DSBA/MBAD 6201 Assignment

Data Exploration and Multiple Linear Regression

(Due on Feb. 13, 2024)

The Data Set and Attribute Description

The data is technical spec of cars. The dataset is downloaded from UCI Machine Learning Repository. This dataset is a modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg".

The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of one multivalued discrete and 5 continuous attributes.

Data Preprocessing

1. This problem consists of three parts:

   a) Generate box-plot for the horsepower and acceleration attributes and identify the cutoff values for outliers. (2 pts)

   b) Generate a scatterplot for acceleration against horsepower. (2 pts)

   c) Comment on how inclusion of the outliers would affect a predictive model of 'mpg' as a function of 'acceleration'. (2 pts)

2. 'mpg' has a somewhat longish tail and is not precisely normally distributed, so we will take a log transformation, (use df['lmpg'] = df['mpg'].apply(np.log)), and then predict 'lmpg' instead. (You should convince yourself that this is a better idea by looking at the histograms to assess normality; however, there is no need to submit such plots.) (2 pts)

Regression Analysis and Assessment

3. Try to fit an MLR to this dataset, with 'lmpg' as the dependent variable. Use all the available variables in your model. (4 pts)

4. Report the coefficients obtained by your model. Would you drop any of the variables used in your model (based on the t-scores or p-values)? (5 pts)

5. Report the MSE obtained on X_train. Score your model (i.e., predict) on X_test. Also report how much the MSE changes. (3 pts)

6. (Bonus Question) Use the stepwise regression to reach your final model. Try different model selection criteria (i.e., AIC, BIC, Adj R^2) and see if you can come up with the same model even with the different criteria. Determine the best model if you get different models with different criteria. (Consider a model that gives the lowest MSE on the test set as the best model). (2 pts)