Assignment 1: Linear Regression Using SAS Enterprise Miner

Daanish Ahmed

DATA 640 9041

Fall 2017

dsahmed2334@yahoo.com

Professor Sounak Chakraborty

UMUC

July 31, 2017

Introduction

Linear regression is a powerful tool for building predictive models. It involves fitting a mathematical equation to a dataset with a continuous target and using the outcome to predict the value of the target variable. It is important to limit the number of inputs used in these models, because using too many can produce a weaker model. To handle this issue, there are three variable selection methods for filtering out insignificant variables: forward selection, stepwise selection, and backward elimination. In this assignment, I will implement each of these methods by creating several linear regression models using SAS Enterprise Miner for a dataset containing vehicle fuel economy data (Kuhn & Johnson, 2014). My goal is to produce a model that can accurately predict the value of a vehicle's fuel economy. I will create one model for each of the three variable selection methods, as well as one that keeps all inputs in the model. Next, I will build a custom model by manually removing variables that I find to be unimportant. Finally, I will recreate the first four models described, but this time using R-squared values to remove insignificant inputs prior to model creation (see Figure 1 in Appendix). I will then compare all nine models to determine how the removal of certain variables can impact the model's accuracy.

The dataset used for this assignment contains 15 columns and 1107 observations (see Figure 2 in Appendix). Some of the possible inputs include the engine displacement, transmission, number of cylinders, number of gears, air aspiration method, intake valves per cylinder, exhaust valves per cylinder, and car line class description (such as two-seater, subcompact, van, etc.). The target variable is fuel economy (FE), which is a continuous numeric variable. The ID variable (labeled as "VAR1") is not useful for the analysis and was therefore rejected, leaving us with 14 variables remaining. Four of these variables are categorical, while the rest are numeric. Two of the categorical variables—transmission and car line class description—have more than ten levels (see Figure 3 in Appendix). Creating dummy variables for these cases may result in a significantly more complicated model. I plan on addressing this issue by using R-squared values to remove the least important inputs before building my last four regression models. Out of the numeric variables, there are four binary, two interval, and four nominal variables (see Figure 3 in Appendix). To reduce the number of dummy variables in my model, I set all four nominal numeric variables to the interval type. According to the descriptive statistics, there are no missing values in any of the variables (see Figure 3 in Appendix). The figure also reveals that the engine display and fuel economy (FE) variables have skewness values of 0.65 and 0.60 respectively. These skewness values seem relatively low, but I will examine the distributions of each variable to determine if it is necessary to apply any transformations to them.

Data Preparation

SAS Enterprise Miner uses the SEMMA process (Sample, Explore, Modify, Model, Assess) to allow users to effectively develop predictive models. As such, I followed this process during the development of my linear regression models (see Figure 1 in Appendix). I used nodes for sampling (Data Partition), exploring (Graph Explore, StatExplore, Variable Selection), modification (Replacement, Transform Variables), modeling (Regression), and assessment (Model Comparison). My dataset was first imported as a .csv file into SAS Enterprise Miner using the File Import node, and it was converted into a .sas7bdat file using the Save Data node. I then used the StatExplore node to generate a bar plot showing the worth of each input variable (see Figure 4 in Appendix). According to this image, the four most important variables are engine displacement, number of cylinders, drive description, and car line class description. Likewise, the least significant variables are transmission creeper gear, valve lift, air aspiration method, valve timing, number of gears, transmission lockup, and intake valves per cylinder. Since these variables have the lowest impact towards the target variable, they will be excluded when I build my custom regression model that manually excludes unimportant inputs.

I then used the Graph Explore node to examine the distribution of the dependent variable "fuel economy" (see Figure 5 in Appendix). As we see, the variable has a mostly normal distribution. However, one of the assumptions of linear regression is that all input variables must be normally distributed (Abrams, n.d.). Thus, I proceeded to examine all variable distributions in the model (see Figure 6 in Appendix). Based on this image, we see that engine displacement has a slight positive skew, while car line class description and intake valves per cylinder have spikes in their distributions. Therefore, my next step is to apply transformations using the Transform Variables node. According to Abrams (n.d.), transformations are useful not only for normalizing data but also for handling outliers and nonlinearity between variables. For interval inputs, I selected "maximum normal" as the default method. According to SAS Enterprise Miner's reference help page, this method performs multiple types of transformations on each variable and selects the best transformation for maximizing that variable's normality. This allows us to easily handle skewness in our dataset without having to manually choose a transformation for each individual variable. For class inputs, I selected "dummy indicators" as the default method. Creating dummy variables is not only helpful for handling distributions with spikes, but it can also be used to convert non-numeric variables into numeric form (Knode, 2016).

Next, I sampled the dataset by splitting it using the Data Partition node. I allocated 60% of data to the training set, 20% to the validation set, and 20% to the test set. As mentioned earlier, there are no missing values in any of the variables. Therefore, it is not necessary to impute or replace such values. However, I chose to ensure that there are no more outliers in my data by using the Replacement node. I set the interval variable default limits method to "standard deviations from the mean" and used a cutoff value of 3.0 to eliminate all values that are more than 3 standard deviations away from the mean. Finally, I addressed the possibility of the variables "exhaust valves per cylinder" and "intake valves per cylinder" being correlated. By examining the dataset, I found that these variables almost always have the same value. Having highly correlated input variables will cause the model to experience multicollinearity, which will impact the quality of the model unless we remove all but one of the correlated variables (Abrams, n.d.). Since "intake valves per cylinder" has a lower predictive worth (see Figure 4 in Appendix), I decided to exclude this variable from every linear regression model that I would build.

Model Development

With data preparation complete, it is now possible to create linear regression models. As stated earlier, I will begin by creating one model for forward selection, one for backward elimination, and one for stepwise selection. According to Knode (2016), forward selection involves starting with no variables in the model and adding them one at a time according to their significance level until satisfying the stopping criterion. He describes backward elimination as including all variables in the model and removing the least significant variable during every iteration until reaching the stopping criterion. Finally, stepwise selection begins with no variables, like forward selection, but it allows for both the removal and addition of variables during each iteration (Knode, 2016). I built these models using the Regression node, setting the regression type to "linear regression" and selection criterion to "validation error" while choosing the selection models "forward," "backward," and "stepwise" for their respective models. According to SAS Enterprise Miner Reference Help, validation error involves choosing the model with the lowest error rate in the validation set. Once these models have been created, we can examine their t-test results (see Figures 7, 9, and 11 in Appendix). These figures provide us with useful information such as the intercepts and coefficients for each variable, all of which can be used to build the linear regression equation. Another important statistic is the p-value, which can be used to determine a variable's significance. Variables that have p-values above 0.05 are not statistically significant and should ideally be removed from the model (Frost, 2013). I will explore these results further in the "results" section of this paper.

Next, I will build a regression model that preserves all input variables. Due to potential issues with having too many inputs in the model, I expect that this model will have a higher error than most other models. This model was also built using the Regression node, but the main difference is that I set the selection model to "none." I then created a fifth model, in which I manually excluded input variables deemed unimportant. When using the StatExplore node during the data preparation phase, I found that the variables with the lowest worth are transmission creeper gear, valve lift, air aspiration method, valve timing, number of gears, transmission lockup, and intake valves per cylinder (see Figure 4 in Appendix). Thus, I removed these variables from the model by clicking "edit variables" and setting their usage to "no." This model functions similarly to the previous one in that neither uses a variable selection method; the only difference is that this model excludes insignificant inputs instead of keeping all variables in the model.

The final set of models that I will build are similar to my first four models (forward, backward, stepwise, and no selection). However, the difference is that I will use R-squared values to remove insignificant inputs before creating the models. R-squared is used to measure how close the data points are to the regression line, in which a higher R-squared value often corresponds to a better fit on the data points (Frost, 2013). I implemented this technique by using the Variable

Selection node and setting the target model to “R-Square.” According to SAS Enterprise Miner’s reference help page, this selection method will use a forward stepwise least squares regression to ensure the highest possible R-squared value. Next, I set the algorithm to bypass interval variables during variable selection while choosing to keep them in the model. This is meant to prevent the algorithm from removing these variables. I also set the maximum number of variables to 30 to prevent the models from containing too many inputs. Finally, I set the “stop R-Square” option to 0 to prevent the algorithm from eliminating too many variables. After this step, I duplicated the existing “forward,” “backward,” “stepwise,” and “no selection” model nodes and connected them to the Transform Variables node. As such, these last four regression models will behave identically to the first four models—the only difference is that the newest models will perform R-squared variable selection prior to the model’s execution.

Results

I will now evaluate the important results and accuracy measures of each model to determine which ones have the lowest error. First, I will examine the t-test results of the first three models. One of the most useful metrics is the p-value which shows each variable’s significance towards predicting the dependent variable. Another important statistic is the regression coefficient, which determines the strength of a variable’s relationship to the dependent variable and whether that relationship is positive or negative (Frost, 2013). For the first forward selection model, we see that all 22 remaining inputs have p-values under 0.05, which means that all of them are significant (see Figure 7 in Appendix). We also see that the number of cylinders and engine displacement variables have the strongest regression coefficients at -19.5 and -19.9 respectively. This indicates that both variables have a strong negative relationship with fuel economy. For the first backward regression model, we see that there are 32 remaining inputs and five of them have p-values greater than 0.05 (see Figure 9 in Appendix). This model not only has more inputs than the previous one, but it also contains insignificant variables that have yet to be removed by the backward elimination model. Once again, the number of cylinders and engine displacement have some of the strongest regression coefficients at -17.2 and -20.3 respectively. Number of gears is the strongest variable in this model, with a coefficient of -25.7. However, this variable was not present in the previous model—most likely meaning that it was not added during the forward selection process.

For the first stepwise model, we see that the 22 remaining inputs all have p-values less than 0.05 (see Figure 11 in Appendix). We also note that the variables with the strongest coefficients are the number of cylinders and engine displacement with -19.5 and -19.9 respectively. It is not surprising that engine displacement has the strongest impact on the model, since the dataset description suggests a strong relationship between fuel economy and engine displacement (Kuhn & Johnson, n.d.). Interestingly, the t-test results for the current stepwise model are identical to those of the first forward regression model. This suggests that no variables were removed during the stepwise selection process, causing the model to behave like a forward selection model.

Another useful metric is the f-test results, which shows how well the model fits the data and allows for multiple coefficients to be assessed at the same time (Frost, 2013). By examining the f-test results of the first three models (see Figures 8, 10, and 12 in Appendix), we see that although the f-values differ from the t-values of the t-test results, the p-values are identical in each model. Since nearly all variables have p-values under 0.05, we can reject the null hypothesis and verify that these models have better fits than a model with zero predictors (Frost, 2013).

However, if we want to examine the accuracy of the entire model rather than individual variables, we need to look at the adjusted R-squared value. Adjusted R-squared measures the distance between the regression line and the data points, but it is different from normal R-squared in that it can evaluate models with different numbers of input variables (Frost, 2013). By examining our results, we find that the adjusted R-squared values for the forward, backward, and stepwise models are 80.32%, 81.23%, and 80.32% (see Figures 8, 10, and 12 in Appendix). Furthermore, the adjusted R-squared value for the model that keeps all variables is 81.1%, while the model with manually-removed variables has an adjusted R-squared of 79.69% (see Figures 13 and 14 in Appendix). From these figures alone, we find that the first backward regression model has the highest R-squared and therefore has the closest fit to the data, while the model with manually-removed inputs has the worst fit. If we examine the four models that had variable selection prior to model creation, then we find that their adjusted R-squared values are significantly worse. The values for the second forward, backward, stepwise, and "no selection" models are 75.63%, 75.60%, 75.63%, and 75.60% (see Figures 15, 16, 17, and 18 in Appendix).

Finally, we can compare the mean squared error of all nine models to determine which model has the lowest error. I used the Model Comparison node to compare the models, setting the selection statistic to "mean squared error." According to SAS Enterprise Miner Reference Help, mean squared error is useful for measuring the error of linear models such as regression. By looking at the error rates for my first five models, we see that the model which keeps all inputs has the highest validation and test error at 12.03 and 13.66 (see Figure 19 in Appendix). This is not surprising, since having more inputs often produces weaker predictive models. What is surprising is that this model also has the second-highest adjusted R-squared value (81.1%). Figure 19 also reveals that the model with custom variable selection has the lowest validation and test error at 8.42 and 12.50. Interestingly, this model also has the lowest adjusted R-squared out of the five original models (79.69%). Finally, by examining the final four models—all of which involved variable selection prior to model building—we find that their test errors (15.85 to 16.13) are by far the highest while their validation errors (9.89 to 10.27) are somewhere in between (see Figure 20 in Appendix). It is interesting to note that although each of my models have higher test errors than training errors, the error rates are still relatively close together and there does not appear to be any overfitting. In fact, most of my models have validation errors that are lower than either the training or test errors. And by looking at the score rankings plot of the first five models for validation and test data (see Figure 21 in Appendix), we see that these models have nearly identical plots. This may suggest that even the weakest of these models is still relatively accurate.

Conclusion

In this analysis, I developed nine distinct linear regression models to predict the value of a vehicle's fuel economy. Through the exploration of p-values and regression coefficients, I found that the most significant predictors are the engine displacement and number of cylinders—both of which have a strong negative relationship with fuel economy. By examining the adjusted R-squared and mean squared errors, it is evident that the weakest models are those which underwent variable selection prior to model implementation. Although these four models have decent mean squared errors for validation data (9.89 to 10.27), this may be explained by selecting the "validation error" selection criterion for the Regression nodes. The significantly worse test errors (15.85 to 16.13) and R-squared values (75.60-75.63%) make it clear that these models provide a weaker fit to the data. Thus, it appears that performing variable selection prior to model creation is unnecessary since several of the regression models already use variable selection methods such as forward, backward, or stepwise selection. This may lead to the removal of many important input variables, weakening the model's ability to predict the dependent variable.

When excluding these four models, the regression model with the next highest error is the one that uses no form of variable selection at all. This model has a validation error of 12.03 and a test error of 13.66. Since this model has the largest number of inputs, it is not surprising that its error rate is higher than that of other models. However, its adjusted R-squared value is 81.1%, which is the second highest of all the models. I would therefore recommend further research to understand why this is the case, since it is unclear whether this model is truly weaker than the others. One possibility may be that the dataset is relatively optimized for linear regression and thus requires minimal variable selection. This is evidenced by the score rankings plot (see Figure 21 in Appendix), which suggests that even the weakest of my models is still fairly accurate. Another area for future research is the fact that the forward and stepwise selection models have identical results. It is possible that these two models behaved similarly because there were no cases where variables had to be removed during stepwise selection.

One final area for future research is to understand why my custom model—which manually excluded irrelevant variables—has lower validation and test errors (8.42 and 12.50) than any other model. Although this model excluded some variables, it did not use any variable selection method. Additionally, its adjusted R-squared value is 79.69%, which is far from the highest value. Ultimately, it is too early to conclude that my custom model will always produce better results than the forward, backward, or stepwise methods. I may experiment with using similar models with different datasets to see if I can produce comparable results. Furthermore, I may try combining the most effective models by using either a forward or stepwise model while manually excluding only a handful of insignificant variables. Although there is currently no concrete answer for which model is really the "best," I believe that further analysis will either provide a clear answer to this question or yield an even more effective model than those used in my analysis.

References

Abrams, D. R. (n.d.). Introduction to Regression. Retrieved October 3, 2017, from http://dss.princeton.edu/online_help/analysis/regression_intro.htm

Frost, J. (2013, December 12). Regression Analysis Tutorial and Examples. Retrieved October 4, 2017, from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples

Knode, S. (2016, August 19). *Regression Models*. Lecture presented at UMUC. Retrieved October 3, 2017.

Kuhn, M., & Johnson, K. (2014, July 25). AppliedPredictiveModeling: Functions and Data Sets for 'Applied Predictive Modeling'. Retrieved October 6, 2017, from https://cran.r-project.org/web/packages/AppliedPredictiveModeling/index.html

Kuhn, M., & Johnson, K. (n.d.). Applied Predictive Modeling. Retrieved October 6, 2017, from http://appliedpredictivemodeling.com/data/

Appendix

Relevant SAS Enterprise Miner Output Images

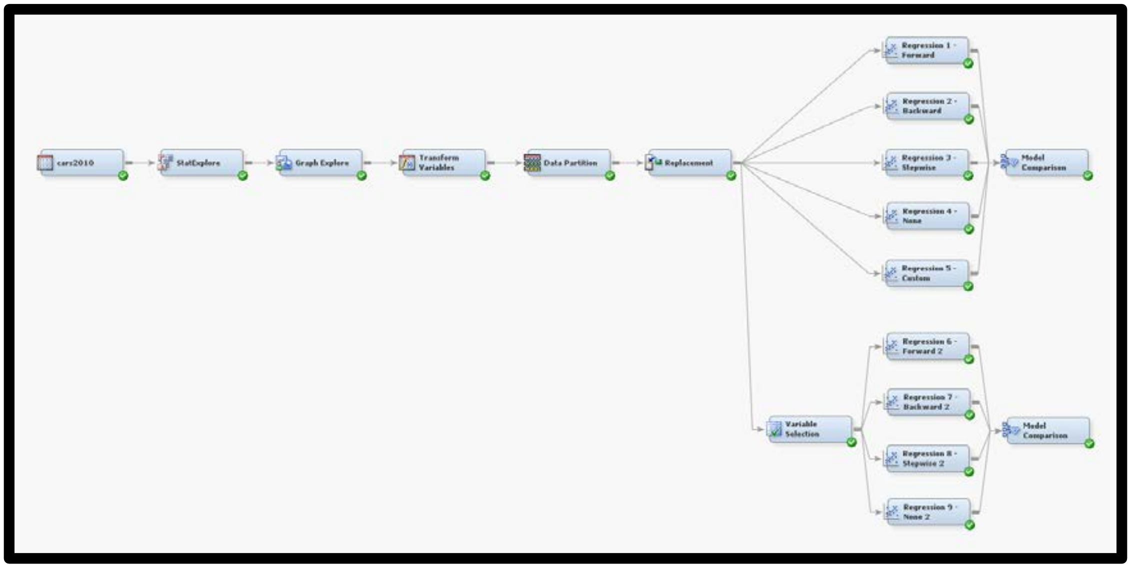
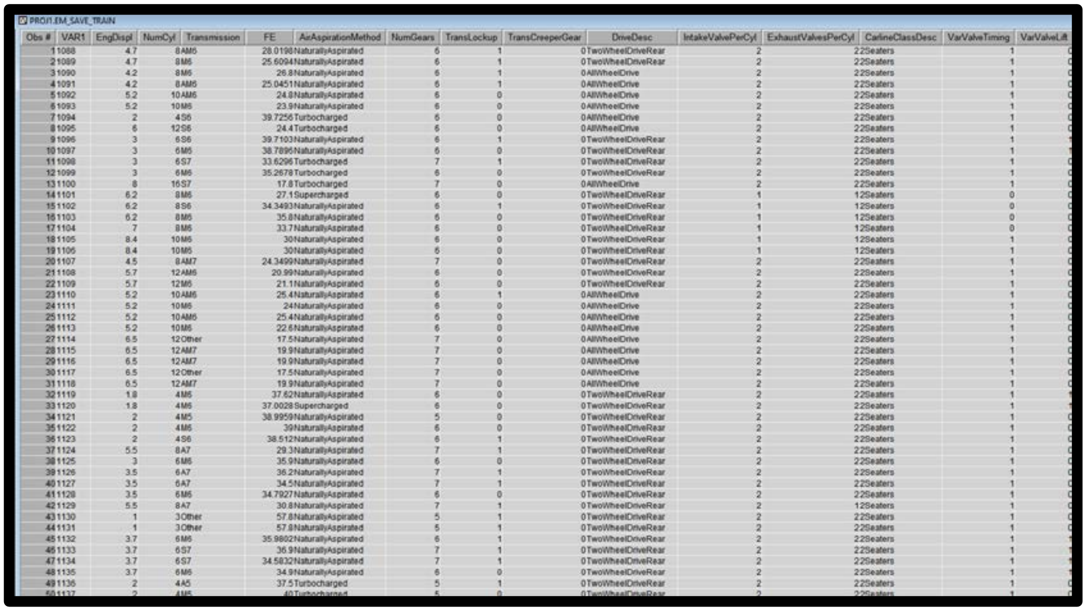
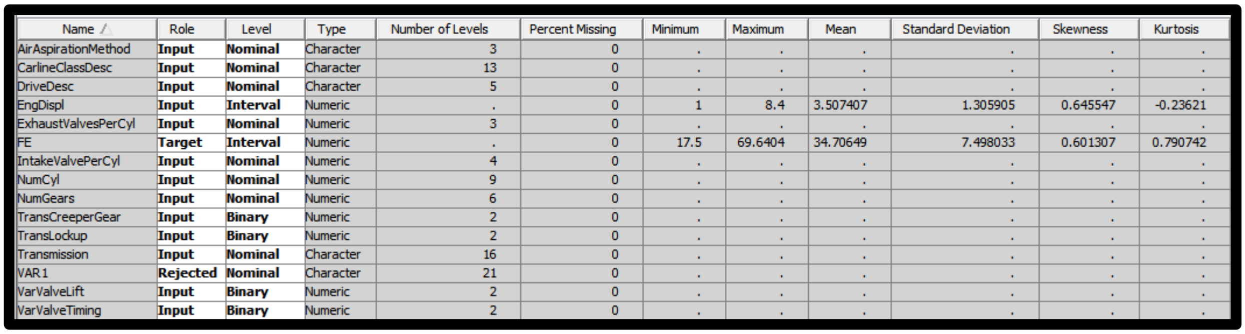


Figure 1. Process Flow Diagram for "Cars2010" Linear Regression Models.


PROJ1.EM_SAVE_TRAIN

Obs #	VAR1	EngDispl	NumCyl	Transmission	FE	AirAspirationMethod	NumGears	TransLockup	TransCreeperGear	DriveDesc	IntakeValvePerCyl	ExhaustValvesPerCyl	CarlineClassDesc	VarValveTiming	VarValveLift
1	1088	4.7	8	AM6	28.0198	NaturallyAspirated	6	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
2	1089	4.7	8	M6	25.6094	NaturallyAspirated	6	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
3	1090	4.2	8	M6	26.8	NaturallyAspirated	6	1	0	AllWheelDrive	2	2	2Seaters	1	0
4	1091	4.2	8	AM6	25.0451	NaturallyAspirated	6	1	0	AllWheelDrive	2	2	2Seaters	1	0
5	1092	5.2	10	AM6	24.8	NaturallyAspirated	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
6	1093	5.2	10	M6	23.9	NaturallyAspirated	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
7	1094	2	4	S6	39.7256	Turbocharged	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
8	1095	6	12	S6	24.4	Turbocharged	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
9	1096	3	6	S6	39.7103	NaturallyAspirated	6	1	0	TwoWheelDriveRear	2	2	2Seaters	1	1
10	1097	3	6	M6	38.7896	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	1
11	1098	3	6	S7	33.6296	Turbocharged	7	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
12	1099	3	6	M6	35.2678	Turbocharged	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
13	1100	8	16	S7	17.8	Turbocharged	7	0	0	AllWheelDrive	2	2	2Seaters	1	0
14	1101	6.2	8	M6	27.1	Supercharged	6	0	0	TwoWheelDriveRear	1	1	2Seaters	0	0
15	1102	6.2	8	S6	34.3493	NaturallyAspirated	6	1	0	TwoWheelDriveRear	1	1	2Seaters	0	0
16	1103	6.2	8	M6	35.8	NaturallyAspirated	6	0	0	TwoWheelDriveRear	1	1	2Seaters	0	0
17	1104	7	8	M6	33.7	NaturallyAspirated	6	0	0	TwoWheelDriveRear	1	1	2Seaters	0	0
18	1105	8.4	10	M6	30	NaturallyAspirated	6	0	0	TwoWheelDriveRear	1	1	2Seaters	1	0
19	1106	8.4	10	M6	30	NaturallyAspirated	6	0	0	TwoWheelDriveRear	1	1	2Seaters	1	0
20	1107	4.5	8	AM7	24.3499	NaturallyAspirated	7	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
21	1108	5.7	12	AM6	20.99	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
22	1109	5.7	12	M6	21.1	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
23	1110	5.2	10	AM6	25.4	NaturallyAspirated	6	1	0	AllWheelDrive	2	2	2Seaters	1	0
24	1111	5.2	10	M6	24	NaturallyAspirated	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
25	1112	5.2	10	AM6	25.4	NaturallyAspirated	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
26	1113	5.2	10	M6	22.6	NaturallyAspirated	6	0	0	AllWheelDrive	2	2	2Seaters	1	0
27	1114	6.5	12	Other	17.5	NaturallyAspirated	7	0	0	AllWheelDrive	2	2	2Seaters	1	0
28	1115	6.5	12	AM7	19.9	NaturallyAspirated	7	0	0	AllWheelDrive	2	2	2Seaters	1	0
29	1116	6.5	12	AM7	19.9	NaturallyAspirated	7	0	0	AllWheelDrive	2	2	2Seaters	1	0
30	1117	6.5	12	Other	17.5	NaturallyAspirated	7	0	0	AllWheelDrive	2	2	2Seaters	1	0
31	1118	6.5	12	AM7	19.9	NaturallyAspirated	7	0	0	AllWheelDrive	2	2	2Seaters	1	0
32	1119	1.8	4	M6	37.62	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	1
33	1120	1.8	4	M6	37.0028	Supercharged	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	1
34	1121	2	4	M5	38.9959	NaturallyAspirated	5	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
35	1122	2	4	M6	39	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
36	1123	2	4	S6	38.512	NaturallyAspirated	6	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
37	1124	5.5	8	A7	29.3	NaturallyAspirated	7	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
38	1125	3	6	M6	35.9	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
39	1126	3.5	6	A7	36.2	NaturallyAspirated	7	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
40	1127	3.5	6	A7	34.5	NaturallyAspirated	7	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
41	1128	3.5	6	M6	34.7927	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	0
42	1129	5.5	8	A7	30.8	NaturallyAspirated	7	1	0	TwoWheelDriveRear	2	1	2Seaters	1	0
43	1130	1	3	Other	57.8	NaturallyAspirated	5	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
44	1131	1	3	Other	57.8	NaturallyAspirated	5	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
45	1132	3.7	6	M6	35.9802	NaturallyAspirated	6	1	0	TwoWheelDriveRear	2	2	2Seaters	1	1
46	1133	3.7	6	S7	36.9	NaturallyAspirated	7	1	0	TwoWheelDriveRear	2	2	2Seaters	1	1
47	1134	3.7	6	S7	34.5832	NaturallyAspirated	7	1	0	TwoWheelDriveRear	2	2	2Seaters	1	1
48	1135	3.7	6	M6	34.9	NaturallyAspirated	6	0	0	TwoWheelDriveRear	2	2	2Seaters	1	1
49	1136	2	4	A5	37.5	Turbocharged	5	1	0	TwoWheelDriveRear	2	2	2Seaters	1	0
50	1137	2	4	M5	40	Turbocharged	5	0	0	TwoWheelDriveRear	2	2	2Seaters	1	

Figure 2. "Cars2010" Fuel Economy Dataset.



Name	Role	Level	Type	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
AirAspirationMethod	**Input**	**Nominal**	Character	3	0	.	.	.	.	.	.
CarlineClassDesc	**Input**	**Nominal**	Character	13	0	.	.	.	.	.	.
DriveDesc	**Input**	**Nominal**	Character	5	0	.	.	.	.	.	.
EngDispl	**Input**	**Interval**	Numeric	.	0	1	8.4	3.507407	1.305905	0.645547	-0.23621
ExhaustValvesPerCyl	**Input**	**Nominal**	Numeric	3	0	.	.	.	.	.	.
FE	**Target**	**Interval**	Numeric	.	0	17.5	69.6404	34.70649	7.498033	0.601307	0.790742
IntakeValvePerCyl	**Input**	**Nominal**	Numeric	4	0	.	.	.	.	.	.
NumCyl	**Input**	**Nominal**	Numeric	9	0	.	.	.	.	.	.
NumGears	**Input**	**Nominal**	Numeric	6	0	.	.	.	.	.	.
TransCreeperGear	**Input**	**Binary**	Numeric	2	0	.	.	.	.	.	.
TransLockup	**Input**	**Binary**	Numeric	2	0	.	.	.	.	.	.
Transmission	**Input**	**Nominal**	Character	16	0	.	.	.	.	.	.
VAR1	**Rejected**	**Nominal**	Character	21	0	.	.	.	.	.	.
VarValveLift	**Input**	**Binary**	Numeric	2	0	.	.	.	.	.	.
VarValveTiming	**Input**	**Binary**	Numeric	2	0	.	.	.	.	.	.

Figure 3. Descriptive Statistics of “Cars2010” Dataset.

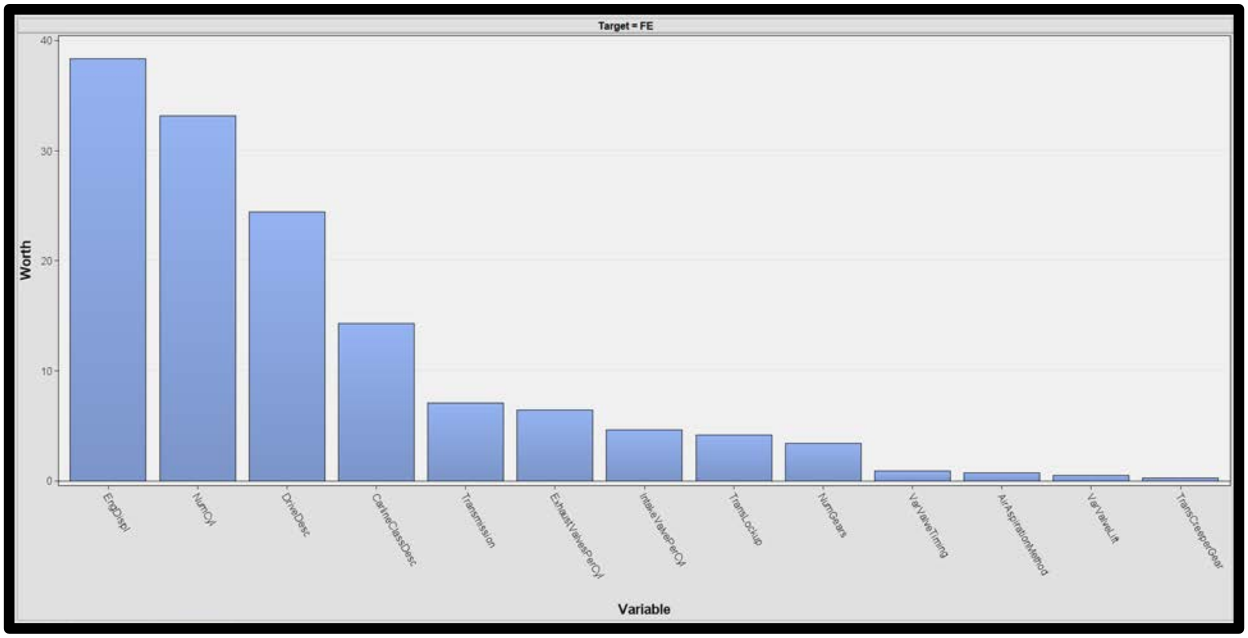


Figure 4. StatExplore Plot of Input Variable Worth.

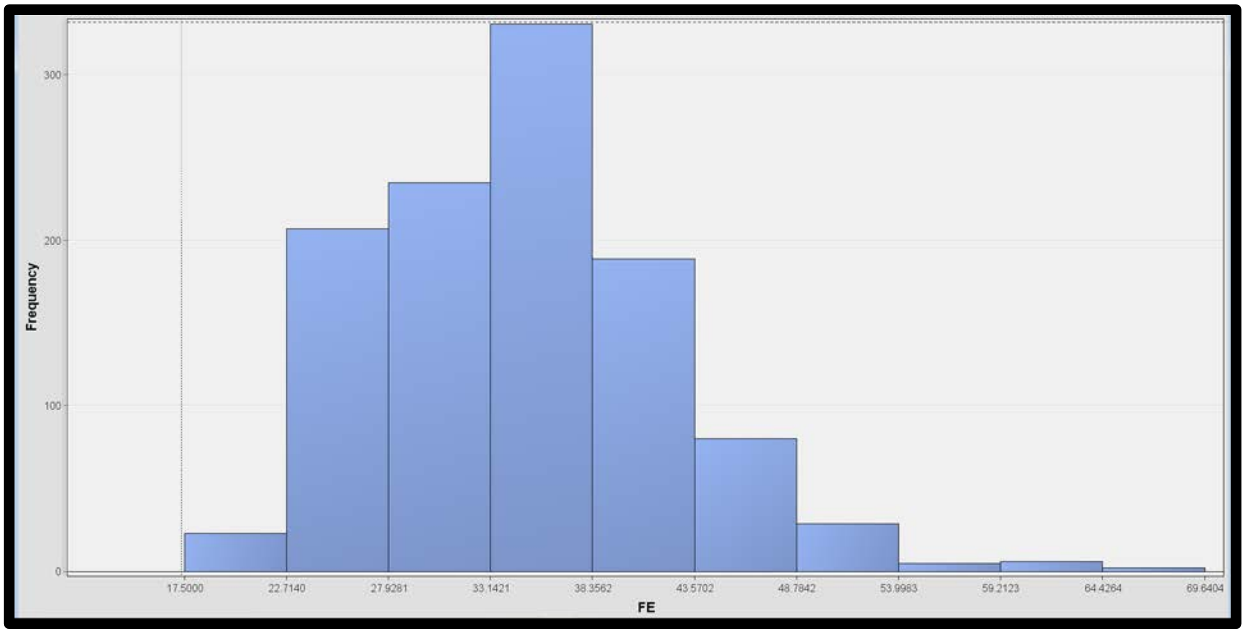
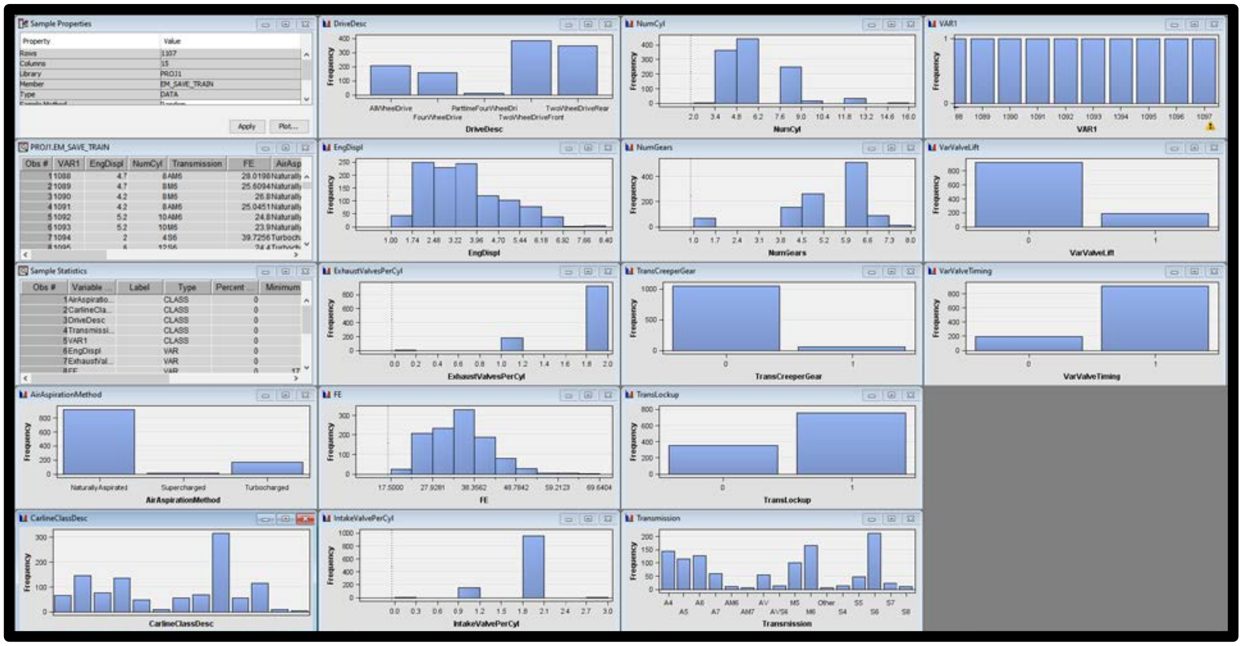


Figure 5. Graph Explore Distribution of Fuel Economy.


Sample Properties

Property	Value
Rows	1107
Columns	15
Library	PROJ1
Member	EM_SAVE_TRAIN
Type	DATA

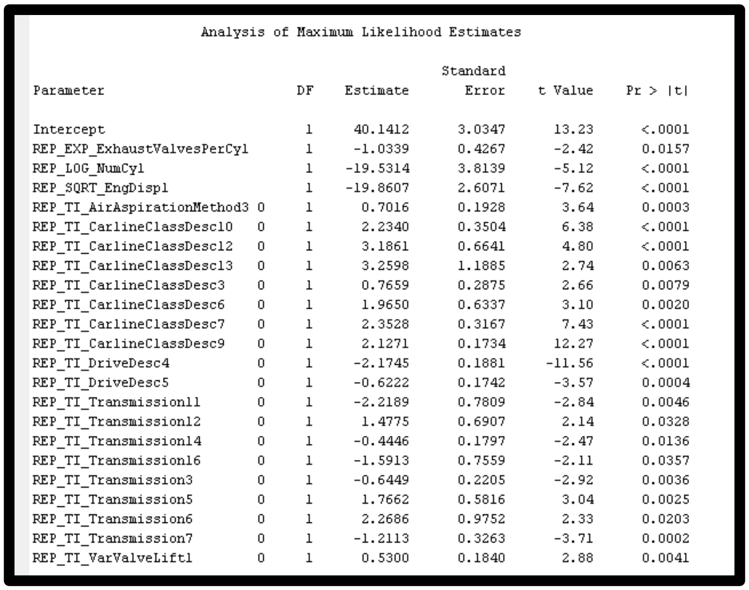
PROJ1.EM_SAVE_TRAIN

Obs #	VAR1	EngDispl	NumCyl	Transmission	FE	AirAsp
1	1088	4.7	8	AM6	28.0198	Naturally
2	1089	4.7	8	M6	25.6094	Naturally
3	1090	4.2	8	M6	26.8	Naturally
4	1091	4.2	8	AM6	25.0451	Naturally
5	1092	5.2	10	AM6	24.8	Naturally
6	1093	5.2	10	M6	23.9	Naturally
7	1094	2	4	S6	39.7256	Turboch.

Sample Statistics

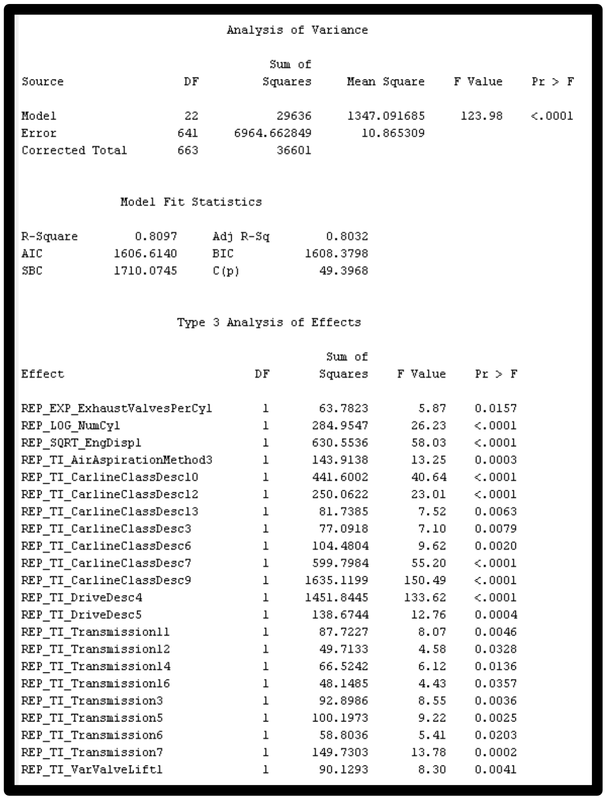
Obs #	Variable ...	Label	Type	Percent ...	Minimum
1	AirAspiratio...		CLASS	0	
2	CarlineCla...		CLASS	0	
3	DriveDesc		CLASS	0	
4	Transmissi...		CLASS	0	
5	VAR1		CLASS	0	
6	EngDispl		VAR	0	
7	ExhaustVal...		VAR	0	

Figure 6. Initial Distributions of Variables in “Cars2010” Dataset.



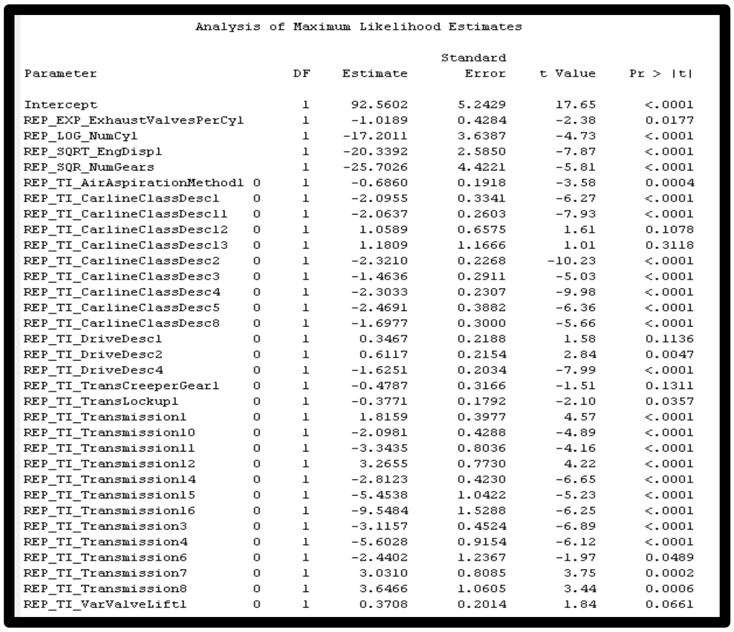
```
                    Analysis of Maximum Likelihood Estimates

                                                     Standard
Parameter                        DF     Estimate        Error    t Value    Pr > |t|

Intercept                         1      40.1412       3.0347      13.23      <.0001
REP_EXP_ExhaustValvesPerCyl       1      -1.0339       0.4267      -2.42      0.0157
REP_LOG_NumCyl                    1     -19.5314       3.8139      -5.12      <.0001
REP_SQRT_EngDispl                 1     -19.8607       2.6071      -7.62      <.0001
REP_TI_AirAspirationMethod3 0     1       0.7016       0.1928       3.64      0.0003
REP_TI_CarlineClassDesc10   0     1       2.2340       0.3504       6.38      <.0001
REP_TI_CarlineClassDesc12   0     1       3.1861       0.6641       4.80      <.0001
REP_TI_CarlineClassDesc13   0     1       3.2598       1.1885       2.74      0.0063
REP_TI_CarlineClassDesc3    0     1       0.7659       0.2875       2.66      0.0079
REP_TI_CarlineClassDesc6    0     1       1.9650       0.6337       3.10      0.0020
REP_TI_CarlineClassDesc7    0     1       2.3528       0.3167       7.43      <.0001
REP_TI_CarlineClassDesc9    0     1       2.1271       0.1734      12.27      <.0001
REP_TI_DriveDesc4           0     1      -2.1745       0.1881     -11.56      <.0001
REP_TI_DriveDesc5           0     1      -0.6222       0.1742      -3.57      0.0004
REP_TI_Transmission11       0     1      -2.2189       0.7809      -2.84      0.0046
REP_TI_Transmission12       0     1       1.4775       0.6907       2.14      0.0328
REP_TI_Transmission14       0     1      -0.4446       0.1797      -2.47      0.0136
REP_TI_Transmission16       0     1      -1.5913       0.7559      -2.11      0.0357
REP_TI_Transmission3        0     1      -0.6449       0.2205      -2.92      0.0036
REP_TI_Transmission5        0     1       1.7662       0.5816       3.04      0.0025
REP_TI_Transmission6        0     1       2.2686       0.9752       2.33      0.0203
REP_TI_Transmission7        0     1      -1.2113       0.3263      -3.71      0.0002
REP_TI_VarValveLift1        0     1       0.5300       0.1840       2.88      0.0041
```

Figure 7. T-Test Results of Regression Model 1 (Forward Selection).



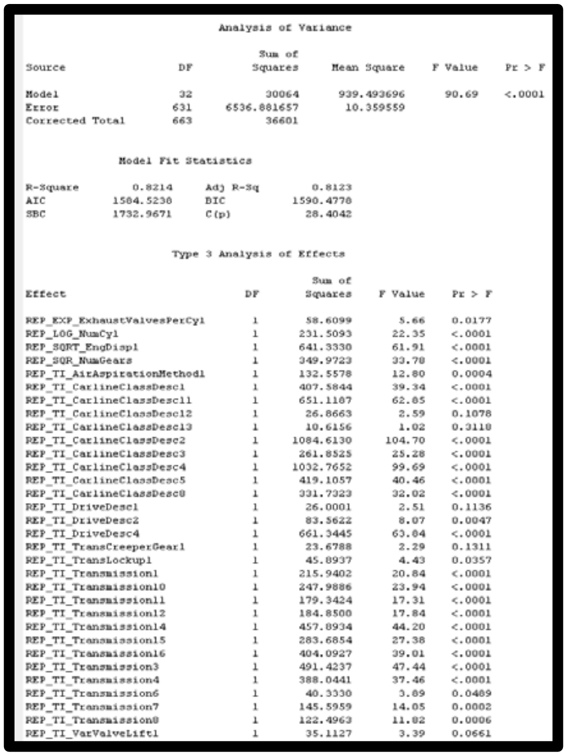
```
                              Analysis of Variance

                                     Sum of
Source                   DF         Squares     Mean Square    F Value    Pr > F

Model                    22           29636     1347.091685     123.98    <.0001
Error                   641     6964.662849       10.865309
Corrected Total         663           36601


                Model Fit Statistics

R-Square        0.8097     Adj R-Sq        0.8032
AIC          1606.6140     BIC          1608.3798
SBC          1710.0745     C(p)           49.3968


                       Type 3 Analysis of Effects

                                               Sum of
Effect                           DF           Squares    F Value    Pr > F

REP_EXP_ExhaustValvesPerCyl       1           63.7823       5.87    0.0157
REP_LOG_NumCyl                    1          284.9547      26.23    <.0001
REP_SQRT_EngDispl                 1          630.5536      58.03    <.0001
REP_TI_AirAspirationMethod3       1          143.9138      13.25    0.0003
REP_TI_CarlineClassDesc10         1          441.6002      40.64    <.0001
REP_TI_CarlineClassDesc12         1          250.0622      23.01    <.0001
REP_TI_CarlineClassDesc13         1           81.7385       7.52    0.0063
REP_TI_CarlineClassDesc3          1           77.0918       7.10    0.0079
REP_TI_CarlineClassDesc6          1          104.4804       9.62    0.0020
REP_TI_CarlineClassDesc7          1          599.7984      55.20    <.0001
REP_TI_CarlineClassDesc9          1         1635.1199     150.49    <.0001
REP_TI_DriveDesc4                 1         1451.8445     133.62    <.0001
REP_TI_DriveDesc5                 1          138.6744      12.76    0.0004
REP_TI_Transmission11             1           87.7227       8.07    0.0046
REP_TI_Transmission12             1           49.7133       4.58    0.0328
REP_TI_Transmission14             1           66.5242       6.12    0.0136
REP_TI_Transmission16             1           48.1485       4.43    0.0357
REP_TI_Transmission3              1           92.8986       8.55    0.0036
REP_TI_Transmission5              1          100.1973       9.22    0.0025
REP_TI_Transmission6              1           58.8036       5.41    0.0203
REP_TI_Transmission7              1          149.7303      13.78    0.0002
REP_TI_VarValveLift1              1           90.1293       8.30    0.0041
```

Figure 8. R-Squared and F-Test Results of Regression Model 1 (Forward Selection).



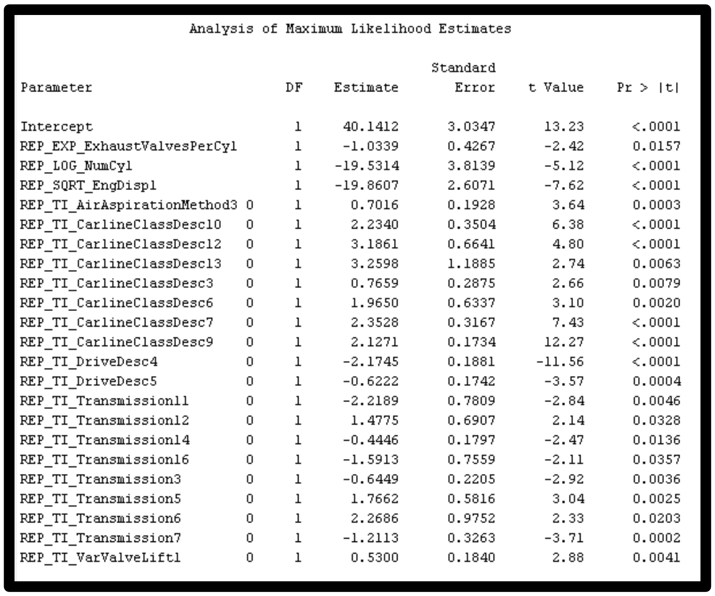
```
                    Analysis of Maximum Likelihood Estimates

                                                       Standard
Parameter                         DF     Estimate         Error     t Value     Pr > |t|

Intercept                          1      92.5602        5.2429       17.65       <.0001
REP_EXP_ExhaustValvesPerCyl        1      -1.0189        0.4284       -2.38       0.0177
REP_LOG_NumCyl                     1     -17.2011        3.6387       -4.73       <.0001
REP_SQRT_EngDispl                  1     -20.3392        2.5850       -7.87       <.0001
REP_SQR_NumGears                   1     -25.7026        4.4221       -5.81       <.0001
REP_TI_AirAspirationMethod1 0      1      -0.6860        0.1918       -3.58       0.0004
REP_TI_CarlineClassDesc1    0      1      -2.0955        0.3341       -6.27       <.0001
REP_TI_CarlineClassDesc11   0      1      -2.0637        0.2603       -7.93       <.0001
REP_TI_CarlineClassDesc12   0      1       1.0589        0.6575        1.61       0.1078
REP_TI_CarlineClassDesc13   0      1       1.1809        1.1666        1.01       0.3118
REP_TI_CarlineClassDesc2    0      1      -2.3210        0.2268      -10.23       <.0001
REP_TI_CarlineClassDesc3    0      1      -1.4636        0.2911       -5.03       <.0001
REP_TI_CarlineClassDesc4    0      1      -2.3033        0.2307       -9.98       <.0001
REP_TI_CarlineClassDesc5    0      1      -2.4691        0.3882       -6.36       <.0001
REP_TI_CarlineClassDesc8    0      1      -1.6977        0.3000       -5.66       <.0001
REP_TI_DriveDesc1           0      1       0.3467        0.2188        1.58       0.1136
REP_TI_DriveDesc2           0      1       0.6117        0.2154        2.84       0.0047
REP_TI_DriveDesc4           0      1      -1.6251        0.2034       -7.99       <.0001
REP_TI_TransCreeperGear1    0      1      -0.4787        0.3166       -1.51       0.1311
REP_TI_TransLockup1         0      1      -0.3771        0.1792       -2.10       0.0357
REP_TI_Transmission1        0      1       1.8159        0.3977        4.57       <.0001
REP_TI_Transmission10       0      1      -2.0981        0.4288       -4.89       <.0001
REP_TI_Transmission11       0      1      -3.3435        0.8036       -4.16       <.0001
REP_TI_Transmission12       0      1       3.2655        0.7730        4.22       <.0001
REP_TI_Transmission14       0      1      -2.8123        0.4230       -6.65       <.0001
REP_TI_Transmission15       0      1      -5.4538        1.0422       -5.23       <.0001
REP_TI_Transmission16       0      1      -9.5484        1.5288       -6.25       <.0001
REP_TI_Transmission3        0      1      -3.1157        0.4524       -6.89       <.0001
REP_TI_Transmission4        0      1      -5.6028        0.9154       -6.12       <.0001
REP_TI_Transmission6        0      1      -2.4402        1.2367       -1.97       0.0489
REP_TI_Transmission7        0      1       3.0310        0.8085        3.75       0.0002
REP_TI_Transmission8        0      1       3.6466        1.0605        3.44       0.0006
REP_TI_VarValveLift1        0      1       0.3708        0.2014        1.84       0.0661
```

Figure 9. T-Test Results of Regression Model 2 (Backward Elimination).



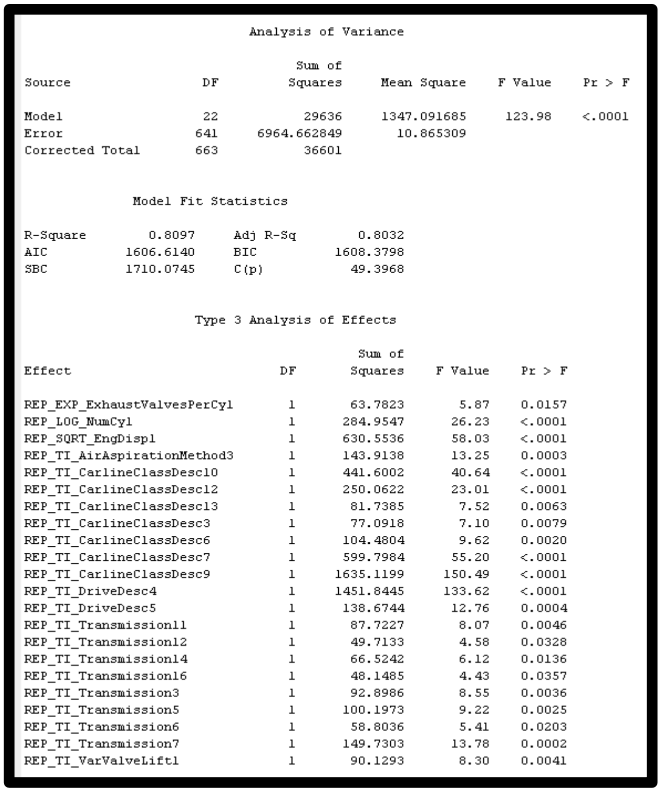
```
                         Analysis of Variance

                                       Sum of
Source                  DF            Squares    Mean Square    F Value    Pr > F

Model                   32              30064     939.493696      90.69    <.0001
Error                  631        6536.881657      10.359559
Corrected Total        663              36601


            Model Fit Statistics

R-Square         0.8214     Adj R-Sq          0.8123
AIC           1584.5238     BIC            1590.4778
SBC           1732.9671     C(p)             28.4042


                   Type 3 Analysis of Effects

                                              Sum of
Effect                             DF        Squares    F Value    Pr > F

REP_EXP_ExhaustValvesPerCyl         1        58.6099       5.66    0.0177
REP_LOG_NumCyl                      1       231.5093      22.35    <.0001
REP_SQRT_EngDispl                   1       641.3330      61.91    <.0001
REP_SQR_NumGears                    1       349.9723      33.78    <.0001
REP_TI_AirAspirationMethod1         1       132.5578      12.80    0.0004
REP_TI_CarlineClassDesc1            1       407.5844      39.34    <.0001
REP_TI_CarlineClassDesc11           1       651.1187      62.85    <.0001
REP_TI_CarlineClassDesc12           1        26.8663       2.59    0.1078
REP_TI_CarlineClassDesc13           1        10.6156       1.02    0.3118
REP_TI_CarlineClassDesc2            1      1084.6130     104.70    <.0001
REP_TI_CarlineClassDesc3            1       261.8525      25.28    <.0001
REP_TI_CarlineClassDesc4            1      1032.7652      99.69    <.0001
REP_TI_CarlineClassDesc5            1       419.1057      40.46    <.0001
REP_TI_CarlineClassDesc8            1       331.7323      32.02    <.0001
REP_TI_DriveDesc1                   1        26.0001       2.51    0.1136
REP_TI_DriveDesc2                   1        83.5622       8.07    0.0047
REP_TI_DriveDesc4                   1       661.3445      63.84    <.0001
REP_TI_TransCreeperGear1            1        23.6788       2.29    0.1311
REP_TI_TransLockup1                 1        45.8937       4.43    0.0357
REP_TI_Transmission1                1       215.9402      20.84    <.0001
REP_TI_Transmission10               1       247.9886      23.94    <.0001
REP_TI_Transmission11               1       179.3424      17.31    <.0001
REP_TI_Transmission12               1       184.8500      17.84    <.0001
REP_TI_Transmission14               1       457.8934      44.20    <.0001
REP_TI_Transmission15               1       283.6854      27.38    <.0001
REP_TI_Transmission16               1       404.0927      39.01    <.0001
REP_TI_Transmission3                1       491.4237      47.44    <.0001
REP_TI_Transmission4                1       388.0441      37.46    <.0001
REP_TI_Transmission6                1        40.3330       3.89    0.0489
REP_TI_Transmission7                1       145.5959      14.05    0.0002
REP_TI_Transmission8                1       122.4963      11.82    0.0006
REP_TI_VarValveLift1                1        35.1127       3.39    0.0661
```

Figure 10. R-Squared and F-Test Results of Regression Model 2 (Backward Elimination).



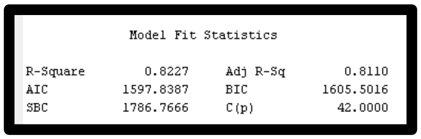
```
                    Analysis of Maximum Likelihood Estimates

                                                     Standard
Parameter                        DF     Estimate        Error    t Value    Pr > |t|

Intercept                         1      40.1412       3.0347      13.23      <.0001
REP_EXP_ExhaustValvesPerCyl       1      -1.0339       0.4267      -2.42      0.0157
REP_LOG_NumCyl                    1     -19.5314       3.8139      -5.12      <.0001
REP_SQRT_EngDispl                 1     -19.8607       2.6071      -7.62      <.0001
REP_TI_AirAspirationMethod3 0     1       0.7016       0.1928       3.64      0.0003
REP_TI_CarlineClassDesc10   0     1       2.2340       0.3504       6.38      <.0001
REP_TI_CarlineClassDesc12   0     1       3.1861       0.6641       4.80      <.0001
REP_TI_CarlineClassDesc13   0     1       3.2598       1.1885       2.74      0.0063
REP_TI_CarlineClassDesc3    0     1       0.7659       0.2875       2.66      0.0079
REP_TI_CarlineClassDesc6    0     1       1.9650       0.6337       3.10      0.0020
REP_TI_CarlineClassDesc7    0     1       2.3528       0.3167       7.43      <.0001
REP_TI_CarlineClassDesc9    0     1       2.1271       0.1734      12.27      <.0001
REP_TI_DriveDesc4           0     1      -2.1745       0.1881     -11.56      <.0001
REP_TI_DriveDesc5           0     1      -0.6222       0.1742      -3.57      0.0004
REP_TI_Transmission11       0     1      -2.2189       0.7809      -2.84      0.0046
REP_TI_Transmission12       0     1       1.4775       0.6907       2.14      0.0328
REP_TI_Transmission14       0     1      -0.4446       0.1797      -2.47      0.0136
REP_TI_Transmission16       0     1      -1.5913       0.7559      -2.11      0.0357
REP_TI_Transmission3        0     1      -0.6449       0.2205      -2.92      0.0036
REP_TI_Transmission5        0     1       1.7662       0.5816       3.04      0.0025
REP_TI_Transmission6        0     1       2.2686       0.9752       2.33      0.0203
REP_TI_Transmission7        0     1      -1.2113       0.3263      -3.71      0.0002
REP_TI_VarValveLift1        0     1       0.5300       0.1840       2.88      0.0041
```

Figure 11. T-Test Results of Regression Model 3 (Stepwise Selection).



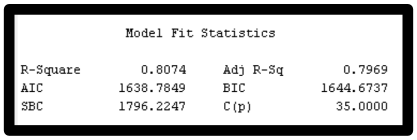
```
                              Analysis of Variance

                                      Sum of
Source                 DF            Squares     Mean Square    F Value    Pr > F

Model                  22              29636     1347.091685     123.98    <.0001
Error                 641        6964.662849       10.865309
Corrected Total       663              36601


              Model Fit Statistics

R-Square        0.8097     Adj R-Sq        0.8032
AIC          1606.6140     BIC          1608.3798
SBC          1710.0745     C(p)           49.3968


                       Type 3 Analysis of Effects

                                             Sum of
Effect                          DF          Squares    F Value    Pr > F

REP_EXP_ExhaustValvesPerCyl      1          63.7823       5.87    0.0157
REP_LOG_NumCyl                   1         284.9547      26.23    <.0001
REP_SQRT_EngDispl                1         630.5536      58.03    <.0001
REP_TI_AirAspirationMethod3      1         143.9138      13.25    0.0003
REP_TI_CarlineClassDesc10        1         441.6002      40.64    <.0001
REP_TI_CarlineClassDesc12        1         250.0622      23.01    <.0001
REP_TI_CarlineClassDesc13        1          81.7385       7.52    0.0063
REP_TI_CarlineClassDesc3         1          77.0918       7.10    0.0079
REP_TI_CarlineClassDesc6         1         104.4804       9.62    0.0020
REP_TI_CarlineClassDesc7         1         599.7984      55.20    <.0001
REP_TI_CarlineClassDesc9         1        1635.1199     150.49    <.0001
REP_TI_DriveDesc4                1        1451.8445     133.62    <.0001
REP_TI_DriveDesc5                1         138.6744      12.76    0.0004
REP_TI_Transmission11            1          87.7227       8.07    0.0046
REP_TI_Transmission12            1          49.7133       4.58    0.0328
REP_TI_Transmission14            1          66.5242       6.12    0.0136
REP_TI_Transmission16            1          48.1485       4.43    0.0357
REP_TI_Transmission3             1          92.8986       8.55    0.0036
REP_TI_Transmission5             1         100.1973       9.22    0.0025
REP_TI_Transmission6             1          58.8036       5.41    0.0203
REP_TI_Transmission7             1         149.7303      13.78    0.0002
REP_TI_VarValveLift1             1          90.1293       8.30    0.0041
```

Figure 12. R-Squared and F-Test Results of Regression Model 3 (Stepwise Selection).



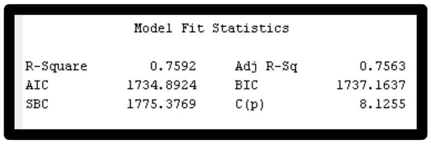
```
                 Model Fit Statistics

R-Square          0.8227      Adj R-Sq          0.8110
AIC            1597.8387      BIC            1605.5016
SBC            1786.7666      C(p)             42.0000
```

Figure 13. R-Squared Results of Regression Model 4 (Keep All Variables).



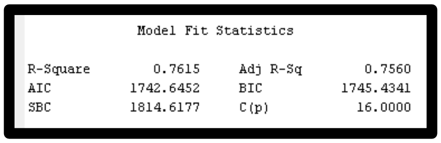
```
                 Model Fit Statistics

R-Square          0.8074      Adj R-Sq          0.7969
AIC            1638.7849      BIC            1644.6737
SBC            1796.2247      C(p)             35.0000
```

Figure 14. R-Squared Results of Regression Model 5 (Custom Variable Selection).



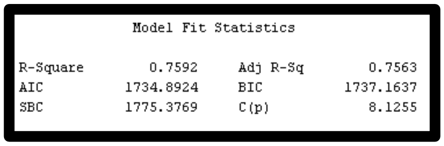
```
                 Model Fit Statistics

R-Square          0.7592      Adj R-Sq          0.7563
AIC            1734.8924      BIC            1737.1637
SBC            1775.3769      C(p)              8.1255
```

Figure 15. R-Squared Results of Regression Model 6 (Forward Selection 2).



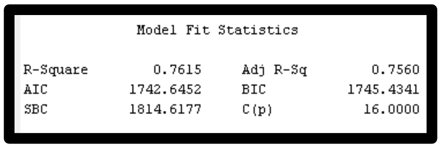
```
                 Model Fit Statistics

R-Square          0.7615      Adj R-Sq          0.7560
AIC            1742.6452      BIC            1745.4341
SBC            1814.6177      C(p)             16.0000
```

Figure 16. R-Squared Results of Regression Model 7 (Backward Elimination 2).



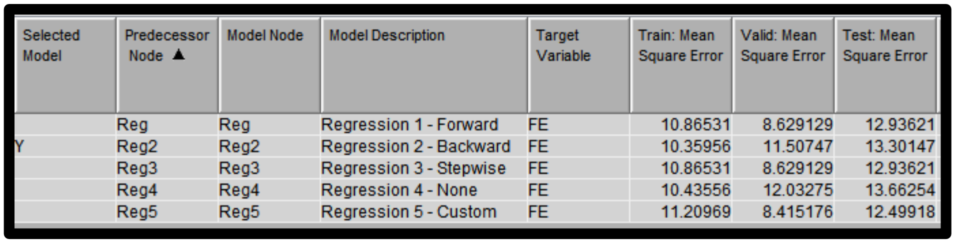
```
                 Model Fit Statistics

R-Square          0.7592      Adj R-Sq          0.7563
AIC            1734.8924      BIC            1737.1637
SBC            1775.3769      C(p)              8.1255
```

Figure 17. R-Squared Results of Regression Model 8 (Stepwise Selection 2).



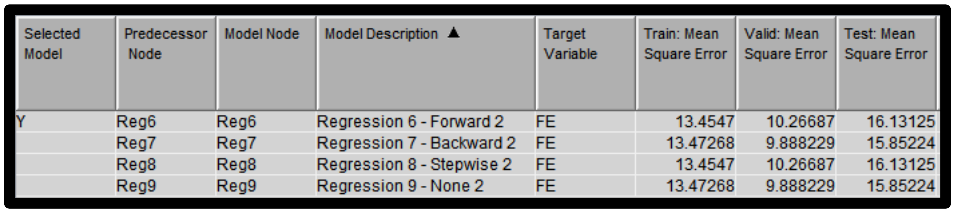
```
                    Model Fit Statistics

R-Square         0.7615    Adj R-Sq          0.7560
AIC           1742.6452    BIC            1745.4341
SBC           1814.6177    C(p)             16.0000
```

Figure 18. R-Squared Results of Regression Model 9 (Keep All Variables 2).



Selected Model	Predecessor Node ▲	Model Node	Model Description	Target Variable	Train: Mean Square Error	Valid: Mean Square Error	Test: Mean Square Error
	Reg	Reg	Regression 1 - Forward	FE	10.86531	8.629129	12.93621
Y	Reg2	Reg2	Regression 2 - Backward	FE	10.35956	11.50747	13.30147
	Reg3	Reg3	Regression 3 - Stepwise	FE	10.86531	8.629129	12.93621
	Reg4	Reg4	Regression 4 - None	FE	10.43556	12.03275	13.66254
	Reg5	Reg5	Regression 5 - Custom	FE	11.20969	8.415176	12.49918

Figure 19. Model Comparison for Models without Variable Selection Node.



Selected Model	Predecessor Node	Model Node	Model Description ▲	Target Variable	Train: Mean Square Error	Valid: Mean Square Error	Test: Mean Square Error
Y	Reg6	Reg6	Regression 6 - Forward 2	FE	13.4547	10.26687	16.13125
	Reg7	Reg7	Regression 7 - Backward 2	FE	13.47268	9.888229	15.85224
	Reg8	Reg8	Regression 8 - Stepwise 2	FE	13.4547	10.26687	16.13125
	Reg9	Reg9	Regression 9 - None 2	FE	13.47268	9.888229	15.85224

Figure 20. Model Comparison for Models with Variable Selection Node.

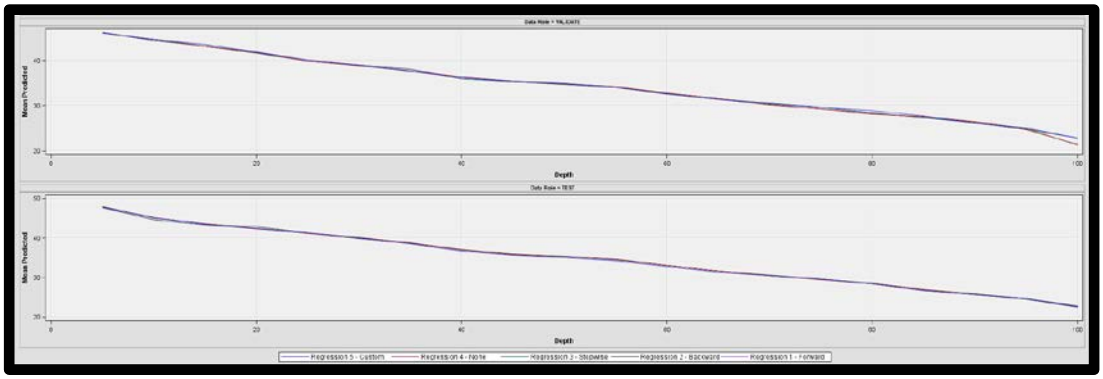


Figure 21. Score Rankings Plot of First Five Models for Validation and Test Data.