Programming Practices 2

Instructions: Problem 1 is based on the contents covered on 18th and 19th July, whereas problem 2 is based on what will be covered in the class on 25th July.

1. Use the Advertising data uploaded in the github repository to do the following
    a. Take a random sample of size 150 and use this as a training data. Rest of the data can be used for testing.
    b. From the training data select random samples of size 100 and fit a linear regression model to predict the sales with respect to all the input features.
    c. Derive the ANOVA table of the model parameters and verify whether the sales figure truly depends on at least any one of the predictor variables taken.
    d. Apply the model on the test data set and derive the Mean Square Error and the R^2 score.
    e. Repeat (b) 30 times by selecting random samples of size 100 from the training data and derive the estimates different model parameters, $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, obtained for each sample set. Find the variance of each of these parameters. Think what inferences can be drawn about the models from these observations.

2. Use the Credit.csv data uploaded in the github repository to do the following
    a. Take random samples of size 300 and use this as a training data. Rest of the data can be used for testing.
    b. Run a forward selection method to select the quantitative features (Limit, Rating, Cards, Age, Education and Balance), that can be used to predict the *Income* of a person. Derive the adjusted R^2, C_p, and AIC scores of the different models to identify the best model.
    c. Now repeat (b) by including the qualitative predictors (like Gender, Student, Married, Ethnicity and Balance) also and report on the best model that predicts the income of a person.
    d. Check if Limit and Balance is correlated (Correlation Coefficient >0.6). In that case include the interaction of these parameters in your model and print the ANOVA table and the R^2 obtained using the test data.