STAT303-2 Assignment 2 Template v1

January 18, 2022

1 STAT303-2: Assignment 2

1.1 Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

d. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is well-organized, clear, and concise (3 points)

1. It's easy enough to identify different sections of the homework assignment (e.g., if there are different sections of an assignment, they're clearly distinguishable by section headers or the like)

2. It's clear which code/markdown blocks correspond to which questions.
3. There aren't excessively long outputs of extraneous information (e.g., no printouts of entire data frames without good reason)

This assignment is **due at 11:59pm on Wednesday, January 26th**. Good luck!

Submissions will be graded with a maximum of **43 points** – 38 points for code & answers, 5 points for anonymity and proper formatting.

1.2 Part 1

The dataset *infmort.csv* gives the infant mortality rates of different countries in the world.

This part is worth 17 points overall.

(1a) Make one plot including the following subplots: (i) a boxplot of mortality against region, (ii) a boxplot of income against region, and (iii) a scatter plot of mortality against income. Be sure to include appropriate axis labels.

What trends do you see in these plots? Mention the trend separately for each subplot.

(3 points for code and plots, 2 points for written answers)

(1b) Europe seems to have the lowest infant mortality, but it also has the highest per capita annual income. We want to see if Europe still has the lowest mortality if we remove the effect of income from the mortality. We will answer this question with the following steps.

(1b-i) Within one visualization, use subplots to plot: (1) mortality against income, (2) log(mortality) against income, (3) mortality against log(income), and (4) log(mortality) against log(income). Be sure to include appropriate axis labels.

(2 points for code)

(1b-ii) Based on the plots from (1b-i), postulate (describe) an appropriate model to predict mortality as a function of income. Fit the corresponding model and print the model summary.

(1 point for answer, 2 points for code)

(1b-iii) Update the model developed in the previous question (1b-ii) by adding *region* as a predictor. Print the model summary.

(2 points for code)

(1b-iv) Use the model developed in the previous question (1b-iii) to compute *adjusted_mortality* for each observation in the data, where adjusted mortality is the mortality after removing the estimated effect of income. Make a boxplot of adjusted mortality against region. Be sure to include appropriate axis labels.

(3 points for code)

(1b-v) Does Europe still has the lowest mortality after removing the effect of income from mortality? After adjusting for income, do more African / Asian / American countries seem to do better than European countries with regard to mortality?

(2 points for answers)

1.3 Part 2

The dataset *soc_ind.csv* contains the GDP per capita of some countries along with several social indicators.

This part is worth 21 points overall.

(2a) For a simple linear regression model predicting *gdpPerCapita*, which predictor will provide the best model fit *(ignore categorical predictors)*? Let that predictor be P.

(1 point for code, 1 point for answer)

(2b) Make a scatterplot of *gdpPerCapita* vs P. Does the relationship between *gdpPerCapita* and P seem linear or non-linear?

(1 point for code, 1 point for answer)

(2c) If the relationship identified in (2b) is non-linear, investigate transformation(s) of the predictor P in the model that might improve the model fit. To do so, use scatterplots displaying values of *gdpPerCapita* against corresponding values of P under different transformation(s).

When you've settled on an optimal model, report the predictors included in that model. Fit your new model and report the change in the R-squared value of this transformed model as compared to the simple linear regression model with only P.

(4 points for code, 2 points for answers)

(2d) Plot the regression curve of the transformed model (developed in the previous question (2c)) over the scatterplot in (2b) to visualize model fit. Also include the regression line of the simple linear regression model with only P on the same plot. Be sure to include a legend to distinguish the two models.

(3 points for code)

(2e) Develop a model to predict *gdpPerCapita* with P and *continent* as predictors. For a given value of P, which continents **do not** have a signficant difference between their mean *gdpPerCapita* and that of Africa? Consider a significance level of 5%.

(1 point for code, 1 point for answer)

(2f) The model developed in (e) has a limitation. It assumes that the increase in mean *gdpPerCapita* with a unit increase in P does not depend on the *continent*. Eliminate this limitation by including interaction of *continent* with P in the model developed in (e). Print the model summary of the model with interactions.

(2 points for code)

(2g) Use the model developed in (2f) to plot regression lines for Africa, Asia, Europe, and Oceania. Put *gdpPerCapita* on the vertical axis and P on the horizontal axis. Use a legend to distinguish among the regression lines of the continents.

(3 points for code)

(2h) Based on the plot develop in the previous question, which continent has the highest increase in mean *gdpPerCapita* for a unit increase in P, and which one has the least?

(1 point for answers)