LINEAR REGRESSION ANALYSIS

In this Assignment we got an opportunity to Understand Supervised Learning of Machine Language. And in this assignment I have taken couple of data sets one for Regression Analysis and the other for Classification.

Linear Regression Analysis:

The Dataset was downloaded from Internet and has been enclosed in the webpage for peers to experiment on.
The dataset taken for Linear Regression Assignment consists of data related to parameters of Used Cars. The data set contains Used Car Price, Car brands, year of manufacturing, engine volume and km run.

We want to predict the price of the used car on the basis of available data.

Potential 1st Regressor Variable: Brand of the Car.
We have AUDI, BMW, Benz, Skoda, Duster & python.
In general we know, BMW is expensive than Skoda or Duster or maruthi..

Potential 2nd Regressor Variable: km Driven.
More km run/driven car is relatively cheaper than the lesser driven car.

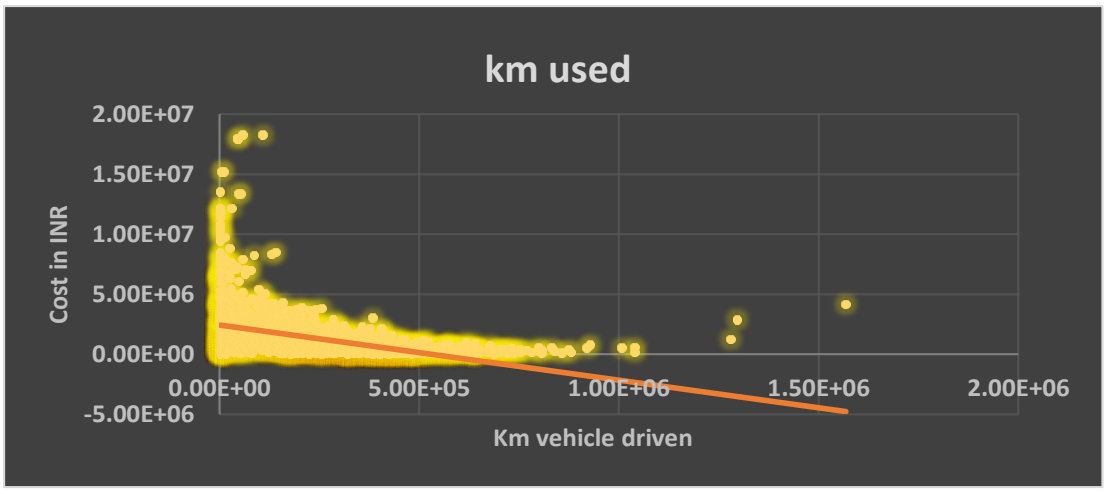
Potential 3rd Regressor Variable: Engine Volume.
More the engine volume the cost will be higher

Potential 4th Regressor Variable: Year of manufacturing.
Older the car Cheaper the price.

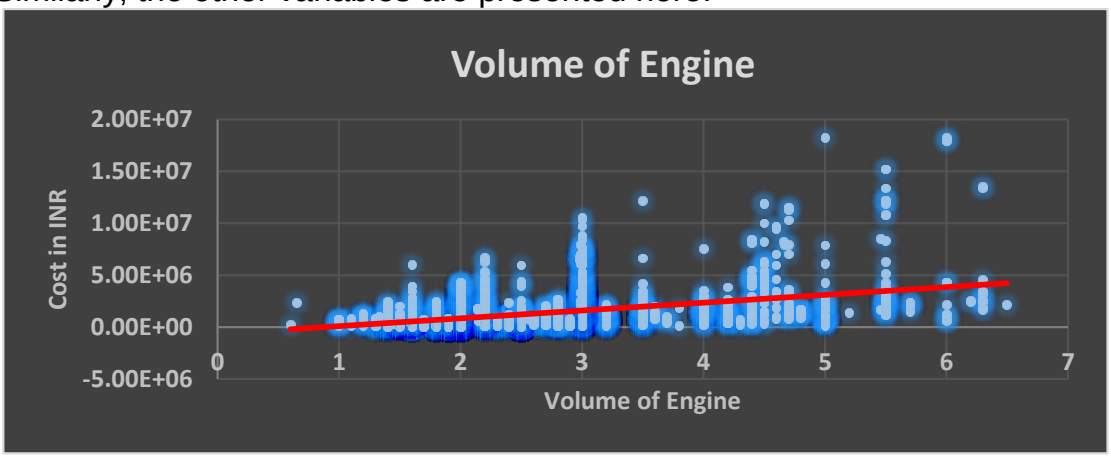
With this basis we start looking over training the model and predict the price of the car using Matlab Regression Analysis & also performed the same with Excel & Python.

Let's Start doing it.
The data used was already screened and null elements are removed to keep data intact. This data is imported on to excel.

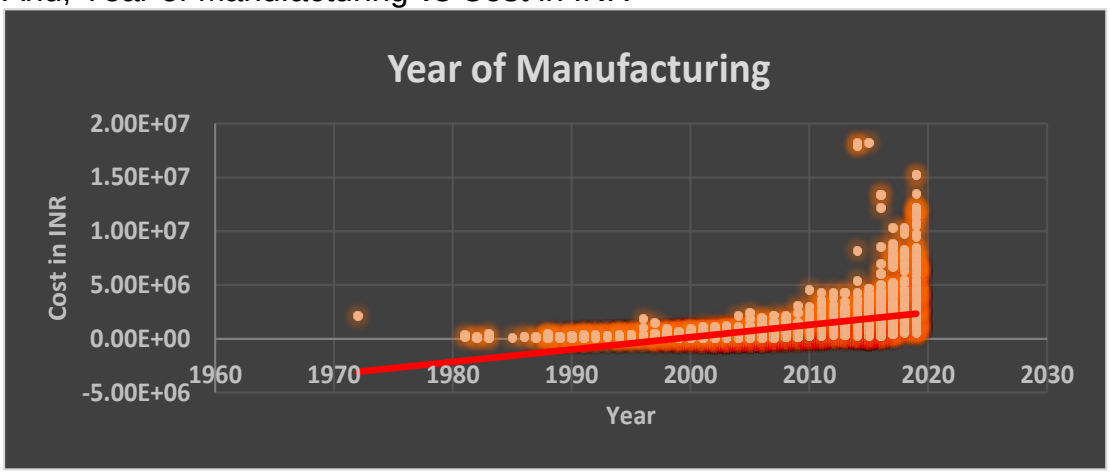
At first, scatter plot was done using Excel to visualize the linearity of data,
Cost of the Used car vs km run.



Similarly, the other variables are presented here.

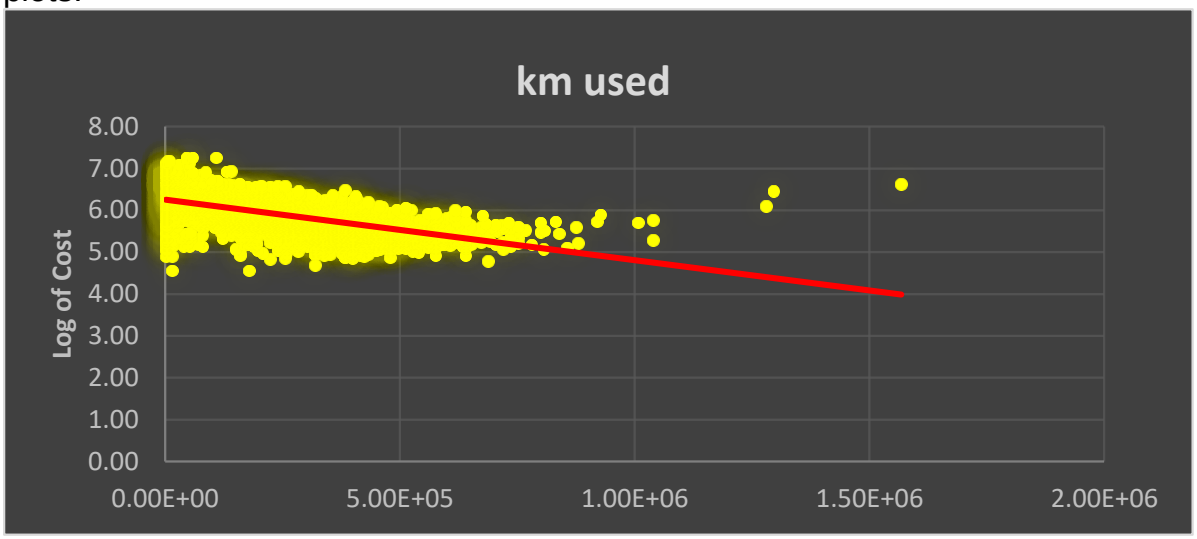


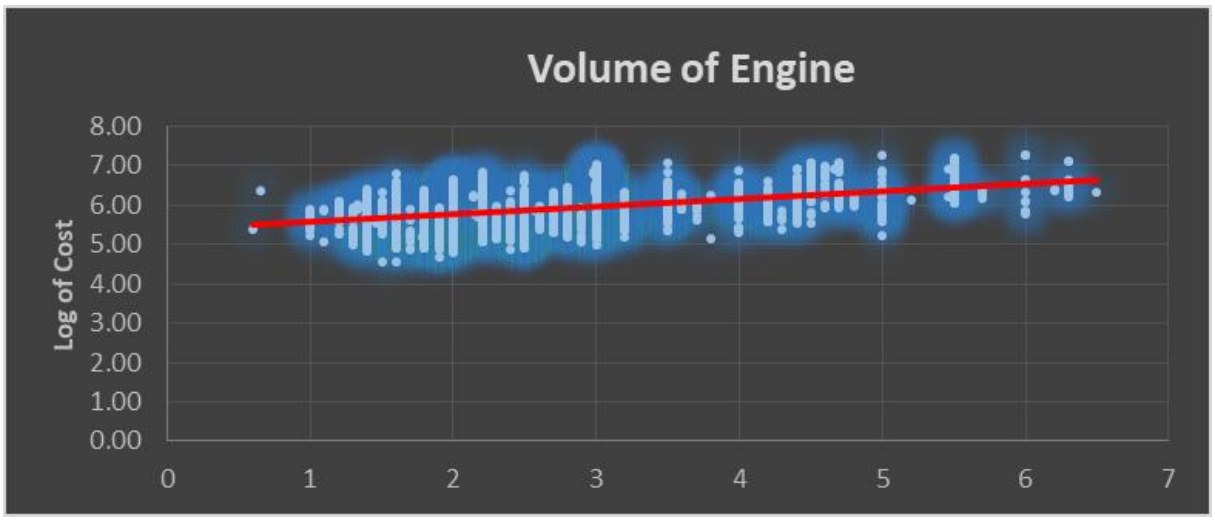
And, Year of Manufacturing vs Cost in INR

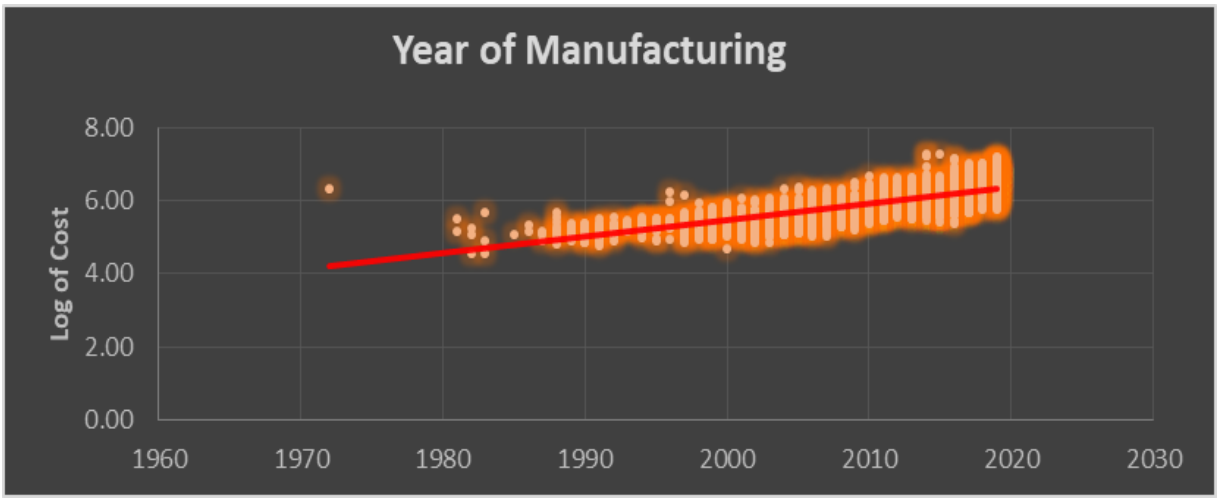


These are not linear enough to perform Linear regression on these data.

As the scatter plot are exponential, Log transformation is performed over these plots.







The data is more linearised in the Log transformed.

Now its time to check no endogeneity exception, by checking residuals with independent X.
And we can check normality and homoscedasticity, as log transformation is performed, homoscedasticity is also performed. And No autocorrelation is also established as these are not time series. Finally to check multicollinearity shall be checked by performing correlation check with probable combinations of variables.

Correlations			
	Km Driven	Year of Manufacturing	Engine volume
Km Driven	1	-0.50	-0.03
Year of Manufacturing	-0.50	1	0.04
Engine volume	-0.04	0.03	1

The correlations reveals that they are far from strong.
Now, we can perform Linear Regression for the set of variables.

At first,
Log of Cost is checked with Log of km run.

Regression analysis

Log of Cost and log km Run

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.52
R Square	0.27
Adjusted R Square	0.27
Standard Error	0.35
Observations	4003.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.00	178.24	178.24	1488.87	0.000
Residual	4001.00	478.98	0.12		
Total	4002.00	657.22			

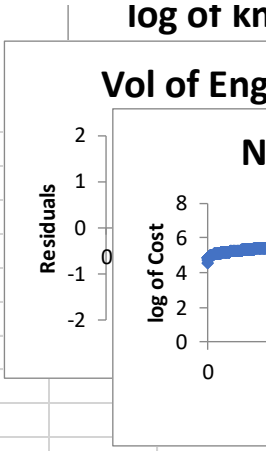
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.77	0.02	258.80	0.000
log Km	-0.34	0.01	-38.59	0.000

The model is significant as F is 0.00 and also the log of km run and Intercept. And, Adjusted R-Square is 0.27 only, which is not very perfect model.

Similarly, the R2 of Volume of Engine is 0.20

Regression analysis (km Run Check)

Log of Cost and Vol of Engine



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.44791208
R Square	0.20062523
Adjusted R Square	0.20040157
Standard Error	0.36702156
Observations	3576

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	120.8	120.8	896.99426	4.9E-176
Residual	3574	481.4	0.135		
Total	3575	602.3			

	Coefficients	*ndard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5.40427089	0.017	324.2	0	5.371585	5.436957	5.371585	5.436957
Vol of Engine	0.18853405	0.006	29.95	4.88E-176	0.176192	0.200876	0.176192	0.200876

Last option left is comparison with Year.

Regression analysis (km Run Check)

Log of Cost and Year of manufacturing

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.74733341
R Square	0.55850722
Adjusted R Square	0.55838369
Standard Error	0.27275842
Observations	3576

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	336.4	336.4	4521.2626	0
Residual	3574	265.9	0.074		
Total	3575	602.3			

	Coefficients	*ndard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-84.529948	1.344	-62.87	0	-87.1658	-81.894	-87.1658	-81.894
Year	0.04499093	7E-04	67.24	0	0.043679	0.046303	0.043679	0.046303

Here R2 is 56% which is very remarkable relationship.

Now it is time to take all three variables into the account, and regression is performed.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8753861
R Square	0.76630082
Adjusted R Square	0.76610455
Standard Error	0.19850274
Observations	3576

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	3	461.5	153.8	3904.2023	0
Residual	3572	140.7	0.039		
Total	3575	602.3			

	Coefficients	*ndard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-71.2605046	1.147	-62.13	0	-73.5092	-69.0118	-73.5092	-69.0118
log of km run	-0.12333625	0.006	-20.29	9.714E-87	-0.13525	-0.11142	-0.13525	-0.11142
Vol of Engine	0.17725103	0.003	52.02	0	0.17057	0.183932	0.17057	0.183932
Year	0.03848987	6E-04	68.39	0	0.037386	0.039593	0.037386	0.039593

The adjusted R square is 77% for the three variables. Almost it is the stronger model.

Now, we are including categorical variable of Brand names with binary coding.

In excel, it is done by creating a dummy variable and assigning 0 &1 to all variables except one of the set. This is done to avoid introducing of multicollinearity model.

The result of regression is as follows.

Regression Statistics	
Multiple R	0.89326114
R Square	0.79791547
Adjusted R Square	0.79746224
Standard Error	0.18471774
Observations	3576

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	8	480.6	60.07	1760.5037	0
Residual	3567	121.7	0.034		
Total	3575	602.3			

	Coefficients	*ndard E*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-76.349861	1.097	-69.62	0	-78.5001	-74.1997	-78.5001	-74.1997
log of km run	-0.10778739	0.006	-18.9	5.092E-76	-0.11897	-0.09661	-0.11897	-0.09661
Vol of Engine	0.13939054	0.004	38.15	2.18E-267	0.132227	0.146554	0.132227	0.146554
Year	0.04101346	5E-04	76.17	0	0.039958	0.042069	0.039958	0.042069
Audi_D	0.03770818	0.012	3.231	0.0012437	0.014828	0.060588	0.014828	0.060588
BMW_D	0.10836499	0.01	10.65	4.497E-26	0.088407	0.128323	0.088407	0.128323
Benz_D	0.10349928	0.01	10.37	7.355E-25	0.083938	0.123061	0.083938	0.123061
Duster_D	-0.16038924	0.011	-14.68	2.013E-47	-0.18181	-0.13897	-0.18181	-0.13897
skoda_D	0.0405992	0.011	3.798	0.0001485	0.019639	0.061559	0.019639	0.061559

Here, the model is significant and Adjusted R square is 80% which shows the very rigid model and very high explanatory model.
And even, each variable included in model is significant.

The model is,
Log of Cost = -76.35 – 0.108 x log of Km run + 0.139 x Vol of Engine + 0.041 x Year of manufacturing + 0.038 x Audi + 0.108 x BMW + 0.103 x Benz - 0.160 x Duster + 0.041 x skoda dummies.

Change of Log of cost = 0.139 for Change volume of engine.
Therefore, Log of Cost1 – log of cost2 = 0.139
Exponential of (log of cost1/cost2)= e^0.139
Cost1/Cost2 =1.139
Cost1 = 1.139 times Cost2

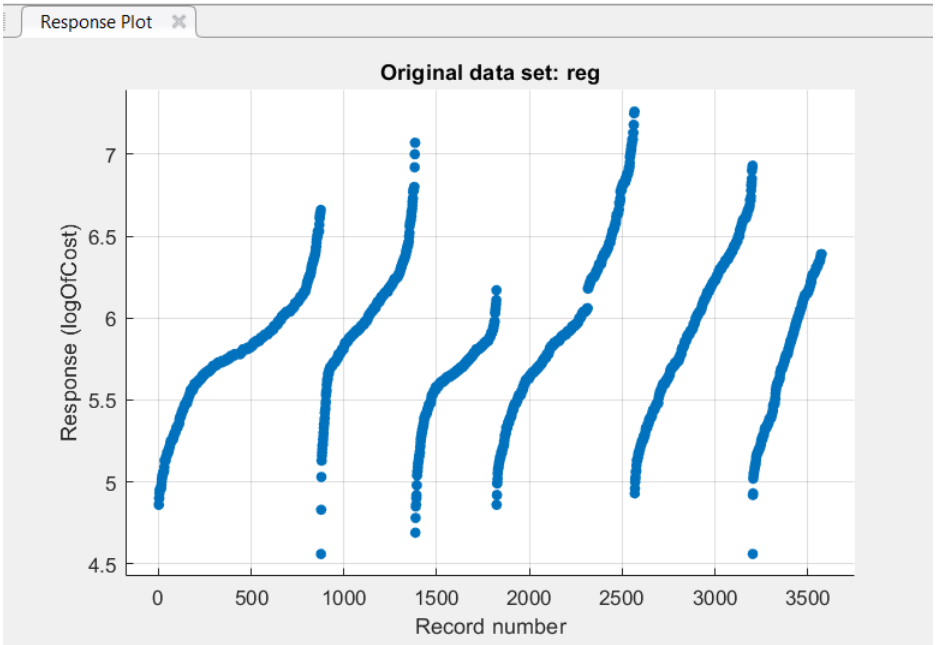
Log of km run coefficient -0.108 represents, that every increase in log of km run, then the log of cost will decrease by 0.108. Similarly, each increase in engine volume the log of cost will be increased by 0.139.

If all dummies are null, then,
Log of Cost = -76.35 – 0.108 x log of Km run + 0.139 x Vol of Engine + 0.041 x Year of manufacturing
Which is the cost of the car Maruthi, which is the benchmark considered in the analysis.
All the brands of the cars are evaluated keeping Maruthi as a benchmark brand.
This shows,
That,
BMW is 11% higher than Maruthi,
Benz is 10% higher than Maruthi,
Audi is 3.8 % higher than Maruthi,
Skoda is 4% higher than Maruthi,
Duster is 16% lower cost than Maruthi.

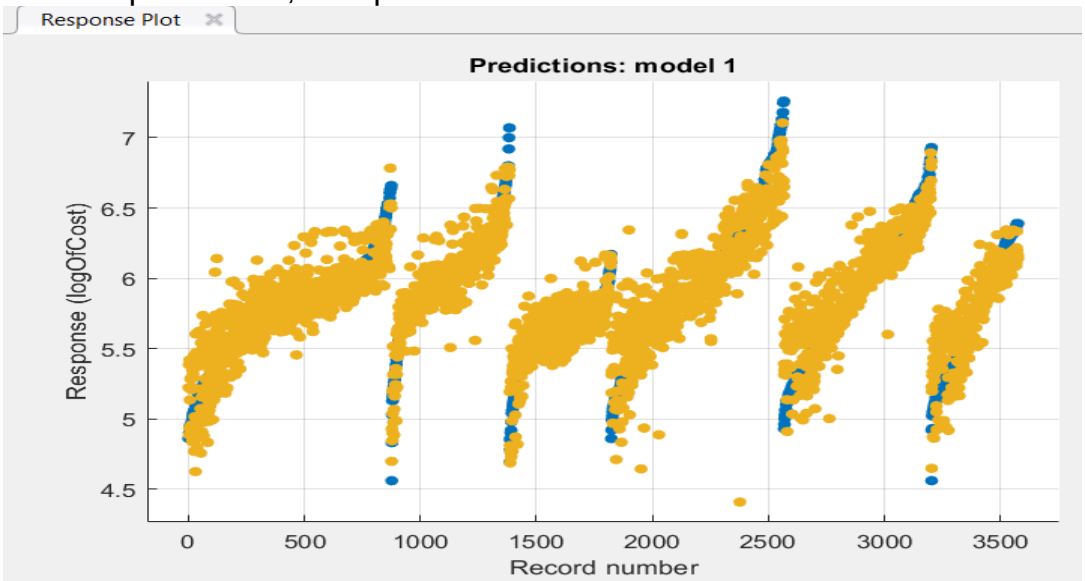
This is how the linear regression helps in understanding combination of things in a better way.

Now the same problem is attempted with Matlab.

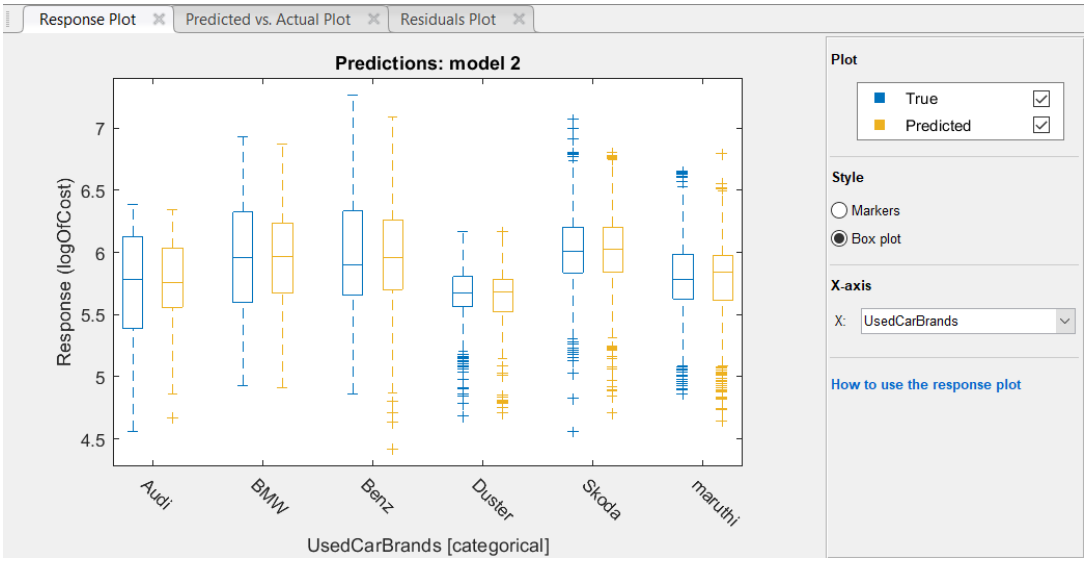
In Matlab, using the regression app, the linear regression are performed at ease.
For the same logarithmic data set, we have performed Linear Regression and images and results are presented here.
Original response plot,



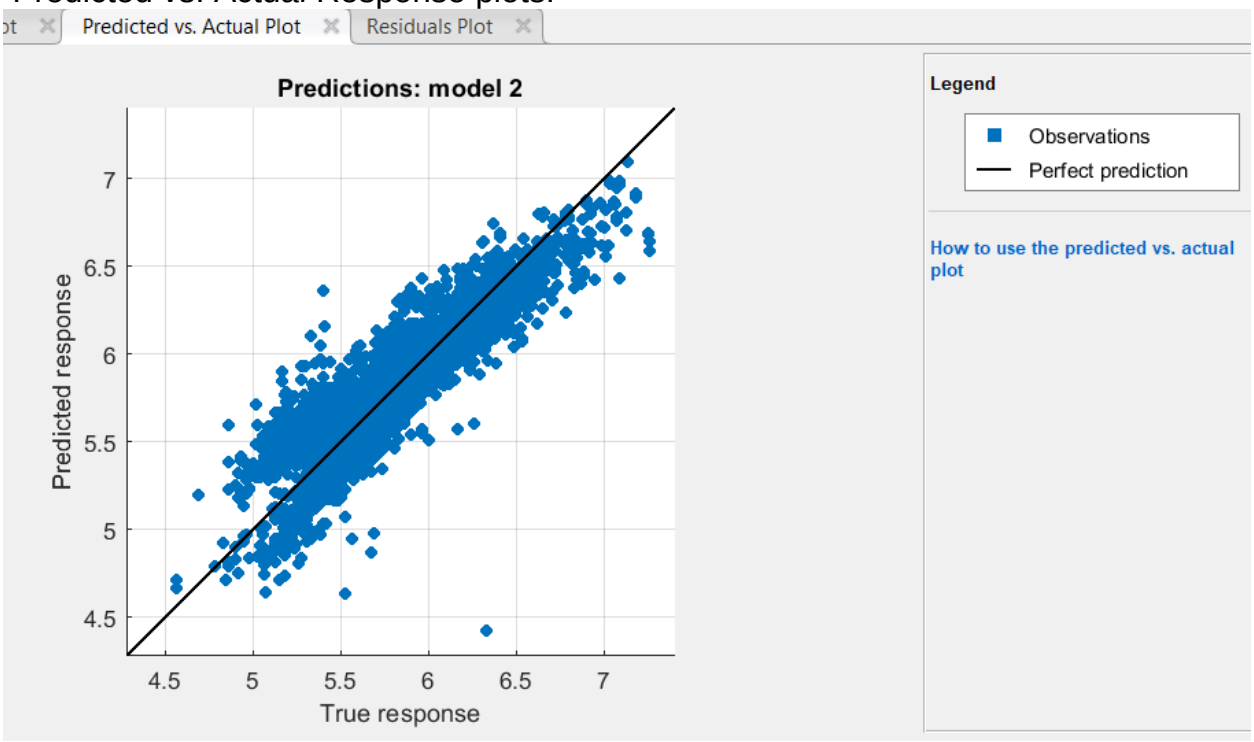
The Response Plot, with predicted and true values



Response Plot of log of Cost vs Car brands, with actual and predicted data representation.



Predicted vs. Actual Response plots.



And the Linear coefficients are

trainedModel5.LinearModel.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	-76.2926	1.0954	-69.6492	0
2 UsedCarBrands_BMW	0.0707	0.0121	5.8200	6.4051e-09
3 UsedCarBrands_Benz	0.0658	0.0119	5.5297	3.4374e-08
4 UsedCarBrands_Duster	-0.1984	0.0136	-14.6146	5.0528e-47
5 UsedCarBrands_Skoda	0.0027	0.0128	0.2127	0.8316
6 UsedCarBrands_maruthi	-0.0379	0.0117	-3.2493	0.0012
7 logOfKmRun	-0.1078	0.0057	-18.9022	5.0552e-76
8 VolOfEngine	0.1393	0.0037	38.1287	3.9688e-267
9 Year	0.0410	5.3838e-04	76.1614	0

Intercept, log of km run, vol of engine and everything is in-line with the Excel model.

Further, the same problem have been checked with python using Google Colab and the same has been enclosed in the assignment webpage and all three way of calculation is in conformance with each other.