HW1 - Predicting Credit Card Charges (due date: 2021 04 26)

IMPORTANT: For grading to be done correctly make sure you follow the instructions precisely. Write your answers in a Word document and submit your writeup with your Excel solution file.

A credit card corp. recorded its customer data when they apply for credit cards. This data consists of the customer's annual household income, number of years of post-high school education, and number of members of the customer's household information. The company has records of the credit card charges accrued by each customer over the past year. The data is sampled for analysis with 5,000 customers and the data is put into **dats501_hw1_DataFile_CreditCustomers.xlsx**

The analysis requires applying multiple regression to data to predict annual credit card charges (y) given a customer's annual household income (x1), number of members of the household (x2), and number of years of post-high school education (x3).

1. Randomly sample 50 observations from this data. Develop an estimated multiple linear regression model that can be used to predict credit card charges. Then, from the remaining data select 9 other sets of 50 observations. Develop 9 more regression equations.

2. Create a table with 'Regression Parameter Estimates and the corresponding p values' for 10 Multiple Regression Models developed in part 1. Discuss your findings.

3. Randomly sample 3,000 observations from the entire data set, and fit multiple linear regression using same independent variables. Discuss your findings.

BIG QUESTION (BQ1). How does regression with large sample differ from regression with small sample? Research the concept of statistical inference and comment on the use of inference with estimates generated from very large samples.

4. To increase the variation in the dependent variable explained by the model, the data analyst decides to augment the original regression with a new independent variable, number of hours per week spent watching television (designate as x4). Is this a better model (compared to the one in part 3)? Discuss your findings.

5. Predict credit charges for the remaining 2000 observations not used in modeling (in parts 3 and 4), using the model in parts 3 and 4, separately. Calculate SSE for both. Which model produces a higher predictive performance?

BIG QUESTION (BQ2). Learn the difference between explanatory vs. predictive modeling. Briefly summarize the objectives in both and discuss how models in parts 3 and 4 perform in the two objectives.

Note: Some of the material in this assignment may require thinking and doing some research.