Problem Set #03 Statistics 506

Due: Oct 10, 10am on Canvas

Instructions

- **Review the attribution of sources discussions.**
- Submit the output of your Quarto document on Canvas. The output should include a link to your GitHub repository for this assignment.
- Unless otherwise explicitly stated, all problems should be solved in **R**.
- Your output file should include all code required to solve the problem. Code folding may be useful to make the output more readable.
- Use a consistent and readable code style for your code. Lack of a consistent and readable code style will negatively affect your grade.
- Some of these exercises may require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is for you to learn to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.
- Be sure to properly document any functions you write using roxygen, and add comments as appropriate to make it clear what you are doing.
- If submitting an HTML file, please make sure to make it self-contained.

Problem 1 - Vision

This problem will require you to do things in Stata we have not covered. Use the Stata help, or online resources, to figure out the appropriate command(s). Use citation as necessary.

a. Download the file VIX_D from this location, and determine how to read it into Stata. Then download the file DEMO_D from this location. Note that each page contains a link to a documentation file for that data set. Merge the two files to create a single Stata dataset, using the **SEQN** variable for merging. Keep only records which matched. Print our your total sample size, showing that it is now 6,980.

b. Without fitting any models, estimate the proportion of respondents within each 10-year age bracket (e.g. 0-9, 10-19, 20-29, etc) who wear glasses/contact lenses for distance vision. Produce a nice table with the results.

   (Hint: One approach might be to try and find a way to produce this table with a single command. Another might be to estimate each proportion separately and then combine the results somehow. Yet another approach might be to manually do the calculations in Mata. Or any other approach that produces a single nice table.)

c. Fit three logistic regression models predicting whether a respondent wears glasses/contact lenses for distance vision. Predictors:

   1. age
   2. age, race, gender
   3. age, race, gender, Poverty Income ratio

   Produce a table presenting the estimated odds ratios for the coefficients in each model, along with the sample size for the model, the pseudo-R^2, and AIC values.

d. From the third model from the previous part, discuss whether the *odds* of men and women being wears of glasess/contact lenses for distance vision differs. Test whether the *proportion* of wearers of glasses/contact lenses for distance vision differs between men and women. Include the results of the test and its interpretation.

Problem 2 - Sakila

Load the "sakila" database discussed in class into SQLite. It can be downloaded from https://github.com/bradleygrant/sakila-sqlite3.

a. Aside from English, what language is most common for films? Answer this with a single SQL query.

For each of the following questions, solve them in two ways: First, use SQL query or queries to extract the appropriate table(s), then use regular R to answer the question. Second, use a single SQL query to answer the question.

b. What genre of movie is the most common in the data, and how many movies are of this genre?

c. Identify which country or countries have exactly 9 customers.

Problem 3 - US Records

Download the “US - 500 Records” data from https://www.briandunning.com/sample-data/ and import it into R. This is entirely fake data - use it to answer the following questions.

a. What proportion of email addresses are hosted at a domain with TLD “.net”? (E.g. in the email, “angrycat@freemail.org”, “freemail.org” is the domain, with TLD (top-level domain) “.org”.)

b. What proportion of email addresses have at least one non alphanumeric character in them? (Excluding the required “`@`” and “`.`” found in every email address.)

c. What is the most common area code amongst all phone numbers?

d. Produce a histogram of the log of the apartment numbers for all addresses. (You may assume any number after the street is an apartment number.)

e. Benford’s law is an observation about the distribution of the leading digit of real numerical data. Examine whether the apartment numbers appear to follow Benford’s law. Do you think the apartment numbers would pass as real data?

f. Repeat your analysis of Benford’s law on the *last* digit of the street number. (E.g. if your address is “123 Main St #25”, your street number is “123”.)