CptS 475/575: Data Science, Fall 2024

Assignment 5 - Part 1: Linear Regression & Logistic Regression

Release Date: Tuesday, Oct. 15, 2024 **Due Date:** Wednesday, Oct. 23, 2024 (11:59 pm)

General instruction: This is the first part of Assignment 5. This assignment assesses your understanding of linear and logistic regression.

Your solution will be submitted as a PDF/HTML file, which must **include your full, functional code and relevant results as stated in each part**. You are encouraged to use R Markdown or Quarto to prepare your file if you work with R. You are free to use Python to solve the problems; you can use Jupyter notebook to prepare your file in that case.

1) (18 points) This question involves the use of multiple linear regression on the redwine (*winequality-red.csv*) data set available on Canvas in the Datasets for Assignments module. This is the same dataset used in Assignment 2.
   a. (6 points) Perform a multiple linear regression with pH as the response and all other variables except citric_acid as the predictors. Show a printout of the result (including coefficient, error, and t-values for each predictor). Comment on the output by answering the following questions:
      i) Which predictors appear to have a statistically significant relationship to the response? How do you determine this?
      ii) What does the coefficient for the free_sulfur_dioxide variable suggest, in simple terms?
   b. (6 points) Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?
   c. (6 points) Fit at least 3 linear regression models (exploring interaction effects) with alcohol as the response and some combination of other variables as predictors. Do any interactions appear to be statistically significant?

2) (30 points) This problem involves the Boston data set, which can be loaded from library MASS in R and is also made available in the Datasets for Assignments module on Canvas (*boston.csv*). We will now try to predict per capita crime rate (crim) using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.
   a. (6 points) For each predictor, fit a simple linear regression model to predict the response. Include the code, but not the output for all the models in your solution.
   b. (6 points) In which of the models is there a statistically significant association between the predictor and the response? Considering the meaning of each variable, discuss the relationship between crim and each of the predictors nox, chas, rm, dis and medv. How do these relationships differ?
   c. (6 points) Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?
   d. (6 points) How do your results from (a) compare to your results from (c)? You can present this comparison as a plot or as a table or any other form of comparison you deem fit.

e. (6 points) Is there evidence of non-linear association between the predictors age and tax and the response crim? To answer this question, for each predictor (age and tax), fit a model of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

Hint: use the poly() function in R. Use the model to assess the extent of non-linear association.

3) (12 points) Suppose we collect data for a group of students in a statistics class with variables:

X_1 = hours studied,

X_2 = undergrad GPA,

X_3 = PSQI score (a sleep quality index), and

Y = receive an A.

We fit a logistic regression and produce estimated coefficient, $\beta_0 = -8$, $\beta_1 = 0.1$, $\beta_2 = 1$, $\beta_3 = -.04$.

a. (4 points) Estimate the probability that a student who studies for 32 h, has a PSQI score of 11 and has an undergrad GPA of 3.0 gets an A in the class. Show your work.
b. (4 points) How many hours would the student in part (a) need to study to have a 65 % chance of getting an A in the class? Show your work.
c. (4 points) How many hours would a student with a 3.0 GPA and a PSQI score of 3 need to study to have a 60 % chance of getting an A in the class? Show your work.