CptS 475/575: Data Science, Fall 2019

Assignment 4: Linear Regression

Release Date: Wed. Oct. 2, 2019 **Due Date:** Thu. Oct. 10, 2019 (6 pm)

General instruction: This assignment has **three problems.** The first two problems are for both CptS 475 and CptS 575 students. The third problem is only for CptS 575 students. For CptS 475 students, Problem 1 and Problem 2 each carry 50% of the total weight. For CptS 575 students, Problem 1 and Problem 2 each carry 45%, and Problem 3 carries 10% of the total weight.

Your solution will be submitted as a PDF file. You are encouraged to use R Markdown or a similar tool (like Jupyter) to prepare your file. If you are using R, most of the required functionalities are builtin, though you may wish to use ggplot2 for more attractive graphics. If you are using Python, you will want to look into libraries such as scikit-learn, Pandas and pyplot for the functionalities described.

1. This question involves the use of multiple linear regression on the Auto data set from the course webpage (https://scads.eecs.wsu.edu/index.php/datasets/). Ensure that you remove missing values from the dataframe, and that values are represented in the appropriate types.
    a. Produce a scatterplot matrix which includes all the variables in the data set.
    b. Compute the matrix of correlations between the variables. You will need to exclude the name variable, which is qualitative.
    c. Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Show a printout of the result (including coefficient, error and t values for each predictor). Comment on the output:
        i. Which predictors appear to have a statistically significant relationship to the response, and how do you determine this?
        ii. What does the coefficient for the displacement variable suggest, in simple terms?
    d. 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?
    e. Fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?
    f. Try transformations of the variables with X^3 and log(X). Comment on your findings.
2. This problem involves the Boston data set, which we saw in the lab. We will now try to predict per capita crime rate 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. For each predictor, fit a simple linear regression model to predict the response. Include the code, but not the output for all models in your solution. 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 nox, chas, medv and dis in particular. How do these relationships differ?

b. 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 $H_0 : \beta_j = 0$?

c. How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. What does this plot tell you about the various predictors?

d. Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, 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. Again, include the code, but not the output for each model in your solution, and instead describe any non-linear trends you uncover.

3. An important assumption of the linear regression model is that the error terms are uncorrelated (independent). But error terms can sometimes be correlated, especially in time-series data.

a. What are the issues that could arise in using linear regression (via least squares estimates) when error terms are correlated? Comment in particular with respect to
   i) regression coefficients
   ii) the standard error of regression coefficients
   iii) confidence intervals

b. What methods can be applied to deal with correlated errors? Mention at least one method.