Regularization Methods for Linear Regression

Mathilde Mougeot

ENSIIE

2018-2019

Variable selection
Linear model

Regression illustration

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
```

The laws

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

for the 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 diagonal of $(X^TX)^{-1}$

for the 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**

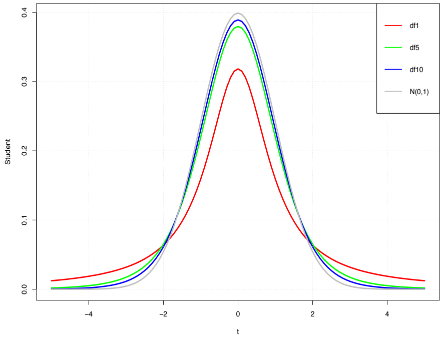
- Student Statistics : T
- Significativity test (bilateral)

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

- 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

Student laws



Regression illustration

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
```

Example : Impact of dependance...

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

	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	

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

	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	***

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

	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	***

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

Global significativity of the model

Test of the model with a risk α

$$\begin{array}{rl} H_0 : & \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\ H_1 : & \exists j = 2, \ldots, p, \beta_j \neq 0 \end{array}$$

Statistics

$$F = \frac{n-p}{p-1} \frac{||\hat{Y} - \bar{\hat{Y}}||^2}{||Y - \hat{Y}||^2} \sim Fisher(p-1, n-p)$$

Remark : $\frac{n-p}{p-1} \frac{||\hat{Y} - \bar{\hat{Y}}||^2}{||Y - \hat{Y}||^2} = \frac{SSE/(p-1)}{SSR/(n-p)}$ (E :Estimated ; R : Residuals)

Decision rule

- si $F_{obs} > q_\alpha^F$, H_0 is rejected, and there exist a coefficient which is not zero. **At least one covariable is "useful" to explain the target**
- si $F_{obs} \leq q_\alpha^F$, H_0 is accepted, all the coefficients are supposed to be null
  The covariable are not "useful" to explain the model

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**
  - si $\frac{n-p}{p-1}\frac{R^2}{1-R^2} > f_{p-1,n-p}(1-\alpha)$
  - si pvalue $< \alpha$

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

Regression result illustration

Model :

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

R output :

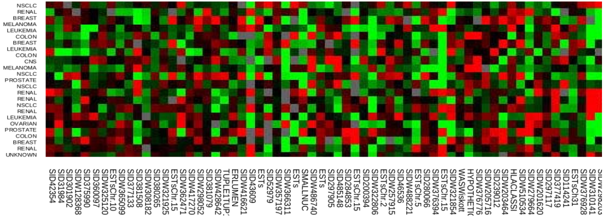
```
##
## Call:
## lm(formula = "cons~.", data = tab)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065302 -0.011873  0.002737  0.015953  0.078986
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1973151  0.2702162   0.730  0.47179
## 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 model
model selection

High dimentional 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.



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).

Linear modeling towards parsimonious models

1. Linear model
   - 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 $X^T X$ is invertible and :

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

Available algorithms to compute the solution :

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

"Optimality" result

Gauss-Markov theorem :

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

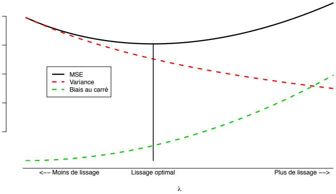
is optimal for the quadratic risk for in the non biased estimator family (BLUE : *best linear unbiased estimator*).

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

Generally

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

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



Linear model
model selection

Model selection in the linear Gaussian framework

Objective : Find the "most simple" models with a high power prediction 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*)

- $$\mathrm{RSS}(\mathcal{M}) \stackrel{def}{=} \|\mathbf{Y} - \mathbf{X}_{\mathcal{M}}(\mathbf{X}_{\mathcal{M}}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{X}_{\mathcal{M}}^T Y\|^2 \,,$$
- $$\hat{\mathcal{M}} \stackrel{def}{=} \underset{\mathcal{M} \subset \{1,\ldots,p\}}{\arg\min} \mathrm{RSS}(\mathcal{M}) + \text{penalty}$$
- 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 variable 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

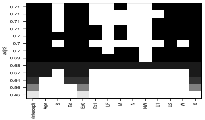
The value of R^2 mechanically increases with the number of variables.

Therefore, it is then not useful for model selection

- $R^2 = \frac{Var\hat{Y}}{VarY} = \frac{SSE}{SST} \in [0,1]$
- SSE : Sum Squared Estimated ; SST : Sum Squared Total

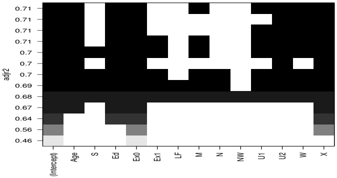
The Adjusted R-squared :

- Its expression uses a penalization which depends of the number of variables
- $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p} = 1 - \frac{RSS}{SST}\frac{n-1}{n-p}$
- 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 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 shows 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)

Evaluation of the predictive power of a model

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

The "rich man" approach : data sampling

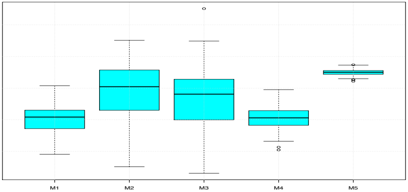
- Cross Validation
    - 50% to train the models (*training set*) ;
    - 25% to test and select the best model associated with the lowest RMSE error (*testing set*) ;
    - 25% to evaluate the best model (*evaluation set*).
- K Fold
- Leave one out

These approaches are extremely used for model selection the 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

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

Polynomial regression

Illustration of over-fitting.

Variables

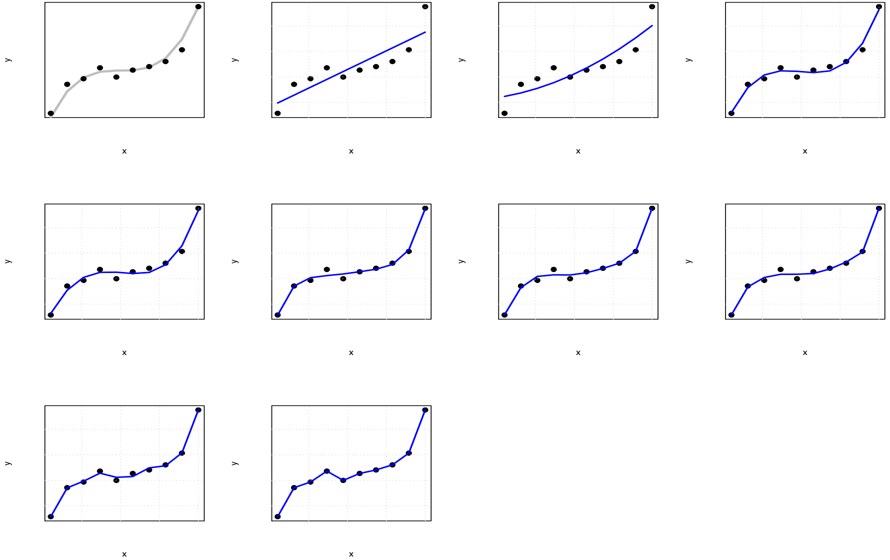
- Y :Target variable, $Y \in \mathbb{R}$
- X : Explanatory variable, $X \in \mathbb{R}$

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



Akaike criteria (AIC, 1973)

For the linear model, several criteria are introduced to penalized the Log-likelihhod.

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 (number of degrees of freedom)

Bayesien Information Criteria (BIC, Schwarz, 1976)

For the linear model, several criteria are introduced to penalized the Log-likelihhod.

BIC general expression

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

BIC vs AIC comparison

- The penality appears to be stronger ($\log n \gg 2$) ;
- 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

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

For the Gaussian Linear Model

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

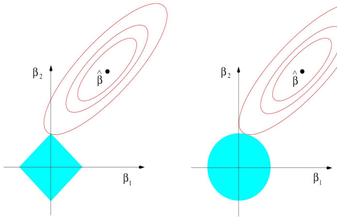
Linear model selection

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

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

Ridge, Lasso are alternative OLS method with Penalized coefficients...

Ordinary Least Square with a penalization on the coefficients



Penalized regression methods

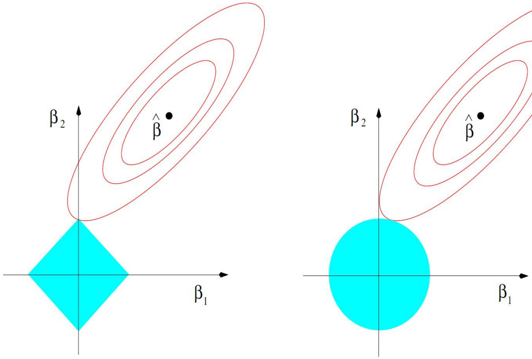
In this case, a constraint on the β coefficients is introduced in the OLS model :

- 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.

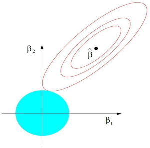
- ℓ_1 penalization induce sparse models. The value of "non useful" coefficients equal zero.
- ℓ_2 penalization helps to compute a solution in degenarative cases.

Penalized regression methods



Lasso et Ridge penalized methods

Ridge regression



Ridge Regression

Three different points of view :

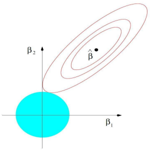
1. It's a solution to a penalized Least Square problem with smoothing properties
2. It induces a "contraction" of the original OLS coefficient values
3. It introduces a Gaussian "Apriori" in a Bayesian estimation

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 :

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

- 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 parameter value to all the coefficients of the model.
- The intercept should be never penalized. In practice, data are 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}

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

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

$\boxed{\textbf{Shrinkage estimator}}$

Ridge Regression. Gaussian apriori

We consider $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, σ^2 known.

We have : $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$

The likelihood is
$$\begin{aligned} L(Y/\{\beta,\sigma\}) &\propto exp\{-\tfrac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\} \\ &\propto exp\{-\tfrac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})\} \end{aligned}$$

Some similarities are observed with $\beta \sim \mathcal{N}_n(\hat{\beta}, \sigma^2(X^T X)^{-1})$

Ridge Regression. Interprétation bayésienne.

A priori Gaussien sur :

$\beta \sim \mathcal{N}_p(0, \sigma_\beta^2)$ et $\pi(\beta) \propto exp\{-\frac{\beta^T\beta}{2\sigma_\beta^2}\}$ avec $k = \sigma^2/\sigma_\beta^2$.

La densità a posteriori de β est

$$\begin{aligned} p(\beta/Y, \sigma) &= L(Y/\beta, \sigma)\pi(\beta) \\ &\propto exp\{-\frac{1}{2\sigma^2}[(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) + k\beta^T\beta]\} \\ &\propto exp\{-\frac{1}{2\sigma^2}[(\beta - \hat{\beta}(k))^T (X^T X + kI_p)(\beta - \hat{\beta}(k))]\} \end{aligned}$$

En posant : $\beta - \hat{\beta} = \beta - \hat{\beta}(k) + \hat{\beta}(k) - \hat{\beta}$ et $\beta = (\beta - \hat{\beta}(k)) + \beta$

la densità a posteriori de β est $\mathcal{N}(\hat{\beta}_{RR}^k, \sigma^2(X^T X + kI_p)^{-1})$

Ridge : Estimateur de Bayes avec un apriori Gaussien sur β

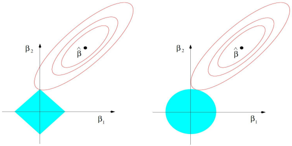
Si σ_β^2 grand (k petit), alors peu d'apriori sur β, l'estimateur Ridge est similaire à celui des MC0.

Ridge Regression

How to choose k ?

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

Lasso regression



lasso (gauche), ridge (droite)

Lasso Regression

- ℓ_1 Penalized OLS :

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

- 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 $\Phi(\beta)$:

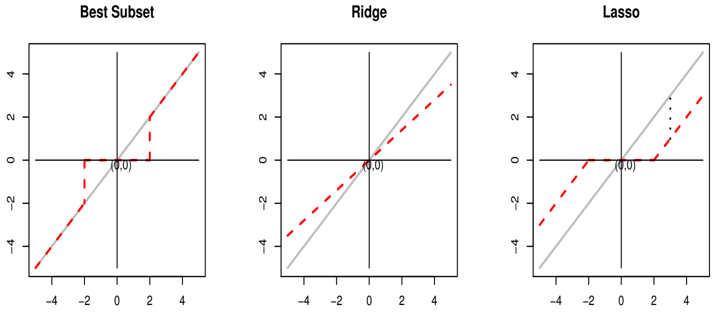
$\rightarrow$ The LARS algorithm is used in practice to compute the LASSO solution

Ridge et Lasso Regression

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

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

Ridge and Lasso Regression

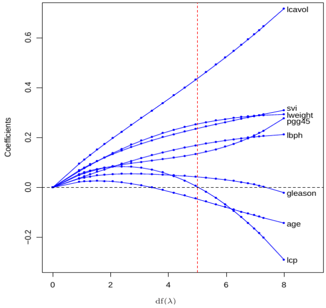


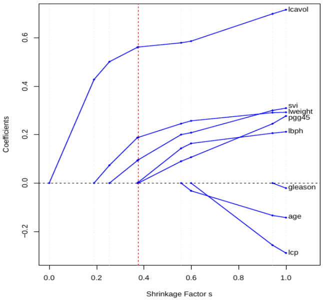
Best Subset, Ridge and Lasso Regression

Ridge and Lasso Regression

Regularization paths.

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

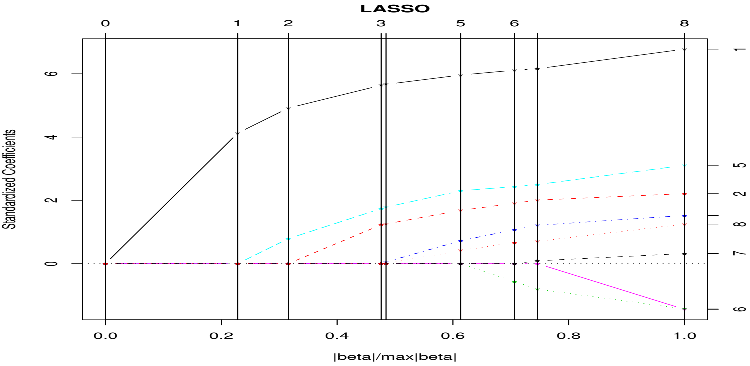




Ridge (left) et Lasso (right) Regression

Application

Study : Prostate cancer data $n = 97$ observations

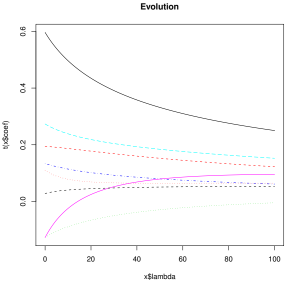


Lasso regularization path

Ridge Regression. Application

Study : Prostate cancer data $n = 97$ observations

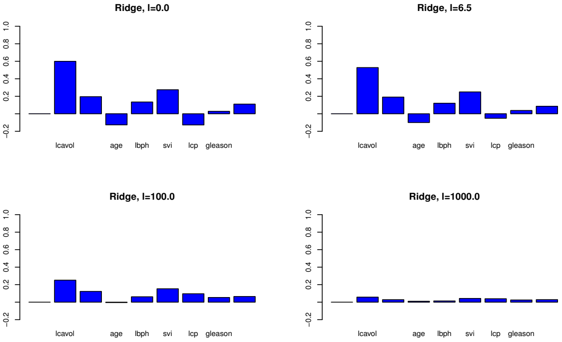
Y		lpsa
X	8	lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45



Ridge Regression. Application

Application : cancer data

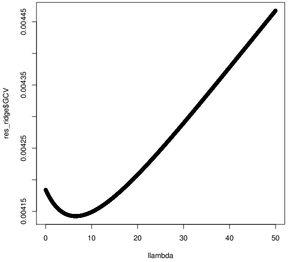
Values of the coefficients for several k penalized values



Ridge Regression. Application

Application : cancer data

Cross-validation error given the penalized coefficient value



Ridge Regression. Algorithme

```
library(MASS); # PROSTATE DATA
tab0 = read.table('prostate.data'); names(data)
tab=tab0[,1:(ncol(tab0)-1)]; names(tab);
tab=data.frame(scale(tab));
#Utilisation de la fonction solve pour calculer les coeffs de
régression
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));
#Utilisation de la fonction solve pour calculer les coeffs de
Ridge
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)
```