Linear Regression models & Regularization part2

Mathilde Mougeot

ensIIE & ENS Paris-Saclay, France

2025

Regularization Methods for Linear Regression

1. Statistical tests for the Linear Model
   - Significativity of a coefficient : Student test
   - Global significativity of the model : Fischer test
   - Impact of correlation and multicolinearity
2. Towards parsimonious model
   - Greedy method for model selection
   - Penalized the Log-likelihood. Information criteria (AIC, BIC)
3. Predictive power of a model
   - Cross validation
4. Penalized OLS regression methods
   - Ridge, ℓ_2 penalization
   - Lasso - ℓ_1 penalization

Outline

Example

Regression model :

$$consommation = \beta_1 + \beta_2 income + \beta_3 price + \beta_4 temp + \epsilon$$

R outputs :

```
##
## Call:
## lm(formula = "cons~.", data = tab)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065302 -0.011873  0.002737  0.015953  0.078986
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1973151  0.2702162   0.730  0.47179
## income       0.0033078  0.0011714   2.824  0.00899 **
## price       -1.0444140  0.8343573  -1.252  0.22180
## temp         0.0034584  0.0004455   7.762  3.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03683 on 26 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.6866
## F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

Law of the estimated coefficients and variance

With an assumption of normality of the residuals, we have :

- Coefficients : $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$
$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 S_{jj}}} \sim \mathcal{N}(0,1)$ with $S_{j,j}$ j^{th} term of the diagnonal of $(X^TX)^{-1}$

- Residual Variance : $\frac{n-p}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}$ with $\hat{\sigma}^2 = \frac{||\hat{\epsilon}||^2}{n-p}$

- We then have : $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 S_{jj}}} / \sqrt{\frac{n-p}{\sigma^2}\hat{\sigma}^2/(n-p)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 S_{jj}}} \sim T(n-p)$

Recall : Student theorem.
$U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(d)$, U and V are independant, then we have $Z = \frac{U}{\sqrt{V/d}}$ follows a Student law of parameter d.

Significativity test of $\hat{\beta}_j$, σ^2 **unknown**

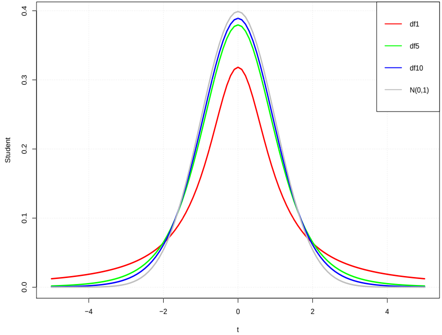
- Student Statistics : T
- Significativity test (bilateral)

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

- Decision with a risk α, **Reject H_0 if**
  - $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 S_{j,j}}} > t_{n-p}(1-\alpha/2)$ with $S_{j,j}$ j^{th} term of diagonal of $(X^T X)^{-1}$
  - pvalue $< \alpha$
- Conclusion (if H_0 is rejected) :
  - β_j is significatively different of zero
  - X_j is significatly involved in the model

Not appropriate if there exists collinearity between the variables

Illustrations of Student laws.



Example

Regression model :

$$consommation = \beta_1 + \beta_2 income + \beta_3 price + \beta_4 temp + \epsilon$$

R output :

```
##
## Call:
## lm(formula = "cons~.", data = tab)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065302 -0.011873  0.002737  0.015953  0.078986
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1973151  0.2702162   0.730  0.47179
## income       0.0033078  0.0011714   2.824  0.00899 **
## price       -1.0444140  0.8343573  -1.252  0.22180
## temp         0.0034584  0.0004455   7.762  3.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03683 on 26 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.6866
## F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

Global significativity of the model

- Fisher Statistic
- Significativity test (bilateral)
  - $H_0 : \beta_2 = \ldots = \beta_p = 0$
  - $H_1 : \exists \beta_j \neq 0$
- Decision with a rish α, **Reject H_0 if**
  - if $\frac{n-p}{p-1}\frac{R^2}{1-R^2} = \frac{ESS/(p-1)}{RSS/(n-p)} > f_{p-1,n-p}(1-\alpha)$
  - if pvalue $< \alpha$

$\rightarrow$ The linear model has globally an added value

Example

Regression model :

$$consommation = \beta_1 + \beta_2 income + \beta_3 price + \beta_4 temp + \epsilon$$

R output :

```
##
## Call:
## lm(formula = "cons~.", data = tab)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065302 -0.011873  0.002737  0.015953  0.078986
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1973151  0.2702162   0.730  0.47179
## income       0.0033078  0.0011714   2.824  0.00899 **
## price       -1.0444140  0.8343573  -1.252  0.22180
## temp         0.0034584  0.0004455   7.762  3.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03683 on 26 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.6866
## F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

Linear Regression model

- **Framework**
  - Target : Y $(N, 1)$ vector. Design matrix : X (n, p) matrix
  - Linear model : $Y = X\beta + \epsilon$
  - $\beta_{OLS} = \arg\min_\beta ||Y - X\beta||_2^2$

- **if X^TX is invertible, the solution is :**
  - $\hat{\beta}_{MCO} = (X^TX)^{-1}X^TY$
  - If $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$
  - $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 S_{jj}}} \sim T(n-p)$
    with $\hat{\sigma}^2 = \frac{||\hat{\epsilon}||^2}{n-p}$ and $S_{j,j}$ j^{th} term of the diagnonal of $(X^TX)^{-1}$

- **if $X^TX = I_p$, (independent variables) the solution is equal to :**
  - $\hat{\beta}_{MCO} = X^TY \qquad \hat{\beta}_j =< X_j, Y >, 1 \leq j \leq p.$
  - The estimation of the coefficients does not depend on the others

Impact of dependance for testing coefficients

Illustration : $n = 100; X = cbind(((1:n)/n)^3, ((1:n)/n)^4);$
$Y = X\% * \%c(1,1) + rnorm(n)/4;$

$\boxed{\text{Model I : } Y = \alpha_0 + \beta_1 X_1 + \epsilon}$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.11	0.03	-3.833	0.000224	***
X[, 1]	2.01	0.07	25.731	< 2e-16	***

$\boxed{\text{Model II : } Y = \gamma_0 + \gamma_2 X_2 + \epsilon}$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.03	0.02	-1.315	0.192	
X[, 2]	2.12	0.08	25.377	<2e-16	***

$\boxed{\text{Model III : } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon}$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.08	0.03	-2.31	0.0226	*
X1	1.24	0.62	1.98	0.0497	*
X2	0.82	0.66	1.24	0.2169	

Impact of Multicolinearity:

Framework: Y: target variable. X_1, X_2: covariables.

The model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Estimation of the coefficients: $\beta = (\beta_0, \beta_1, \beta_2)$. $\hat{\beta} = \underset{\beta}{\mathrm{ArgMin}} \; \|Y - X\beta\|_2^2$.

$$E(\beta) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \right)^2 .$$

Derivative Computation.

$$\begin{cases}
\dfrac{\partial E(\beta)}{\partial \beta_0} = 0 \Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 \\[2ex]
\dfrac{\partial E(\beta)}{\partial \beta_1} = 0 \Rightarrow 2 \displaystyle\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \right)(x_{1i}) . \\
\qquad \Rightarrow \displaystyle\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \beta_1 (x_{1i} - \bar{x}_1) - \beta_2 (x_{2i} - \bar{x}_2) \right] (x_{1i} - \bar{x}_1) \\[2ex]
\dfrac{\partial E(\beta)}{\partial \beta_2} = 0 \Rightarrow 2 \displaystyle\sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \right] (x_{2i}) \\
\qquad \Rightarrow \displaystyle\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \beta_1 (x_{i1} - \bar{x}_1) - \beta_2 (x_{2i} - \bar{x}_2) \right] (x_{2i} - \bar{x}_2)
\end{cases}$$

$$\Rightarrow \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_1, x_2) & \mathrm{Var}(x_2) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(x_1, y) \\ \mathrm{Cov}(x_2, y) \end{bmatrix}$$

$$\begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(x_1, y) \\ \mathrm{Cov}(x_2, y) \end{bmatrix}.$$

$$\begin{bmatrix} \hat\beta_1 \\ \hat\beta_2 \end{bmatrix} = S_{xx}^{-1} \cdot S_{xy} \quad \text{with} \quad S_{xx} = \begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix} \qquad S_{xy} = \begin{bmatrix} \mathrm{Cov}(x_1, y) \\ \mathrm{Cov}(x_2, y) \end{bmatrix}$$

$$S_{xx}^{-1} = \frac{1}{S_1^2 S_2^2 - S_{12}^2} \begin{bmatrix} S_2^2 & -S_{12} \\ -S_{12} & S_1^2 \end{bmatrix}$$

$$S_{xx}^{-1} = \frac{1}{S_1^2 S_2^2 (1-\rho^2)} \begin{bmatrix} S_2^2 & -S_{12} \\ -S_{12} & S_1^2 \end{bmatrix} \qquad \rho = \frac{S_{12}}{S_1 \cdot S_2}$$

Test of significativity for the coefficients β_1 and β_2

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0. \end{cases}$$

α: level of the Test.

. Test statistics

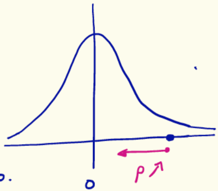
$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot V_{jj}}} \qquad T \sim \text{Student}(n-p).$$

V_{jj} : j^{th} element of S_{xx}^{-1}.

. test for β_1

$$V_{11} = \frac{s_2^2}{s_1^2 \cdot s_2^2 (1-\rho^2)} = \frac{1}{s_1^2 [1-\rho^2]}.$$

$$T = \frac{s_1 \cdot \sqrt{1-\rho^2} \cdot \hat{\beta}_1}{\hat{\sigma}}$$



. Remark

- if $\rho \to 1$ the $T \to 0$.
  (The corelation between $X_1 . X_2$ increases)
- if $\rho \to 1$, The statistical Test tends to keep H_0.
  (H_0 is not rejected,
  conclusion: β_1 is not significatively different of zero.)

For the General Framework.

$$V_{jj} = \frac{1}{1 - R_j^2}$$

R_j^2 = determination coefficient for the model. x_j explained by the other variables.

VIF: Variance Inflation Factor.

Linear Regression model

If $X^T X$ non inversible.

Use of the Pseudo inverse to compute the coefficients

$X^T X$ is non invertible with the rank k, $k < p$:

$$
\begin{aligned}
X^T X &= U \Sigma^2 U^T \\
&= U \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \vdots & 0 & 0 \\ 0 & 0 & \sigma_k^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} U^T \\
&= U_k \Sigma_k^2 U_k^T
\end{aligned}
$$

$$
(X^T X)^{*-1} = U_k {\Sigma_k^2}^{-1} U_k^T \quad \text{avec } \Sigma_k^2 = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \sigma_k^2 \end{pmatrix}
$$

$\hat{\beta} = (X^T X)^{*-1} X^T Y$

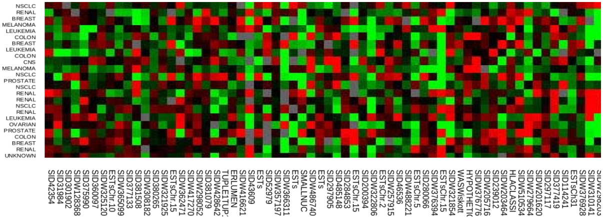
$\rightarrow$ **No unique solution for the coefficients**

Outline

High dimensional modeling. Illustration

First example : genetics

- We study the production of a given molecule and Y_i is the concentration of the production for the i^{th} experiment.
- For each experiment, we can measure the expression of the p genes. $X_{i,1}, \ldots, X_{i,p}$ ($p \gg 1$). In this case, there is a huge number of inputs.
- $p >> n$



Main objectives :

Selection of the *important* variables

- What does *important* means ?
- *screening* : at least, all the important variables are selected.
- *selection* : Only the important variables are selected.
- $\rightarrow$ Need of **interpretability** and **parsimony**.

Estimation of the variable parameters

- Modeling vs prediction. Both objectives are different.

Accurate target prediction for futur observed inputs

- How can we measure accuracy ? Be careful not to be to optimistic.
- Bootstrap sampling (bootstrap) or cross-validation (simple or K fold).
- Information criteria(AIC, BIC, C_p).

Illustration of over-fitting for polynomial regression

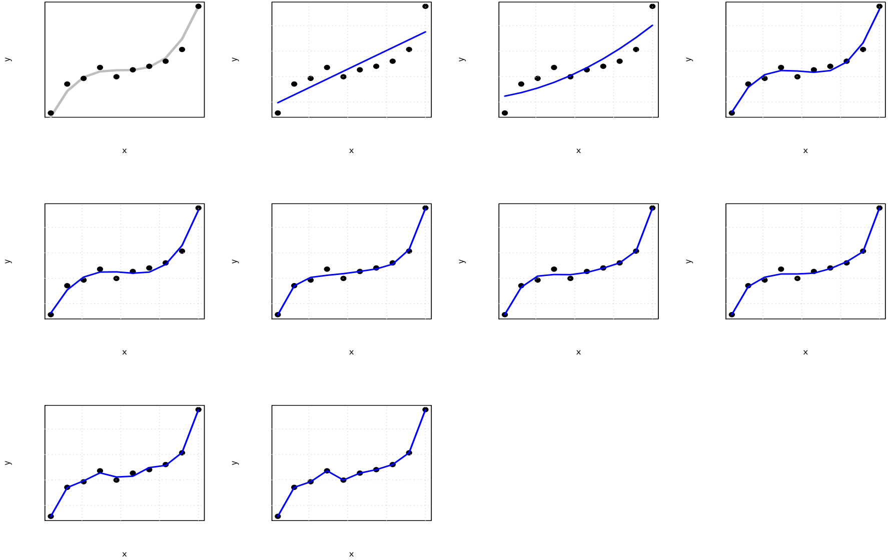
Variables

- Y :Target variable, $Y \in \mathbb{R}$
- X : Explanatory variable, $X \in \mathbb{R}$

Model : $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_{p-1} X^{p-1}$

Goal :
$\rightarrow$ Given a set of data, we aim to recover the appropriate expression, p ? β_j ?

Polynomial regression with different orders : 1,2,... p...



Linear modeling towards parsimonious models

1. Linear model (Gaussian assumption on the residuals)
    - Estimation and prediction
    - Tests of significativity of the coefficients
    - Search of parsimonious models
    - Estimation and selection of parsimonious models based on penalized likelihood
2. Penalized Ordinary Least Square (OLS)
    - Ridge regression : OLS with ℓ_2 penalized coefficents
    - Lasso regression : OLS with ℓ_1 penalized coefficents

Linear Model

Model

Observations $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$
$\forall i, Y_i = X_i\beta + \epsilon_i$ with matrix notation : $Y = X\beta + \epsilon$
$\beta \in \mathbb{R}^p$, ϵ_i iid $\mathcal{N}(0,1)$, X known.

Independant columns

If X is of full rank then X^TX is invertible and :

$$\hat{\beta}^{\text{MCO}} = \arg\min_{\alpha\in\mathbb{R}^p} \|Y - X\alpha\|^2 = (X^TX)^{-1}X^TY$$

Available algorithms to compute the solution :

- Choleski en $p^3 + Np^2/2$
- QR en Np^2

"Optimality" result

Gauss-Markov theorem :

$$\hat{\beta}^{\mathrm{MCO}} \stackrel{def}{=} \arg\min_{\alpha \in \mathbb{R}^p} \|Y - X\alpha\|^2 = (X^T X)^{-1} X^T Y \ .$$

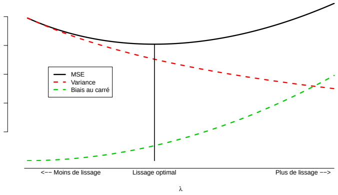
is optimal for the quadratic risk for in the non biased estimator family (BLUE : *best linear unbiased estimator*).

- The BLUE of $\beta^{(i)}$ est $\hat{\beta}^{(j)} := (\hat{\beta}^{\mathrm{MCO}})^{(j)}$

Generally

MSE $= \mathbb{E}[(\hat{\beta} - \beta)^2]$:

$$\mathtt{MSE} = \mathtt{biais}^2 + \mathtt{variance}$$



Model selection in the linear Gaussian framework

Objective : Find the "most simple" models with a high predictive power among all the linear possible models :

$$Y = X_{\mathcal{M}}\beta + \epsilon$$

where $\mathcal{M} \subset \{1, \ldots, p\}$ et $\mathbf{X}_{\mathcal{M}} = [X_{i,j_k}]_{i=1,\ldots,n; j_k \in \mathcal{M}}$.

Best subset family (*best subset*)

- $$\mathrm{RSS}(\mathcal{M}) \overset{def}{=} \|\mathbf{Y} - \mathbf{X}_{\mathcal{M}}(\mathbf{X}_{\mathcal{M}}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{X}_{\mathcal{M}}^{T} Y\|^2 \,,$$
- $$\hat{\mathcal{M}} \overset{def}{=} \underset{\mathcal{M} \subset \{1,\ldots,p\}}{\arg\min}\ \mathrm{RSS}(\mathcal{M}) + \text{penalty}$$
- 2^p models to test ! Condition : $(\mathbf{X}^T\mathbf{X})$ invertible.
- "Smart" algorithms (type *branch and bound* cf. Furnival & Wilson, 1974), can be used up to $p \sim 50$. (RSS : Residual Sum of Square)

Linear models and model (variable subset) selection

$$Y = X\beta + \epsilon \text{ avec } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Several approaches :

Exhaustive method : Best Subset

Incremental approaches :

1. Forward regression
2. Backward regression
3. Stepwise regression

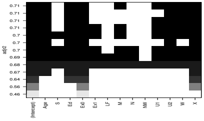
Criteria to penalized the number of variables

The R-squared :

- $R^2 = \frac{Var\hat{Y}}{VarY} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \in [0, 1]$
  TSS : Total Sum Squared, ESS : Estimated SS, RSS : Residual.
- The value of R^2 mechanically increases with the number of variables. Therefore, it is not useful for model selection

The Adjusted R-squared introduces a penalization of the number of variables :

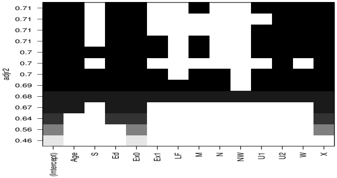
- $R^2_{adj} = 1 - \frac{RSS}{TSS}\frac{n-1}{n-p} = 1 - (1 - R^2)\frac{n-1}{n-p}$
  Recall that :
  - $RSS/(n-p)$ Non biased estimator of the residual error,
  - $TSS/(n-1)$ Non biased estimator of the variance
- R^2_{adj} can take negative values



Best subset selection. R software outputs :

Best subset method

- The number of initial p variables is not too large, typically $p < 30$
- All or most of the models are implemented (2^p)
  (Furnival, Wilson 1974)
- For a given p, the model providing the largest R^2 value is selected
- Between two models characterized with a different numebr of inputs, the model with the largest adjusted R-squared is selected (R^2_{adj}).



Best subset selection. R outputs

Incremental methods ("Greedy" method)

Forward selection (*step by step*)

- First step : the model is resume to the intercept $\mathcal{M}_0$ nul ;
- At step k, the variable which may increased the most the R^2 index is added to the previous $\mathcal{M}_k$.
- This step by step process ends when the variable which should be integrated has a non significative coefficient in the current model.

Backward selection (*step by step*)

- First step : Full model ;
- At step k, the variable which showed the lowest Z score leaves the $\mathcal{M}_k$ model.
- This step by step process ends when all the variables of the model showed significative coefficients.

Stepwise selection (*step by step*)

- First step : the model is resume to the intercept $\mathcal{M}_0$ nul ;
- Etape k
  - At step k, the variable which may increased the most the R^2 index is added to the previous $\mathcal{M}_k$.
  - Non significative regressors are drop.
- This step by step process ends when the variable which should be integrated shows a non significative coefficient in the current model.

Limitations

- Instability (cf Breiman, 1996)
- Globally not optimal (partial exploration) ("Greedy" method)
- based on a Student Test which used a Gaussian framework.

Akaike criteria (AIC, 1973)

For variable selection and linear model, several criteria are introduced to penalized the Log-likelihood.

AIC general expression :

$$-2\mathbb{E}(\log f_{\hat{\beta}}(\mathbf{X}, Y)) \simeq -2\mathbb{E}(\log \text{lik}) + 2\frac{p}{n} \simeq -2\log \text{lik} + 2\frac{p}{n} \stackrel{def}{=} \texttt{AIC}$$

with loglik $= \sum \log(f_{\hat{\beta}}(\mathbf{X}, Y))$ et $\hat{\beta}$: Maximum Likelihood Estimation (MLE)

Gaussian Linear model

- The OLS estimator is the same than the MLE.
- p is the number of parameters of the model (degree of freedom)

$\rightarrow$ Find the model which minimizes AIC criteria

Bayesien Information Criteria (BIC, Schwarz, 1976)

For variable selection and linear model, several criteria are introduced to penalized the Log-likelihood.

BIC general expression

$$\texttt{BIC} \stackrel{def}{=} -2\text{loglik} + \log n \frac{p}{n}$$

BIC vs AIC comparison

→ Find the model which minimizes BIC criteria

- The penality appears to be stronger tan AIC ($\log n \gg 2$) ;
- BIC will lead to more parsimonious models (with less variables)
- Bayesian framework

C_p of Mallows (1968)

For the linear model, several criteria are introduced to penalized the number of parameters.

Expression of the Mallows C_p index

$$C_p = \hat{\mathbb{E}}(Y - X\hat{\beta})^2 = n^{-1}\sum(Y_i - \mathbf{X}_i\hat{\beta})^2 + \frac{2p}{n}\underbrace{\hat{\sigma}^2}_{\text{for the complete model}}.$$

For the Gaussian Linear Model

- The OLS estimator is the same than the MLE.
- p is the number of parameters of the model (degree of freedom)

$\rightarrow$ Find the model which minimizes Mallows criteria.

Linear model selection

Regarding :

- Best Subset method
- Forward, Backward, Stepwise methods
- AIC, BIC, Mallows criteria

All of these criteria are defined in the linear model framework, with Gaussian assumptions for the residuals (MLE).

$\rightarrow$ Ridge, Lasso are alternative OLS method with Penalized coefficients...

Outline

Evaluation of the predictive power of a model : a Machine Learning view

Idea

- if we use the same data to first compute the parameters of a model then to evaluate its ability to predict by the computation of the RMSE prediction, we are **over optimistic** .
- $\hat{\beta} = \hat{\beta}((X_i, Y_i))$ and new observations observations (X_i, Y_i')

$$\frac{1}{n}\mathbb{E}_{(\mathbf{X},\mathbf{Y}')}[\|\mathbf{Y}' - \mathbf{X}\hat{\beta}\|^2 | (\mathbf{X},\mathbf{Y})] = \underbrace{\frac{1}{n}\sum (Y_i - \mathbf{X}_i\hat{\beta})^2}_{= n^{-1}\|\hat{\epsilon}\|^2 = \text{erreur résiduelle}} + \text{Terme} > 0 \ .$$

Evaluation of the predictive power of a model : a Machine Learning view

The "rich man" approach : data sampling

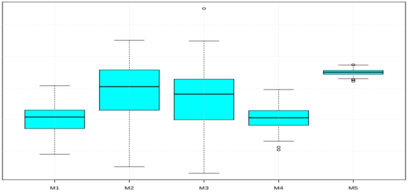
- Cross Validation
  - 50% to train the models (*training set*) ;
  - 25% to test and select the best model associated with the lowest RMSE error (*validation set*) ;
  - 25% to evaluate the best model (*test set*).
- K Fold
- Leave one out

These approaches are extremely used for model selection in the Machine learning community, even when the model is not a linear model.

Sometimes, we are "poor" of data and we need other approaches....

Model selection in practice : a Machine Learning view

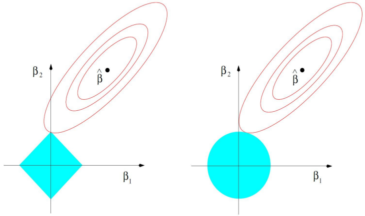
For a given problem, several models are implemented and the model, which shows the best predictive power, i.e. the lowest error on a test data set, is finally selected.



Model comparisons and selection based on K fold cross validation

Outline

Ordinary Least Square with a penalization on the coefficients



Penalized regression methods

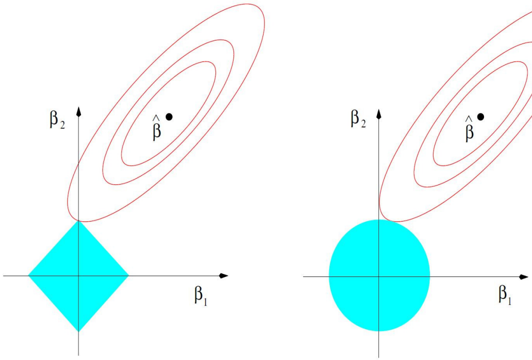
In this case, a constraint on the β coefficients is introduced in the OLS model :

- Ridge : $E(\beta) = ||Y - X\beta||^2$ under the constraint $\sum_j \beta_j^2 \leq c$
- Lasso : $E(\beta) = ||Y - X\beta||^2$ under the constraint $\sum_j |\beta_j|^1 \leq c$

$\rightarrow$ ℓ_1 or ℓ_2 penalizations induce different properties in the final computed estimation.

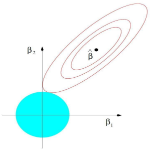
- ℓ_1 penalization induce sparse models. The value of "non useful" coefficients equal zero.
- ℓ_2 penalization helps to compute a solution in degenarative cases.

Penalized regression methods



Lasso et Ridge penalized methods

Ridge regression



Ridge Regression

Several points :

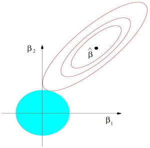
1. It's a solution to a penalized Least Square problem with smoothing properties
2. It induces a "contraction" of the original OLS coefficient values
3. It introduces a Gaussian "Apriori" in a Bayesian estimation

Ridge Regression. ℓ_2 Penalized OLS.

when $p >> n$ then $(X^T X)$ is a non inversible matrix.

The Ridge regression brings regularization in the variance-covariance matrix. In this case, the quadratic error is defined by :

$$E(\beta) = (Y - X\beta)^T(Y - X\beta) \quad \text{under the constraint} \quad ||\beta||^2 \leq c$$



Illustration

Ridge Regression. ℓ_2 Penalized OLS.

- The quadratic error is defined by :

$$E(\beta) = (Y - X\beta)^T(Y - X\beta) \quad \text{under the constraint} \quad ||\beta||^2 \leq c$$

- With the help of the Lagrange multiplier, we write :

$$\begin{aligned} \Phi(\beta) &= (Y - X\beta)^T(Y - X\beta) + k\sum_{j=1}^{p} \beta_j^2 \\ &= (Y - X\beta)^T(Y - X\beta) + k\beta^T\beta \qquad \text{with } k \geq 0 \end{aligned}$$

- $\hat{\beta}_{RR}$ minimizes $\Phi(\beta)$:

$$\hat{\beta}_{RR} = (X^TX + kI_p)^{-1}X^TY$$

Ridge Regression, in practice.

Remarque :

- Data scaling is essential (for all the variables X_j, $1 \leq j \leq p$) in order to apply the same penalization value to all coefficients.
- The intercept should be never penalized. In practice, data are often centered before any computation.
  $\Phi(\beta) = (Y - X\beta)^T(Y - X\beta) + k\sum_{j=2}^{p}\beta_j^2$

R instructions, as an example :

- modridge=lm.ridge(Y ∼ X,data=Z,lambda=5) ;
  print(summary(modridge)) ;
- Output fields :
  coef / lambda / scales / ym / xm / GCV
- modridge\$coef ; values of the coefficients in the "rescaling framework"
- coef(modridge) ; values of the coefficients in the initial framework

Ridge Regression. OLS coefficient shrinkage

Ridge and OLS comparison

To simplify the computations, we present the comparison in the particulary case when $X^T X$ is the identity matrix.
In this case, the variables are orthogonal with unit variance :

- **Estimation of** $\hat{\beta}_{RR} = (X^T X + kI_p)^{-1} X^T Y$
- **In the case where** $X^T X = I_p$
  For each j^{th} coefficients of β_{RR}

$$\begin{aligned} \beta^j_{RR} &= \frac{1}{1+k}\beta^j_{MC0} \\ ||\beta^j_{RR}||^2 &= \left(\frac{1}{1+k}\right)^2 ||\beta^j_{MC0}||^2 \end{aligned}$$

$\rightarrow$ The shrinkage of each coeffcient is proportional to $1/(1+k)$

Shrinkage estimator

Ridge Regression

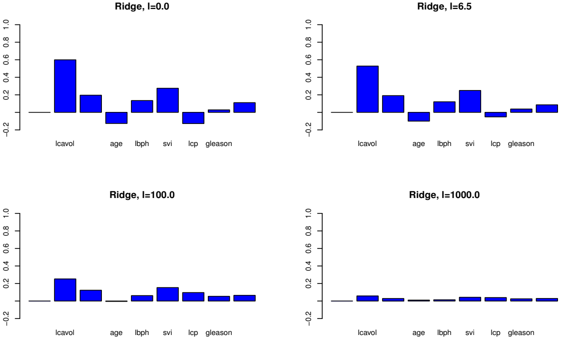
How to choose k ?

- biais-variance trade-off
- K-fold cross-validation

Ridge Regression. Application

Application : cancer data

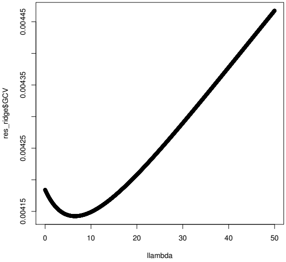
Values of the coefficients for several k penalized values



Ridge Regression. Application

Application : cancer data

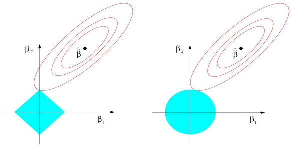
Cross-validation help to chose the k parameter value



Ridge Regression Algorithm

```
library(MASS); # PROSTATE DATA
tab0 = read.table('prostate.data'); names(data)
tab=tab0[,1:(ncol(tab0)-1)]; names(tab);
tab=data.frame(scale(tab));
# --- solve function to compute the reg. coeffs ---
X=as.matrix(cbind( rep(1,nrow(tab)),tab[,-ncol(tab)])); dim(X)
Y=tab[,ncol(tab)];
betasolve=solve(t(X)%*%X,t(X)%*%matrix(Y,nrow=nrow(tab),1));
# --- solve function to compute the ridge. coeffs ---
lambda=100; Id=diag(rep(1,ncol(X)));Id[1,1]=0; S=t(X)%*%X +
lambda*Id*nrow(tab);
betaridgesolve=solve(S,t(X)%*%matrix(Y,nrow=nrow(tab),1));
print(betaridgesolve)
# --- lambda tabaux=cbind( rep(1,nrow(tab)),tab); ---
names(tabaux)[1]='cst'; names(tabaux)
resridge = lm.ridge('lpsa .',data=tab,model=F, lambda
=nrow(tab)*100);
attributes(resridge)
reridge$coef; coef(resridge);
```

Lasso regression



lasso (gauche), ridge (droite)

Lasso Regression

- ℓ_1 Penalized OLS :

$$E(\beta) = (Y - X\beta)^T (Y - X\beta) \quad \text{constrain} \quad |\beta| \leq c$$

- Lagrange multiplier :

$$\Phi(\beta) \quad = \quad (Y - X\beta)^T (Y - X\beta) + k \sum_{j=1}^{p} |\beta_j| \quad \text{under the constraint}$$

- $\hat{\beta}_{Lasso}$ minimise $\Phi(\beta)$:

$\rightarrow$ The LARS algorithm is used in practice to compute the LASSO solution

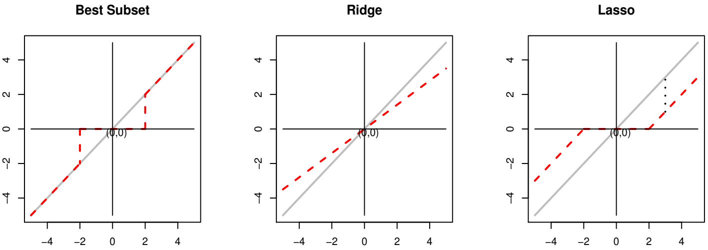
Ridge et Lasso Regression Comparison

For orthogonal variables and unitary variances : $X^TX = I_p$

Estimation	Expression
Best Subset (taille M)	$\hat{\beta}^j_{MCO} 1\{rang(\lvert\hat{\beta}^j_{MCO}\rvert) \leq M\}$
Ridge	$\frac{\hat{\beta}^j_{MCO}}{1+\lambda}$ $\quad (\lambda = k)$
Lasso	$\text{Sign}(\hat{\beta}^j_{MCO})(\lvert\beta^j_{MCO}\rvert - \lambda/2)_+$ **Soft Thresholding**

Ridge et Lasso Regression Comparison

Illustration with independent variables, $X^TX = I_p$

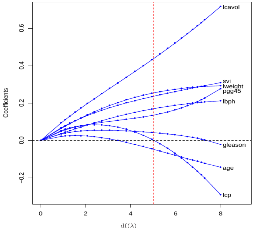


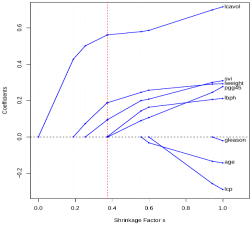
Best Subset, Ridge and Lasso Regression

t

Ridge and Lasso Regression

Regularization paths.





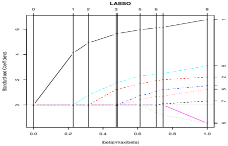
Ridge (left) et Lasso (right) Regression

Evolution of the values of the coefficients for different values of the penalized coefficient.

The LARS Algorithm for computing Lasso solution

Least Angle Regression, proposed in 2004 for High dimentional regression by Efron, Hastie, Johnston, Tibshirani.

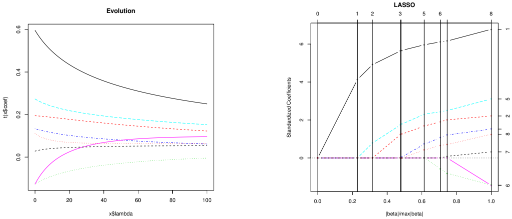
1. Start with all coefficients β equal to zero.
2. Find the predictor x_j most correlated with Y
3. Coefficient computation :
   - Increase the coefficient β_j in the direction of the sign of its correlation with y
   - Take residuals $r = y - \hat{y}$ along the way.
   - Stop when some other predictor x_k has as much correlation with r as x_j
4. Increase (β_j, β_k) in their joint least squares direction, until some other predictor x_m has as much correlation with the residual r.
5. Continue until : all predictors are in the model



Ridge Regression. Application

Study : Prostate cancer data $n = 97$ observations

Y		lpsa
X	8	lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45



Lasso regularization path