Lab Project

STAT 6106 - Statistical Inference for Data Analytics

Release date: Wednesday 31st March 2021

Due date: Presentation - Friday 16th April 2021, 4:00 PM
Technical report - Sunday 18th April 2021 8:00 PM

Total: 10% (5% each component)

OBJECTIVE:

The objective of this project is to build and apply a significant predictive model in the statistical software *R* using one of the following three statistical methods covered in class:

(1) Multiple regression
(2) Logistic regression
(3) Principal component analysis

DATA:

You may use a data set of your choice which can either be a built-in data set, a manually entered data set or a data set obtained from online then imported into *R*. Note that the data set should be suited to the method chosen, for example a data set which contains one or more binary and/or categorical variables can only be used in logistic regression.

BACKGROUND:

You are to formulate a problem or objective around your data set and aim to solve this objective. For example in lab 3 (*gold.csv* data set) the objective may be to reduce mining costs by coming up with a significant model which investigates the relationship between gold deposits and factors of the soil. An application of this model can therefore be to predict the probability of gold deposited under specific soil conditions and use this probability to decide whether to mine for gold.

The project is divided into two parts:
1. Presentation
2. Technical report

DESCRIPTION OF PRESENTATION

Each student has to prepare and present a 5 minute presentation on Friday 16th April 2021 (during class time 4PM - 6PM). The presentation titled ID.pdf or ID.pptx **must** be submitted through myelearning before 4 PM (TT time) on the said day.

The presentation should include but not be limited to the following sections:

1. Introduction - provide an introduction to the problem and the objective/s of the study.

2. Data description - provide a description of the data set such as where it was obtained, what variables it contains, a brief description of these variables (e.g. height is a continuous variable within the range 150 cm - 170 cm) etc. You may also include figures or tables to describe the data.

3. Model Building

Method - process of building and evaluating the model and any hypothesis tests used in reducing the model (if required).

Results - results of the hypothesis tests performed (if any), the final model and any related figures (such as regression plots).

4. Application - example of an application related to the objective of the study.

5. Conclusion - whether the objective/s were achieved, what are the possible limitations of the study and suggestions of how the study can be improved or extended.

Order of presentations:

1. Ken Manohar
2. Calvin Liverpool
3. Celine Ganar
4. Kelsey Pierre
5. Karishma Harrykissoon
6. Antonio Rajkumar
7. Aadidev Sooknanan
8. Ismaeel Hosein
9. Danica De Freitas
10. Daniel Rudder
11. Gabrielle Gibbons
12. Christopher Gift
13. Odion Hillocks
14. Thais Montanez
15. Nathaniel Edwards
16. Ronnell Roberts-Reid
17. Anthony Vidal
18. Sabina Gooljar

DESCRIPTION OF REPORT

In addition to the presentation each student has to prepare a short report showing the computational element of the presentation. The report entitled ID.pdf **must** be submitted on myelearning by Sunday 18th April 8:00 PM (TT time) separate from the presentation.

The technical report should contain the following:

1. Overview of the project in no more than 200 words.

The data set used should be mentioned in this overview.

− for manually entered data state that the data was entered as shown in the R codes
− for built-in data in *R* provide the name of the data set
− for a separate data file provide a link to the data set.

If the data set used cannot be shared (for privacy reasons) then provide an image showing the first 2 − 3 rows/columns of data along with the variable names.

2. All **R codes** used and their related R output. These can be provided by either:

(i) Print-screen and crop images of the R console or
(ii) Copy R codes with output from the R console

If using the print-screen option ensure that the console is set wide enough to capture all of the codes that were run. Graphs can be copied from the R Graphics window.

Please note: if clarification is needed while marking, you will be asked to provide your R script.

3. A **conclusion** which summarizes what was achieved for example: We obtained a significant model by reducing the initial model of $\hat{y} = 1.02 - 0.03x_1 + 0.22x_2 - 4.32x_3 + 8.97x_4$ to $y = 0.19 + 0.27x_2 + 3.60x_4$. We also found the reduced model to be better than the full model using the partial F-test, comparing their adjusted R^2 values etc.
You may also include any other important findings obtained from the output.

4. A list of the **references** used. A minimum of one reference should be provided.

END