Multiple Linear Regression
Analyzing Hybrid Vehicle Pricing

Damon Resnick
12/12/2016

Introduction:

In this report, I will analyze some basic data relating to the sale price and performance of hybrid cars from 1997 to 2013. There are 153 observations in the data set with 7 variables consisting of four continuous variables and three categorical variables. Those variables include:

Continuous variables:

msrp: Manufacturer's suggested retail price in 2013 $ US dollars, *response* variable
accelerate: acceleration in km/hour/second, *explanatory* variable
mpg: miles per gallon, *explanatory* variable
mpgmpge: Max of mpg or MPGe = 33.7*driverange/batterycapacity, *explanatory* variable

- Since only a few cars can be run as fully electric plug ins this variable has only a few observations that are different than mpg.

Categorical *explanatory* variables:

vehicle: name of the vehicle
carclass: the type of car, classes include: C = Compact (32), M = Midsize (56), TS = Two Seater (8), L = Large (8), PT = Pickup Truck (6), MV = Minivan (4), SUV = Sport Utility Vehicle (39), (No. of observations for each level)
year: 1997-2013, with 1998 and 1999 missing. Presumably this is because the Prius was the only hybrid available to the public from 1997 to 2000.

Data can be found here:
http://www.stat.ufl.edu/~winner/datasets.html
http://www.stat.ufl.edu/~winner/data/hybrid_reg.txt
http://www.stat.ufl.edu/~winner/data/hybrid_reg.csv

Questions:

Two questions of interest in this analysis are:

1) What is the best model to predict future sales prices of hybrid vehicles based on this data set?

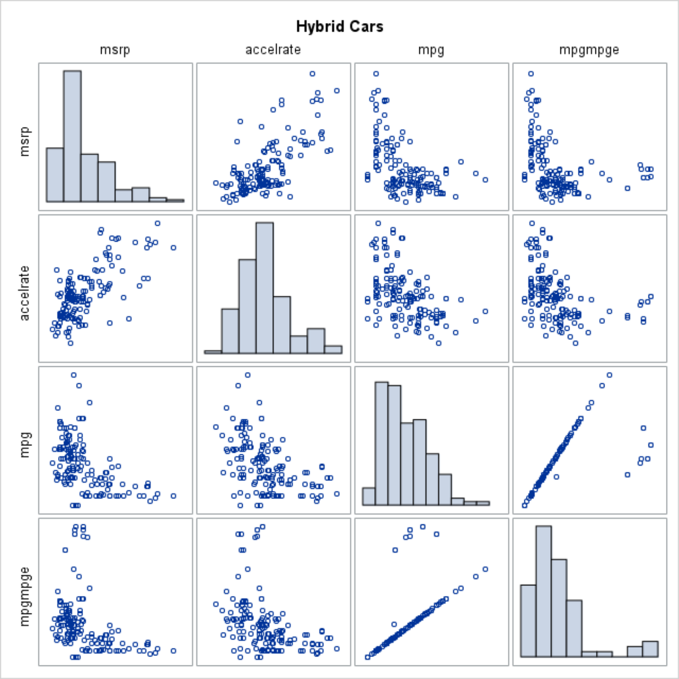
2) What is the best model to better understand the factors that have an effect on the sales price of a hybrid vehicle and do these factors change depending on the type of car?

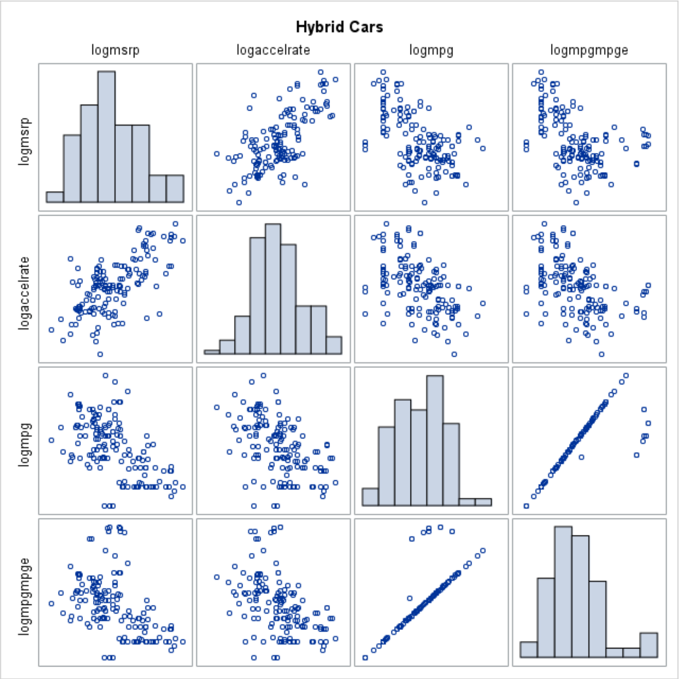
To find a good answer to question one, I will attempt to find the model that best fits the data taking into account the number of observations as well as the number of parameters in the model. I will mainly use the adjusted R squared value to compare different models.

To find a good answer to question two, I will attempt to find a model that fits the data well in a way that gives a clear understanding to the way those variables relate to each other and the response variable.

Exploratory Data Analysis:

The figures below are graph matrixes showing untransformed and natural logarithm (log) transformed data for the four continuous variables plotted versus each other as well as a histogram for each. After the log transformation, the histograms appear more normal and the data looks to be more linearly correlated with each other. You can see that *accelerate* seems to have a positive linear relationship with *msrp* while mpg and *mpgmpge* seem to have a negative linear relationship. This would make basic sense because in general the more power a car has the more expensive it is to construct and cars that have better gas mileage are generally smaller and therefore cost less to make.





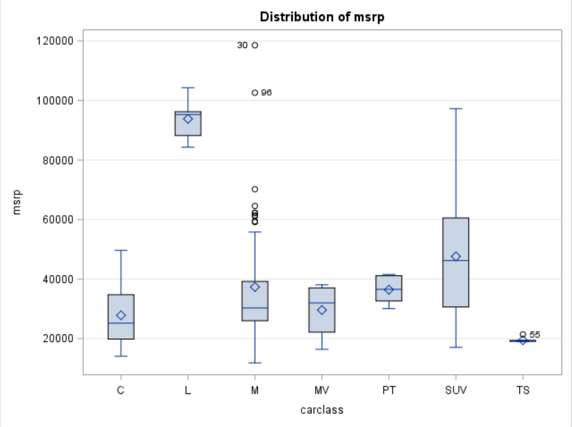
You can also see that *mpg* and *mpgmpge* are nearly identical with only 7 points that are not the same. This is because *mpgmpe* is simply the maximum value of mpg or mpge and there are few cars that can be run as fully electric.

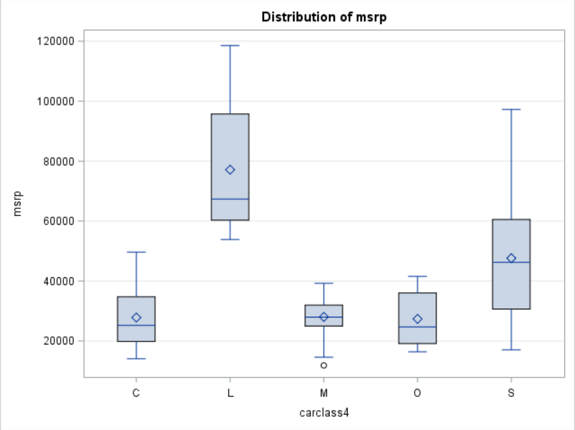
The Model:

It would be nice if the answers to the two questions above came from the same model. However, since I am unsure how likely this is I will try to answer question *two* first, and then see if adding additional terms increases the fit in a useful way.

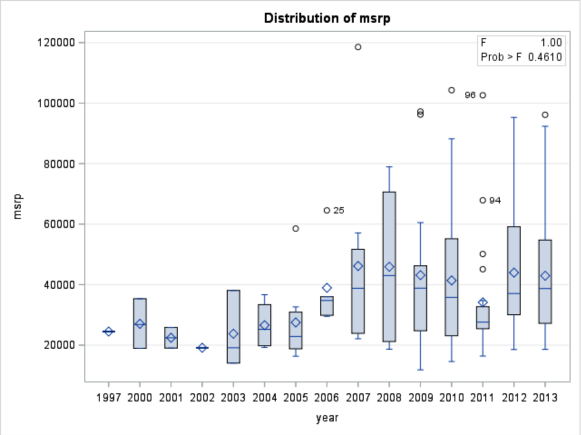
First off after experimenting with proc glm and proc glmselect I noticed several things about the data. There are 7 levels to carclass and 14 for year. After trying several models with different interaction terms the R^2 values were very high (in the 0.8 range), but very few of the interaction terms were significant. Those complicated models did not tell me much about the data or about the relationship between carclass, price, and mpg, so I felt that all the levels were simply increasing the parameters and giving better fits but not giving useful fits. Also, some of the levels have very few data points (1-8) and it seems like that is very few points to be making conclusions. After comparing the means of the response variable *msrp* grouped by carclass, I decided to group the categorical variables into fewer levels of 30-40 observations each. There were some natural groupings and some levels stayed the same. Originally I made carclass 4 levels keeping C, M, and SUV and grouping the other levels into Other. This made sense because the other 4 levels had only just a few observations each but together became comparable with C, M, and SUV. After looking at that I realized that there were several values in the midsize level that were much different than the rest of the midsize. Several of the data points have very high *msrp* and these turned out to be luxury vehicles which had similar *msrp* to all the vehicles in the large level, which are also luxury vehicles, so I grouped those together. I settled on groupings of C, M, L (for large and luxury), SUV, and Other (TS, MV, PT). Since I am trying to model price it makes sense that these factors would affect price. Grouping the carclass this way increased the adjusted R^2 by about 0.1 and reduced the number of parameters!

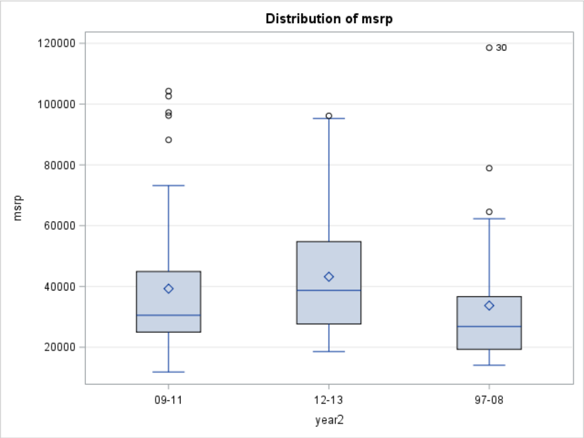
You can see in the figures below the original 7 groupings for carclass comparing the distributions of the *msrp* for each grouping compared to the new 5 groupings taking into account luxury vehicles and combining the groupings in Other with very few data points.





I also grouped year into three ranges, 97-08, 09-11, and 12-13. This gave them all roughly a third of the total observations (40-50), and seemed like natural groupings.





Looking more closely at a model, since the log data looks to fit the assumptions better, most especially the residual plots below, I will apply different models that include the categorical groups above to that log data set. Trying some different combinations of variables with categorical interaction terms using proc glm and proc glmselect the *year* category does not appear to have a significance in modeling msrp. A good model was found:

$$\log(\text{msrp}) = A + B \cdot \log(\text{accelerate}) + C \cdot \log(\text{mpg}) \cdot (\text{carclass})$$

or another way to write it using SUV as the reference:

$$\log(\text{msrp}) = \beta_0 + \beta_1 \log(\text{accelrate}) + \beta_2 \log(\text{mpg}) + \beta_3(\text{Compact}) + \beta_4(\text{Large/Luxury}) + \beta_5(\text{Midsize}) + \beta_6(\text{Other}) + \beta_7 \log(\text{mpg}) \cdot (\text{Compact}) + \beta_8 \log(\text{mpg}) \cdot (\text{ Large/Luxury}) + \beta_9 \log(\text{mpg}) \cdot (\text{Midsize}) + \beta_{10} \log(\text{mpg}) \cdot (\text{Other})$$

In a sense the model is fitting 5 linear simultaneous equations to the data. One for each carclass level.

Fitting the Model:

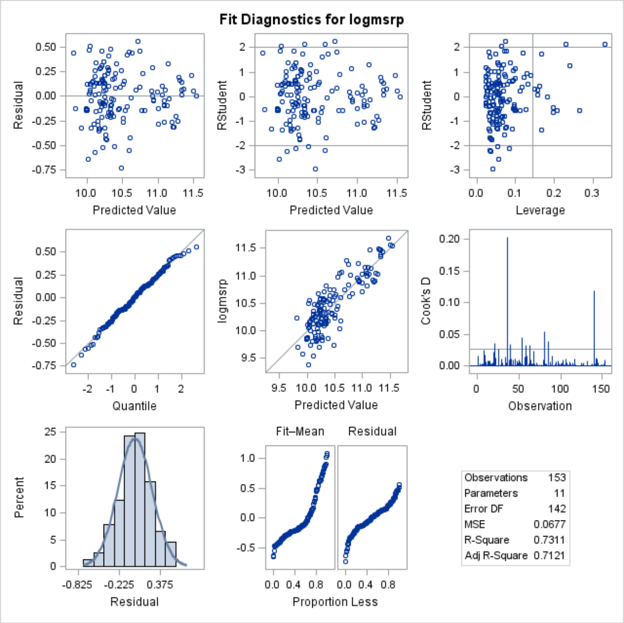
This model gives an R^2 of 0.7311 and an adjusted R^2 of 0.7121. You can see that this model with SUV as the reference calculates most of the parameters p-values to be less than 0.05. The one exception is the Large/Luxury carclass, which have similar prices to high end SUVs, the reference, so it is not too surprising that they may share a similar role in the model.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	26.14768317	2.61476832	38.61	<.0001
Error	142	9.61782293	0.06773115		
Corrected Total	152	35.7655061			

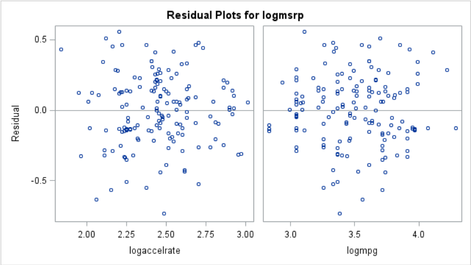
Parameter	Estimate		Standard	t Value	Pr > |t|	95% Confidence Limits	
			Error				
Intercept	13.95995327	B	0.88111876	15.84	<.0001	12.21814801	15.70175853
logaccelrate	0.48858312		0.12851736	3.8	0.0002	0.23452858	0.74263765
logmpg	-1.39025334	B	0.23885695	-5.82	<.0001	-1.86242838	-0.9180783
carclass4 C	-5.38982344	B	1.23122756	-4.38	<.0001	-7.82372762	-2.95591926
carclass4 L	-0.9875393	B	1.36161887	-0.73	0.4695	-3.67920246	1.70412386
carclass4 M	-6.53767538	B	1.03538547	-6.31	<.0001	-8.58443677	-4.490914
carclass4 O	-3.40872282	B	0.90705134	-3.76	0.0002	-5.20179187	-1.61565377
carclass4 S	0	B	.	.	.	.	.
logmpg*carclass4 C	1.5234784	B	0.34836912	4.37	<.0001	0.8348185	2.2121383
logmpg*carclass4 L	0.43803643	B	0.41681834	1.05	0.2951	-0.38593464	1.2620075
logmpg*carclass4 M	1.83470847	B	0.30436942	6.03	<.0001	1.23302765	2.43638929
logmpg*carclass4 O	0.97155471	B	0.27164052	3.58	0.0005	0.43457275	1.50853667
logmpg*carclass4 S	0	B	.	.	.	.	.

Checking the fit:

You see that R^2 is 0.7311 and adjusted R^2 is 0.7121. The residual plots look fairly random and evenly scattered. The RStudent and Cook's D plots do suggest a few high influence and high leverage points. The qq-plot and histogram suggest a fairly normal distribution of the residuals. The residual plots for the untransformed data don't meet these basic assumptions as well, which is why the log-log transformation was used for this model. However, the same model applied to the untransformed data gives roughly the same p-values for the parameters and a slightly higher adj. R^2.



Observations	153
Parameters	11
Error DF	142
MSE	0.0677
R-Square	0.7311
Adj R-Square	0.7121



Interpreting the Coefficients:

Looking closer at the parameter estimates using this model:

$log(msrp) = \beta_0 + \beta_1 log(accelrate) + \beta_2 log(mpg) + \beta_3(Compact) + \beta_4(Large/Luxury) + \beta_5(Midsize) + \beta_6(Other) + \beta_7 log(mpg)\cdot(Compact) + \beta_8 log(mpg)\cdot(Large/Luxury) + \beta_9 log(mpg)\cdot(Midsize) + \beta_{10} log(mpg)\cdot(Other)$

The model gives us these five equations; one for each carclass level:

SUV: $log(msrp) = 13.96 + 0.489\cdot log(accel) - 1.39 log(mpg)$

C: $log(msrp) = 8.57 + 0.489\cdot log(accel) + 0.133\cdot log(mpg)$

L: $log(msrp) = 12.97 + 0.489\cdot log(accel) - 0.952\cdot log(mpg)$

M: $log(msrp) = 7.42 + 0.489\cdot log(accel) + 0.444 \cdot log(mpg)$

O: $log(msrp) = 10.55 + 0.489\cdot log(accel) - 0.419\cdot log(mpg)$

	Parameter	Estimate		Standard	t Value	Pr > |t|	95% Confidence Limits	
				Error				
SUV	Intercept	13.96	B	0.88112	15.84	<.0001	12.21815	15.70176
	logaccelrate	0.48858		0.12852	3.8	0.0002	0.234529	0.742638
	logmpg	-1.3903	B	0.23886	-5.82	<.0001	-1.86243	-0.91808
C	Intercept	8.57013	B	1.04391	8.21	<.0001	6.506516	10.63374
	logaccelrate	0.48858		0.12852	3.8	0.0002	0.234529	0.742638
	logmpg	0.13323	B	0.25738	0.52	0.6055	-0.37556	0.642007
L	Intercept	12.9724	B	1.20891	10.73	<.0001	10.58263	15.3622
	logaccelrate	0.48858		0.12852	3.8	0.0002	0.234529	0.742638
	logmpg	-0.9522	B	0.34407	-2.77	0.0064	-1.63238	-0.27205
M	Intercept	7.42228	B	0.76149	9.75	<.0001	5.916962	8.927594
	logaccelrate	0.48858		0.12852	3.8	0.0002	0.234529	0.742638
	logmpg	0.44446	B	0.18952	2.35	0.0204	0.06981	0.819101
O	Intercept	10.5512	B	0.61271	17.22	<.0001	9.340021	11.76244
	logaccelrate	0.48858		0.12852	3.8	0.0002	0.234529	0.742638
	logmpg	-0.4187	B	0.1355	-3.09	0.0024	-0.68657	-0.15083

Above I have constructed a table with all the parameter estimates as well as their corresponding Confidence intervals.

Interpreting the parameter estimates:

Starting with the reference SUV,

Intercepts:

The intercept of a log-log equation is equal to the value of the response when the explanatory variables are equal to one. Therefore, log(*msrp*) = intercept when *accelrate* = *mpg* = 1. This means *msrp* = $e^{(\text{intercept})}$. These values are:

SUV: $e^{13.96}$ = 1155396, p-value < 0.0001, CI: (202430, 6594579)
C: $e^{8.57}$ = 5272, p-value < 0.0001, CI: (669, 41512)
L: $e^{12.97}$ = 430376, p-value < 0.0001, CI: (39444, 4695886)
M: $e^{7.42}$ = 1673, p-value < 0.0001, CI: (371, 7537)
O: $e^{10.55}$ = 38224, p-value < 0.0001, CI: (11385, 128340)

These values are largely meaningless in this analysis since we do not except the linear relationship to have relevance at small values of the explanatory variables. They are therefore only useful for predicting values within the range of the groupings.

Slopes:

First let's look at the parameter estimates for the *accelrate* term:

The value of 0.4886 is the same for each grouping. For a log-log model this means that for vehicles which get the same miles per gallon, it is estimated that a doubling of *accelrate* is associated with a $2^{0.4886}$ = 1.4 times increase in the *msrp*. Cleary if you want a powerful car you have to pay more to get it. If you want your hybrid to have twice the acceleration of your neighbor's hybrid then you have to pay 1.4 times or 140% as much as they did! The p-value is 0.0002 with a 95% confidence interval of ($2^{0.235}$, $2^{0.743}$) or (118%, 167%).

Next let's look at the parameter estimates for the *mpg* terms:

The value of the slope for the *mpg* terms are all different, and in fact have different signs in some cases. This is very interesting and gets to the heart of one of the major questions. Let's first interpret and look at each slope estimate, p-value, and confidence interval and then we will interpret the values as a whole and how all the estimates paint a picture to help understand the data.

SUV: The value of the slope is -1.39, with a p-value < 0.0001 and CI: (-1.86, -0.92). For a log-log model this means that for vehicles with the same *accelrate*, it is estimated that a doubling of *mpg* is associated with a $2^{-1.39}$= 0.38 times change in the *msrp*. This represents a 62% decrease in *msrp* for a doubling of *mpg*, with CI: (72%, 47%).

C: The value of the slope is 0.133, with a p-value =0.6055 and CI: (-0.376, 0.642). For a log-log model this means that for vehicles with the same *accelrate*, it is estimated that a doubling of *mpg* is associated with a $2^{0.133}$ = 1.1 times change in the *msrp*. This represents a 10% increase in *msrp* for a doubling of *mpg*, with CI: (-23%, 56%).

L: The value of the slope is -0.952, with a p-value < 0.0001 and CI: (-1.632, -0.272). For a log-log model this means that for vehicles with the same *accelrate*, it is estimated that a doubling of *mpg* is associated with a $2^{-0.952}$= 0.52 times change in the *msrp*. This represents a 48% decrease in *msrp* for a doubling of *mpg*, with CI: (68%, 17%).

M: The value of the slope is 0.444, with a p-value < 0.0204 and CI: (0.07, 0819). For a log-log model this means that for vehicles with the same *accelrate*, it is estimated that a doubling of *mpg* is associated with a $2^{0.444}$= 1.36 times change in the *msrp*. This represents a 36% increase in *msrp* for a doubling of *mpg*, with CI: (5%, 76%).

O: The value of the slope is -0.419, with a p-value < 0.0001 and CI: (-1.86, -0.92). For a log-log model this means that for vehicles with the same *accelrate*, it is estimated that a doubling of *mpg* is associated with a $2^{-0.419}$= 0.748 times change in the *msrp*. This represents a 25% decrease in *msrp* for a doubling of *mpg*, with CI: (38%, 10%).

Looking broadly at the mpg slope estimates we see that SUV, L, and O levels have negative slopes while C and M levels have positive slopes. It should be pointed out that the evidence points to all but the C slopes as being significantly different than zero. This is a very interesting result. It means that for some vehicle types you pay less for better mpg. This seems strange at first until you look at the particular levels of SUV and Large/Luxury which don't usually get good gas mileage. In fact these vehicles probably have the negative slope because when you pay more for these vehicles you are generally paying for more options, like heated seats, sound proofing, and other options that increase the weight of the vehicle and so decrease the mpg while heavily increasing the price.

On the other hand, most people buy hybrid vehicles because they want a green vehicle that gets good gas mileage. This is seen for compact and midsized vehicles! The more expensive compact and midsize vehicles generally have slightly better gas mileage. It should be noted again that the p-value for the compact vehicles is above 0.05 and so may not be much different than zero. In this case you might attribute the increase in price probably also equates with an increase to extra options and quality while the mpg also increase. However because of the potential confounder of extra options and quality, which data is not included in this data set and so can only be speculated on, it is hard to know but can perhaps be assumed that the slope would be more significant if weight was taken into account.

Variable selection:

Variable selection was done by hand comparing different models with varying adjusted R2 values as well as using the automatic variable selection procedure glmselect in SAS. Forward, backward, and stepwise with select = CV was used as well as 5 fold internal cross validation. After the intensive and clever =) changes to the carclass and year levels all the proc glmselect procedures agreed with the model that was so painstakingly constructed! This probably took me 20 hours! It was almost too much fun…

```
proc glmselect data = hybrid8;
class carclass4 year2;
model logmsrp = logaccelrate year2 logmpg
carclass4/ selection = Stepwise(select=CV)
cvmethod=random(5) stats =adjrsq;
run;
quit;
```

Parameter Estimates				
Parameter	DF	Estimate	Standard	t Value
			Error	
Intercept	1	10.324371	0.556123	18.56
logaccelrate	1	0.556477	0.144038	3.86
logmpg	1	-0.322445	0.103562	-3.11
carclass4 C	1	-0.196905	0.090043	-2.19
carclass4 L	1	0.398304	0.088559	4.5
carclass4 M	1	-0.293365	0.075338	-3.89
carclass4 O	1	-0.270569	0.095175	-2.84
carclass4 S	0	0	.	.

```
proc glmselect data = hybrid8;
class carclass4 year2;
model logmsrp = logaccelrate
logmpg|carclass4/ selection =
Stepwise(select=CV) cvmethod=random(5)
stats =adjrsq;
run;
quit;
```

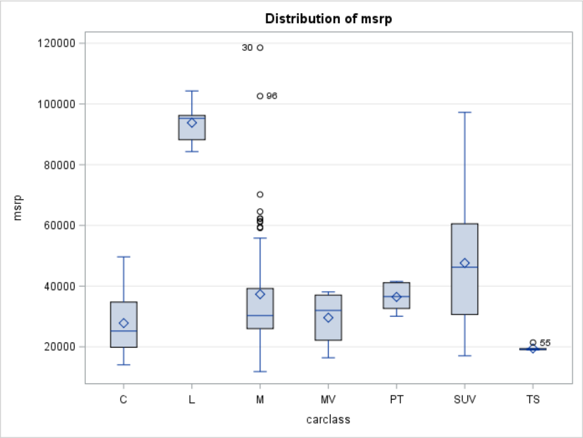
Parameter Estimates				
Parameter	DF	Estimate	Standard	t Value
			Error	
Intercept	1	13.959953	0.881119	15.84
logaccelrate	1	0.488583	0.128517	3.8
carclass4 C	1	-5.389823	1.231228	-4.38
carclass4 L	1	-0.987539	1.361619	-0.73
carclass4 M	1	-6.537675	1.035385	-6.31
carclass4 O	1	-3.408723	0.907051	-3.76
carclass4 S	0	0	.	.
logmpg*carclass4 C	1	0.133225	0.257375	0.52
logmpg*carclass4 L	1	-0.952217	0.344071	-2.77
logmpg*carclass4 M	1	0.444455	0.18952	2.35
logmpg*carclass4 O	1	-0.418699	0.135505	-3.09
logmpg*carclass4 S	1	-1.390253	0.238857	-5.82

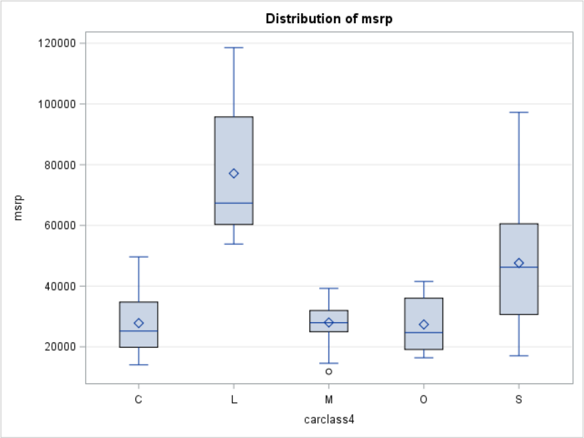
It should be noted that when using other interaction terms proc glmselect can choose a slight different model. However, these models generally have smaller adj. R^2 values and have many more interaction terms with p-values greater than 0.05. It should also be noted interacting both categorical variables with both continuous variables does increase adj. R^2 but it is uncertain if it is in useful way.

In order to find the best way to understand the data and group the carclass levels a hypothesis test was done comparing the means and medians of the response variable *msrp* and log(*msrp*) per *carclass* level. A Tukey-Kramer multiple-comparison procedure option for proc glm was used to make this comparison. Tukey was chosen because of the difference in the number of observations for each level.

Comparisons significant at the 0.05 level are				
indicated by ***.				
carclass	Difference	Simultaneous 95% Conf		
Comparison	Between	Limits		
	Means			
L - SUV	46205	26380	66031	***
L - M	56494	37132	75857	***
L - PT	57407	30537	84277	***
L - MV	64213	33941	94486	***
L - C	66004	45906	86101	***
L - TS	74366	49369	99362	***
SUV - L	-46205	-66031	-26380	***
SUV - M	10289	216	20362	***
SUV - PT	11202	-9978	32382	
SUV - MV	18008	-7349	43365	
SUV - C	19798	8375	31222	***
SUV - TS	28160	9415	46906	***
M - L	-56494	-75857	-37132	***
M - SUV	-10289	-20362	-216	***
M - PT	913	-19834	21659	
M - MV	7719	-17277	32715	
M - C	9509	-1090	20108	
M - TS	17871	-384	36126	
PT - L	-57407	-84277	-30537	***
PT - SUV	-11202	-32382	9978	
PT - M	-913	-21659	19834	
PT - MV	6806	-24369	37982	
PT - C	8597	-12838	30032	
PT - TS	16959	-9125	43042	
MV - L	-64213	-94486	-33941	***
MV - SUV	-18008	-43365	7349	
MV - M	-7719	-32715	17277	
MV - PT	-6806	-37982	24369	
MV - C	1790	-23780	27361	
MV - TS	10152	-19424	39728	
C - L	-66004	-86101	-45906	***
C - SUV	-19798	-31222	-8375	***
C - M	-9509	-20108	1090	
C - PT	-8597	-30032	12838	
C - MV	-1790	-27361	23780	
C - TS	8362	-10671	27395	
TS - L	-74366	-99362	-49369	***
TS - SUV	-28160	-46906	-9415	***
TS - M	-17871	-36126	384	
TS - PT	-16959	-43042	9125	
TS - MV	-10152	-39728	19424	
TS - C	-8362	-27395	10671	

Comparisons significant at the 0.05 level				
are indicated by ***.				
carclass4	Difference	Simultaneous 95% Con		
Comparison	Between	Limits		
	Means			
L - S	29508	19432	39584	***
L - M	49111	39196	59027	***
L - C	49307	38925	59688	***
L - O	49760	37857	61662	***
S - L	-29508	-39584	-19432	***
S - M	19603	11502	27704	***
S - C	19798	11133	28464	***
S - O	20252	9812	30691	***
M - L	-49111	-59027	-39196	***
M - S	-19603	-27704	-11502	***
M - C	195	-8283	8674	
M - O	649	-9636	10933	
C - L	-49307	-59688	-38925	***
C - S	-19798	-28464	-11133	***
C - M	-195	-8674	8283	
C - O	453	-10282	11188	
O - L	-49760	-61662	-37857	***
O - S	-20252	-30691	-9812	***
O - M	-649	-10933	9636	
O - C	-453	-11188	10282	





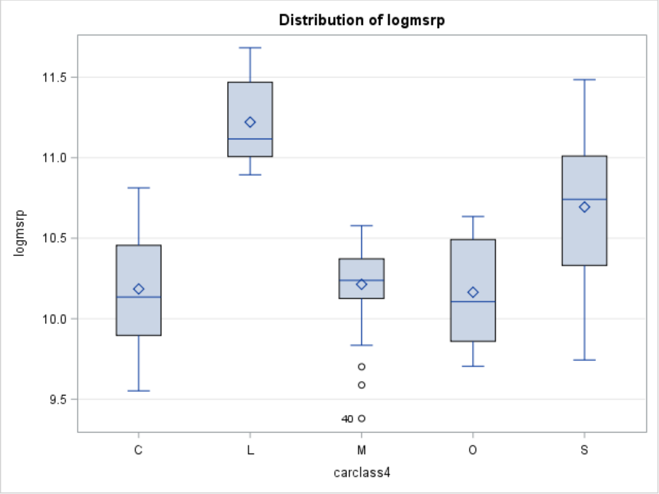
The above graphs and tables represent the tests of different means with the two different groupings. The one on the left represents the original levels and groupings of the carclass variable while the one on the right represents the changes made to these levels. You can see a change from 7 levels to 5 with a merging of L and M. These tests were instrumental providing the information to help group the levels properly to better understand the relationship between vehicle type *accelrate* and *mpg* and how they relate to *msrp*.

Looking specifically at the log(*msrp*) with respect to the adjusted carclass with 5 levels. We will do a six step hypothesis test but first check the assumptions.

Analysis Variable : logmsrp								
carclass4	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
C	33	10.18	0.32	9.55	9.9	10.13	10.46	10.81
L	20	11.22	0.26	10.89	11.01	11.12	11.47	11.68
M	43	10.21	0.25	9.38	10.13	10.24	10.37	10.58
O	18	10.16	0.33	9.7	9.86	10.11	10.49	10.63
S	39	10.69	0.41	9.74	10.33	10.74	11.01	11.48

The variances and number of observations are slightly different so we should be a bit careful. The histograms, boxplots, and qq-plots don’t give evidence against normality however it cannot be assured with the sparsity of observations for levels L and O.

1) H_0: all the group medians are equal
   H_a: one of the group medians is different
2) Critical Value: F_{crit} = 2.873 (alpha = 0.05, dfn = 4, dfd = 148)
3) F_{stat} = 49.18 (from table below)
4) P-value < 0.0001 (from table below)
5) REJECT H_0
6) There is sufficient evidence to suggest that there is a difference in the median of the distributions of the 5 groupings of the log(*msrp*). In fact, using the Tukey comparison option we have a table showing the levels which there is sufficient evidence for them to be significantly different from each other. Individual level comparisons are shown in the table with CI’s.



Comparisons significant at the 0.05 level are				
indicated by ***.				
carclass4	Difference	Simultaneous 95% Confidence		
Comparison	Between	Limits		
	Means			
L - S	0.52701	0.28236	0.77166	***
L - M	1.00779	0.76703	1.24855	***
L - C	1.0365	0.78442	1.28857	***
L - O	1.05659	0.76759	1.34559	***
S - L	-0.52701	-0.77166	-0.28236	***
S - M	0.48078	0.28408	0.67748	***
S - C	0.50949	0.29909	0.71989	***
S - O	0.52958	0.27611	0.78305	***
M - L	-1.00779	-1.24855	-0.76703	***
M - S	-0.48078	-0.67748	-0.28408	***
M - C	0.02871	-0.17716	0.23457	
M - O	0.0488	-0.20092	0.29852	
C - L	-1.0365	-1.28857	-0.78442	***
C - S	-0.50949	-0.71989	-0.29909	***
C - M	-0.02871	-0.23457	0.17716	
C - O	0.02009	-0.24056	0.28074	
O - L	-1.05659	-1.34559	-0.76759	***
O - S	-0.52958	-0.78305	-0.27611	***
O - M	-0.0488	-0.29852	0.20092	
O - C	-0.02009	-0.28074	0.24056	

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	20.41103976	5.10275994	49.18	<.0001
Error	148	15.35446634	0.10374639		
Corrected Total	152	35.7655061			

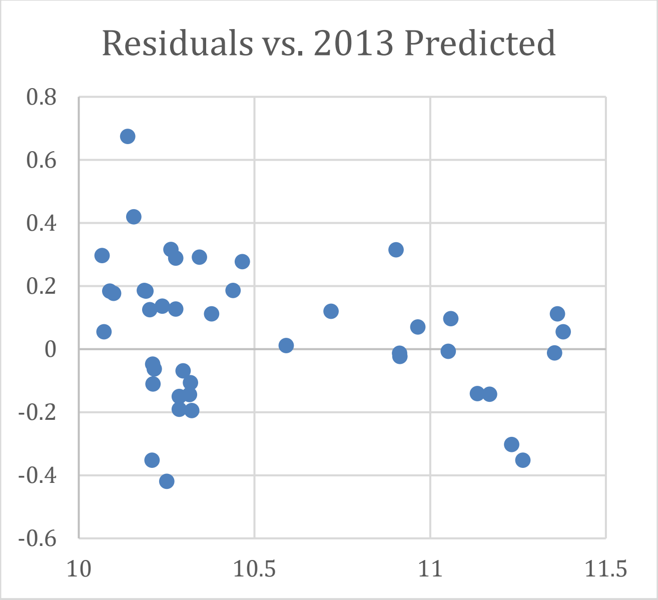
This is not an experimental study so no strong conclusions can be drawn. However, we can make inferences about this data, which does encompass nearly all the hybrid vehicles before 2014, and indeed can infer that there is a significant difference between L and all levels, SUV and all levels, but the M, C, and O levels do not have significantly different medians.

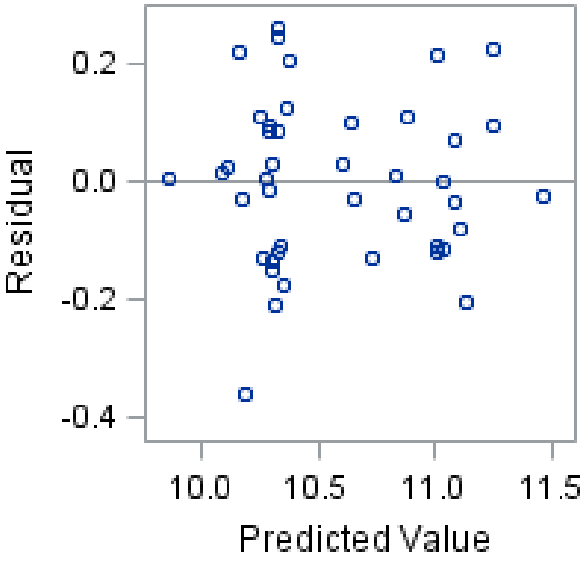
In the regression model the p-values and CI's are a bit different. The L and SUV levels are not significantly different, while the M and C levels are also not significantly different. The M and O are significantly different per the regression. So there are many similarities between the two tests, but since the regression is more powerful in this case, we factor in other variables with the regression, we gain more understanding from the data using the multiple regression!

Prediction:

So the last thing I want to look at is seeing how good this model is at predicting the future. I took all the data before 2012 and run the model just over that data and compared it to the model over the original data. What I found was the adj. R^2 was a bit lower 0.6859 from 0.7121, and the parameter estimates were all a bit different but very similar. Next, using the 1997-2012 data I had SAS calculate the predicted 2013 response *logmsrp* values given the explanatory variables being the same.

Using proc glm and the cli option I was able to have SAS calculate the predicted 2013 logmsrp values as well as the 95% Confidence Limits for Individual Predicted Value. Taking these values for the 43 observations that make up the 2013 data I was able to make a residual plot, (actual – predicted) vs. predicted. The figure below and on the left shows this residuals plots. All the predicted values fall within their respected prediction intervals.





Even though there are not as many points evenly distributed along the horizontal axis this residuals graph for the 2013 data looks fairly random distributed. Running the model again on just the 2013 data the adj. R^2 is 0.8664 with very good looking residual plots. Comparing the predicted residual (above left) to the actual residual plot of the model on the 2013 data (above right) we can see that the predictive model does a good job, but the outliers in the 1997-2012 set probably are not the best training set for the model. However, the goodness of fit of the model for just the 2013 data suggests that the model was very sensibly chosen. Now that hybrid technology is mainstream and better understood, it will probably be much more easy to make models that predict future prices.

Conclusion:

Analysis of hybrid vehicle data was performed using multiple regression. It was found that vehicles of different types model the price of hybrid vehicles differently. In fact, it is seen that compact and midsized vehicles are valued for their fuel efficiency while SUV and Large Luxury hybrid vehicles, while also valued for their fuel efficiency, are more expensive because of other vehicle options. Additional study of potential confounding variables such as weight would be needed to further illuminate these conclusions. Prediction is very possible with this model and future models should probably stick with data from the last few years to predict future year prices.

SAS code below:

```
data hybrid;
infile 'C:\Users\hp\Desktop\SMU\Exp Stats 1\Live\Unit
13\Project\hybrid_reg.csv' dlm = ',' firstobs = 2;
input carid $ vehicle $ year $ msrp accelrate mpg mpgmpge carclass $
carclass_id $;
if vehicle = "C-Max FW" then carclass = "C";
if vehicle = "C-Max FW" then carclass_id = "1";
run;

proc print data = hybrid; run;

data hybrid2;
set hybrid;
logmsrp=log(msrp);
logaccelrate=log(accelrate);
logmpg=log(mpg);
logmpgmpge=log(mpgmpge);
run;

proc print data = hybrid2; run;

proc sgscatter data = hybrid2;
matrix msrp accelrate mpg mpgmpge/diagonal=(histogram);
Title Hybrid Cars;
run;

proc sgscatter data = hybrid2;
matrix logmsrp logaccelrate logmpg logmpgmpge/diagonal=(histogram);
Title Hybrid Cars;
run;

proc univariate data=hybrid2;
   var logmsrp;
   histogram logmsrp;
run;

proc glm data = hybrid2 plots = all;
class year carclass (ref="C");
model msrp = accelrate mpg mpgmpge carclass year/ solution;
run;
quit;

proc glm data = hybrid2 plots = all;
class year (ref="1997") carclass (ref="SUV");
model msrp = accelrate|year mpg|carclass/ solution;
run;
quit;

*remove data point 38 because cooks d was very high for this
observation;
data hybrid3;
set hybrid2;
if _n_ = 38 then delete;
run;

data hybrid4;
      set hybrid2;
       if year eq "1997" then year2 = "97-08";
       if year eq "2000" then year2 = "97-08";
       if year eq "2001" then year2 = "97-08";
       if year eq "2002" then year2 = "97-08";
       if year eq "2003" then year2 = "97-08";
       if year eq "2004" then year2 = "97-08";
       if year eq "2005" then year2 = "97-08";
       if year eq "2006" then year2 = "97-08";
       if year eq "2007" then year2 = "97-08";
       if year eq "2008" then year2 = "97-08";
       if year eq "2009" then year2 = "09-11";
       if year eq "2010" then year2 = "09-11";
       if year eq "2011" then year2 = "09-11";
       if year eq "2012" then year2 = "12-13";
       if year eq "2013" then year2 = "12-13";
;

data hybrid5;
      set hybrid4;
       if carclass eq "C" then carclass2 = "C";
       if carclass eq "M" then carclass2 = "M";
       if carclass eq "SUV" then carclass2 = "SUV";
       if carclass eq "TS" then carclass2 = "O";
       if carclass eq "PT" then carclass2 = "O";
       if carclass eq "MV" then carclass2 = "O";
       if carclass eq "L" then carclass2 = "L";
;

data hybrid7;
      set hybrid5;
       if carclass eq "C" then carclass3 = "C";
       if carclass eq "M" then carclass3 = "M";
       if carclass eq "SUV" then carclass3 = "SUV";
       if carclass eq "TS" then carclass3 = "TS";
       if carclass eq "PT" then carclass3 = "O";
       if carclass eq "MV" then carclass3 = "O";
       if carclass eq "L" then carclass3 = "L";
;


proc print data = hybrid7; run;

data hybrid9;
set hybrid8;
if _n_ = 40 then delete;
run;

proc print data = hybrid9; run;

*Putting Luxury vehicles in with Large vehicles!;
data hybrid8;
      set hybrid7;
      carclass4 = carclass2;
       if msrp > 50000 and carclass = "M" then carclass4 = "L";
;

proc print data = hybrid8; run;

proc glm data = hybrid8 plots = all;
class year2 carclass4 (ref="M");
model logmsrp = logaccelrate logmpg|carclass4/ solution clparm;
run;
quit;
```

```
proc glm data = hybrid8 plots = all;
class carclass4 (ref="S");
model msrp = accelrate mpg|carclass4/ solution clparm;
run;
quit;

proc glm data = hybrid5 plots = all;
class carclass2 (ref="C");
model msrp = accelrate/ solution;
run;
quit;

proc glmselect data = hybrid5 plots = all;
class carclass2;
model logmsrp = logaccelrate logmpgmpge|carclass;
run;
quit;

proc reg data = hybrid2;
model msrp = accelrate mpg mpgmpge/ stb clb VIF scorr1 scorr2;
run;
quit;

proc reg data = hybrid2;
model logmsrp = logaccelrate logmpg/ stb clb VIF scorr1 scorr2;
run;
quit;

proc corr data = hybrid2;
var accelrate mpg mpgmpge;
run;
quit;

proc glmselect data = hybrid3;
class carclass year;
model logmsrp = logaccelrate|year logmpg|carclass/ selection =
Forward(stop=CV) cvmethod=random(5) stats =adjrsq;
run;
quit;

proc glmselect data = hybrid8;
class carclass4 year2;
model logmsrp = logaccelrate year2 logmpg carclass4/ selection =
Backward(stop=CV) cvmethod=random(5) stats =adjrsq;
run;
quit;

proc glmselect data = hybrid8;
class carclass4 year2;
model logmsrp = logaccelrate year2 logmpg carclass4/ selection =
Stepwise(stop=CV) cvmethod=random(5) stats =adjrsq;
run;
quit;


proc glmselect data = hybrid8;
class carclass4 year2;
model logmsrp = logaccelrate logmpg|carclass4/ selection =
Stepwise(select=CV) cvmethod=random(5) stats =adjrsq;
run;
quit;
```

```
proc glm data = hybrid5 plots = all;
class carclass2;
model logmsrp = carclass2;
run;
quit;

proc glm data = hybrid7 plots = all;
class carclass3;
model logmsrp = carclass3;
run;
quit;


proc glm data=hybrid8 plots = all;
class carclass4;
model logmsrp=carclass4;
means carclass4/ hovtest = bf tukey cldiff;
run;


proc glm data=hybrid8;
class year2;
model msrp=year2;
means year2/ hovtest = bf tukey cldiff;
run;


data fcritval;
F=finv(.975,4,148)
;

proc print data=fcritval;
run;

proc means data=hybrid8 maxdec=2 mean std min q1 median q3 max;
class carclass4;
var logmsrp;
run;

proc univariate data =hybrid8;
class carclass4;
var logmsrp;
histogram  / normal(mu=est sigma=est color=yellow w=2.5);
qqplot;
run;

data hybrid10;
set hybrid8;
if year = "2013" then delete;
run;

proc print data = hybrid10;
run;

proc glm data = hybrid10 plots = all;
class year2 carclass4 (ref="S");
model logmsrp = logaccelrate logmpg|carclass4/ solution clparm;
Title Hybrid Cars "1997-2012";
run;
quit;
```

```
data hybrid11;
set hybrid8;
if year ne "2013" then delete;
run;

proc print data = hybrid11;
run;

proc glm data = hybrid11 plots = all;
class year2 carclass4 (ref="S");
model logmsrp = logaccelrate logmpg|carclass4/ solution clparm;
Title Hybrid Cars "2013";
run;
quit;

data hybrid12;
set hybrid8;
if year = "2013" then msrp = ".";
if year = "2013" then logmsrp = ".";
run;

proc print data = hybrid12;
run;

proc glm data = hybrid12 plots = all;
class year2 carclass4 (ref="S");
model logmsrp = logaccelrate logmpg|carclass4/ solution clparm;
Title Hybrid Cars "1997-2012 predicted 2013";
run;
quit;

proc glm data = hybrid12 plots = all;
class year2 carclass4 (ref="S");
model logmsrp = logaccelrate logmpg|carclass4/ solution cli;
Title Hybrid Cars "1997-2012 predicted 2013";
run;
quit;
```