Practice Class Test 2

Data Analysis

Instructions

You are expected to complete a series of tasks described below. Please follow these instructions carefully before you start your work.

1. **Do NOT** open RStudio until you have downloaded the two files described in Instructions 2. and 3.
2. Go to the **Practice Class Test 2 Files** folder in the **Week 9: GLM** section of the **Data Analysis Moodle page**.
3. Download these files in the **Practice Class Test 2 Files** folder to the **same folder** on your **M: drive**:
    - `.csv` - which contains the data set;
    - `PracticeClassTest2Template.Rmd` - a R Markdown template for this practice class test which includes the tasks described below. It loads the R packages necessary to complete the task and also reads the data set into R (assuming you've saved the `.csv` files in the **same** folder as the `.Rmd` file).
4. Open RStudio and open `PracticeClassTest2Template.Rmd` then save it as `YourStudentNumberPracticeClassTest2.Rmd` in the **same folder** as the `.csv` files are saved on your **M: drive**.
5. **Before you start to work**, compile `YourStudentNumberPracticeClassTest2.Rmd` (using `Knit`) and check that the `YourStudentNumberPracticeClassTest2.pdf` file is produced as you expected. It is wise to periodically compile and check the `.pdf` file as you work through the tasks so you can fix any bugs in your code as you go.
6. Unlike in reports, you **are required** to **include** your R code in the `.pdf` file, hence `echo=TRUE` is set as the default in the `.Rmd` template.
7. Before answering each question, ensure that the data is in `tidy` format (and if it isn't modify it so that it is) and produce graphical summaries relevant to the question(s) asked of the data. Note that you **don't** need to label plots in this practice class test (unlike in reports) but you should comment briefly about how the graphical summaries informally inform your subsequent analysis.
8. When you are ready to submit your completed tasks, click on the **Practice Class Test 2 .pdf Upload** link under **Data Analysis > Week 9: GLM** and upload and submit the file `YourStudentNumberPracticeClassTest2.pdf`.
9. Also upload and submit the R Markdown file `YourStudentNumberPracticeClassTest2.Rmd` using the **Practice Class Test 2 .Rmd Upload** link. Please note that only the `.pdf` file will be marked. The `.Rmd` file will only be considered if there was a problem compiling the `.pdf` file.

PLEASE TURN OVER

Examination Conditions

- You have the full two hours to complete the class test and you can submit your completed tasks anytime within that time.
- You must work on your own - *NO communication* by any means with anyone is permissible.
- You are required to use `tidyverse` and `infer` functions for the analysis and `RMarkdown` to produce your answers
- You may consult ANY resources (hardcopy or online), e.g. `tidyverse` "cheat sheets" and/or the online tutorials from the course.

Question 1

Large samples of iron ore were mined from quarries "x" and "y", and each of the two samples were broken down into 10 smaller sub-samples for analysis. Each of these 20 sub-samples were sent to a chemical laboratory, and the percentage of iron in each sub-sample was measured. These data are stored in `practice1.csv`. Using bootstrapping, is there evidence in these data that the population mean iron percentage in each quarry is 35%, and are the population mean percentages different between the two quarries?

Question 2

(a) Using the data contained in `practice2.csv` build a regression model that adequately describes the `response` in terms of the potential explantory variables `X1`, `X2`, `X3` and `X4`. Your chosen model will therefore be the one that you believe best represents the `response`. Use only the theoretical confidence intervals generated under standard assumptions (which you should check) to identify the correct model. Note you **do not** need to consider interaction terms.

(b) Construct a table of **all** the possible linear models (without interactions or transformations) that could be fitted to the `response` variable in `practice2.csv`. In the table include the R^2_{adj} and AIC values for model comparisons. Do these measures lead you to the same conclusion about the model that best represents the data as in part (a)? Note: you are **not** required to check assumptions for each of these models in this task.