PROJECT 1.1

Simple Regression Analysis on Fuel Economy Data

Part 1 (R & Excel Analytics)

1. Introduction

You are provided with two datasets, “FE2010.csv” and “FE2011.csv”. You are required to work on “FE2010.csv” only for any kind of experiment. The datasets contain different estimates of fuel economy for passenger cars and trucks. 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. 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 were not in the 2010 data set.

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.

2. Objective

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

3. Prerequisites

N/A

4. Associated Data Files

https://drive.google.com/file/d/0B3nTkXniPMACMUc0RVJXODhlUDA/view?usp=sharing

5. Problem Statement

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).
2. Fit a Simple Linear Regression Model using the selected input variable. Use the formulas discussed in the class to calculate the coefficients.
3. Observe the relationship between the Input variable and FE and analyse 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.

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.

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.

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

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

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.

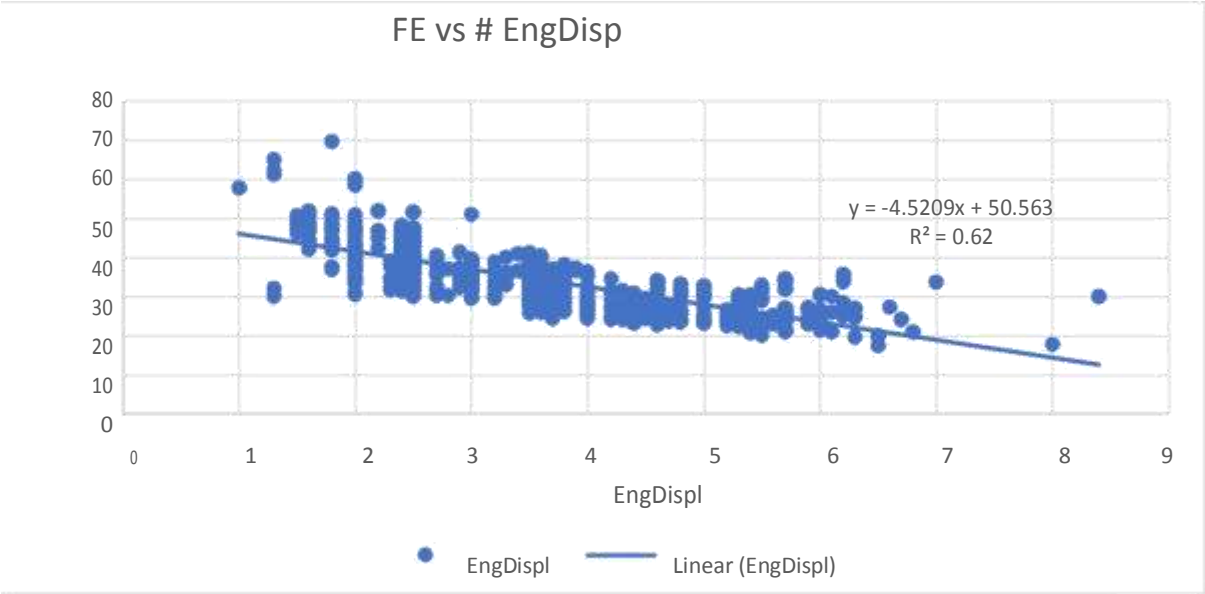
8. You should be able to repeat all the points asked under “Use Excel” using Data Analysis tool. You may need to do the random sampling separately here as well.

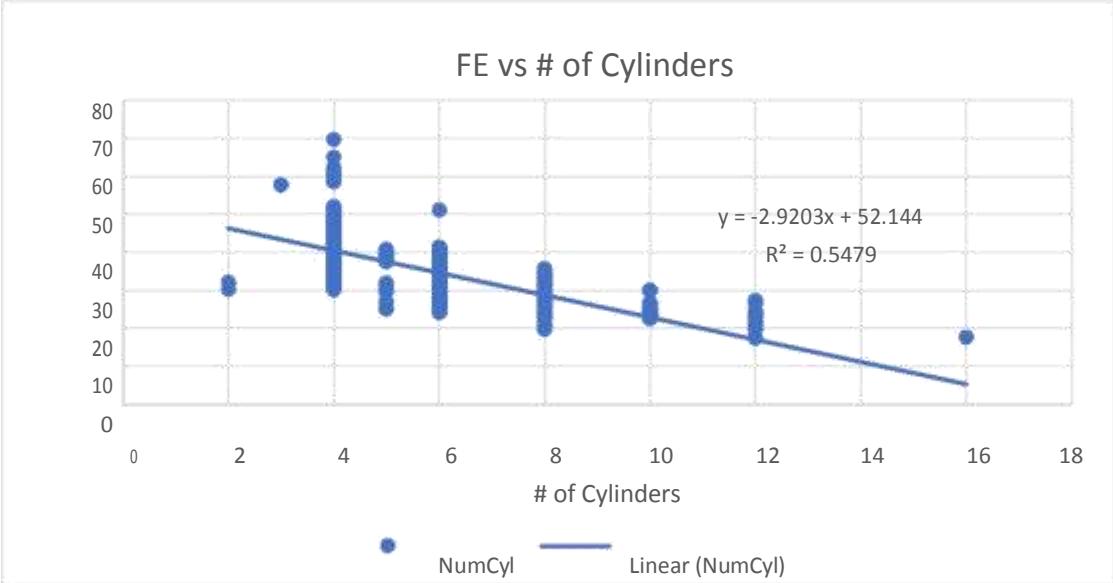
APPROACH

FE 2010 data consists of many input variables and a response variable given as Fuel efficiency collected based on the various vehicle input data. For all these variable correlation coefficient is found in excel and is found as given below

EngDispl	-0.79
Numcyl	-0.74
NumGears	-0.21
TransLockup	-0.27
TransCreeperGear	-0.07
IntakeValvePerCyl	0.28
ExhaustValvesPerCyl	0.34
VarValveTiming	0.12
VarValveLift	0.10

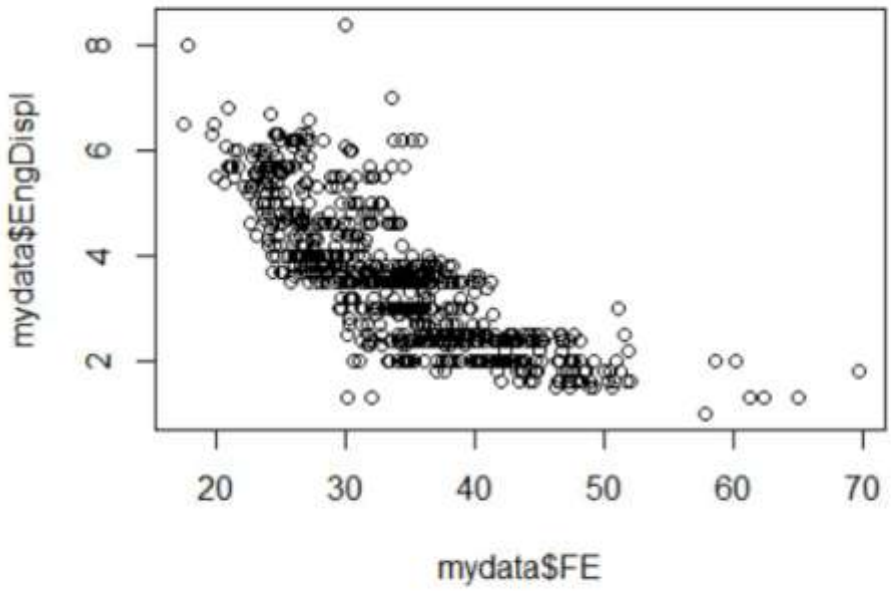
Based on the above findings EngDispl and Numcyl are having very good correlation and as per the problem statement we have plotted these variable with suitable charting in Excel


FE vs # EngDisp
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
y = -4.5209x + 50.563
R² = 0.62
EngDispl
EngDispl
Linear (EngDispl)


FE vs # of Cylinders
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
18
y = -2.9203x + 52.144
R² = 0.5479
# of Cylinders
NumCyl
Linear (NumCyl)

Based on this graph these two input variables appear to be linear. Further the whole data is analysed in R to find the Linear relationship

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.




```
cor(mydata$FE,mydata$EngDispl)
```

```
## [1] -0.7873938
```

```r
cor(mydata$FE,mydata$VarValveLift)
```

```
## [1] 0.09621127
```

```r
cor(mydata$FE,mydata$VarValveTiming)
```

```
## [1] 0.1249528
```

```r
cor(mydata$FE,mydata$ExhaustValvesPerCyl)
```

```
## [1] 0.3356529
```

```r
cor(mydata$FE,mydata$IntakeValvePerCyl)
```

```
## [1] 0.280344
```

```r
cor(mydata$FE,mydata$TransCreeperGear)
```

```
## [1] -0.06962168
```

```r
cor(mydata$FE,mydata$TransLockup)
```

```
## [1] -0.2719389
```

```r
cor(mydata$FE,mydata$NumGears)
```

```
## [1] -0.2112849
```

```r
cor(mydata$FE,mydata$NumCyl)
```

```
## [1] -0.740218
```

```r
mod=lm(mydata$FE~mydata$EngDispl)
mod
```

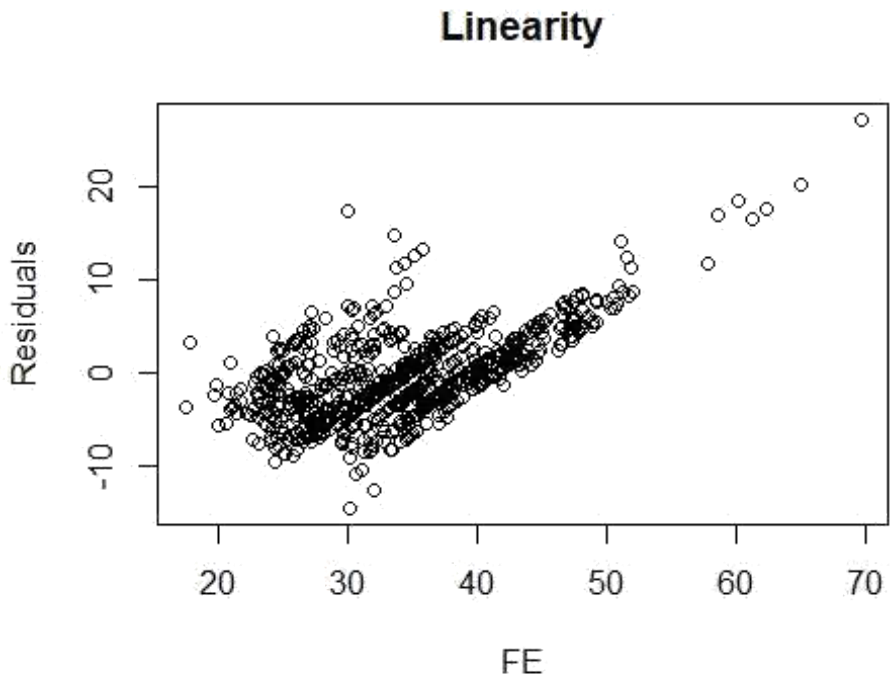
```
##
## Call:
## lm(formula = mydata$FE ~ mydata$EngDispl)
```

```
##
## Coefficients:
##     (Intercept)  mydata$EngDispl
##           50.563           -4.521
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = mydata$FE ~ mydata$EngDispl)
##
## Residuals:
##     Min       1Q  Median      3Q      Max
## -14.486   -3.192  -0.365   2.671  27.215
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      50.5632     0.3985   126.89  <2e-16 ***
## mydata$EngDispl  -4.5209     0.1065   -42.46  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.624 on 1105 degrees of freedom
## Multiple R-squared:   0.62,  Adjusted R-squared:  0.6196
## F-statistic:  1803 on 1 and 1105 DF,  p-value: < 2.2e-16
```

```
#Assumption1 Linearity
plot(mydata$FE,mydata$error,xlab="FE",ylab="Residuals",main="Linearity")
```




```
fit<-lm(FE~EngDispl+NumCyl+NumGears+TransLockup+TransCreeperGear+IntakeValvePerCyl+ExhaustValvesPerCyl+VarValveTimi
ng+VarValveLift,data=FE2010)
fit
```

```
##
## Call:
## lm(formula = FE ~ EngDispl + NumCyl + NumGears + TransLockup +
##     TransCreeperGear + IntakeValvePerCyl + ExhaustValvesPerCyl +
##     VarValveTiming + VarValveLift, data = FE2010)
##
## Coefficients:
##         (Intercept)             EngDispl               NumCyl
```

```
##             54.3472              -3.8610              -0.4888
##            NumGears          TransLockup     TransCreeperGear
##             -0.1725              -1.4450              -0.9138
##   IntakeValvePerCyl  ExhaustValvesPerCyl       VarValveTiming
##             -0.3737              -1.1105               1.6870
##        VarValveLift
##              0.6235
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = FE ~ EngDispl + NumCyl + NumGears + TransLockup +
##     TransCreeperGear + IntakeValvePerCyl + ExhaustValvesPerCyl +
##     VarValveTiming + VarValveLift, data = FE2010)
##
## Residuals:
##      Min        1Q   Median      3Q       Max
## -17.1153   -2.7142  -0.3535   2.4191  25.6521
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          54.3472     1.0973   49.530  < 2e-16 ***
## EngDispl             -3.8610     0.2805  -13.765  < 2e-16 ***
## NumCyl               -0.4888     0.1845   -2.649  0.00819 **
## NumGears             -0.1725     0.1065   -1.620  0.10555
## TransLockup          -1.4450     0.3000   -4.817 1.66e-06 ***
## TransCreeperGear     -0.9138     0.6681   -1.368  0.17167
## IntakeValvePerCyl    -0.3737     0.9892   -0.378  0.70566
## ExhaustValvesPerCyl  -1.1105     0.9598   -1.157  0.24752
## VarValveTiming        1.6870     0.3796    4.444 9.71e-06 ***
## VarValveLift          0.6235     0.3719    1.676  0.09393 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.489 on 1097 degrees of freedom
## Multiple R-squared:  0.6445, Adjusted R-squared:  0.6415
## F-statistic: 220.9 on 9 and 1097 DF,  p-value: < 2.2e-16
```

```
vif(fit)
```

```
##            EngDispl              NumCyl            NumGears
##            7.363137            6.750388            1.214238
##         TransLockup    TransCreeperGear   IntakeValvePerCyl
##            1.075253            1.137623            6.693985
## ExhaustValvesPerCyl      VarValveTiming        VarValveLift
##            7.073284            1.153276            1.057688
```

```
vif(fit)>5
```

```
##            EngDispl              NumCyl            NumGears
##                TRUE                TRUE               FALSE
##         TransLockup    TransCreeperGear   IntakeValvePerCyl
##               FALSE               FALSE                TRUE
## ExhaustValvesPerCyl      VarValveTiming        VarValveLift
##                TRUE               FALSE               FALSE
```

Based on the above Vif(fit)>5 the yellow highlighter like EngDispl, Numcyl etc are better linearity and fit for further analysis in Excel

Excel Analysis

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.
One example (small part of excel data)for the data Numcyl is given below to show how we calculate MAPE values

For the entire 2010 data we have calculated Beta coefficient and the intercept from the scatter plot to predict the Fuel efficiency as given below in predicted Y

Beta coefficient calculation for 2010

	FE	Engdisp	Numcyl	NumGears	TransLockup	TransCreeperGear	IntakeValvePerCyl	ExhaustValvesPerCyl	VarVa
std dev	7.498033	1.305905	1.900575	1.396624	0.466603	0.215506	0.353046	0.374035	0.381
Variance (VARA(B3:B1109))	56.22049	1.705388	3.60892	1.948796	0.217522	0.046401	0.124529	0.139776	0.145
Covariance (COVAR(A3:A1109,B3:B1109)		-7.70297	-10.539	-2.21056	-0.95055	-0.1124	0.741443	0.940497	0.357
Beta coeff by cal Formula=(covariance/variance)		-4.51685	-2.92026	-1.13432	-4.36989	-2.42232	5.953976	6.728613	2.453
Beta coeff by graph		(-)4.5209	(-)2.9203						

Excel Instructions:

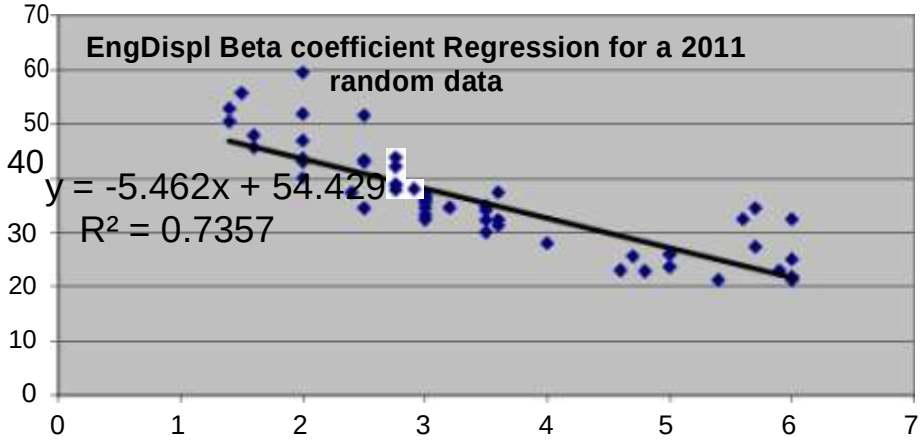
There are four ways that you can calculate a Beta using Excel. The first is to use the "=slope" formula. In this formula, the X variable series is the return on the market and the Y variable series if the return on FE.
This gives:
-5.0180

A second alternative is to calculate the Beta directly as the covariance between the two return series, divided by the variance of market returns. Using the summary statistics at the bottom of the page, this gives:
-5.4164

A third method is to use the full regression procedure in Excel. Go to "Tools" - "Data Analysis" - "Regression" and click OK. Again, enter the

market return series as the X variable and the HD return series as the Y variable. When you click OK, you should get full regression output similar to that shown in the worksheet labelled " FERegr Output".

Finally, you can create a regression Beta in Excel using the chart functions. Remember that Beta is just the slope in a regression where EngDispl are on the X axis and FUEL EFICENCY (FE are on the Y axis.
To create this type of graph, highlight the two columns of data, with market returns on the left. After highlighting the return series, click on the chart wizard icon (or choose "Insert" - "Chart"). Under chart types, select "X-Y scatterplot" and click Next. Click Next twice more to get to step 4 of 4. In step 4, select "as new sheet" and click finish. When the new chart comes up, select the "Chart" tab at the top of the page, and then select "Add Trendline". Select the "Options" tab, and click the buttons for "display equation" and "display R square", then click OK. This should add both a regression line and a regression equation to your chart. The results should look similar to those shown in the worksheet labelled "FE Chart".



Based on the calculation of Beta coefficient as given above and with the intercept from the graph,
One example (small part of FE 2011 excel data)for the data Numcyl is given below to show how we calculate MAPE values

X	Y	Predicted Y				
NumCyl	FE	Predicted FE	Error		Dis Y and their mean	Square F
12	22.9258	17.1004	5.8254	33.93529	-11.80486408	139.3548
8	26.7678	28.7816	-2.0138	4.05539	-7.962864082	63.4072
8	24.301	28.7816	-4.4806	20.07578	-10.42966408	108.7779
10	24.3325	22.941	1.3915	1.936272	-10.39816408	108.1218
10	23.0667	22.941	0.1257	0.0158	-11.66396408	136.0481
6	32.8579	34.6222	-1.7643	3.112754	-1.872764082	3.507245
4	52.2	40.4628	11.7372	137.7619	17.46933592	305.1777
4	55.6446	40.4628	15.1818	230.4871	20.91393592	437.3927
8	26	28.7816	-2.7816	7.737299	-8.730664082	76.2245
12	25	17.1004	7.8996	62.40368	-9.730664082	94.68582
8	26.8	28.7816	-1.9816	3.926739	-7.930664082	62.89543

Abs v of error	Error 2	ABS values of error/Actual value
5.8254	33.93529	0.254098
2.0138	4.05539	0.075232
4.4806	20.07578	0.184379
1.3915	1.936272	0.057187
0.1257	0.0158	0.005449
1.7643	3.112754	0.053695
11.7372	137.7619	0.224851
15.1818	230.4871	0.272835
2.7816	7.737299	0.106985
7.8996	62.40368	0.315984
1.9816	3.926739	0.073940

No of rows is calculate in n = COUNT(D3:D247) MSE is calculated (L248/C250)(Error square/n)
RMSE is the square root of MSE (SQRT(C253)
MAPE is calculated ((M248/C250)*100) M248 = sum of (ABS values of error/Actual value) and C 250 is the no row count ie 245
Given below is calculated from the prediction of FE using the input variable Numcyl beta coefficient and intercept. Similarly we have calculated the same with EngDispl variable

	n	245
MAD		4.470793061

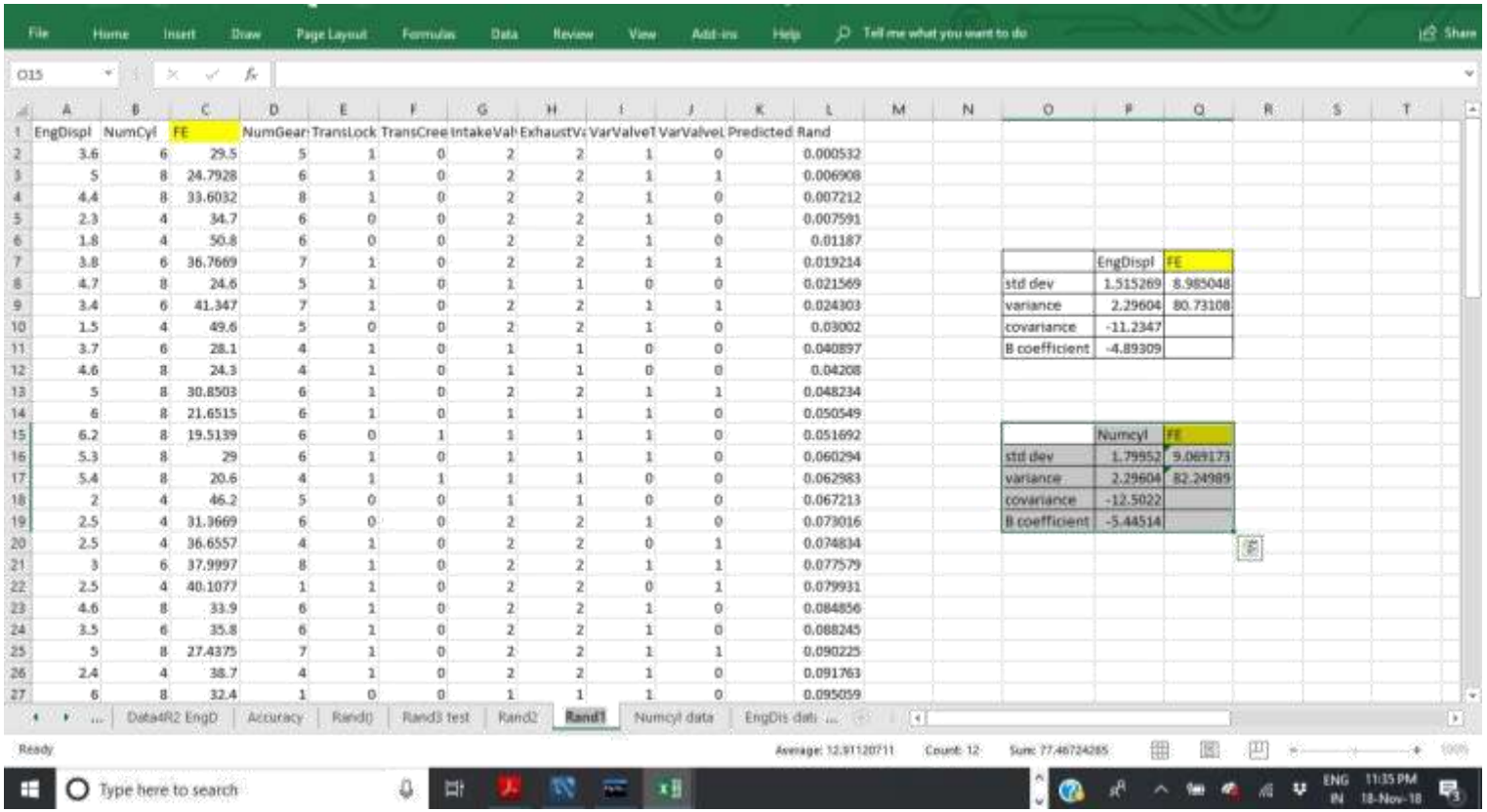
MSE		34.67015725	
RMSE		5.888136993	
MAPE		13.37649118	Mean absolute percentage error

1 - (SUM(E3:E247)) sum of all values of Error square) /sum of Square of Dis Y and their mean would provide the R square by calculation.

0.589277463	R 2 by calculation	1-(E248/G248) by graph 0.5479

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

FE 2011 data is divided into 3 sets using random function in Excel like Rand() , Rand1,Rand2 and Rand3 is used for testing based on these two data, Hence Fe2011 data is divided into 3 equal parts approximately 82 values in each



EngDispl	NumCyl	FE	NumGear	TransLock	TransCree	IntakeVal	ExhaustV	VarValveT	VarValveL	Predicted	Rand
3.6	6	29.5	5	1	0	2	2	1	0		0.000532
5	8	24.7928	6	1	0	2	2	1	1		0.006908
4.4	8	33.6032	8	1	0	2	2	1	0		0.007212
2.3	4	34.7	6	0	0	2	2	1	0		0.007591
1.8	4	50.8	6	0	0	2	2	1	0		0.01187
3.8	6	36.7669	7	1	0	2	2	1	1		0.019214
4.7	8	24.6	5	1	0	1	1	0	0		0.021569
3.4	6	41.347	7	1	0	2	2	1	1		0.024303
1.5	4	49.6	5	0	0	2	2	1	0		0.03002
3.7	6	28.1	4	1	0	1	1	0	0		0.040897
4.6	8	24.3	4	1	0	1	1	0	0		0.04208
5	8	30.8503	6	1	0	2	2	1	1		0.048234
6	8	21.6515	6	1	0	1	1	1	0		0.050549
6.2	8	19.5139	6	0	1	1	1	1	0		0.051692
5.3	8	29	6	1	0	1	1	1	0		0.060294
5.4	8	20.6	4	1	1	1	1	0	0		0.062983
2	4	46.2	5	0	0	1	1	0	0		0.067213
2.5	4	31.3669	6	0	0	2	2	1	0		0.073016
2.5	4	36.6557	4	1	0	2	2	0	1		0.074834
3	6	37.9997	8	1	0	2	2	1	1		0.077579
2.5	4	40.1077	1	1	0	2	2	0	1		0.079931
4.6	8	33.9	6	1	0	2	2	1	0		0.084856
3.5	6	35.8	6	1	0	2	2	1	0		0.088245
5	8	27.4375	7	1	0	2	2	1	1		0.090225
2.4	4	38.7	4	1	0	2	2	1	0		0.091763
6	8	32.4	1	0	0	1	1	1	0		0.095059

	EngDispl	FE
std dev	1.515269	8.985048
variance	2.29604	80.73108
covariance	-11.2347	
B coefficient	-4.89309	

	Numcyl	FE
std dev	1.79952	9.069173
variance	2.29604	82.24989
covariance	-12.5022	
B coefficient	-5.44514	

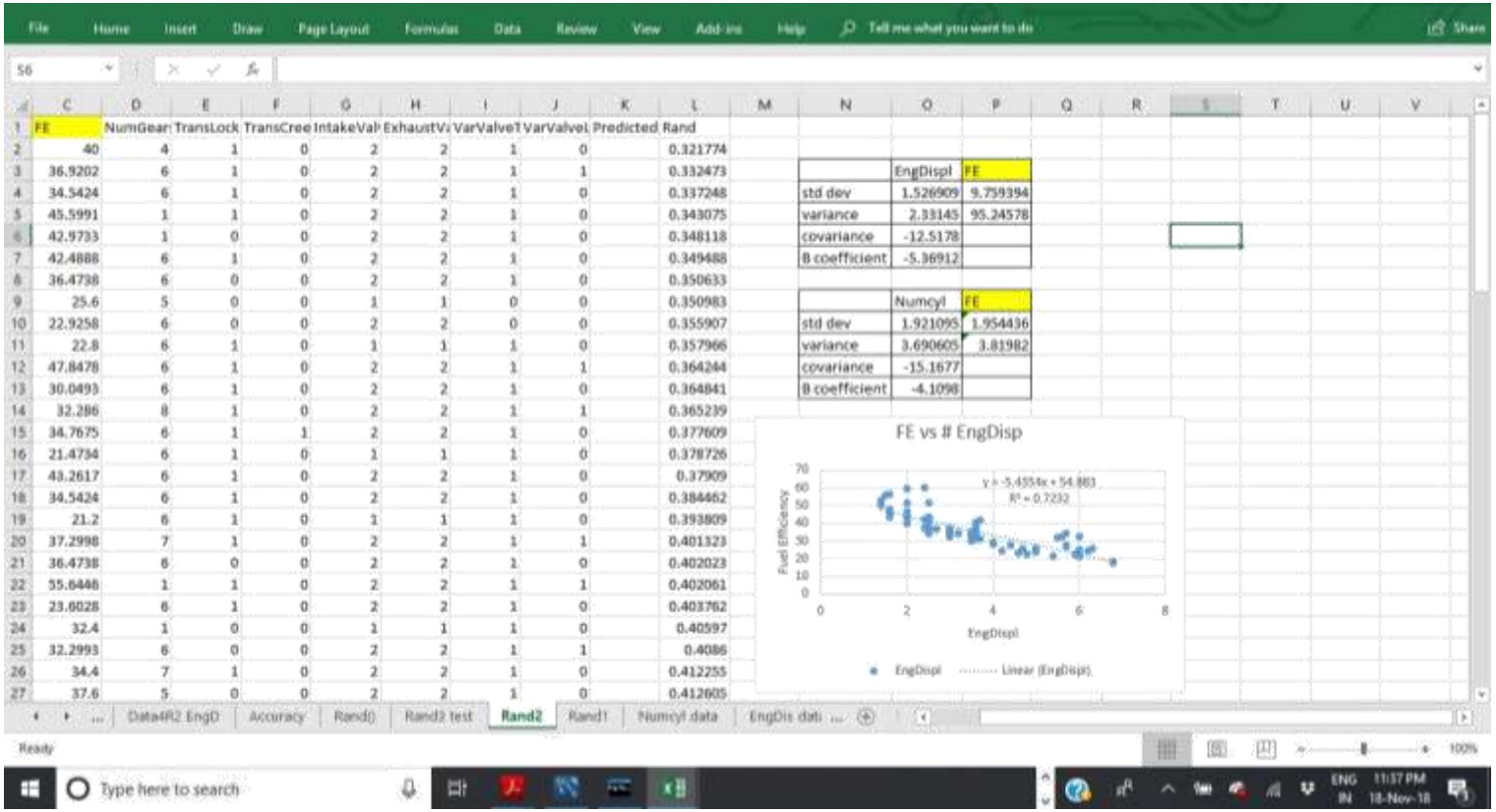
Data4R2 EngD | Accuracy | Rand0 | Rand3 test | Rand2 | Rand1 | Numcyl data | EngDis data ...

Ready

Average: 12.91120711 Count: 12 Sum: 77.46724265

Type here to search

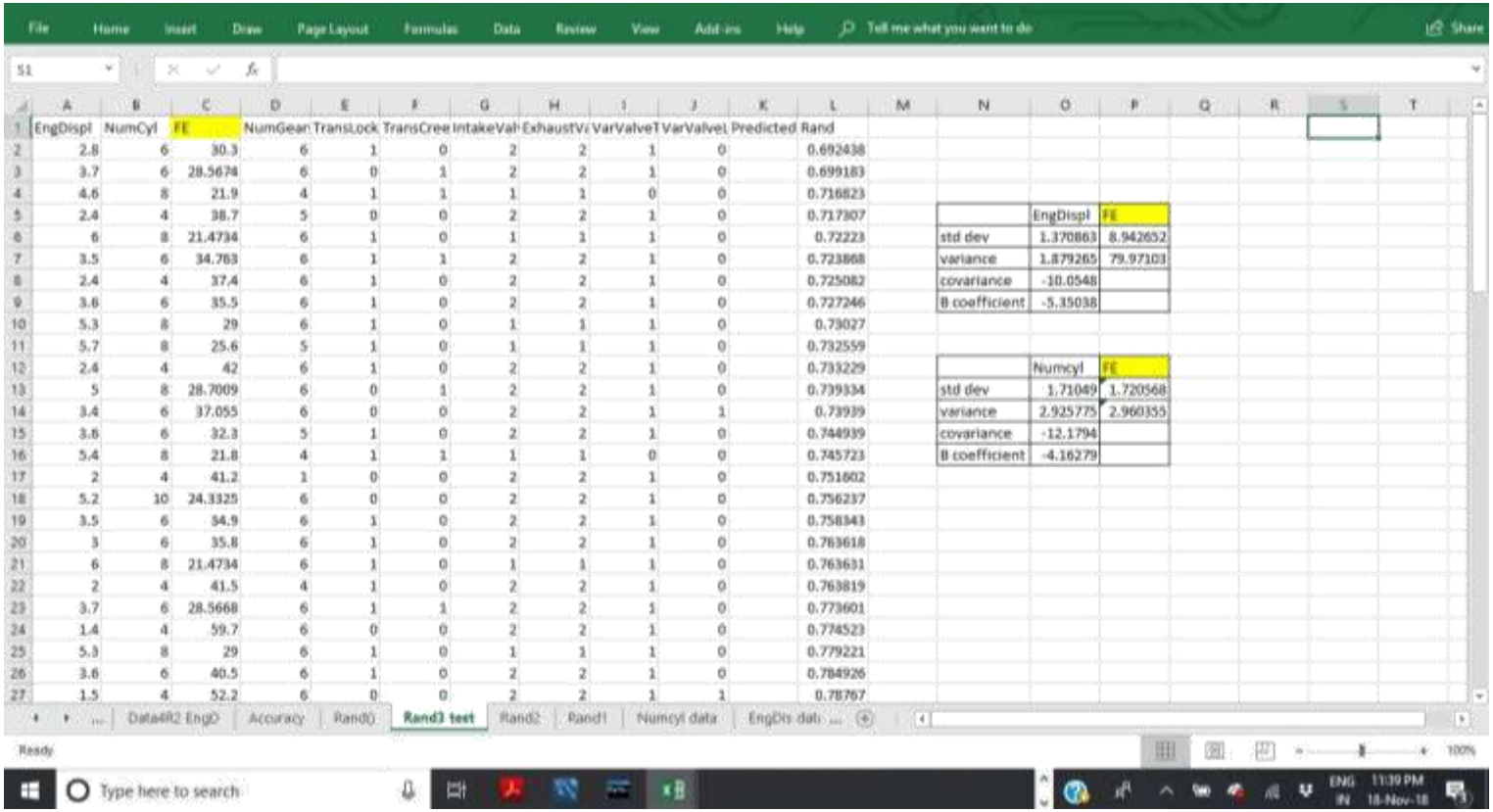
ENG IN 11:35 PM 18-Nov-18



FE	NumGear	TransLock	TransCree	IntakeVal	ExhaustV	VarValveT	VarValveL	Predicted	Rand
40	4	1	0	2	2	1	0		0.321774
36.9202	6	1	0	2	2	1	1		0.332473
34.5424	6	1	0	2	2	1	0		0.337248
45.5991	1	1	0	2	2	1	0		0.343075
42.9733	1	0	0	2	2	1	0		0.348118
42.4888	6	1	0	2	2	1	0		0.349488
36.4738	6	0	0	2	2	1	0		0.350633
25.6	5	0	0	1	1	0	0		0.350983
22.9258	6	0	0	2	2	0	0		0.355907
22.8	6	1	0	1	1	1	0		0.357966
47.8478	6	1	0	2	2	1	1		0.364244
30.0493	6	1	0	2	2	1	0		0.364841
32.286	8	1	0	2	2	1	1		0.365239
34.7675	6	1	1	2	2	1	0		0.377609
21.4734	6	1	0	1	1	1	0		0.378726
48.2617	6	1	0	2	2	1	0		0.37909
34.5424	6	1	0	2	2	1	0		0.384462
21.2	6	1	0	1	1	1	0		0.393809
37.2998	7	1	0	2	2	1	1		0.401323
36.4738	6	0	0	2	2	1	0		0.402023
55.6448	1	1	0	2	2	1	1		0.402061
23.6028	6	1	0	2	2	1	0		0.403762
32.4	1	0	0	1	1	1	0		0.40597
32.2993	6	0	0	2	2	1	1		0.4086
34.4	7	1	0	2	2	1	0		0.412255
37.6	5	0	0	2	2	1	0		0.412605

	EngDispl	FE
std dev	1.526909	9.759394
variance	2.33145	95.24578
covariance	-12.5178	
B coefficient	-5.36912	

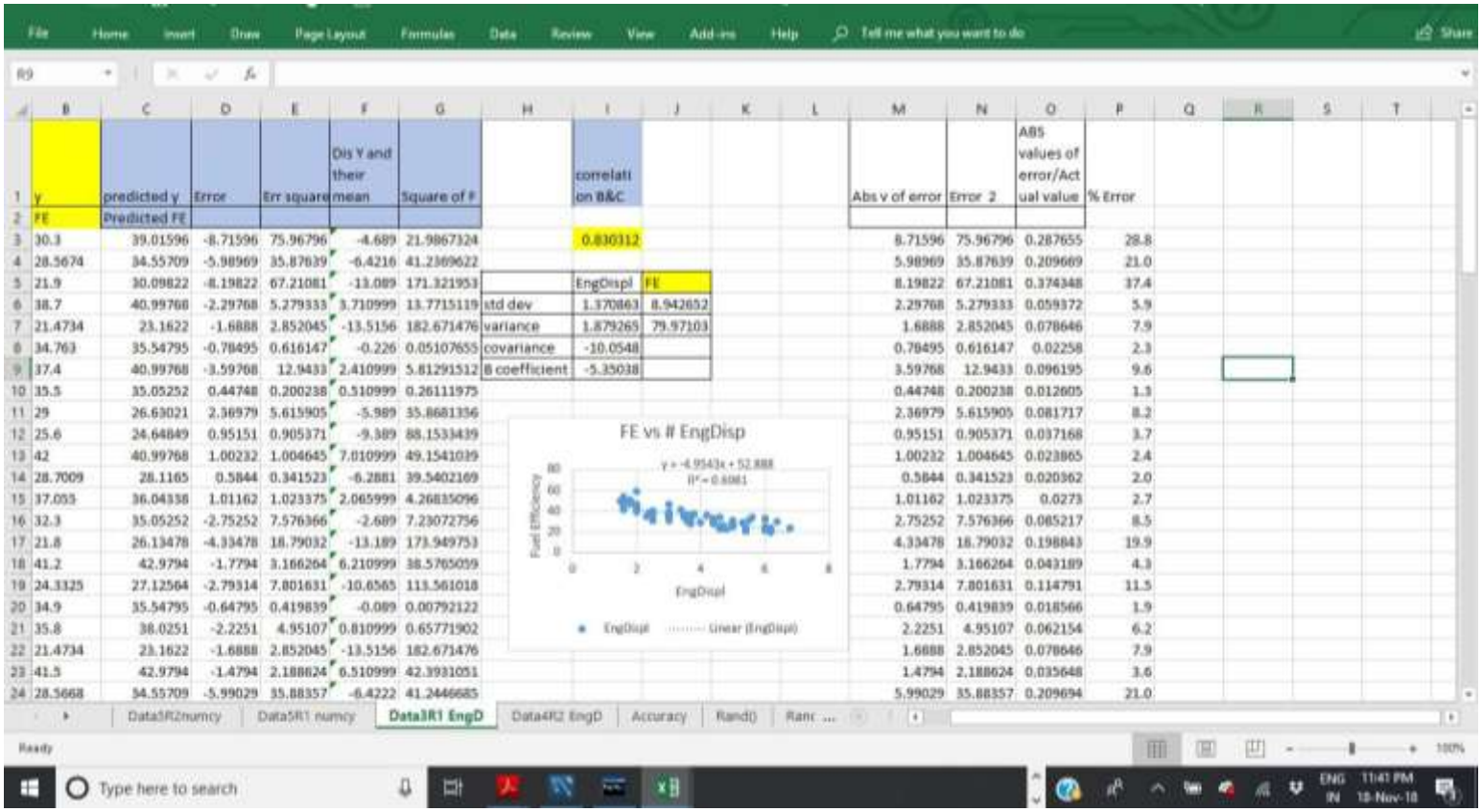
	Numcyl	FE
std dev	1.921095	1.954436
variance	3.690605	3.81982
covariance	-15.1677	
B coefficient	-4.1098	



EngDispl	NumCyl	FE	NumGear	TransLock	TransCree	IntakeVal	ExhaustVa	VarValveT	VarValveL	Predicted	Rand
2.8	6	30.3	6	1	0	2	2	1	0		0.692438
3.7	6	28.5674	6	0	1	2	2	1	0		0.699183
4.6	8	21.9	4	1	1	1	1	0	0		0.716823
2.4	4	38.7	5	0	0	2	2	1	0		0.717307
6	8	21.4734	6	1	0	1	1	1	0		0.72223
3.5	6	34.763	6	1	1	2	2	1	0		0.723868
2.4	4	37.4	6	1	0	2	2	1	0		0.725082
3.6	6	35.5	6	1	0	2	2	1	0		0.727246
5.3	8	29	6	1	0	1	1	1	0		0.73027
5.7	8	25.6	5	1	0	1	1	1	0		0.732559
2.4	4	42	6	1	0	2	2	1	0		0.733229
5	8	28.7009	6	0	1	2	2	1	0		0.739334
3.4	6	37.055	6	0	0	2	2	1	1		0.73939
3.6	6	32.3	5	1	0	2	2	1	0		0.744939
5.4	8	21.8	4	1	1	1	1	0	0		0.745723
2	4	41.2	1	0	0	2	2	1	0		0.751602
5.2	10	24.3325	6	0	0	2	2	1	0		0.756237
3.5	6	34.9	6	1	0	2	2	1	0		0.758343
3	6	35.8	6	1	0	2	2	1	0		0.763618
6	8	21.4734	6	1	0	1	1	1	0		0.763631
2	4	41.5	4	1	0	2	2	1	0		0.763819
3.7	6	28.5668	6	1	1	2	2	1	0		0.773601
1.4	4	59.7	6	0	0	2	2	1	0		0.774523
5.3	8	29	6	1	0	1	1	1	0		0.779221
3.6	6	40.5	6	1	0	2	2	1	0		0.784926
1.5	4	52.2	6	0	0	2	2	1	1		0.78767

	EngDispl	FE
std dev	1.370863	8.942652
variance	1.879265	79.97103
covariance	-10.0548	
B coefficient	-5.35038	

	Numcyl	FE
std dev	1.71049	1.720568
variance	2.925775	2.960355
covariance	-12.1794	
B coefficient	-4.16279	



	y	predicted y	Error	Err square	Dis Y and their mean	Square of F		correlation B&C		Abs v of error	Error_2	ABS values of error/Actual value	% Error
2	FE	Predicted FE											
3	30.3	39.01596	-8.71596	75.96796	-4.689	21.9867324		0.830312		8.71596	75.96796	0.287655	28.8
4	28.5674	34.55709	-5.98969	35.87639	-6.4216	41.2369622				5.98969	35.87639	0.209669	21.0
5	21.9	30.09822	-8.19822	67.21081	-13.089	171.321953		EngDispl	FE	8.19822	67.21081	0.374348	37.4
6	38.7	40.99768	-2.29768	5.279333	3.710999	13.7715119	std dev	1.370863	8.942652	2.29768	5.279333	0.059372	5.9
7	21.4734	23.1622	-1.6888	2.852045	-13.5156	182.671476	variance	1.879265	79.97103	1.6888	2.852045	0.078646	7.9
8	34.763	35.54795	-0.78495	0.616147	-0.226	0.05107655	covariance	-10.0548		0.78495	0.616147	0.02258	2.3
9	37.4	40.99768	-3.59768	12.9433	2.410999	5.81291512	B coefficient	-5.35038		3.59768	12.9433	0.096195	9.6
10	35.5	35.05252	0.44748	0.200238	0.510999	0.26111975				0.44748	0.200238	0.012605	1.3
11	29	26.63021	2.36979	5.615905	-5.989	35.8681356				2.36979	5.615905	0.081717	8.2
12	25.6	24.64849	0.95151	0.905371	-9.389	88.1533439				0.95151	0.905371	0.037168	3.7
13	42	40.99768	1.00232	1.004645	7.010999	49.1541039				1.00232	1.004645	0.023865	2.4
14	28.7009	28.1165	0.5844	0.341523	-6.2881	39.5402169				0.5844	0.341523	0.020362	2.0
15	37.053	36.04338	1.01162	1.023375	2.065999	4.26835096				1.01162	1.023375	0.0273	2.7
16	32.3	35.05252	-2.75252	7.576366	-2.689	7.23072756				2.75252	7.576366	0.085217	8.5
17	21.8	26.13478	-4.33478	18.79032	-13.189	173.949753				4.33478	18.79032	0.198843	19.9
18	41.2	42.9794	-1.7794	3.166264	6.210999	38.5765059				1.7794	3.166264	0.043189	4.3
19	24.3325	27.12564	-2.79314	7.801631	-10.6565	113.561018				2.79314	7.801631	0.114791	11.5
20	34.9	35.54795	-0.64795	0.419839	-0.089	0.00792122				0.64795	0.419839	0.018566	1.9
21	35.8	38.0251	-2.2251	4.95107	0.810999	0.65771902				2.2251	4.95107	0.062154	6.2
22	21.4734	23.1622	-1.6888	2.852045	-13.5156	182.671476				1.6888	2.852045	0.078646	7.9
23	41.5	42.9794	-1.4794	2.188624	6.510999	42.3931051				1.4794	2.188624	0.035648	3.6
24	28.5668	34.55709	-5.99029	35.88357	-6.4222	41.2446685				5.99029	35.88357	0.209694	21.0

	n	82
MAD		3.870526098
MSE		25.37700219

Prediction of FE in Rand3 test case done and calculated the following

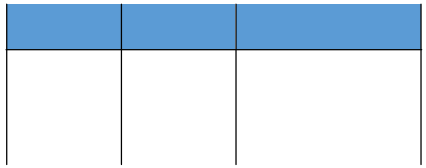
Average % Error	11.5

Average Test Accuracy	88.5

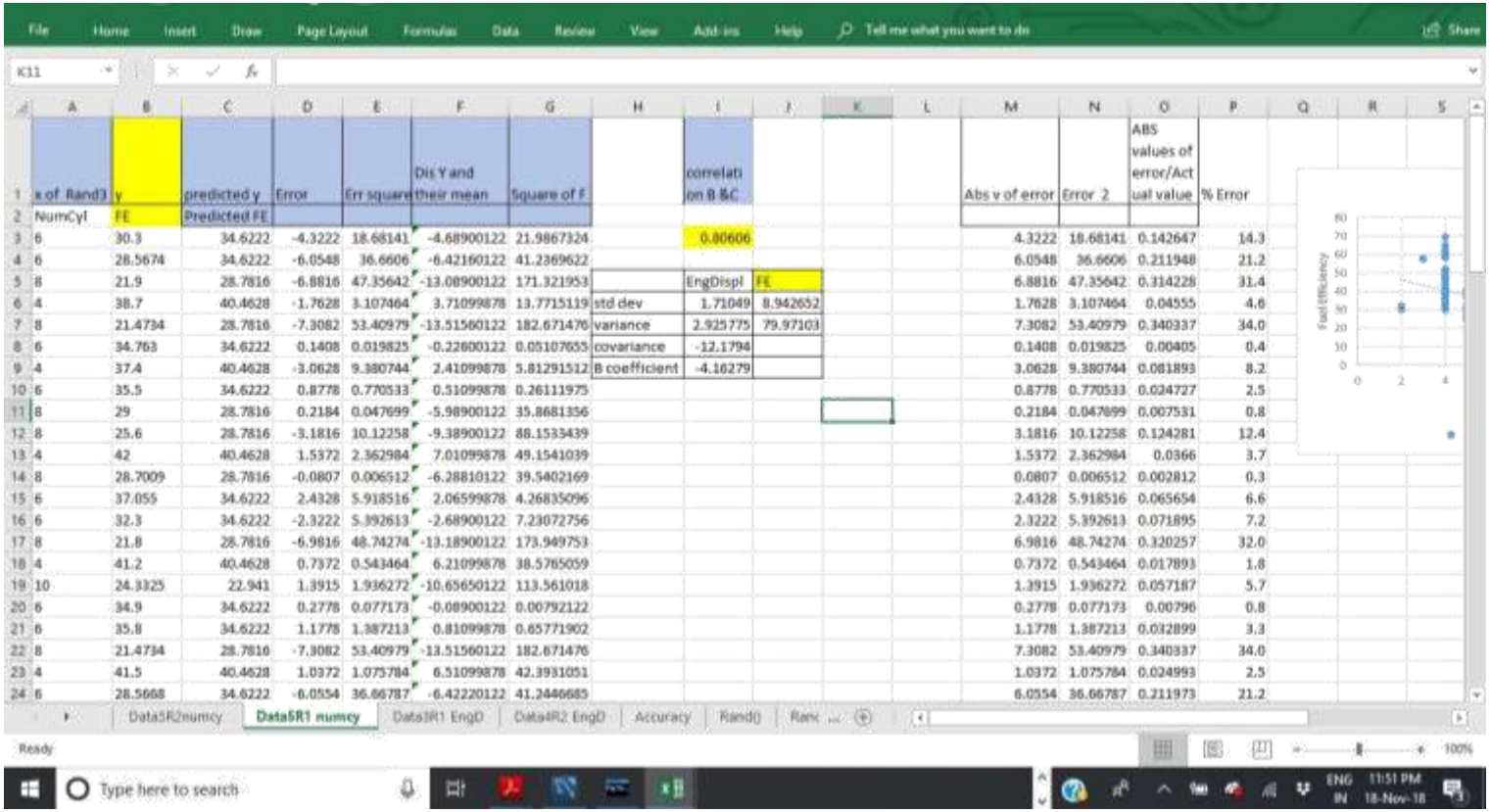
RMSE	5.03755915
MAPE	11.50998577

Mean absolute percentage error

R square by calculation	0.678755
As per plot	0.6981



In the similar way we have predicted using Rand2 Beta coefficient values of EngDispl values and Numcyl values



x of Rand3	y	predicted y	Error	Err square	Dis Y and their mean	Square of F		correlation B &C			Abs v of error	Error 2	ABS values of error/Actual value	% Error
NumCyl	FE	Predicted FE												
6	30.3	34.6222	-4.3222	18.68141	-4.68900122	21.9867324		0.80606			4.3222	18.68141	0.142647	14.3
6	28.5674	34.6222	-6.0548	36.6606	-6.42160122	41.2369622					6.0548	36.6606	0.211948	21.2
8	21.9	28.7816	-6.8816	47.35642	-13.08900122	171.321953		EngDispl	FE		6.8816	47.35642	0.314228	31.4
4	38.7	40.4628	-1.7628	3.107464	3.71099878	13.7715119	std dev	1.71049	8.942652		1.7628	3.107464	0.04555	4.6
8	21.4734	28.7816	-7.3082	53.40979	-13.51560122	182.671476	variance	2.925775	79.97103		7.3082	53.40979	0.340337	34.0
6	34.763	34.6222	0.1408	0.019825	-0.22600122	0.05107655	covariance	-12.1794			0.1408	0.019825	0.00405	0.4
4	37.4	40.4628	-3.0628	9.380744	2.41099878	5.81291512	B coefficient	-4.16279			3.0628	9.380744	0.081893	8.2
6	35.5	34.6222	0.8778	0.770533	0.51099878	0.26111975					0.8778	0.770533	0.024727	2.5
8	29	28.7816	0.2184	0.047699	-5.98900122	35.8681356					0.2184	0.047699	0.007531	0.8
8	25.6	28.7816	-3.1816	10.12258	-9.38900122	88.1533439					3.1816	10.12258	0.124281	12.4
4	42	40.4628	1.5372	2.362984	7.01099878	49.1541039					1.5372	2.362984	0.0366	3.7
8	28.7009	28.7816	-0.0807	0.006512	-6.28810122	39.5402169					0.0807	0.006512	0.002812	0.3
6	37.055	34.6222	2.4328	5.918516	2.06599878	4.26835096					2.4328	5.918516	0.065654	6.6
6	32.3	34.6222	-2.3222	5.392613	-2.68900122	7.23072756					2.3222	5.392613	0.071895	7.2
8	21.8	28.7816	-6.9816	48.74274	-13.18900122	173.949753					6.9816	48.74274	0.320257	32.0
4	41.2	40.4628	0.7372	0.543464	6.21099878	38.5765059					0.7372	0.543464	0.017893	1.8
10	24.3325	22.941	1.3915	1.936272	-10.65650122	113.561018					1.3915	1.936272	0.057187	5.7
6	34.9	34.6222	0.2778	0.077173	-0.08900122	0.00792122					0.2778	0.077173	0.00796	0.8
6	35.8	34.6222	1.1778	1.387213	0.81099878	0.65771902					1.1778	1.387213	0.032899	3.3
8	21.4734	28.7816	-7.3082	53.40979	-13.51560122	182.671476					7.3082	53.40979	0.340337	34.0
4	41.5	40.4628	1.0372	1.075784	6.51099878	42.3931051					1.0372	1.075784	0.024993	2.5
6	28.5668	34.6222	-6.0554	36.66787	-6.42220122	41.2446685					6.0554	36.66787	0.211973	21.2

	n	82
MAD		4.06664878
MSE		32.66996772
RMSE		5.715764841
MAPE		11.8571359

Mean absolute percentage error

R square by calculation	0.586434

As per plot	0.5479

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.

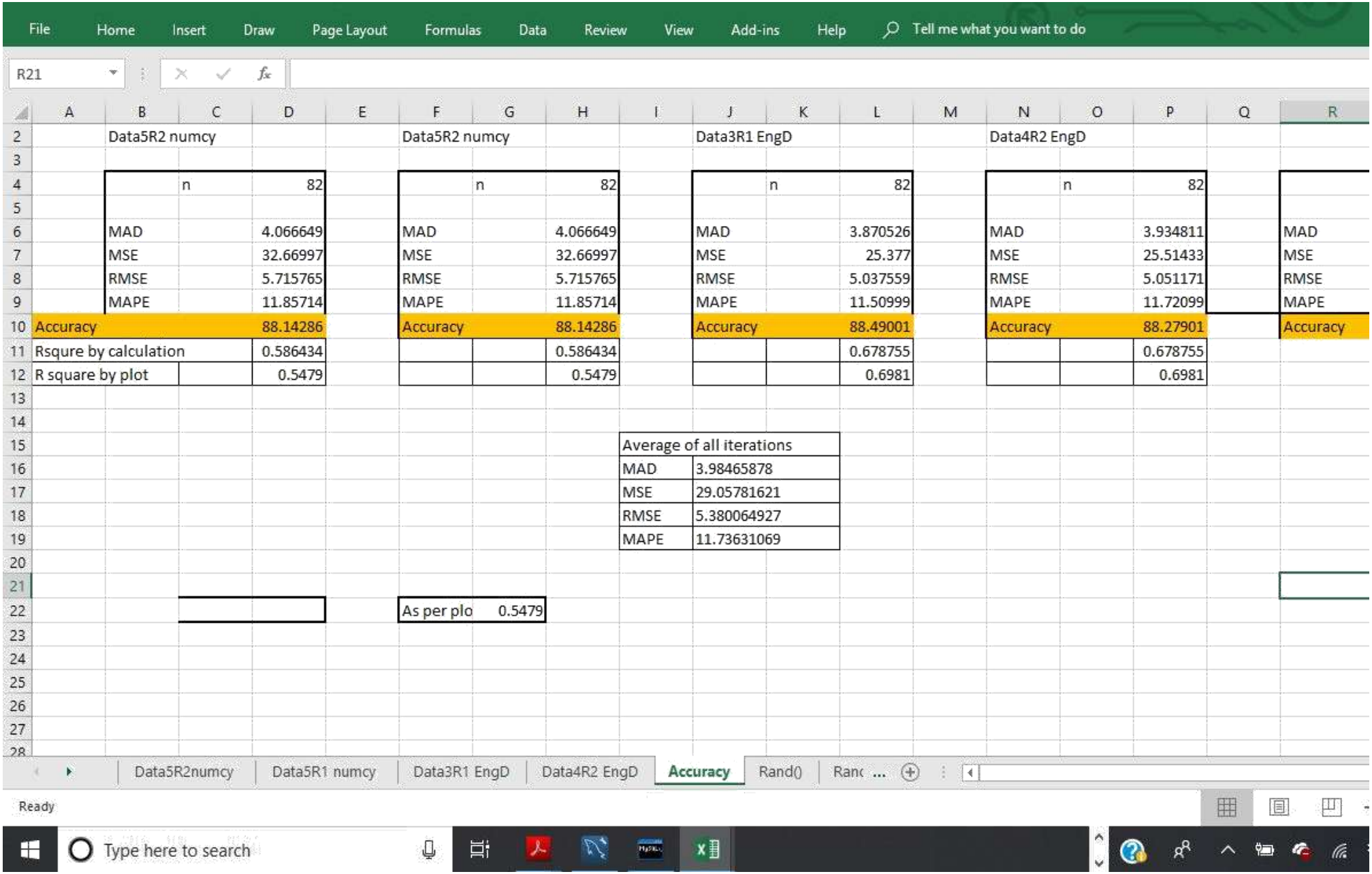
c. Iterate this process 3 time to cover all possible selection of 2 parts for modeling and the 3rd part for testing. There are 3 possible combination in this way. So you would end up with creating 3 models on three different dataset.

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.

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

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

For the requirement of iterations in the question the same is repeated with Rand 1 and Rand2 variables and the consolidate finding is given below



		Data5R2 numcy				Data5R2 numcy				Data3R1 EngD				Data4R2 EngD			
			n	82			n	82			n	82			n	82	
		MAD		4.066649		MAD		4.066649		MAD		3.870526		MAD		3.934811	MAD
		MSE		32.66997		MSE		32.66997		MSE		25.377		MSE		25.51433	MSE
		RMSE		5.715765		RMSE		5.715765		RMSE		5.037559		RMSE		5.051171	RMSE
		MAPE		11.85714		MAPE		11.85714		MAPE		11.50999		MAPE		11.72099	MAPE
	Accuracy			88.14286		Accuracy		88.14286		Accuracy		88.49001		Accuracy		88.27901	Accuracy
	Rsqure by calculation			0.586434				0.586434				0.678755				0.678755	
	R square by plot			0.5479				0.5479				0.6981				0.6981	

Average of all iterations	
MAD	3.98465878
MSE	29.05781621
RMSE	5.380064927
MAPE	11.73631069

As per plo	0.5479

Based on the findings the MAPE values of all tests are found to be vey close though the test is done on various input variable and the accuracy is found to be around 88.3%

The average of the all the model of Engdispl Beta coefficient is given below and further the average Beta coefficient is applied and predicted the FE as given in the screenshot below. And the MAPE values and accuracy is calculated as per the requirement in the question(Test the final Accuracy by implementing the model on 2011 dataset.) and the same is given below. Relevant excel sheets are attached separated in the submission,

	Eng displ
	-4.8939
	-5.36912
	-5.35080
Average	-5.20460667

		245
MAD		4.025664653
MSE		31.19592662
RMSE		5.585331379
MAPE		11.17662112

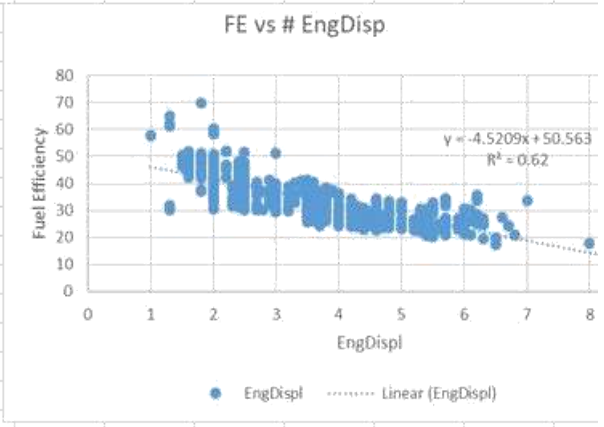
Mean absolute percentage error

0.630435	R 2 by calculation

File Home Insert Draw Page Layout Formulas Data Review View Add-ins Help Tell me what you want to do

C3 =50.563-5.0246*A3

	A	B	C	D	E	F	G	H	I	J	K
1	x	y	predicted y	Error	Err square	Dis Y and their mean	Square of F		Abs v of error	Error 2	ABS values of error/Actual value
2	EngDispl	FE	Predicted FE								
3	3.6	29.5	32.47444	-2.9744	8.8472933	-5.2307	27.359847		2.97444	8.84729	0.10083
4	5	24.793	25.44	-0.6472	0.4188678	-9.9379	98.761143		0.6472	0.41887	0.0261
5	4.4	33.603	28.45476	5.14844	26.506434	-1.1275	1.2711753		5.14844	26.5064	0.15321
6	2.3	34.7	39.00642	-4.3064	18.545253	-0.0307	0.0009403		4.30642	18.5453	0.1241
7	1.8	50.8	41.51872	9.28128	86.142158	16.0693	258.22356		9.28128	86.1422	0.1827
8	3.8	36.767	31.46952	5.29738	28.062235	2.03624	4.1462567		5.29738	28.0622	0.14408
9	4.7	24.6	26.94738	-2.3474	5.5101929	-10.131	102.63035		2.34738	5.51019	0.09542
10	3.4	41.347	33.47936	7.86764	61.899759	6.61634	43.775901		7.86764	61.8998	0.19028
11	1.5	49.6	43.0261	6.5739	43.216161	14.8693	221.09715		6.5739	43.2162	0.13254
12	3.7	28.1	31.97198	-3.872	14.992229	-6.6307	43.965706		3.87198	14.9922	0.13779
13	4.6	24.3	27.44984	-3.1498	9.921492	-10.431	108.79875		3.14984	9.92149	0.12962
14	5	30.85	25.44	5.4103	29.271346	-3.8804	15.057225		5.4103	29.2713	0.17537
15	6	21.652	20.4154	1.2361	1.5279432	-13.079	171.06453		1.2361	1.52794	0.05709
16	6.2	19.514	19.41048	0.10342	0.0106957	-15.217	231.54991		0.10342	0.0107	0.0053
17	5.3	29	23.93262	5.06738	25.67834	-5.7307	32.840511		5.06738	25.6783	0.17474
18	5.4	20.6	23.43016	-2.8302	8.0098056	-14.131	199.67567		2.83016	8.00981	0.13739
19	2	46.2	40.5138	5.6862	32.33287	11.4693	131.54567		5.6862	32.3329	0.12308
20	2.5	31.367	38.0015	-6.6346	44.017917	-3.3638	11.314909		6.6346	44.0179	0.21152
21	2.5	36.656	38.0015	-1.3458	1.8111776	1.92504	3.7057633		1.3458	1.81118	0.03671
22	3	38	35.4892	2.5105	6.3026103	3.26904	10.686596		2.5105	6.30261	0.06607
23	2.5	40.108	38.0015	2.1062	4.4360784	5.37704	28.912515		2.1062	4.43608	0.05251
24	4.6	33.9	27.44984	6.45016	41.604564	-0.8307	0.6900028		6.45016	41.6046	0.19027
25	3.5	35.8	32.9769	2.8231	7.9698936	1.06934	1.1434793		2.8231	7.96989	0.07886
26	5	27.438	25.44	1.9975	3.9900063	-7.2932	53.190242		1.9975	3.99001	0.0728
27	2.4	38.7	38.50396	0.19604	0.0384317	3.96934	15.755628		0.19604	0.03843	0.00507
28	6	32.4	20.4154	11.9846	143.63064	-2.3307	5.4319951		11.9846	143.631	0.3699
29	4.4	30.548	28.45476	2.09324	4.3816537	-4.1827	17.494679		2.09324	4.38165	0.06852



Fe2011Engdipl | Data5R2numcy | Data5R1 numcy | Data3R1 EngD | Data4R2 EngD | Accuracy ...

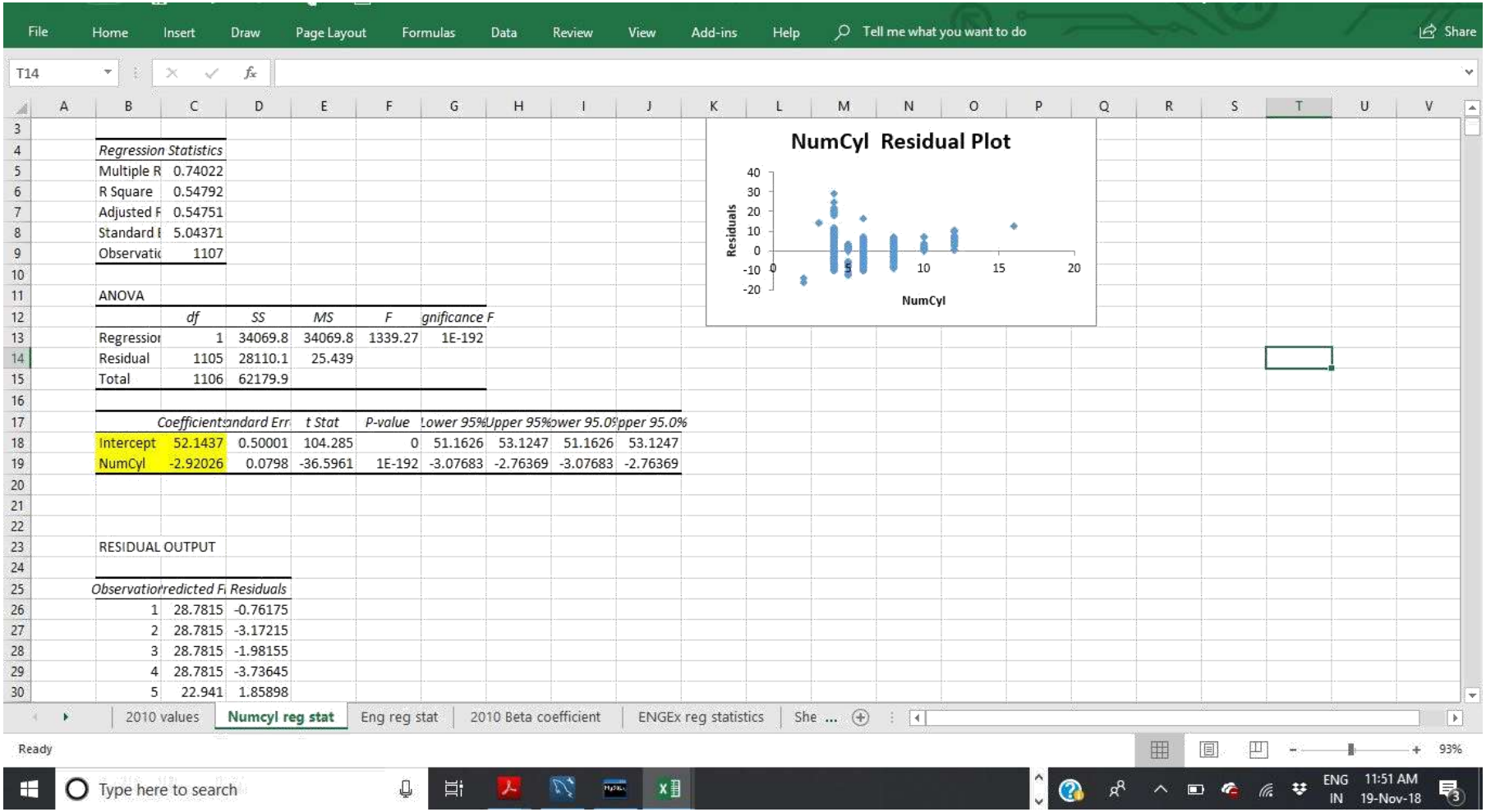
Ready

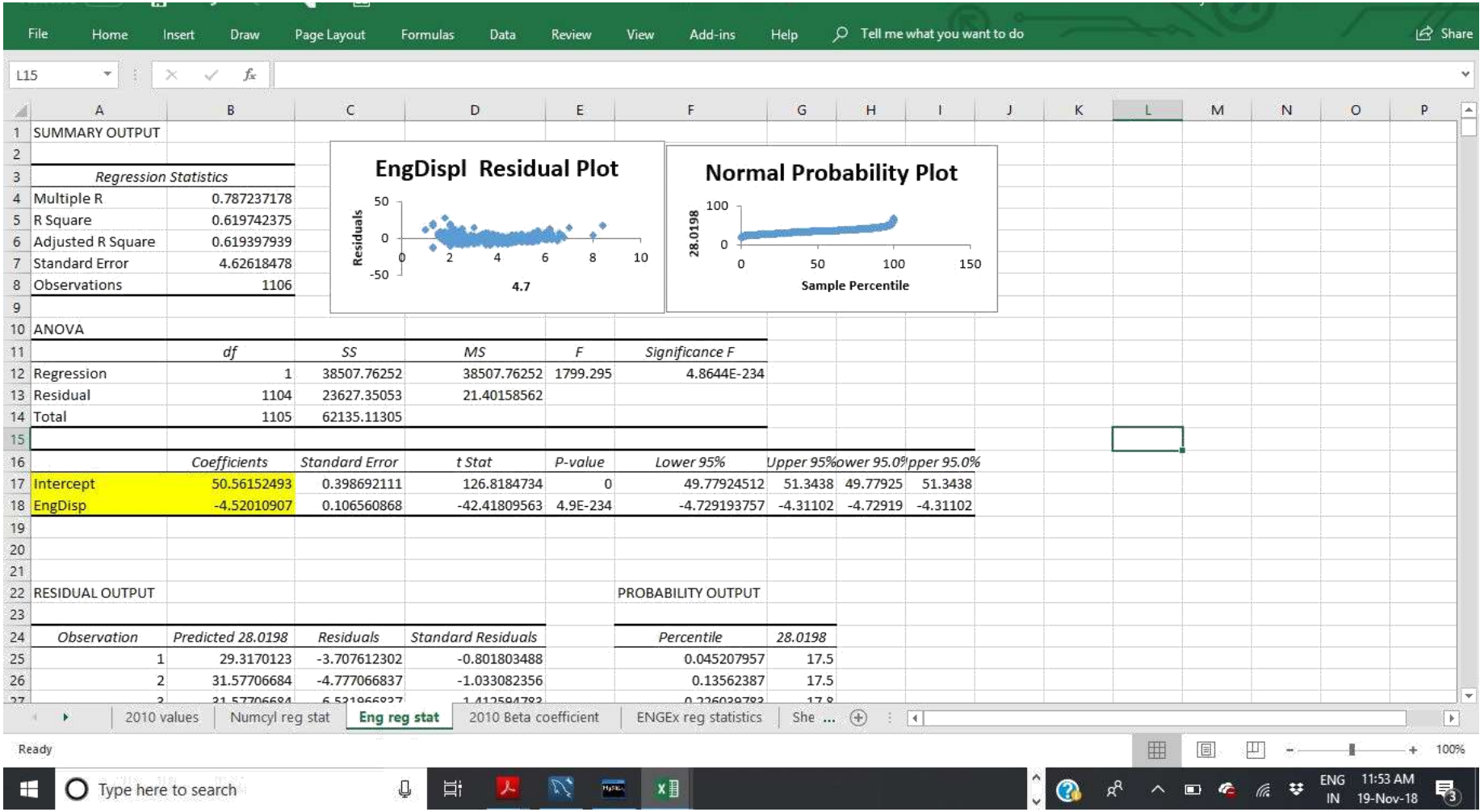
Type here to search

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.

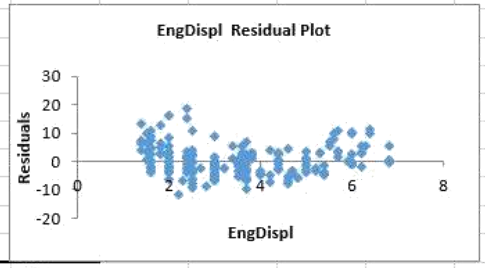
8. You should be able to repeat all the points asked under “Use Excel” using Data Analysis tool. You may need to do the random sampling separately here as well.

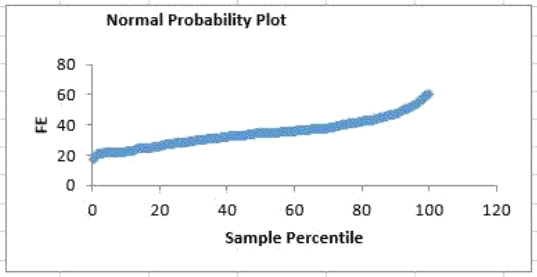

File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Tell me what you want to do
Share
T14
NumCyl Residual Plot
Residuals
NumCyl
40
30
20
10
0
-10
-20
0
5
10
15
20
Regression Statistics
Multiple R 0.74022
R Square 0.54792
Adjusted F 0.54751
Standard E 5.04371
Observatic 1107
ANOVA
df SS MS F gnificance F
Regressior 1 34069.8 34069.8 1339.27 1E-192
Residual 1105 28110.1 25.439
Total 1106 62179.9
Coefficients tandard Err t Stat P-value Lower 95% Upper 95% ower 95.0% pper 95.0%
Intercept 52.1437 0.50001 104.285 0 51.1626 53.1247 51.1626 53.1247
NumCyl -2.92026 0.0798 -36.5961 1E-192 -3.07683 -2.76369 -3.07683 -2.76369
RESIDUAL OUTPUT
Observatior redicted Fi Residuals
1 28.7815 -0.76175
2 28.7815 -3.17215
3 28.7815 -1.98155
4 28.7815 -3.73645
5 22.941 1.85898
2010 values
Numcyl reg stat
Eng reg stat
2010 Beta coefficient
ENGEx reg statistics
She ...
Ready
93%
Type here to search
ENG
IN
11:51 AM
19-Nov-18

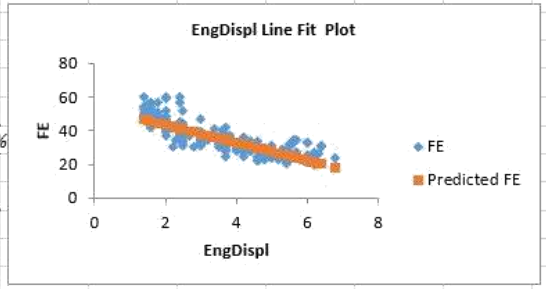


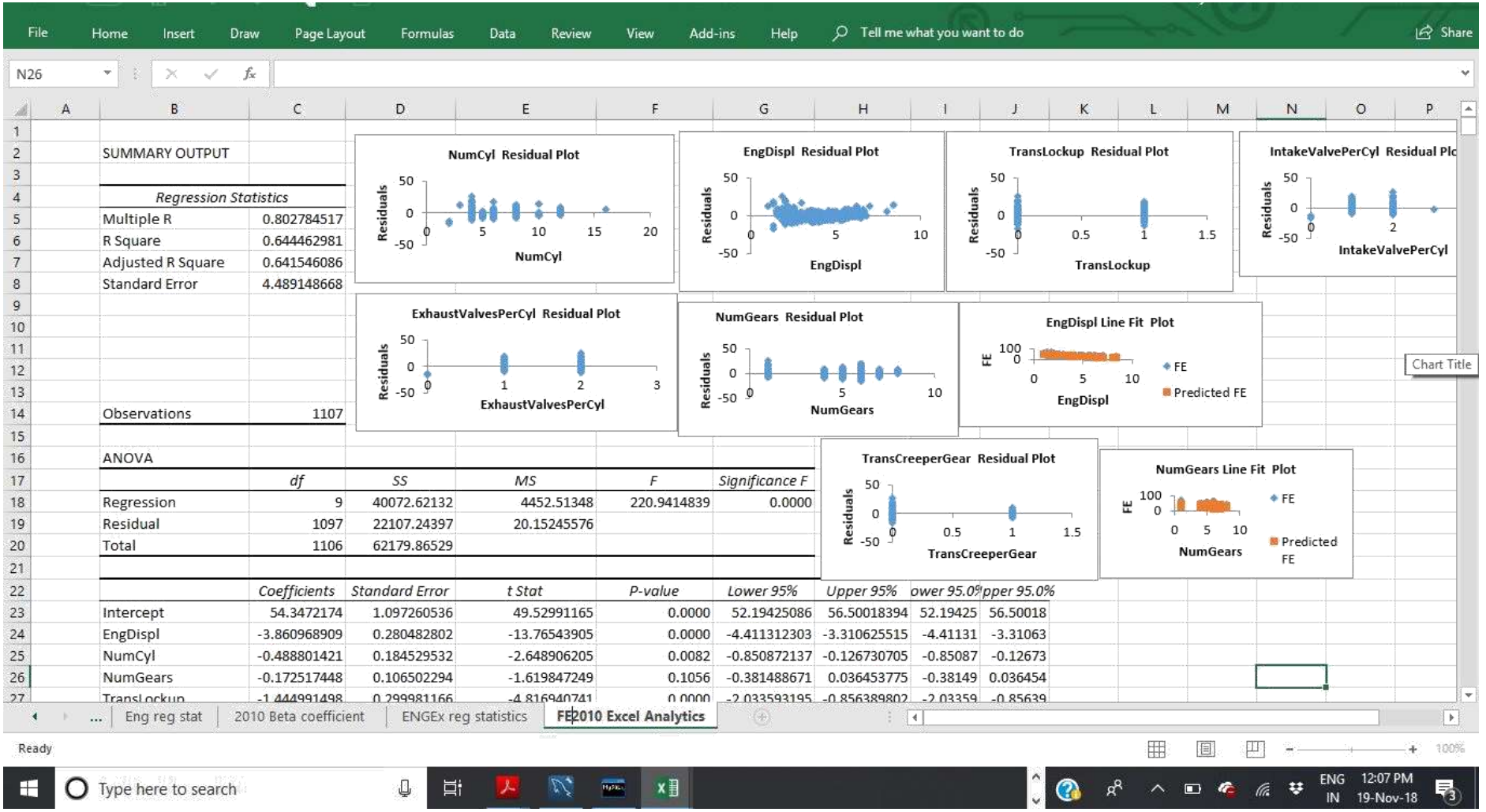
	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.787237178							
5	R Square	0.619742375							
6	Adjusted R Square	0.619397939							
7	Standard Error	4.62618478							
8	Observations	1106							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	38507.76252	38507.76252	1799.295	4.8644E-234			
13	Residual	1104	23627.35053	21.40158562					
14	Total	1105	62135.11305						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	50.56152493	0.398692111	126.8184734	0	49.77924512	51.3438	49.77925	51.3438
18	EngDisp	-4.52010907	0.106560868	-42.41809563	4.9E-234	-4.729193757	-4.31102	-4.72919	-4.31102
19									
20									
21									
22	RESIDUAL OUTPUT					PROBABILITY OUTPUT			
23									
24	*Observation*	*Predicted 28.0198*	*Residuals*	*Standard Residuals*		*Percentile*	*28.0198*		
25	1	29.3170123	-3.707612302	-0.801803488		0.045207957	17.5		
26	2	31.57706684	-4.777066837	-1.033082356		0.13562387	17.5		
27	3	31.57706684	6.521966837	1.412594783		0.226039783	17.8		

SUMMARY OUTPUT									
Regression Statistics									
Multiple F	0.837773								
R Square	0.701864								
Adjusted I	0.700637								
Standard E	5.037217								
Observati	245								
ANOVA									
	df	*SS*	*MS*	*F*	*gnificance F*				
Regressio	1	14515.31	14515.31	572.0648	8.44E-66				
Residual	243	6165.773	25.37355						
Total	244	20681.09							
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
Intercept	53.78837	0.859332	62.59321	6.9E-152	52.09568	55.48107	52.09568	55.48107	
EngDispl	-5.2421	0.219171	-23.9179	8.44E-66	-5.67382	-4.81038	-5.67382	-4.81038	
RESIDUAL OUTPUT					PROBABILITY OUTPUT				
)bservatio	*redicted F*	*Residuals*	*dard Residuals*		*Percentile*	*FE*			
1	34.91681	-5.41681	-1.07757		0.204082	17.7			
2	27.57787	-2.78507	-0.55404		0.612245	18.6			

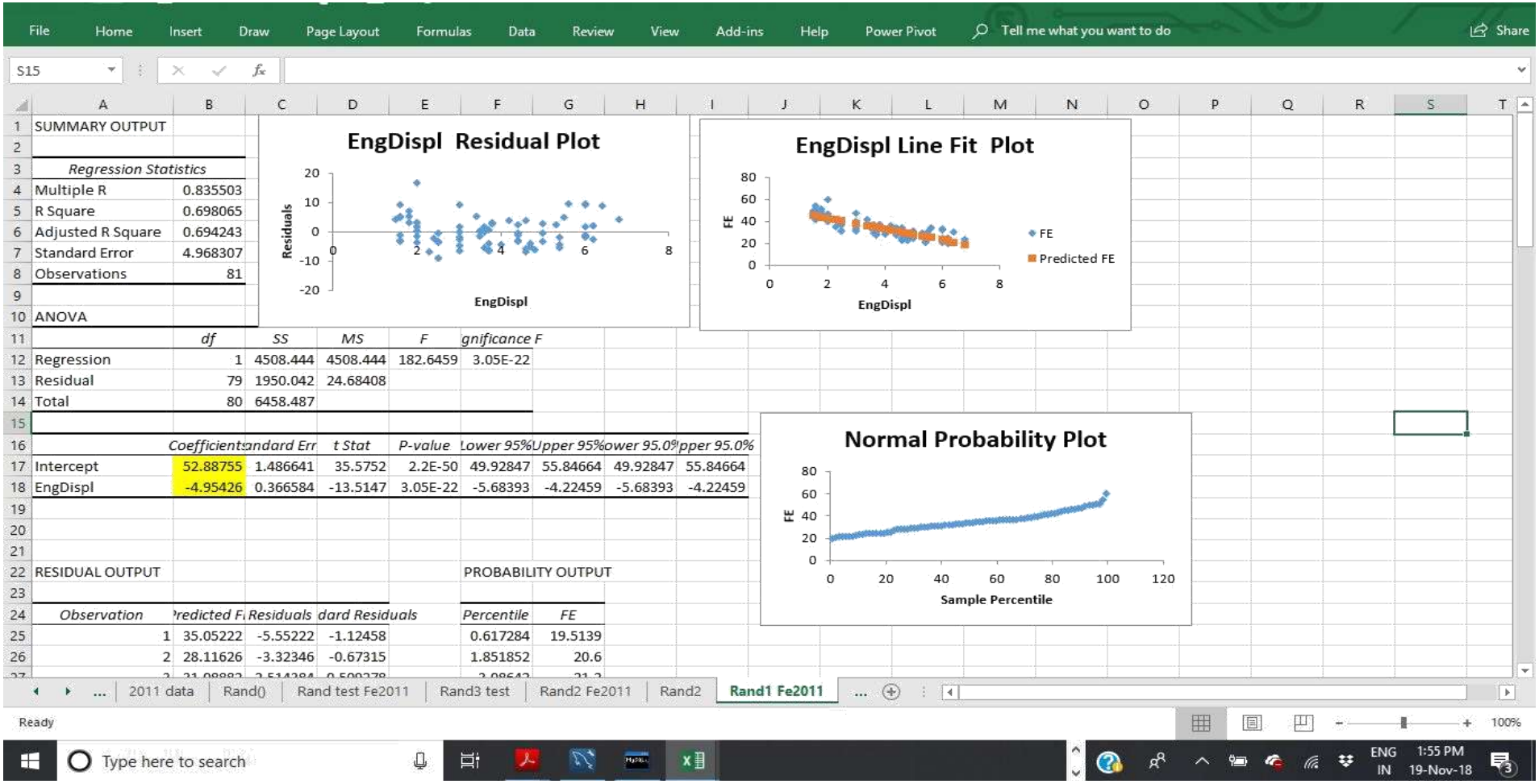


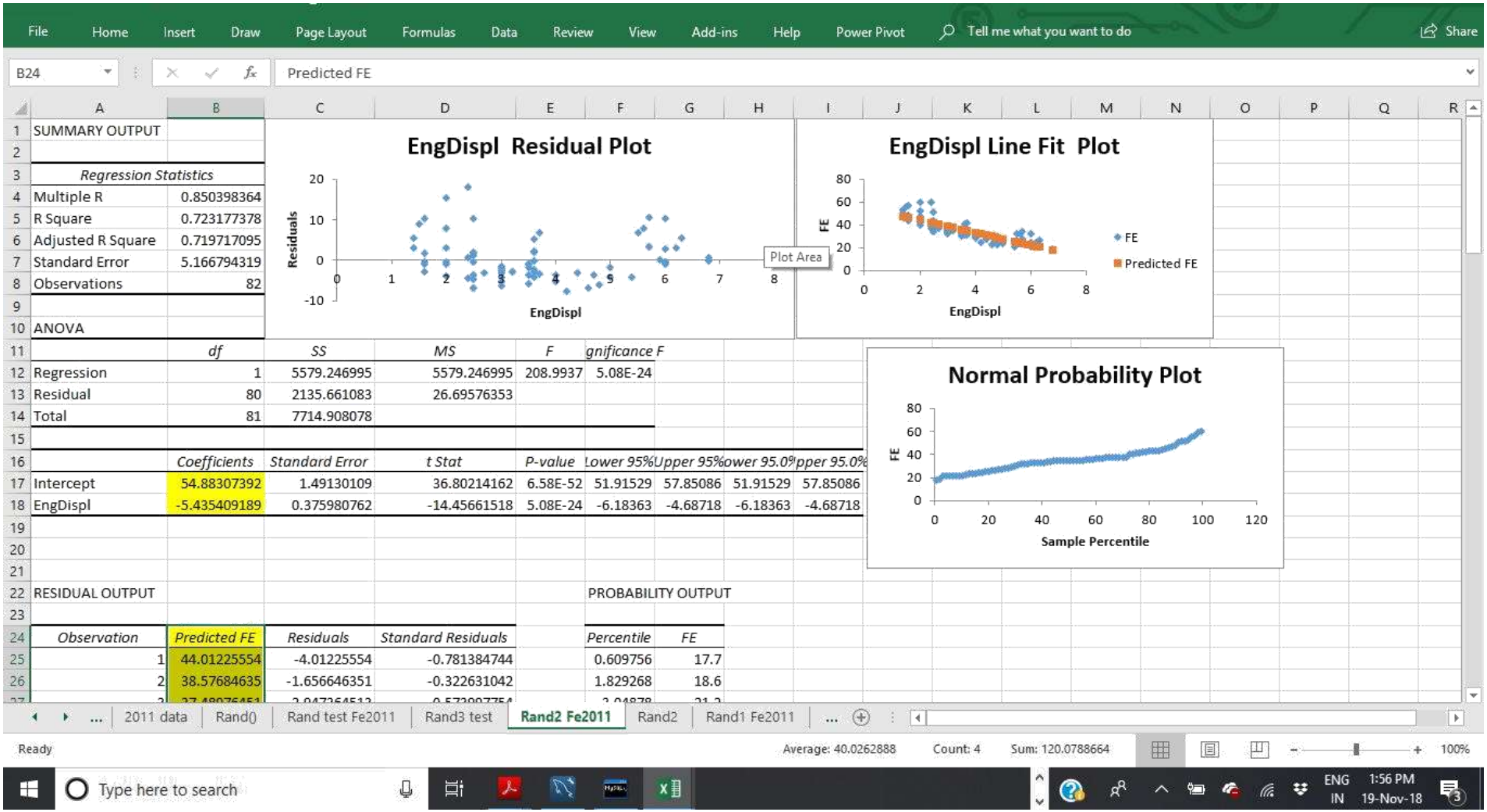


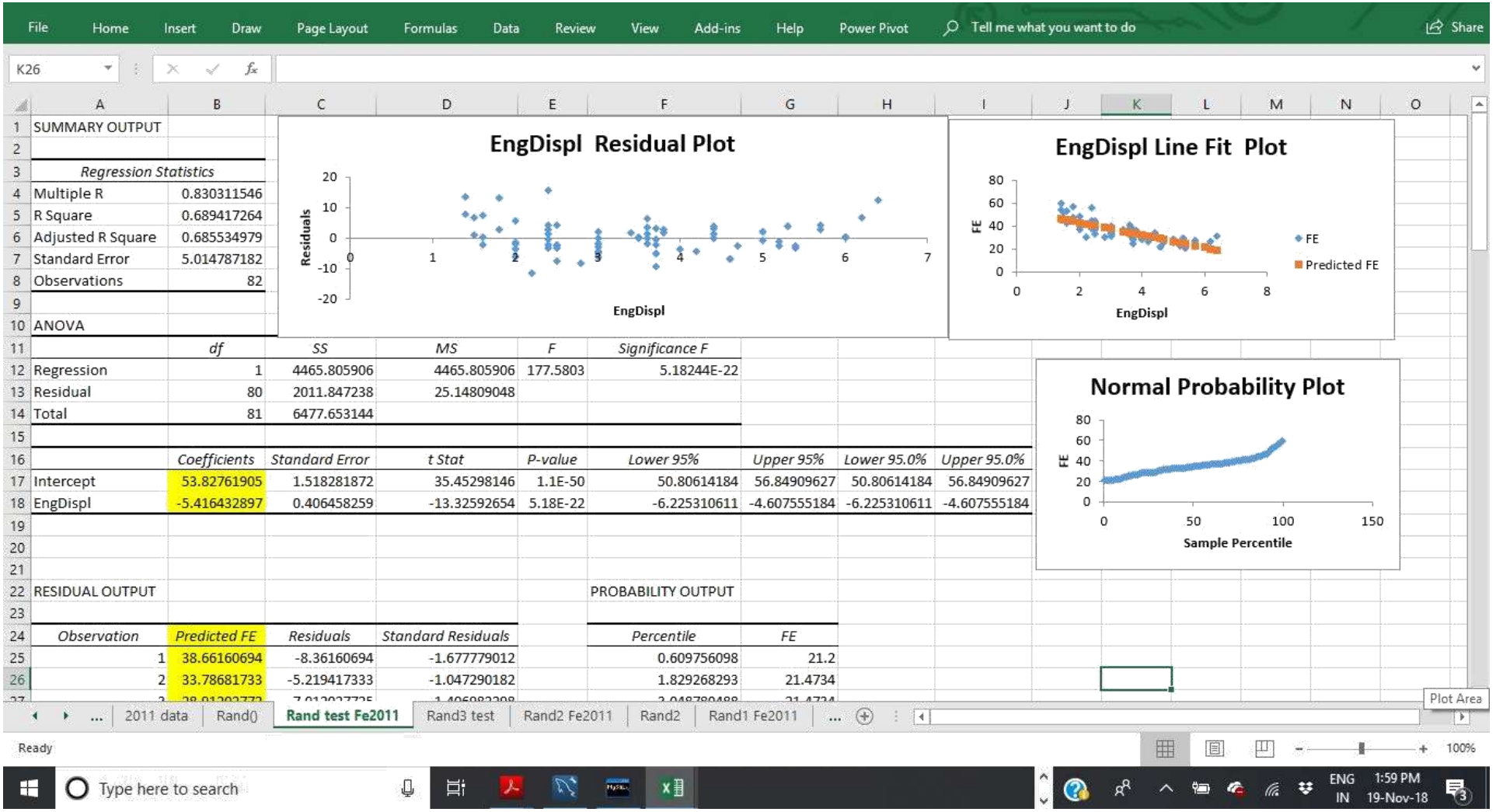




2	SUMMARY OUTPUT									
3										
4	*Regression Statistics*									
5	Multiple R	0.802784517								
6	R Square	0.644462981								
7	Adjusted R Square	0.641546086								
8	Standard Error	4.489148668								
9										
10										
11										
12										
13										
14	Observations	1107								
15										
16	ANOVA									
17		*df*	*SS*	*MS*	*F*	*Significance F*				
18	Regression	9	40072.62132	4452.51348	220.9414839	0.0000				
19	Residual	1097	22107.24397	20.15245576						
20	Total	1106	62179.86529							
21										
22		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
23	Intercept	54.3472174	1.097260536	49.52991165	0.0000	52.19425086	56.50018394	52.19425	56.50018	
24	EngDispl	-3.860968909	0.280482802	-13.76543905	0.0000	-4.411312303	-3.310625515	-4.41131	-3.31063	
25	NumCyl	-0.488801421	0.184529532	-2.648906205	0.0082	-0.850872137	-0.126730705	-0.85087	-0.12673	
26	NumGears	-0.172517448	0.106502294	-1.619847249	0.1056	-0.381488671	0.036453775	-0.38149	0.036454	
27	TransLockup	-1.444991498	0.299981166	-4.816940741	0.0000	-2.033593195	-0.856389802	-2.03359	-0.85639	


File Home Insert Draw Page Layout Formulas Data Review View Add-ins Help Power Pivot Tell me what you want to do Share
S15
SUMMARY OUTPUT
Regression Statistics
Multiple R 0.835503
R Square 0.698065
Adjusted R Square 0.694243
Standard Error 4.968307
Observations 81
ANOVA
df SS MS F gnificance F
Regression 1 4508.444 4508.444 182.6459 3.05E-22
Residual 79 1950.042 24.68408
Total 80 6458.487
Coefficients tandard Err t Stat P-value Lower 95% Upper 95% ower 95.0% pper 95.0%
Intercept 52.88755 1.486641 35.5752 2.2E-50 49.92847 55.84664 49.92847 55.84664
EngDispl -4.95426 0.366584 -13.5147 3.05E-22 -5.68393 -4.22459 -5.68393 -4.22459
RESIDUAL OUTPUT
PROBABILITY OUTPUT
Observation redicted F Residuals dard Residuals Percentile FE
1 35.05222 -5.55222 -1.12458 0.617284 19.5139
2 28.11626 -3.32346 -0.67315 1.851852 20.6
EngDispl Residual Plot
Residuals
EngDispl
EngDispl Line Fit Plot
FE
Predicted FE
EngDispl
Normal Probability Plot
FE
Sample Percentile
2011 data Rand() Rand test Fe2011 Rand3 test Rand2 Fe2011 Rand2 Rand1 Fe2011
Ready
100%
Type here to search
ENG IN 1:55 PM 19-Nov-18


File Home Insert Draw Page Layout Formulas Data Review View Add-ins Help Power Pivot Tell me what you want to do Share
B24
Predicted FE
EngDispl Residual Plot
Residuals
EngDispl
Plot Area
EngDispl Line Fit Plot
FE
Predicted FE
EngDispl
Normal Probability Plot
FE
Sample Percentile
SUMMARY OUTPUT
Regression Statistics
Multiple R 0.850398364
R Square 0.723177378
Adjusted R Square 0.719717095
Standard Error 5.166794319
Observations 82
ANOVA
df SS MS F gnificance F
Regression 1 5579.246995 5579.246995 208.9937 5.08E-24
Residual 80 2135.661083 26.69576353
Total 81 7714.908078
Coefficients Standard Error t Stat P-value Lower 95% Upper 95% ower 95.0% pper 95.0%
Intercept 54.88307392 1.49130109 36.80214162 6.58E-52 51.91529 57.85086 51.91529 57.85086
EngDispl -5.435409189 0.375980762 -14.45661518 5.08E-24 -6.18363 -4.68718 -6.18363 -4.68718
RESIDUAL OUTPUT
PROBABILITY OUTPUT
Observation Predicted FE Residuals Standard Residuals
1 44.01225554 -4.01225554 -0.781384744
2 38.57684635 -1.656646351 -0.322631042
Percentile FE
0.609756 17.7
1.829268 18.6
2011 data Rand() Rand test Fe2011 Rand3 test Rand2 Fe2011 Rand2 Rand1 Fe2011
Ready
Average: 40.0262888 Count: 4 Sum: 120.0788664
100%
Type here to search
ENG IN 1:56 PM 19-Nov-18


File Home Insert Draw Page Layout Formulas Data Review View Add-ins Help Power Pivot Tell me what you want to do Share
K26
SUMMARY OUTPUT
Regression Statistics
Multiple R 0.830311546
R Square 0.689417264
Adjusted R Square 0.685534979
Standard Error 5.014787182
Observations 82
ANOVA
df SS MS F Significance F
Regression 1 4465.805906 4465.805906 177.5803 5.18244E-22
Residual 80 2011.847238 25.14809048
Total 81 6477.653144
Coefficients Standard Error t Stat P-value Lower 95% Upper 95% Lower 95.0% Upper 95.0%
Intercept 53.82761905 1.518281872 35.45298146 1.1E-50 50.80614184 56.84909627 50.80614184 56.84909627
EngDispl -5.416432897 0.406458259 -13.32592654 5.18E-22 -6.225310611 -4.607555184 -6.225310611 -4.607555184
RESIDUAL OUTPUT
Observation Predicted FE Residuals Standard Residuals
1 38.66160694 -8.36160694 -1.677779012
2 33.78681733 -5.219417333 -1.047290182
PROBABILITY OUTPUT
Percentile FE
0.609756098 21.2
1.829268293 21.4734
EngDispl Residual Plot
Residuals
EngDispl
EngDispl Line Fit Plot
FE
Predicted FE
EngDispl
Normal Probability Plot
FE
Sample Percentile
2011 data Rand() Rand test Fe2011 Rand3 test Rand2 Fe2011 Rand2 Rand1 Fe2011
Plot Area
Ready
100%
Type here to search
ENG IN 1:59 PM 19-Nov-18

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
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.

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”.

My SQL Part is submitted separately as MYSQL submission for Project 1.1 as part 2

Acknowledgement

This is a quite interesting project and I have gained a lot of knowledge about Excel analytics, MYSQL and finding the linear relationship in R, Excel graphs are very much interesting. I thank the institute Acadgild and the Mentors Mr. Mohit & Ms. Puja who taught us the R Excel, MYSQL and other subjects to understand the Analytics. I once again thank Acadgild for enlighten me on Machine learning through online teaching and various coding support through the support coordinators. Thank you Acadgild.