DSC 423: Data Analysis and Regression
Assignment 03: Multiple Regression

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." Your submission must be submitted as a PDF.

1. Short Essay (20 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. (10 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. (10 pts) Define 'interaction term'. From your own experience, identify an instance in which you believe an interaction term would be appropriate.
2. BANKING (30 pts.) 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:
    - Median age of the population (Age)
    - Median years of education (Education)
    - Median income (Income) in $
    - Median home value (HomeVal) in $
    - Median household wealth (Wealth) in $
    - Average bank balance (Balance) in $

    a. (5 pts.) 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 (5 in total) into your submission. Describe the relationships.
    b. (5 pts.) 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 for the bank data. Paste them into your submission. Describe which variables appear to be strongly associated? Interpret any correlation values you deem important.
    c. (5 pts.) Fit a single 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.
    d. (5 pts.) Which of the five predictors have a significant (a=.05) effect on balance? Explain.
    e. (5 pts.) 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. 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 the remaining predictors are significant. Present the final regression model.

f. (5 pts.) Interpret each of the regression coefficients for the final model. Discuss the adj-R^2 for the final model. Is this a good model? Explain.