Coding Quiz 4

April 20, 2023

Instructions:

- Drop your .R file as per the links given below:

  BS: https://www.dropbox.com/request/AiDC7YaCebyMdwWvXJje.

  MSc (221252 - 221372): https://www.dropbox.com/request/OZbZC9osbgV7WSBLFSz4.

  MSc (221273 - 221466): https://www.dropbox.com/request/fLttNg7ghSxu7zv3NGQk.

  NO other files should be submitted through this link.

- Accepted format of the file: 111111.R (where '111111' is your roll number).

  Use your full name (e.g., Subhajit Dutta) and IITK email address (e.g., duttas@iitk.ac.in) while submitting your .R file.

- If you submit multiple files, only the first .R file will be considered.

  Moreover, **grading will be based solely on this .R file**.

- Output should be printed **strictly in the order of the questions** given in page 2.

- Total marks: 25

- Time: 40 minutes (6:30pm to 7:10pm).

  Needless to say, R codes dropped **after 7:10pm will NOT be graded**.

- Only text written below in Red should be printed when the R code is executed.

- Please **avoid spamming** by NOT uploading incorrect and/or multiple files.

Read the set of instructions given above again, before moving to page 2.

Question 1: In an experiment, two different methods were employed and the changes (in appropriate units) observed are as follows:

Method 1: $79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02$

Method 2: $80.02, 79.74, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97$

Test whether the two methods differ with respect to their performance at $\alpha = 0.10$ and 0.05, and interpret the results.
Comment on the likely relative spread of the two samples.

Question 2: Read the data directly from the link given below:

http://stat4ds.rwth-aachen.de/data/GSS2018.dat.

Do a cross classification of the 2016 vote for President (PRES16) by sex (1 = male, 2 = female).

(a) Form a contingency table, and calculate both the conditional distributions.

(b) Conduct a χ^2-squared test. Show the estimated expected frequencies, report the standardized residuals and form a mosaic plot.

Question 3: Read the data directly from the following link:

https://stat4ds.rwth-aachen.de/data/Iris.dat.

The last column (species) has two categories, namely, I.versicolor and I.virginica. Let n and m denote the number of occurrences of I.versicolor and I.virginica, respectively. Create a new response variable which takes the value $-(n+m)/n$ for I.versicolor, and the value $(n+m)/m$ for I.virginica.

Drop the first (record) and last (species) columns from this data set. Create a new data set by appending the new response variable with the remaining data columns.

Run a regression on this new data set, extract the coefficient vector and project the observations on a subspace using this estimated vector.

Now, use two different colors corresponding to the variable 'species' (recall the last column which you had dropped earlier) to create a plot for the projected observations. What do you infer from this visualization?

Output marks: (2+1) + (2+2) + (1+1+2+1) = 12

R code marks: 3 + 4 + 6 = 13