MSDS 6372 Project 1

Introduction

The purpose of this project is to build a multiple linear regression model to predict the manufacturer's suggested retail price (MSRP) of existing cars from data scraped via Edmunds.com, a popular car dealership site. This is not meant to predict a specific car's price based on its particular mileage and condition, but to predict its base MSRP. We will perform initial exploratory data analysis prior to building the model to ensure appropriate assumptions are met and conditions are optimized for the model selection process. We will also compare several models against one another in order to identify the best model selection technique. The secondary objective of our analysis is to utilize the two-way ANOVA technique on two of the data categorical variables to investigate a potential interaction between the two independent variables on the dependent variable. We will use SAS software to perform statistical analyses for this project.

Data Description

This dataset was scraped by Sam Keene from Edumunds.com and Twitter.com and posted on the popular data site kaggle.com. The data includes 11,914 observations and 15 explanatory variables (8 categorical and 7 continuous). Our continuous response variable will be the MSRP (in U.S. dollars) of the particular year and model of the car. The name and type of each variable along with an example is shown in Display 1.1.

Display 1.1 Structure of the car dataset

Variable	Type	Example
Make	string	BMW
Model	string	1 Series M
Year	numeric	2011
Engine Fuel Type	string	premium unleaded (required)
Engine HP	numeric	335
Engine Cylinders	numeric	6
Transmission Type	string	MANUAL
Driven_Wheels	string	rear wheel drive
Number of Doors	numeric	2
Market Category	string	Factory Tuner,Luxury,High-Performance
Vehicle Size	string	Compact
Vehicle Style	string	Coupe
highway MPG	numeric	26

city mpg	numeric	19
Popularity	numeric	3916
MSRP	numeric	46135

No description of the variables was provided, however most variables are self-explanatory. The only non-intuitive metric is *popularity,* which was calculated by unknown means. With the data being so large, statistical significance of specific variables will be found much more readily than is practically significant. This, coupled with the potential of overfitting the data led us to take a 3,000 simple random sample of the data. We then removed the records that contained any null values. This reduced our observations to 2,981.

Exploratory Analysis

Based on the summary statistics table (Display 1.2), the dataset contains cars that were produced from 1990 to 2017. The minimum value of Engine Cylinders is zero, which is not a concern as electric cars are included in this data. The range and the standard deviation of our response variable (MSRP) is quite large. The minimum value of MSRP is $2,000 and the maximum value is $1,500,000. The fact that the standard deviation is higher relative to the mean of the response variable, suggests that the data is right-skewed and a log transformation may be necessary to correct the skewness.

Display 1.2 Summary statistics of the continuous variables

Variable	N	Mean	Std Dev	Minimum	Maximum
Year	2981	2010.37	7.6074666	1990.00	2017.00
Engine HP	2981	249.4226770	108.9349652	55.0000000	707.0000000
Engine Cylinders	2981	5.6309963	1.7291515	0	12.0000000
Number of Doors	2981	3.4280443	0.8879914	2.0000000	4.0000000
highway MPG	2981	26.3264005	6.6227038	12.0000000	109.0000000
city mpg	2981	19.3267360	6.2466033	7.0000000	128.0000000
Popularity	2981	1513.13	1420.88	2.0000000	5657.00
MSRP	2981	40709.85	60734.89	2000.00	1500000.00

Next, we ran summary statistics on the response variable by make of the car (Display 1.3) and Engine Fuel Type (Display 1.4) to obtain more information on the sample size. Several brands, such as Alfa Romeo, Genesis, HUMMER, have a very low number of observations (less than 5). Also engine fuel type, namely electric and natural gas, only have two and one observation(s) respectively. Our final model may not be able to accurately predict MSRP of a car brand and/or engine fuel type that have so few observations.

Display 1.3 Summary statistics of the response variable (MSRP) by make of the car.

Analysis Variable : MSRP						
Make	N Obs	N	Mean	Std Dev	Minimum	Maximum
Acura	67	67	35175.49	24614.29	2000.00	156000.00
Alfa Romeo	1	1	53900.00	.	53900.00	53900.00
Aston Martin	27	27	203616.93	70191.97	98200.00	305650.00
Audi	74	74	54388.27	38015.69	2000.00	173500.00
BMW	91	91	61350.65	27281.92	4697.00	137000.00
Bentley	19	19	236613.16	43365.57	177500.00	340990.00
Buick	44	44	29814.55	12012.71	2000.00	49625.00
Cadillac	88	88	53737.98	17802.89	2000.00	97460.00
Chevrolet	267	267	28075.82	17109.99	2000.00	92345.00
Chrysler	52	52	25811.96	12381.64	2000.00	42770.00
Dodge	149	149	22714.05	18062.09	2000.00	118795.00
FIAT	21	21	22110.00	3125.56	16995.00	27495.00
Ferrari	18	18	255440.89	109731.30	150694.00	643330.00
Ford	205	205	27644.06	16880.08	2000.00	149995.00
GMC	127	127	31253.42	16742.65	2000.00	70220.00
Genesis	2	2	47975.00	9298.45	41400.00	54550.00
HUMMER	4	4	35947.50	2051.24	33390.00	38365.00
Honda	114	114	25949.61	9122.06	2000.00	42870.00
Hyundai	69	69	24433.46	10508.73	2000.00	68500.00
Infiniti	84	84	40649.10	15716.90	2000.00	88850.00
Kia	49	49	24675.00	9868.36	11820.00	54500.00
Lamborghini	11	11	394236.36	377716.33	191900.00	1500000.00
Land Rover	47	47	62824.89	34938.42	24975.00	186495.00
Lexus	58	58	47617.95	20163.45	4282.00	120440.00
Lincoln	37	37	39483.16	17464.48	2000.00	71260.00
Lotus	9	9	66748.89	12879.12	43995.00	79980.00
Maserati	18	18	107269.11	32123.33	69800.00	182009.00
Maybach	5	5	606490.00	433574.89	366000.00	1380000.00
Mazda	114	114	19295.78	9812.51	2000.00	36625.00
McLaren	1	1	229000.00	.	229000.00	229000.00
Mercedes-Benz	102	102	72824.43	69048.26	2000.00	495000.00
Mitsubishi	57	57	21374.98	9560.32	2000.00	38195.00
Nissan	130	130	28210.70	16466.06	2000.00	115710.00
Oldsmobile	39	39	9511.46	12450.55	2000.00	35870.00
Plymouth	26	26	4771.46	9077.34	2000.00	44625.00
Pontiac	46	46	18603.35	10166.63	2000.00	34020.00
Porsche	45	45	99606.60	86903.23	3047.00	440000.00
Rolls-Royce	4	4	399578.75	86946.85	319400.00	474990.00
Saab	30	30	25574.60	18863.43	2000.00	51330.00
Scion	9	9	22633.33	5227.01	15665.00	29990.00
Spyker	1	1	209990.00	.	209990.00	209990.00
Subaru	63	63	26100.48	7819.30	2000.00	37645.00
Suzuki	96	96	17576.57	7799.37	2000.00	31749.00
Toyota	186	186	28667.14	13243.39	2000.00	65080.00
Volkswagen	200	200	28851.05	10261.72	2000.00	96600.00
Volvo	75	75	26798.21	18724.84	2000.00	51050.00

Display 1.4 Summary statistics of the response variable (MSRP) by engine fuel type.

Analysis Variable : MSRP						
Engine Fuel Type	N Obs	N	Mean	Std Dev	Minimum	Maximum
diesel	35	35	41600.43	18676.44	2098.00	85200.00
electric	2	2	26260.00	1060.66	25510.00	27010.00
flex-fuel (premium unleaded re	18	18	126025.83	81772.37	28900.00	267000.00
flex-fuel (unleaded/E85)	208	208	35436.30	12937.41	2234.00	82570.00
natural gas	1	1	29390.00	.	29390.00	29390.00
premium unleaded (recommended)	388	388	39720.88	14605.95	16995.00	115900.00
premium unleaded (required)	530	530	100990.92	120925.51	21885.00	1500000.00
regular unleaded	1799	1799	22924.97	14564.60	2000.00	89000.00

Objective 1 – Predictive Least Square Regression Model

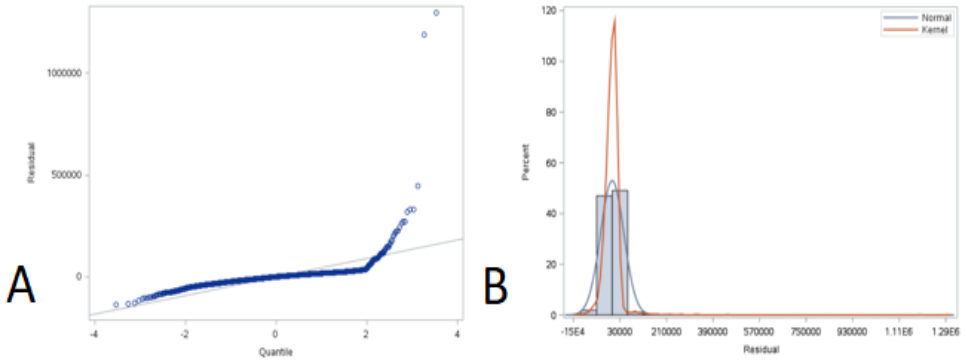
The objective is to build the most predictive least squares regression model for MSRPs of cars using the following model selection techniques: forward selection, backward selection, stepwise selection and lasso selection. To aid in our final determination of the best predictive model, we will provide a table specifying the predictor(s) selected by each model selection technique, the adjusted R^2 and the CV Press for each of the four models. The model with the lowest CV Press will be the best predictive model. Later we discuss why we chose CV Press.

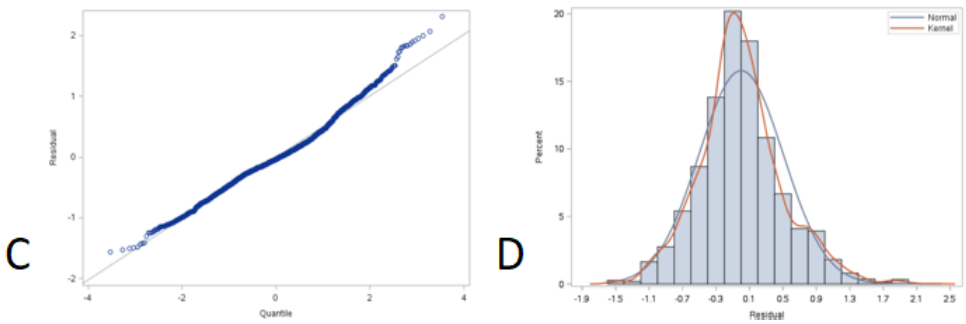
Model Assumptions

In this section, we will assess whether the model's assumptions required for least squares regression analysis are met. We will also check for multicollinearity, and identify and address outliers/influential observations.

Based on the Q-Q plot and the histogram of the residuals for MSRP of the original data (Display 1.5), there is a strong evidence that the residuals are right-skewed. Therefore, we take a log of MSRP to correct the skewness. After the log transformation, the normality of the residuals improves.

Display 1.5 Q-Q plot and histogram of residuals for MSRP before log transformation (panel A, B) and after log base e transformation (panel C, D) from the regression of MSRP on all the continuous variables

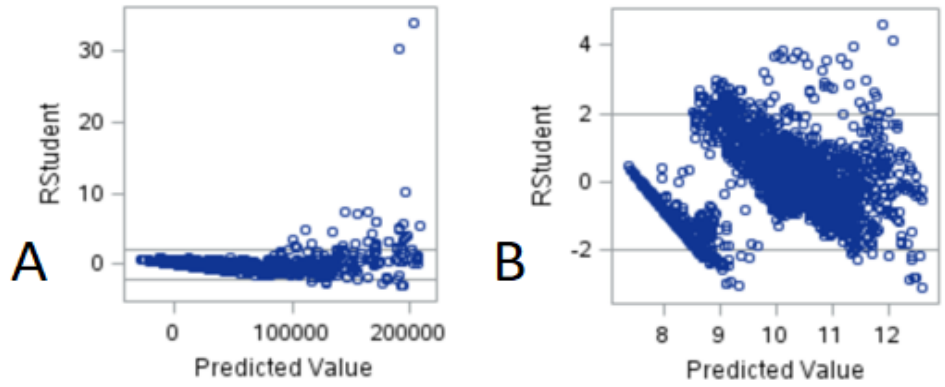




The studentized residual plot of the original data in Display 1.6 (A) shows strong evidence of non-constant variance, the residuals have a funnel shape with two extreme outlying observations. There are two data points that are far distant from the upper line.

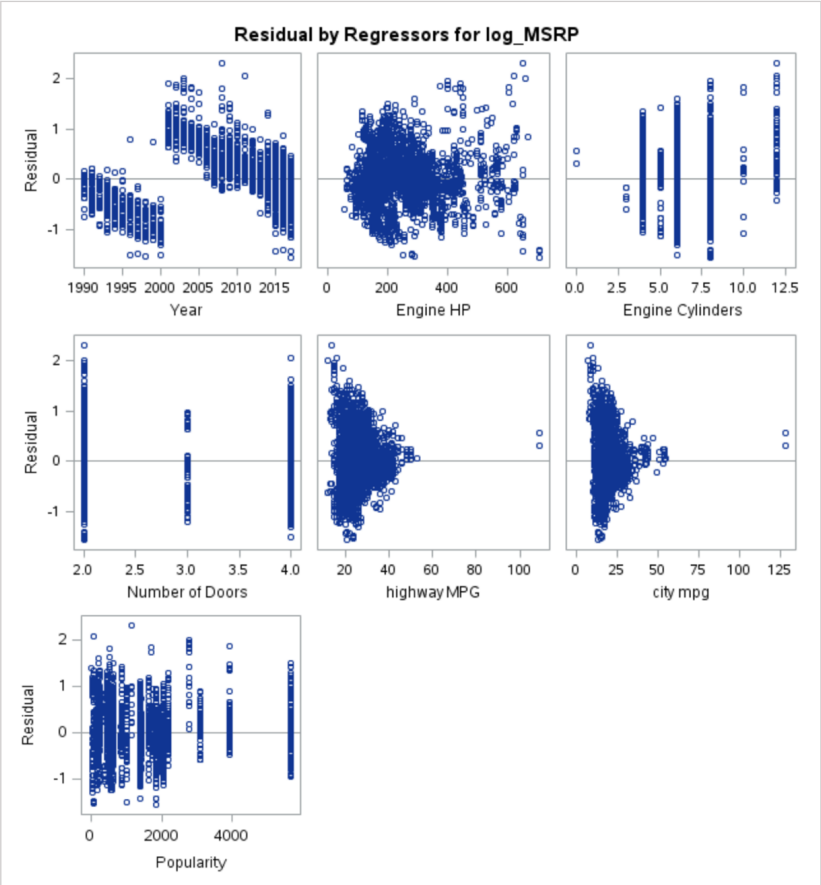
With the log transformation, the residuals are more randomly and equally spread out and there are no extreme outliers as shown in Display 1.6 (B).

Display 1.6 Studentized residual plot before log transformation (A) and after log transformation (B) from the regression of MSRP on all the continuous variables

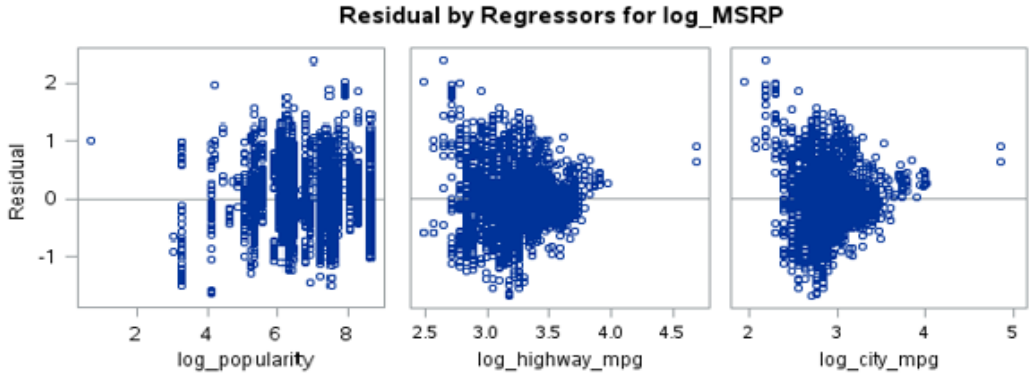


We then examine the residual plots of log transformed MSRP vs the continuous variables. The residuals vs year plot in display 1.7, shows an anomalous break from negative to positive at the year 2000. No transformation we conducted was able to correct this. The residuals for engine horsepower, engine cylinders, and number of doors all look randomly distributed. Both highway and city mpg values have a funnel shape with decreasing residuals as mpg values increase. The two high values are due to electric cars. We will take a log on these two variables and rerun the residual diagnostics to check the effect. We will also take a log of popularity since there are some high values that create a right-skewed distribution. After taking the log of the popularity, highway mpg, and city mpg (Display 1.8), the residuals look much more normally distributed with constant variance.

Display 1.7 Residual plots of log_MSRP vs the continuous variables.

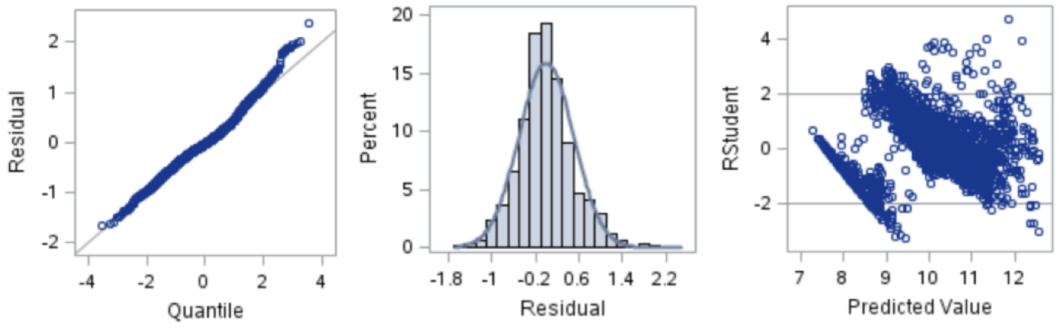


Display 1.8 Residual plots for log transformed popularity, highway mpg, and city mpg continuous variables.



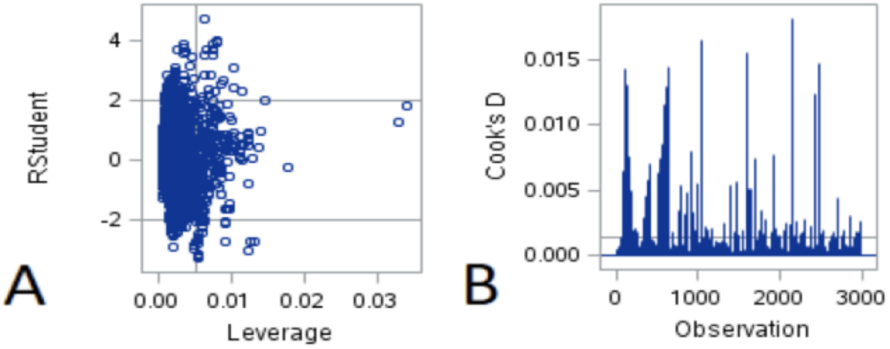
The residuals still look normally distributed with constant variance after the log transformation of the three continuous variables as shown in Display 1.9. Additionally, we cannot find any reasons to consider that these records are dependent on each other, thus we will assume that the data are independent. We feel comfortable moving forward now that we have addressed the three assumptions of residual normality, constant variance, and independence of observations.

Display 1.9 Studentized residual plot, QQ plot, and histogram of residuals from the regression of log_MSRP on the continuous variables (popularity, highway mpg, and city mpg are log transformed).



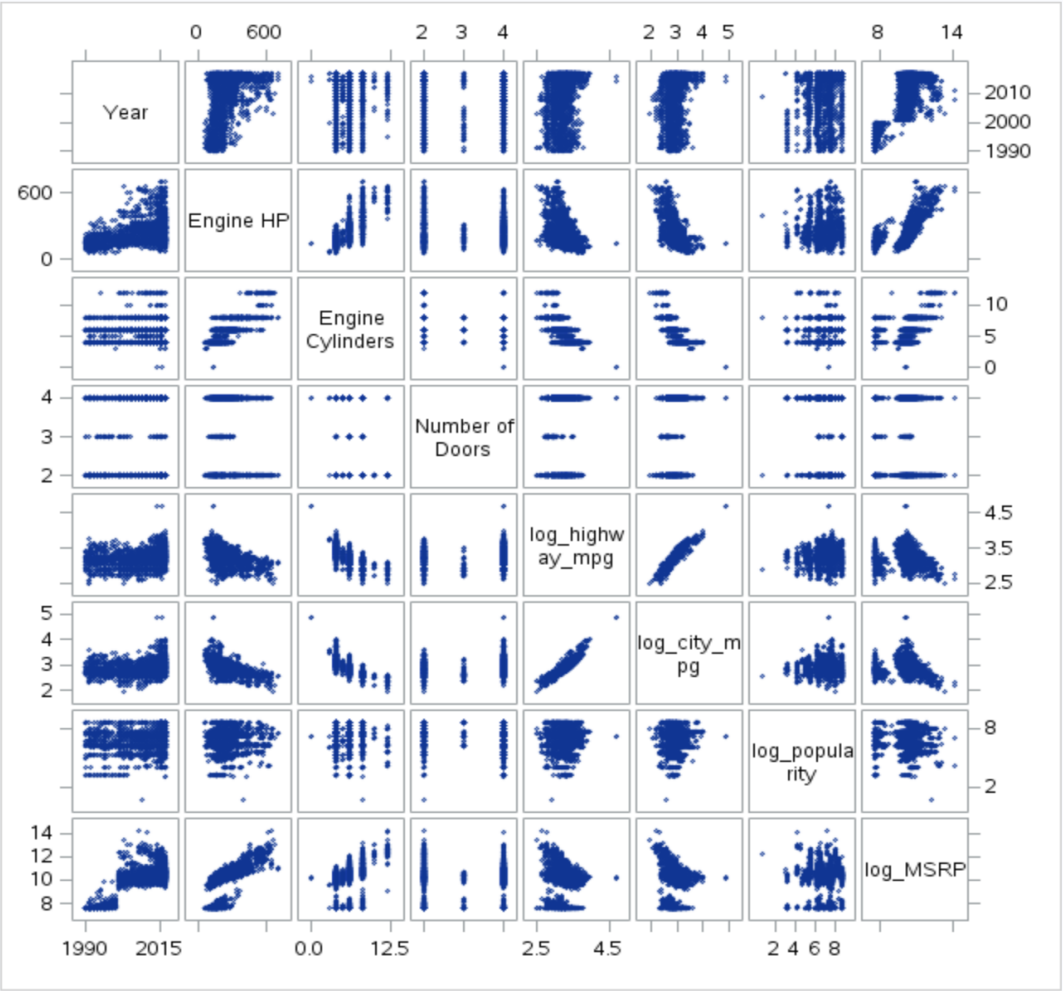
Least squares regression analysis is not resistant to outliers. Fitting a model with extreme outliers and leverage points may result in poorly-fit least squares regressions and less accurate conclusions. Examining the leverage plot and the Cook's D plot (Display 1.10) of the log transformed data, there are two data points with high leverage. However, since the largest Cook's D value is much less than 1 and there is no reason to believe that there are any recording errors, we will keep those two data points and proceed to the next step.

Display 1.10 RStudent by leverage plot (A) and Cook's D (B) plot from the regression of log_MSRP on the continuous variables (popularity, highway mpg, and city mpg are log transformed).



The scatterplot matrix of the log transformed response variable and the continuous variables in Display 1.11 shows a correlation between Engine HP and Engine Cylinders. There is even a stronger correlation between log_highway_mpg and log_city_mpg. We ran a proc reg to obtain variance inflation factors (VIFs) to check for multicollinearity. If the VIFs between two predictors in each pair are high, one of them can be removed to simplify the model.

Display 1.11 Scatterplot matrix of log_MSRP and the continuous variables



The VIFs for Engine HP and Engine Cylinders are small (Display 1.12). The VIF for log_city_mpg is the highest and it's above 10. Since the log_city_mpg is intercorrelated with log_highway_mpg, we have decided to drop log_city_mpg when we use different model selection methods in the next section. After the variable is removed, the VIF for log_highway_mpg drops from 9.50 to 2.57 (Display 1.13).

Display 1.12 Parameter estimates and VIFs from the regression of log_MSRP and the continuous variables

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-191.61414	3.54311	-54.08	<.0001	0
Year	1	0.10038	0.00182	55.15	<.0001	2.25442
Engine HP	1	0.00342	0.00018208	18.80	<.0001	4.62590
Engine Cylinders	1	0.06519	0.01145	5.69	<.0001	4.61167
Number of Doors	1	-0.03748	0.01121	-3.34	0.0008	1.16605
log_highway_mpg	1	-0.05367	0.11785	-0.46	0.6488	9.50164
log_city_mpg	1	-0.15657	0.11540	-1.36	0.1750	11.41668
log_popularity	1	-0.07889	0.00891	-8.86	<.0001	1.06609

Display 1.13 Parameter estimates and VIFs from the regression of log_MSRP and the continuous variables (after removing log_city_mpg)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-190.94019	3.50862	-54.42	<.0001	0
Year	1	0.10003	0.00180	55.51	<.0001	2.20870
Engine HP	1	0.00348	0.00017703	19.66	<.0001	4.37133
Engine Cylinders	1	0.06754	0.01132	5.96	<.0001	4.50648
Number of Doors	1	-0.03767	0.01122	-3.36	0.0008	1.16587
log_highway_mpg	1	-0.19019	0.06135	-3.10	0.0020	2.57431
log_popularity	1	-0.08020	0.00886	-9.05	<.0001	1.05359

When we perform the overall F-test (Display 1.14) after the removal of log_city_mpg, there is a strong evidence that at least one regression coefficient is different from 0 (p-value <0.001) at the 0.05 significance level. Given that the model assumptions are met, we refer to the individual t-tests in Display 1.13 and it indicates that the remaining six continuous variables are all statistically significant.

Display 1.14 Overall F-Test on the regression of log_MSRP and the continuous variables (after removing log_city_mpg)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2961.89422	493.64904	1947.16	<.0001
Error	2974	753.97711	0.25352		
Corrected Total	2980	3715.87133			

Model Selection

We include the six continuous variables that are statistically significant from the individual t-tests and all the categorical variables in the model selection. We choose to use the CV (cross validation) as our stop criterion over AIC (Akaike Information Criterion) and SBC (Schwarz Bayesian Criterion) because AIC tends to choose more complex models by including more variables and SBC heavily penalizes for parameters that are highly correlated. Since we had two variables that are highly correlated and we removed one of them for a VIF >10, this criteria is less appropriate. We use “CVMETHOD=SPLIT(5)” option to do five-fold cross validation and “STATS=PRESS” option to obtain the CV Press from each step of the process.

Display 1.15 Summary of Fit Table

Selection Method	Predictors	Adjusted R^2	CV PRESS
Forward Selection	Model	0.9612	635.92
Backward Selection	Make, Model, Year, Engine Fuel Type, Engine HP, Engine Cylinders, Transmission Type, Driven_Wheels, Number of Doors, Vehicle Size, Vehicle Style, log_highway_mpg	0.9797	427.28
Stepwise Selection	Model	0.9612	635.92
Lasso Selection	Make, Model, Year, Engine Fuel Type, Engine HP, Engine Cylinders, Transmission Type, Driven_Wheels, Market Category, Vehicle Style, log_highway_mpg, log_popularity	0.8658	326.52

Forward and the stepwise variable selection methods yield the same explanatory variable, Model, with the highest CV Press value as shown in Display 1.15. Backward variable selection results in more predictors and a lower CV Press value. Lastly, Lasso variable selection has the lowest CV Press value of 326.52. The regression model selected by the forward/stepwise selection method is quite simple and easy to interpret, however, it will have lower predictability than the model selected by Lasso, which has high predictability but is difficult to interpret.

Now that we know which explanatory variables are statistically significant according to the Lasso model we want to perform analysis on the parameter estimates. To do this we must run our model on a different statistical function in SAS (from glmselect to glm) to obtain confidence intervals on our parameter estimates.
Once parameter estimates were generated in glm, we observed this function included all categorical variables in contrast to glmselect which only included categorical variables that were statistically significant. When we included Make in the model, a total of 721 variations were now included causing a decrease the statistical significance of all other variables, and completely removing log_popularity as an estimate.

Parameter Interpretation

It would not be practical to discuss each categorical variable levels' estimate and confidence interval, so we've taken partial screenshots and will discuss a few samples for the categorical variables with a large number of variations.

Display 1.16 Top partial parameter estimate and the 95% Confidence Intervals on the regression model selected by Lasso

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-61.08463286	B	6.41060719	-9.53	<.0001	-73.65620407	-48.51306164
Make Acura	-0.00374380	B	0.23862460	-0.02	0.9875	-0.47170042	0.46421282
Make Alfa Romeo	0.11523134	B	0.26452107	0.44	0.6632	-0.40350976	0.63397244
Make Aston Martin	1.16608484	B	0.22326055	5.22	<.0001	0.72825799	1.60391169

Make [Categorical]
Acura and Alfa Romeo yields a non-significant p-value of 0.9875, however, since glm includes it as an estimate we will continue with its interpretation.

If a car has a Make of Acura then it's MSRP, holding all other variables constant, results in a e^(-0.00374) = 0.996 multiplicative change in the median MSRP. That translates to a 0.4% decrease in the median MSRP. The 95% confidence interval is contained between a multiplicative change of [e^-0.472, e^0.464] = [0.624, 1.587]. This is practically insignificant and considering it isn't statistically significant, it should be dropped from our final model.

If a car has a Make of Aston Martin (very significant at p-value of < 0.001) then it's MSRP, holding all other variables constant, results in a e^(1.166) = 3.209 multiplicative change in the median MSRP. That translates to a 220.9% increase in the median MSRP. The 95% confidence interval is contained between a multiplicative change of [e^0.728, e^1.604] = [2.071, 4.973].

Model [Categorical]

Model 1 Series	-0.21272057	B	0.12683147	-1.68	0.0936	-0.46144442	0.03600328
Model 100	-2.60116875	B	0.20325358	-12.80	<.0001	-2.99976077	-2.20257673
Model 124 Spider	0.11150964	B	0.21205764	0.53	0.5990	-0.30434766	0.52736694

There are 721 different model variations.

If a car is a Model 1 Series it's effect (coefficient/beta) is not found to be significant at a p-value of 0.0930 at α=0.05. The same is true for a Model 124 Spider with a p-value of 0.5990 at α=0.05. then it's natural log of the MSRP will decrease by .02127. Both should be dropped from the final model.

If a car is a Model 100 (very significant at p-value of < 0.001) then it's MSRP, holding all other variables constant, results in a e^(-2.6011) = 0.0742 multiplicative change in the median MSRP. That translates to a 92.58% decrease in the median MSRP. The 95% confidence interval is contained between a multiplicative change of [e^-2.99976, e^-2.2025] = [0.05, 0.11]. Along with being statistically significant, this has enormous practical significance.

Year [Continuous]

Year	0.03553698		0.00321620	11.05	<.0001	0.02922983	0.04184413

For every one year increase in age of a car (very significant at p-value of < 0.001) it is associated with a multiplicative change of e^(0.0355) = 1.0362 in the median MSRP when we hold all other variables constant. This translates to a 3.62% increase in the median MSRP for each year a car was made closer to the present year. The 95% confidence interval is contained between a multiplicative change of [e^0.0292, e^0.0418] = [1.03, 1.04].

Engine Fuel Type [Categorical]

Engine Fuel Type diesel	0.24810586	B	0.09940794	2.50	0.0126	0.05316114	0.44305057
Engine Fuel Type electric	0.00000000	B	.	.	.	.	.
Engine Fuel Type flex-fuel (premium unleaded re	0.28426569	B	0.11888050	2.39	0.0169	0.05113414	0.51739723
Engine Fuel Type flex-fuel (unleaded/E85)	-0.01146168	B	0.03052923	-0.38	0.7074	-0.07133126	0.04840789
Engine Fuel Type natural gas	0.40850424	B	0.16196814	2.52	0.0117	0.09087537	0.72613312
Engine Fuel Type premium unleaded (recommended)	0.13506161	B	0.02462414	5.48	<.0001	0.08677225	0.18335097
Engine Fuel Type premium unleaded (required)	0.24236653	B	0.02934062	8.26	<.0001	0.18482788	0.29990518
Engine Fuel Type regular unleaded	0.00000000	B	.	.	.	.	.

Engine HP [Continuous]

Engine HP	0.00117546		0.00018360	6.40	<.0001	0.00081542	0.00153551

For every one horsepower increase in the engine of a car (very significant at p-value of < 0.001) it is associated with a multiplicative change of e^(0.00118) = 1.0012 in the median MSRP when we hold all other variables constant. This translates to a 0.12% increase in the median MSRP for every one horsepower increase in the engine of a car. The 95% confidence interval is contained between a multiplicative change of [e^0.000815, e^0.000154] = [1.0008, 1.0015]. Although an increase of 0.12% increase in the median MSRP for every horsepower added to a car's engine seems insignificant, with a range from 55-707 HP we believe this proves to be practically significant. An example would be a $30,000 car increasing its horsepower by 50, would add $1,853.

Engine Cylinders [Continuous]

Engine Cylinders	0.01077114		0.00888261	1.21	0.2254	-0.00664817	0.02819045

For every additional engine cylinder it is associated with a e^(.0108) = 1.0108 multiplicative change in the median MSRP, when we hold all other variables constant. This translates to a 1.08% increase in median MSRP per every additional engine cylinder in the car. With a p-value of 0.2254 we would exclude this variable.

Transmission Type [Categorical]

Transmission Type AUTOMATED	0.44687098	B	0.09212815	4.85	<.0001	0.26620235	0.62753961
Transmission Type AUTOMATIC	0.46627817	B	0.08917653	5.23	<.0001	0.29139785	0.64115849
Transmission Type DIRECT_DR	0.51527345	B	0.20860094	2.47	0.0136	0.10619494	0.92435195
Transmission Type MANUAL	0.39189221	B	0.08860961	4.42	<.0001	0.21812364	0.56566077
Transmission Type UNKNOWN	0.00000000	B	.	.	.	.	.

As an example, if a vehicle is Automatic, it would have a multiplicative effect of e^(0.466) = 1.5941 on the median MSRP when we hold all other variables constant (p-value < 0.0001). This translates to a 59.41% increase in the median MSRP. It's interesting to note the similarity in the parameter estimate of Automated and Automatic, with the possibility this category covers the same transmission type.

Driven Wheels [Categorical]

Driven_Wheels all wheel drive	0.05626054	B	0.01939402	2.90	0.0038	0.01822775	0.09429332
Driven_Wheels four wheel drive	0.05391106	B	0.01526629	3.53	0.0004	0.02397298	0.08384914
Driven_Wheels front wheel drive	0.02686275	B	0.02366187	1.14	0.2564	-0.01953954	0.07326504
Driven_Wheels rear wheel drive	0.00000000	B	.	.	.	.	.

As an example, if a vehicle has all-wheel drive, it would have a multiplicative effect of e^(0.0562) = 1.7552 on the median MSRP when all other variables are held constant (significant at p-value = 0.0038). This translates to a 75.52% increase in median MSRP.

Market Category [Categorical]

Market Category Exotic,Luxury,High-Performance	-0.03256209	B	0.14193928	-0.23	0.8186	-0.31091322	0.24578904
Market Category Exotic,Luxury,High-Performance,Hybrid	-0.60181667	B	0.24862210	-2.42	0.0156	-1.08937899	-0.11425436
Market Category Exotic,Luxury,Performance	0.00000000	B	.	.	.	.	.

As an example, if a vehicle is in the market category of Exotic, Luxury, High-Performance, Hybrid, this would result in a multiplicative effect of e^(-0.602) = 0.5478 on the median MSRP when all other variables are held constant (p-value = 0.0156). This translates to a 45.22% decrease in the median MSRP. This category is seen to have redundancy due to the categorization of how cars are described and malleable in interpretation. Due to this, in custom models this category would likely need to be parsed better than the data was already, or dropped entirely.

Vehicle Style [Categorical]

Vehicle Style Cargo Van	-0.01163744	B	0.19359103	-0.06	0.9521	-0.39128065	0.36800576
Vehicle Style Convertible	0.10669875	B	0.04263441	2.50	0.0124	0.02309021	0.19030729
Vehicle Style Coupe	-0.00783407	B	0.04078322	-0.19	0.8477	-0.08781232	0.07214417

Whether a vehicle is a convertible has a statistical significance (p-value = 0.0124) and a multiplicative change of e^(.10669) = 1.1125 on the median MSRP. For a convertible as the vehicle style there is an increase in the median MSRP by 11.25%. This level of the categorical variable is both statistically and practically significant.

Log_highway_mpg [Continuous]

log_highway_mpg	-0.21970078		0.07686778	-2.86	0.0043	-0.37044293	-0.06895863

For a doubling of a car's highway MPG, it is associated with a 2^(-0.21970078) = 0.8587 multiplicative increase in the median MSRP. This translates to a 14.13% decrease in median MSRP as a car's highway MPG doubles. This is found to be statistically significant with a p-value of 0.0043 at α=0.05. This is likely due nearly all of the high MSRP luxury high performance cars having some of the lowest highway MPG out of the entire data set.

Log_popularity [Continuous]

log_popularity	0.00000000	B	.	.	.	.	.

Intercept

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-61.08463286	B	6.41060719	-9.53	<.0001	-73.65620407	-48.51306164

The intercept estimates the difference of log(MSRP) for all dummy variables holding continuous variables at 0. Oddly when including all Models, Makes and Market Category in the model, the log(popularity) degrees of freedom reduces to 0 and has no parameter estimate. Additionally there are multiple categorical variables that have multiple dummy variables. When Make, Model, and Market Category are taken out of the model the resulting parameter estimates resemble a more typical model with one dummy variable per categorical variable and log(popularity) returns a non-zero parameter estimate that is statistically significant.

When all variables (categorical and continuous) are held at zero, the median MSRP is e^(61.08) = 2.96x10^-27. The y-intercept is very statistically significant with a p-value of < 0.001. This makes practical sense in that if we have no car, the median MSRP will be zero.

Conclusion/Discussion

Based on our analysis, we conclude that the best model for predicting the MSRP of cars is the one selected by Lasso. Although it is a complex least square regression model, it has the highest predictability with the lowest CV Press value of the models we tested. Alternatively, the model generated by forward and stepwise variable selection could be used to predict the MSRP with easier interpretation but with low predictable power due to its high CV Press.

For future studies it would be important to research more on how to handle categorical variables with a large number of levels, hundreds in our case. The large number of categorical variables included individual parameter estimates that were not statistically significant but were included in the model produced by glm. Our final model would exclude them.

For future models it would be worth attempting two different models, one for each century. This is because of the anomalous gap in MSRP between cars made in this century (>=2000) and the last (<2000) in which no transformation was evident that could correct for this.

Objective 2 - Two-way ANOVA

We are going to run a two-way analysis of variance (ANOVA) on our continuous response variable, log transformed MSRP, with relation to two explanatory categorical variables, Year and Engine Fuel Type. The goal is to identify any potential interaction between the explanatory variables and to determine if they are statistically significant in predicting the response variable log(MSRP). We wanted to test the null hypothesis that there is no difference in the populations log_MSRP when grouped by year and engine fuel type.

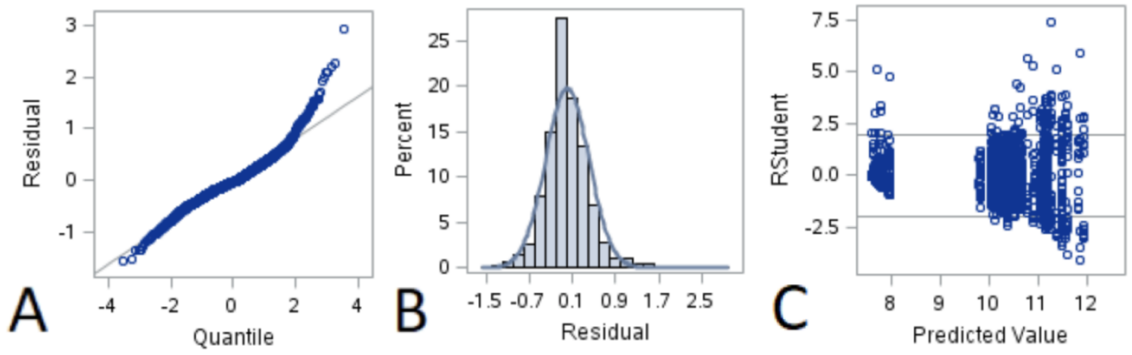
Model Assumption

In this section, we will assess whether the model's assumptions required for two-way ANOVA are met and whether if there are outliers/influential observations that need to be corrected.

Based on the Q-Q plot and the histogram of the residuals for log_MSRP (Display 2.1), there is mild evidence that the residuals are right-skewed. However, since the sample size is large, the data is robust to the normality assumption because of the central limit theorem and we will assume that the assumption is met.

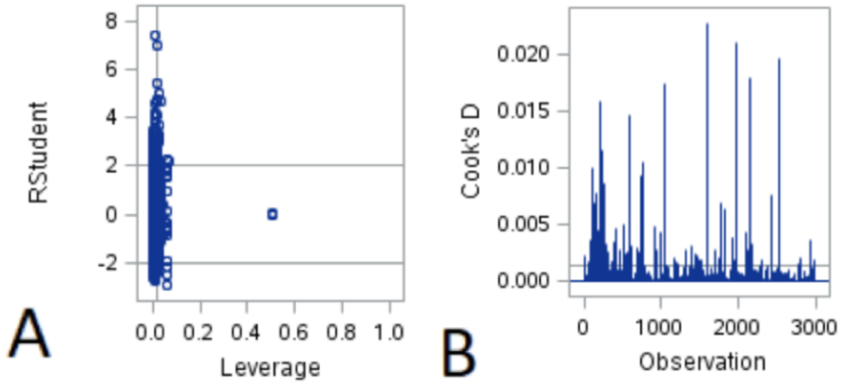
The studentized residual plot in Display 2.1 shows evidence of non-constant variance, there are few residuals on the lower end and another dense cloud of residuals on the right which has a funnel shape. Since our objective here is to predict the MSRP of cars using the data, and not for testing, it's not important if the residual assumption is not met.
Additionally, there is no reason to believe that these records are dependent on each other, thus we will assume that the data are independent.

Display 2.1 Q-Q plot (A), histogram of residuals (B) and studentized residual plots (C) from the regression of log_MSRP on the two categorical variables.



Examining the leverage plot and the Cook's D plot (Display 2.2), there is one data point with high leverage. However, since the largest Cook's D value is relatively small and there is no reason to believe that there are any recording errors, we will keep the data point and proceed to the next step.

Display 2.2 RStudent by leverage plot (A) and Cook's D (B) plot from the regression of log_MSRP on the two categorical variables



Analysis

We fit the two-way ANOVA model on the log_MSRP with two categorical variables, year and engine fuel type. Display 2.3 shows that there is a strong evidence that at least one year of car produced and/or at least one engine type group has a different log_MSRP than the others (p-value < .0001) at the significant level of 0.05.

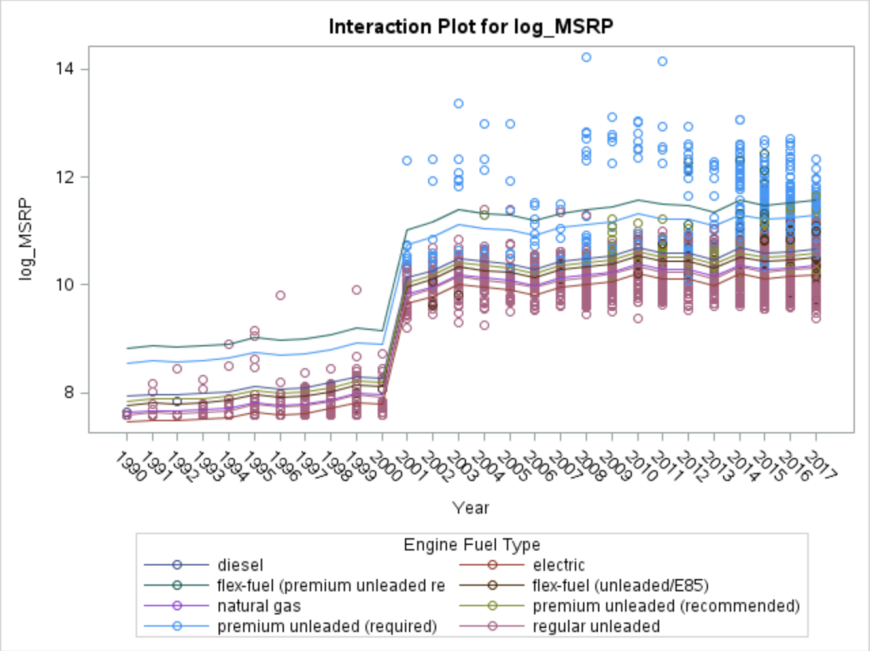
Display 2.3 ANOVA table from the regression of log_MSRP on year and engine fuel type.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	34	3188.897602	93.791106	524.33	<.0001
Error	2946	526.973733	0.178878		
Corrected Total	2980	3715.871335			

Next, we check the interaction plot to see if there is a significant interaction between year and engine fuel type. In our case, an interaction would be one in which the engine fuel type is depended on the year of

car produced. Since all the lines in the interaction plot (Display 2.4) are close to parallel, there is no evidence of an interaction effect. Additionally, we do not have replicate observations at each interaction level, which means we cannot access the interaction term. Thus, we will proceed to fit the two-way ANOVA model without the interaction term.

Display 2.4 Interaction Plot for two categorical variables (year and engine fuel type).



Both year and engine fuel type are statistically significant predictors of MSRP. There is a strong evidence that, after accounting for the years of car produced, there is a significant difference (Display 2.5, p-value < 0.0001) in median MSRP between at least one pair of engine fuel types. Similarly, if we account for the engine fuel types, there is a strong evidence that there is a significant difference (Display 2.5, p-value < 0.0001) in median MSRP between at least one pair of years of car manufactured.

Display 2.5 Type 3 Tests of Fixed Effects.

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Year	27	2946	426.01	<.0001
Engine Fuel Type	7	2946	296.15	<.0001

Next we will run pairwise comparisons to determine which pairs are statistically significant. Since we are doing multiple comparisons, we will apply Bonferroni adjustment to adjust the p-values and the confidence intervals in order to reduce Type I error.

The table below indicates that there are statistically significant differences between some pairs of year the car was produced and not the others. For instance, the difference in the median MSRP between 1990 and 2001 is estimated to be a multiplicative factor of e^(-2.1825) = 0.113. This translates to a 88.72% decrease in the median MSRP. We are 95% confident that the median MSRP of 1990 made car is estimated to decrease by between 92.5% and 83% (e^-2.5906, e^-1.7943 = 0.075, 0.17) compared to the median MSRP of 2001 made car (after a Bonferroni adjustment).

Similarly, for different pairs of engine fuel type, there are some pairs that are statistically significant. For instance, we are 95% confident that the median MSRP of diesel car is estimated to be between 13% and 77% (e^0.1179, e^0.5734 = 1.13, 1.77) more than the median MSRP of car with regular unleaded engine fuel type (after a Bonferroni adjustment).

Display 2.6 Partial differences of Least Squares Means table

Differences of Least Squares Means																
Effect	Engine Fuel Type	Year	_Engine Fuel Type	_Year	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Year		1990		1991	-0.04171	0.1044	2946	-0.40	0.6895	Bonferroni	1.0000	0.05	-0.2463	0.1629	-0.4411	0.3577
Year		1990		1992	-0.02980	0.1053	2946	-0.28	0.7773	Bonferroni	1.0000	0.05	-0.2364	0.1768	-0.4330	0.3734
Year		1990		1993	-0.04848	0.1004	2946	-0.48	0.6291	Bonferroni	1.0000	0.05	-0.2453	0.1483	-0.4326	0.3356
Year		1990		1994	-0.08540	0.1014	2946	-0.84	0.3998	Bonferroni	1.0000	0.05	-0.2842	0.1134	-0.4735	0.3027
Year		1990		1995	-0.1975	0.1077	2946	-1.83	0.0668	Bonferroni	1.0000	0.05	-0.4087	0.01365	-0.6098	0.2147
Year		1990		1996	-0.1466	0.1071	2946	-1.37	0.1710	Bonferroni	1.0000	0.05	-0.3566	0.06333	-0.5565	0.2632
Year		1990		1997	-0.1701	0.1022	2946	-1.67	0.0959	Bonferroni	1.0000	0.05	-0.3705	0.03018	-0.5612	0.2209
Year		1990		1998	-0.2487	0.1077	2946	-2.31	0.0210	Bonferroni	1.0000	0.05	-0.4599	-0.03754	-0.6610	0.1635
Year		1990		1999	-0.3675	0.1114	2946	-3.30	0.0010	Bonferroni	0.3722	0.05	-0.5860	-0.1490	-0.7940	0.05896
Year		1990		2000	-0.3386	0.1174	2946	-2.88	0.0040	Bonferroni	1.0000	0.05	-0.5689	-0.1083	-0.7881	0.1109
Year		1990		2001	-2.1925	0.1040	2946	-21.08	<.0001	Bonferroni	<.0001	0.05	-2.3964	-1.9885	-2.5906	-1.7943
Year		1990		2002	-2.3336	0.1006	2946	-23.20	<.0001	Bonferroni	<.0001	0.05	-2.5309	-2.1364	-2.7186	-1.9487
Year		1990		2003	-2.5626	0.1005	2946	-25.50	<.0001	Bonferroni	<.0001	0.05	-2.7597	-2.3655	-2.9473	-2.1779
Engine Fuel Type	diesel		electric		0.4783	0.3079	2946	1.55	0.1204	Bonferroni	1.0000	0.05	-0.1254	1.0820	-0.4843	1.4409
Engine Fuel Type	diesel		flex-fuel (premium unleaded re		-0.8926	0.1236	2946	-7.22	<.0001	Bonferroni	<.0001	0.05	-1.1349	-0.6503	-1.2789	-0.5063
Engine Fuel Type	diesel		flex-fuel (unleaded/E85)		0.1607	0.07770	2946	2.07	0.0387	Bonferroni	1.0000	0.05	0.008360	0.3130	-0.08222	0.4036
Engine Fuel Type	diesel		natural gas		0.3038	0.4292	2946	0.71	0.4790	Bonferroni	1.0000	0.05	-0.5377	1.1454	-1.0381	1.6458
Engine Fuel Type	diesel		premium unleaded (recommended)		0.08094	0.07497	2946	1.08	0.2804	Bonferroni	1.0000	0.05	-0.06605	0.2279	-0.1534	0.3153
Engine Fuel Type	diesel		premium unleaded (required)		-0.6208	0.07418	2946	-8.37	<.0001	Bonferroni	<.0001	0.05	-0.7662	-0.4753	-0.8527	-0.3889
Engine Fuel Type	diesel		regular unleaded		0.3456	0.07284	2946	4.75	<.0001	Bonferroni	<.0001	0.05	0.2028	0.4884	0.1179	0.5734

Conclusion/Discussion

We have found that we could not assess the interaction term because we do not have replicate observations at each interaction. Additionally, when viewing the mean plot (Display 2.4) without the interaction term, we do not see significant deviation from a parallel pattern among the groups which suggests that this is an additive model.

When running our model without the interaction term, we find that we reject the null hypothesis (Display 2.3: p-value < 0.001, α=0.05). There is a difference in at least one of the groups.

Given the large sample size overall in the two-way ANOVA analysis, we might detect the small differences that are statistically significant although there is no actual practical significance.

However, we do find that both year and engine fuel type are statistically significant. More recently produced cars commanding a higher median log(MSRP), and flex fuel premium commanding a higher median log(MSRP) relative to other fuel types. The two-way ANOVA model is a less complicated and easily interpretable model that can be used to predict the MSRP of cars. However, the predictability level will not be as strong as the model selected by LASSO in the previous section.

Appendix

SAS Code of Analysis:

```
/* Importing car data */
proc import datafile= "/home/schew0/Swee_SMU/car_data.csv"
    dbms= csv replace
    out= car_data_set;
    getnames= yes;
    guessingrows=60;
run;

/* Generate simple random sample of n=3000 */
proc surveyselect data=car_data_set seed=1212
method=srs n=3000 out=car_data1;
run;

/* Remove all records that have null values */
data car_data2;
set car_data1;
if cmiss(of _all_) then delete;
run;

/* Run proc means to get summary statistics on continous variables */
proc means data=car_data2;
run;

proc means data=car_data2;
class Make;
var MSRP;
run;

proc means data=car_data2;
class 'Engine Fuel Type'n;
var MSRP;
run;

/* To obtain residual plots before log transformation of MSRP */
proc reg data=car_data2 PLOTS=ALL PLOTS;
model MSRP= Year 'Engine HP'n 'Engine Cylinders'n 'Number of Doors'n 'highway MPG'n 'city mpg'n
Popularity;
title 'Regression of (MSRP)';
run;

data car_data3;
set car_data2;
log_MSRP=log(MSRP);
run;

/* To obtain residual plots after log transformation of MSRP */
proc reg data=car_data3 PLOTS=ALL PLOTS;
```

```
model log_MSRP= Year 'Engine HP'n 'Engine Cylinders'n 'Number of Doors'n 'highway MPG'n 'city mpg'n
Popularity;
title 'Regression of (MSRP)';
run;

/* Take log of skewed variables highway_mpg, city_mpg, and popularity */
data car_data4;
set car_data3;
log_highway_mpg=log('highway MPG'n);
log_city_mpg=log('city mpg'n);
log_popularity=log(popularity);
run;

/* matrix scatterplot */
/* proc sgscatter data = car_data4; */
/* title "Scatterplot Matrix of Car Variables"; */
/* matrix Year 'Engine HP'n 'Engine Cylinders'n 'Number of Doors'n 'highway MPG'n 'city mpg'n Popularity
log_MSRP; */
/* run; */
/* title; */

/* To obtain residual plots and VIFs after log transformation of the predictors - highway_mpg city_mpg
popularity */
proc reg data=car_data4 PLOTS=ALL PLOTS;
model log_MSRP= Year 'Engine HP'n 'Engine Cylinders'n 'Number of Doors'n log_highway_mpg
log_city_mpg log_popularity /clb VIF;
title 'Regression of (MSRP)';
run;

/* To obtain residual plots and VIFs after removing log_city_mpg */
proc reg data=car_data4 PLOTS=ALL PLOTS;
model log_MSRP= Year 'Engine HP'n 'Engine Cylinders'n 'Number of Doors'n log_highway_mpg
log_popularity /clb VIF;
title 'Regression of (MSRP)';
run;

/*matrix scatterplot after log transformation*/
proc sgscatter data = car_data4;
/* title "Scatterplot Matrix of Car Variables"; */
matrix Year 'Engine HP'n 'Engine Cylinders'n 'Number of Doors'n log_highway_mpg log_city_mpg
log_Popularity log_MSRP;
run;
title;

/* Forward selection */
proc glmselect data = car_data4 plots(stepaxis = number) = (criterionpanel ASEPlot) seed = 1;
class Make Model 'Engine Fuel Type'n 'Transmission Type'n 'Driven_Wheels'n 'Market Category'n
'Vehicle Size'n 'Vehicle Style'n;
model log_MSRP=  Make Model Year 'Engine Fuel Type'n 'Engine HP'n 'Engine Cylinders'n
'Transmission Type'n 'Driven_Wheels'n 'Number of Doors'n 'Market Category'n 'Vehicle Size'n 'Vehicle
```

```
Style'n log_highway_mpg log_popularity / selection = forward(choose=CV stop=CV) cvmethod=split(5)
cvdetails=cvpress stats=press;
run; quit;
;
run;

/* Backward selection */
proc glmselect data = car_data4 plots(stepaxis = number) = (criterionpanel ASEPlot) seed = 1;
class Make Model 'Engine Fuel Type'n 'Transmission Type'n 'Driven_Wheels'n 'Market Category'n
'Vehicle Size'n 'Vehicle Style'n;
model log_MSRP=  Make Model Year 'Engine Fuel Type'n 'Engine HP'n 'Engine Cylinders'n
'Transmission Type'n 'Driven_Wheels'n 'Number of Doors'n 'Market Category'n 'Vehicle Size'n 'Vehicle
Style'n log_highway_mpg log_popularity / selection = backward(choose=CV stop=CV) cvmethod=split(5)
cvdetails=cvpress stats=press;
run; quit;
run;

/* Stepwise selection */
proc glmselect data = car_data4 plots(stepaxis = number) = (criterionpanel ASEPlot) seed = 1;
class Make Model 'Engine Fuel Type'n 'Transmission Type'n 'Driven_Wheels'n 'Market Category'n
'Vehicle Size'n 'Vehicle Style'n;
model log_MSRP=  Make Model Year 'Engine Fuel Type'n 'Engine HP'n 'Engine Cylinders'n
'Transmission Type'n 'Driven_Wheels'n 'Number of Doors'n 'Market Category'n 'Vehicle Size'n 'Vehicle
Style'n log_highway_mpg log_popularity / selection = stepwise(choose=CV stop=CV) cvmethod=split(5)
cvdetails=cvpress stats=press;
run; quit;
run;

/* Lasso selection */
proc glmselect data = car_data4 plots(stepaxis = number) = (criterionpanel ASEPlot) seed = 1;
class Make Model 'Engine Fuel Type'n 'Transmission Type'n 'Driven_Wheels'n 'Market Category'n
'Vehicle Size'n 'Vehicle Style'n;
model log_MSRP= Make Model Year 'Engine Fuel Type'n 'Engine HP'n 'Engine Cylinders'n 'Transmission
Type'n 'Driven_Wheels'n 'Number of Doors'n 'Market Category'n 'Vehicle Size'n 'Vehicle Style'n
log_highway_mpg log_popularity / selection = lasso(choose=CV stop=CV) cvmethod=split(5)
cvdetails=cvpress stats=press;
run;

/* to obtain parameter estimates and CIs*/
proc glm data = car_data4 plots=all ;
class Make Model 'Engine Fuel Type'n 'Transmission Type'n 'Driven_Wheels'n 'Market Category'n
'Vehicle Style'n;
model log_MSRP= Make Model Year 'Engine Fuel Type'n 'Engine HP'n 'Engine Cylinders'n
'Transmission Type'n 'Driven_Wheels'n'Market Category'n 'Vehicle Style'n log_highway_mpg
log_popularity/ solution CLPARM cli clm;
output out=residual_data residual=resid ;
run;

/* Proc glm 2-Way anova Analysis without interaction */
proc glm data= car_data4 plots =(DIAGNOSTICS RESIDUALS);
class Year 'Engine Fuel Type'n;
```

```
model log_MSRP=  Year 'Engine Fuel Type'n;
lsmeans Year / pdiff tdiff cl adjust=bon;
run;

/* Proc mixed 2-Way anova Analysis without interaction */
proc mixed data= car_data4 plots =(ResidualPanel);
class Year 'Engine Fuel Type'n;
model log_MSRP=  Year 'Engine Fuel Type'n;
lsmeans Year 'Engine Fuel Type'n / cl adjust=bon;
run;
```

Output from Forward/Stepwise Selection:

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 1.

Effects:	Intercept Model

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	720	3606.59134	5.00915	103.59
Error	2260	109.27999	0.04835	
Corrected Total	2980	3715.87133		

Root MSE	0.21990
Dependent Mean	10.10774
R-Square	0.9706
Adj R-Sq	0.9612
AIC	-5430.48638
AICC	-4968.12500
PRESS	185.96654
SBC	-4087.47622
CV PRESS	635.91666

Output from Backward Selection:

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 1.

Effects:	Intercept Make Model Year Engine Fuel Type Engine HP Engine Cylinders Transmission Type Driven_Wheels Number of Doors Vehicle Size Vehicle Style log_highway_mpg

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	759	3659.67437	4.82171	190.56
Error	2221	56.19697	0.02530	
Corrected Total	2980	3715.87133		

Root MSE	0.15907
Dependent Mean	10.10774
R-Square	0.9849
Adj R-Sq	0.9797
AIC	-7335.00196
AICC	-6812.35032
PRESS	115.13288
SBC	-5757.99125
CV PRESS	427.27528

Output from Lasso Selection:

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 94.

Effects:	Intercept Make_Bentley Make_Cadillac Make_Ferrari Make_Land Rover Make_Maybach Make_Nissan Make_Oldsmobile Make_Plymouth Make_Pontiac Make_Porsche Make_Suzuki Model_350Z Model_360 Model_456M Model_850 Model_9-3 Model_900 Model_9000 Model_Aerostar Model_Alero Model_Aurora Model_B-Series Model_B-Series P Model_Blazer Model_C/K 1500 S Model_Cabrio Model_Camaro Model_Camry Sola Model_Carrera GT Model_Cutlass Model_Cutlass Su Model_Dakota Model_DeVille Model_E-150 Model_Eighty-Eig Model_Esteem Model_Explorer S Model_F-150 Heri Model_F-250 Model_Freelander Model_Freestyle Model_Grand Am Model_Grand Voya Model_J30 Model_LSS Model_LeSabre Model_Legend Model_Mark VIII Model_Montero Sp Model_NSX Model_New Yorker Model_Passport Model_Phaeton Model_RSX Model_RX 300 Model_Ram Van Model_Ramcharger Model_Reventon Model_S-10 Model_S-10 Blaze Model_S70 Model_Seville Model_Sidekick Model_Silhouette Model_Tundra Model_Venture Model_Windstar Model_XL-7 Model_XLR Model_Z3 Year Engine Fuel Type_diesel Engine Fuel Type_flex-fuel (premium unleaded re Engine Fuel Type_flex-fuel (unleaded/E85) Engine Fuel Type_premium unleaded (required) Engine Fuel Type_regular unleaded Engine HP Engine Cylinders Transmission Type_MANUAL Transmission Type_UNKNOWN Driven_Wheels_all wheel drive Market Category_Exotic,Factory Tuner,High-Performance Market Category_Exotic,Factory Tuner,Luxury,High-Perfor Market Category_Exotic,High-Performance Market Category_Exotic,Luxury,High-Performance Market Category_Exotic,Luxury,Performance Market Category_Factory Tuner,High-Performance Vehicle Style_4dr SUV Vehicle Style_Cargo Van Vehicle Style_Convertible Vehicle Style_Extended Ca Vehicle Style_Wagon log_highway_mpg log_popularity

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	94	3232.96521	34.39325	205.54
Error	2886	482.90613	0.16733	
Corrected Total	2980	3715.87133		

Root MSE	0.40906
Dependent Mean	10.10774
R-Square	0.8700
Adj R-Sq	0.8658
AIC	-2252.99180
AICC	-2246.53410
SBC	-4665.99046
CV PRESS	326.51648

Proc GLMSELECT without statistically insignificant factor levels:

The GLMSELECT Procedure

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Model R-Square	CV PRESS
0	Intercept		1	0.0000	3717.0033
1	Year		2	0.1773	1518.2922
2	Engine HP		3	0.7096	792.5008
3	Engine Fuel Type_premium unleaded (required)		4	0.7759	722.4622
4	Market Category_Exotic,High-Performance		5	0.7838	691.9585
5	Engine Fuel Type_regular unleaded		6	0.7884	689.7628
6	Transmission Type_MANUAL		7	0.7994	666.4859
7	Engine Cylinders		8	0.8065	651.8826
8	Vehicle Style_Convertible		9	0.8103	636.0717
9	Make_Ferrari		10	0.8131	630.4678
10	Market Category_Exotic,Luxury,High-Performance		11	0.8143	620.1809
11	Market Category_Exotic,Luxury,Performance		12	0.8160	609.8666
12	Make_Bentley		13	0.8228	608.1808
13	Vehicle Style_4dr SUV		14	0.8230	598.4225
14	Market Category_Exotic,Factory Tuner,High-Performance		15	0.8234	591.0001
15	Make_Maybach		16	0.8250	587.2035
16	Model_850		17	0.8266	582.1722
17	Model_Montero Sp		18	0.8273	576.0991
18	Model_Tundra		19	0.8275	569.7854
19	Make_Land Rover		20	0.8276	565.1370
20	log_popularity		21	0.8285	563.2608
21	Model_C/K 1500 S		22	0.8287	559.1729
22	Model_Passport		23	0.8304	552.4808
23	Driven_Wheels_all wheel drive		24	0.8316	550.9353
24	Model_Cabrio		25	0.8316	545.5019
25	Model_Carrera GT		26	0.8318	541.0700
26	Model_900		27	0.8320	536.7662
27	Model_F-250		28	0.8330	532.4372
28	Model_Mark VIII		29	0.8330	527.0579
29	Model_Windstar		30	0.8339	520.5804
30	Model_S70		31	0.8341	516.1692
31	Vehicle Style_Cargo Van		32	0.8348	511.4913
32	Make_Plymouth		33	0.8352	508.1285
33	Model_Venture		34	0.8358	502.8342
34	Engine Fuel Type_flex-fuel (unleaded/E85)		35	0.8381	498.3220
35	Model_Sidekick		36	0.8385	494.1102
36	Model_Reventon		37	0.8394	493.8248
37	Model_B-Series P		38	0.8410	490.2101
38	Market Category_Exotic,Factory Tuner,Luxury,High-Perfor		39	0.8412	486.7649
39	Model_RX 300		40	0.8413	482.7278
40	Market Category_Factory Tuner,High-Performance		41	0.8417	480.5854
41	Model_LS5		42	0.8418	476.2620
42	Make_Suzuki		43	0.8420	468.8216

43	Model_Explorer S		44	0.8424	464.0237
44	log_highway_mpg		45	0.8427	458.1950
45	Model_B-Series		46	0.8435	454.2522
46	Make_Cadillac		47	0.8435	448.7116
47	Model_Silhouette		48	0.8443	445.3071
48	Make_Pontiac		49	0.8446	439.6220
49	Make_Nissan		50	0.8455	436.8955
50	Model_J30		51	0.8462	433.7039
51	Make_Oldsmobile		52	0.8462	431.3048
52	Model_350Z		53	0.8463	422.4395
53	Model_NSX		54	0.8473	419.8041
54	Model_Esteem		55	0.8487	417.3011
55	Model_Cutlass Su		56	0.8492	415.8162
56	Model_E-150		57	0.8493	413.0932
57	Model_Grand Voya		58	0.8495	411.3360
58	Model_Ram Van		59	0.8497	409.3922
59	Model_9-3		60	0.8498	406.8011
60	Model_DeVille		61	0.8501	404.4565
61	Model_456M		62	0.8503	403.3250
62	Engine Fuel Type_flex-fuel (premium unleaded re		63	0.8511	401.7450
63	Model_9000		64	0.8511	399.3563
64	Transmission Type_UNKNOWN		65	0.8520	396.0350
65	Model_S-10		66	0.8526	392.5605
66	Model_Legend		67	0.8536	389.9858
67	Model_Grand Am		68	0.8537	388.6850
68	Model_F-150 Heri		69	0.8552	385.8082
69	Model_Camaro		70	0.8562	383.9849
70	Model_Seville		71	0.8572	381.2032
71	Engine Fuel Type_diesel		72	0.8576	378.5628
72	Make_Porsche		73	0.8578	375.0929
73	Model_Ramcharger		74	0.8586	371.4527
74	Model_New Yorker		75	0.8591	369.1920
75	Model_Aurora		76	0.8598	365.1243
76	Model_S-10 Blaze		77	0.8598	362.6873
77	Model_Blazer		78	0.8604	360.3660
78	Vehicle Style_Extended Ca		79	0.8610	359.6057
79	Model_Freestyle		80	0.8620	356.6836
80	Model_Dakota		81	0.8627	354.7257
81	Model_Freelander		82	0.8629	353.3618
82	Model_Z3		83	0.8636	351.1114
83	Model_Camry Sola		84	0.8636	347.8799
84	Vehicle Style_Wagon		85	0.8645	346.1439
85	Model_Cutlass		86	0.8646	344.9198
86	Model_XL-7		87	0.8655	343.8718
87	Model_Phaeton		88	0.8658	341.2872
88	Model_Eighty-Eig		89	0.8666	340.3675
89	Model_LeSabre		90	0.8667	338.0752
90	Model_360		91	0.8670	337.2309
91	Model_RSX		92	0.8683	334.6012
92	Model_Alero		93	0.8689	330.5920
93	Model_XLR		94	0.8700	328.5450
94	Model_Aerostar		95	0.8700	328.5165*