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

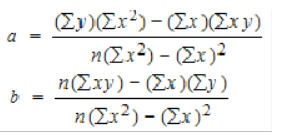
|  |  |  |  |
| --- | --- | --- | --- |
| -0.78739 | correlation between FE and Engine displacement is best | | |
| -0.74022 | correlation between FE and NumCyl |  |  |
| -0.21128 | correlation between FE and NumGears |  |  |
| -0.27194 | correlation between FE and TransLockup |  |  |
| -0.06962 | correlation between FE and TransCreeperGear | |  |
| 0.280344 | correlation between FE and IntakeValvepercyl | |  |
| 0.335653 | correlation between FE and Exhaust valve per Cyl | |  |
| 0.124953 | correlation between FE and Var valve Timing | |  |
| 0.096211 | correlation between FE and Varvalvelift |  |  |

Ans: our best input variable is **Engine displacement** as has maximum value **(0.78)** ignoring sign

2. Fit a Simple Linear Regression Model using the selected input variable. Use the formulas

discussed in the class to calculate the coefficients.

Ans:

 **a = 50.52 b= -4.52**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a = (38420.08 \* 15504.37) - (3882.7 \* 126227.7)  ((1107 \* 15504.37) - (15075359.29) | | | | | | | | |
|  |  |  |  |  |  |  |  |  |

b = (1107 \* 126227.7) - (3882.7 \* 38420.08)

(1107 \* 15504.37) - (15075359.29)

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.

Ans:

-0.82119 correlation between FE & Log Trans of Engine displacement has increased from (-78739) -0.80854 correlation between FE & square root Trans of Engine displ has increased from (-78739) -0.81378 correlation between FE & cube root Trans of Engine displ has increased from (-78739).

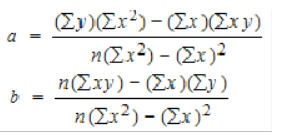
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| Note: Among transformation with Log, square root and cube root, **Transformation with log has max value** | | | | | | | | | |

5. 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.

Ans:

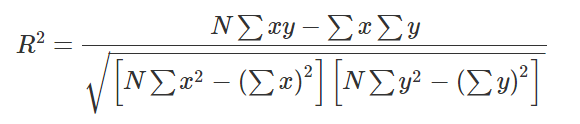
 **a = 54.21 b= -16.44**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a = (38420.08 \* 1711.95) - (1312.83 \* 43014.48)  ((1107 \* 1711.95) - (1723537.94) | | | | | | | | |
|  |  |  |  |  |  |  |  |  |

b = (1107 \* 43014.48) - (1312.83 \* 38420.08)

(1107 \* 1711.95) - (1723537.94)

MAPE = mean(abs(y-yhat/y)) = 319.089815



R square = (1107 \* 43014.48) - (1312.83 \* 38420.08)

Sqrt([1107 \* 1711.95 - (1723537.94)] \* [1107 \* 1395606.07 – (1476102808)])

R square = 0.62

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.

**Ans: a, b, MAOR are not consistent**

e. Compute the Beta coefficients by taking average of the three models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **TESTING** | **TRAINING** | **MAPE** | **a** | **b** |
| **Model 1** | **A,B** | **C** | 0.097 | 51.95 | -4.59 |
| **Model 2** | **A,C** | **B** | 0.108 | 50.56 | -4.52 |
| **Model 3** | **B.C** | **A** | 0.1 | 48.6 | -4.21 |
| **SUM** |  |  |  | 151.11 | -13.32 |
| **AVGERAGE** |  |  |  | 50.37 | -4.44 |

f. Test the final Accuracy by implementing the model on 2011 dataset.

|  |  |
| --- | --- |
| **MAPE =** | **0.113376** |

7. Use Data Analysis feature of Excel to bypass the co-efficient calculation formulas and

compute the Regression Model directly.

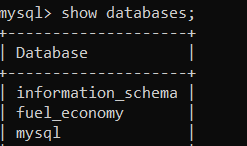
8. You should be able to repeat all the points asked under “Use Excel” using Data Analysis

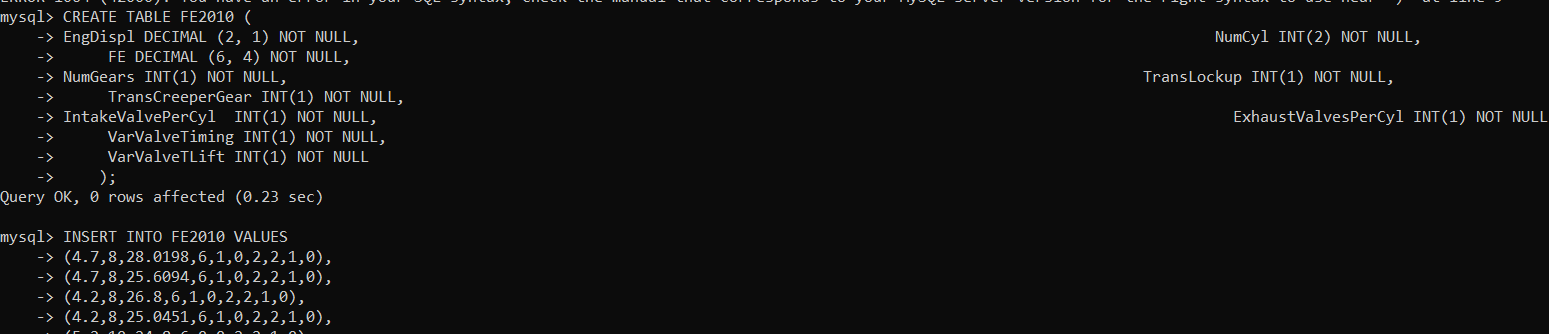
tool. You may need to do the random sampling separately here as well.

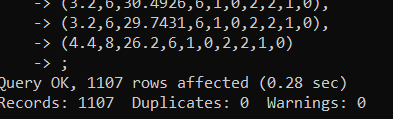
**Use MySQL**

9. Upload the 2010 and 2011 dataset into a MySQL database named “fuel\_economy”. The

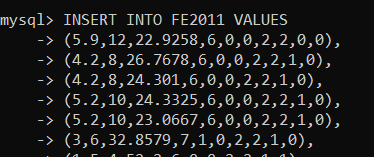
table name should be “fe2010” and “fe2011” respectively.

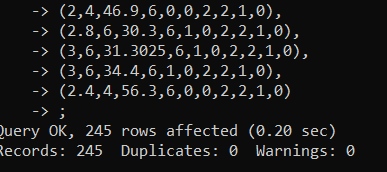






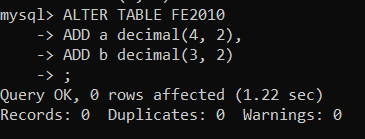






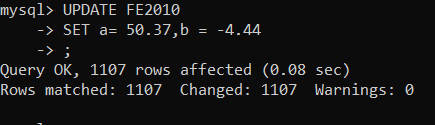
10. You have already calculated the beta coefficients for the full 2010 dataset. Insert two

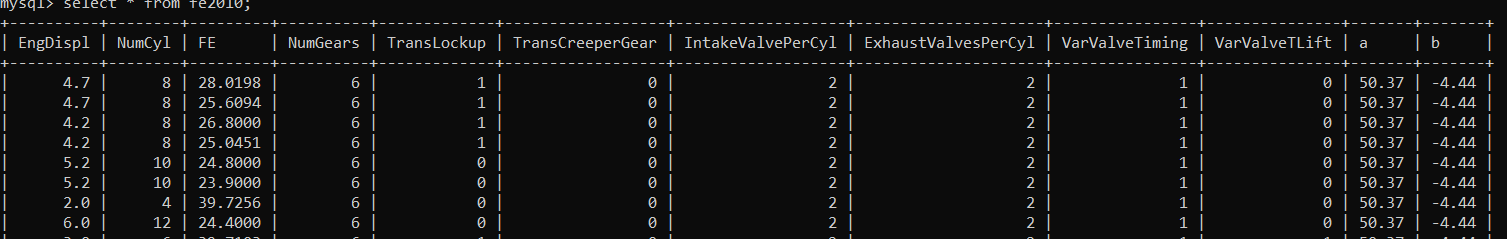
additional columns for the beta coefficients in the “fe2010” table.



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.

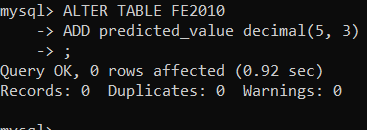




11. Once point 10 is done, Calculate the Predicted value for “feb2011” table by using the

input variable from “feb2011” and beta coefficients from “feb2010” table. Insert the

predicted values in an additional column in table “feb2010”.



SELECT (unit\_cost \* quantity) AS claim\_value

FROM `xx\_non\_part\_usage`

WHERE request\_id = request\_id

GROUP BY request\_id

INSERT INTO fe2010 (predicted\_value)

SELECT (50.37+(-4.44\*EngDispl)) AS predicted\_values

from fe2011

UPDATE fe2010

SET fe2010.predicted\_value = (50.37+(-4.44\*EngDispl))

from fe2011

where fe2011.EngDisp = fe2010.EngDisp