



Institut de Mathématiques de Toulouse, INSA Toulouse

Supervised Learning- Part I

Data Mining
February 2021

Béatrice Laurent-Philippe Besse

Introduction

In the framework of **Supervised learning**, we have a **Learning sample** composed with observation data of the type **input/output** :

$$d_1^n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$$

with $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \mathcal{Y}$ for $i = 1 \dots n$.

Objectives : From the learning sample, we want to

- **Estimate** the link between the input vector \mathbf{x} (explanatory variables) and the output y (variable to explain) :

$$y = f(x^1, x^2, \dots, x^p)$$

- **Predict** the output y associated to a new entry \mathbf{x} ,
- **Select** the important explanatory variables among x^1, \dots, x^p .

Introduction

quantitative output

$$\mathcal{Y} \subset \mathbb{R}$$



real regression

qualitative output

\mathcal{Y} finite



classification

The explanatory variables X^1, \dots, X^p can be **qualitatives** or **quantitatives**

Choice of the model

- Importance of the principle of **parcimony** : "it is necessary to determine a model that provides an adequate representation of the data, with as few parameters as possible".
- Bias-variance **trade-off**
- **Robustness**

Methods studied in this course :

Part I : linear methods for regression and classification

- Estimation of a prediction error
- Linear model, model selection, variable selection, Ridge regression, Lasso.
- Logistic regression
- Support Vector Machine

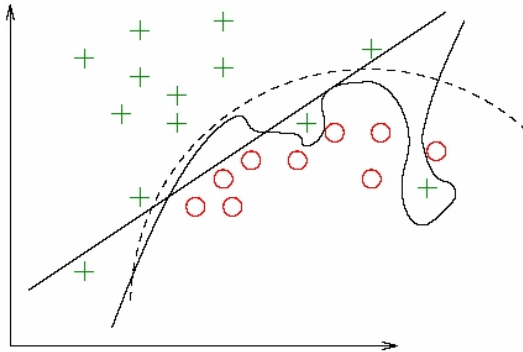
Part II : nonlinear algorithms

- Classification And Regression Trees (CART)
- Bagging, Random Forests, Boosting
- Neural networks, Introduction to deep learning
- Algorithms for Anomaly detection

Part I-0 : Estimation of a prediction error

- The **generalization** performance of a learning procedure is related to its prediction capacity on a **new data set**, independent of the learning sample that was used to build the learning algorithm.
- Evaluating this performance is crucial to choose a learning method or model among several possible ones.
- It is also important to measure the quality of the ultimately chosen procedure.

Bias, variance, Model complexity



Supervised Classification : Complexity of the models

Loss functions

- In a **regression framework**, one generally considers the \mathbb{L}_2 -loss :

$$\ell(Y, \hat{f}(\mathbf{X})) = (Y - \hat{f}(\mathbf{X}))^2,$$

or the \mathbb{L}_p -loss ($p \geq 1$) :

$$\ell(Y, \hat{f}(\mathbf{X})) = |Y - \hat{f}(\mathbf{X})|^p.$$

- In **supervised classification**, generally one considers the 0-1 loss :

$$\ell(Y, \hat{f}(\mathbf{X})) = \mathbb{1}_{Y \neq \hat{f}(\mathbf{X})}.$$

Training error

- Given a loss function ℓ , the *training error* of a prediction rule \hat{f} is defined by

$$\overline{err} = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \hat{f}(\mathbf{X}_i)),$$

where $(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$ is the learning (or training) sample, used to train the algorithm \hat{f} .

Evaluation of the risk - generalization error

- Given a loss function ℓ , the *risk* - or *generalization error* - of a prediction rule \hat{f} is defined by

$$R_P(\hat{f}) = \mathbb{E}_{(\mathbf{X}, Y) \sim P}[\ell(Y, \hat{f}(\mathbf{X}))].$$

- Note that, in the above definition, (\mathbf{X}, Y) is **independent of the training sample** $D^n = (\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$ that was used to build the prediction rule \hat{f} .

Evaluation of the risk - generalization error

- It is important to estimate the generalization error of a learning algorithm \hat{f} : when the model becomes more and more complex, it is able to capture more complex underlying structures in the "true " model : **the bias decreases**, but at the same time, the estimation error increases, due to **the increase of the variance**.
- The "optimal" model is the one realizing **the best compromise between the bias term and the variance term** to give the smallest generalization error.
- The training error is not a good estimate of the generalization error : it decreases as the complexity of the model increases. Hence minimizing the training error leads to select the most complex model, this leads to **overfitting**.

Training error in a regression model

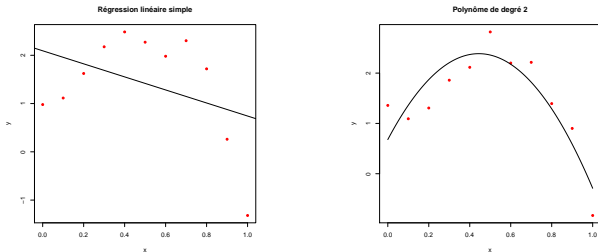


FIGURE – Polynomial regression : adjusted model, on the left : $y = \beta_0 + \beta_1 x + \epsilon$, $R^2 = 0.03$, on the right : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, $R^2 = 0.73$.

In a regression model, with the \mathbb{L}_2 loss, the training error is equal to the

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{SSR}{n}.$$

Training error in a regression mode

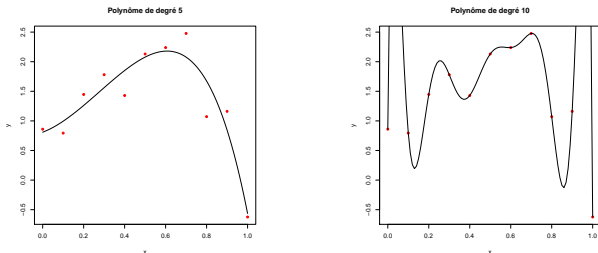


FIGURE – Polynomial regression : adjusted model, on the left :

$y = \beta_0 + \beta_1x + \dots + \beta_5x^5 + \epsilon$, $R^2 = 0.874$, on the right :

$y = \beta_0 + \beta_1x + \dots + \beta_{10}x^{10} + \epsilon$, $R^2 = 1$.

The training error is equal to 0 for the polynomial of degree $n - 1$ (which has n coefficients) and passes through all the training points.

Training error and test error

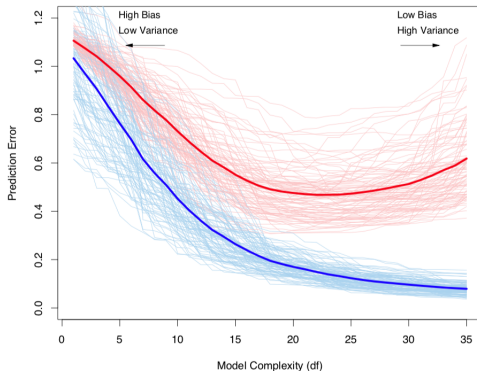


FIGURE – Behavior of training error (in blue) and test error (in red) as the complexity of the model increases. Source : "The elements of Statistical Learning", T. Hastie, R. Tibshirani, J. Friedman.

Evaluation of the risk - generalization error

- An accurate evaluation of the generalization error has two objectives :
 - **Model selection** : selecting, among a collection of models (or prediction rules), the one with the smallest risk, realizing the best bias/variance trade-off.
 - **Model assessment** : Once the final model has been chosen, evaluating its generalization error on a **new data set**.

Evaluation of the risk - generalization error

- If we have enough data, the recommended approach is to divide randomly the dataset in three parts : the learning sample, the validation sample, and the test sample.
 - **The learning sample** is used to train the models (generally by minimizing the training error).
 - **The validation sample** is used for model selection : we estimate the generalization error of each model with the validation sample and we select the model with the smallest generalization error.
 - **The test sample** is used to evaluate the risk of the final selected model.
- It is generally recommended to take 50% of the data for the learning sample, 25% of the data for the validation sample and 25% of the data for the test sample.

Cross-validation

- Often, taking only 50% of the data set to train the models may lead to bad performances, especially if we do not have too much data.
- In this case, it is recommended to gather the learning and validation samples and to use **K fold cross-validation** to estimate the generalization error.
- We split randomly the data into K subsamples, with (almost) the same size. ($K = 10$ generally).
- Each of the K folds will be successively used as a test sample.
- When the fold k is the test sample, we train a model with the $K - 1$ other folds, and we evaluate the loss function of this model on each element the fold k .
- This is done for $k = 1, \dots, K$, and we compute a global estimation of the generalization error.

Cross-validation

- More precisely, assume that we have a n -sample $(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$ and a collection of models $(\hat{f}_m, m \in \mathcal{M})$. We split the data into K folds.
- Let for all $i \in \{1, \dots, n\}$, $\tau(i) \in \{1, \dots, K\}$ denote the number of the fold containing the observation (\mathbf{X}_i, Y_i) .
- Let $\hat{f}_m^{(-k)}$ denote the model m trained with all the data, except the fold k .
- The cross-validation estimate of the generalization error of the model m is

$$CV(m) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \hat{f}_m^{(-\tau(i))}(\mathbf{X}_i)).$$

- $CV(m)$ estimates the generalization error of the model m and we select the model which minimizes $CV(m)$.
- The selected model is then fitted with all the data.

Cross-validation

- When the number of folds $K = n$, the method is called **leave-one-out** cross-validation.
- This method has a low bias to estimate the generalization error, but a high variance since all the estimators $\hat{f}_m^{(-i)}$ are highly correlated.
- The computation time is also high for the **leave-one-out** method.
- The choice of $K = 5$ or 10 is often recommended.

In the next two sections, we will see two methods for supervised learning : the linear model for regression and the logistic model for classification.

Part I-1 :

- Linear model, model selection, variable selection, penalized criterion.
 - Linear model
 - Least square estimation
 - Confidence intervals and prediction intervals
 - Testing a submodel
 - Determination coefficient, Diagnosis on the residuals
 - Model selection, variable selection, penalized criterion

The Linear model

We have a quantitative variable Y *to explain* which is related with p variables $\mathbf{X}^1, \dots, \mathbf{X}^p$ called *explanatory variables*.

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

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

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 variables $\mathbf{X}^1, \dots, \mathbf{X}^p$.

The Linear model

The linear model is defined by :

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

with the following assumptions :

- 1 The random variables ε_i are independent and identically distributed (i.i.d.) ; they are independent of $(\mathbf{X}^1, \dots, \mathbf{X}^p)$, $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$.
- 2 We have $E(\mathbf{Y}|\mathbf{X}^1, \dots, \mathbf{X}^p) = \beta_0 + \beta_1 \mathbf{X}^1 + \beta_2 \mathbf{X}^2 + \cdots + \beta_p \mathbf{X}^p$ and $\text{Var}(\mathbf{Y}|\mathbf{X}^1, \dots, \mathbf{X}^p) = \sigma^2$.
- 3 The unknown parameters β_0, \dots, β_p are supposed to be constant.
- 4 It is sometimes assumed that the errors are Gaussian : $\varepsilon = [\varepsilon_1 \cdots \varepsilon_n]' \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. The variables ε_i are then i.i.d. $\mathcal{N}(0, \sigma^2)$.

The Linear model

- The explanatory variables are given in the matrix $\mathbf{X}(n \times (p + 1))$.
- The regressors \mathbf{X}^j can be quantitative variables, nonlinear transformation of quantitative variables (such as log, exp, square ..), interactions : $\mathbf{X}^j = \mathbf{X}^k \cdot \mathbf{X}^l$.
- They can also correspond to qualitative variables : in this case the variables \mathbf{X}^j are indicator variables coding the different levels of a factor.
- The response variable is given in the vector \mathbf{Y} .
- We set $\beta = [\beta_0 \ \beta_1 \ \cdots \ \beta_p]'$, which leads to the matricial formulation of the linear model :

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

Example

We consider the **Ozone data set** .

The data frame has 1041 observations of the following components :

JOUR	type of the day ; public holiday(1) or not (0)
O3obs	Ozone concentration observed the next day at 17h., generally the maximum of the day
MOCAGE	Prediction of this pollution obtained by a deterministic model of fluid mechanics
TEMPE	Temperature forecast by MétéoFrance for the next day 17h
RMH2O	Moisture ratio
NO2	Nitrogen dioxide concentration
NO	Concentration of nitric oxide
STATION	Location of the observation : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache and Plan de Cuques
VentMOD	Wind force
VentANG	Orientation of the wind.

- We denote by Y the variable (**O3obs**) to explain.
- We set X^1, \dots, X^p for the explanatory variables (**MOCAGE** , **TEMPE**, **JOUR** ..). The variables are quantitative (**MOCAGE** , **TEMPE** , ...), or qualitative (**JOUR**, **STATION**).
- We consider the linear model :

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

- 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 are set in the reference.
This is an analysis of covariance model (mixing quantitative and qualitative variables).

Least square estimation

- The unknown parameters of the model are the vector β and σ^2 .
- β is estimated by **minimizing the residuals sum of square**.
- We minimise with respect to the parameter $\beta \in \mathbb{R}^{p+1}$ the criterion :

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

Lemma

Let $h : \beta \mapsto \beta' A \beta$ where A is a symmetric matrix.

Then $\nabla h(\beta) = 2A\beta$.

Let $g : \beta \mapsto \beta' z = z' \beta = \langle z, \beta \rangle$ where $z \in \mathbb{R}^p$.

Then $\nabla g(\beta) = z$.

Least square estimation

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

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

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

Least square estimation

We make the additional assumption that the matrix $\mathbf{X}'\mathbf{X}$ is invertible. Under this assumption, the estimation of β is given by :

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

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

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the "*hat matrix*".

Geometrically, it corresponds to the matrix of orthogonal projection in \mathbb{R}^n onto the subspace $\text{Vect}(\mathbf{X})$ generated by the columns of \mathbf{X} .

Least square estimation

- If $\mathbf{X}'\mathbf{X}$ is not invertible, the application $\beta \mapsto \mathbf{X}\beta$ is not injective, hence the model is not identifiable and β is not uniquely defined.
- In this case, the predicted values $\hat{\mathbf{Y}}$ are still defined as the projection of \mathbf{Y} onto the space generated by the columns of \mathbf{X} .
- In practice, if $\mathbf{X}'\mathbf{X}$ is not invertible (which is necessarily the case in high dimension when $p > n$), we have to remove variables from the model or to consider other approaches to reduce the dimension (*Ridge*, *Lasso*, *PLS* ...).

Least square estimation

- We define the vector of residuals as :

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

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

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

Properties of the least square estimator

THEOREM

— 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 $\hat{\boldsymbol{\beta}}$ is a Gaussian vector :

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

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

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

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

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

Confidence intervals

One can easily deduce from the first theorem that

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

This allows to build **confidence intervals** and **tests of significance** for the parameters β_j .

The following interval is a 0.95 **confidence interval** for β_j :

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

In order to test $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, we reject the null hypothesis at the level 5% if 0 does not belong to the previous confidence interval.

Test of significance

- We recall the linear model

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

- We want to test if the variable X^j is significant in the model or not, which is equivalent to test the nullity of the parameter β_j .
- We test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$.
- Under the hypothesis H_0 ,

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{j,j}}} \sim \mathcal{T}_{(n-(p+1))}.$$

Test of significance

- The p-value of the test is defined as

$$\mathbb{P}_{H_0}(|T_j| > |T_j|_{obs}) = \mathbb{P}(|\mathcal{T}_{(n-(p+1))}| > |T_j|_{obs}),$$

where $|T_j|_{obs}$ is the observed value for the variable $|T_j|$ with our data.

- If the p-value is very small, then it is unlikely that $|T_j|_{obs}$ is obtained from a Student distribution with $n - (p + 1)$ degrees of freedom, hence we will reject the hypothesis H_0 , and conclude that the variable X^j is significant.
- We fix some level α (generally 5%) for the test.
- If $\text{p-value} < \alpha$, we reject the nullity of β_j and conclude that the variable X^j is significant in the model.
- One easily prove that the probability to reject H_0 when it is true (i.e. to conclude that the variable X^j is significant when it is not) is less than the level α of the test.

Example

We consider the **Ozone data set** .

The data frame has 1041 observations of the following components :

JOUR	type of the day ; public holiday(1) or not (0)
O3obs	Ozone concentration observed the next day at 17h., generally the maximum of the day
MOCAGE	Prediction of this pollution obtained by a deterministic model of fluid mechanics
TEMPE	Temperature forecast by MétéoFrance for the next day 17h
RMH2O	Moisture ratio
NO2	Nitrogen dioxide concentration
NO	Concentration of nitric oxide
STATION	Location of the observation : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache and Plan de Cuques
VentMOD	Wind force
VentANG	Orientation of the wind.

We first consider a simple linear regression model with the single variable $X = \text{MOCAGE}$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

For the least square estimation, we obtain the following results :

Coefficients	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.78887	3.42998	11.02	<2e-16 ***
MOCAGE	0.61006	0.02573	23.71	<2e-16 ***

Residual standard error : 33.04 on 1039 degrees of freedom

Multiple R-squared : 0.3511, Adjusted R-squared : 0.3505

F-statistic : 562.1 on 1 and 1039 DF, p-value : < 2.2e-16

We consider here a linear regression model with all the variables :

$$Y_i = \beta_0 + \beta_1 X_i^1 + \dots + \beta_p X_i^p + \varepsilon_i, \quad i = 1, \dots, n.$$

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)	-33.43948	6.98313	-4.789	1.93e-06 ****
JOUR1	0.46159	1.88646	0.245	0.806747
MOCAGE	0.37509	0.03694	10.153	< 2e-16 ***
TEMPE	3.96507	0.22135	17.913	< 2e-16 ***
...

Residual standard error : 27.83 on 1028 degrees of freedom

Multiple R-squared : 0.5445, Adjusted R-squared : 0.5391

F-statistic : 102.4 on 12 and 1028 DF, p-value : < 2.2e-16

Prediction

As mentioned above, the vector of predicted values is

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

Based on the n previous observations, we may be interested with the prediction of the response of the model for a new point

$\mathbf{X}_0' = (1, X_0^1, \dots, X_0^p)$:

$$Y_0 = \beta_0 + \beta_1 X_0^1 + \beta_2 X_0^2 + \dots + \beta_p X_0^p + \varepsilon_0,$$

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

The predicted value is

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0^1 + \dots + \hat{\beta}_p X_0^p = \mathbf{X}_0' \hat{\boldsymbol{\beta}}.$$

Prediction

- We derive from Theorem 1 that

$$\mathbb{E}(\hat{Y}_0) = \mathbf{X}_0' \boldsymbol{\beta} = \beta_0 + \beta_1 X_0^1 + \beta_2 X_0^2 + \dots + \beta_p X_0^p$$

and that $\hat{Y}_0 \sim \mathcal{N}(\mathbf{X}_0' \boldsymbol{\beta}, \sigma^2 \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0)$. Let $t = t_{n-(p+1), 0.975}$.

- **Confidence interval for the mean response $\mathbf{X}_0' \boldsymbol{\beta}$ at the new observation point \mathbf{X}_0 :**

$$\left[\mathbf{X}_0' \hat{\boldsymbol{\beta}} - t \hat{\sigma} \sqrt{\mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0}, \mathbf{X}_0' \hat{\boldsymbol{\beta}} + t \hat{\sigma} \sqrt{\mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0} \right].$$

- **Prediction interval for the response Y_0 at the new observation point \mathbf{X}_0 is :**

$$\left[\mathbf{X}_0' \hat{\boldsymbol{\beta}} - t \hat{\sigma} \sqrt{1 + \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0}, \mathbf{X}_0' \hat{\boldsymbol{\beta}} + t \hat{\sigma} \sqrt{1 + \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0} \right].$$

We consider here a simple linear regression model with the single variable $X = \text{MOCAGE}$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

For the least square estimation, we obtain the following results :

Coefficients	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.78887	3.42998	11.02	<2e-16 ***
MOCAGE	0.61006	0.02573	23.71	<2e-16 ***

Residual standard error : 33.04 on 1039 degrees of freedom

Multiple R-squared : 0.3511, Adjusted R-squared : 0.3505

F-statistic : 562.1 on 1 and 1039 DF, p-value : < 2.2e-16

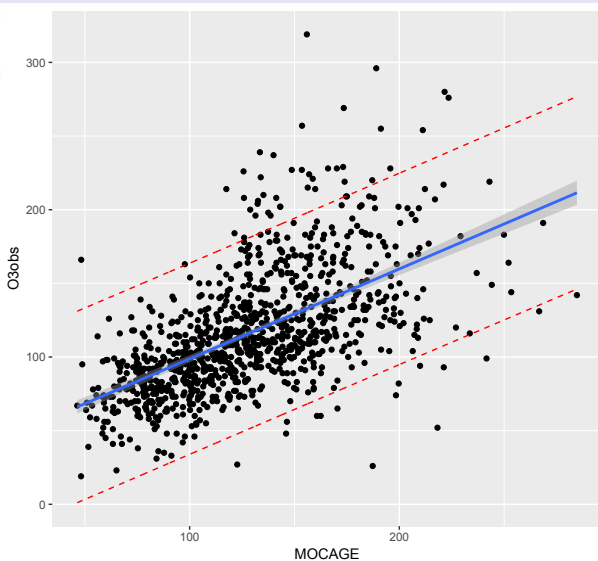


FIGURE – Simple linear regression model : confidence and prediction intervals

We consider here a linear regression model with all the variables :

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)} + \varepsilon_i, \quad i = 1, \dots, n.$$

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)	-33.43948	6.98313	-4.789	1.93e-06 ****
JOUR1	0.46159	1.88646	0.245	0.806747
MOCAGE	0.37509	0.03694	10.153	< 2e-16 ***
TEMPE	3.96507	0.22135	17.913	< 2e-16 ***
...

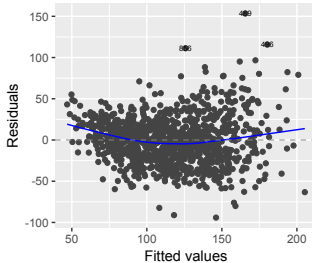
Residual standard error : 27.83 on 1028 degrees of freedom

Multiple R-squared : 0.5445, Adjusted R-squared : 0.5391

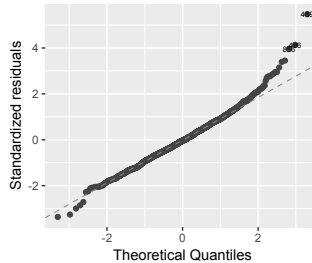
F-statistic : 102.4 on 12 and 1028 DF, p-value : < 2.2e-16

Diagnosis on the residuals

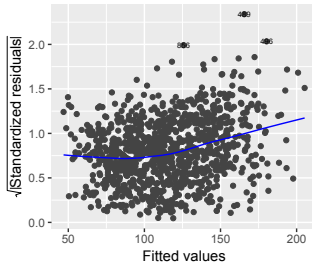
Residuals vs Fitted



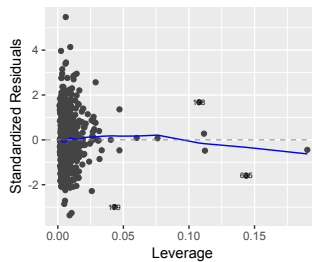
Normal Q-Q



Scale-Location



Residuals vs Leverage



Measures for goodness-of-fit

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

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

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

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}\|^2,$$

hence

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

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

Note that $0 \leq R^2 \leq 1$.

Determination coefficient and Model selection

The model is well adjusted to the n training data 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.

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 :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_k X_i^k + \varepsilon_i.$$

When k increases, the model is more and more complex, hence

$\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ decreases, and R^2 increases as shown in Figures 5 and 6.

Determination coefficient and Model selection

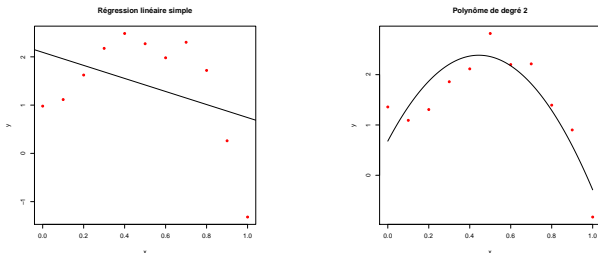


FIGURE – Polynomial regression : adjusted model, on the left :

$y = \beta_0 + \beta_1 x + \epsilon$, $R^2 = 0.03$, on the right : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, $R^2 = 0.73$.

Determination coefficient and Model selection

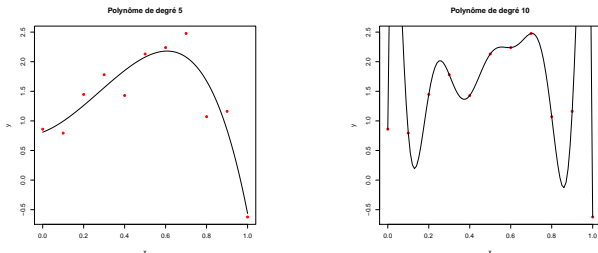


FIGURE – Polynomial regression : adjusted model, on the left :

$y = \beta_0 + \beta_1x + \dots + \beta_5x^5 + \epsilon$, $R^2 = 0.874$, on the right :

$y = \beta_0 + \beta_1x + \dots + \beta_{10}x^{10} + \epsilon$, $R^2 = 1$.

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.

Model selection

- The best model is the one that realizes the best trade-off between the bias term and the variance term.
- 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 :

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

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.

Ozone data

We first consider a simple linear regression model with the single variable $X = \text{MOCAGE}$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

For the least square estimation, we obtain the following results :

Coefficients	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.78887	3.42998	11.02	<2e-16 ***
MOCAGE	0.61006	0.02573	23.71	<2e-16 ***

Residual standard error : 33.04 on 1039 degrees of freedom

Multiple R-squared : 0.3511, Adjusted R-squared : 0.3505

F-statistic : 562.1 on 1 and 1039 DF, p-value : < 2.2e-16

Ozone data

We consider here a linear regression model with all the variables :

$$Y_i = \beta_0 + \beta_1 X_i^1 + \dots + \beta_p X_i^p + \varepsilon_i, \quad i = 1, \dots, n.$$

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)	-33.43948	6.98313	-4.789	1.93e-06 ****
JOUR1	0.46159	1.88646	0.245	0.806747
MOCAGE	0.37509	0.03694	10.153	< 2e-16 ***
TEMPE	3.96507	0.22135	17.913	< 2e-16 ***
...

Residual standard error : 27.83 on 1028 degrees of freedom

Multiple R-squared : 0.5445, Adjusted R-squared : 0.5391

F-statistic : 102.4 on 12 and 1028 DF, p-value : < 2.2e-16

Model selection

- 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. 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)
 - Adding some constraints on the parameters of the model by *shrinking* them (*Ridge* or *Lasso* penalization)

Variable selection

- 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".
- Maximizing the R^2 is not a good criterion since this 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 Mallows'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.

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

Variable selection

The Mallows'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 the next Figure.

Variable selection

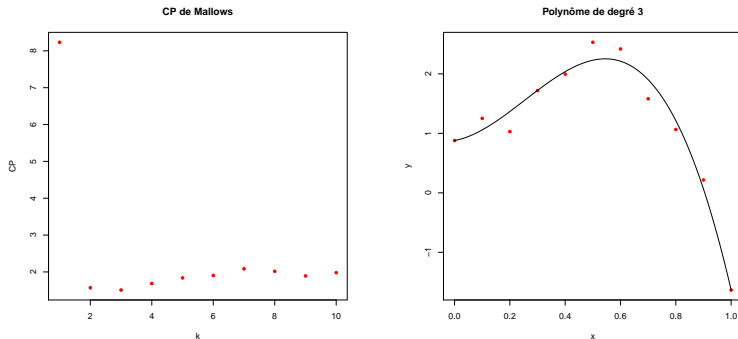


FIGURE – Mallows' C_p in function of the degree of the polynomial. Selected model : polynomial with degree 3.

Variable selection

- 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 :
 - **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.

All those algorithms stop when the criterion can no more be reduced.

Variable selection

Applications of the **Stepwise Algorithm** to the **Ozone** data. We apply the StepAIC algorithm, with the option **both** of the software R in order to select a subset of variables, and we present here an intermediate result :

```
Start: AIC=6953.05
O3obs ~ MOCAGE + TEMPE + RMH20 + NO2 + NO + VentMOD + VentANG
      Df Sum of Sq  RSS   AIC
- VentMOD  1    1484  817158 6952.9
<none>                 815674 6953.0
- RMH20    1    4562  8202354 6956.9
- VentANG  1   12115  827788  6966.4
- NO2      1   21348  837022  6977.9
- NO       1   21504  837178  6978.1
- MOCAGE   1   225453  1041127  7205.1
- TEMPE    1   268977  1084651  7247.7
Step: AIC= 6952.94
O3obs ~ MOCAGE + TEMPE + RMH20 + NO2 + NO + VentANG
```

Ridge regression

The principle of the Ridge regression is

- to consider all the explanatory variables
- 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 β .

Model and estimation

We consider the linear model

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

where

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1 & X_1^1 & X_1^2 & \cdot & X_1^p \\ 1 & X_2^1 & X_2^2 & \cdot & X_2^p \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_n^1 & X_n^2 & \cdot & X_n^p \end{pmatrix},$$
$$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix}.$$

\mathbf{X} is the matrix $\tilde{\mathbf{X}}$ 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 (note that the parameter β_0 is not penalized).

Definition

The *ridge* estimator of $\tilde{\beta}$ in the model $\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\beta} + \epsilon$, is defined by

$$\hat{\beta}_{\text{Ridge}} = \operatorname{argmin}_{\tilde{\beta} \in \mathbb{R}^{p+1}} \left(\|\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\beta}\|^2 + \lambda \sum_{j=1}^p \beta_j^2 \right),$$

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

Assume that \mathbf{X} and \mathbf{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

$$\hat{\beta}_0 = \bar{Y}, \hat{\beta}_R = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}.$$

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

Remarks :

- ① $\mathbf{X}'\mathbf{X}$ is a nonnegative symmetric matrix. 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 $\hat{\beta}_0 = \bar{\mathbf{Y}}$, adding a constant to \mathbf{Y} does not modify the values of $\hat{\beta}_j$ for $j \geq 1$.
- ③ The *ridge* estimator is not invariant by normalization of the vectors $\mathbf{X}^{(j)}$, 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 β is not too large : $\hat{\beta}_R = \arg \min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 ; \|\beta\|^2 < c \right\}$. 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 next Figure, 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.

Choice of the penalty term

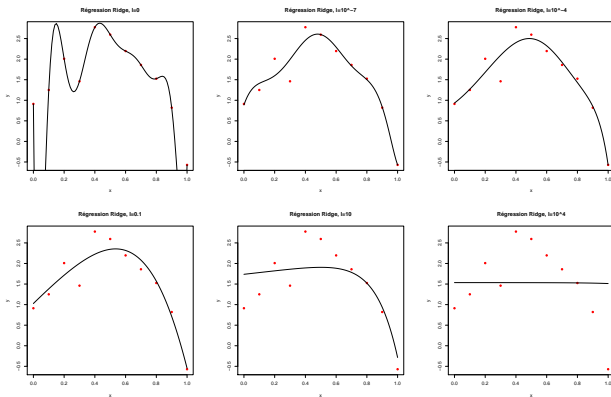
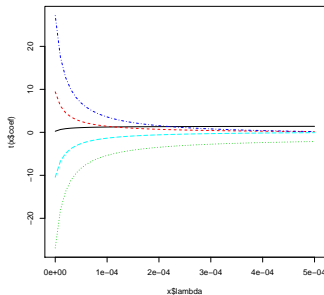
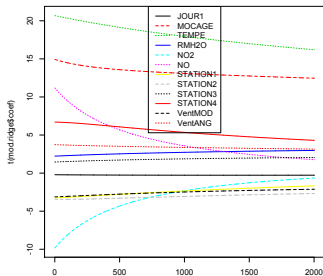


FIGURE – Ridge penalisation for the polynomial model

Choice of the penalty term

- For each regularization method, the choice of the parameter λ is determinant for the model selection. We see in next Figure 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 principle is the following :

- We split the data into K sub-samples. For all l from 1 to K :
 - We compute the Ridge estimator associated to a regularization parameter λ from the data of all the subsamples, except the l -th (that will be a "test" sample).
 - We denote by $\hat{\beta}_{\lambda}^{(-l)}$ 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 l -th sub-sample.
- We compute the criterion :

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{x}_i \hat{\beta}_{\lambda}^{(-\tau(i))})^2.$$

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

Application to the Ozone data : The value of λ selected by cross-validation is 5.4. We show the obtained value in the next Figure.

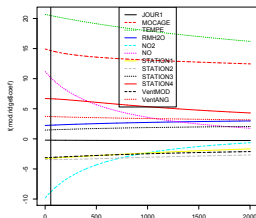


FIGURE – Selection of the regularization parameter by CV

The LASSO regression

- LASSO is the abbreviation of **Least Absolute Shrinkage and Selection Operator**.
- The Lasso estimator is introduced in the paper by Tibshirani, R. (1996) : 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.

Definition

The Lasso estimator of β in the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, is defined by :

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

where λ is a nonnegative tuning parameter.

Model and estimation

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

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

where t is suitably chosen, and $\hat{\beta}_{0\text{Lasso}} = \bar{Y}$.

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 $j = 1, \dots, p$.

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

— If $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the solution β of the minimization of the Lasso criterion

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

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

$$\beta_j = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)\mathbb{1}_{|\hat{\beta}_j| \geq \lambda},$$

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

The obtained estimator corresponds to a soft thresholding of the least square estimator.

The coefficients $\hat{\beta}_j$ are replaced by $\phi_\lambda(\hat{\beta}_j)$ where

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

- The LASSO is equivalent to the minimization of the least square criterion 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 $\sum_{j=1}^p |\beta_j|$:

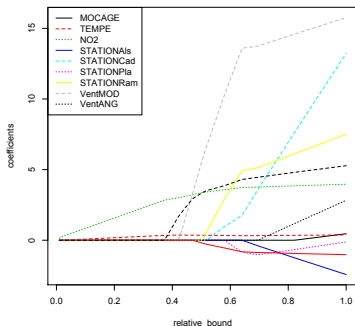
$$\sum_{j=1}^p |\beta_j| \leq \kappa \sum_{j=1}^p |\hat{\beta}_j^{(0)}|,$$

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

For $\kappa = 1$ we recover the least square estimator and for $\kappa = 0$, all the $\hat{\beta}_j$, $j \geq 1$, vanish.

Applications

We represent in the next Figure the values of the coefficients in function of κ for the Ozone 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 next Figure gives a geometric interpretation of the minimization problems for both the Ridge and Lasso estimators. This explains why the Lasso solution is sparse.

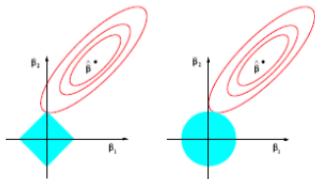


Figure 3.12: Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the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

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

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

- 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.

Conclusion

- Linear models are quite general since they can incorporate new variables defined as functions of the initial variables : X_j^2 , $\sin(X_j)$, $\log(X_j)$...
- We have seen the importance of model/variable selection to avoid overfitting
- The next step will be to consider linear models for classification (such as logistic regression models, linear SVM) and nonlinear models for regression or classification.

References

- Giraud C. (2015) *Introduction to High-Dimensional Statistics* Vol. 139 of Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL.
- Hastie, T. and Tibshirani, R. and Friedman, J, (2009), *The elements of statistical learning : data mining, inference, and prediction*, Springer.
- McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models*. 2nd edition. Chapman et Hall.