Using aggregation functions for data analysis

Total Marks 100, Weighting 20%

The provided zip file contains the data file [RedWine.txt] and the R code [AggWaFit718.R] to use with the following tasks, include these in your R working directory. You can use the R script [template.R] to organise your code.

Clarification and related resources are provided on

https://d2l.deakin.edu.au/d2l/le/content/1193535/viewContent/6065224/View

Red wine quality Dataset

The given dataset, "RedWine.txt", is used to model wine quality based on physicochemical tests. The dataset provides the 1,599 red wine samples from the north of Portugal. It is a modified version of the data used in the study [1]. This dataset includes 5 variables, denoted as X1, X2, X3, X4, X5, and Y, described as follows:

X1 - citric acid

X2 - chlorides

X3 - total sulfur dioxide

X4 - pH

X5 – alcohol

Y - quality (score between 0 and 10)

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Assignment Tasks

1. Understand the data

(i) Import the txt file (RedWine.txt) and save it to your R working directory.

(ii) Assign the data to a matrix, e.g. using

```
the.data <- as.matrix(read.table("RedWine.txt "))
```

(iii) The variable of interest is quality (Y). To investigate Y, generate a subset of 440 data, e.g. using:

```
my.data <- the.data[sample(1:1599,440),c(1:6)]
```

[The following tasks are based on the 440 sample data]

(iv) Using scatter plots and histograms to understand the relationship between each of the variables X1, X2, X3, X4, X5 and the variable of interest Y.

2. Transform the data

Choose any four from the five variables (X1, X2, ..., X5). Make appropriate transformations to the chosen four variables and the variable of interest Y individually, so that the values can be aggregated in order to predict the variable of interest. Assign your transformed data along with your transformed variable of interest to an array.

[All the following tasks are based on the saved transformed data]

3. Build models and investigate the importance of each variable

(i) Import AggWaFit718.R file to your working directory and load into the R workspace using, source("AggWaFit718.R")

(ii) Evaluating the following fitting functions on the transformed data:

- A weighted arithmetic mean (WAM)
- Weighted power means (WPM) with P=2
- An ordered weighted averaging function (OWA)

4. Use your model for prediction

Using your best fitting model based on Q3, predict the wine quality for the input:

X1=1; X2= 0.075; X3=41; X4=3.53; X5=9.3.

[Apply the same pre-process as Q2 for the new input]

5. Summarising your data analysis procedures in up to 20 slides for a 5-minutes presentation. The slides should include the following contents:

- What kinds of the data distribution you have identified in the raw data.

- Explain the transformations applied for the selected four variables and the variable of interest.

- Include two tables - one with the error measures and correlation coefficients, and one summarising the weights/parameters and any other useful information learned for your data.

- Explain the importance of each of the variables (the four variables that you have selected).

- Which fitting function is the best fitting model on your selected data.

- Give your prediction result and comment on whether you think it is reasonable.

- Discuss the best conditions (in terms of your chosen four variables) under which a higher quality wine will occur.

- Comment the implications and the limitations of the fitting model you used for prediction.

The 5-minutes presentation can be using a simple and accessible platform such as YouTube or PowerPoint Audio.

Submission requirements

Submit to the SIT718 CloudDeakin Dropbox. Your final submission must include the following **TWO** files:

1. The slides with audio (a link to YouTube/Dropbox is acceptable).
2. The R code file (that you have written to produce your results) named "name-code.R" (where “name” is replaced with your surname or first name).

Your assignment will not be assessed if the code is missing, or the outputs of the code are inconsistent with the slides.

Following Harvard style for code citation and reference in your R script with comments:

https://www.deakin.edu.au/students/studying/study-support/referencing#tab__harvard-other-sources

You must cite all the datasets and packages you used for this assessment. You will lose some scores for inappropriate citations/references.