ISEP
IIF.1105 - IF.2301 - Data Science
Academic year 2025-2026

DATA SCIENCE PROJECT
Principal Component Analysis (PCA) and Linear Regression

1 Instructions to read carefully

In this project you will perform Principal Component Analysis and Linear regression with real data. You will work in groups of **3 students**. You will have to prepare a presentation and pass an oral defense. You can use any programming language that performs PCA and Linear regression, however, the use of Python is highly recommended. The instructions are the following :

About the oral defense

The defense will last about 15 minutes per group and it will consist in 10 minutes of oral presentation plus 5 minutes of questions. You should prepare a presentation with the following (minimal) content :

- a cover page with the first name, last name and the student identification number of all the members.
- a table of contents,
- a short introduction,
- the main body of the presentation (results, figures, tables, interpretations, comments or any other element that might help you answer the questions). In this part, you should answer all the questions referred to as [*graded question*] . If necessary, you can use up to three significant digits in your numerical results.
- the conclusion
- the references

It is not necessary to include your code in the presentation. However, you should have it at hand, in case, you have any related questions.

You will find in the hyperplanning the date of the your oral defense. You should submit the presentation file in pdf format one day before the oral defense. To this end, in moodle you will find a deposit box to upload the file. The file name must have the following format :

LastNameStudent1_ LastNameStudent2_ LastNameStudent3.pdf

Just one deliver per group must be done. There is no report to submit, only the presentation ! The language of the presentation can be either French or English.

About the evaluation

The oral defense is divided in 2 parts, an oral presentation and questions. The quality of the oral presentation will be appreciated and it **should not exceed 10 minutes**. It must be clear, explicit and well understandable. During the second part, in turn each member of the group will be asked some questions. The quality of the answers in terms of comments, interpretations and reasoning will be taken into account for the final mark. The evaluation is individual.

2 Data analysis

2.1 The dataset

The dataset *prostate.txt* contains information about patients with prostate cancer. The prostate is a gland in the male reproductive system. Cancer develops from the tissues of the prostate when cells mutate and multiply uncontrollably. These can then spread (metastasize) by migrating from the prostate to other parts of the body. Like most types of cancer, the earlier it is detected, more chances of recovery the patient has.

The *Prostate-specific antigen* is a protein normally secreted by prostate cells, however a cancer cell secretes 10 times more than a normal cell. Hopefully, this property might be useful for detecting prostate cancer. However, the level of the prostate-specific antigen can be increased by many other factors such as the prostate volume, infections and/or inflammation) or even decreased by certain treatments.

The objective of this study is to better understand the factors influencing the level of prostate-specific antigen. A study has been conducted on men with prostate cancer and who have undergone radical prostatectomy, i.e. complete surgical removal of the prostate. Before the surgery, the level of *prostate-specific antigen* was determined by a blood test. The tissues removed during the operation were examined in order to characterize the cancer more precisely.

The dataset *prostate.txt* contains the following variables :

- volume of cancer (`vol`).
- prostate weight (`wht`).
- patient age (`age`).
- quantity of benign hyperplasia (`bh`). Benign hyperplasia is a benign tumor of the prostate.
- capsular penetration (`pc`). The higher the variable, the more the cancer crosses the capsule surrounding the prostate to reach the neighboring structures.
- level of prostate-specific antigen (`psa`).

2.2 Preliminary analysis : descriptive statistics

Import the dataset *prostate.txt* and name it `prostate`. Get familiar with the data by performing the following tasks and answering the questions :

1. [*graded question*] How many observations are there ? How many variables ? Are there any missing values in the dataset ?
2. [*graded question*] Calculate descriptive statistics for all the variables. You can use graphics of your choice to help you describe the data (boxplot, scatter plot, etc.). Interpret the results.

By plotting the scatter plots between `psa` and each of the other variables you will notice that these plots display large accumulations of points in small areas. To solve this, it is usual to consider the logarithm rather than the original values. Perform a logarithmic transformation of all the variables dataset except `age` and change the names of the transformed variables by preceding the name with letter l (example `lvol` instead of `vol`). Visualize again the scatter plots and interpret the results.

Warning : Thereafter, all the analysis will be performed using the transformed values.

2.3 Principal Component Analysis (PCA)

1. *[graded question]* **Theoretical question** If two variables are perfectly correlated in the dataset, would it be suitable to include both of them in the analysis when performing PCA ? Justify your answer. In contrast, what if the variables are completely uncorrelated ?

Practical application :

1. *[graded question]* Calculate the variance of each variable and interpret the results. Do you think it is necessary to standardize the variables before performing *PCA* for this dataset ? Why ?
2. *[graded question]* Perform PCA using the appropriate function with the appropriate arguments and options considering your answer to the previous question. Analyze the output of the function. Interpret the values of the two first principal component loading vectors.
3. *[graded question]* Calculate the percentage of variance explained *(PVE)* by each component ? Plot the PVE explained by each component, as well as the cumulative PVE. How many components would you keep ? Why ?
4. *[graded question]* Display both the principal component scores and the loading vectors along with the correlation circle. Interpret the results.

2.4 Linear Regression

In this part, you will perform linear regression to predict the target variable (`lpsa`) as a function of the other variables in the datasets.

[graded question] **theoretical question :** Let us suppose that we fit a linear regression model to explain Y as a linear function of two variables X_1 and X_2. Let us denote R^2 the associated coefficient of determination. Interpret R^2. What is the range of values that can be taken by R^2 ? If we denote r_1 and r_2 the coefficient of correlation between X_1 and Y and the coefficient of correlation between X_2 and Y respectively. What is the relationship between R^2 and r_1 and r_2 ?

2.4.1 Simple linear regression

[graded question] Calculate the correlation between the target and all the other variables in the dataset. Which variable is the most correlated with the target `lpsa` ? Comment on the results.

[graded question] Fit a simple linear regression model using as target variable `lpsa`, denoted Y, and as feature variable the most correlated variable to it that you identified in the previous question, denoted X :

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

Then, answer the following questions :

1. Which variable is correlated the most with `lpsa` ? What is the value of the coefficient of correlation ?
2. What are the coefficient estimates ? Interpret coefficient estimate $\hat{\beta}_1$.
3. Give the general expression of a $1 - \alpha$ confidence interval for the parameter β_1. Calculate the 95% confidence interval for this coefficient. Interpret the results.

4. Elaborate the zero slope hypothesis test for coefficient β_1 and conclude if there is an impact of the predictor on the log of the level of the prostate-specific antigen (`lpsa`). Is β_1 significantly non zero ?
5. What is the value of the coefficient of determination R^2 ? Interpret this result. Is this model suitable to predict the level of the prostate-specific antigen ?

2.4.2 Feature selection for multiple linear regression

Now you are going to fit multiple linear regression models in order to predict the target variable `lpsa` as a function of two or more other predictors.

In some practical situations it is suitable to select only a subset of the predictors instead of considering all the available variables, since some variables can have no or just little statistical significance to predict the target. The *best subset selection* method consists in fitting a separate least squares regression for each possible combination of the available features. In Python you can use the function `combinations()` of the module `itertools` to get all the possible combinations of k predictors for $k \in \{1, ...5\}$.

Perform the following tasks and answer the questions :

1. [*graded question*] Use Best Subset Selection method to select the best model for any possible number of features ranging from 1 to 5. Select the best model. That is, the model for which the adjusted coefficient of determination $\bar{R}^2$ is the highest.
2. [*graded question*] How many features did you keep ? Which ones ?
3. [*graded question*] Why is it more appropriate to use the adjusted coefficient of determination $\bar{R}^2$ instead of the coefficient of determination R^2 when comparing two models with different numbers of predictors ?
4. [*graded question*] For the selected model, what are the values of the coefficient estimates ? Interpret them. What is the value of the coefficient of determination R^2 ? Interpret this value.
5. [*graded question*] For the selected model, perform the zero slope hypothesis test for all the coefficients except β_0 and conclude.
6. [*graded question*] For the selected model, make a prediction of the level of prostate-specific antigen (`psa`) given the following conditions volume `vol` $= 7.2$; prostate weight equal to `wht` $=$ 22 ; quantity of benign hyperplasia `bh` $= 1.5$ and capsular penetration `pc` $= 0.26$ for a patient aged 67 years old. Do not forget to perform a logarithmic transformation of the original values previously.

3 References

— Stamey TA, Kabalin JN, McNeal JE, Johnstone IM, Freiha F, Redwine EA, Yang N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. J Urol. 1989 May ;141(5) :1076-83.