MSDS Project 1

**Introduction:**

Imagine your dream car. What is the make and model? What year is it? The biggest question of them all is: how much will it cost you? From this analysis we use multiple linear regression techniques to predict the most accurate manufacturer's suggested retail price (MSRP) of a hybrid vehicle in 2013.

In order to do so, we approach predicting MSRP from a few different perspectives: a minimalistic approach using simple linear regression and simple two-way ANOVA, and another with greater complexity using all variables with forward, backward, and LASSO variable selection. We hope to leave you with a greater sense of what aspects drive MSRP when looking to buy a hybrid vehicle!

**Data Description:**

We leveraged a dataset from the University of Florida statistical department website which attached to this submission as hybrid\_reg.csv. The dataset contains 153 observations and 11 variables. The response variable is the MSRP in US dollars for a variety of hybrid vehicles in 2013. Other variables are vehicle make and model, model year, acceleration rate, fuel economy (mpg) and mpge for full electric vehicles. Details on the variables are highlighted in the table below.

|  |  |
| --- | --- |
| Variable Name | Description/Calculation |
| carid | Vehicle ID |
| year | Model Year |
| vehicle | Vehicle Make and Model |
| msrp | Manufacturer's Suggested Retail Price in 2013 (US Dollars) |
| accelrate | Acceleration Rate in km/hour/second |
| mpg | Fuel Economy in Miles/gallon |
| mpgmpge | Max of mpg and mpge  MPGe = 33.7\*driverange/batterycapacity. |
| carclass | Model classes: C = Compact, M = Midsize, TS = 2 Seater, L = Large, PT = Pickup Truck, MV = Minivan, SUV = Sport Utility Vehicle |
| carclass\_id | Model Class ID |

**Exploratory Analysis:**

To start, we removed the variable *car\_id*, which was a unique identifier for every record in the data. We also removed the variable *carclass\_id* which was a numerical identifier for the different types of model classes (compact, midsize, etc.) The reasoning behind removing the carclass\_id variable is that conveys the exact same information as the *carclass* variable, so due to the nature and explainability of the *carclass* variables, we chose to leverage this field instead.

After removing unnecessary variables from the dataset, we chose to create a scatterplot with a matrixed comparison of the continuous variables in order to assess correlations and relationships between the inputs. (Figure 1) In analyzing the matrix, we found there to be inherent positive correlations with *msrp* and *accelrate*. Another visual observation is an exponential relationship to between *msrp* and *mgp*, along with a very strong linear relationship between *mpg* and *mpgmpge*. Upon further investigation, it was found mpg and mpgmpge shared the exact same information aside from 7 observations. One possible explanation for the near identical fields is that the *mpgmpge* field contains information for fully electric vehicles. In 2013, the amount of car models that operated as fully electric vehicles was a very small percentage compared to the hybrid models. Because only 7 observations out of 150+ are fully electric vehicles, it was a conscious decision to leverage the mpg variable over the mpgmpge variable, and we removed *mpgmpge* from the dataset. Lastly, in order to handle the curvature in the scatterplot we experimented between logging both *msrp* and *mpg*. After analyzing the assumptions needed for linear regression (linearity, constant variance, normality, etc.) it was found that a log to *mpg* resulted in the the most linearly correlated variables. Further model specific analysis related to outlier detection can be found under Objective 1 Analysis. (Figure 2)

Now that we have a better understanding of the continuous variables, it is time to explore the relationship between our categorical variables and our response, *msrp*. (Figure 3A)An informative box plot illustrated the differences in price between the classes. Car classes L and TS stood out most. Those two car classes are large vehicles (L) and two-seaters (TS). This makes practical sense as large vehicles are more likely to have higher base prices (larger engines, more material, etc.) explaining why *carclass* L was on the higher end of the spectrum. In contrast, two-seater vehicles do not require large engines or an excess of materials, landing them on the lower end of the spectrum. The other categorical variable for use in this dataset is vehicle. Vehicle is a complex categorical as it holds 109 levels.  While the graphic displayed in Figure 4 is hard to digest at first glance, there is useful information that it provides. Hybrid Civic’s and Sonata’s have the largest ranges, and the means and ranges of the levels do not often overlap. This is an indication that it is likely the vehicle make and model are going to be instrumental in accurately predicting MSRP.

**Objective 1:**

Initially we built a slew of intuitive models that would be easier to interpret. We started with all the variables, performed some VIF analysis, and then manually removed and added variables with intuition and interpretation of model results. This was an iterative process. In doing so, we built roughly 10 or so different intuitive models. Documented below you will only see the model that returned the best adjusted R-squared, but the process in arriving at that model is documented in this analysis. After manually creating the intuitive model, we then built models built using feature selection techniques like Forward, Backward, and LASSO. These models were far more complex with upwards of 75+ model variables.

*The Intuitive Model:*

The initial intuitive model we built began with the variables: *year*, *accelrate*, *mpg*, *carclass*. We decided to remove the variable *vehicle* from our initial exploration intuitive exploration because it the volume of levels contained within the factor would make the model extremely complex and difficult to interpret. Sticking with the raw variables mentioned above surfaced an adjusted R-squared of 0.6691. In addition, due to the curvature found in the initial data exploration, we then found that logging the *mpg* variable resulted in an adjusted R-squared of 0.6875. We also attempted some sqrt transformations on the response and mpg variable to assess what difference that would make. It was found that the transformation was indeed helpful in improving the assumptions needed and in the adjusted R-squared but provided the hurdle of interpretation.

We then performanced a lack of fit test to see if interaction terms in the model would be beneficial. We also had a hunch on including the interaction term accelrate and log(mpg), because they looked to be collinear in the initial exploratory scatterplots. From an intuitive sense, it also seemed that controlling for model year, the MSRP of a hybrid car would change based on the carclass. Thus, we wanted to test the interaction year\*carclass. It was found that the interactions accelrate\*log(mpg) and year\*carclass were influential in predicting MSRP. The adjusted R-squared with interaction terms is 0.6869. Upon that finding, we then took another look at the outliers and leverage points with this model. By going through an iterative process, we found the following rows to be outliers and/or have leverage in the model: 30, 37, 36. After removing these data points, we were able to achieve a final adjusted-R squared of 0.751. This means that model is to account for 75% of the variance in the MSRP in. It is that final model below that we will continue to elaborate on below.

msrp~year+accelrate+log(mpg)+accelrate\*log(mpg)+carclass+carclass\*year

*Assumptions:*

There are four main assumptions to check when validating a linear regression: normality, linearity, constant variance (equal spread), and independence.  To address the linearity assumption, we don’t see any curvature in the residuals. (Figure 5, Figure 7) Looking at Figure 2, with the logged mpg variable, all continuous inputs meet the linearity assumption. To address linearity, the qq-plot gives no indication of a normality violation. (Figure 6) For the most part, the residual plot indicates a random cloud around 0. (Figure 5) For the constant variance assumption, looking back at the residual plots, we see there is equal distribution or spread. Thus, meeting the requirement for constant variance. For this analysis, we will assume the data is independent. It is important to note as well, that this is a rather large sample population of more than 150 observations. This allows flexibility in stringent requirements for a multiple linear regression due to the power of the central limit theorem.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient Estimate (US Dollars)** | **P-Value** |
| (Intercept) | 81727.5 | 0.92763 |
| year | -125.7 | 0.77447 |
| accelrate | 23326.6 | 1.73e-06 \*\*\* |
| log(mpg) | 49272.6 | 0.00414 \*\* |
| carclassL | 10284769.6 | 0.02899 \* |
| carclassM | -2638768.0 | 0.07288 |
| carclassMV | 3983858.5 | 0.22557 |
| carclassPT | -1794110.5 | 0.47797 |
| carclassSUV | -760444.7 | 0.62793 |
| carclassTS | 564208.6 | 0.78796 |
| accelrate:log(mpg) | -5855.7 | 4.40e-05 \*\*\* |
| year:carclassL | -5099.9 | 0.02939 \* |
| year:carclassM | 1311.2 | 0.07312 . |
| year:carclassMV | -1982.3 | 0.22607 |
| year:carclassPT | 891.0 | 0.47889 |
| year:carclassSUV | 378.7 | 0.62765 |
| year:carclassTS | -285.5 | 0.78494 |

*Interpretation:*

In order to interpret the outcome of the model, we need to look at the coefficients above and provide an explanation for what they are telling us. Looks like the factor *carclassC* is being used as the reference. Even though we can see that statistically this is not significant, because at least one other level in the *carclass* variable in the model is significant, we need to include it in the overall model. This is also true for the year variable. Because the interaction of *year* with *carclass* is statistically significant, we must include it in the overall model. To best understand the other coefficients, we need to interpret the intercept coefficient first. The intercept is the mean estimate for compact cars, keeping all other variables constant. The classclassL coefficient can be interpreted as the mean additional cost to the mean compact car cost (from the intercept) for an estimate for a large car. You can extrapolate these examples to all *carclass* coefficients. In fact, this type of interpretation can be applied to all other variables that do not have a log transformation. The logged variables however, need to be transformed back from their log in order to be properly interpreted in the context of other variables. Because this is a linear-log model, when interpreting the coefficient of mpg, a one unit increase in the mpg would result in a 49272.6 log(2) = 34152.2 change in the mean of y.

The confidence intervals for the parameter estimates are in the table pictured in Figure 12 of the appendix. The confidence intervals show that *accelrate*, *log(mpg),* *carclassL*, *accelrate\*log(mpg),* and *year\*carclassL* do not contain zero at the 95% level. This indicates that we can reject the null hypothesis and say that these estimates are significantly different from zero. (Figure 12)

Because the interaction term between *year* and *carclassL* is significant, both year and *carclass* by themselves need to stay in the model. The only variable not included in our model that wasn’t deemed redundant was the vehicle variable. Adding the *vehicle* variable to the model increases the adjusted R-squared to 0.9845. Even though this adjusted R-squared is really high, we chose to leave the variable out of the mode because we believe it was overfitting. Vehicle has 109 levels. There are only 153 observation in the data set. Many levels of vehicle only have one data point in them. If *vehicle* was included in the model, the model would overfit to the specific vehicles. Furthermore, the model (including *vehicle*) would be very good at explaining this particular data set but wouldn’t necessarily be good at explaining a different data set with similar information. It is likely that vehicle type could be an important variable in predicting the retail price, but to use it, more data points would need to be added for each of the levels.

*Other Model Building Approaches:*

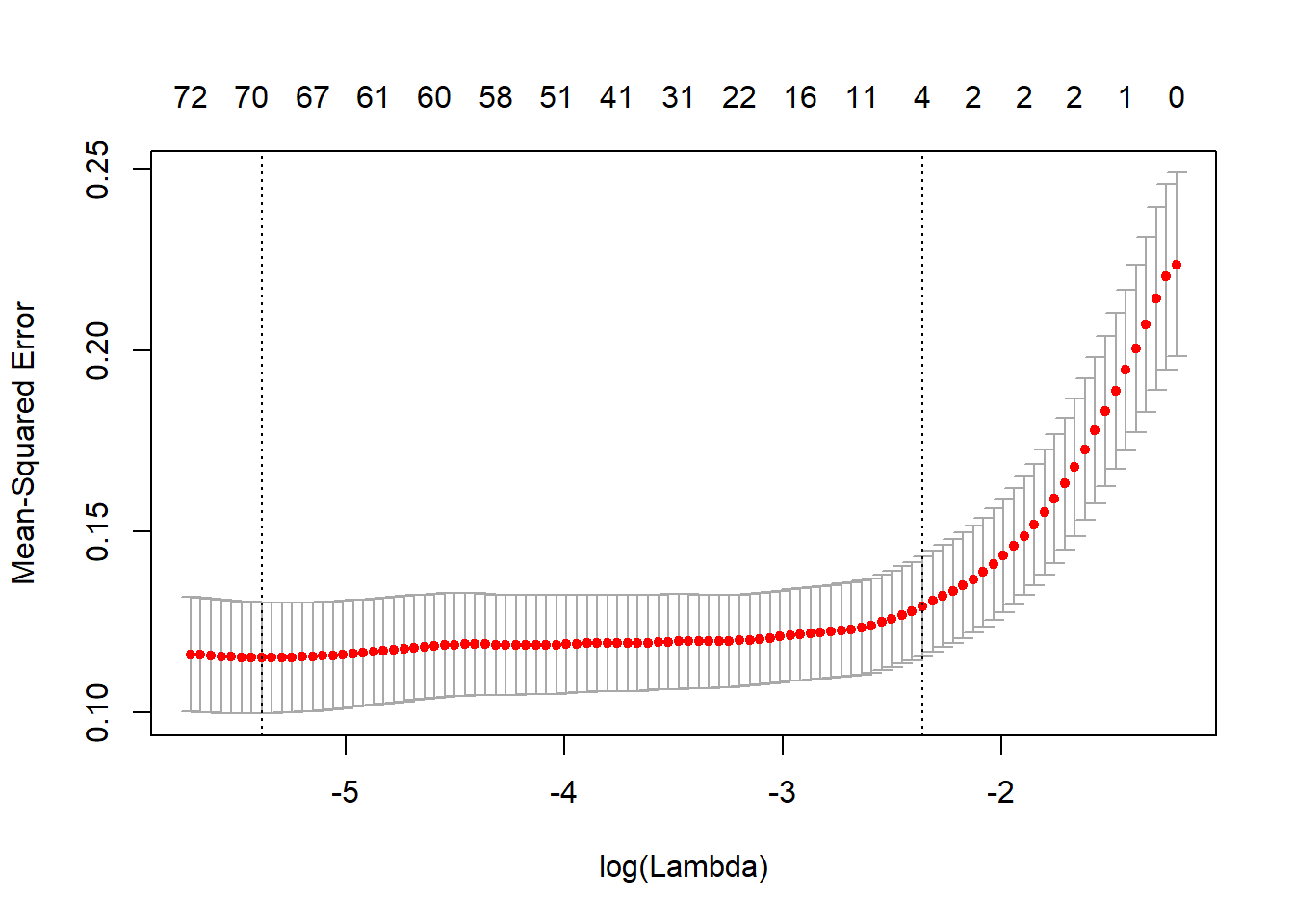
When we initially built our intuitive model, we did not split our dataset into train and test partitions. Because of this, there wasn’t a fair way to compare our intuitive model to the LASSO and forward models that we built. In order to create a fair comparison, we ended up using the same model structure from our intuitive work, but this time instead of feeding it the entire dataset, we performed a 70-30 split on our original dataset to create a train and test set. We then used our train set to fit our intuitive model structure and created a prediction vector from our test set. From there we were able to calculate the MSE of our intuitive model and compare it to those of forward and LASSO.

Unfortunately, what we found was that our MSE for both our initiative and LASSO models were so far from 0 that we were shocked! We thought to ourselves: how could a model that produced an adjusted R-squared of 0.75 lead to errors greater than 100!? (The MSE for the intuitive model was: 2333345037, for LASSO it was: 1.44x10^8) This took us back to the drawing board. Did we violate an assumption? Did we approach the problem incorrectly? How could we be so far off? In doing that analysis, we found that if we logged the response variable *msrp* we got results that were much more reasonable. Looking over the assumptions again with a logged *msrp (Figure 2B, 13-16),* we found there to be even an even stronger linear correlation between the variables with a logged *msrp*. When rerunning our intuitive model with a logged response, what we found again was another shock -- only two variables were statistically significant with a logged response: *accelrate* and the interaction *log(msg)\*accelrate*. Rerunning the analysis with the two statistically significant variables, we found was much less explainability in the adjusted R-squared at 0.57. However, when we used this model in our train and test split, we achieved an MSE of 0.05151259! (Figure 18) Wow, what a difference in the mean squared error. The assumptions of an MLR model (linearity, constant variance, normality, and independence) were met with the new model outlined below:

msrp~log(mpg)+accelrate+accelrate\*log(accelrate)

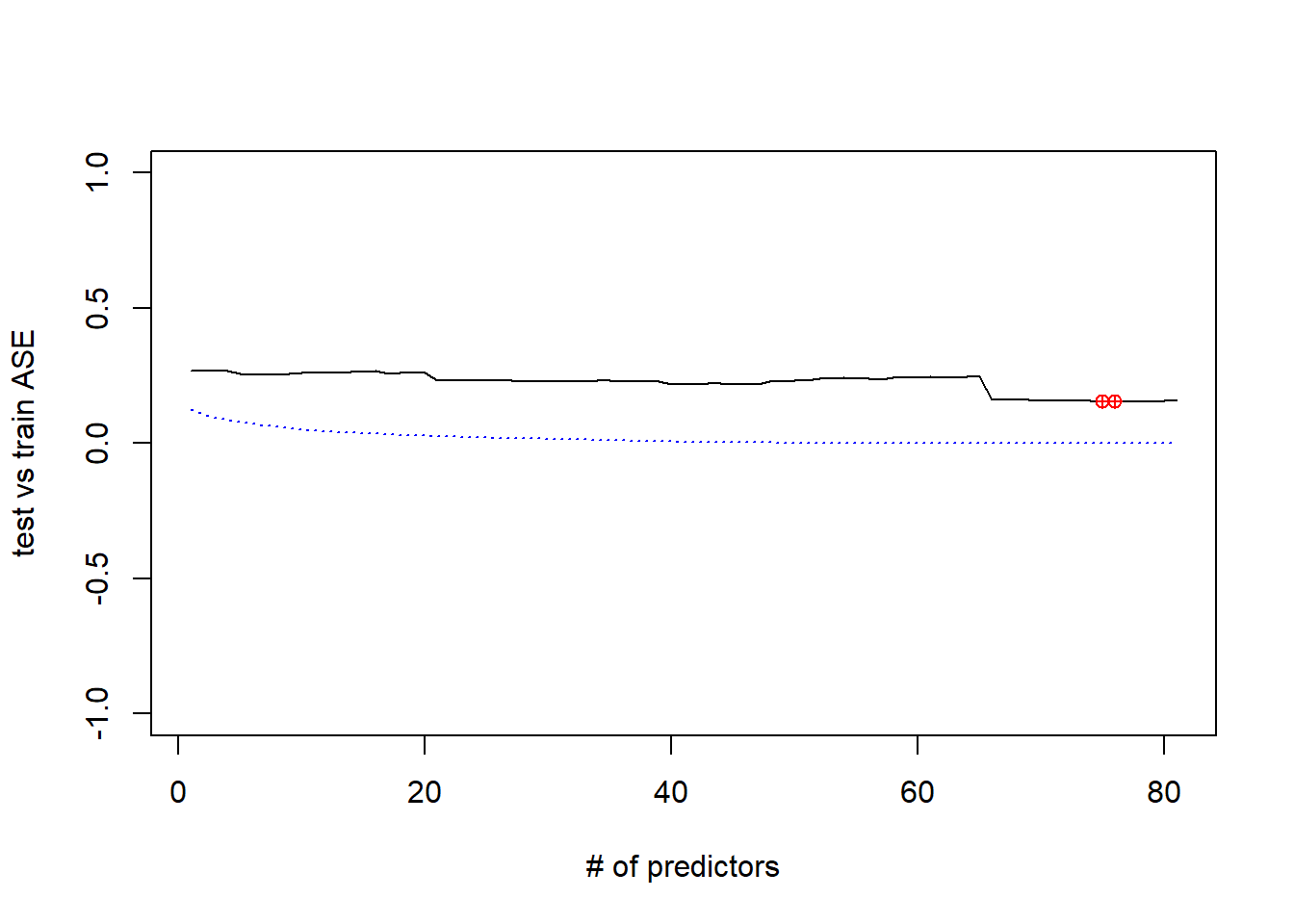
The coefficients of model are described in Figure 17. This taught our team a lesson on the importance of test and training set splits, and how relying solely on the adjusted R-squared metric could leave you in hot water when trying to leverage the insights gleaned from a training set on separate data points.

The next step in this discovery led us to compare this model to the feature selection methods like LASSO and forward linear regression models. For this model, we wanted to find a model that might be hard to interpret but would be better at prediction. The LASSO model we built leveraged both logged *msrp* and logged mpg, in addition to all other potential variables including vehicle. As a reminder, we left vehicle out of our intuitive model because of the complexity and number of levels.  What the LASSO approach discovered that 70 variables were significant in minimizing the ASE. When we separated the data set into a training and test set, the test MSE of the model was 0.759. This was not as small as the test MSE of the intuitive model! The figure below is an output of the LASSO model that helps illustrate the number of variables and the MSE at each stage of the model building process.



After analyzing the differences in the MSE’s, we immediately reflected on why our simple intuitive model is performing better than the LASSO. Our analysis brought us back to the concept of overfitting. The LASSO model is attempting to leverage as many variables as possible to lower the MSE to the training set. That training set is likely to not include variables for all *carclass* levels, and even more likely vehicle levels. Because of this, when the model encounters levels from these factors it has never seen before in the training set, it makes an incorrect prediction that skews the average mean squared error to be larger.

We then wanted to compare the results of a forward selection model. It came to the conclusion of having roughly 75 variables and resulted in a test MSE of 0.3. Below is the graphical representation of the results of the forward model.

****

*Conclusion:*

In conclusion, the intuitive model with only 2 predictors (*log(mpg) and accelrate\*log(mpg*)) leveraging a logged response gave the best MSE. We used a train and test split in order to validate the MSE across an intuitive, forward, and LASSO models.

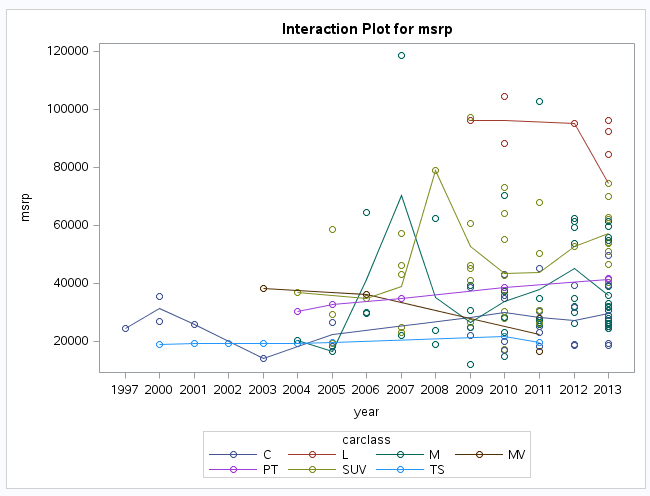
**Objective 2**

For the second objective, we ran a 2-way ANOVA on the data using *year* and *carclass* as the categorical variables. Below are tables that illustrate the summary statistics. We can see immediately within the *carclass* variable the sample sizes are not equal. This is something we may need to consider later on in our analysis. We also see that for the year variable, we have a range of years from 1997 to 2013. This is important to discuss because there is a visible trend within the dataset, there are very few hybrid cars for the years 1997 to 2009, and then these values double or triple over the remaining years from 2009 to 2013. We believe this is likely due to the technological advancements made during these years in hybrid car technology. This is also likely due to the financial benefit of leveraging a hybrid vehicle due oil and gas premiums in that began in 2008.

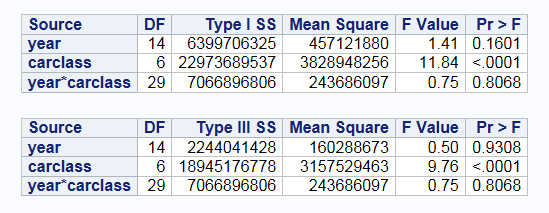
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Car Class | C | L | M | MV | P | SUV | TS |
| Frequency | 32 | 8 | 56 | 4 | 6 | 39 | 8 |
| Proportion | 0.209 | 0.052 | 0.366 | 0.026 | 0.039 | 0.255 | 0.052 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Value | 1997 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
| Frequency | 1 | 3 | 2 | 1 | 3 | 4 | 8 | 5 | 8 | 4 | 14 | 23 | 20 | 14 | 43 |
| Proportion | 0.007 | 0.02 | 0.013 | 0.007 | 0.02 | 0.026 | 0.052 | 0.033 | 0.052 | 0.026 | 0.092 | 0.15 | 0.131 | 0.092 | 0.281 |

*General Model:*



Looking at the interaction plot above, it’s not entirely clear of the data conforms to an additive model. Car class levels that are higher tending to stay higher and vise-versa, but there does seem to be some cross-over. The overall model has a p-value of 0.0002, meaning that the model as a whole is significant (Figure 9).



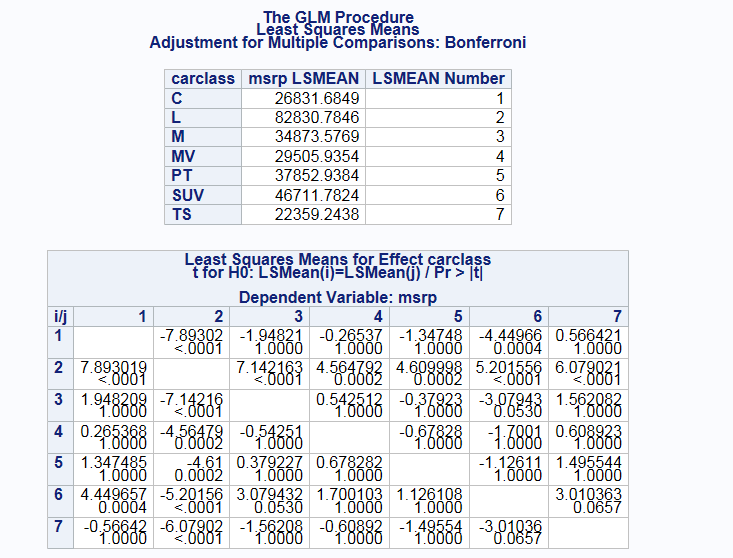
Looking at the Type III SS table, we can see that the interaction term between year and car class is not significant. This means we fail to reject the null hypothesis that the interaction term is not equal to zero and that we have an additive model. The main effect for *year* is not significant as well, however *carclass* is. Next, we verified the assumptions for a two-way ANOVA and tested contrasts between the different levels of *carclass* to see which were significant.

*Assumptions:*  
 The assumptions for a two-way ANOVA are independence, constant variance, and normality.  The residual plots for the ANOVA are shown in Figure 10. The qq-plot does not indicate that there are any normality violations. There may be instances of unequal standard deviations between the different categories. Looking at Figure 3A with boxplots of the different levels of car class, they do seem to have pretty different spreads. We tried taking the log of *msrp* to see if that would help, but it doesn’t seem like it improved the different spreads all that much (Figure 3B) The difference in spread for *carclass* could be a result of not having enough data points. Levels with smaller spreads tend to have smaller sample sizes. The appearance of unequal spread could be due to the data being sparse. For this analysis, we will continue with the assumption that there are constant variances and that the data is independent.

Just to ensure our analysis was robust and accurate, we also reran this analysis using a logged response (*msrp*) but found that it resulted in the same conclusions. The rest of the analysis you see below is referred to using an unlogged response. (Figure 19)

*Pairwise Analysis:*

Since the model is additive, we did not need to analyze the interaction term between year and car class. We focused our pairwise analysis on the main effect of *carclass* only, because the year main effect was not significant. We ran pairwise analysis of all of the combinations in car class with a Bonferroni adjustment. The results are shown below.



Car class L was significantly different from every other car class. The only other significantly different pair at an family-wise alpha level of 0.05 was car class C and SUV. The 95% confidence intervals for each of the pairs give a visual indication of which combinations are significant (Figure 11).

*Conclusion:*

The linear model that we built had only one level of the *carclass* and *year* interaction term that was significant, and that was with other variables included. It makes sense compared to the linear model that *year* and the interaction term were not significant in the 2-way ANOVA. Car class L represents large vehicles, so it is understandable that on average they would be more expensive than the other types of cars. However, levels L, MV, P, and TS have smaller sample sizes than the other two so that could be preventing true differences between the levels from being detected.

**Appendix**

**Sources:**

D-J. Lim, S.R. Jahromi, T.R. Anderson, A-A. Tudorie (2014).   
"Comparing Technological Advancement of Hybrid Electric Vehicles (HEV) in  
Different Market Segments," Technological Forecasting & Social Change,  
<http://dx.doi.org/10.1016/j.techfore.2014.05.008>

Description of dataset: <http://users.stat.ufl.edu/~winner/data/hybrid_reg.txt>

Download of dataset: <http://users.stat.ufl.edu/~winner/data/hybrid_reg.csv>

**Important Figures:**

Figure 1:

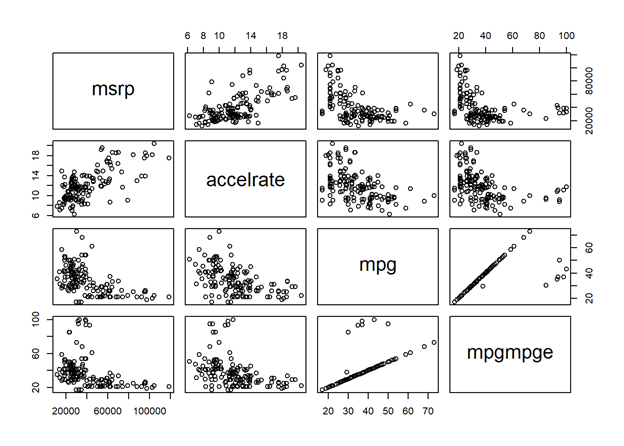


Figure 2A:

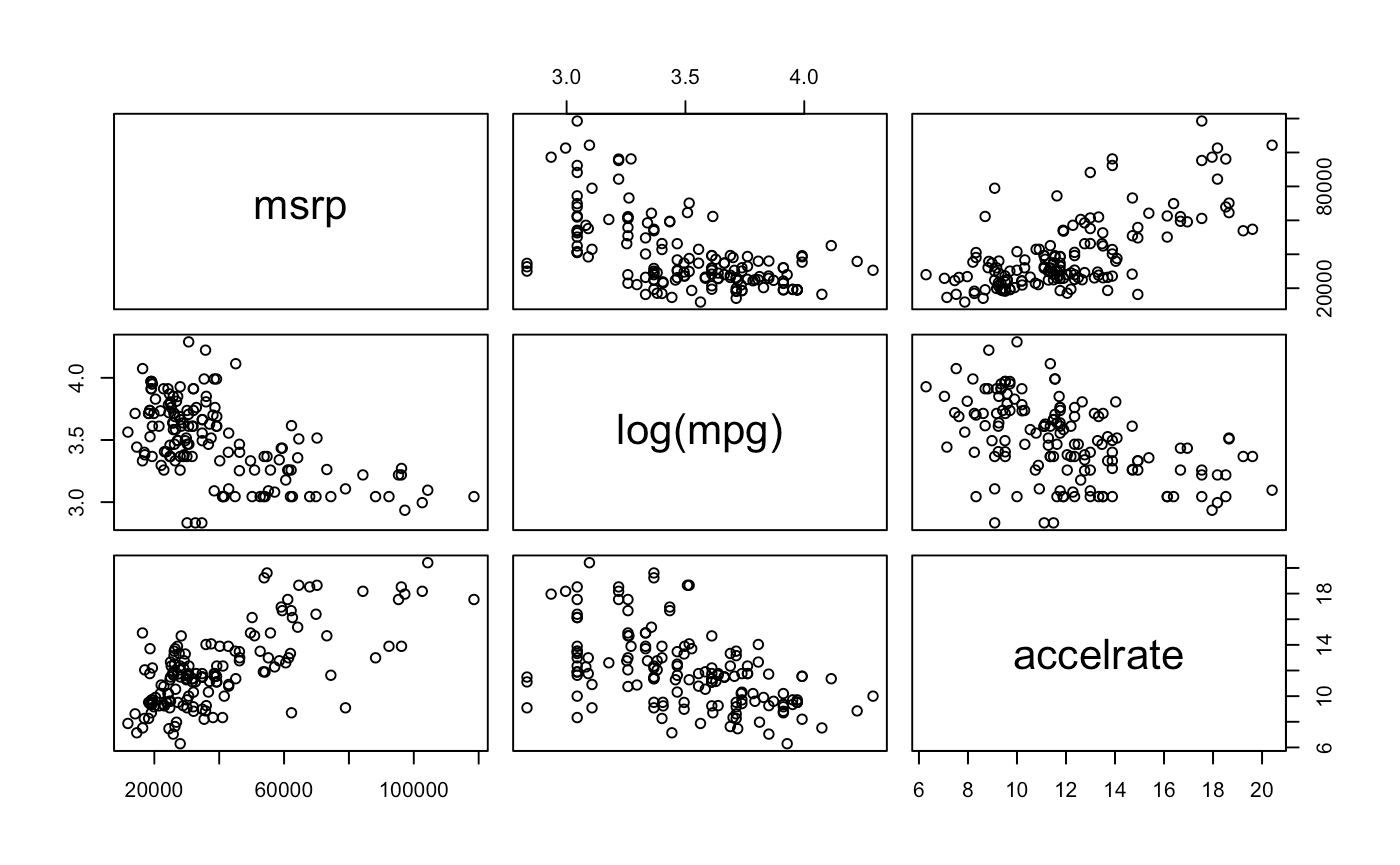


Figure 2B:

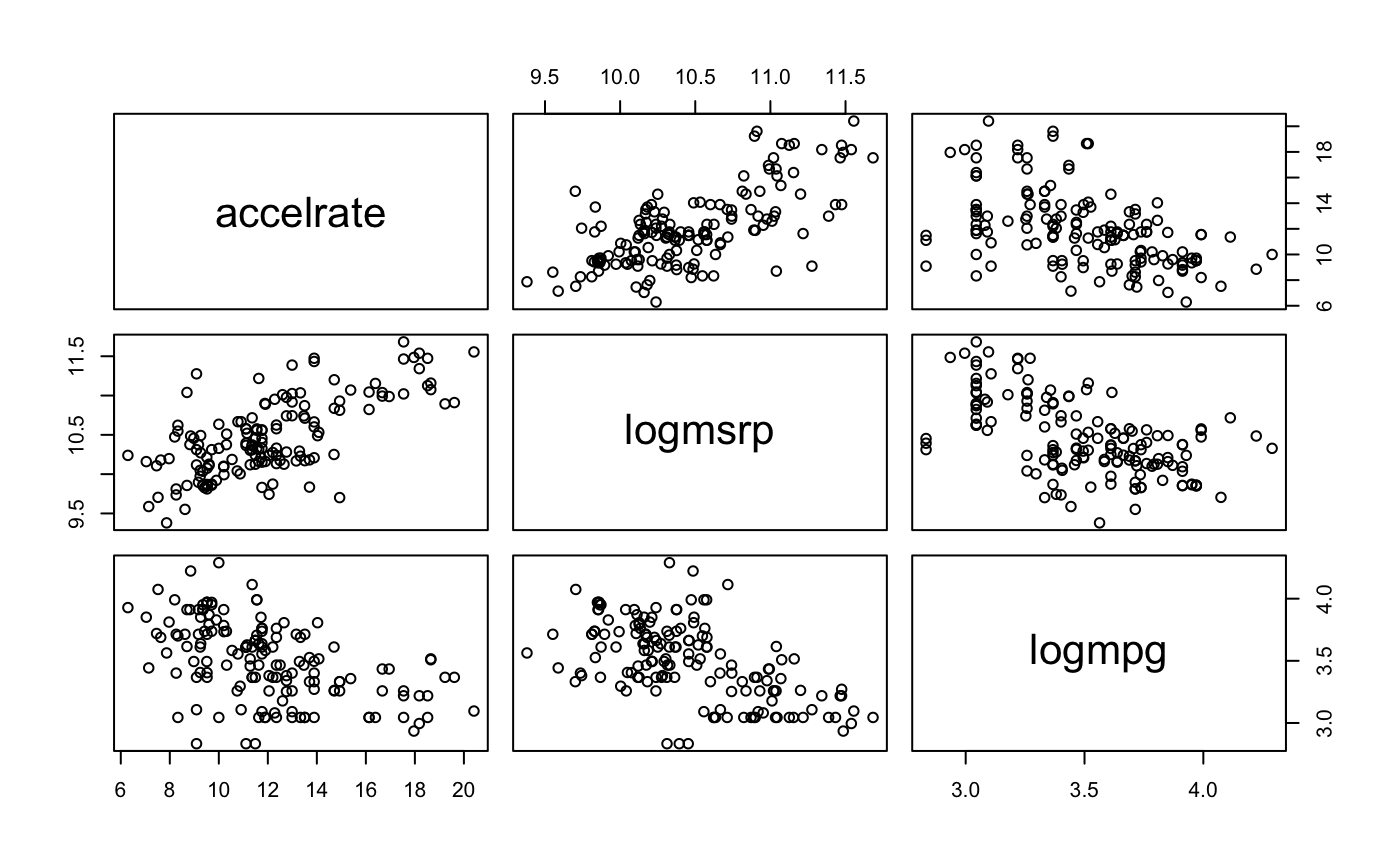


Figure 3A:

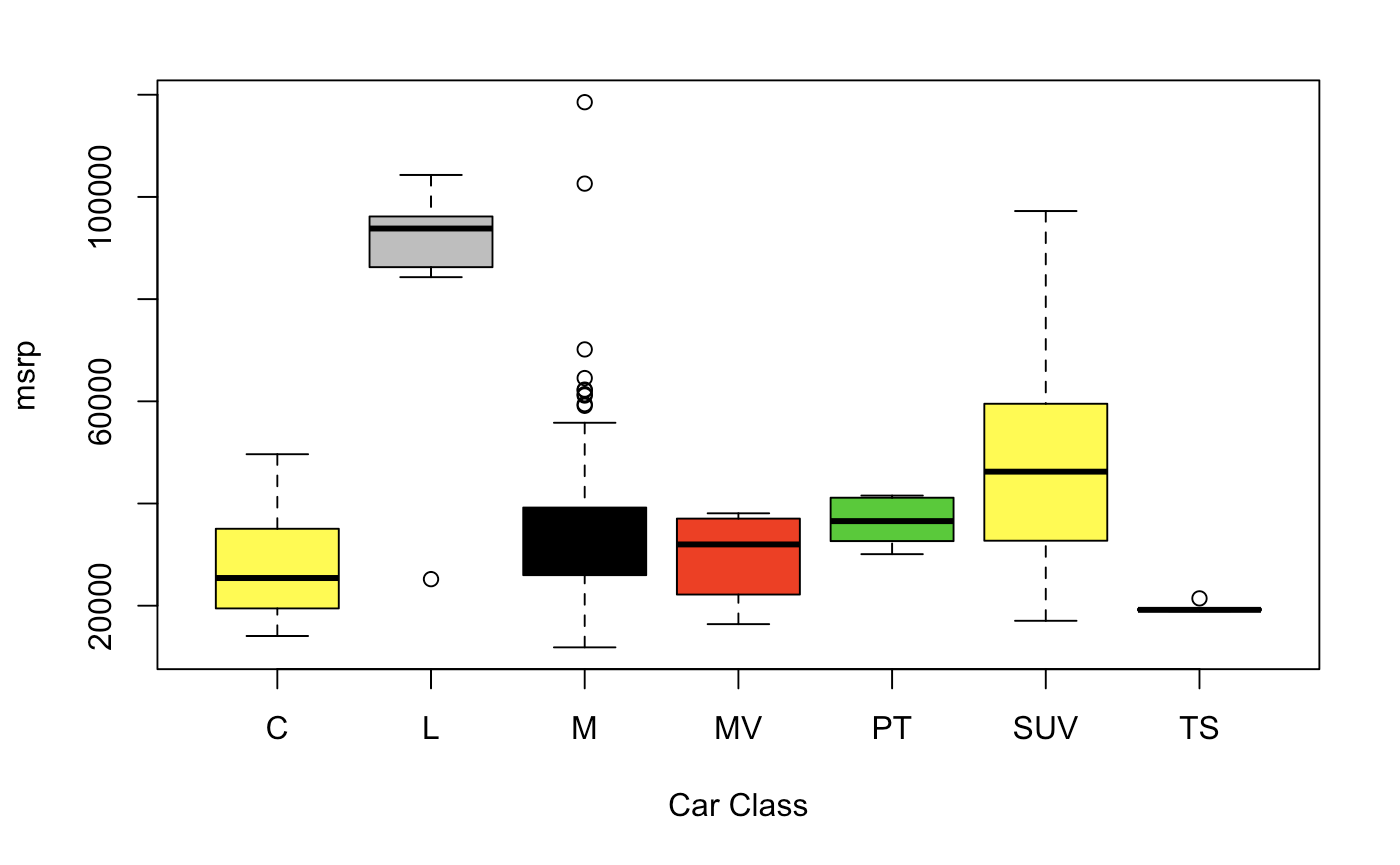


Figure 3B:

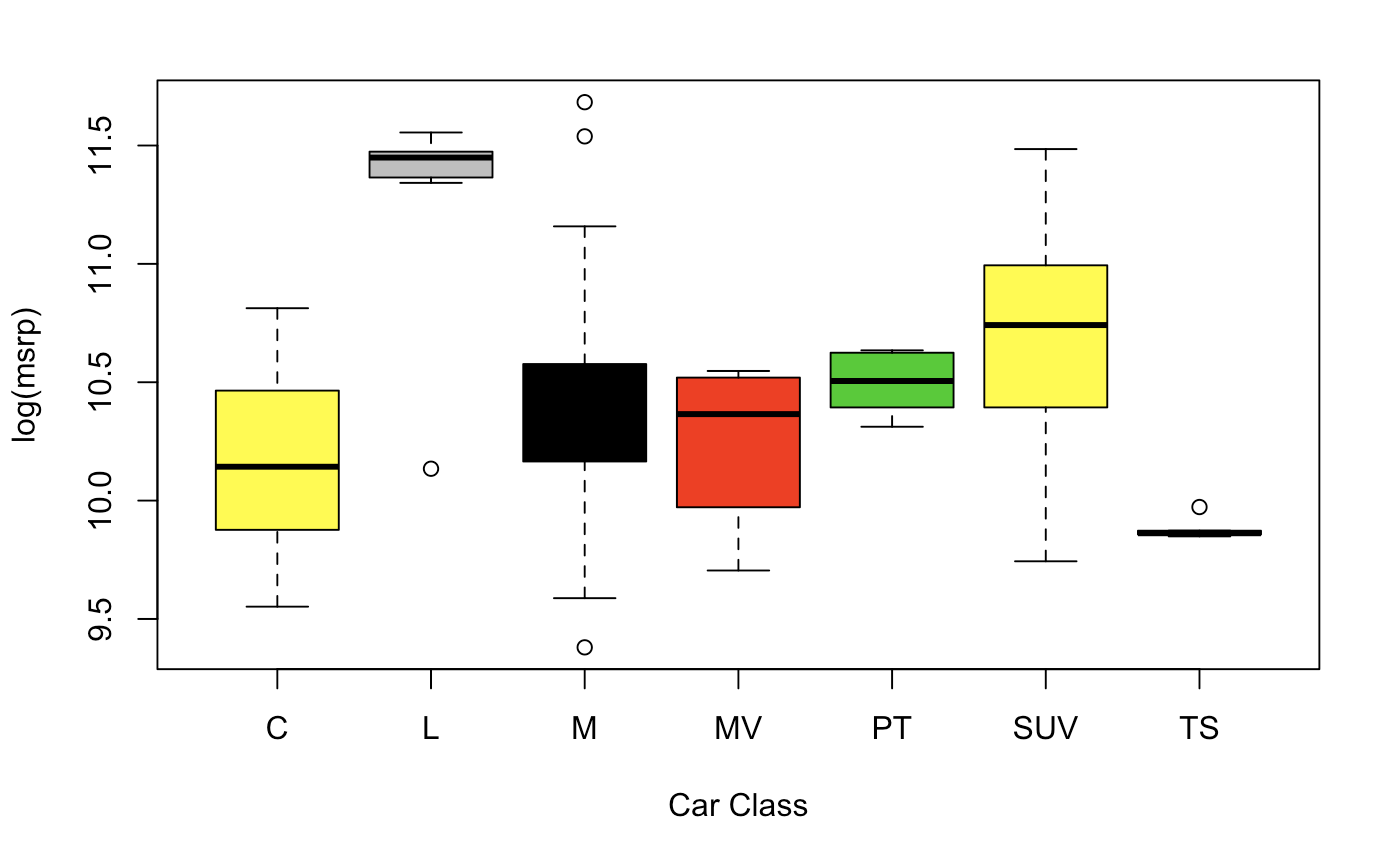


Figure 4:

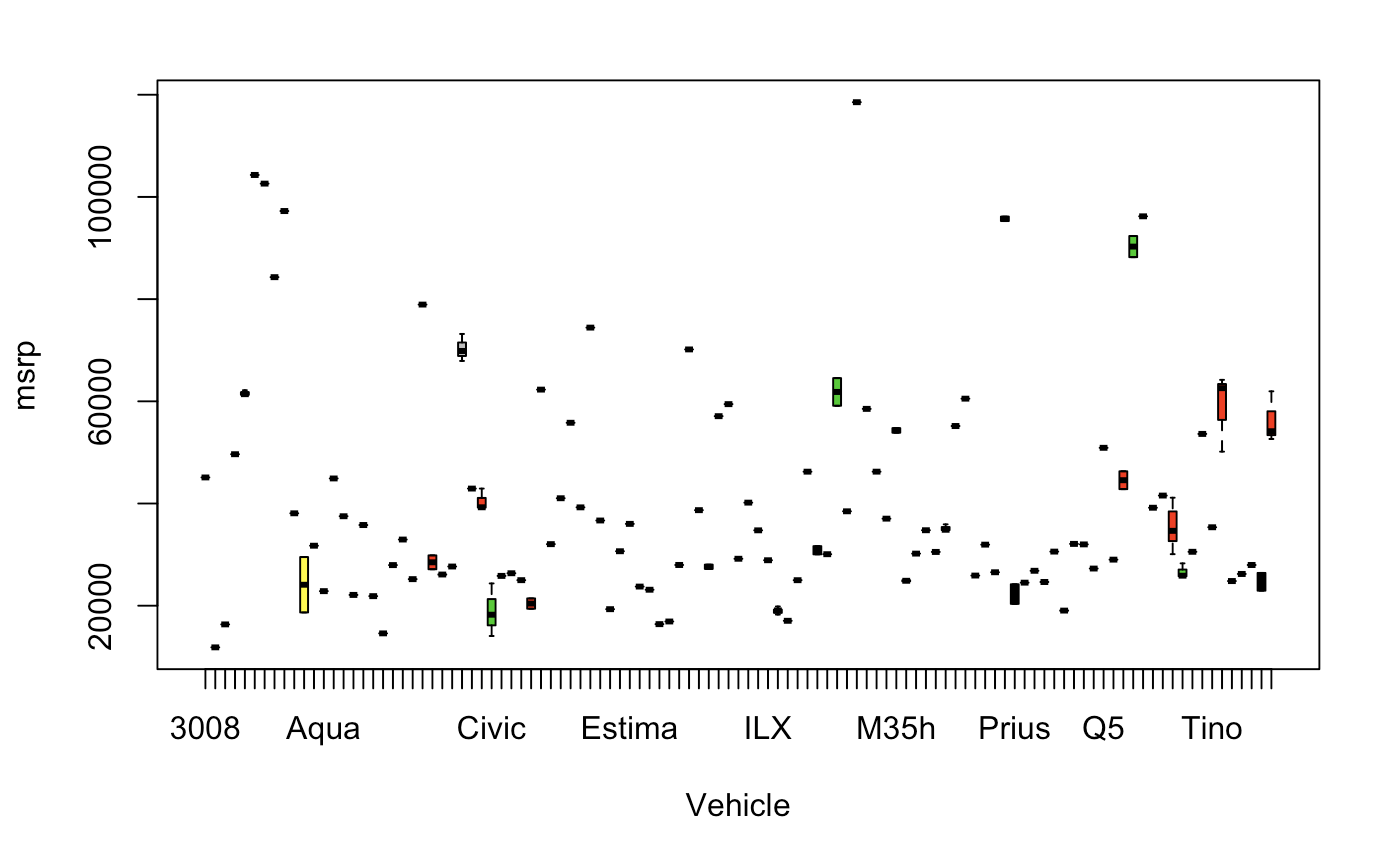


Figure 5:

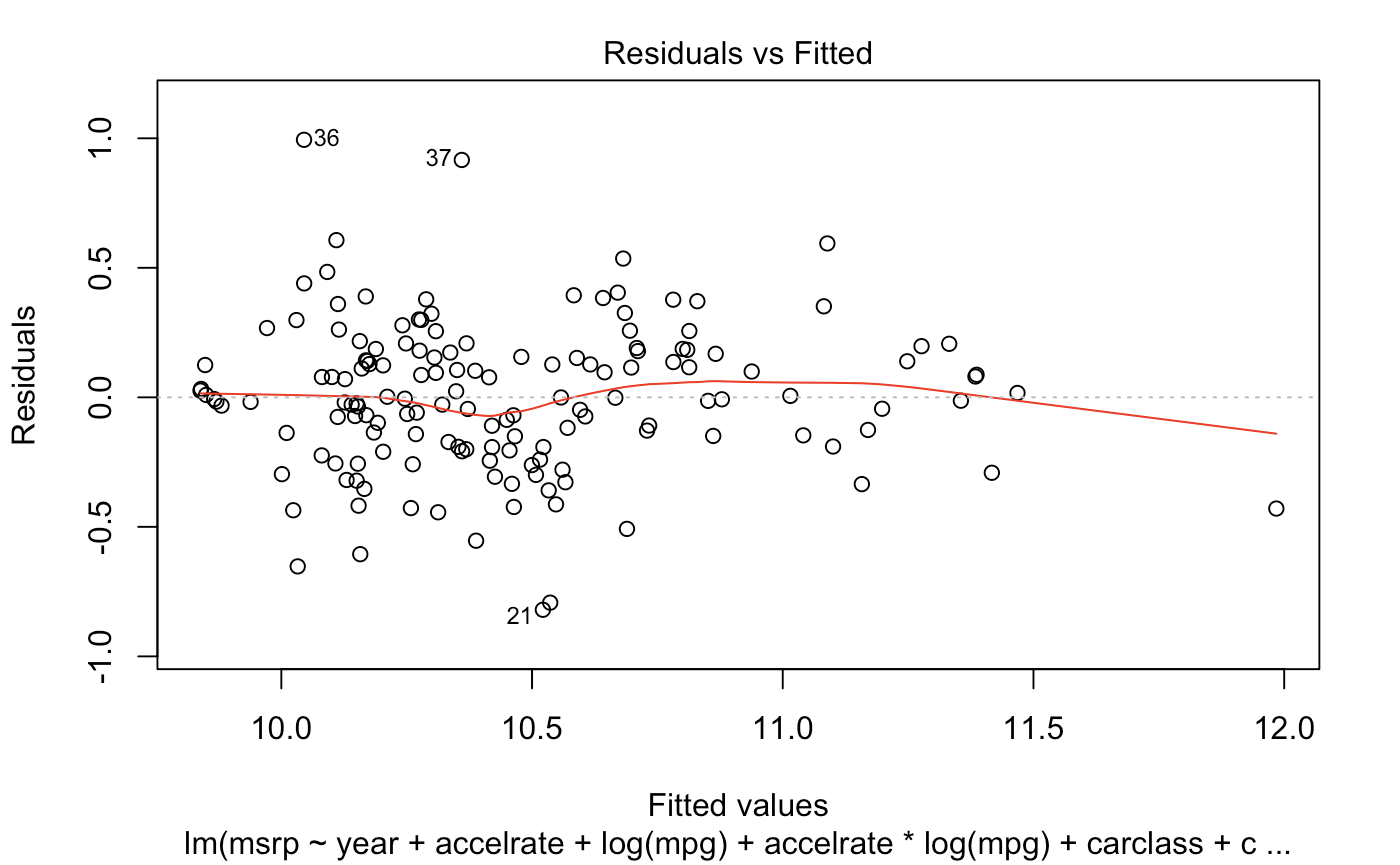


Figure 6:

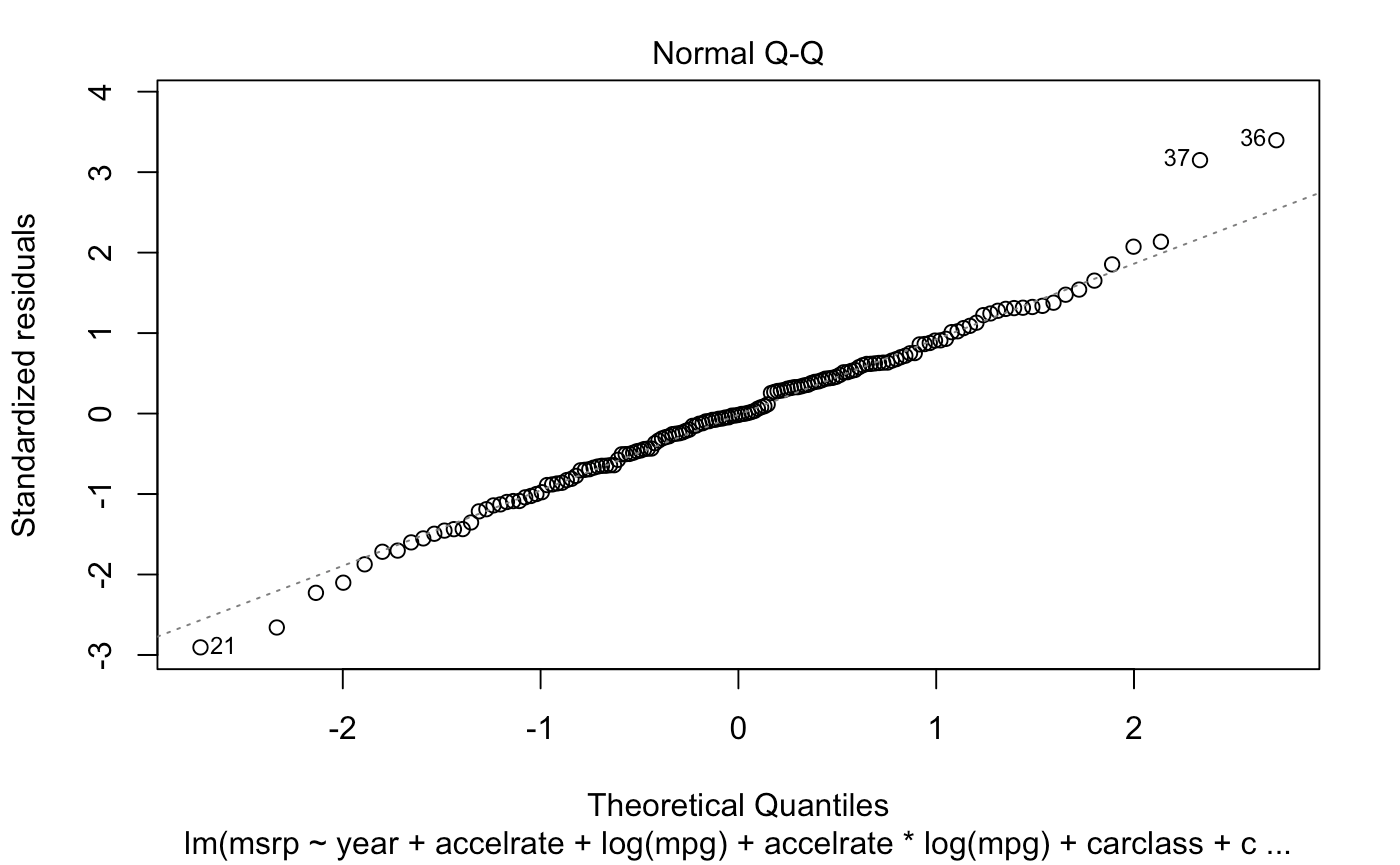


Figure 7:

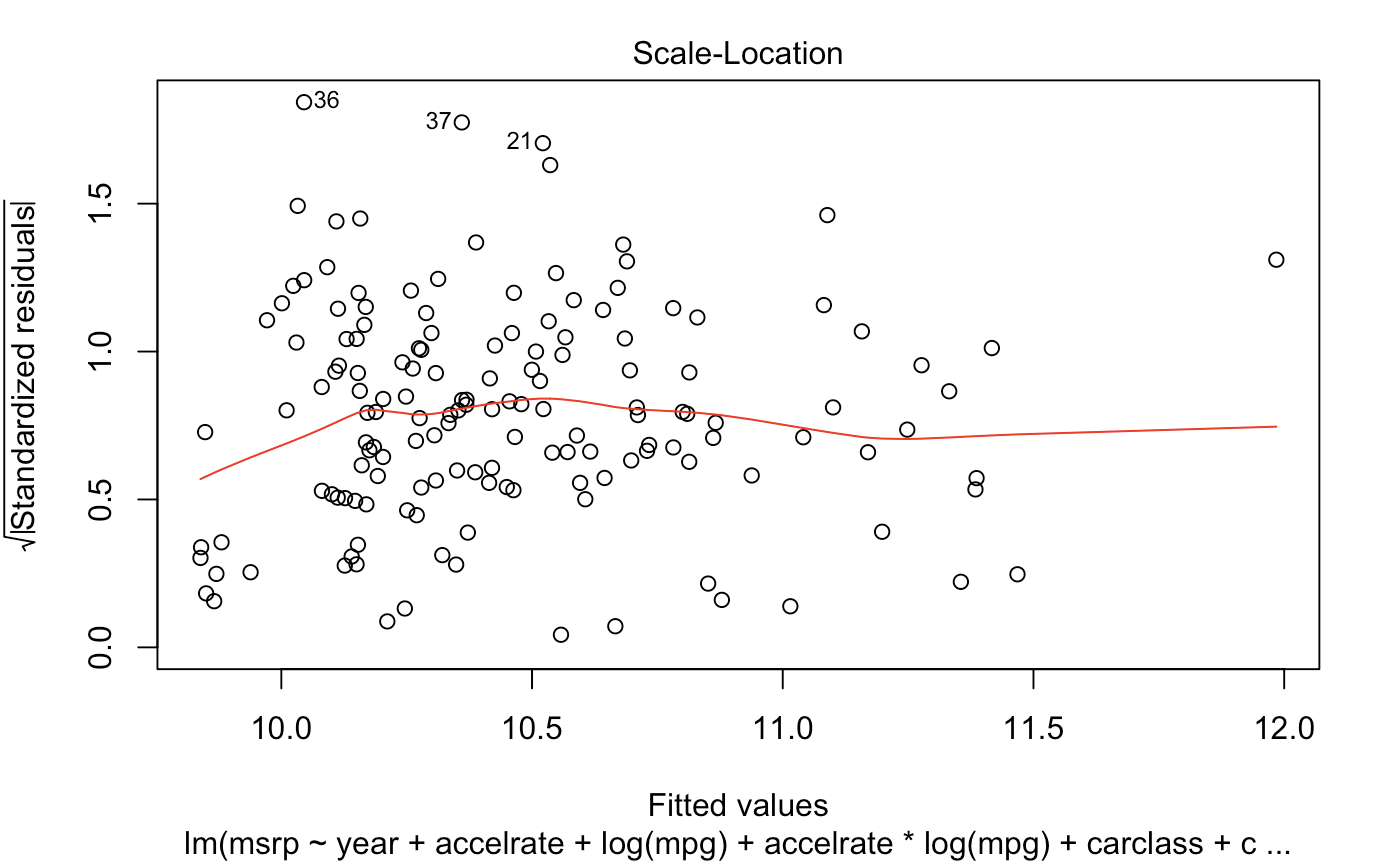
****

Figure 8:

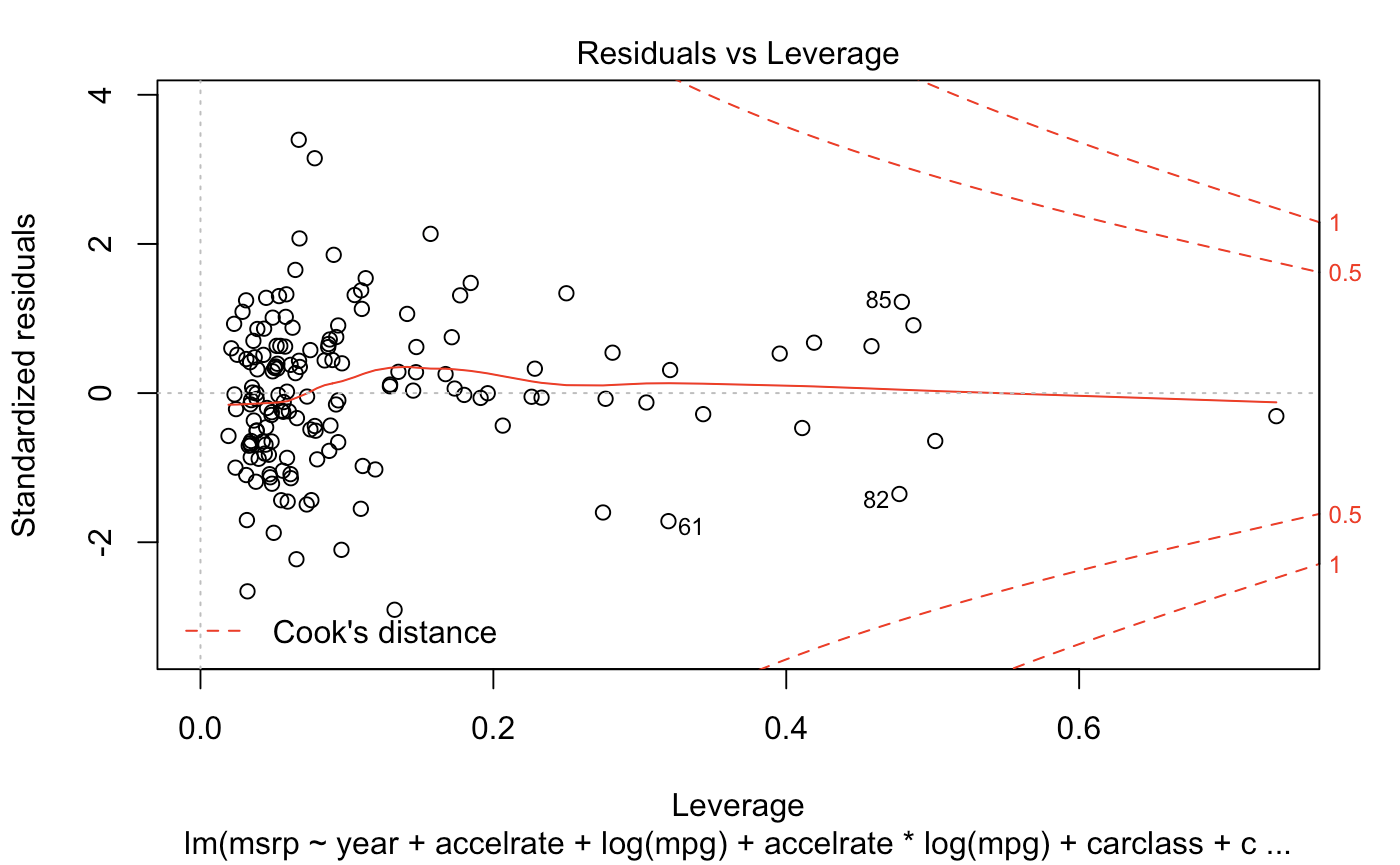


Figure 9:

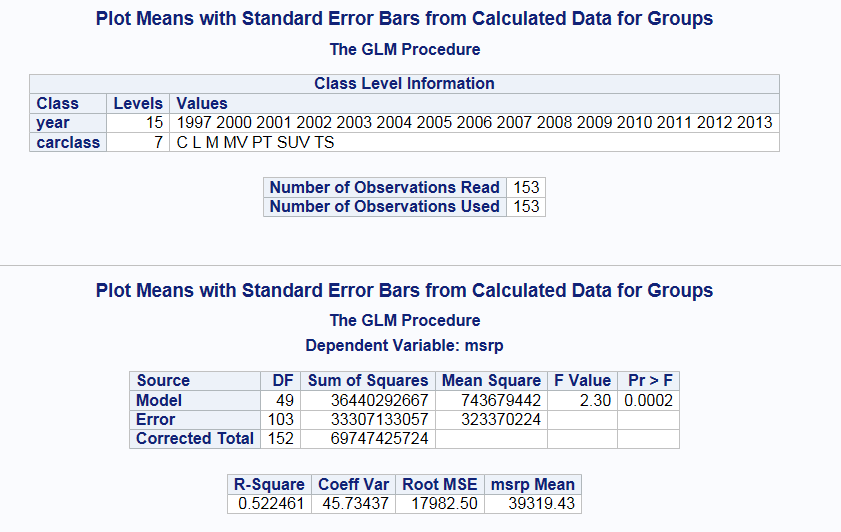
****

Figure 10:

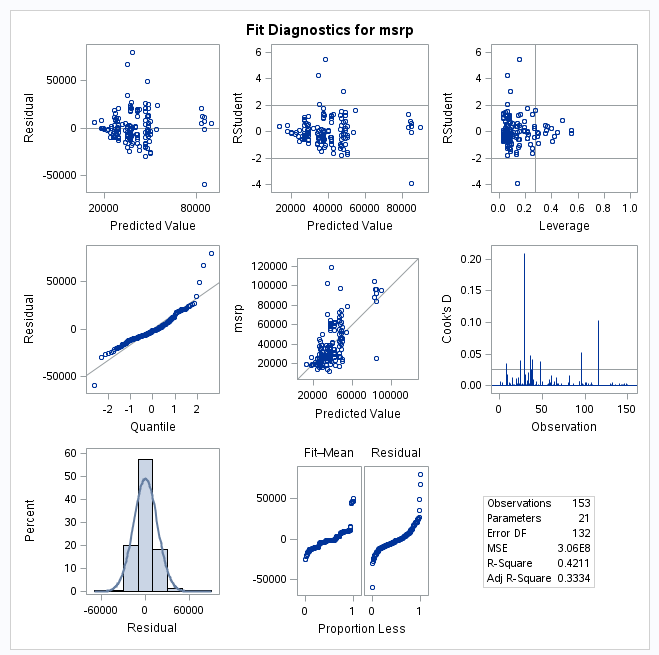
****

Figure 11:

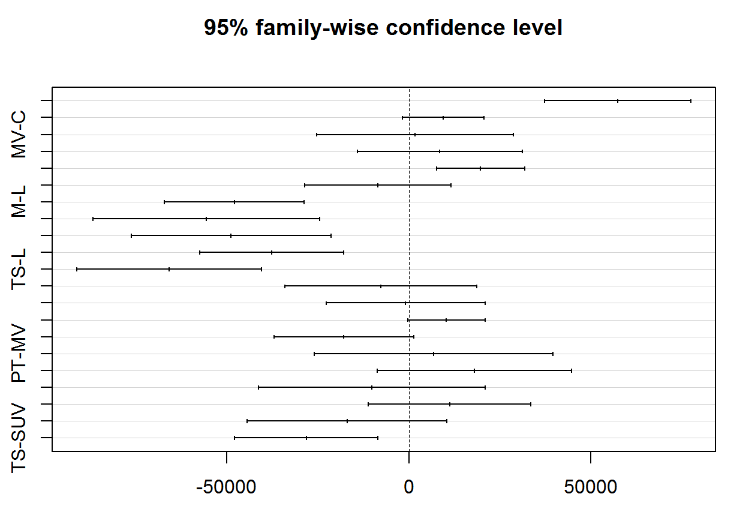
****

Figure 12:

|  |  |  |
| --- | --- | --- |
|  | Confidence Interval Lower Limits (2.5%) | Confidence Level Upper Limit (97.5%) |
| (Intercept) | -1694801.1006 | 1858256.0813 |
| Year | -991.4201 | 740.0808 |
| Accelrate | 14108.8072 | 32544.3058 |
| log(mpg) | 15872.8516 | 82672.2832 |
| Car Class L | 1069483.9397 | 19500055.2276 |
| Car Class M | -5525762.6082 | 248226.7059 |
| Car Class MV | -2488313.7098 | 10456030.7932 |
| Car Class PT | -6781192.7817 | 3192971.7585 |
| Car Class SUV | -3856851.1440 | 2335961.7322 |
| Car Class TS | -3576622.9973 | 4705040.1343 |
| accelrate\*log(mpg) | -8596.6475 | -3114.6725 |
| Year\*car class L | -9680.9279 | -518.9414 |
| Year\*Car Class M | -124.5486 | 2746.9976 |
| Year\*Car Class MV | -5206.3466 | 1241.6960 |
| Year\*Car Class PT | -1590.9123 | 3372.9309 |
| Year\*Car Class SUV | -1162.1582 | 1919.6350 |
| Year\*Car Class TS | -2350.3975 | 1779.4774 |

Figure 13:

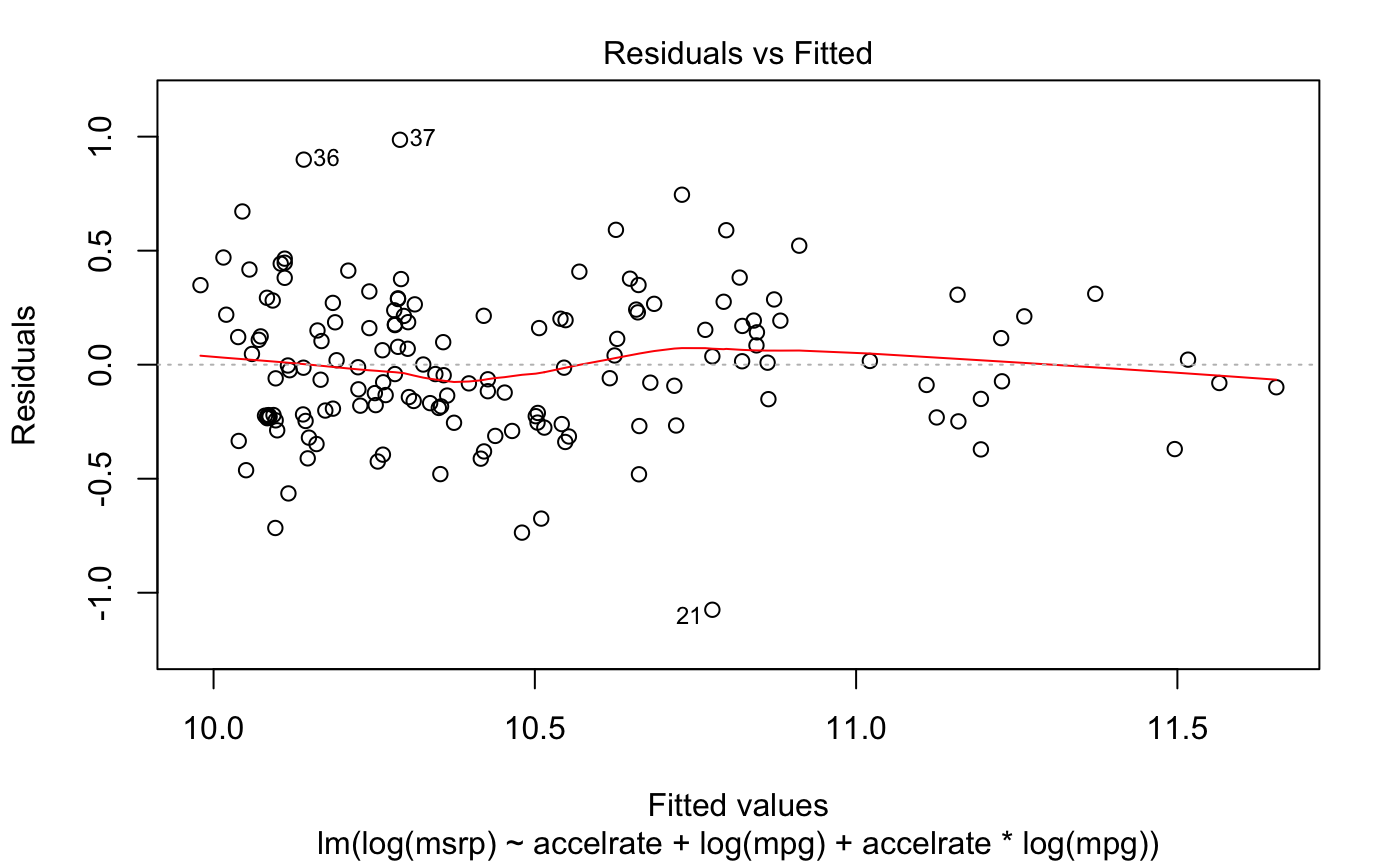


Figure 14:

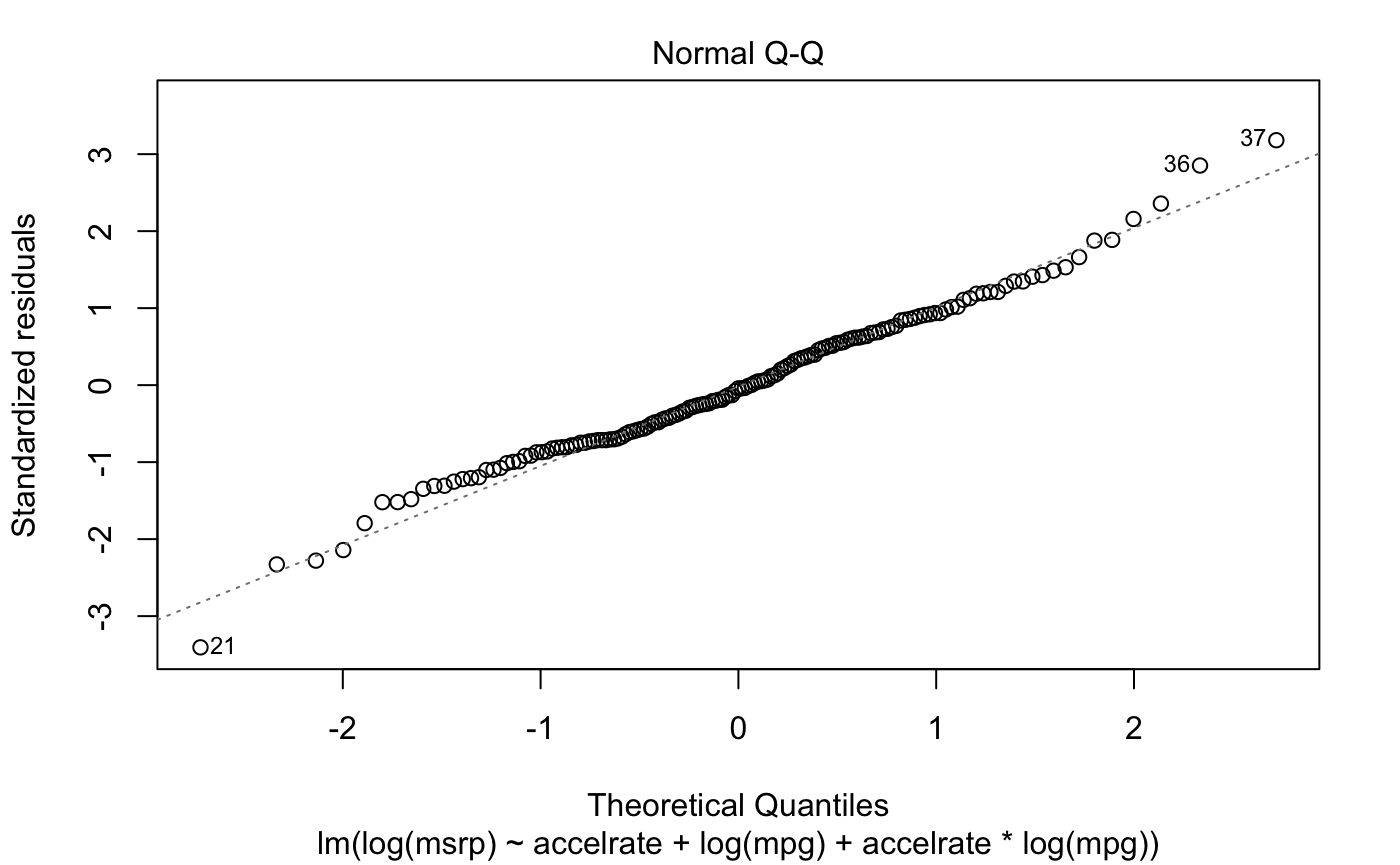


Figure 15:

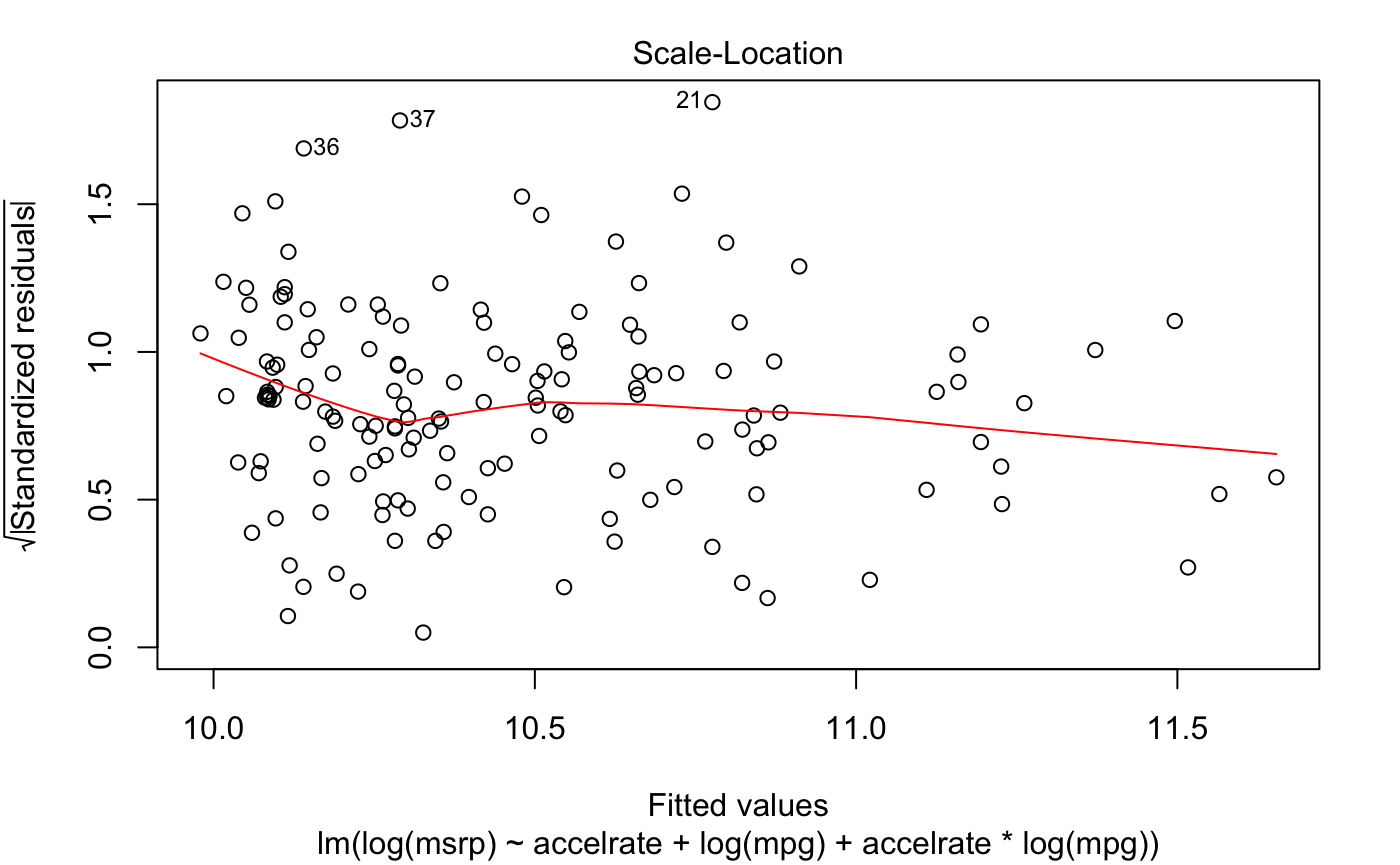


Figure 16:

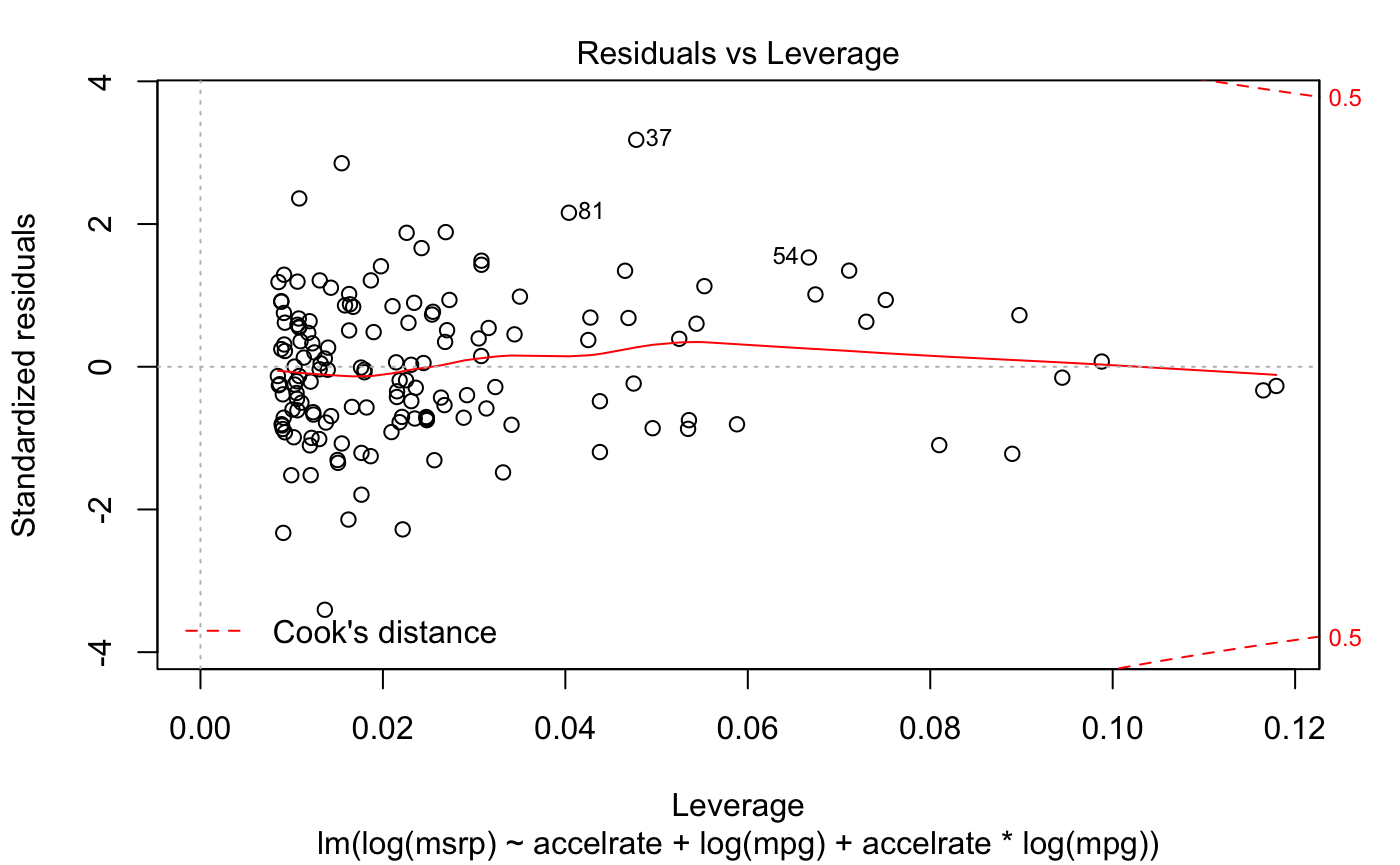


Figure 17:

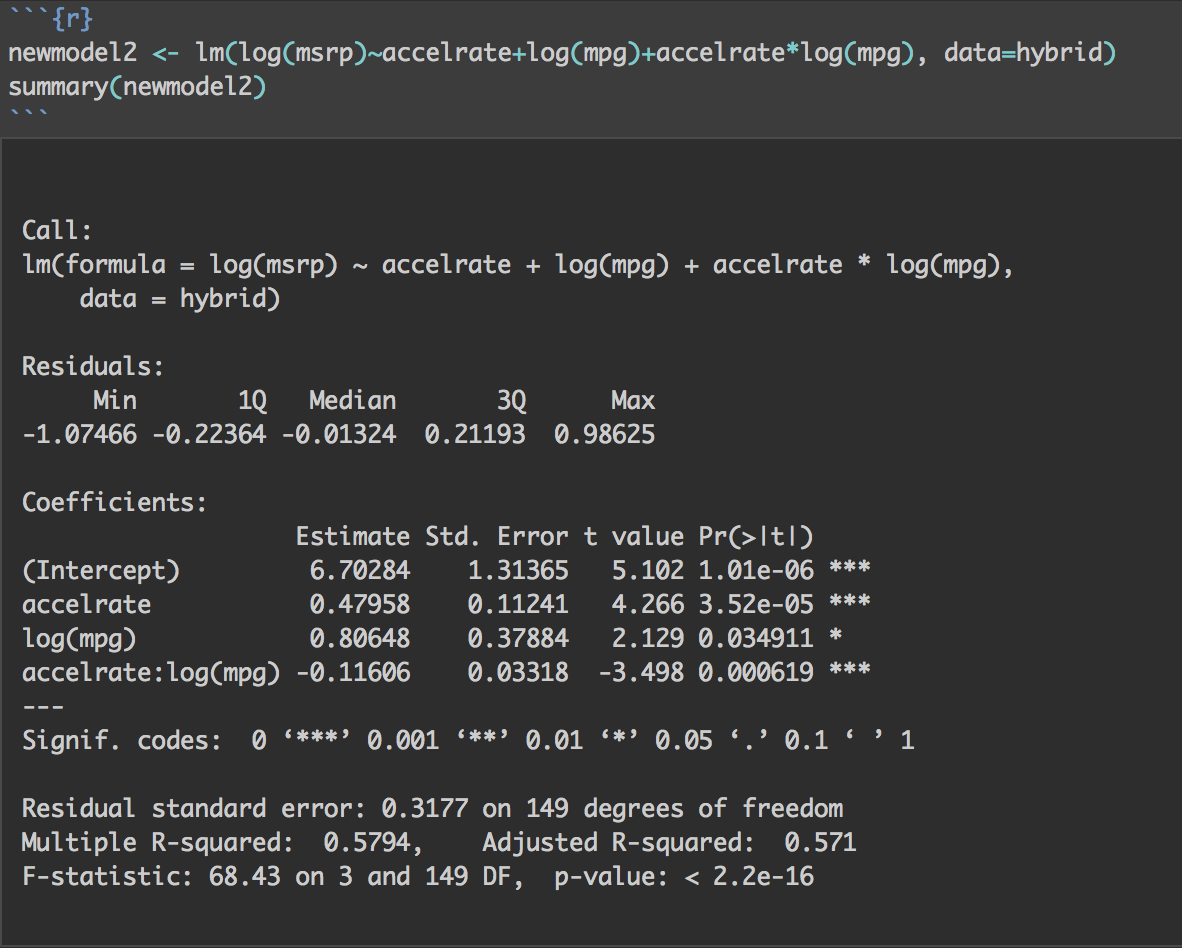


Figure 18:

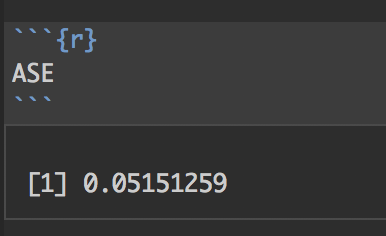
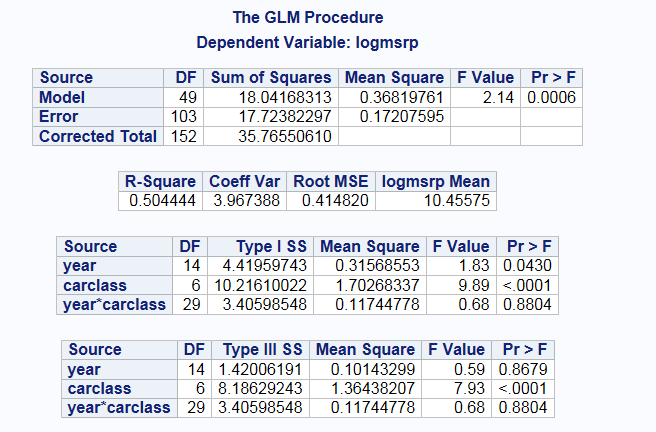
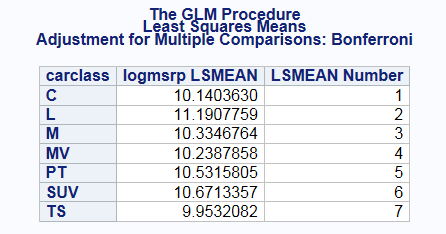


Figure 19:





**Code:**

Linear Regression:

# Read in data

hybrid <- read.csv("hybrid\_reg.csv", header=TRUE)

# Find scatter plots with year, car model, and car class removed because they are categorical

pairs(hybrid[,-c(1,2,3,8,9)])

# Categorical Car Class Box Plot

plot(as.factor(carclass), msrp, xlab="Car Class",ylab="log(msrp)",title="Hybrid Car Dataset",col=c(7,32,57,82,107))

# Categorical Vehicle Box Plot

plot(as.factor(vehicle), msrp, xlab="Vehicle",ylab="msrp",title="Hybrid Car Dataset",col=c(7,32,57,82,107))

# Logging msrp and mpg for assumptions

hybrid$logmsrp <- log(hybrid$msrp)

hybrid$logmpg <- log(hybrid$mpg)

pairs(hybrid[,c(5,10,11)])

pairs(hybrid[,c(5,6,10)])

# Removing columns that aren’t needed

hybrid2 <- hybrid[,-c(1,7,9,10,11)]

hybrid2$mpg <- log(hybrid2$mpg)

str(hybrid2)

#Lasso model

library(glmnet)

set.seed(123)

index<-sample(1:dim(hybrid2)[1],100,replace=F)

train<-hybrid2[index,]

test<-hybrid2[-index,]

x=model.matrix(msrp~.,train)[,-1]

y=train$msrp

xtest<-model.matrix(msrp~.,test)[,-1]

ytest<-test$msrp

grid=10^seq(10,-2, length =100)

lasso.mod=glmnet(x,y,alpha=1, lambda =grid)

cv.out=cv.glmnet(x,y,alpha=1)

plot(cv.out)

bestlambda<-cv.out$lambda.min

lasso.pred=predict (lasso.mod ,s=bestlambda ,newx=xtest)

testMSE\_LASSO<-mean((ytest-lasso.pred)^2)

testMSE\_LASSO

# Forward model

library(leaps)

forward=regsubsets(msrp~.,data=train,method="forward",nvmax=117)

par(mfrow=c(1,3))

bics<-summary(forward)$bic

plot(1:81,bics,type="l",ylab="BIC",xlab="# of predictors")

index<-which(bics==min(bics))

points(index,bics[index],col="red",pch=10)

adjr2<-summary(forward)$adjr2

plot(1:81,adjr2,type="l",ylab="Adjusted R-squared",xlab="# of predictors")

index<-which(adjr2==max(adjr2))

points(index,adjr2[index],col="red",pch=10)

rss<-summary(forward)$rss

plot(1:81,rss,type="l",ylab="train RSS",xlab="# of predictors")

index<-which(rss==min(rss))

points(index,rss[index],col="red",pch=10)

predict.regsubsets =function (object , newdata ,id ,...){

 form=as.formula (object$call [[2]])

 mat=model.matrix(form ,newdata )

 coefi=coef(object ,id=id)

 xvars=names(coefi)

 mat[,xvars]%\*%coefi

}

testASE<-c()

for (i in 1:81){

  predictions<-predict.regsubsets(object=forward,newdata=test,id=i)

  testASE[i]<-mean(((test$msrp)-predictions)^2)

}

par(mfrow=c(1,1))

plot(1:81,testASE,type="l",xlab="# of predictors",ylab="test vs train ASE",ylim=c(-1,1))

index<-which(testASE==min(testASE))

points(index,testASE[index],col="red",pch=10)

rss<-summary(forward)$rss

lines(1:81,rss/100,lty=3,col="blue")

forwardfinal=regsubsets(msrp~.,data=train,method="forward",nvmax=76)

coef(forward,75)

# The intuitive model (reminder that only mpg is logged here, it was done it previous dataset)

easy <- lm(msrp~year+accelrate+mpg+accelrate\*mpg+carclass+year\*carclass, data=hybrid2)

summary(easy)

plot(easy)

confint(easy)

# The new intuitive model w/ MSE calculated

loggednewmodel <- lm(msrp~accelrate+mpg+accelrate\*mpg, data=hybridlogged)

summary(newloggedmodel)

plot(newmodel)

hybridtrain2 <- hybridlogged [1:100, ]

hybridtest2 <-hybridlogged [101:150, ]

loggedtrainmodel<-lm(msrp~accelrate+mpg+accelrate\*mpg, data=hybridlogged)

pred<-predict(loggedtrainmodel, newdata=hybridtest2)

ASE<-mean((hybridtest2$msrp-pred)^2)

ASE

Two-way Anova:

proc glm data=Hybrid PLOTS=(DIAGNOSTICS RESIDUALS);

class year carclass;

model msrp=  year carclass ;

lsmeans carclass / pdiff tdiff adjust=bon;

estimate 'C vs L' year -1 1 0;

Run;

# Removed the Interaction, rerun the code

```{r modelfit2 }

model.fit<-aov(msrp~carclass+year,data=Hybrid)

par(mfrow=c(1,2))

plot(model.fit$fitted.values,model.fit$residuals,ylab="Resdiduals",xlab="Fitted")

qqnorm(model.fit$residuals)

```

```{r}

library(car)

Anova(model.fit,type=3)

```

proc glm data=Hybrid PLOTS=(DIAGNOSTICS RESIDUALS);

class year carclass;

model msrp=  year carclass ;

lsmeans carclass / pdiff tdiff adjust=bon;

run;

n;

data Hybrid2;

set Hybrid;

logmsrp=log(msrp);

run;

proc glm data=Hybrid2 PLOTS=(DIAGNOSTICS RESIDUALS);

class year carclass;

model logmsrp=  year carclass ;

lsmeans carclass / pdiff tdiff adjust=bon;

run;

proc glm data=Hybrid2;

class carclass;

model logmsrp=carclass;

means carclass /tukey;

run;

proc glm data=Hybrid2 PLOTS=(DIAGNOSTICS RESIDUALS);

class year carclass;

model logmsrp=  year carclass year\*carclass ;

lsmeans carclass / pdiff tdiff adjust=bon;

run;