**PROJECT-1**

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 answer:

**Use Excel and Functions**

1. Find the best input variable for predicting FE using suitable statistical test(s).

ANSWER: The data given shows that the Fuel Efficiency is a continuous variable and the other variables are discrete or binary like dummy 1 and 0. Considering the requirement to predict using a linear regression, and going by the assumptions of linear regressions, we are constrained to use Pearson’s correlation coefficient, which abides by the assumptions of a linear association between two variables. Else Spearman’s rank correlation could also be used. Anyway the following are the results of both types of correlation to identify the best input variable to predict Fuel Efficiency.

|  |  |  |
| --- | --- | --- |
| **VARIABLE-PAIR** | **PEARONS Correlation** | **SPEARMAN RANK Correlation** |
| FE&ED | -0.787393826 | -0.781683753 |
| FE&NC | -0.740217981 | -0.70836013 |
| FE&NG | -0.211284876 | -0.415598271 |
| FE&TL | -0.271938867 | -0.897387156 |
| FE&TCG | -0.069621679 | -0.730469869 |
| FE&IVPC | 0.280344032 | -0.551681306 |
| FE&EVPC | 0.335652854 | -0.445681813 |
| FE&VVT | 0.124952779 | -0.664585866 |
| FE&VVL | 0.096211275 | -0.2753457 |

Fuel Efficiency seems to be highly correlated to Engine Displacement, as per both Rank and Spearman’s correlation coefficients. But using the Spearman’s correlation it is found that Fuel Efficient is also highly correlated with the presence of Trans Lock Up. Since Trans Lock Up Data is a dummy variable with binary values of 1 and 0, the correlation is difficult to interpret. It shows negative relationship and we cannot say whether the presence of absence of Trans Lock Up facility increases fuel efficiency or not. The literature on automobiles says that Trans Lock Ups now are increasing fuel efficiency. But for linear regression this variable is not amenable since the independent variable should have some variation in itself to effectively predict the continuous dependent variable. So we are going by Engine Displacement as a variable which can effectively predict the fuel efficiency of an automobile.

1. Fit a Simple Linear Regression Model using the selected input variable. Use the formulas discussed in the class to calculate the coefficients.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Slope | -4.52093 |  |  |  |  |
| Intercept | 50.56323 |  |  |  |  |
| Equation | Fuel Efficiency = 50.5632 - 4.52093 \* Engine Displacement | | | | |
| Mean of Squared Error | 21.38374 |  |  |  |  |
| Standard Error of Regression | 4.624255 |  |  |  |  |
| Standard Error of Slope | 0.106476 |  |  |  |  |
| Standard Error of Intercept | 0.398479 |  |  |  |  |
| t-Stat for Slope | -42.4595 |  |  |  |  |
| t-Stat for Intercept | 126.8904 |  |  |  |  |
| R-Squared | 0.619989 |  |  |  |  |
| MAPE | 0.103938 |  |  |  |  |

1. Observe the relationship between the Input variable and FE and analyze if they maintain a linear relationship using a suitable chart in Excel.

The above chart shows that there is a downward sloping linear trend line which is fitting fairly between the thickly packed scatter points of Fuel Efficiency and Engine Displacement. Though there are some outliers one the upper side of the line, barring them, there seems to a negative linear relationship between the two variables, which is also highlighted by the negative slope coefficient, negative pearson’s correlation coefficient. In case we would like to fit a trend line to cover scatter plot most appropriately, then a decaying exponential trend line is visible, which is more like an inverse relationship between fuel efficiency and the engine displacement.

1. 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.

Taking the input variable as raw, and inverse and natural log transformations the following correlations have emerged. Log transformation of both dependent and independent variables generates the higher correlation coefficient

|  |  |  |  |
| --- | --- | --- | --- |
| Both Linear | Invrse Indep | Both log forms | Log  Indep |
| -0.78739 | 0.814941 | -0.82819 | -0.82119 |

The below line scatter plot is drawn after transforming the engine displacement into its inverse values.

The below line scatter plot is drawn after transforming the engine displacement into its log values.

The below line scatter plot is drawn after transforming the both the variables into their log forms

1. Calculate the MAPE (Mean Absolute percentage Error) and R2 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.

Two models are estimated using an inverse of the engine displacement and log of engine displacement

Log of Engine Displacement Model

|  |  |  |  |
| --- | --- | --- | --- |
| Slope | -16.4473 |  |  |
| Intercept | 54.21195 |  |  |
| Equation | Fuel Efficiency = 54.2119 - 16.4473 \* Engine Displacement | | |
| Mean of Squared Error | 18.32452 |  |  |
| Standard Error of Regression | 4.280715 |  |  |
| Standard Error of Slope | 0.343828 |  |  |
| Standard Error of Intercept | 0.427575 |  |  |
| t-Stat for Slope | -47.8358 |  |  |
| t-Stat for Intercept | 126.7892 |  |  |
| R-Squared | 0.674354 |  |  |
| MAPE | 0.09662 |  |  |

Inverse of Engine Displacement Model

|  |  |  |  |
| --- | --- | --- | --- |
| Slope | 48.68371 |  |  |
| Intercept | 18.75657 |  |  |
| Equation | Fuel Efficiency = 18.7565 - 48.6837 \* Engine Displacement | | |
| Mean of Squared Error | 18.89994 |  |  |
| Standard Error of Regression | 4.347406 |  |  |
| Standard Error of Slope | 1.041509 |  |  |
| Standard Error of Intercept | 0.365385 |  |  |
| t-Stat for Slope | 46.74345 |  |  |
| t-Stat for Intercept | 51.33374 |  |  |
| R-Squared | 0.664129 |  |  |
| MAPE | 0.098719 |  |  |

6. Use a random sampling method to divide the dataset in to 3 parts. Use rand() function.

a. Take 2 parts for modelling and 1 part for testing at a time randomly.

b. Check the modelling 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 modelling 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **ITER1** | **ITER2** | **ITER3** | **MEAN** |
| **TRAINING ACCURACY** | **MAPE** | 0.096313 | 0.09902 | 0.094614 | 0.096649221 |
|  | **RSQ** | 0.673069 | 0.655025 | 0.681192 | 0.669761937 |
| **Beta 1** | **SLOPE** | -16.089 | -16.2362 | -16.4602 | -16.26181203 |
| **Beta 0** | **INTERCEPT** | 53.67433 | 53.98871 | 54.23083 | 53.9646227 |
| **Number of Obs.** |  | 772 | 740 | 755 |  |
| **TEST ACCURACY** | **MAPE** | 0.097861 | 0.092358 | 0.100966 | 0.09706178 |

When we compare the MAPE and the slope and intercept coefficients there is some amount of consistency observed since whatever way the data for FE2010 is taken, by random 2/3 and 1/3 training to testing proportion, the slope and intercept coefficients are almost similar. The MAPE for the third iteration is relatively higher than the two iterations, due to the number of observations used in the sample

Using the average slope coefficient and average intercept extracted out of all the three iterations, we get the following result and MAPE accuracy for the FE2011 data. R square Values do not make any sense while calculating for the testing FE2011 data or for the 1/3 portion of the test data from the previous FE2010 calculations. Since the Beta coefficient is external to the regression equation, so interpreting R squared for the test data is not meaningful. So, only MAPE is calculated to test for the accuracy of the regression.

Using the equation – Fuel Efficiency = 53.9646 - 16.2618 \* Engine Displacement we get the the MAPE of 0.1072, which is around 10.72% error, in prediction using this model, which is slightly higher than the training and testing models using FE 2010 data.

**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.

**The same results are replicated using the EXCEL function on regression. The details are in the excel file attached named “FE2010-Function-Regressions”**

**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.

mysql>CREATE DATABASE fuel\_economy

mysql>USE fuel\_economy

mysql>CREATE TABLE fe2010 (EngDispl Decimal(10.4), NumCyl Decimal(10,4), FE Decimal (10,4), NumGears INT, TransLockup INT, TransCreeperGear INT, IntakeValvePerCyl INT, Exhaust ValvesPerCyl INT, VarValveTiming INT, VarValveLift INT)

mysql>CREATE TABLE fe2011 (EngDispl Decimal(10.4), NumCyl Decimal(10,4), FE Decimal (10,4), NumGears INT, TransLockup INT, TransCreeperGear INT, IntakeValvePerCyl INT, Exhaust ValvesPerCyl INT, VarValveTiming INT, VarValveLift INT)

mysql>LOAD DATA INFILE ‘c:/ProgramData/MySQL/MySQL Server 8.0/Uploads/FE2010.csv’

INTO TABLE fe2010

FIELDS TERMINATED BY ‘|’

ENCLOSED BY ‘”’

LINES TERMINATED BY ‘\n’

IGNORE 1 ROWS;

mysql>LOAD DATA INFILE ‘c:/ProgramData/MySQL/MySQL Server 8.0/Uploads/FE2011.csv

INTO TABLE fe2011

FIELDS TERMINATED BY ‘|’

ENCLOSED BY ‘”’

LINES TERMINATED BY ‘\n’

IGNORE 1 ROWS;

NOTE: There is something unique happening here. The decimal values from the source CSV files are not getting uploaded as decimals, rather they are getting rounded off. I could not figure it out as to how to get them as it is. There is something which MySQL is doing, like overriding the input format with its own round off format.

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.

mysql>ALTER TABLE fe2010 ADD COLUMN Intercept Decimal (10,4);

mysql>ALTER TABLE fe2010 ADD COLUMN Slope Decimal(10,4);

mysql>UPDATE fe2010 SET Intercept = 53.9646

mysql>UPDATE fe2010 SET Slope = -16.2618

mysql>ALTER TABLE fe2011 ADD COLUMN Intercept Decimal (10,4);

mysql>ALTER TABLE fe2011 ADD COLUMN Slope Decimal(10,4);

mysql>UPDATE fe2011 SET Intercept = 53.9646

mysql>UPDATE fe2011 SET Slope = -16.2618

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”.

mysql>ALTER TABLE fe2011 ADD COLUMN logEngDispl Decimal (10,4);

mysql>UPDATE fe2011 SET logEngDispl = LN(EngDispl);

mysql>ALTER TABLE fe2011 ADD COLUMN PREDFE2011 Decimal(10,4);

mysql>UPDATE fe2011 SET PREDFE2011 = Intercept + ((Slope)\*(logEngDispl));

In the above approach I have populated the new columns with the intercept and slope values in the fe2011 testing data set, which contains only 245 records. Later I added a new column to store the predicted values of FE 2011 based on the Engine Displacement value of fe2011 and the intercept and slope coefficients already estimated from the training data set.

I tried to add the column in fe2010, which has 1107 records and populate it with the fe2011.PREDFE2011 column, but the querry is getting executed with the INSERT INTO command, but the data is not getting populated.

mysql> ALTER TABLE fe2010 ADD COLUMN PREDFE2011 DECIMAL(10,4);

Query OK, 0 rows affected (0.24 sec)

Records: 0 Duplicates: 0 Warnings: 0

mysql> INSERT INTO fe2010 (PREDFE2011) (SELECT PREDFE2011 FROM fe2011);

Query OK, 245 rows affected (0.08 sec)

Records: 245 Duplicates: 0 Warnings: 0

mysql> SELECT PREDFE2011 FROM fe2010 LIMIT 5;

+------------+

| PREDFE2011 |

+------------+

| NULL |

| NULL |

| NULL |

| NULL |

| NULL |

+------------+

5 rows in set (0.00 sec)