HW 4 GV207 2022
Instructor: Ryan Bakker
Due Week 12

Write an R code file named (`gv207-HW4.R`) to complete the following tasks.

Rules–2pts each

- Submit two files, and two files only. That is, submit (1) the coversheet and (2) your R code file (`gv207-HW4.R`). Don't submit your graph or other outputs.

- Make sure that you delete your name from your R code file.

- Execute everything before you submit (e.g., CTRL + A & CTRL + Return on a Windows PC; Command + A & Command + Enter on a Mac machine), and make sure your file runs without an error.

- Your file must have a proper header.

- Add comments and annotations to everything you do. Try to make your code file look like my code file. Don't copy and paste all the questions into your R code file, but do show me the question number for each question.

Tasks

1. You are going to use the following 6 R packages: `ggplot2`, `gmodels`, `Hmisc`, `stargazer`, `effects`, and `gridExtra`. Load all the packages. [2 points]

2. Load Titanic passenger survival dataset available on Moodle (`titanic.csv`), and store it as an object named `td`. [2 points]

3. The unit of observation is individual passengers. How many passengers does the dataset have? That is, how many rows are there in the dataset? Provide a command to get the answer. Also, write your answer in a comment line. [2 points]

4. The dataset contains various information on passengers, including their name (`name`) and whether or not they survived (`survived`). The dummy variable `survived` is coded as 1 if a passenger survived and 0 otherwise. Create a simple frequency table of the `survived` variable (that is, there is no need to change the column names or to obtain relative percentages for this task; one line of command is sufficient) to see how many passengers survived and how many did not make it. Provide a command to create a frequency table of this variable (no need to write a comment). [2 points]

5. Calculate the survival percentage. Provide command(s) to calculate it. Also, write your answer in a comment line. [2 points]

6. The data set contains a variable named `pclass`, which is coded as 1 (= passenger has a 1st class ticket; for those of you who don't know, 1st class tickets are more expensive than 2nd class tickets, 2nd class more expensive than 3rd class), 2 (= 2nd class ticket), and 3 (= 3rd class ticket). This ordinal (ordered categorical) variable could be used as a proxy for socio-economic class of passengers. Create a simple frequency table of this variable (that is, there is no need to change the column names; one line of command is sufficient) to see the distribution of this variable. Provide a command to create a frequency table of this variable (no need to write a comment). [2 points]

7. In tasks 7–11, we will analyze the relationship between socio-economic class of passengers and their survival using one of the three bivariate hypothesis testing methods you have learned in Weeks 6 & 7. The question we ask here is: how does socio-economic class of passengers influence the likelihood of passenger survival? Let's say we hypothesize that socio-economic class of passengers is positively associated with passenger survival. First, what are the dependent and independent variables in our investigation? Provide your answers in a comment line. [2 points]

8. Second, create a two-way frequency table (a.k.a, cross tabulation) of the two variables, `pclass` and `survived`. Provide a command to create such a table (no need to write a comment). Make sure that (1) values of the dependent variable are shown in rows and the independent variable in columns, (2) your table shows column percentages but not row percentages, cell percentages, or χ^2 contributions, and (3) your table produces a χ^2 test statistic. [6 points]

9. Read the table you produced the above and answer a few questions. (a) What is the survival percentage among the 1st class passengers? (b) What is the survival percentage among the 2nd class passengers? (c) What is the survival percentage among the 3rd class passengers? Provide your answers in comment lines. [3 points]

10. Would you say that the relationship between `survived` and `pclass` is consistent with our hypothesis I described in task 7? Why or why not? There is no need to comment on statistical significance, but do comment on the pattern observed in the sample. Provide your answers in a comment line. [3 points]

11. Fill in the blanks of the following statements that summarize the results. Provide your answers in a comment line. You only need to write four options, such as (a), (b), (c), etc., in the correct order. [8 points]

    Since the test statistic produces a p-value smaller than ____, we can ____ the null hypothesis of no association at ____% confidence level. We thus ____ support for our hypothesis.

    (a) 173   (b) 172   (c) 128   (d) 127   (e) 99.9   (f) 99
    (g) 95   (h) 90   (i) 0.1   (j) 0.05   (k) 0.01   (l) 0.001
    (m) accept   (n) reject   (o) find   (p) do not find

12. The dataset contains a variable named `fare`, which is the price of the ticket each passenger has. It is shown in pre-1970 GBP. (Note: £ 1 in 1911 is equivalent in purchasing power to about £ 112 in 2018.) Do female passengers tend to have a more expensive ticket compared with male passengers? Explore the relationship between the `female` variable (coded as "Female" for female passengers and "Male" for male passengers) and `fare`. Choose an appropriate bivariate statistical testing method for these two variables from the three methods you have learned in Weeks 6 & 7, and perform the test. Provide command(s) to perform the analysis. (Hint: I am not asking you to run a regression.) [5 points]

13. Interpret the results of the bivariate test you performed above and answer the question (do female passengers tend to have a more expensive ticket?). Comment on the observed pattern in the sample as well as the statistical significance, and draw a conclusion (i.e., answer the question posed here). Your answers must have up to three sentences. [6 points]

14. The dataset contains a variable named `age` (age of the passenger). Do older passengers tend to have a more expensive ticket compared with younger passengers? Explore the relationship between `age` and `fare` graphically. That is, create a plot that shows the relationship between these two variables using the `ggplot` function. Provide commands to create the plot. [5 points]

15. Perform an appropriate bivariate statistical test (again, choose one from the three methods covered in Weeks 6 & 7) to explore the relationship between `age` and `fare`. Provide command(s) to perform the analysis. (Hint: I am not asking you to run a regression.) [5 points]

16. Interpret the results of the bivariate test you performed above and answer the question (do older passengers tend to have a more expensive ticket?). Comment on the observed pattern in the sample as well as the statistical significance, and draw a conclusion (i.e., answer the question posed here). Your written answers must have up to three sentences. [6 points]

17. Regress `fare` on `age`, and produce a regression table using the `stargazer` function. [4 points]

18. Create a plot that illustrates the estimated effect of `age` on `fare` based on the model you estimated above. [3 points]

19. Judging from the numerical and graphical results of the regression analysis, would you say that `age` and `fare` are positively related? Comment on the observed pattern in the sample as well as the statistical significance, and draw a conclusion. [6 points]

20. Graphically explore the relationship between `age` and `fare`, holding constant the `female` variable. That is, create a plot using the `ggplot` function that shows the relationship between `age` and `fare` for female and male passengers separately. Try to have one plot that has two panels (one for female and one for male). [5 points]

21. Regress `fare` on `age` and `female`, and produce a regression table using the `stargazer` function that summarizes the results of this model and the model you estimated in task 17. [4 points]

22. Which one of the two regression models performs better? Fill in the blanks of the following statement. Provide your answers in a comment line. You only need to write four options, such as (a), (b), (c), etc., in the correct order. [8 points]

    Since _____ is _____ for the _____ model, the _____ model fits the data better.

    (a) standard error   (b) p value   (c) R^2 (d) Adjusted R^2 (e) the number of stars
    (f) smaller   (g) greater   (h) first   (i) second

23. Create a plot that illustrates the estimated effect of `female` on `fare` based on the second regression model you estimated. [3 points]

24. Create a plot that illustrates the estimated effect of `age` on `fare` for male and female passengers separately based on the second regression model you estimated. Try to have one plot that has two panels (one for female and one for male). [6 points]

25. It may be the case that there the effect of age is conditional on sex. To test this, estimate a model that includes an interaction term between age and female. That is, include age, female, and the interaction as indepedent variables. Produce a stargazer table that presents the results. [4pts]

26. Produce a graph that displays the relationship between age and fare for men and for women. Include a brief description of what this graph tells us about the effect of age on fare for the 2 groups. (Is the effect of age different between men and women? If so, how is it different?). [4pts]

27. . Write out the implied regression equation for men and for women (2 equations). Be sure to comment these out. [4pts]

End of file