DSC 423

Assignment 2

Based on Modules 3 and 4

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work." You submission must be submitted as a PDF.

1) Short Essay (10 pts.) For each of these questions, your audience are persons that are not experts in statistics. Write with complete sentences and paragraphs. Cite any references that you use.
   a. (5 pts.) When building a model, you make four assumptions about the residuals. Explain what they are and how you can verify that your assumptions are correct.
   b. (5 pts) Define 'interaction term'. From your own experience, identify an instance in which you believe an interaction term would be appropriate.
2) Banking (20 pts.)
   a. Use the Banking dataset for this question, found under content on the D2L. This dataset consists of data acquired from banking and census records for different zip codes in the bank's current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The fields in the dataset:
      i. Median age of the population (Age)
      ii. Median years of education (Education)
      iii. Median income (Income) in $
      iv. Median home value (HomeVal) in $
      v. Median household wealth (Wealth) in $
      vi. Average bank balance (Balance) in $
   b. Load the data into R.
   c. In R, you can create a scatterplot by using the plot command, i.e. plot(x, y). Create scatterplots to visualize the associations between bank balance and the other five variables. Paste them into your submission. Describe the relationships.
   d. In R, you can compute correlations between two variables by using the cor command, i.e. cor(x,y) where x and y are the names of your variables, or you can compute pair-wise correlations by using cor(D), where D is the name of your dataframe. Compute correlations found in the bank data. Interpret the correlation values. Paste them into your submission. Describe which variables appear to be strongly associated?
   e. Fit a regression model of balance vs the other five variables. Present the estimated regression model and evaluate it. Recall that you can build a linear regression model by using the lm command and display the model by using the summary command.
   f. Which of the five predictors have a significant effect on balance? (a=.05) Explain.
   g. A good model should only contain significant independent variables, so remove the variable with the largest p-value (>0.05) and refit the regression model of balance versus the remaining four predictors. Present the new regression model.

h. Analyze if all four predictors have a significant association with balance? (a=.05) If not continue to remove one insignificant variable at a time until all of the remaining predictors are significant.

i. Interpret each of the regression coefficients for the final model.

j. Discuss the adj-R^2 for the final model.

k. Are there any influential points in your data set? Explain what impact an influence point might have.

3) WATEROIL (20 pts.) In the oil industry, water that mixes with crude oil during production and transportation must be removed. Chemists have found that the oil can be extracted from the water/oil mix electrically. Researchers at the University of Bergen (Norway) conducted a series of experiments to study the factors that influence the voltage (y) required to separate the water from the oil (Journal of colloid and interface science, Aug. 1995). The seven independent variables investigated in the study are listed in the table. (Each variable was measured at two levels - a "low" level and a "high" level.) Sixteen water/oil mixtures were prepared using different combinations of independent variables; then each emulsion was exposed to a high electric field. In addition, three mixtures were tested when all independent variables were set to 0. The variables are given in the table below.

Experiment number
y: voltage (kw/cm)
x1: disperse phase volume (%)
x2: salinity (%)
x3: temperature (^{0}C)
x4: time delay (hours)
x5: surfactant concentration (%)
x6: span:triton
x7: solid particles (%)

a. Use R to perform a regression analysis on the WATEROIL dataset (found on the D2L). Consider interaction terms and second-order terms. Evaluate the t-tests, F-Test and adj-R^2 accordingly.

b. (5 pts.) Paste your final model into your submission.

c. (15 pts.) Describe your model. Assume your audience is a fellow DSC423 student. Your description should begin by reporting basic facts about your model; but should also include an analysis of the findings.