ISEP
IF.1205 - Data Science
May, 2024

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 Python, R or any other language that performs PCA and Linear regression. 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 authors.
- 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 R or Python code in the presentation. However, you should have your code 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 purpose of this study is predicting the age of abalone from physical measurements. Traditionally the age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope which is a so much time-consuming task. In contrast, other characteristics ans measurements easier to obtain can be used to predict the age.

The *abalone* dataset (file *abalone_ data.csv*) contains the following variables describing abalones :

Name	Unit of measurement	Description
`Sex`	-	M or F, or I (infant)
`Length`	mm	Longest shell measurement
`Diameter`	mm	perpendicular to length
`Height`	mm	with meat in shell
`Whole weight`	grams	whole abalone
`Shucked weight`	grams	weight of meat
`Viscera weight`	grams	gut weight (after bleeding)
`Shell weight`	grams	after being dried

The last variable, `Rings`, that is the number of rings which can be used to estimate the age by adding 1.5 units. That is, if the number of rings is 5, the age of the abalone will be approximately 6,5 years.

2.2 Preliminary analysis : descriptive statistics

Import the datafile *abalone_ data.csv* into a dataframe. Get familiar with the data and answer the questions :

1. [*graded question*] How many observations abalones are described ? How many variables are there ?
2. [*graded question*] Are there any missing values in the dataset ? If any, delete the observations with missing values.
3. [*graded question*] Calculate descriptive statistics for all the variables (mean, max, quartiles, etc.). Interpret the output statistics. Are there any qualitative variables ? If so, what are their categories ? You can use graphics of your choice to help you describe the data (boxplots, scatter plots, histograms, etc.). Interpret the graphics.

2.3 Principal Component Analysis (PCA)

Theoretical question

1. [*graded 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 : You are going to perform PCA with the *abalone* dataset. In this section you are going to consider all the variables except the `Sex`.

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*] Use a biplot with a correlation circle to display both the principal component scores and the loading vectors in a single plot. Interpret the results.

2.4 Linear Regression

[*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 ?

Remind that the purpose is to predict the age of the abalone. The age is calculated by adding 1.5 units to the number of rings. Create the new variable `Age` for all the observations by adding 1.5 units to `Rings`.

2.4.1 Simple Linear regression

[*graded question*] Calculate the correlation coefficient between `Age` and each of the other variables (except `Sex` ans `Rings` of course). Comment on the results. Which variable is the most correlated with the target `Age` ?

[*graded question*] Fit a simple linear regression model using as target variable `Age`, 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. What are the coefficient estimates ? Interpret coefficient estimate $\hat{\beta}_1$.
2. 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.
3. Elaborate the zero slope hypothesis test for coefficient β_1 and conclude if there is an impact of the predictor on the `Age`. Is β_1 significantly non zero ?
4. What is the value of the coefficient of determination R^2 ? Interpret this result. Is this model suitable to predict the `Age` of an abalone ?

2.4.2 Multiple linear regression

Now you are going to fit multiple linear regression models in order to predict the target variable as a function of two or more 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 [1]. 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 4. Plot the curve $\bar{R}^2$ versus the number of features. Then, 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. Perform the F-test and conlcude.

2.5 Multiple linear regression with a qualitative variable (Bonus)

In this part you are going to include the variable `Sex` in your model.

1. [*graded question*] How many observations are in each category of the variable `Sex` ?
2. [*graded question*] Plot a boxplot of the target variable `Age` versus the `Sex`. Comment on the output.
3. [*graded question*] Perform multiple linear regression by adding the `Sex` as explanatory variable to the model selected in the previous section. Interpret the coefficient estimates of the variable `Sex` (for each category),perform the zero slope test and conclude.
4. [*graded question*] For the fitted model make a prediction for an infant abalone with the following characteristics : `Length` = 0.4 mm, `Diameter` = 0.35 mm , `Height` = 0.12 mm , `Whole weight` = 0.40 g, `Shucked weight` =0.15g, `Viscera weight` = 0.08 g and `Shell weight` = 0.13 g.

3 Conclusion

Do not forget to draw your conclusions !

4 References

— Nash,Warwick, Sellers,Tracy, Talbot,Simon, Cawthorn,Andrew, and Ford,Wes. (1995). Abalone. UCI Machine Learning Repository. `https://doi.org/10.24432/C55C7W`.

1. In `R` the `regsubsets()` function of the `leaps` library performs best subset selection by identifying the best model that contains a given number of predictors, where best means the one that minimizes the RSS (residual sum of squares). In Python you will need to write the code by using a `for` loop and the function `combinations()` of the module `itertools`. Alternatively you can use the `R` function for this part even if you used Python in the rest of the project