S520 Final Exam-ish

Due: May 10, 2020 by 11:59pm (But maybe sooner would be great. . .)

Oh My Heart. . . (Or spotifyr: Why Did You Fail Me.)

The following data is a modified form of a famous dataset from the UCI Machine Learning Data Repository. Researchers collected data related to various physical aspects of many patients. Their objective was to create a way of classifying whether their patients had heart disease or not without performing invasive procedures. To that end, they collected the following information from the patients.

The data file is available online as `HeartAbridged.csv`.

- `age`: The person's age in years
- `RestBP`: The person's resting blood pressure (mm Hg on admission to the hospital)
- `Chol`: The person's cholesterol measurement in mg/dl
- `MaxHR`: The person's maximum heart rate achieved during controlled exercise
- `AHD`: Whether or not a person has heart disease. + `Yes`: The do have a heart disease. + `No`: They do not have a heart disease.

Exam Guidelines

This exam will be a quasi-report. It doesn't have to be extensively detailed. The objectives listed in the *Required Analyses* are objectives for your write-up.

- Do not provide graphs, statistics, confidence intervals, or hypothesis tests without discussing them.
- Assume you are explaining everything like you are giving a presentation to the class and your grade depends on it. They know the material, but do not assume they are experts (maybe they are but that's beside point), so explain what you are doing!

Here is the general structure of your write-up.

Exploratory Data Analysis: Find the two variables with the the highest linear correlation. Describe their distributions via graphs and summary statistics. Create a scatterplot and describe the relation between the two variables.

Statistical Analyses: Perform the requested statistical analyses on the variables you chose. You are to provide a written statement that describes what analyses you are performing and the conclusions that arise from the analyses. If it is a hypothesis test, state the type of test, the conclusion you can make from the hypothesis test. If it is a confidence interval, write a sentence interpreting the confidence interval. To the best of your ability, comment on how the results make sense or do not make sense.

Conclusion: Summarize your results.

Required Analyes

1. Get confidence itervals for the mean value of each of the variables you chose. Make sure to provide interpretations.
2. Next you will do linear regression. Choose which variable will be your predictor variable and which will be your response variable. Justify your choice.

- Create a scatterplot of the data with the regression line on it. Describe how well the variables follow a linear pattern or not and report the correlation between the two.
- Create the linear regression model between the two variables and report the results. This included giving the resulting linear regression equation. Interpret the slope and intercept, and explain why they may or may not make sense.
- Perform a hypothesis test to assess whether the predictor variable is a "good" predictor in a linear model.
- Provide and interpret a confidence interval for the slope.
- Provide predictions of the value of the response variable at the 20th, 40th, 60th, and 80th, percentiles of the predictor variable.
- Check the regression assumptions: Normality of the residuals, the residuals have a consistent mean of zero, the residuals have constant variability.

Submission Guidelines

The main body of the report should not contain any actual R code, just the relevant output. R code is to be at the end of the report in an appendix.

Your submission should be a PDF.

Grading

This midterm is graded on a 100 point scale. You will be graded on the following criteria.

- Your code runs correctly.
- Requested calculations are performed correctly.
- All directions have been followed.
- All analyses are performed and conclusions are correct.
- The writing is clear and organized.