BAN 600: Week Four Assignment

Due Saturday, September 28th, 2019, at 11:59 PM, Pacific Time

If you have questions about the following instructions or about your assignment, please send me an email with a description of your question and what you've tried in attempt to answer it. Be sure to include your data file and R script. Please do NOT submit late assignments; they will not be accepted after answers are posted.

Data Description

This data was downloaded from Kaggle.com, a site that houses open source datasets. This specific dataset is titled: "Student Alcohol Consumption: Social, gender and study data from secondary school students."

Source Information
P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Fabio Pagnotta, Hossain Mohammad Amran. Email:fabio.pagnotta@studenti.unicam.it, mohammadamra.hossain '@' studenti.unicam.it University Of Camerino
https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION

Visit Kaggle for a complete description:
https://www.kaggle.com/uciml/student-alcohol-consumption

Work in R

1. Download the Wk4_Assignment4.R file (R script) and save it to your local computer.
2. Download the student-mat.csv (data file) and save it on your local computer.
3. Copy the file path of where you saved your data.
4. Open the R script with Rstudio. Read the instructions on how to import the data.
5. Install and load the required packages (see R/Rstudio install instructions)
6. Using what you've learned from the E-Textbook, write (and run!) the code you would use to respond to the questions below. Submit your R script with your assignment.

Adding Explanatory Variables to Empty Models (100 points)

As the lead data scientist, you are asked to complete an analysis for a client to understand the leading causes of student **absences** (number of school absences, 0 to 93) in math classes. The client you are completing the report for hypothesizes that students who are in a romantic relationship (yes or no) will tend to have more absences. Use an explanatory model to test this research question, and write up the results of your analysis in an APA-style report to deliver to your client. Write as if your reader is unfamiliar with the dataset and sampling design.

1. **Introduction and research questions**
    a. Introduce your purpose and scope.
        i. Create a story of the DGP that might be responsible for the variation in the **absences** outcome variable (do not just talk about the researchers' predictions about relationship status).
    b. Describe the research question.
        i. Give a rationale for studying the relationship between romantic relationship and absences. Why do you think romantic relationship would explain variation in absences?
        ii. Include the word equation.
2. **Data Description**
    a. What are the frequencies of the categorical demographic variables?
    b. What are the five-number summaries of your quantitative demographic variables?
    c. Provide descriptive statistics of your outcome and explanatory variable.
    d. Create a histogram of the absences variable. Reference the figure in your report. What do you see? Describe the shape, center, skewness, and weirdness.
    e. Provide a visualization of the research question. Reference the figure in your report. What do you see?
3. **Research Question: Absences explained by Romantic**
    a. **Empty model:** Fit the empty model for your outcome.
        i. Now that you have estimates of your parameter(s), put them into a model statement (see Ch. 7.2).
        ii. Based on the model output and model statement, interpret the estimated parameter(s). What do the numbers mean?
    b. **Explanatory Model:** Add your explanatory variable to the model.
        i. How is X_i coded?
        ii. Now that you have estimates of your parameter(s), put them into a model statement (see Ch. 7.2).
        iii. Based on the model output and model statement, interpret the estimated parameter(s). What do the numbers mean?
    c. **Comparing the two models**
        i. Create an APA-style table of the supernova() table in your document. Discuss your findings in reference to the table, including a discussion of the F ratio, degrees of freedom, and the difference in the number of parameters between the models.
        ii. What is the proportion reduction in error (PRE), and what does it mean? What can you say about the strength of the model from PRE?
        iii. What is Cohen's d, and what does it mean? What can you say about the strength of the model from Cohen's d?
    d. **Conclusion**
        i. What you have learned from your models in relation to the research question? Was the research question supported?
        ii. Do you believe there are any limitations to your analysis? For example, sample size, bias, missing data, mistakes, measurement error, sampling error, unrepresentative sample, etc.?