FINAL PROJECT SUBMISSION

Data Science using MS-Excel

OCTOBER 21, 2018

Contents

Problem Statement

Perform the Simple Regression Analysis on Fuel Economy Data using Excel.

Methods used

We performed the simple regression using both Manual computation and Data Analysis in Excel.

Approach

- During the 1st approach, the Excel sheet 2010 was used as train data to build the model during the first iteration of the exercise. The equation was tested on the excel sheet 2011 to validate the fitment.
- During the second iteration, the input sheet was split using a random function and the generated equation was tested and validated on the 2011 sheet.

First Approach

The whole of the 2010 data was used as a train data.

Using the Data Analysis

We chose the variables of interest by adhering to Hypothesis testing principles.

Ho: There is no relationship between the Fuel economy and the variable and slope is zero.

Ha: The slope is a non-zero number and there is relationship between the Fuel economy (Dependent variable) and the independent variables.

The first iteration of regression analysis depicted the Engine Displacement, Number of cylinders, Transmission Lockup and Variable Valve timing exhibiting the relatively stronger significance as the corresponding p-value observed was less than p-value of 5%.

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	54.34722	1.097261	49.52991	5.0196E-282
EngDispl	-3.86097	0.280483	-13.7654	6.91523E-40
NumCyl	-0.4888	0.18453	-2.64891	0.008191264
NumGear:	-0.17252	0.106502	-1.61985	0.105552505
TransLock	-1.44499	0.299981	-4.81694	1.66227E-06
TransCree	-0.91375	0.668076	-1.36774	0.171674472
IntakeVal	-0.37372	0.98923	-0.37779	0.705658016
ExhaustVa	-1.1105	0.959809	-1.157	0.247523127
VarValveT	1.687012	0.379592	4.444277	9.71335E-06
VarValveL	0.623536	0.371933	1.676472	0.093930612

The second iteration of simple regression method resulted into the below output indicating a significant relationship with Engine displacement among the rest of the independent variables.

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	50.736815	0.552	91.915	0
EngDispl	-3.45210302	0.25058	-13.777	5.8819E-40
NumCyl	-0.69775192	0.17126	-4.0742	4.9475E-05
TransLockup	-1.45369601	0.30001	-4.8455	1.4434E-06
VarValveTiming	1.4980054	0.36234	4.13426	3.8312E-05

This helped us conclude the best fit line of **y = -4.5209x+50.563** intercept and slope.

Using the Excel Function

We plotted the scatter plot between Fuel economy and individual independent variable and got to see the highest co-relation between the Engine Displacement compared with others. We plotted the trend line and displaced the equation on graph. We see the similar nos. populated using the Regression Analysis output using Data Analysis pack of the Excel.

Determining R-Squared

Later on, we populated the table to find the SSE (Sum Squared Error), SSR (Sum Squared Residual) and the SST (Sum Squared Total).This helped us find the R-squared value (about 62%) which signifies the variation in Fuel Economy, which is attributed to the Engine Displacement.

Second Approach

We used a random function in Excel to split the data into 70:30 ratio and build the equation of using this partitioned data. We tested the equation on the test data.

Output after 1st iteration:

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	53.95883909	1.345975	40.08903	1.1219E-189
EngDispl	-3.646417913	0.346582	-10.5211	2.92312E-24
NumCyl	-0.609946569	0.228981	-2.66374	0.007891353
NumGears	-0.131006684	0.138917	-0.94306	0.34595002
TransLockup	-1.600930855	0.374021	-4.28033	2.10487E-05
TransCreeperGear	-1.101362706	0.800568	-1.37573	0.169311829
IntakeValvePerCyl	0.007682001	1.252615	0.006133	0.995108396
ExhaustValvesPerCyl	-1.083039741	1.21196	-0.89363	0.371804491
VarValveTiming	1.279559551	0.468906	2.728821	0.006503013
VarValveLift	0.652365684	0.463709	1.406844	0.159882578

Output after 2nd iteration:

The output below re-established the relatively stronger relation between the engine displacement and the Fuel Economy.

ANOVA					
	df	*SS*	*MS*	*F*	*g*
Regression	4	27784.1	6946.02	322.710272	
Residual	765	16465.9	21.524		
Total	769	44249.9			
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lc*
Intercept	51.2969866	0.68112	75.3123	0	
EngDispl	-3.37619441	0.30703	-10.996	3.24E-26	
NumCyl	-0.74453925	0.21007	-3.5442	0.00041767	
TransLockup	-1.63809259	0.37165	-4.4077	1.1943E-05	
VarValveTiming	1.17208706	0.44385	2.64072	0.00844172	

We replicated the same steps as of the first approach.

Validation on 2011 Excel sheet

The best-fit line equation was later validated on the Excel 2011 data to gauge the accuracy of the established regression equation. We could see a decent level of R-squared value with this projected equation in both the approaches.

Brief Findings

We followed both "Manual Excel Functions" and "Data Analysis Tab" function to establish the Simple Linear Regression Model. We used the "train, test and validate" approach in both the approaches using 2010 and 2011 excel sheets. The Hypothesis Testing concept was explored to gauge the best-fitted variable among the list of variables. This involved the iterative running of Simple Regression model to get the best-related independent variable. The R-squared computation was also performed to ensure a decent level of explanation in variation is exhibited by the independent variable from the regression line. The equation thus, can be used as a good prediction tool for Fuel Economy for the given set of independent variable, especially the Engine Displacement.