Travaux Pratiques - Modèles de Régression régularisée

October 3th 2023

Goal of the practical session

- Model selection for linear models in R. Ridge and Lasso regression

Remarks

- The work has to be carried out by a team of 2 students and `R studio` is used to perform the practical sessions.
- A report should be written **only for exercice IV** , automatically generated using a `R markdown` file format for 'R studio'.
- The 'R markdown file' and the corresponding pdf file have to be uploaded **before next practical session on the ENSIIE project web site in the folder MRR2023TP2**.

I. Tests of significativity and model selection

a) Analyze and study the following instructions. Specify the underlaying theoretical model.

```
n=100; X=cbind(((1:n)/n)^3,((1:n)/n)^4 ); Y=X%*%c(1,1)+rnorm(n)/4;
res=summary(lm(Y~X)); print(res); print(res$coef[2,4]);
```

Compare the results provided by a multiple regression model and the results computed independently using two simple models. Conclusion.

```
reg1=lm(Y~X[,1]);print(summary(reg1));
reg2=lm(Y~X[,2]);print(summary(reg2));
```

b) Execute the previous instructions several times (2 or 3 times) and describe the behaviour of the estimators of the coefficients. Compute the empirical correlation matrix. Instruction `cor()`.

```
cor(X[,1],X[,2])
```

II Model selection in a linear regression framework

Following table details several criteria used in model selection. We denote $RSS = \sum_i (y_i - \hat{y}_i)^2$;

Notation	Definition	Criteria	Objective	R Instruction
R^2	$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$	R-squared	-	`lm()`
R^2_{adj}	$R^2_{adj} = 1 - \frac{n-1}{n-p}(1 - R^2)$	Adjusted R-squared	Max. R^2_{adj}	`lm()`
$\hat{\sigma}^2_p$	$\hat{\sigma}^2_p = \frac{RSS}{n-p}$	Non biased residual est.	Min. $\hat{\sigma}_p$	Fonction `lm()`
AIC	$\simeq nlog(\frac{RSS}{n}) + 2p$	Aikaike Information (1971)	Min. AIC	`extractAIC()`
BIC	$\simeq nlog(\frac{RSS}{n}) + log(n)p$	Bayesian Information C (1978)	Min. BIC	`extractAIC(,k=log(n))`
C_p	$= \frac{RSS(p)}{\sigma^2} - (n - 2p)$	C_p Mallows (1973)	Min. C_p	`regsubsets()`

The `step()` function is used to compare and select parcimonious models (models based on few variables). The fonction starts from the global model and withdraw step by step one variable. The procedure stops when the coefficient of the variable which should be removed is significative ($\alpha = 0.1$ threshold).

Applications

Analyze the files "USCrimeinfo.txt" et "UsCrime.txt". The target variable, Y, is stored in the first column.

• Upload the file in the R environment using `tab=read.table()`. What is the number of available observations? Provide a scatterplot of all joint distributions. Conclusion.

• Compute the empirical correlation matrix. Conclusion. Use the `corrplot()` function of `corrplot()` library to highlight potential linear relations between variables.

A. Multiple regression model.

The goal is now to study the opportunity to use a linear model to explain the target variable Y. Specify the model.

a) What can you briefly say on the results provided by a linear model on the USCrime data set using the function `reg=lm("R~.",data=tab)` where Y denotes the target variable and X the explanatory variables ($p = 14$).

b) Does the linear model **globally** have a interest? Justify your answer.

c) What can you say about the significativity of the coefficients? Justify your answer.

d) Compute the Residual Sum of Square (RSS) in this case with $p = 14$ variables ?

B. Model selection.

The goal of this section is to find a sparse model based on a small subset of variables of size p_0 to explain the target variable Y. Prior writing any R instruction, read carefully the help of the R `step()` function.

a) Backward regression. Study and implement the following instruction

```
regbackward=step(reg,direction='backward'); summary(regbackward)
```

Comment the successive removed variables. What is the final model ? How may variables are selected ?

b) Forward regression.

```
regforward=step(lm(R~1,data=tab),list(upper=reg),direction='forward');
summary(regforward);
```

Comment the successive added variables. Compute the AIC criteria using the instruction `AIC()`. What is the final model ? How may variables are selected ? Compare this model with the model computed with the Backward regression method.

c) Stepwise regression:

```
regboth=step(reg,direction='both')
summary(regboth)
```

Comment the added and removed variables for the stepwise regression. Compare the selected models obtained with all the previous selection methods.

d) Remarks. Use the `formula(s0)` function where $s0$ denotes the R output object computed with the `step()` function. Note that the instruction `reg0=lm(formula(s0),data=tab);` let you use the computed selected model for further applications and that `summary(reg0)` provides you detailed information on this model.

III RIDGE and LASSO penalized regression.

A. Simulated data. Illustration.

a) Execute and comments the results using the following instructions.

```
rm(list=ls()); n=10000; p=5;
X=matrix(rnorm(n*(p)),nrow=n,ncol=p); X=scale(X)*sqrt(n/(n-1));
beta=matrix(10*rev(1:p),nrow=p,ncol=1); print(beta)
```

```
epsi=rnorm(n,1/n^2); Y=X%*%beta +epsi;
Z=cbind(Y,data.frame(X)); Z=data.frame(Z);
```

b) Considering a linear model, provide an estimation of the coefficients using X and Y data with the help of the `lm()` function. Conclusion.

Execute `t(X)\%*\%Y/n` and comment the result. The `lars()` function of R `lars` library can be used to implement a LASSO regression as the `glmnet()` function of R `glmnet` library. Upload the library in your R environment and read carefully the help of the function.

In this section, the goal is now to implement and study a linear model with a ℓ_1 penalization on the coefficients using X and Y data.

c) Execute:

```
library(lars);
modlasso=lars(X,Y,type="lasso"); attributes(modlasso);
```

What do the fields `modlasso\(meanx` and `modlasso\)normx` store ? What are they for ?

d) Comment the following graphs:

```
par(mfrow=c(1,2));
plot(modlasso); plot(c(modlasso$lambda,0),pch=16,type="b",col="blue"); grid()
```

e) Execute and comment the results? Why is it possible to guess, before any computation, the results computed with the LASSO in this situation? Justify carefully.

```
print(coef(modlasso));
coef=predict.lars(modlasso,X,type="coefficients",mode="lambda",s=2500);
coeflasso=coef$coefficients;
par(mfrow=c(1,1)); barplot(coeflasso,main='lasso, l=1',col='cyan');
```

B. Applications

The data studied in this section are indicators of development used for economy, demography, sociology in the United states over a period of 15 years. Our goal, in this application, is to identify the indicators which best explained the CO2 emissions observed in the athmosphere. For this purpose, the RIDGE and the lasso regression are both used and study.

a) Describe the content of the files "usa_indicators_info.txt" and "usa_indicators.txt".

b) Upload the data in the R environment using `tab=read.table()`. What are the number of observations and the number of variables? Can you use, in this situation, a multiple linear model ? Justify your answer.

c) What is the variable used for the CO2 emission ? Plot the temporal evolution on this indicator on a graph.

d) As the data correspond to various indicators, the units may also be very different. Explain how it can be a difficulty either for regular linear models or penalized linear models ? Scale the variables of the data set using function `scale(tab, center=FALSE)`.

e) Use the `lm()` function to estimate the parameters of a linear model. Conclusion.

RIDGE. Regression with ℓ_2 penalization.

a) Recall the definition of the Ridge regression.

The function `lm.ridge` of the `MASS` library is used to compute a Ridge regression. Upload the MASS library in your R environment, and read carefully the help of the function `lm.ridge()`.

b) Compute a Ridge regression for values of the penalization parameter equaled to $\lambda = 0$, $\lambda = 100$ without using the `Year` variable in your model. Print the computed coefficients using the instruction `coef()`. Plot the five largest coefficients. What do they represent? What are the differences between `coef(resridge)` and `resridge$coef` instructions ?

c) Compute ridge regression models for different values of λ starting from 0 to 10 with an increment of 0.01 ($\lambda = seq(0, 10, 0.01)$). Plot the performances computed by cross-validation given the values of λ (field `$GCV` of the ridge R object, GCV for Generalized Cross Validation). Plot the evolution of the values of the coefficients given λ using the instruction `plot(resridge)`. Conclusion. What model may you advice ? Print the corresponding value for the regularization parameter λ. Print and store automatically the parameters of the best model with the help of the functions `which.min()` and `coef()` #`coefridge=...`

d) Compute the mean quadratic error between the observed target and the estimated target $\hat{Y}_{ridge}$ using matrix computation where X denotes the input matrix: `Yridge=as.matrix(X)\%*\%as.vector(coefridge)`.

LASSO. Regression with ℓ_1 penalization

a) Compute a lasso regression using the instruction `reslasso=lars(X,Y,type="lasso")` where X denotes the input matrix and Y the target matrix.

Execute both following instructions: `plot(reslasso)` and `plot(reslasso$lambda)`. Comment.

b) Plot the values of the model coefficients for $\lambda = 0$ with the help of the instruction:

`coef=predict.lars(reslasso,X,type="coefficients",mode="lambda",s=0)`. Conclusion.

c) Plot the values of the coefficients for $\lambda = 100$. Conclusion.

Compare these results with the results already obtained with the ridge regression. Conclusion.

d) Compute the mean quadratic error between the observed target and the estimated target ($\hat{Y}_{lasso}$): `pY=predict.lars(reslasso,X,type="fit",mode="lambda",s=0.06)`.

e) How can you chose lambda ?

IV. Cookies Study

As a data scientist, you are now asked to study the `Cookies.csv` dataset. The aim of this study is to explain a fat indicator (the target variable, Y) given some spectra information.

The dataset

Each cookie is characterized by a fat indicator (scalar value) and a vector of size $p = 700$. In the dataset, each row stores one cookie information: the fat indicator is available in the first colum and the remaining values correspond to the spectra information. $n = 32$ cookies have been analysed.

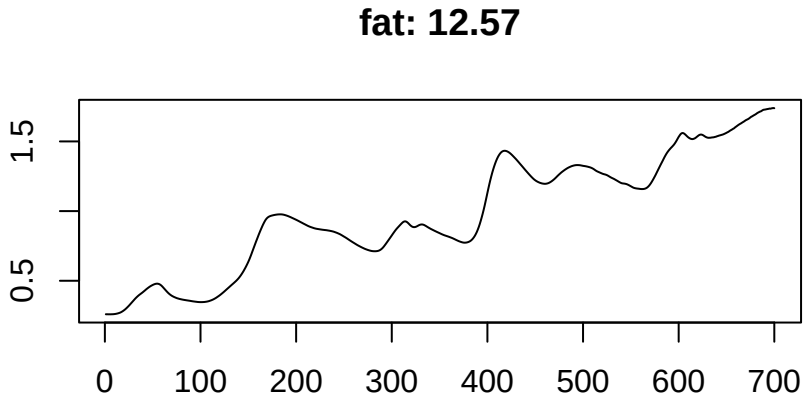

Figure 1: First cookie information

Regression model

Study and comment a linear regression model with ℓ_1 or ℓ_2 penalisation (or not) to explain the fat given the values of the spectra.