STAT40150 Multivariate Analysis Assignment

Dr. Garrett Greene

Trimester 2 2023/2024

• There is a total of 150 marks for this assignment and it is worth 40% of your final module mark.

• Answer all questions and carry out all analyses in R.

• Due date: 5PM Friday March 29th 2024. Please submit your assignment by uploading (i) a pdf file to BrightSpace containing your answers to the questions, and (ii) a *fully commented* R script which clearly indicates how you obtained your answers. It should be possible for the grader of your assignment to copy and paste this code into R, thereby reproducing your work. Students may submit a single pdf containing answers and code generated using R Markdown if they wish but this will gain no additional marks. Assignments submitted by email will not be graded.

• If submitting a single pdf generated using Rmarkdown, the pdf should be no longer than 20 pages. If submitting a pdf containing answers to the questions only, it should be no longer than 10 pages. Your submission should be as concise as possible, both in terms of commentary, output included and code.

• Full reasoning should be provided for any decisions made throughout the analyses.

• Late assignments will be graded according to UCD's Late Submission of Coursework Policy, as detailed on the module's Brightspace page.

• Your pdf should include solutions to the questions posed below; these solutions may include text, necessary output and plots from R. In your R script, all code should be fully commented to clearly illustrate how you arrived at your solution for each question. NB: where relevant, if R code is not provided, marks will not be awarded.

• Any plots included should be clearly labelled.

• *While discussion of the problems is encouraged plagiarism, in any form, is not permitted.* Students should familiarise themselves with the plagiarism policy detailed on the module's Brightspace page.

- This assignment is based on a dataset of human facial temperature measurements made using an Infrared Thermograph (IRT), a thermal imaging sensor which assesses temperature by measuring infrared radiation. IRTs are of interest in screening for infectious diseases, since they can provide a rapid temperature measurement without physical contact, which reduces risk of infection. Since measurements may be used to screen for illness, it is important that these tools are properly validated and calibrated against a "gold standard" measurement.
- The first 10 columns in the dataset contain the IRT temperature measurements. The 11th column contains the Oral Temperature measurement. The next four columns contain demographic covariates relating to the subjects. The final column indicates whether the patient's temperature is considered to be elevated (Oral Temp >37.8 Celsius). Background information relating to these data and how they were acquired can be found here. However only the subset of data included in the Temperature Data CSV file on Brightspace should be analysed.

1. Load the data set Temperature_Data.csv from Brightpsace into R. Use the `set.seed` function in R to set the seed to your student number. Using `R` random number functions, select a random subset of 1000 observations from the dataset. In this way you will achieve an (almost certainly) unique dataset for your own analysis. Ensure that you include the code used in this step in the R code you submit with your assignment so that your work can be reproduced. [0 marks]

2. The variable Oral_Temp contains the average of several oral temperature measurements, and is assumed to be the most accurate measure of true temperature. Remove from the dataset any record/observation which has a missing/NA value for Oral_Temp. Then, visualise the facial and oral temperature measurements using suitable plots. Comment on the plots. Very low values of oral temperature may indicate measurement error. Remove any observations with Oral_Temp more than 4 standard deviations below the mean. [10 marks]

3. Use hierarchical clustering and k-means clustering on the facial temperature meaurements to determine if there are clusters of similar profiles in the data. Motivate any decisions you make. Compare the hierarchical clustering and k-means clustering solutions. Comment on/explore any clustering structure you uncover, considering the data generating context. [25 marks]

4. The pattern of facial temperature measurements for an individual may be affected by factors such as the shape and structure of the face, the age of the subject, skin pigment, etc. As a result, it may be possible to classify a measurement as belonging to a particular type of subject based on the temperature pattern. Perform a linear or quadratic discriminant analysis to classify subjects by gender, using the facial temperature data. Assess how well your classifier performs using an appropriate method. Produce a plot showing the linear decision boundary for your LDA. Compare the performance of LDA and QDA in classifying the gender, and comment on your results. [30 marks]

5. Apply principal components analysis to the facial temperature data, motivating any decisions you make in the process. Plot the cumulative proportion of the variance explained by the principal components. How many principal components do you think are required to represent the data? Explain your answer. [10 marks]

6. Derive the principal component scores for each subject from first principles (i.e., you should not use an inbuilt function such as predict(...)). Plot the principal component scores for the subjects. Comment on any structure you observe. [15 marks]

7. We would like to be able to predict the oral temparature based on the facial temperature measurements. Since the predictors are highly correlated, regression models applied to them may be overfitted. Principal components regression (PCR) is one approach to "regularising" regression for such highly correlated data. Research the principal components regression method and how it works e.g., see An Introduction to Statistical Learning with Applications in R by James et al. (2021), The Elements of Statistical Learning by Hastie et al. (2017), and/or the peer-reviewed journal article The pls Package: Principal Component and Partial Least Squares Regression in R by Mevik and Wehrens (2007).
   In your own words, write a maximum 1 page synopsis of the PCR method. Your synopsis should (i) explain the method's purpose, (ii) provide a general description of how the method works, (iii) detail any choices that need to be made when using the method and (iv) outline the advantages and disadvantages

of the method. [30 marks]

8. Divide your data into training and test sets, and use the function `pcr` in the `pls` R package to perform PCR on the training data. Use your fitted model to predict the Oral Temperature in the test set. Motivate any decisions you make and evaluate the performance of your model. [30 marks]