SIMPLE REGRESSION ANALYSIS ON FUEL ECONOMY DATA

Submitted by:

Pranoy Ray Chowdhury

Data Analytics with R, Excel and Tableau

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.

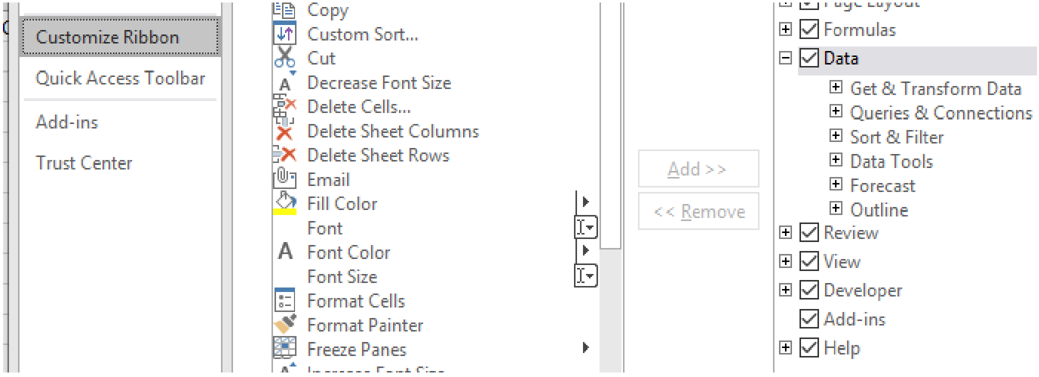
Thirdly, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates

APPROACH

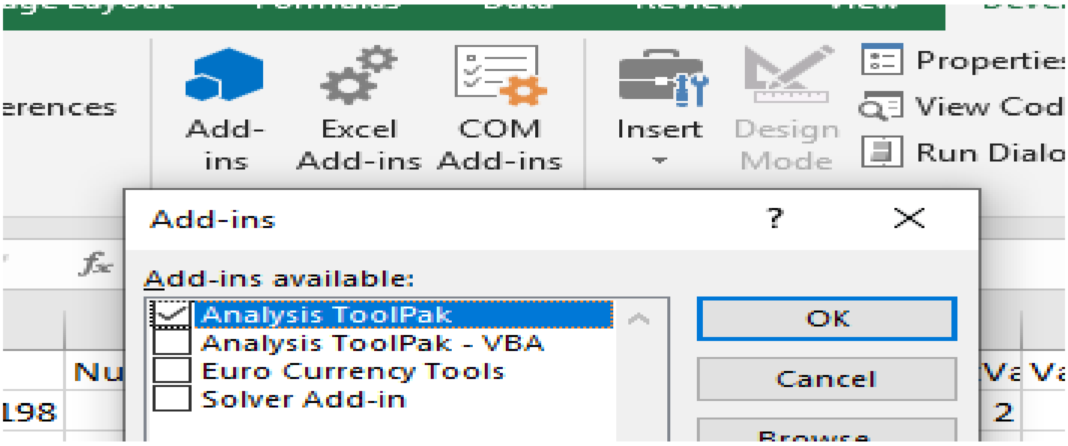
Two datasets, “FE2010.csv” and “FE2011.csv” containing different estimates of fuel economy for passenger cars and trucks were given. The regression should be done on the fe2010 dataset and the coefficient and intercept obtained should be employed on the fe2011 dataset for predicting the fuel economy.

The first sheet contains the following variables: -
EngDispl = Engine Displacement (cu.in.) [continuous variable]
NumCyl = No of cylinder [continuous variable]
FE = Fuel Economy [continuous variable]
NumGears = Number of Gears
TransLockup = Transmission Lockup [Binomial Variable]
TransCreeperGear = Transmission Creeper Gear [Binomial Variable]
IntakeValvePerCyl = Intake Valve Per Cylinder [factorial Variable]
ExhaustValvesPerCyl = Exhaust Valve Per Cylinder [factorial Variable]
VarValveTiming = Variable Valve Timing [Binomial Variable]
VarValveLift = Variable Valve Lift [Binomial Varaible]

We used Data Analysis Package from Add “Developer” tab and excel add ins to get the “Data Analysis” option under DATA tab:



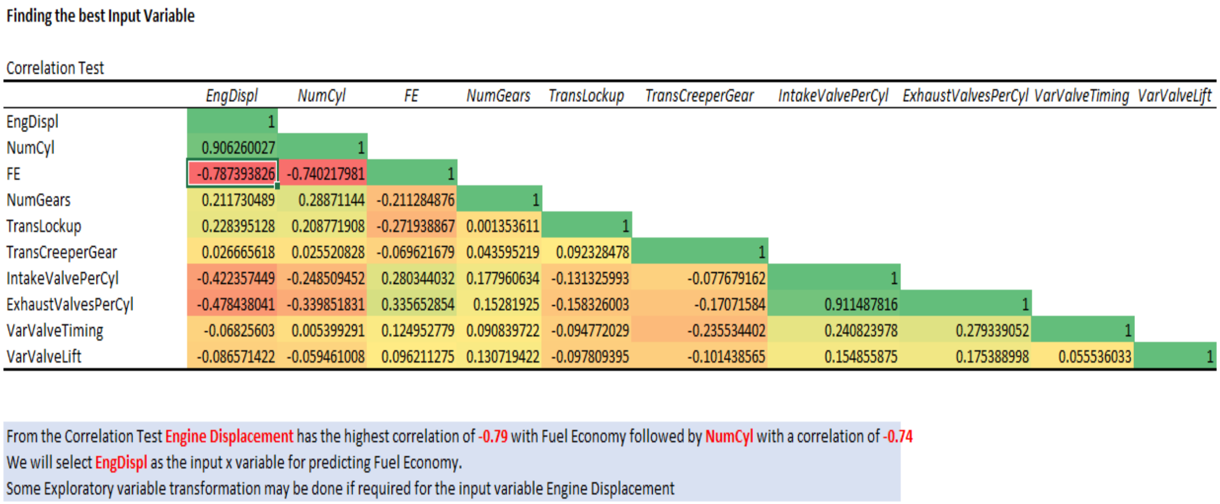
☑ Formulas
☑ Data
⊞ Get & Transform Data
⊞ Queries & Connections
⊞ Sort & Filter
⊞ Data Tools
⊞ Forecast
⊞ Outline
☑ Review
☑ View
☑ Developer
☑ Add-ins
☑ Help



Add-ins available:
☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Euro Currency Tools
□ Solver Add-in

First, for predicting fuel economy (FE), the correlation of all the 9 variables with the former (FE) in the fe2010 dataset was found out, and the most correlated variable was chosen as the input variable. As the correlation coefficient value goes towards 0, the relationship between the two

variables will be weaker. In the Figure below It shows that EngDispl and NumCyl are the most strongly related to FE Negatively.
Engine displacement was the most correlated variable found with R value of 0.79.

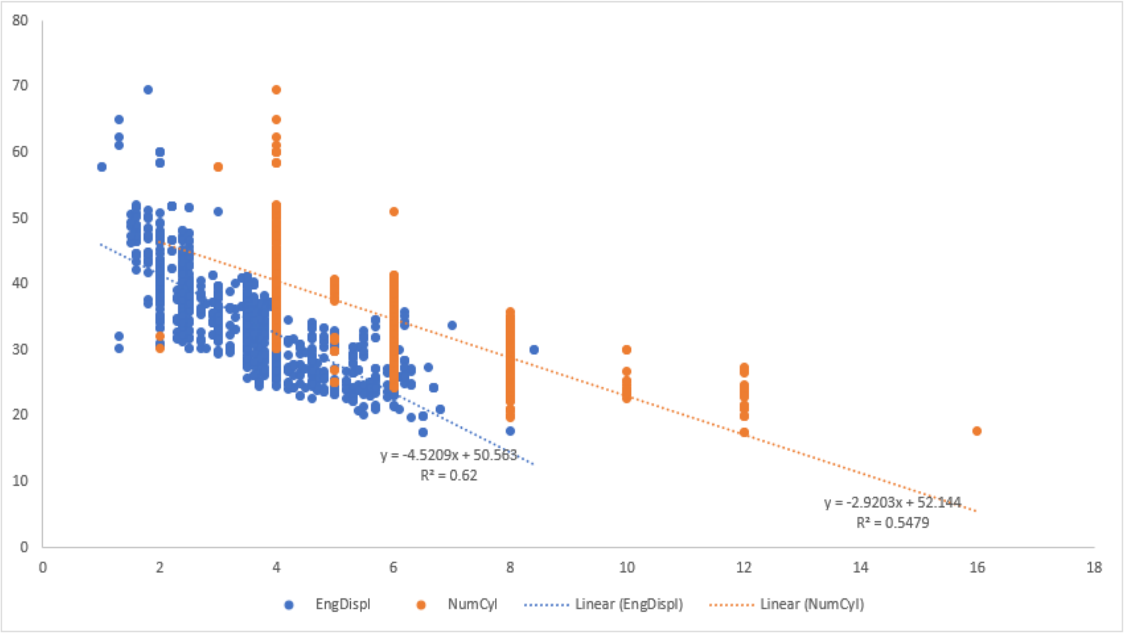


Finding the best Input Variable

Correlation Test

	EngDispl	*NumCyl*	*FE*	*NumGears*	*TransLockup*	*TransCreeperGear*	*IntakeValvePerCyl*	*ExhaustValvesPerCyl*	*VarValveTiming*	*VarValveLift*
EngDispl	1									
NumCyl	0.906260027	1								
FE	-0.787393826	-0.740217981	1							
NumGears	0.211730489	0.28871144	-0.211284876	1						
TransLockup	0.228395128	0.208771908	-0.271938867	0.001353611	1					
TransCreeperGear	0.026665618	0.025520828	-0.069621679	0.043595219	0.092328478	1				
IntakeValvePerCyl	-0.422357449	-0.248509452	0.280344032	0.177960634	-0.131325993	-0.077679162	1			
ExhaustValvesPerCyl	-0.478438041	-0.339851831	0.335652854	0.15281925	-0.158326003	-0.17071584	0.911487816	1		
VarValveTiming	-0.06825603	0.005399291	0.124952779	0.090839722	-0.094772029	-0.235534402	0.240823978	0.279339052	1	
VarValveLift	-0.086571422	-0.059461008	0.096211275	0.130719422	-0.097809395	-0.101438565	0.154855875	0.175388998	0.055536033	1

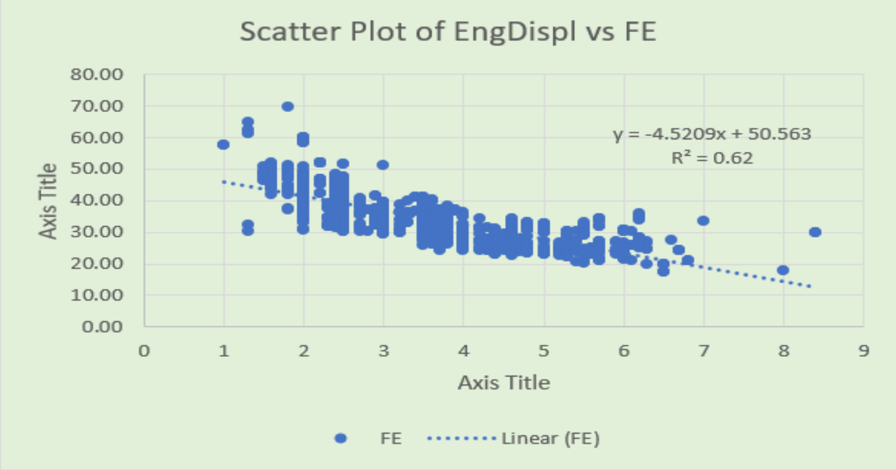
From the Correlation Test Engine Displacement has the highest correlation of -0.79 with Fuel Economy followed by NumCyl with a correlation of -0.74
We will select EngDispl as the input x variable for predicting Fuel Economy.
Some Exploratory variable transformation may be done if required for the input variable Engine Displacement



Correlation Plot Between FE and EngDispl and FE and NumCyl. FE vs EngDispl had a higher R2 value of 0.62 compared to FE and NumCyl with R2 value of 0.55

Exploratory data analysis was done on the input we selected (EngDispl) variable using to model the linear regression.

	EngDispl	Log10 (EngDispl)
Skewness (-1 to 1):	0.6455	-0.0627
Kurtosis (-2 to 2):	-0.2362	-0.6469



Scatter Plot of EngDispl vs FE showing a linear relationship with a R2 value of 0.62

Applying a Log10(x) does reduce skewness as can be seen but since without log10 it is within acceptable range so we decided not to build a model based on log10.

Before training and implementing the coefficient and intercept, cross validation was done. The dataset was divided into three parts randomly. Two parts were used for modelling and the third part was used for testing.

We used a random sampling method to divide the dataset in to 3 parts. Use rand() function.:-
We generated random number using the formula = rand()
Then arranged the observation (FE and EngDispl) randomly using the random generated numbers.
Total observations consist of 1107 values.
We divided the dataset into three subset and each subset consist of 369 values. We now have three samples: sample1, Sample2, Sample3 of 369 observations each.
This method is cross validation.
So, we created three models on three different datasets:

Training was done Sample1 and Sample2 and testing was done on Sample3.
Similarly, training was done on Sample2 and Sample3 and Testing was done on Sample1 and training was done on Sample3 and Sample1, testing is done on Sample2.

The average model accuracy and average test accuracy was calculated from the three models. We will observe if they are consistent and compute the Beta coefficients by taking average of the three models. Then we will test the final Accuracy by implementing the model on 2011 dataset.
We will use Data Analysis feature of Excel to bypass the co-efficient calculation formulas and compute the Regression Model directly.

For calculating the MAPE we will use the following formula in Excel:

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

Where A_t is the actual value and F_t is the forecast value.
n is the number of observations.

For Calculating R2 we will use the following formula in Excel:
R2 = SSR/SST
Where SST = (Y – mean Y) ^2
SSR = (Y predict – Mean y predict) ^2

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:

Model1:
The following values were obtained from model 1(training on sample 1 and Sample 2) and testing on Sample 3.

	Model1
Intercept	50.63
beta	-4.57
R2 Training	0.61
MAPE	9.46
R2 Test	0.69

Model1:
The following values were obtained from model 2(training on sample 2 and Sample 3) and testing on Sample 1.

	Model2
Intercept	51.14
beta	-4.62
R2 Training	0.63
MAPE	11.15
R2 Test	0.68

Model1:
The following values were obtained from model 3(training on sample 1 and Sample 3) and testing on Sample 2.

	Model3
Intercept	49.91
beta	-4.37
R2 Training	0.62
MAPE	10.69
R2 Test	0.51

The Averages of all the three models are taken for the final intercept and beta:

	SUMMARY			
	Model1	**Model2**	**Model3**	**Average**
Intercept	50.63	51.14	49.91	**50.56**
beta	-4.57	-4.62	-4.37	**-4.52**
R2 Training	0.61	0.63	0.62	**0.62**
MAPE	9.46	11.15	10.69	**10.43**
R2 Test	0.69	0.68	0.51	**0.63**

From the above table we can conclude that the **R2** and **MAPE** values are consistent with each other. Average **R2** and **MAPE** for all the models are **0.63** and **10.43** respectively which are acceptable.

We will take the average values of intercept and beta from the three models for predicting FE in FE2011 dataset.
The final intercept and beta values are **50.56** and **-4.52** respectively.

Then 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.

- The R squared and MAPE was found on the fe2011 test dataset as
  R squared = 0.522
  MAPE = 11.217

The test results were found within acceptable limits.

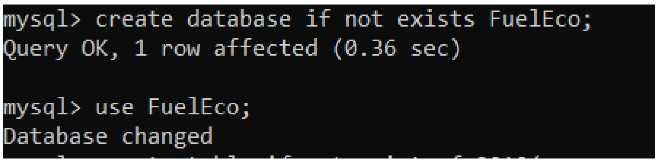
SQL Commands:

Below are the command used in SQL:-

Import the excel file into SQL using the following command:-

Create the database using the following command.
Create database if not exists FuelEco;

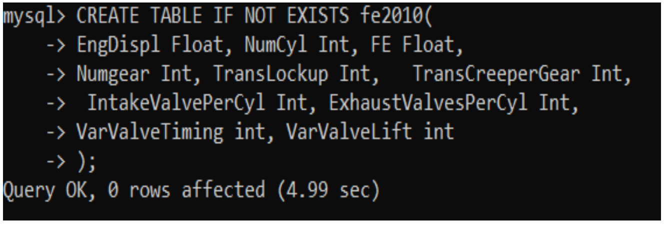
To use the created database, execute the following command: -
use FuelEco:



```
mysql> create database if not exists FuelEco;
Query OK, 1 row affected (0.36 sec)

mysql> use FuelEco;
Database changed
```

Create command is used to create the structure of the table .

CREATE TABLE IF NOT EXISTS fe2010(
EngDispl Float, NumCyl Int, FE Float,
Numgear Int, TransLockup Int,   TransCreeperGear Int,
 IntakeValvePerCyl Int, ExhaustValvesPerCyl Int,
VarValveTiming int, VarValveLift int
);

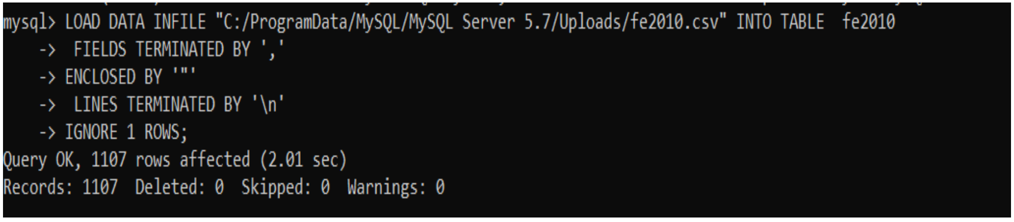


```
mysql> CREATE TABLE IF NOT EXISTS fe2010(
    -> EngDispl Float, NumCyl Int, FE Float,
    -> Numgear Int, TransLockup Int,   TransCreeperGear Int,
    ->  IntakeValvePerCyl Int, ExhaustValvesPerCyl Int,
    -> VarValveTiming int, VarValveLift int
    -> );
Query OK, 0 rows affected (4.99 sec)
```

The LOAD DATA INFILE statement allows you to read data from a text file and import the file's data into a database table.

The following statement imports data from the C:/ProgramData/MySQL/MySQL Server 5.7/Uploads/fe2010.csv file into the fe2010 table.

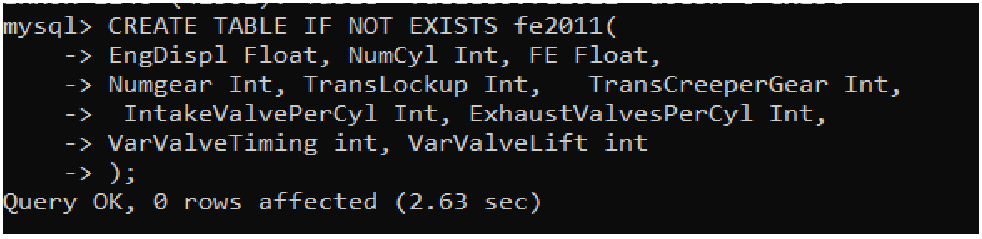
LOAD DATA INFILE "C:/ProgramData/MySQL/MySQL Server 5.7/Uploads/fe2010.csv" INTO TABLE fe2010
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;



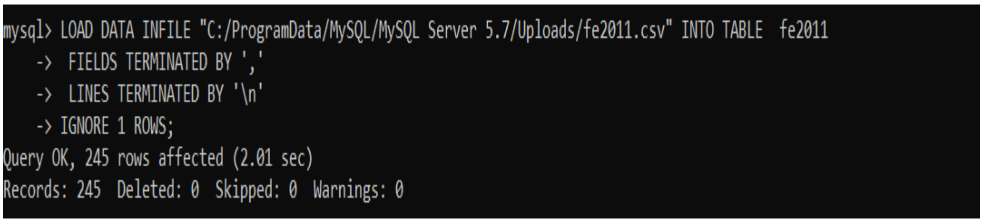
```
mysql> LOAD DATA INFILE "C:/ProgramData/MySQL/MySQL Server 5.7/Uploads/fe2010.csv" INTO TABLE  fe2010
    ->  FIELDS TERMINATED BY ','
    -> ENCLOSED BY '"'
    ->  LINES TERMINATED BY '\n'
    -> IGNORE 1 ROWS;
Query OK, 1107 rows affected (2.01 sec)
Records: 1107  Deleted: 0  Skipped: 0  Warnings: 0
```

Similarly, we will import fe2011 excel file by creating the schema and loading it with the following command: -

```
CREATE TABLE IF NOT EXISTS fe2011
EngDispl Float, NumCyl Int, FE Float,
Numgear Int, TransLockup Int, TransCreeperGear Int,
 IntakeValvePerCyl Int, ExhaustValvesPerCyl Int,
VarValveTiming int, VarValveLift int
);
LOAD DATA INFILE "C:/ProgramData/MySQL/MySQL Server 5.7/Uploads/fe2011.csv" INTO
TABLE fe2011
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

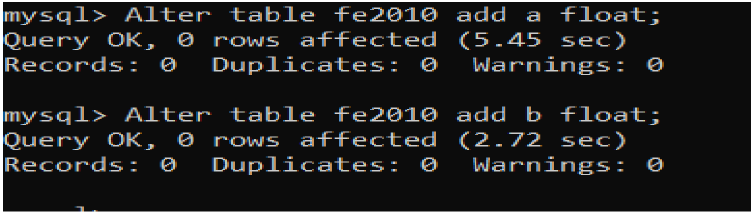


```
mysql> CREATE TABLE IF NOT EXISTS fe2011(
    -> EngDispl Float, NumCyl Int, FE Float,
    -> Numgear Int, TransLockup Int,   TransCreeperGear Int,
    ->  IntakeValvePerCyl Int, ExhaustValvesPerCyl Int,
    -> VarValveTiming int, VarValveLift int
    -> );
Query OK, 0 rows affected (2.63 sec)
```



```
mysql> LOAD DATA INFILE "C:/ProgramData/MySQL/MySQL Server 5.7/Uploads/fe2011.csv" INTO TABLE  fe2011
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 ROWS;
Query OK, 245 rows affected (2.01 sec)
Records: 245  Deleted: 0  Skipped: 0  Warnings: 0
```

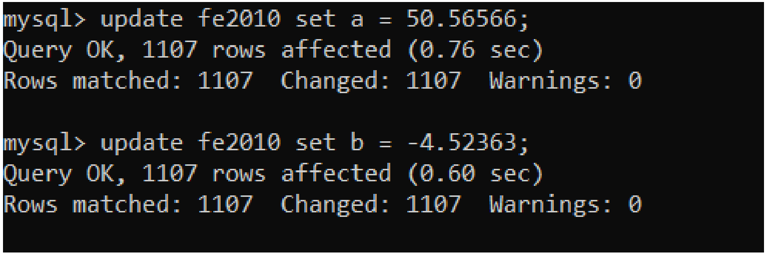
```
Alter table fe2010 add a float;
Alter table fe2010 add b float;
```



```
mysql> Alter table fe2010 add a float;
Query OK, 0 rows affected (5.45 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> Alter table fe2010 add b float;
Query OK, 0 rows affected (2.72 sec)
Records: 0  Duplicates: 0  Warnings: 0
```

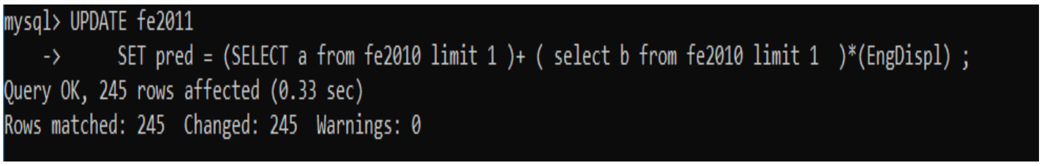
Update the column a and b with the values 50.56566 and -4.52363.

```
update fe2010 set a = 50.56566;
update fe2010 set b = -4.52363;
```



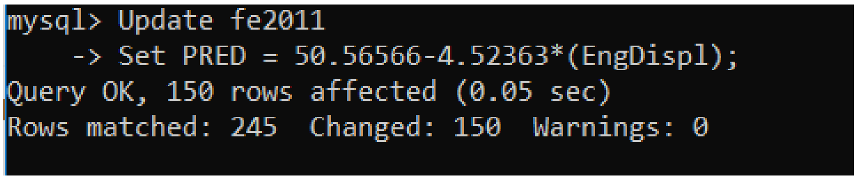
```
mysql> update fe2010 set a = 50.56566;
Query OK, 1107 rows affected (0.76 sec)
Rows matched: 1107  Changed: 1107  Warnings: 0

mysql> update fe2010 set b = -4.52363;
Query OK, 1107 rows affected (0.60 sec)
Rows matched: 1107  Changed: 1107  Warnings: 0
```

Update Pred in 2011 with a and b from 2010 data set



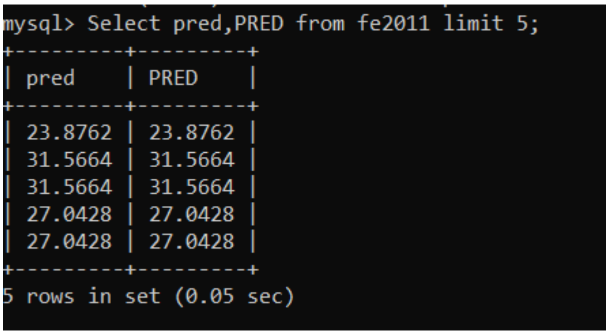
```
mysql> UPDATE fe2011
    ->      SET pred = (SELECT a from fe2010 limit 1 )+ ( select b from fe2010 limit 1  )*(EngDispl) ;
Query OK, 245 rows affected (0.33 sec)
Rows matched: 245  Changed: 245  Warnings: 0
```

We can directly update the value a and b in Pred using the formula a+bx



```
mysql> Update fe2011
    -> Set PRED = 50.56566-4.52363*(EngDispl);
Query OK, 150 rows affected (0.05 sec)
Rows matched: 245  Changed: 150  Warnings: 0
```

Compare both the column



```
mysql> Select pred,PRED from fe2011 limit 5;
+---------+---------+
| pred    | PRED    |
+---------+---------+
| 23.8762 | 23.8762 |
| 31.5664 | 31.5664 |
| 31.5664 | 31.5664 |
| 27.0428 | 27.0428 |
| 27.0428 | 27.0428 |
+---------+---------+
5 rows in set (0.05 sec)
```

Files Attached:

The following supporting files are also attached

- Fe2010 excel workbook with all calculated sheets
- Fe2011 and F2010 datasets