SIMPLE REGRESSION ANALYSIS

On

FUEL ECONOMY DATA

Submitted by

DEEPAK JOY

Business Analytics

TABLE OF CONTENTS

OBJECTIVE

The project aims to perform Simple Regression Analysis on Fuel Economy Data.

INTRODUCTION

Linear regression is the most basic type of regression and commonly used predictive analysis. The overall idea of regression is to examine two things: (1) Does a set of predictor variables do a good job in predicting an outcome variable? Is the model using the predictors accounting for the variability in the changes in the dependent variable? (2) Which variables in particular are significant predictors of the dependent variable? And in what way do they--indicated by the magnitude and sign of the beta estimates--impact the dependent variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

Three major uses of regression analysis are (1) causal analysis, (2) forecasting an effect, and (3) trend forecasting. Other than correlation analysis, which focuses on the strength of the relationship between two or more variables, regression analysis assumes dependence or causal relationship between one or more independent variables and one dependent variable.

Firstly, the regression might be used to identify the strength of the effect that the independent variables have on a dependent variable..

Secondly, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable change with a change in one or more independent variables.

Thirdly, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates

APPROACH

Two datasets, “FE2010.csv” and “FE2011.csv” containing different estimates of fuel economy for passenger cars and trucks were given. The regression should be done on the fe2010 dataset and the coefficient and intercept obtained should be employed on the fe2011 dataset for predicting the fuel economy.

First, for predicting fuel economy, the correlation of all the other 9 variables with the former in the fe2010 dataset was found out, and the most correlated variable was chosen as the input variable. **Engine displacement** was the most correlated variable found. Now, dummy values were assigned to the coefficient and intercept variable created and a prediction column was made on the fuel economy of the vehicles using the formula

y =mx + c

where y is the prediction

x is the input variable

c is the intercept

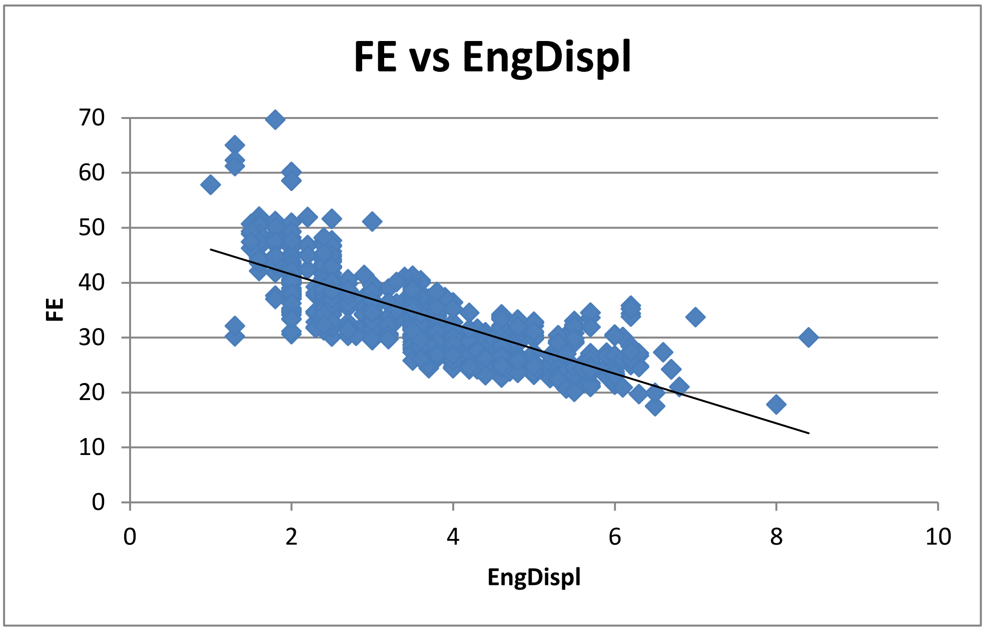
The error column is calculated, which is the difference between the original value and the prediction. An error square column and the sum of squares of error (SSE) was calculated. Now the SSE is fed into the **solver/data analysis tool** to minimize it by changing the coefficient and intercept. This is done in order to reduce the error prediction error and thus obtain optimum coefficients with the least possible error for prediction on the test data. The test accuracy on the data was understood by finding R^2 and Mean Absolute Percentage Error (MAPE).

Before implementing on fe2011 data, the training data was divided into 3 parts and 2 parts of this data were used for modelling and the rest 1 part for testing. This was iterated to cover all possible sections for modelling and testing.

The entire process was repeated using data analysis tool as well. The raw data of fe2010 & fe2011 was imported to MySQL and the coefficients added to 2010’s data were used to predict 2011’s FE using its Engine displacement in 2010 table using update command.

RESULTS AND OBSERVATIONS

- The correlation using FE & EngDispl was the maximum and was obtained as **-0.787**
- The graph between FE and EngDispl is as below



- The coefficients, namely x-variable and intercept obtained in sub-datasets were averaged and was obtained as

X-variable = **-4.532283012**

Intercept = **50.60588291**

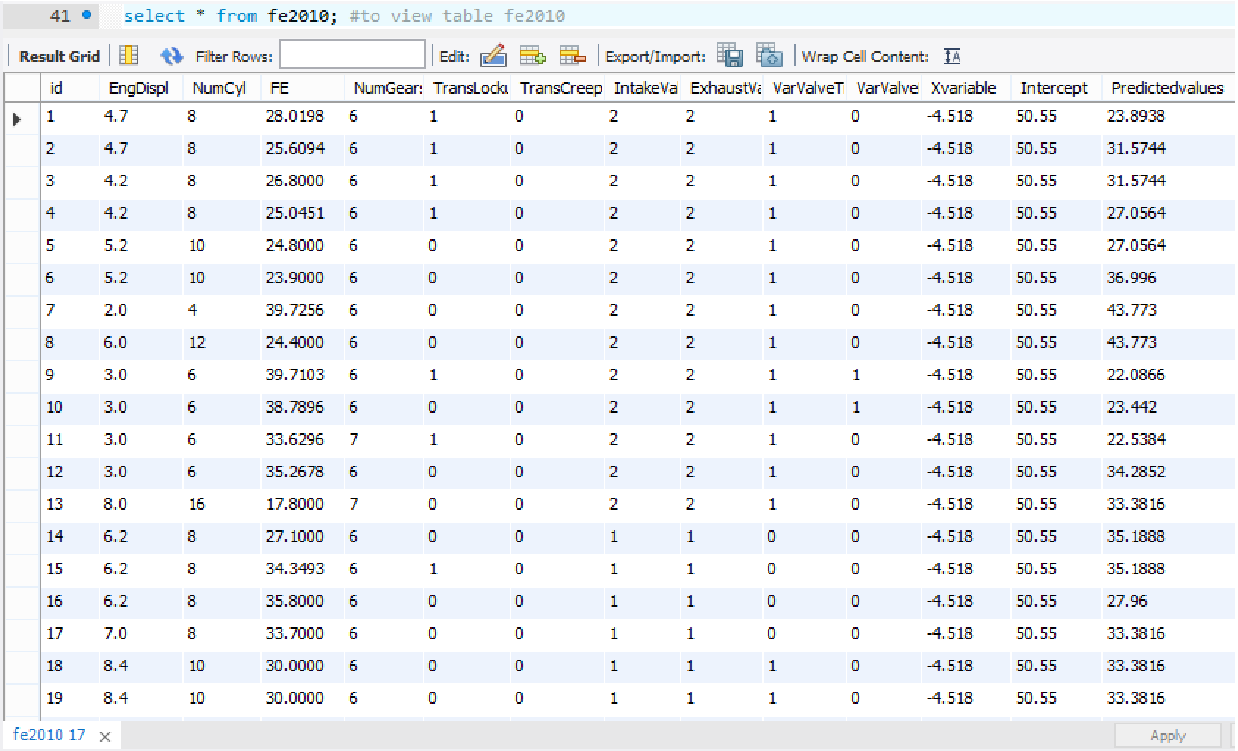
The coefficients were observed to be the same being calculated with **Excel & functions and the Data Analysis Tool.**

- The R squared and MAPE was found on the fe2011 test dataset as

R squared = **0.701864**

MAPE = **11.20899**

- The datas were imported to MySQL and the Predicted values were calculated in fe2010 dataset.(Screenshots attached. & **MySQL code** attached in APPENDIX)



```
select * from fe2010; #to view table fe2010
```

id	EngDispl	NumCyl	FE	NumGear:	TransLocku	TransCreep	IntakeVa	ExhaustV:	VarValveT	VarValve	Xvariable	Intercept	Predictedvalues
1	4.7	8	28.0198	6	1	0	2	2	1	0	-4.518	50.55	23.8938
2	4.7	8	25.6094	6	1	0	2	2	1	0	-4.518	50.55	31.5744
3	4.2	8	26.8000	6	1	0	2	2	1	0	-4.518	50.55	31.5744
4	4.2	8	25.0451	6	1	0	2	2	1	0	-4.518	50.55	27.0564
5	5.2	10	24.8000	6	0	0	2	2	1	0	-4.518	50.55	27.0564
6	5.2	10	23.9000	6	0	0	2	2	1	0	-4.518	50.55	36.996
7	2.0	4	39.7256	6	0	0	2	2	1	0	-4.518	50.55	43.773
8	6.0	12	24.4000	6	0	0	2	2	1	0	-4.518	50.55	43.773
9	3.0	6	39.7103	6	1	0	2	2	1	1	-4.518	50.55	22.0866
10	3.0	6	38.7896	6	0	0	2	2	1	1	-4.518	50.55	23.442
11	3.0	6	33.6296	7	1	0	2	2	1	0	-4.518	50.55	22.5384
12	3.0	6	35.2678	6	0	0	2	2	1	0	-4.518	50.55	34.2852
13	8.0	16	17.8000	7	0	0	2	2	1	0	-4.518	50.55	33.3816
14	6.2	8	27.1000	6	0	0	1	1	0	0	-4.518	50.55	35.1888
15	6.2	8	34.3493	6	1	0	1	1	0	0	-4.518	50.55	35.1888
16	6.2	8	35.8000	6	0	0	1	1	0	0	-4.518	50.55	27.96
17	7.0	8	33.7000	6	0	0	1	1	0	0	-4.518	50.55	33.3816
18	8.4	10	30.0000	6	0	0	1	1	1	0	-4.518	50.55	33.3816
19	8.4	10	30.0000	6	0	0	1	1	1	0	-4.518	50.55	33.3816

- The following supporting files are also attached
  - Fe2010 excel workbook with all calculated sheets
  - Fe2011 excel workbook with all calculations
  - ‘MySQL_miniproject’ .(sql file)
  - ‘MySQL code’ (.txt file)
  - Screenshots of sql outputs

APPENDIX

```
create database fuel_economy; # Creating DB fuel economy
use fuel_economy;

CREATE TABLE `fuel_economy`.`fe2010` ( #Creating table fe2010
          id int auto_increment primary key,
  `EngDispl` DECIMAL(2,1) NOT NULL,
  `NumCyl` INT NOT NULL,
  `FE` DECIMAL(6,4) NOT NULL,
  `NumGears` INT NOT NULL,
  `TransLockup` INT NOT NULL,
  `TransCreeperGear` INT NOT NULL,
  `IntakeValvePerCyl` INT NOT NULL,
  `ExhaustValvesPerCyl` INT NOT NULL,
  `VarValveTiming` INT NOT NULL,
  `VarValveLift` INT NOT NULL);
ALTER TABLE fe2010 ADD        #Adding the Xvariable value to the table fe2010
Xvariable decimal(4,3) DEFAULT -4.518 NOT NULL;
ALTER TABLE fe2010 ADD        #Adding the Intercept value to the table fe2010
Intercept decimal(4,2) DEFAULT 50.55 NOT NULL;

CREATE TABLE `fuel_economy`.`fe2011` ( #Creating table fe2011
          id int auto_increment primary key,
  `EngDispl` DECIMAL(2,1) NOT NULL,
  `NumCyl` INT NOT NULL,
  `FE` DECIMAL(6,4) NOT NULL,
  `NumGears` INT NOT NULL,
  `TransLockup` INT NOT NULL,
  `TransCreeperGear` INT NOT NULL,
  `IntakeValvePerCyl` INT NOT NULL,
  `ExhaustValvesPerCyl` INT NOT NULL,
  `VarValveTiming` INT NOT NULL,
  `VarValveLift` INT NOT NULL);

# Table Data import wizard is used to import the csv files of fe2010 & fe2011 to the
tables

alter table fe2010 add Predictedvalues double;
                                        # Adding new column in fe2010 for 2011's prediction
SET SQL_SAFE_UPDATES = 0;

update fe2010,fe2011
```

```
        #Predicting fe2011's FE using the xvariable and intercept from fe2010 table
set fe2010.Predictedvalues=fe2010.Intercept+ (fe2010.Xvariable*fe2011.EngDispl)
where fe2010.id=fe2011.id;                    #Prediction column is added to fe2010

select * from fe2010; #to view the updated table fe2010
select * from fe2011; #to view table fe2011
```