**PROJECT**

**SIMPLE REGRESSION ANALYSIS**

**ON**

**FUEL ECONOMY DATA**

***Jaydeep Potdar***

CONTENTS

|  |  |
| --- | --- |
| Sr.No. | Contents |
| 1 | Introduction & data sets |
| 2 | Finding the best input variable for predicting FE using suitable statistical test(s). |
| 3 | Simple Linear Regression Model |
| 4 | Observing the relationship between the Input variable and FE and analyzing if they maintain a linear relationship using a suitable chart in Excel . Transformation of input variable and building the Regression model after transformation. |
| 5 | Mean Absolute percentage Error and R2  Implementing the model on the test data and finding the test accuracy as well |
| 6 | Random sampling method to divide the dataset in to 3 parts |
| 7 | Modelling and Testing |
| 8 | Use Data Analysis feature of Excel to bypass the co-efficient calculation formulas and compute the Regression Model directly |
| 9 | Use MySQL |
|  |  |

1. **Introduction and Data Sets**

**Introduction**

We have two datasets, “FE2010.csv” and “FE2011.csv”. The The datasets contain different estimates of fuel economy for passenger cars and trucks

We are required to work on “FE2010.csv” only for any kind of experiment.

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.

We have to 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 are not in the 2010 data set.

We 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. Using 2010 dataset for this purpose.

To use the coefficients from 2010 dataset for 2011 dataset.

**Data Sets**

The data file are in .CSV format . The data is in single column separated with | .

Using ‘text to column’ tab under the ‘data’ tab convert the data points in tabular form.

Save the data file in .xlsx format.

Worksheet named – FE2010 in the file FE2010 SRM MAPE R2.xlxs

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

Using the function “correl” find out the correlation and R2 for all the variable. The data is produced below.

|  |  |  |  |
| --- | --- | --- | --- |
| FE (X) | Variable (Y) | Correlation | R2 |
| FE | EngDispl | -0.787 | 0.620 |
| FE | NumCyl | -0.740 | 0.548 |
| FE | NumGears | -0.211 | 0.045 |
| FE | TransLockup | -0.272 | 0.074 |
| FE | TransCreeperGear | -0.070 | 0.005 |
| FE | IntakeValvePerCyl | 0.280 | 0.079 |
| FE | ExhaustValvesPerCyl | 0.336 | 0.113 |
| FE | VarValveTiming | 0.125 | 0.016 |
| FE | VarValveLift | 0.096 | 0.009 |

Worksheet named “Best input var’ in the file FE2010.xlsx

Using conditional formatting find out maximum value from the R2.

**Comparing the correlation values and R2 it is drawn that EngDispl is the best input variable for predicting the FE.**

1. **Simple Linear Regression Model**

Worksheet “EngDispl SRM MAPE R2” in the file “FE2010 SRM MAPE R2.xlsx”

Simple Regression Model

Y = β0 + β1 (X )

So we have to calculate β0 and β1

β0 = ybar - Beta1 xbar

β1 = sum of (xiyi) - nxbar ybar / sum of (xi)square - n xbar square

The required values are calculated in the worksheet. The calculated values are reproduced below:

n = 1107

ybar = 34.706

xbar = 3.507

sum of (xiyi) = 126227.692

sum of (xi)square = 15504.370

Then

β0 = 50.563

β1 = -4.521

Different Worksheet is prepared for each variable in the file “FE2010 SRM MAPE R2.xlsx”

1. **the relationship between the Input variable and FE**

The relationship is linear and transformation is not required. Hence we continue with the above calculated coefficients.

1. **Mean Absolute percentage Error , R2 and Accuracy**

Worksheet “EngDispl SRM MAPE R2” in the file “FE2010 SRM MAPE R2.xlsx”

MAPE = 1/n[absolutesummation(at - ft)/at]

= 10.39

Correlation = n(summation(xy) - summation(x)\* Summation(y)/square root of [n\*(summation(x square)-summation of X whole square)] \* [n\*(summation(y square)-summation of y whole square)]

= -0.79

RSquared = 0.62

Accuracy = 100-MAPE

=89.61

1. **Random sampling method to divide the dataset in to 3 parts**

Worksheet “RandPartion” in the file “FE2010 modeling and testing”

Use Rand() function in additional column.

Convert the data to ‘values’.

Sort the data for the above column.

There are total 1107 rows. Divide them into three as

Set1 : Row 2 to Row 370 ,

Set2: Row 371 to Row 739 &

Set3: Row 740 to Row 1107

1. **Modeling and testing is done as below in the File “FE2010 modeling and testing.xlsx”**

|  |  |  |  |
| --- | --- | --- | --- |
| Modeling | | Testing | |
| Set | Reference Worksheet | Set | Reference Worksheet |
| Set1 + Set2 | Train Set1 Set2 | Set3 | Test Set3 |
| Set2 + Set3 | Train Set2 Set3 | Set1 | Test Set1 |
| Set1 + Set3 | Train Set1 Set3 | Set2 | Test Set2 |

SRM, Correlation , MAPE and R2 are calculated for each Train Set and coefficients used in Test Set.

The Average values are as below in the worksheet “average” in the file “FE2010 modeling and testing.xlsx”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Train Set Name | Model MAPE | Model Accuracy | Beta0 | Beta1 |
| TrainSet1Set2 | 9.710 | 90.290 | 51.956 | -4.595 |
| TrainSet2Set3 | 10.004 | 89.996 | 48.603 | -4.217 |
| TrainSet1Set3 | 10.671 | 89.329 | 50.324 | -4.535 |
| Average Model | 10.128 | 89.872 | 50.294 | -4.449 |
|  |  |  |  |  |
|  |  |  |  |  |
| Test Name | Test MAPE | Test Accuracy |  |  |
| Test Set3 | 14.439 | 85.561 |  |  |
| Test Set1 | 10.484 | 89.516 |  |  |
| Test Set2 | 9.500 | 90.500 |  |  |
| Average Test | 11.474 | 88.526 |  |  |

**Average Model accuracy is more than Average Test Accuracy**

**Average Model accuracy is more than Test Accuracy of entire data set**

The model is implemented on 2011 dataset. File name FE2011.xlsx

The results are copied below

|  |  |
| --- | --- |
| **MAPE** | **11.27** |
| **Correlation** | **-0.84** |
| **R2** | **0.70** |
| **Accuracy** | **88.73** |

1. **Use Data Analysis feature of Excel to bypass the co-efficient calculation formulas and compute the Regression Model directly**

File name “FE2010 SRM MAPE R2 Direct.xlsx”

Data > Data Analysis > Regression

Give input range for X values and Y values (tick- line fit plots)

System will directly calculate the statistics and predicted Y. Thus we can calculate the MAPE and accuracy. The results are reproduced below:



File name “FE2010 modeling and testing – Direct.xlsx”

Again the data set is divided in three parts using Rand() function as stated above.

Same procedure is applied for Modeling and coefficients are used for Test Sets. Averages are reproduced below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Train Set Name | Model MAPE | Model Accuracy | Beta0 | Beta1 |
| TrainSet1Set2 | 10.401 | 89.599 | 51.380 | -4.722 |
| TrainSet2Set3 | 10.435 | 89.565 | 50.197 | -4.439 |
| TrainSet1Set3 | 10.322 | 89.678 | 50.134 | -4.410 |
| Average Model | 10.386 | 89.614 | 50.570 | -4.524 |
|  |  |  |  |  |
| Test Name | Test MAPE | Test Accuracy |  |  |
| Test Set3 | 10.646 | 89.354 |  |  |
| Test Set1 | 10.205 | 89.795 |  |  |
| Test Set2 | 10.493 | 89.507 |  |  |
| Average Test | 10.448 | 89.552 |  |  |

**Average Model accuracy is more than Average Test Accuracy**

**Average Model accuracy is more than Test Accuracy of entire data set**

The model is implemented on 2011 dataset. File name FE2011 Direct.xlsx

The results are copied below

|  |  |
| --- | --- |
| **MAPE** | **11.21** |
| **Correlation** | **-0.84** |
| **R2** | **0.70** |
| **Accuracy** | **88.79** |

1. **Using MySQL**

SQL File FeulEconomy

Create database FeulEconomy

New Script FuelEconomy

use FuelEconomy;

-- import tables FE2010 & FE2011 from database

alter table fe2010 drop column betazero;

alter table fe2010 add betazero float(12,3);

update fe2010 set betazero = 050.563 where betazero is null;

select \* from fe2010;

alter table fe2010 drop column betaone;

alter table fe2010 add betaone float(12,3);

update fe2010 set betaone = -4.521 where betaone is null;

alter table fe2010 drop column predicted;

alter table fe2010 add predicted float(12,3);

update fe2010 set predicted = betazero + betaone\*EngDispl where predicted is null;

select \* from fe2010;

alter table fe2011 drop column predicted;

alter table fe2011 add predicted float(12,3);

update fe2011 a,fe2010 b set a.predicted=b.betazero+(a.EngDispl\*b.betaone) where a.predicted is null;

select \* from fe2011**;**

***Jaydeep Potdar***