{m}_\lambda = \left\{j, (\hat{\beta}_\lambda)_j \neq 0\right\}$ be the support of $\hat{\beta}_\lambda$.
We can derive from Equation (E) that

- If $\lambda \geq 2\sup_j |X_j^* Y|$, then $\hat{\beta}_\lambda = 0$.
- If $\lambda < 2\sup_j |X_j^* Y|$, then denoting $X_{\hat{m}_\lambda}$ the submatrix obtained from X by keeping only the columns belonging to $\hat{m}_\lambda$, we have the following equation :

$$X_{\hat{m}_\lambda}^* X_{\hat{m}_\lambda} (\hat{\beta}_\lambda)_{\hat{m}_\lambda} = X_{\hat{m}_\lambda}^* Y - \frac{\lambda}{2} sign((\hat{\beta}_\lambda)_{\hat{m}_\lambda}).$$

Computing the Lasso estimator

$\beta \mapsto L(\beta) = \|Y - X\beta\|^2 + \lambda|\beta|_1$ is convex.
Hence a simple and efficient approach to minimize this function is to alternate minimization over each coordinate of β.
This algorithm converges to the Lasso estimator thanks to the convexity of L.
If we assume that the columns of X have norm 1, then we have

$$\frac{\partial R}{\partial \beta_j}(\beta) = -2X_j^*(Y - X\beta) + \lambda \frac{\beta_j}{|\beta_j|}, \quad \forall \beta_j \neq 0.$$

Hence, we can see (after some easy computations) that $\beta_j \mapsto R(\beta_1, \ldots, \beta_{j-1}, \beta_j, \ldots, \beta_p)$ is minimum in

$$\beta_j = R_j \left(1 - \frac{\lambda}{2|R_j|}\right)_+$$

with $R_j = X_j^*(Y - \sum_{k \neq j} \beta_k X_k)$.

The coordinate descent algorithm is summarized as follows :

- Initialise β_{init} arbitrarily
- Iterate until convergence :

$$\forall j = 1, \ldots, p, \beta_j = R_j \left(1 - \frac{\lambda}{2|R_j|}\right)_+$$

  with $R_j = X_j^*(Y - \sum_{k \neq j} \beta_k X_k)$.

- Output β.

This algorithm is implemented in the R package `glmnet`.

Due to its parsimonious solution, this method is widely used to select variables in high dimension settings (when $p > n$). Verzelen (2012)[10] has shown important results on the limitation of the Lasso in ultra high dimension.

7 Elastic Net

Elastic Net is a method that combines Ridge and Lasso regression, by introducing simultaneously the l_1 and l_2 penalties. The criterion to minimize is

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

- For $\alpha = 1$, we recover the LASSO.
- For $\alpha = 0$, we recover the Ridge regression.

In this case, we have two tuning parameters to calibrate by cross-validation.

8 Principal Components Regression and Partial Least Square regression

8.1 Principal Component Regression (PCR)

We denote by $Z^{(1)}, \dots Z^{(p)}$ the principal components associated to the variables $X^{(1)}, \dots X^{(p)}$:

- $Z^{(1)}$ is the linear combination of $X^{(1)}, \dots, X^{(p)}$ of the form $\sum_{i=1}^{p} \alpha_j X^{(j)}$ with $\sum \alpha_j^2 = 1$ with maximal variance.
- $Z^{(m)}$ is the linear combination of $X^{(1)}, \dots, X^{(p)}$ of the form $\sum_{i=1}^{p} \alpha_{j,m} X^{(j)}$ with $\sum \alpha_{j,m}^2 = 1$ with maximal variance and orthogonal to $Z^{(1)}, \dots, Z^{(m-1)}$.

The Principal Component Regression (PCR) consists in considering a predictor of the form :

$$\hat{Y}^{PCR} = \sum_{m=1}^{M} \hat{\theta}_m Z^{(m)}$$

with

$$\hat{\theta}_m = \frac{\langle Z^{(m)}, Y \rangle}{\|Z^{(m)}\|^2}.$$

Comments :

- If $M = p$, we keep all the variables and we recover the ordinary least square (OLS) estimator.
- If one can obtain a good prediction with $M < p$, then we have reduced the number of variables, hence the dimension.
- Nevertheless, interpretation is not always easy : if the variables are interpretable, the principal components (that correspond to linear combination of the variables) are generally difficult to interpret.
- This method is quite similar to the Ridge regression, which shrinks the coefficients of the principal components. Here, we set to 0 the coefficients of the principal components of order greater than M.
- The first principal components are not necessarily well correlated with the variable to explain Y, this is the reason why the PLS regression has been introduced.

8.2 Partial Least Square (PLS) regression

Le principle of this method is to make a regression on linear combinations of the variables X_i's, that are highly correlated with Y.

- We assume that Y has been centered, and that the variables $X^{(j)}$ are also centered and normalized (with norm 1).
- The first PLS component is defined by :

$$W^{(1)} = \sum_{j=1}^{p} \langle Y, X^{(j)} \rangle X^{(j)}.$$

- The prediction associated to this first component is :

$$\hat{Y}^1 = \frac{\langle Y, W^{(1)} \rangle}{\|W^{(1)}\|^2} W^{(1)}.$$

Note that if the matrix X is orthogonal, this estimator corresponds to the ordinary least square (OLS) estimator, and in this case, the following steps of the PLS regression are useless.

- In order to obtain the following directions, we orthogonalize the variables $X^{(j)}$ with respect to the first PLS component $W^{(1)}$:

- We substract to each variables $X^{(j)}$ $(1 \leq j \leq p)$ its orthogonal projection in the direction given by $W^{(1)}$ and we normalize the variables thus obtained.
- We compute the second PLS component $W^{(2)}$ in the same way as the first component by replacing the variables $X^{(j)}$'s by the new variables.
- We iterate this process by orthogonalizing at each step the variables with respect to the PLS components.

The algorithm is the following :

- $\hat{Y}^0 = \bar{Y}$ and $X^{(j),0} = X^{(j)}$. For $m = 1, \ldots, p$
- $W^{(m)} = \sum_{j=1}^{p} \langle Y, X^{(j,m-1)} \rangle X^{(j,m-1)}$.
- $\hat{Y}^m = \hat{Y}^{m-1} + \frac{\langle Y, W^{(m)} \rangle}{\|W^{(m)}\|^2} W^{(m)}$.
- $\forall j = 1, \ldots, p$, $X^{(j),m} = \frac{X^{(j),m-1} - \Pi_{W^{(m)}}(X^{(j),m-1})}{\|X^{(j),m-1} - \Pi_{W^{(m)}}(X^{(j),m-1})\|}$.
- The predictor $\hat{Y}^p$ obtained at step p corresponds to ordinary least square estimator.
- This method is useless if the variables $X^{(j)}$ are orthogonal.
- When the variables $X^{(j)}$ are correlated, PCR and PLS methods present the advantage to deal with new variables, that are orthogonal.
- The choice of the number of PCR or PLS components can be done by cross-validation.
- In general, the PLS method leads to more parcimoneous representations than the PCR method.
- The PLS regression leads to a **reduction of the dimension**.
- If p is large, this is particularly interesting, but can lead to problems of interpretation since the PLS components are linear combinations of the variables.

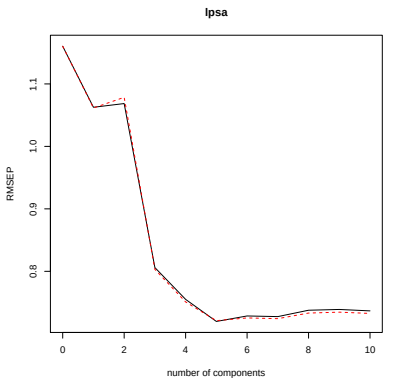

Figure 8: Estimation of the error by cross-validation in function of the number of PLS components

- There exists a sparse version : **sparse PLS** (inspired from the Lasso method), for which we consider linear combinations of the initial variables $X^{(j)}$ with only a few non zero coefficients, hence keeping only a few variables, which makes the interpretation more easy.

Application of the PLS regression

Application to the prostate cancer data.

```
   Data:
X dimension:  97 10
Y dimension:  97 1
Number of components considered:  10


   VALIDATION: RMSEP
Cross-validated using 10 random segments.
          (Intercept)    1 comps    2 comps    3 comps    4 comps    5 comps    6 comps    7 comps
   CV           1.160      1.066      1.031     0.8344     0.7683     0.7501     0.7634     0.7672
adjCV           1.160      1.064      1.075     0.8291     0.7641     0.7451     0.7584     0.7618
          8 comps    9 comps    10 comps
   CV      0.7704     0.7758      0.7730
adjCV      0.7641     0.7695      0.7671
TRAINING: % variance explained
      1 comps    2 comps    3 comps    4 comps    5 comps    6 comps    7 comps    8 comps
   X    93.47      98.49      99.66      99.75      99.93      99.96      99.97      99.98
lpsa    18.20      26.98      57.02      63.09      64.38      66.15      66.46      66.59
      9 comps    10 comps
   X    100.0      100.0
lpsa     66.6       66.6
```

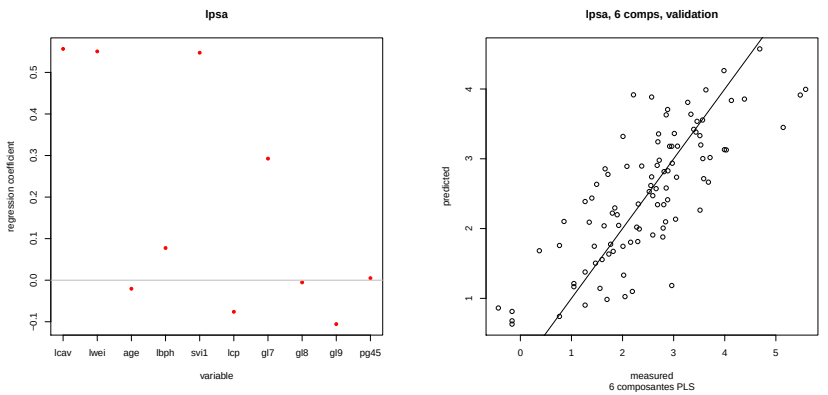

Figure 9: Model with 5 PLS components: estimated coefficients and predicted values in function of the observations

References

[1] P.J. Brown, T. Fearn, and M. Vannucci. Bayesian wavelet regression on curves with applications to a spectroscopic calibration problem. *Journal of the American Statistical Society*, 96:398–408, 2001.

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

[3] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[4] T. Hastie, R. Tibshirani, and J Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2009. Second edition.

[5] Nicole Krr, Anne-Laure Boulesteix, and Gerhard Tutz. Penalized partial least squares with applications to b-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94:60–69, 2008.

[6] C.L. Mallows. Some comments on cp. *Technometrics*, 15:661–675, 1973.

[7] B. G. Osborne, T. Fearn, A. R. Miller, and S. Douglas. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.*, 35:99–105, 1984.

[8] M. Stone and 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*, 52:237–269, 1990.

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

[10] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Statistics*, 6:38–90, 2012.