Multivariate Modeling

DATS 6450

LAB # 5- Multiple Linear Regression and Least Square Estimate

The auto.csv dataset will be used for this LAB. The dependent variable is 'price' and the independent variables are 'normalized-losses', 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-size', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg' and 'highway-mpg'.

In this LAB you want to eliminate features (if applicable) and find the best multiple linear regression model. The multiple linear regression model is:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$$

$\beta_0, \beta_1, \ldots \beta_k$ are unknown values which needs to be estimated using LSE using the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where X and Y are given as :

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \cdots & x_{k,T} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}$$

Matrix X has T rows and (k+1) columns where T is the number of samples and k is number of independent variable.

1- Using Pandas library load the time series dataset from the BB. Split the dataset into training set and test set. Use 80% for training and 20% for testing. Display the training set and testing set array as follow:
   Hint: This can be done using the following library in python.
   "*from sklearn.model_selection import train_test_split*" . Make sure the "Shuffle=False".
2- Plot the correlation matrix using the seaborn package and heatmap function.
3- Using python, construct matrix X and Y using x-train and y-train dataset and estimate the regression model unknown coefficients using the Normal equation (LSE method, above equation). Display the unknown coefficients on the console. Note: You are not allowed to use OLS for this part.
4- Using python, statsmodels package and OLS function, find the unknown coefficients. Compare the results with the step 3. Display the result on the console. Are the unknown coefficient calculated using step 3 and step 4 the same?
5- Perform a prediction for the length of test-set and plot the train, test and predicted values in one graph. Add appropriate x-label, y-label, title, and legend to your graph.
6- Calculate the prediction errors and plot the ACF of prediction errors. Write down your observation.
7- Calculate the forecast errors and plot the ACF of forecast errors. Write down your observation.

8- Calculate the estimated variance of the prediction errors and the forecast errors. Compare the results? What is your observation? Hint: Use need to the following equation to estimate the variance:

$$\hat{\sigma}_e = \sqrt{\frac{1}{T-k-1}\sum_{t=1}^{T} e_t^2}$$

9- Plot the scatter plot between y-test and $\hat{y}_{t+h}$ and display the correlation coefficient between them on the title. Justify the accuracy of this predictor by observing the correlation coefficient between y-test and $\hat{y}_t$.

10- Using a stepwise regression, try to reduce the feature space dimension. You need to use the AIC, BIC and Adjusted R^2 as a predictive accuracy for your analysis. If your analysis recommends an elimination, which feature(s) would you eliminate? You can use the backward or forward or hybrid stepwise regression for feature selection.

11- Perform a complete t-test and F-test analysis on the final model and write down your observations.

Upload the **solution <u>report (as a single pdf</u>**) plus **<u>the .py file</u>** through BB by the due date.