CptS 483-04: Introduction to Data Science, Fall 2017

Assignment 3: Linear Regression and Logistic Regression

Release Date: September 27, 2017 **Due Date:** October 4, 2017 (11:59 pm)

This assignment has **four problems**.

1. This question involves the use of multiple linear regression on the Auto data set from the course webpage. Ensure that you remove missing values from the dataframe, and that values are represented in the appropriate types (num or int for quantitative variables, factor, logi or str for qualitative).
    a. Produce a scatterplot matrix which includes all of the variables in the data set.
    b. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.
    c. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. 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 year variable suggest?
    d. Use the plot() function to 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. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?
    f. Try transformations of the variables with X^2 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. In which of the models is there a statistically significant association between the predictor and the response? Discuss the relationship between crim and zn, nox and ptratio in particular. How do these relationships differ?
    b. Fit a multiple regression model to predict the response using all of 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.

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.

3. This question should be answered using the Weekly data set, which is part of the ISLR package. With the ISLR package loaded, run the command ?Weekly to get a summary of the variables in the dataframe.
   a. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns or correlations?
   b. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
   c. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
   d. Now fit the logistic regression model using a training data period from 1990 to 2007, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2008 to 2010).
4. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set. Ensure that you remove missing values from the dataframe, and that values are represented in the appropriate types (num or int for quantitative variables, factor, logi or str for qualitative).
   a. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to include mpg01 as a new column in the Auto data frame.
   b. Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. List the features you think will be useful in predicting mpg01 with a short justification of why.
   c. Split the data into a training set and a test set by adding a new column to the Auto dataframe. The train column should be true for all rows up to and including 80, and false for all other rows. Note that we select training data this way only to standardize your results; generally this is a poor way to select training data, as the sampling should be random. Perform logistic regression on the training data in order to predict mpg01 using the four variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?