Practical session - Modèles de Régression linéaire

Modèles de Régression Régularisée (MRR)

2025

Goal of the session

- To understand the Ordinary Least Square (OLS) method and the linear model, from a methodological and practical point of view.
- To apply simple or multiple linear models on several data sets using the 'R' language.
- To interprete 'R' outputs of linear model functions.

Warnings and Advices

- The goal of this practical session is not "just to program with R" but more specifically to understand the framework of Modeling, to learn how to developp appropriate models for answering to a given operationnal question and a given data set. The MRR course belongs to the **Data Science courses** and is a preliminary step before using more advanced methods introduced in other 'machine learning' courses. → For each MRR practical session, you should **first understand** the mathematical and statistical backgrounds, **then write your own program with 'R'** to practically answer to the question.

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 exercices IV et V**, automatically generated using a **`R markdown`** file format for 'R studio'.
- The names of the Students should be mentioned in the first two lines of the document (no name, no grade!)
- 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 MRR2025TP1**.
- If you need some help to create a markdown file please refer to the web site https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet.
- In order to start any run with a clean environment, the two following lines should be always put in the beginning of your code: `rm(list=ls()); graphics.off()`.

If you are a beginner in R

Perform the first exercices of the swirl() package as mentionned in the first lesson.

Some preliminary exercices using R

1. **Vector manipulation**. Recall that

$$e^x = \sum_{k \geq 0} \frac{x^k}{k!}.$$

   Create a vector called `exp2` storing the 20 first elements of the previous expression. Remove all the values lower than 10^{-8} and compute an approximation of e^2. Finally, compare this result with the value obtained using directly `exp(2)`.

2. **Data simulation**. Use the fonction `rnorm` (`?rnorm` to get some help) to simulate a vector X of size 100 drawn from a Gaussian law $\mathcal{N}(2, 1)$, with a mean equaled to 2 and a variance equaled to 1. Compute a second vector Y of the same size by multiplying X with the value 9.8 and by adding a Gaussian noise of zero mean and a standard deviation equaled to 1/10.

3. **Read and write a text file**. Store both X and Y vectors in a data frame (`?data.frame` if necessary) and then store this `data.frame` in a text file using the instruction `write.table` on your hard disk. In a second step, upload the data from the text file into the R environment using the instructionn `read.table`. Compare the values of the data before and after the storage. Conclusions.

4. **Read and write a RData file** . Store both X and Y vectors in a `data.frame`. Store the previous `data.frame` using the instruction `save`. Then, use the instruction `load` to upload the data from the previous file to the R environment. Compare the values of both tables. Conclusion.

5. **Scatter Plot**. Plot the values of Y function of X, first using first the `plot` function then the `ggplot2` function.

6. **Histogram**. Draw the histogram of X. How can you modify the numbers of bins? Compare visually two computed histograms using a small and a large number of bins. Conclusion.

7. **Loop for**. We consider here a random variable which follows a given χ^2 distribution with an unkown degree of freedom. The goal is here to be able to recover the apriori unknow degree of freedom using the computation of its empirical mean using n observations. Write a `R` program to evaluate the accuracy of this estimation for $n = 3$ then $n = 100$. Use a loop `for` to solve this exercice.

8. **Loop for**. Same exercice as previously using the 'sapply' instruction. Conclusion.

I. Ordinary Least Square (OLS) / Moindres Carrés Ordinaires (MCO)

Real estate transactions in Paris Study.

The "immo.txt" file contains a set of real estate transactions in Paris. The column variables contains: col 1: "the surface of the appartment in m^2"; col 2: "the prize of the previous sale of the appartment several years ago"; col 3: "the prize of the transaction in K-euros".

Preliminary work

- Upload the "immo.txt" file into the R environment with the help of the `read.table()` function and save the data in a dataframe called `tab`. The name of the variables of the dataframe should be automatically defined using the information provided in the first line of the text file.
- Execute the following instructions: `head(tab)`, `names(tab)`, `tab[,1]`, `tab$surface`, `tab[,c(1,3)]`, `tab$prix`. Conclusions. What are the number of observations of the data set? (instructions: `nrow(tab)` `dim(tab)`)
- Execute the following instruction `plot(tab)`. Compute the correlation matrix using the function `cor()` and comment the results.

First model using Ordinary Lesat Square (OLS)

The goal is now to build a model able to linearly explain the prize of the real restate transaction (here the target variable) using the other available explanatory variables $X_1, \ldots, X_p$, $p = 2$, $n = 20$. The model is here defined by:

$$Y_i = \beta_0 + \sum_{j=1}^{p=2} \beta_j X_{ij} + \epsilon_i$$

The residual ϵ_i is defined by the difference between the observed target and the estimated target using the model.

The $\hat{\beta}_j$, $0 \leq j \leq p$ coefficients are computed by the OLS method.

$$E(\beta) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \sum_{j=1}^{p=2} \beta_j X_{ij}))^2$$

a) Using appropriate matrix notations, recall the value of the estimated coefficients using the OLS method $\hat{\beta}_j$ $0 \leq j \leq p$.

b) Compute with the help of the R software, the OLS estimated values of the coeffcients using the data set with the help of the instruction `modreg=lm(prix~.,data=tab)`. Execute sequencially the following instructions and comment the obtained result: print(modreg); summary(modreg); attributes(modreg); coef(modreg); modreg$res.

Using the appropriate `help()` function, describe the fields `modreg$res`, `modreg$model` of the R object provided by the function `lm()`.

c) Plot the bivariate distribution $(Y_i, \hat{Y}_i)$, $1 \leq i \leq n$ using the `plot()` function. Use the `grid()` function to plot a grid and draw the bissetrice line with the help of the `abline()` function. Conclusions. What is the benefice of such representation ? What are the under estimated values ? the over estimated values ? Plot another graph to visualize residuals of the model $\hat{\epsilon}_i = Y_i - \hat{Y}_i$. Conclusion.

d) Recall the definition and the geometrical interpretation of the R-square? Then compute the R^2 defined by $R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$ with $\bar{Y} = \sum_j Y_i/n$. Conclusion.

e) Using your own matrix computations, compute "by hand" the values of the estimated coefficients: $\hat{\beta} = (X^T X)^{-1} X^T Y$ with the help of the following functions if necessary `solve()`, `matrix()`, `as.matrix()`, `cbind()`.

II. The linear model. Study of ice cream consumptions.

- The aim of this section is to explain, based on a given data set, the consumption of ice cream regarding $p = 3$ chosen explanatory variables: the tax value (income), the ice cream price (price) and the average temperature (temp) over the considered period (for more information see the "Icecreaminfo.txt" file). The "Icecreamdata.txt" file contains the corresponding data set.
- Upload in the `R` environment the data set using the instruction `tab=read.table(..)`. What is the number of available observations ? (functions: size(tab); dim(tab))

a) The aim is now to build a linear model able to explain the ice cream consumption given the other co variables. Present and write formally the linear model to be study. Use then 'R' to study the ability of the linear model for this problem.

b) Estimated coefficients

- What are the values of the estimated coefficients? What can you say about their values? Comments.
- Recall the 'statistical test' used to test the significativity of each coefficient of the model.
- Recall the signification of the labels `***`, `**`, `*` specified for each coefficient.
- Recall how to use the p-value in this situation. Draw your conclusions in this case and add assumtions if needed.
- What are the limits of such approach ?
- Using matrix computations and writing your own instructions, find again 1/ the value of the coefficients, 2/ the value of the statistics of the test and 3/ the associated-pvalue. Use the slides of the lesson if necessary to develop your own code. Read the help of the Student Law function `help(rt)`.
- Using the function `confint()`, compute a confidence interval, for each of the coefficient with a risk of 5%, 1% and 0.001 . Explain the link between the computed intervals and the labels `*`, `**`, `***` previously obtained for each coefficient.

c) Predicted targets.

- Plot the predicted targets given of observed targets for this data set and linear modelby using the appropriate field of the `lm()` function. Conclusion.
- Compute the confidence for the predicted values with a risk of 5% using the 'R' function `predict()` (option confidence) with the appropriate parameters.

d) Residuals.

- Compute the **root mean squared error (RMSE)** of the model and compute a non biased estimated of the residual variance of the model.
- Plot the residuals $\hat{E}$ (`res$residuals`) given the values of the real targets Y_i. Conclusion.
- Study the empirical distribution of the residuals (`qqnorm`, `qqline`). What is your conclusions of using a linear model in this situation i.e. for the prediction of ice cream consumption ?
- Test de normality of the residual distribution using the Shapiro test `shapiro.test()`. Conclusion.

e) Predictive values. Compute an estimated value of icecream consumption fot the following values of the explanatory variables: income=85, price=0.28, temp=50 ? For this question, the parameters of the model are computed using all the data available in the data set. Provide a confidence interval for this prediction.

f) Predictive power of the model. The goal, of this part, is to study the `ability` of the linear model to compute fair predictions for 'new' observations, which do not belong to the initial data set used to estimate the coefficients of the model. For this purpose, a partition of the Ice-cream data set is previously performed.

- **Random Partitionning**.
- Split randomly the initial dataset in two dataframes containing respectively 75% of the observations (the 'Training' data base, TabTrain) and 25% of the remaining observations (the 'Test' data set, TabTest) using the 'R' function `sample()`.
- Use the training data set to estimate the parameters of the model and compute the RMSE on these data. Given this previous model, use the test data set to compute the RMSE to evaluate the performances of the model on the test data set using the 'R' function `predict()` or `predict.lm()`. Be careful not to compute a new estimation of the coefficients using the test data set! The test data set has to be 'independant' of the training set.
- Repeat the two first steps (randomly spliting the computing both RMSE) 10 times and compare the results obtained with the help of 2 boxplots ('R' function `boxplot()`). Conclusion?

III. Curse of high dimension

We decide to successively add new variables in the 'IceCream' dataset, randomly generated $\mathcal{N}(0,1)$ and to study the impact on the modeling.

a) Compute an estimation of the parameters of the linear model with $k = p + 1$ variables (4 variables + intercept), the extra variable is randomly generated $\mathcal{N}(0,1)$. Compute the RMSE $E_P = \sqrt{\sum_j (Y_i - \hat{Y}_i^k)^2/n}$ where $\hat{Y}_i^k$ denotes the predited targets computed with $k = 4$ variables.

b) Add succesively $k = 2, 3, 4...20$ new variables into the first data set, randomly chosen $\mathcal{N}(0,1)$. For each data set (with a new variable), compute an estimation of the coefficients of the model, the $RMSE(k)$ on the training data set, and the R-square ($R^2(k)$) Plot the $RMSE(k)$ and the $R^2(k)$ given the number of adding variables for $k = 1, 2, ...20$. Conclusion.

IV. Application: study your own data using a linear model with transformed data

The grow of several compagnies belonging for example to the GAFAM or BATX is spectacular (GAFAM: Google, Apple, Facebook Amazon Microsoft ; BATX: Baidu, Alibaba, Tencent, Xiaomi) As a data scientist, you are now asked to study the grow of such compagny.

After having gathered about twenty numerical values which illustrate the growth of a companiy during the last years, propose an appropriate statistical model based on the linear model allowing to explain the growth of this company over time. The source of the data and the list of raw data values will be directly saved in the R code. Discuss the results obtained (as comments in the code or document).

V. Electricity Data Set

As a data scientist, you are now asked to study the `Mexico_data.csv` dataset. This dataset contains several meteorological parameters and the Total amount of electricity daily generated in Mexico city. The aim of this study is to explain the amount of generated electricity thanks to several other variables stored in the data file.

The dataset

In the dataset, each row corresponds to one observation. The variables are:

-X0: day of the year

-RH: Relative humidity (%)

-SSRD: Surface solar radiation downward (J.m-2)

-STRD: Surface thermal radiation downward (J.m-2)

-T2M: average daily temperature at 2m (°C)

-T2Mmax: maximum daily temperature at 2m (°C)

-T2Mmin: minimum daily temperature at 2m (°C)

-Covid: COVID-19 stringency index

-Holidays: holidays or public holidays, 1 = holidays, 0 otherwise

-DOW: Day of Week, 0 for Monday, 1 for Tuesday, ...

-TOY: Day of the year (from 1 to 366)

-Total: Electricity generated daily in the territory (GWh)

Data Study

Bring interesting conclusions, by studying this dataset.