BISC 305 R Project 2 - Your Name (EDIT NAME HERE)

04/14/2020

Description

For the second R project you are going to do an ANOVA analysis and a regression analysis. Further, you get to choose your own dataset (this dataset cannot be one of the datasets that is pre-loaded into R). You can do both analyses on the same dataset, or you can select one dataset to do the ANOVA analysis and a different dataset to do the regression analysis. There are many places to find datasets, here are few links (your dataset does not have to be from one of these sites):

- http://archive.ics.uci.edu/ml/about.html (click view all datasets in upper right)
- http://faculty.marshall.usc.edu/gareth-james/ISL/ (click datasets and figures on left)
- http://web.stanford.edu/~hastie/ElemStatLearn/ (click data on left)

In Blackboard/Content/R stuff, there is a document, 'thirdRnotebook.pdf' that will be helpful for this assignment. Note: you cannot do these analyses on all datasets, so the ability to do these analyses will limit your selection of dataset(s). Also, sometimes importing datasets into R is a challenge. We recommend you try to import your dataset into R before doing anything else. We are more than happy to help you with this, but it is important that you don't wait until the last moment.

Part 1 (Due 04/16)

Since different students will be analyzing different datasets, you will have to describe your dataset (if you use two different datasets then you will have to describe them both). Make sure your description includes an explanation of the variables, whether the study was observational or an experiment, how the data was collected, the sample size, and any potential biases or confounding variables. Your description should be typed, about one page in length, and include the reference of where you got the data. Your description should be sufficient such that a person reading your report will not need to go to this reference to understand the dataset.

Important: How to chose a dataset?

1. For ANOVA: The dataset should have one categorical and one numeric continuous variable.

2. For Regression: The dataset should have at least two numeric continuous variables.

It is perfectly okay to pick two different datasets for ANOVA and Regression. You might also be able to find datasets that satisfy the criteria for both ANOVA and Regression.

1. Describe your dataset.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

INCLUDE A LINK TO THE DATASET: https://www.google.com

a. Explaination of the variables.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

b. Study Observational or Experimental?

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

c. How was the data collected?

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

d. Sample size?

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

e. Any potential biases or confounding variables?

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

f. Can you run ANOVA analysis on this dataset? Explain your choice of variables. (Note: it needs to include at least one categorical and one continuous variable).

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

g. Can you run regression analysis on this dataset? If yes, explain your choice of variables for running regression (Note: they both need to be numeric). If not, repeat Parts a-f for a different dataset which is suitable for a regression analysis.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

Part 2 (Due 04/28 - leave as it is while you are turning in Part 1)

1. Do a one-way ANOVA analysis.

a. State the null and alternative hypotheses.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

b. Make a side-by-side boxplot.

WRITE CODE BELOW (AND REMOVE THIS).

c. Show the ANOVA table.

WRITE CODE BELOW (AND REMOVE THIS).

d. Clearly state your conclusion.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

2. Do a regression analysis.

a. State the X and Y variables.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

b. Make a scatter plot and label the X and Y axes.

WRITE CODE BELOW (AND REMOVE THIS).

c. Add the least squares regression line to the plot.

WRITE CODE BELOW (AND REMOVE THIS).

d. What is the value of the correlation coefficient? Also state the p-value and the 95% confidence interval.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

e. What is the estimate of the slope and y-intercept? Also state the two p-values.

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

f. What fraction of the variance in Y is explained by the linear relationship between X and Y (coefficient of determination)?

YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

g. Plot the residuals as a function of the X variable. Do the residuals appear to depend on the X variable?

WRITE CODE BELOW (AND REMOVE THIS).
YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).

3. Propose your own question to answer with your dataset. Clearly state the question you are trying to answer, the statistical test you performed, the results of the test, and your conclusion. Include any relevant plots or tables.

WRITE CODE BELOW (AND REMOVE THIS).
YOUR TEXT ANSWER GOES HERE (AND REMOVE THIS).