ISEP
IIF.1105 - IF.2301 - Data Science
January, 2023

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 recent 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 etc 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 question-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 `Live` dataset contains statistics about posts in Facebook pages of 10 Thai fashion and cosmetics retail sellers from March 2012 to June 2018. Each observation (row) represents a post of different nature (video, photo, status or link). The features (columns) are variables describing each post :

- `status_type`: type of post : link, photo, status or video.
- `status_published` : date and time of the post.
- `num_reactions`, `num_comments`, `num_shares` are the number of reactions, comments and shares.
- `num_likes`, `num_loves`, `num_wows`, `num_hahas`, `num_sads` and `num_angrys` are emoji reactions.

2.2 Preliminary analysis : descriptive statistics

Import the datafile *Live_20210128.csv*. Get familiar with the data and answer the questions :

1. [*graded question*] How many observations are there ? How many variables ?
2. [*graded question*] Are there any missing values in the dataset ? If you think it is appropriate, delete the variables concerning missing values.
3. [*graded question*] Calculate descriptive statistics for all the variables except `status_type` and `status_published`. You can use graphics of your choice to help you describe the data (boxplot, scatter plot, etc.). Interpret the results.

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.
2. [*graded question*] In contrast, what if the variables are completely uncorrelated ?

Practical application : Now you are going to perform PCA using only the variables `num_comments`, `num_shares`, `num_likes` and `num_loves`. For the PCA you will consider only these four features.

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 ?

Practical application

It is well-known that given a post in social media, the number of reactions (likes, shares, comments, etc.) have a big impact on the visibility of the post. Among all reactions the one which Facebook gives most priority to appear on a user's news feed is *shares*. Indeed, sharing implies reposting to share with friends and followers. Furthermore it is possible to share a post leaving a comment. In this part, you are going to perform linear regression using the number of shares `num_shares` as the target variable as a function of the other variables.

[*graded question*] Perform an initial analysis of the variable `num_shares` based on the others by calculating the correlation coefficient between `num_shares` and each of the other variables except `status_type`, `status_published` and `num_reactions`. Which one is the most correlated with `num_shares` ?

[*graded question*] Fit a simple linear regression model using as target variable `num_shares`, 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 number of shares. 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 number of shares ?

Feature selection for multiple linear regression

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

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 `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 (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). 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 6. 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.

3 References

— Nassim Dehouche and Apiradee Wongkitrungrueng. Facebook Live as a Direct Selling Channel, 2018, Proceedings of ANZMAC 2018 : The 20th Conference of the Australian and New Zealand Marketing Academy. Adelaide (Australia), 3-5 December 2018.