ISEP
IIF.1105 - IF.2301 - Data Science
January, 2022

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

You should not include your R or Python code in the presentation. However, 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 personal.

2 Data analysis

2.1 The dataset

It is well-known that the Garment Industry, is a highly labour-intensive industry with lots of manual processes. Satisfying the huge demand for garment products highly depends on the performance of the employees in garment manufacturing companies (see the references).
The file *garments_worker_productivity.csv* contains a dataset about productivity in Garment industry. This dataset is public dataset was dowload from the UCI repository website `https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees`. The file contains the following variables :

- *date* : Date in format MM-DD-YYYY
- *day* : Day of the Week
- *quarter* : A portion of the month. A month was divided into four quarters.
- *department* : Associated department.
- *team_no* : Associated team number.
- *no_of_workers* : Number of workers in each team
- *no_of_style_change* : Number of changes in the style of a particular product
- *targeted_productivity* : Targeted productivity set by the Authority for each team for each day.
- *smv* : Standard Minute Value, it is the allocated time for a task
- *wip* : Work in progress. Includes the number of unfinished items for products
- *over_time* : Represents the amount of overtime by each team in minutes
- *incentive* : Represents the amount of financial incentive that motivates a particular course of action.
- *idle_time* : The amount of time that the production was interrupted
- *idle_men* : The number of workers who were idle due to production interruption
- *actual_productivity* : The actual productivity in percentage that was delivered by the workers. It ranges from 0-1. This is the target variable.

You are tasked to analyze the factors that have a strong impact on productivity and predict the productivity performance of the working teams in their factories.

2.2 Preliminary analysis : descriptive statistics

Import the dataset *garments_worker_productivity.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 so, how many observations contain missing values and which variables are concerned with this missing data ?
3. *[graded question]* **Theoretical question :** Without performing any calculation, suggest at least two methods to deal with missing data.

   IMPORTANT : Now, you will create a new dataset by deleting all the observations containing missing values. Hereafter, you will use this dataset.

4. *[graded question]* Calculate descriptive statistics for the target variable `actual_productivity`. Interpret the results.

5. *[graded question]* Perform an initial analysis of the variable based on the others by calculating the correlation coefficient between `actual_productivity` and each of the other variables. Which ones are the most correlated with the `actual_productivity` ?

2.3 Principal Component Analysis (PCA)

[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.

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 following variables : `targeted_productivity`, `smv`, `over_time`, `incentive`, `no_of_workers` and `actual_productivity`. You will use the appropriate function with the appropriate arguments and options considering the previous question. Analyze the output of the function. Interpret the values of the two first principal component loading vectors ?
3. *[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.
4. *[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 ?

2.4 Linear Regression

[graded question] **theoretical question :** Let us suppose that we fit a simple linear regression model to explain Y as a linear function of X. What is the relationship between, the correlation coefficient between these two variables $r(X, Y)$ and the coefficient of determination R^2 obtained by fitting the model ? What is the range of values that can be taken by R^2 ? How about r ?

Practical application

In this part you are going to fit a linear regression model in order to predict the target variable `actual_productivity`. First of all, consider the following simple linear regression model :

$$\texttt{actual\_productivity} = \beta_0 + \beta_1 \texttt{incentive} + \epsilon \qquad (1)$$

Fit the model given in (1) and answer the following questions :

1. *[graded question]* What are the coefficient estimates ? Interpret coefficient estimate $\hat{\beta}_1$.
2. *[graded question]* 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. *[graded question]* Elaborate the zero slope hypothesis test for coefficient β_1 and conclude if there is a relationship between the real productivity and and the financial incentive. Is β_1 significantly non zero ?
4. *[graded question]* What is the value of the coefficient of determination R^2 ? Interpret this result. Is this model suitable to predict the productivity ?

Feature selection

Now you are going to fit multiple linear regression models in order to predict the target variable `actual_productivity` as a function of some of all the following feature variables : `targeted_productivity`, `smv`, `wip`, `over_time`, `incentive`, `no_of_workers`.

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. Perform the following tasks and answer the questions :

1. [*graded question*] Use Best Subset Selection 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 ?
5. [*graded question*] For the selected model, perform the zero slope hypothesis test for all the coefficients except β_0 and conclude.

3 References

— Imran, A. A., Amin, M. N., Islam Rifat, M. R., and Mehreen, S. (2019). "Deep Neural Network Approach for Predicting the Productivity of Garment Employees". In 6th International Conference on Control, Decision and Information Technologies (CoDIT).

— Rahim, M. S., Imran, A. A., and Ahmed, T. (2021) : "Mining the Productivity Data of Garment Industry". In International Journal of Business Intelligence and Data Mining, 1(1).