Institut de Mathématiques de Toulouse, INSA Toulouse

Linear models for Regression

Formation en Machine Learning
EUR NanoX

Béatrice Laurent - Olivier Roustant

Introduction

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

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

with $\boldsymbol{x}_i \in \mathcal{X}$(p-dimensional), $y_i \in \mathcal{Y}$ for $i = 1 \ldots n$.

Objectives : From the learning sample, we want to

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

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

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

quantitative output

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

$\downarrow$

regression

qualitative output

$\mathcal{Y}$ finite

$\downarrow$

classification

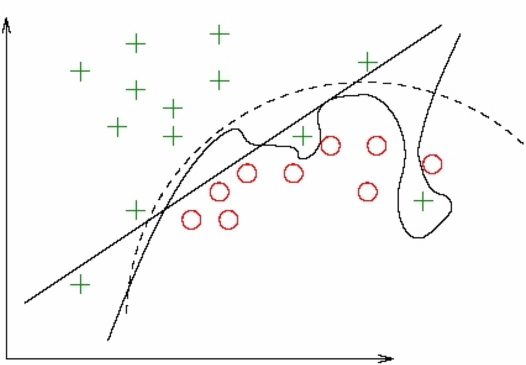
- In this course, we consider supervised learning with linear models for real regression ($\mathcal{Y} \subset \mathbb{R}$) and classification ($\mathcal{Y}$ finite).
- The explanatory variables $X^1, \dots X^p$ can be qualitatives or quantitatives

Complexity of the models

Choice of the model

- Importance of the principle of parsimony : "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

Complexity of the models



Supervised Classification : $x \in \mathbb{R}^2$, $Y \in \{0,1\}$

First step : *Data munging*

1. Extraction
2. Exploration, visualization
3. Cleaning, removal of outliers, transformation of the data, ...
4. Management of missing data

Second step : *Learning*

1. Random Partition of the sample : learning, (validation), test
2. **For** each method that we consider :
   - Learning (estimation) depending on θ (complexity)
   - Optimization of θ : validation set or cross-validation with the learning set
3. Comparison of the methods : prediction error on the test sample
4. Eventual Iteration (*Monte Carlo*)
5. Choice of the method (prevision *vs.* interpretability).
6. Estimation of the selected model with all the sample, exploitation

Possibly : Aggregation of several models

Question : Where to bring the effort ?

- *Data munging*
- Selection of the methods to compare
- Optimization of the parameters
- Optimal Combination of the models

Depending on :

- Goal (allotted time)
- Regularity of the underlying problem
- Structure and properties of the data

Program of the course :

Part I

- Linear models for regression
- Linear models for classification

Part II

- Classification And Regression Trees, Bagging, Random Forests
- Neural networks, Introduction to deep learning

Part I-1 :

- Linear models for regression
  - Linear model
  - Least square estimation
  - Confidence intervals and prediction intervals
  - Determination coefficient, Diagnosis on the residuals
  - Model selection, variable selection

The Linear model

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

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

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

We assume in a first time that $n > p + 1$.

In *the linear model*, the regression function is linear in the input variables $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p$.

The Linear model

The linear model is defined by :

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

with the following assumptions :

1. The random variables ε_i are independent and identically distributed (i.i.d.), $\mathbb{E}(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, independent of $(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p)$.
2. We have
$$E(\mathbf{Y}) = \beta_0 + \beta_1 \boldsymbol{X}^1 + \beta_2 \boldsymbol{X}^2 + \cdots + \beta_p \boldsymbol{X}^p \text{ and } Var(\mathbf{Y}) = \sigma^2.$$
3. The unknown parameters $\beta_0, \ldots, \beta_p$ are estimated from the learning sample, as well as σ^2.
4. It is generally assumed that 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 $\boldsymbol{X}^j$ can be quantitative variables, nonlinear transformation of quantitative variables (such as log, exp, square ..), interactions : $\boldsymbol{X}^j = \boldsymbol{X}^k.\boldsymbol{X}^l$.
- 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.
- The response variable is given in the vector $\mathbf{Y}$.
- We set $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \cdots \ \beta_p]'$, which leads to the matricial formulation of the linear model :

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

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 $X^1, \ldots X^p$ for the explanatory variables (**MOCAGE** , **TEMPE**, **JOUR** ..). The variables are quantitative (**MOCAGE** , **TEMPE** , ...), or qualitative (**JOUR**, **STATION**).
- We consider the linear model :

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

- 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 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 :

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

Lemma

Let $h : \beta \mapsto \beta' A \beta$ *where* A *is a symmetric matrix.*
Then $\bigtriangledown h(\beta) = 2A\beta$.
Let $g : \beta \mapsto \beta' z = z'\beta = \langle z, \beta \rangle$ *where* $z \in \mathbb{R}^p$.
Then $\bigtriangledown g(\beta) = z$.

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 :

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

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

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

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 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 $\widehat{\mathbf{Y}}$ 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 approches to reduce the dimension (*Ridge*, Lasso, PLS ...).

Least square estimation

- We define the vector of residuals as :

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

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

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

(We recall that $\beta \in \mathbb{R}^{p+1}$).

Properties of the least square estimator

Theorem

— *Assuming that*

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

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

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

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

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

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

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

Confidence intervals

One can easily deduce from the first theorem that

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

where $\mathcal{T}_{(k)}$ denotes a Student distribution with k degrees of freedom.
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 :

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

where $t_{k,0.975}$ denotes the 0.975 quantile of the Student distribution with k degrees of freedom.

Test of significance

- We recall the linear model

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

- 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,

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

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 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= **MOCAGE**

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

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 :

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

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

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

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

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

Prediction

We derive from Theorem 1 a :

- **Confidence interval for the mean response $\mathbf{X_0}'\beta$** at the new observation point $\mathbf{X_0}$:

$$\left[\mathbf{X_0}'\widehat{\beta} - t\hat{\sigma}\sqrt{\mathbf{X}_\mathbf{0}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X_0}}, \mathbf{X_0}'\widehat{\beta} + t\hat{\sigma}\sqrt{\mathbf{X}_\mathbf{0}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X_0}}\right].$$

- **Prediction interval for the response** Y_0 at the new observation point $\mathbf{X_0}$ is :

$$\left[\mathbf{X_0}'\widehat{\beta} - t\hat{\sigma}\sqrt{1 + \mathbf{X}_\mathbf{0}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X_0}}, \mathbf{X_0}'\widehat{\beta} + t\hat{\sigma}\sqrt{1 + \mathbf{X}_\mathbf{0}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X_0}}\right].$$

We first consider a simple linear regression model with the single variable $X=$ **MOCAGE**

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

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

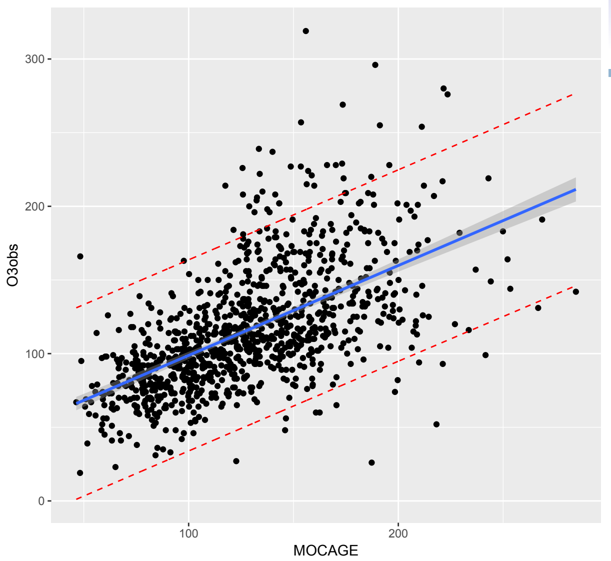
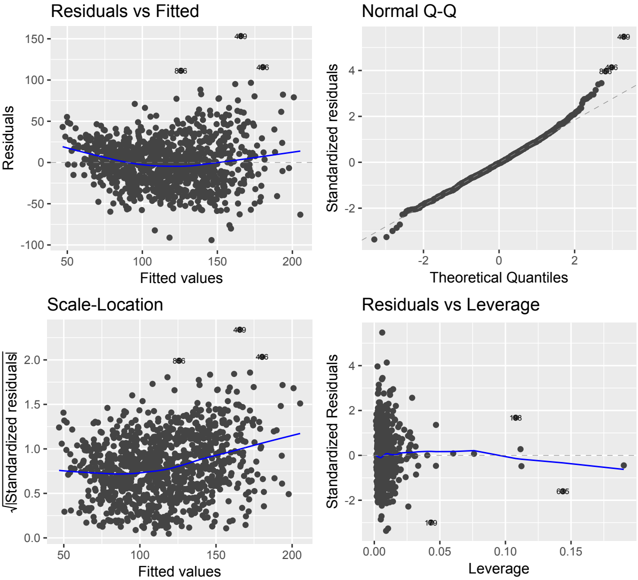


Figure – Simple linear regression model : confidence interval for the mean response (in grey) and prediction intervals (in red).

Diagnosis on the residuals



Measures for goodness-of-fit

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

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

- The determination coefficient also called R^2 is defined as

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

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

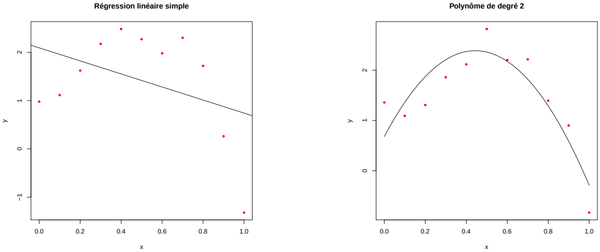
Determination coefficient and Model selection

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 :

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

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 2 and 3.

Determination coefficient and Model selection



Régression linéaire simple

Polynôme de degré 2

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

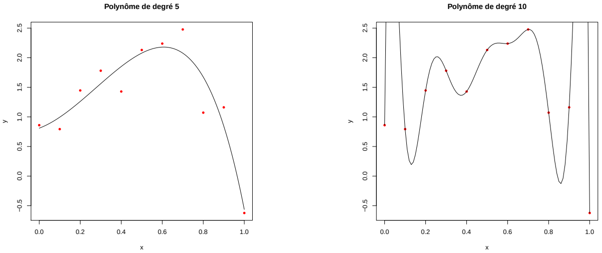


FIGURE – Polynomial regression : adjusted model, on the left : $y = \beta_0 + \beta_1 x + \ldots + \beta_5 x^5 + \epsilon$, $R^2 = 0.874$, on the right : $y = \beta_0 + \beta_1 x + \ldots + \beta_{10} x^{10} + \epsilon$, $R^2 = 1$.

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.

Determination coefficient and 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 or adjusted R^2 defined by :

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

  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= **MOCAGE**

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

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 :

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

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 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.

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

Variable selection

The Mallow's C_P criterion is

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

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

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

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

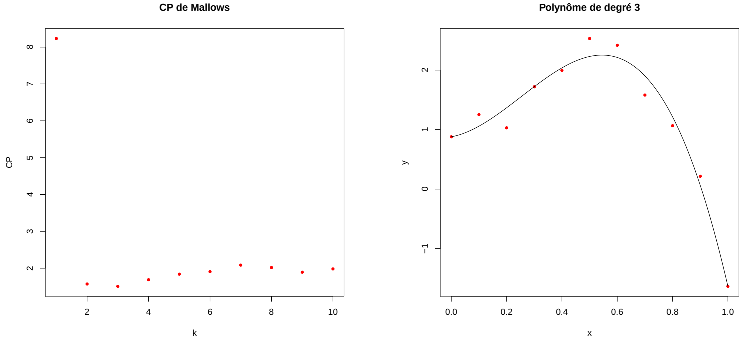


FIGURE – Mallows' C_P in function of the degree of the polynomial. Selected model : polynomial with degree 3.

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 + RMH2O + NO2 + NO + VentMOD + VentANG
                      Df   Sum of Sq   RSS       AIC
         - VentMOD    1    1484        817158    6952.9
         <none>                        815674    6953.0
         - RMH2O      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 + RMH2O + 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

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

where

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

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

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

Definition

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

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

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

Assume that $\boldsymbol{X}$ 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

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

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

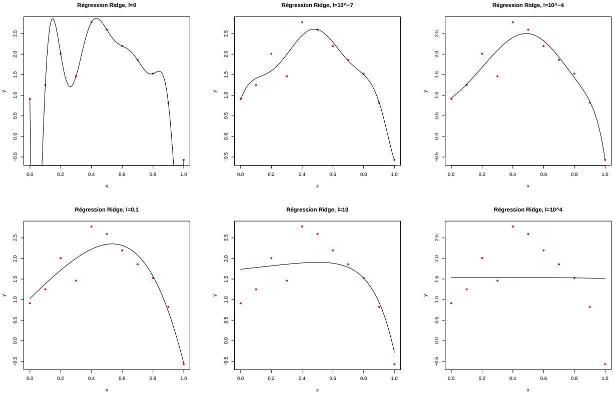
Remarks :

1. $\mathbf{X}'\mathbf{X}$ is a nonnegative symmetric matrix. 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 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



Régression Ridge, l=0

Régression Ridge, l=10^−7

Régression Ridge, l=10^−4

Régression Ridge, l=0.1

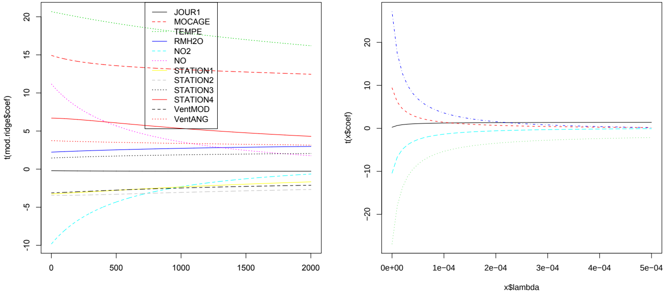
Régression Ridge, l=10

Régression Ridge, l=10^4

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 principe 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 :

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

- 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.

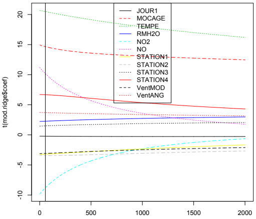


FIGURE – Selection of the regularization parameter by CV

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 :

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

where λ is a nonnegative tuning parameter.

Model and estimation

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

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

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.

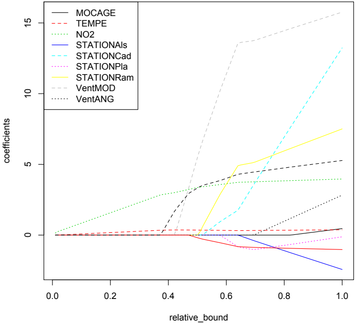
- The LASSO is equivalent to the minimization of the least square criterion under the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$, for some $t > 0$.
- The statistical software R introduces a constraint expressed by a relative bound for $\sum_{j=1}^{p} |\beta_j|$:

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

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.

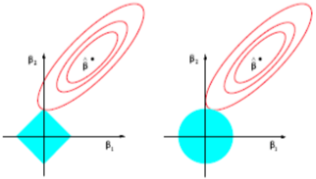


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the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

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 the logistic regression models) or non linear models for regression (trees, neural networks)

References

- Hastie, T. and Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning : data mining, inference, and prediction*, Springer.
- Jobson, J.D. (1991) *Applied Multivariate Data Analysis, Vol 1 : Regression and Experimental Design*, Springer.