Mini Project: Simple Regression Analysis on Fuel Economy Data

Problem Statement:

You are provided with two dataset “FE2010.csv” and “FE2011.csv”. You are required to work on “FE2010.csv” only for any kind of experiments. The datasets contain different estimates of fuel economy for passenger cars and trucks. For each vehicle, various characteristics are recorded such as the engine displacement or number of cylinders. Along with these values, laboratory measurements are made for the city and highway fuel economy (FE) of the car.

Analyze the data on the relationship between fuel economy and engine displacement. The training data consists of model year 2010 data and the test set is comprised of cars from 2011 that were not in the 2010 data set.

You are required to build a Regression Model for fuel economy (FE), by choosing a single input variable which is the best suitable for predicting FE. You will use 2010 dataset for this purpose. All your work will be validated on 2011 dataset.

Below are the points which your final submission should answers.

Use Excel and Functions

1. Find the best input variable for predicting FE using suitable statistical test(s).
2. Fit a Simple Linear Regression Model using the selected input variable. Use the formulas discussed in the class to calculate the coefficients.
3. Observe the relationship between the Input variable and FE and analyze if they maintain a linear relationship using a suitable chart in Excel.
4. Use appropriate transformation of input variable if the relation above is not linear. Build the Regression model after transformation. Please ask the course instructor for help in variable transformation, if you required so.
5. Calculate the MAPE (Mean Absolute percentage Error) and R^2 of the model. Implement the model on the test data and find out the test accuracy as well. The formula and small note for the error calculation are given at the end of the document.
6. Use a random sampling method to divide the dataset in to 3 parts. Use rand() function.
    a. Take 2 parts for modeling and 1 part for testing at a time randomly.
    b. Check the modeling Error statistics (as given in previous point 5) of the model and test on the 3rd part of the data for testing the error.
    c. Iterate this process 3 time to cover all possible selection of 2 parts for modeling and the 3rd part for testing. There are 3 possible combination in this way. So you would end up with creating 3 models on three different dataset.
    d. Calculate the average model accuracy (Use Error formulas from 5.) and average test accuracy. Judge if they are consistent and provide your comment on what you observe.

e. Compute the Beta coefficients by taking average of the three models.
f. Test the final Accuracy by implementing the model on 2011 dataset.

Use Excel Data Analysis tool

7. Use Data Analysis feature of Excel to bypass the co-efficient calculation formulas and compute the Regression Model directly.
8. You should be able to repeat all the points asked under “Use Excel” using Data Analysis tool. You may need to do the random sampling separately here as well.

Use MySQL

9. Upload the 2010 and 2011 dataset into a MySQL database named “fuel_economy”. The table name should be “fe2010” and “fe2011” respectively.
10. You have already calculated the beta coefficients for the full 2010 dataset. Insert two additional columns for the beta coefficients in the “fe2010” table and populate the columns with beta values. You can just take the previously calculate beta values to populate here. Remember the beta values will be constant for each column here.
11. Once point 10. Is done, Calculate the Predicted value for “feb2011” table by using the input variable from “feb2011” and beta coefficients from “feb2010” table. Insert the predicted values in an additional column in table “feb2010”.

Appendix:

MAPE

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation. It usually expresses accuracy as a percentage, and is defined by the formula:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where A_t is the actual value and F_t is the forecast value.

The difference between A_t and F_t is divided by the Actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points *n*. multiplying by 100 makes it a percentage error.

Coefficient of determination (R2)

In statistics, the **coefficient of determination**, denoted R^2 or r^2 and pronounced "R squared", is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.

It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

A power point has been provided for more detail on R^2 and its formula. Please refer to it.