

## 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**

Refer the assignment

##### **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 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.

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**

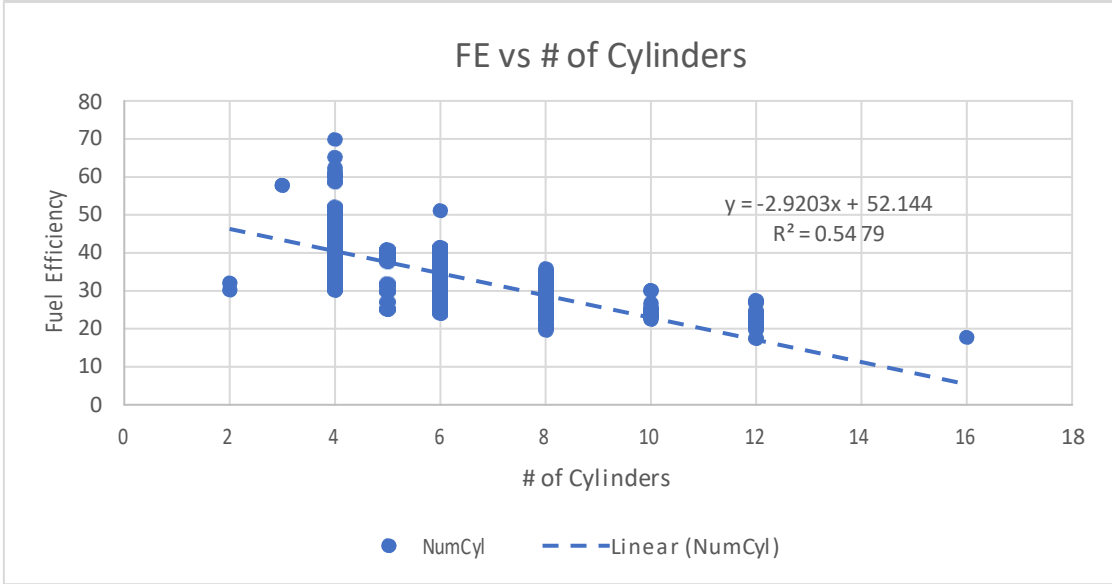
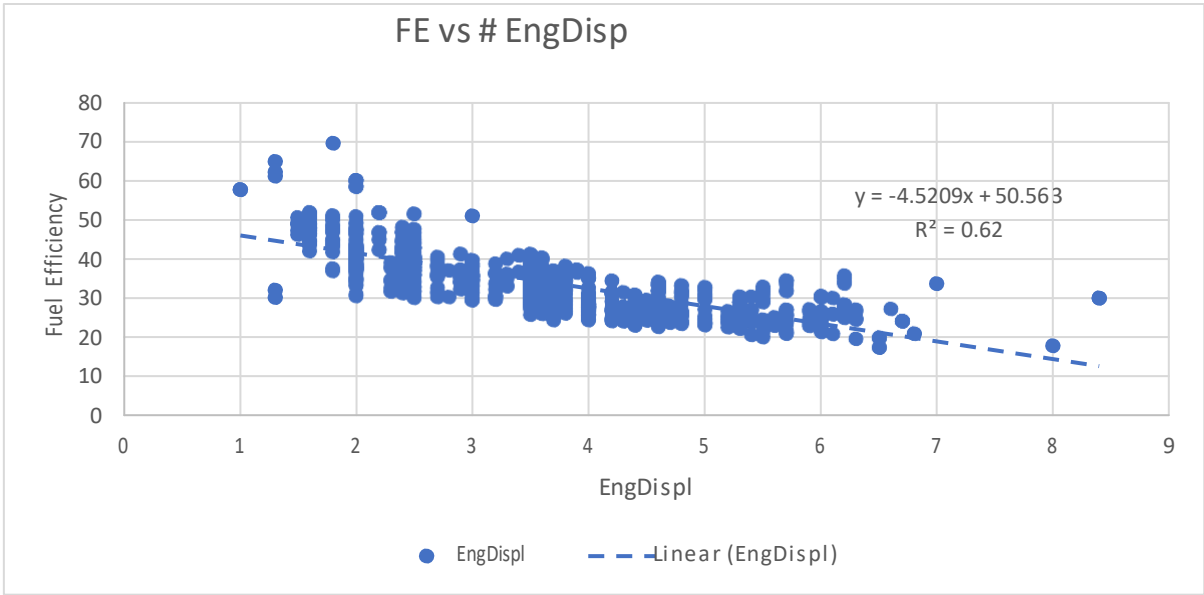
- 7. Use Data Analysis feature of Excel to bypass the co-efficient calculation formulas and compute the Regression Model directly.
- 8. You should be able to repeat all the points asked under “Use Excel” using Data Analysis tool. You may need to do the random sampling separately here as well.

# 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 variables, correlation coefficient is found in excel and is found as given below

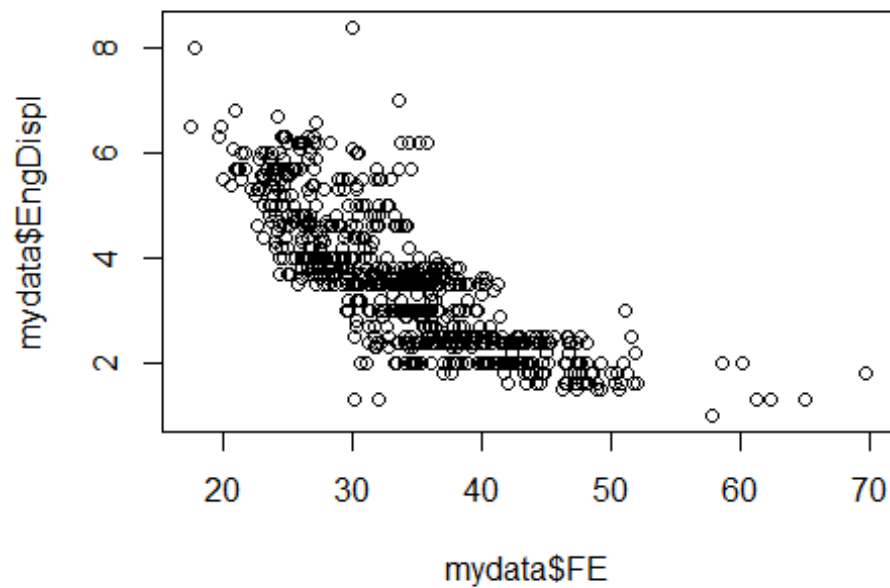
<b>EngDispl</b>	-0.79
<b>Numcyl</b>	-0.74
<b>NumGears</b>	-0.21
<b>TransLockup</b>	-0.27
<b>TransCreeperGear</b>	-0.07
<b>IntakeValvePerCyl</b>	0.28
<b>ExhaustValvesPerCyl</b>	0.34
<b>VarValveTiming</b>	0.12
<b>VarValveLift</b>	0.10

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



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
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
mod
```

```
##
```

```
## Call:
```

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

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

```
summary(mod)
```

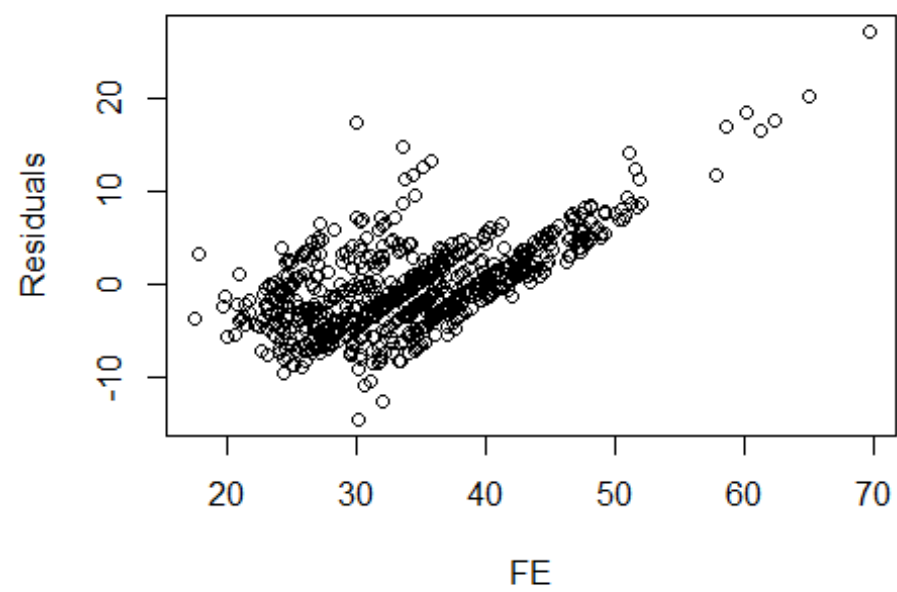
```
##  
## Call:  
## lm(formula = mydata$FE ~ mydata$EngDispl)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -14.486  -3.192  -0.365   2.671  27.215   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    50.5632     0.3985  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")
```



### 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
## Call:
##           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
##
## 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	VarV
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 is 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:

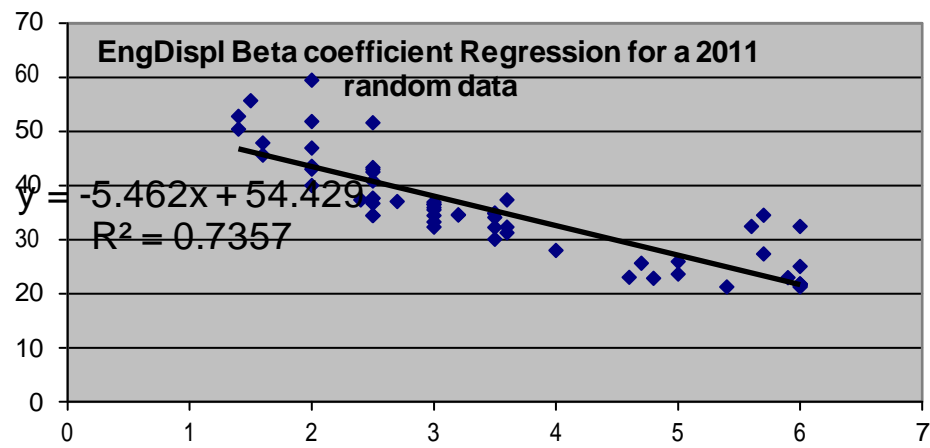
**-5.4164**

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  $n = \text{COUNT}(D3:D247)$  MSE is calculated  $(L248/C250) \times (\text{Error square} / n)$

RMSE is the square root of MSE  $(\text{SQRT}(C253))$

MAPE is calculated  $((M248/C250) \times 100)$   $M248 = \text{sum of (ABS values of error/Actual value)}$  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



Dataset 2011 rand testing - Excel

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

O15

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	EngDispl	NumCyl	FE	NumGear	TransLock	TransCree	IntakeVal	ExhaustVal	VarValveT	VarValveL	Predicted	Rand								
2	3.6	6	29.5	5	1	0	2	2	1	0	0.000532									
3	5	8	24.7928	6	1	0	2	2	1	1	0.006908									
4	4.4	8	33.6032	8	1	0	2	2	1	0	0.007212									
5	2.3	4	34.7	6	0	0	2	2	1	0	0.007591									
6	1.8	4	50.8	6	0	0	2	2	1	0	0.01187									
7	3.8	6	36.7669	7	1	0	2	2	1	1	0.019214									
8	4.7	8	24.6	5	1	0	1	1	0	0	0.021569									
9	3.4	6	41.347	7	1	0	2	2	1	1	0.024303									
10	1.5	4	49.6	5	0	0	2	2	1	0	0.03002									
11	3.7	6	28.1	4	1	0	1	1	0	0	0.040897									
12	4.6	8	24.3	4	1	0	1	1	0	0	0.04208									
13	5	8	30.8503	6	1	0	2	2	1	1	0.048234									
14	6	8	21.6515	6	1	0	1	1	1	0	0.050549									
15	6.2	8	19.5139	6	0	1	1	1	1	0	0.051692									
16	5.3	8	29	6	1	0	1	1	1	0	0.060294									
17	5.4	8	20.6	4	1	1	1	1	0	0	0.062983									
18	2	4	46.2	5	0	0	1	1	0	0	0.067213									
19	2.5	4	31.3669	6	0	0	2	2	1	0	0.073016									
20	2.5	4	36.6557	4	1	0	2	2	0	1	0.074834									
21	3	6	37.9997	8	1	0	2	2	1	1	0.077579									
22	2.5	4	40.1077	1	1	0	2	2	0	1	0.079931									
23	4.6	8	33.9	6	1	0	2	2	1	0	0.084856									
24	3.5	6	35.8	6	1	0	2	2	1	0	0.088245									
25	5	8	27.4375	7	1	0	2	2	1	1	0.090225									
26	2.4	4	38.7	4	1	0	2	2	1	0	0.091763									
27	6	8	32.4	1	0	0	1	1	1	0	0.095059									

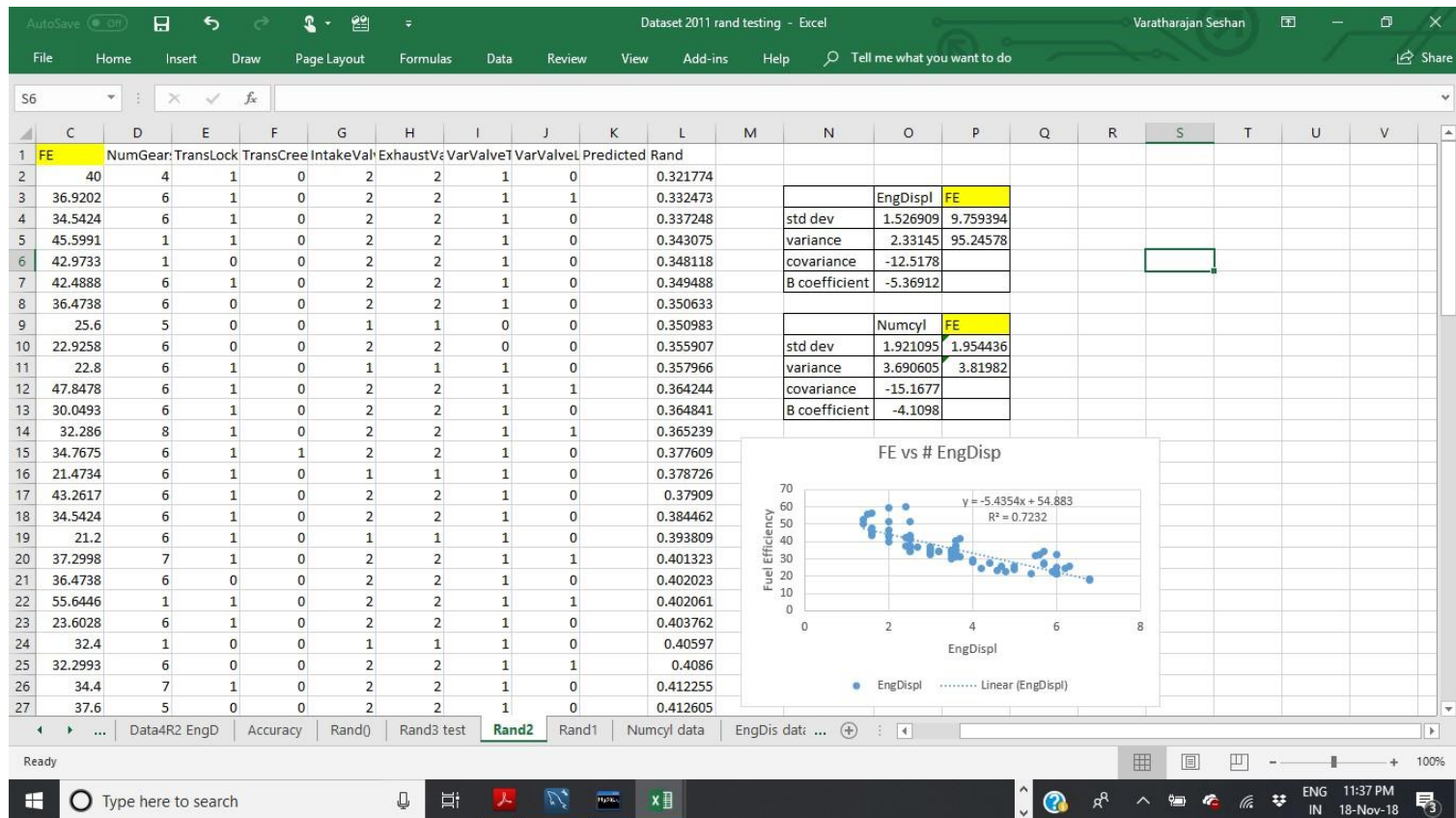
	EngDispl	FE
std dev	1.515269	8.985048
variance	2.29604	80.73108
covariance	-11.2347	
B coefficient	-4.89309	

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

Ready

Average: 12.91120711 Count: 12 Sum: 77.46724265

ENG 11:35 PM  
IN 18-Nov-18



AutoSave Dataset 2011 rand testing - Excel Varatharajan Seshan

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

S1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	EngDispl	NumCyl	FE	NumGear	TransLock	TransCree	IntakeVal	ExhaustVal	VarValveT	VarValveL	Predicted	Rand								
2	2.8	6	30.3	6	1	0	2	2	1	0	0.692438									
3	3.7	6	28.5674	6	0	1	2	2	1	0	0.699183									
4	4.6	8	21.9	4	1	1	1	1	0	0	0.716823									
5	2.4	4	38.7	5	0	0	2	2	1	0	0.717307									
6	6	8	21.4734	6	1	0	1	1	1	0	0.72223									
7	3.5	6	34.763	6	1	1	2	2	1	0	0.723868									
8	2.4	4	37.4	6	1	0	2	2	1	0	0.725082									
9	3.6	6	35.5	6	1	0	2	2	1	0	0.727246									
10	5.3	8	29	6	1	0	1	1	1	0	0.73027									
11	5.7	8	25.6	5	1	0	1	1	1	0	0.732559									
12	2.4	4	42	6	1	0	2	2	1	0	0.733229									
13	5	8	28.7009	6	0	1	2	2	1	0	0.739334									
14	3.4	6	37.055	6	0	0	2	2	1	1	0.73939									
15	3.6	6	32.3	5	1	0	2	2	1	0	0.744939									
16	5.4	8	21.8	4	1	1	1	1	0	0	0.745723									
17	2	4	41.2	1	0	0	2	2	1	0	0.751602									
18	5.2	10	24.3325	6	0	0	2	2	1	0	0.756237									
19	3.5	6	34.9	6	1	0	2	2	1	0	0.758343									
20	3	6	35.8	6	1	0	2	2	1	0	0.763618									
21	6	8	21.4734	6	1	0	1	1	1	0	0.763631									
22	2	4	41.5	4	1	0	2	2	1	0	0.763819									
23	3.7	6	28.5668	6	1	1	2	2	1	0	0.773601									
24	1.4	4	59.7	6	0	0	2	2	1	0	0.774523									
25	5.3	8	29	6	1	0	1	1	1	0	0.779221									
26	3.6	6	40.5	6	1	0	2	2	1	0	0.784926									
27	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

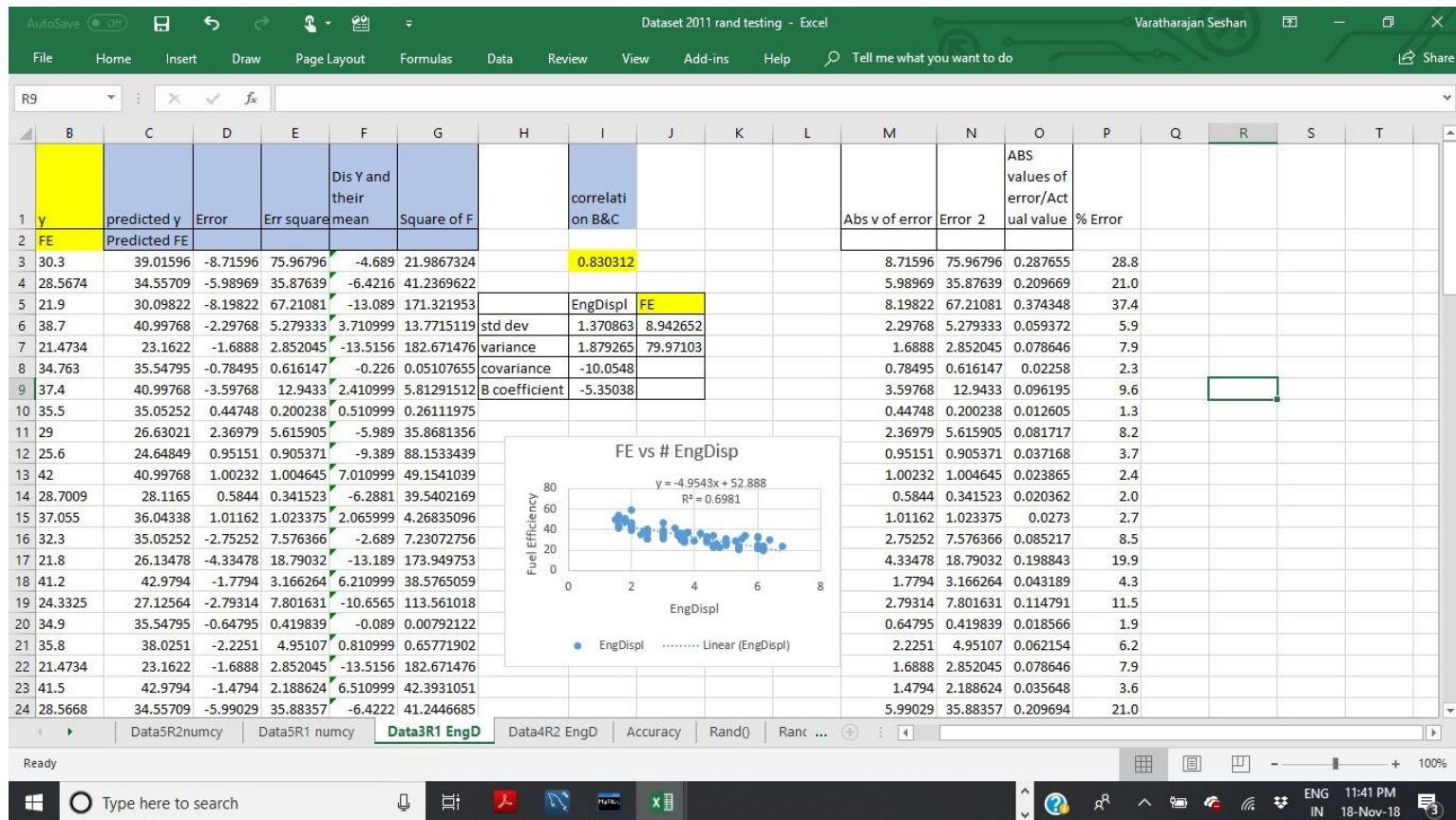
covariance -12.1794

B coefficient -4.16279

Ready

Type here to search

11:39 PM 18-Nov-18



	n	82
MAD		3.870526098
MSE		25.37700219

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

Average % Error

11.5

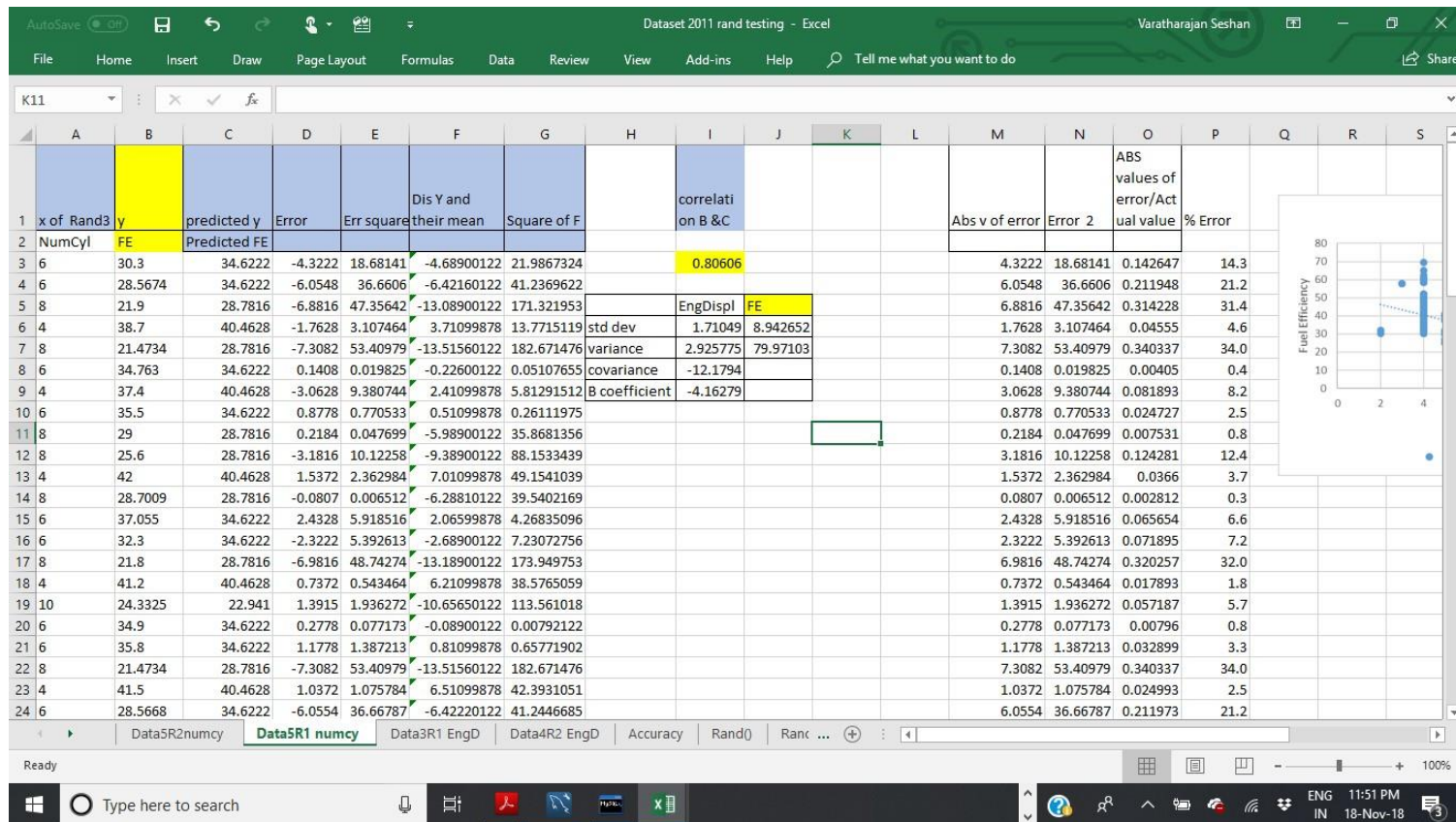


RMSE		5.03755915
MAPE		11.50998577
Mean absolute percentage error		

Average Test Accuracy 88.5

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



	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

13

:f. •

;

Dataset 2011rand testing - Excel

Varatharajan Seshan

File

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View

Add-ins

Help

P

Tell me what you want to do

R21

	A	B	C	D	E	F	G	H		K	L	M	N	O	P	Q	R
2		Dat35R2 numcy				Data5R2 numcy				Data3RIEngD			Data4R2 EngD				
3																	
4		n		82		n		82		n		82		n		82	
5																	
6		MAD		4066649		MAD		4066649		MAD		3870526		MAD		3934811	MAD
7		MSE		3266997		MSE		3266997		MSE		25377		MSE		255433	MSE
8		RMSE		5715765		RMSE		5715765		RMSE		5037559		RMSE		5051171	RMSE
9		MAPE		118574		MAPE		118574		MAPE		1150999		MAPE		1172099	MAPE
10	Accuracy			88.4286		Accuracy		88.4286		Accuracy		88.49001		Accuracy		88.27901	Accuracy
11	Rsquire by calculation			0586434				0586434				0678755				0678755	
12	R square by pbt			05479				05479				06981				06981	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	

Average of all iterations	
MAO	398465878
MSE	290578621
RMSE	5380064927
MAPE	1173631069

As per pbt 054791

Data5R2numcy	Data5R1 numcy	Data3R1 EngD	Data4R2 EngD	Accuracy	RandQ	Rane ... (±)
--------------	---------------	--------------	--------------	----------	-------	--------------

Ready

Q Type here to search





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

The average of 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<sup>2</sup> by calculation

13

:f.

;

Dataset 2011 rand testing - Excel

VaratharaJan Seshar

File

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View

Add-ms

Help

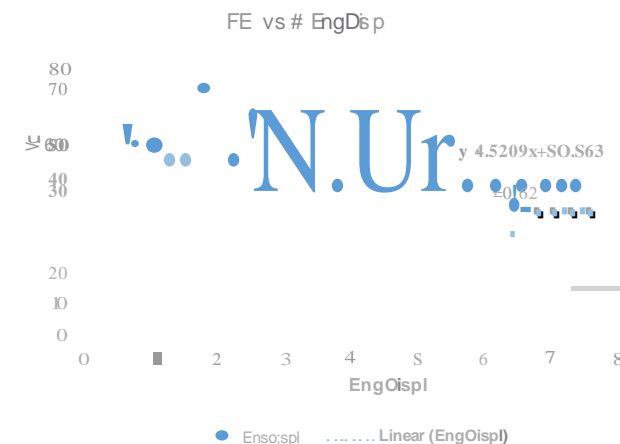
P

Tell me what you want to do

C3

::50.563-5-0246\*A3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	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									
1	EngDispl	FE	Predicted FE																	
3	3.6	29.5	32.47444	-2.97444	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	36.7	30.06122	6.63878	43.86458	10.0703	100.09405		6.63878	43.8645	0.1817									
8	3.8	36.767	31.46952	5.29738	28.062235	2.93624	4.1462567		5.29738	28.0622	0.14498									
9	4.7	24.6	26.94738	-2.3474	5.5101929	-10.131	102.63035		2.34738	5.51019	0.09542									
10	3.4		33.47936		61.899759					61.8998	0.19028									
11	1.5	49.6	43.0261	6.5739	43.216161	14.8693	22.109715		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	17.106453		1.2361	1.52794	0.05709									
16	6.2	19.514	19.41048	0.10342	0.0106957	-15.217	23.154991		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	19.2504	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		IR 4 47.e:																	



Fe2011Engdipl

Data 5R2numcy

Data 5R1 numcy

Data 3R1 EngD

Data 4R2 EngD

Accuracy

®

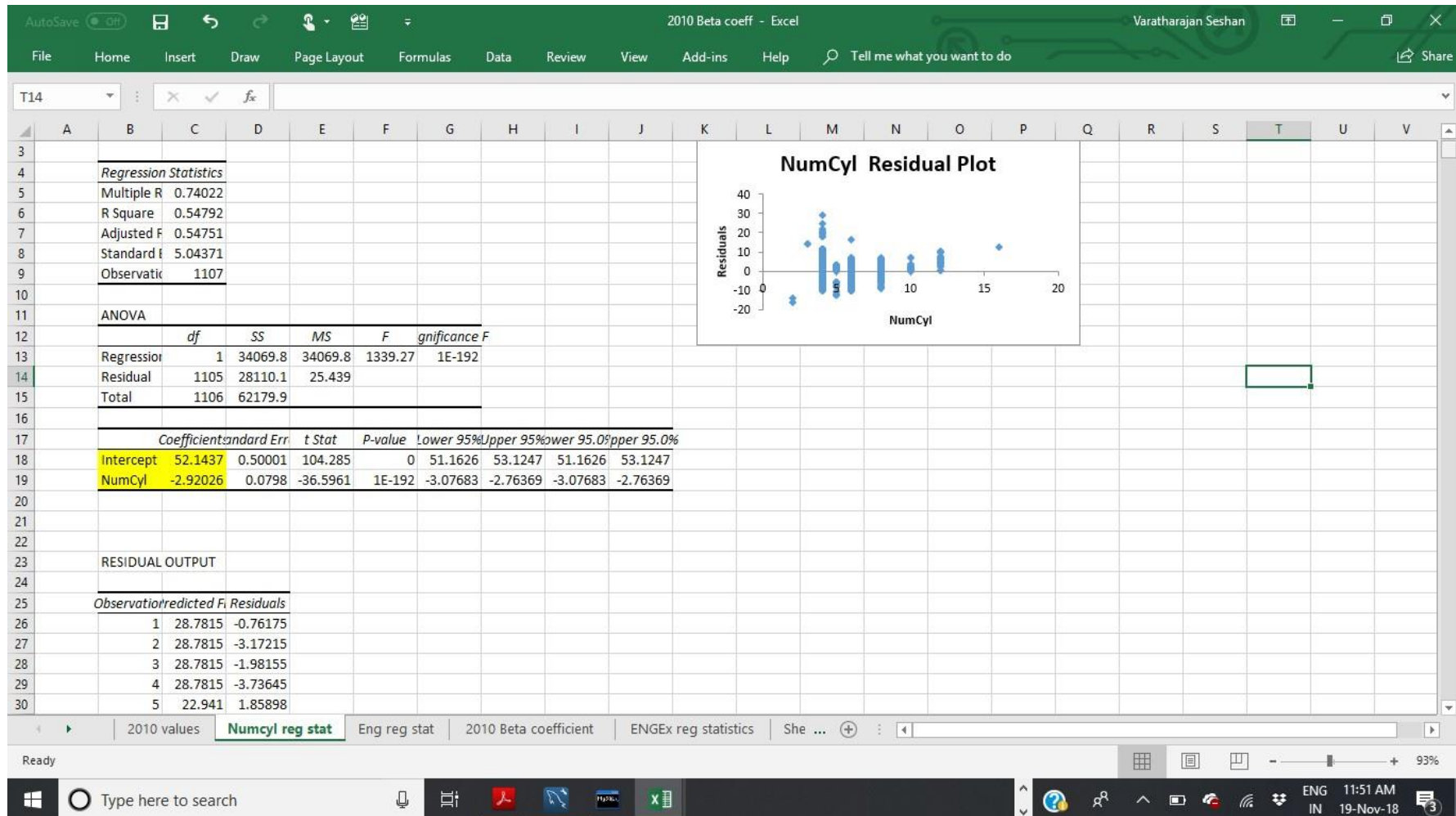
Ready

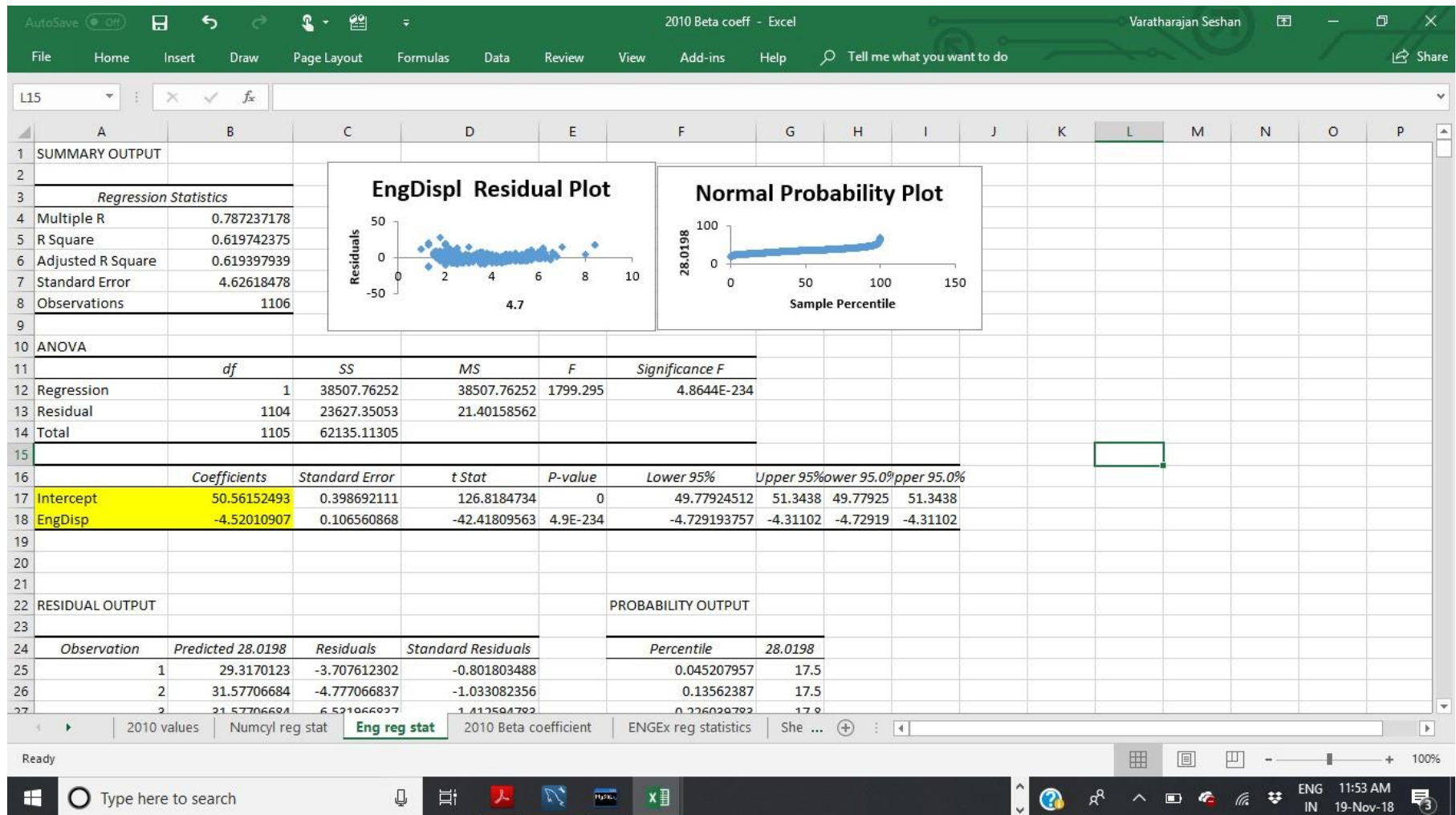
Q Type here to search



## 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.





13

i

O

2.

:

Dataset 2011 rand testing (version 2) - Recovered - Excel

Varatharajan Seshan

III

-

DI

X

File

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View

Add-ins

Help

Power Pivot

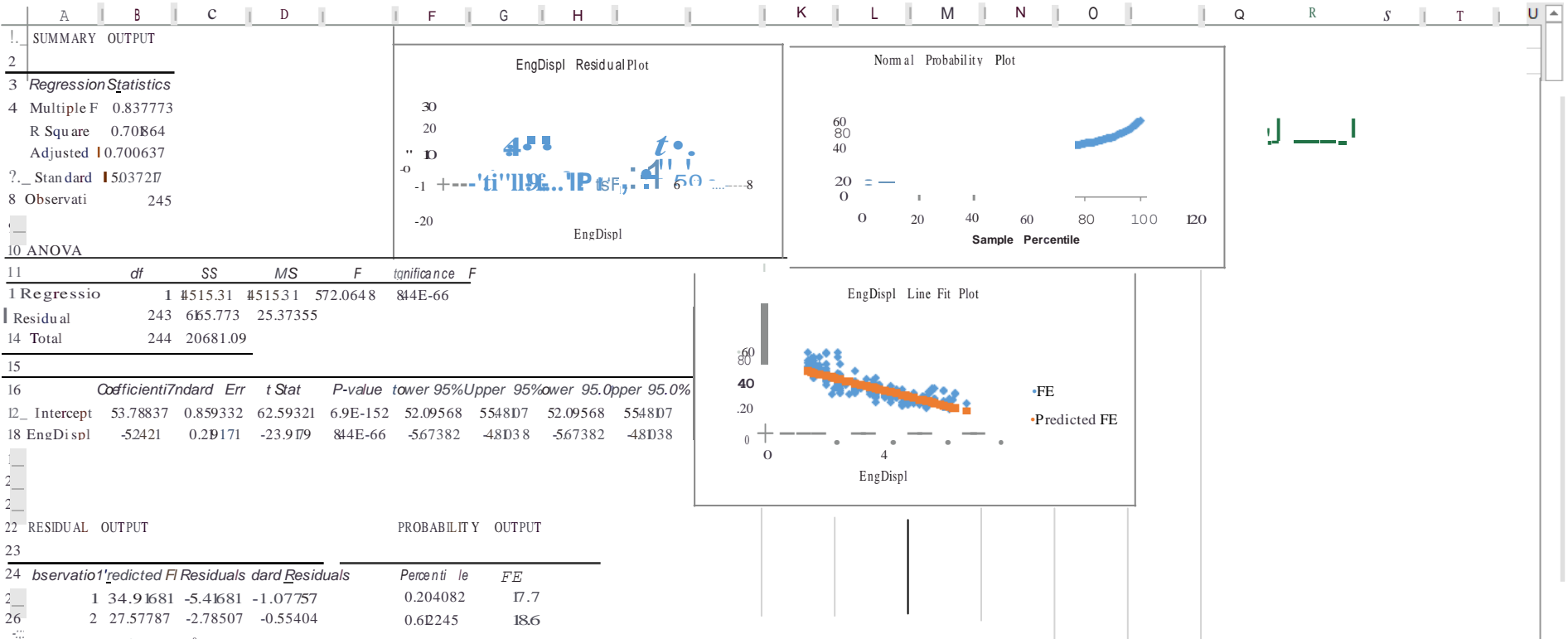
P

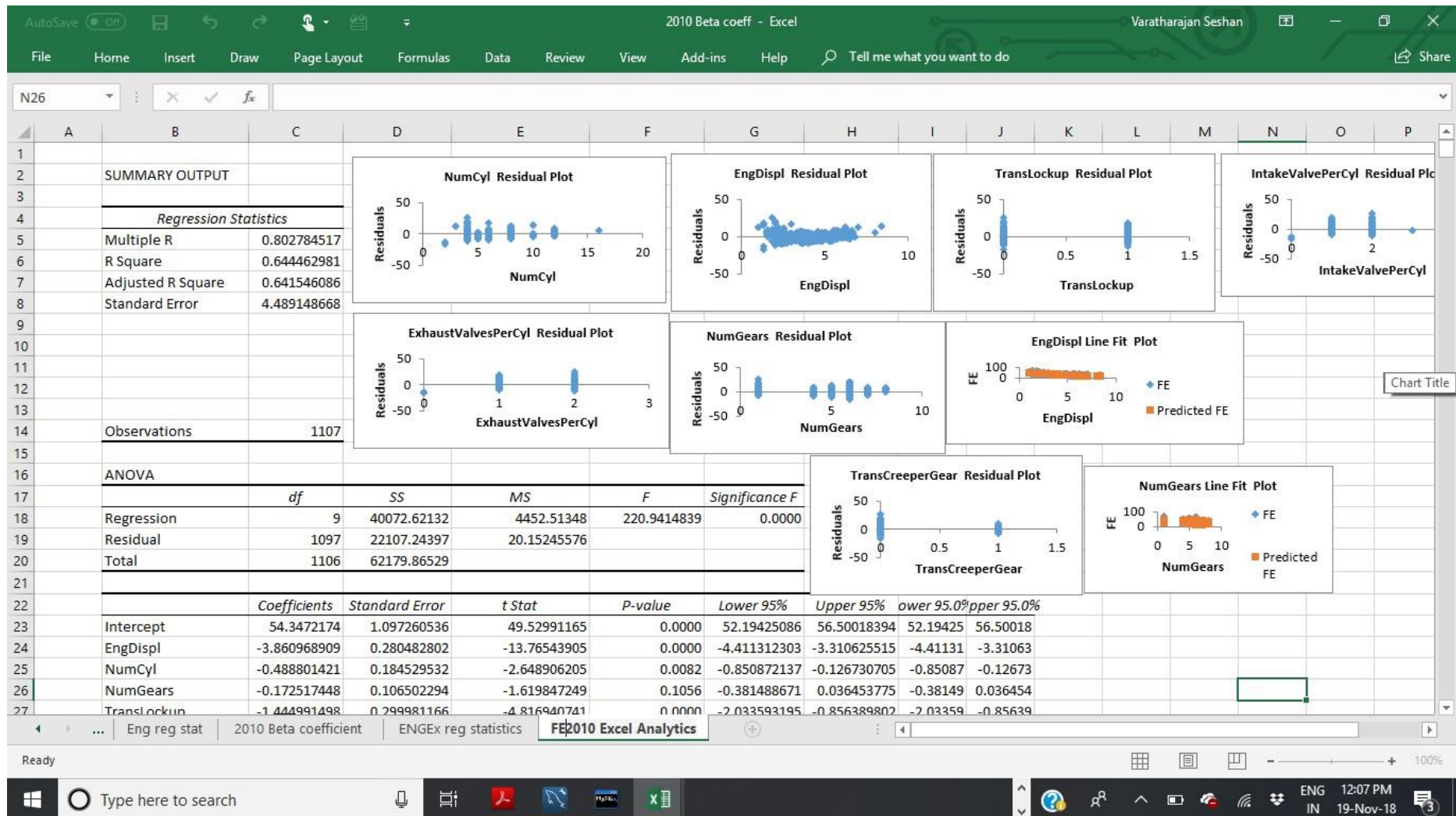
Tell me what you want to do

tB

Share

RS

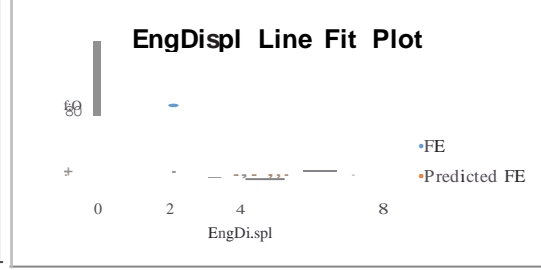
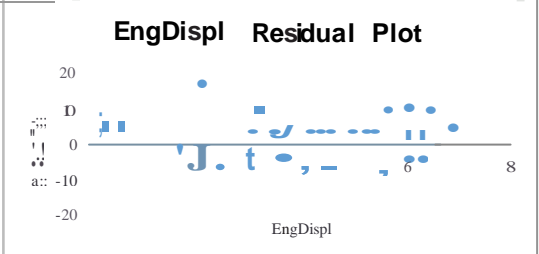




S15 fx

A B C D E F G H I K M N O P Q R S T El

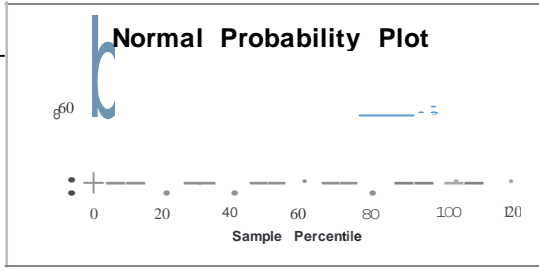
SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.835503
R Square	0.698065
Adjusted R Square	0.694243
Standard Error	4.968307
Observations	81



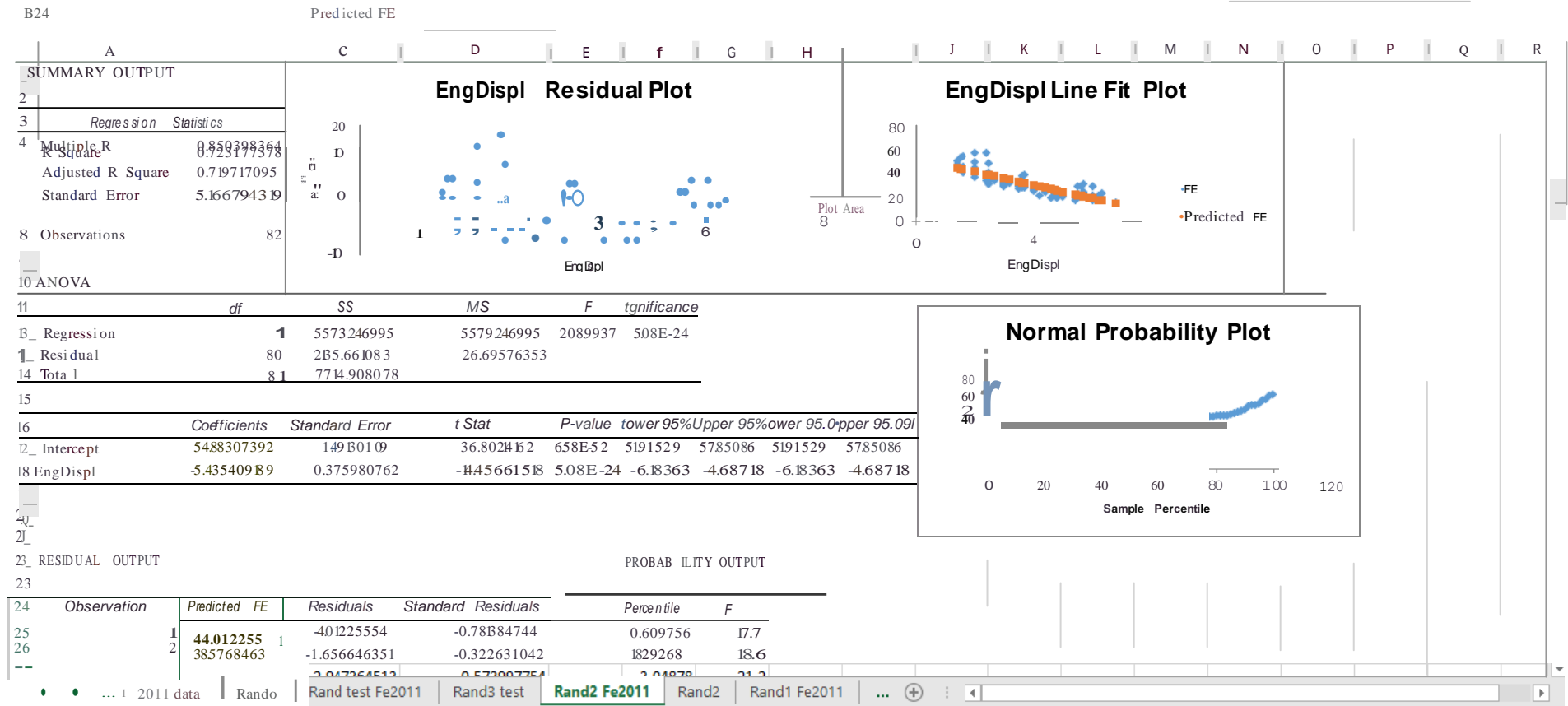
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	4508.444	4508.444	182.6459	3.05E-22
Residual	79	1950.042	24.68408		
Total	80	6458.487			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	528755	1.486641	35.5752	2.2E-50	49.92847	5584664	49.92847	5584664
EngDispl	-485426	0.366584	-13.5147	3.05E-22	-5.68393	-42245.9	-5.68393	-4.22459

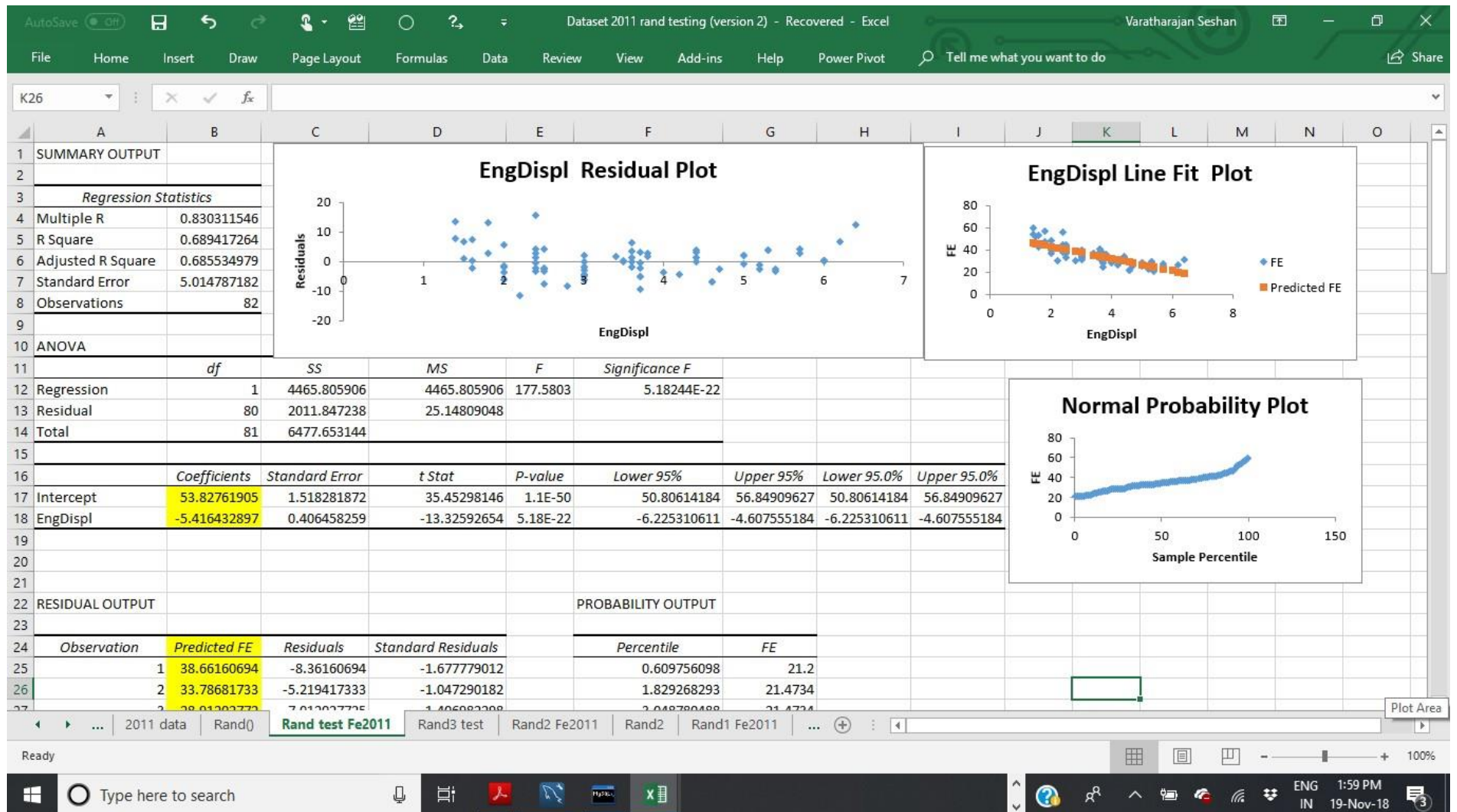
REIDUAL OUTPUT				PROBABILITY OUTPUT	
Observation	Predicted FE	Residuals	Standard Residuals	Percentile	FE
1	3505222	-55222	-11.2458	0.617284	19.5139
2	2811626	-332346	-0.67315	0.51852	206











## 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 “fe2010” table.

Insert the predicted values in an additional column in table “fe2010”.

**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 & Mr. Gaurav, who taught us the R and related subjects to understand the Analytics.

Thank you Acadgild!