PSTAT 126 Final Project

Due Date: Monday, December 11, 5pm

Reed Gilbreth, Isabelle Lambert, Davis Messer, Orla Ruane

Abstract

In our final project, we began with a data set titled "Hybrid", which includes 153 observations of hybrid models. Using this data set, we set out to find the most effective and efficient predictor variables to model our chosen response variable: MSRP. Through analysis, we discovered acceleration rate, MPG, car class, and their interactions produced the optimal model:

$$log(MSRP) = (0.32909) * accelrate + (0.06706) * MPG + (0.49919) * carclass_{id} + (-0.00512) * accelrate * MPG$$
$$+(-0.02409) * accelrate * carclass_{id} + (0.00645) * MPG * carclass_{id}$$

Problem & Motivation

Our study aims to examine and analyze the data contained in the 2013 Hybrid data set, more specifically, the MSRP's of a selection of hybrid vehicles. We seek to identify the set of given predictor variables that best explains the makeup of our chosen response variable, MSRP. Our analysis will prove very valuable in determining the influence of different variables on MSRP. Essentially, we are looking to discover positive and negative correlation between different variables and MSRP, and use this information to create a model, which can be used to determine the value of a hybrid car. This study's practical application is extremely important because it will assist in both the car buying and selling market; offering the fairest price for both buyer and seller. Throughout the project, we worked through statistical analysis in order to find the relationship these factors had on our response variable (MSRP) visible, and ultimately provide us with a useful regression model. Although we cannot conclude that the variables analyzed in this project solely affect the manufacturer's suggested retail price (MSRP), we are able to obtain a useful correlation between predictors and response variable.

Questions of Interest

Which model gives us the best prediction of MSRP? If I have a specific want or need in a vehicle, what price should I expect to pay? If I want to sell my 2002 SUV with an acceleration of 8 km per hour per second and 30 miles per gallon, what price should I expect to sell my car for?

Data

The hybrid dataset contains 153 observations of hybrid models, including data on eight potential predictor variables: Vehicle ID (1 to 154), Vehicle, Year, Acceleration Rate (accelrate), Miles Per Gallon (MPG), Max of MPG and MPG Equivalent (MPGMPGe), Model Class, and Class ID (numerical categories for Model Class).

Regression Methods

To produce a model which will best predict MSRP, we will apply statistical analysis techniques including, but not limited to, backward/forward selection, comparison of AIC/BIC, analysis of plots (Added Variable, QQ, Residuals vs Fitted, etc.) for non-transformed and transformed predictors and response variables. We will then answer our proposed question regarding the 2002 SUV by using our final regression model to calculate a predicted value and confidence interval using the given parameters and the predict() function.

Regression Analysis, Results and Interpretation

We began modeling for MSRP by using each of our predictor variables—we removed Car_id because it is an indexing variable, Vehicle because it is the name of the vehicle, MPGMPGe because it gives the same information as MPG and Carclass was replaced by its numerical counterpart Carclass_id— to build a simple linear model. We then created a similar model using the same variables, but added in two-term interaction terms to improve the model. An ANOVA table told us whether adding these two-term interactions improved the fit:

```
Analysis of Variance Table

Model 1: msrp ~ year + accelrate + mpg + carclass_id
Model 2: msrp ~ ((carid + vehicle + year + accelrate + mpg + mpgmpge +
    carclass + carclass_id) - carid - vehicle - mpgmpge - carclass)^2
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    148 3.2320e+10
2    142 2.2263e+10  6 1.0058e+10 10.692 8.656e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value calculated was $8.656*10^{-10}$. Given such a low value, we can conclude the interactions are significant, thus we continue with our second model. We repeat the process with a new model that includes three-term interaction terms compared to our two-term interaction model.

```
Analysis of Variance Table

Model 1: msrp ~ ((carid + vehicle + year + accelrate + mpg + mpgmpge +
    carclass + carclass_id) - carid - vehicle - mpgmpge - carclass)^2
Model 2: msrp ~ ((carid + vehicle + year + accelrate + mpg + mpgmpge +
    carclass + carclass_id) - carid - vehicle - mpgmpge - carclass)^3
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    142 2.2263e+10
2    138 2.2073e+10  4 189933013 0.2969 0.8796
```

However, this time our p-value is 0.8796, so we do not use the larger model. Moving on, we explored stepwise selection to calculate the best model, beginning with the null model and ending with our complete two-term model.

```
Start:  AIC=3052.47
msrp ~ 1

             Df  Sum of Sq        RSS    AIC
+ accelrate   1 3.3746e+10 3.6002e+10 2953.3
+ mpg         1 1.9727e+10 5.0020e+10 3003.6
+ year        1 3.0696e+09 6.6678e+10 3047.6
<none>                     6.9747e+10 3052.5
+ carclass_id 1 6.6732e+08 6.9080e+10 3053.0

Step:  AIC=2953.29
msrp ~ accelrate

             Df  Sum of Sq        RSS    AIC
+ mpg         1 3.0316e+09 3.2970e+10 2941.8
<none>                     3.6002e+10 2953.3
+ year        1 1.2960e+08 3.5872e+10 2954.7
+ carclass_id 1 4.5353e+05 3.6001e+10 2955.3
- accelrate   1 3.3746e+10 6.9747e+10 3052.5

Step:  AIC=2941.83
msrp ~ accelrate + mpg

               Df  Sum of Sq        RSS    AIC
+ accelrate:mpg  1 6.8552e+09 2.6115e+10 2908.2
<none>                       3.2970e+10 2941.8
+ carclass_id    1 3.8893e+08 3.2581e+10 2942.0
+ year           1 1.1167e+08 3.2858e+10 2943.3
- mpg            1 3.0316e+09 3.6002e+10 2953.3
- accelrate      1 1.7050e+10 5.0020e+10 3003.6
```

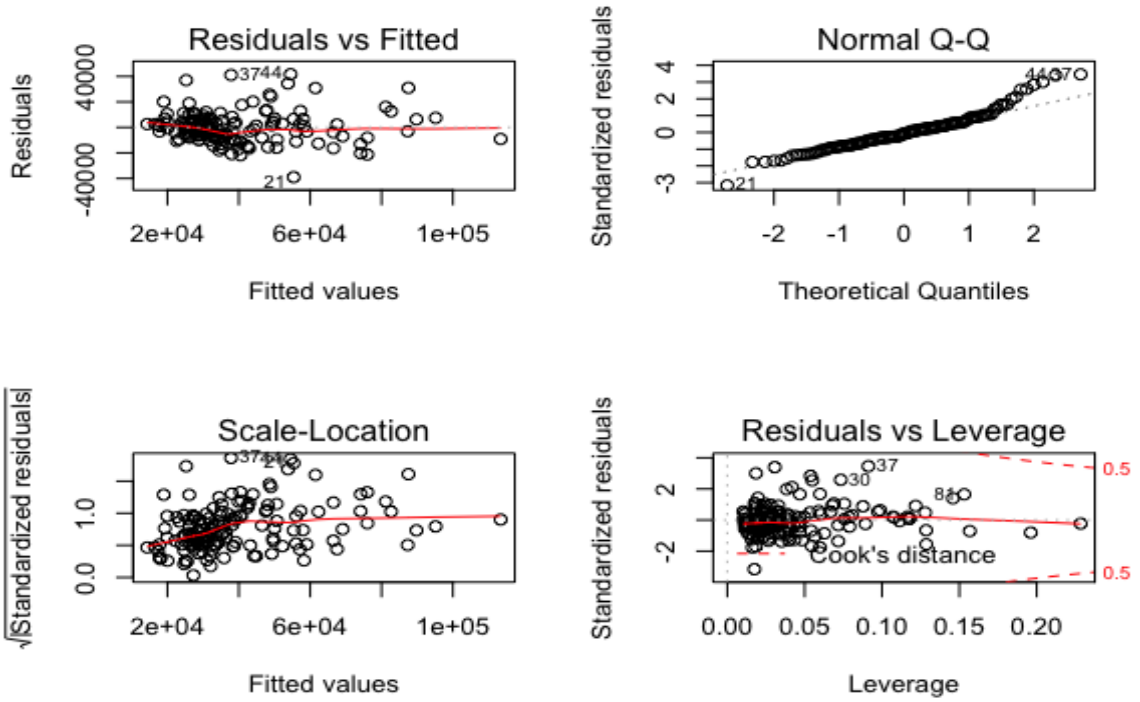
```
Step:  AIC=2908.17
msrp ~ accelrate + mpg + accelrate:mpg

               Df  Sum of Sq        RSS    AIC
+ carclass_id   1  506507024 2.5608e+10 2907.2
<none>                       2.6115e+10 2908.2
+ year          1   56013561 2.6059e+10 2909.8
- accelrate:mpg 1 6855211087 3.2970e+10 2941.8

Step:  AIC=2907.17
msrp ~ accelrate + mpg + carclass_id + accelrate:mpg

                        Df  Sum of Sq        RSS    AIC
+ accelrate:carclass_id  1 1395351546 2.4213e+10 2900.6
<none>                               2.5608e+10 2907.2
- carclass_id            1  506507024 2.6115e+10 2908.2
+ mpg:carclass_id        1   90415934 2.5518e+10 2908.6
+ year                   1    3294640 2.5605e+10 2909.2
- accelrate:mpg          1 6972790934 3.2581e+10 2942.0

Step:  AIC=2900.6
msrp ~ accelrate + mpg + carclass_id + accelrate:mpg + accelrate:carclass_id

                        Df  Sum of Sq        RSS    AIC
+ mpg:carclass_id        1 1561664374 2.2651e+10 2892.4
<none>                               2.4213e+10 2900.6
+ year                   1   24925461 2.4188e+10 2902.4
- accelrate:carclass_id  1 1395351546 2.5608e+10 2907.2
- accelrate:mpg          1 8128103584 3.2341e+10 2942.9
```

```
Step:  AIC=2892.4
msrp ~ accelrate + mpg + carclass_id + accelrate:mpg + accelrate:carclass_id +
    mpg:carclass_id

                        Df  Sum of Sq        RSS    AIC
<none>                                2.2651e+10 2892.4
+ year                   1    4542386 2.2647e+10 2894.4
- mpg:carclass_id        1 1561664374 2.4213e+10 2900.6
- accelrate:carclass_id  1 2866599985 2.5518e+10 2908.6
- accelrate:mpg          1 8938616955 3.1590e+10 2941.3
```

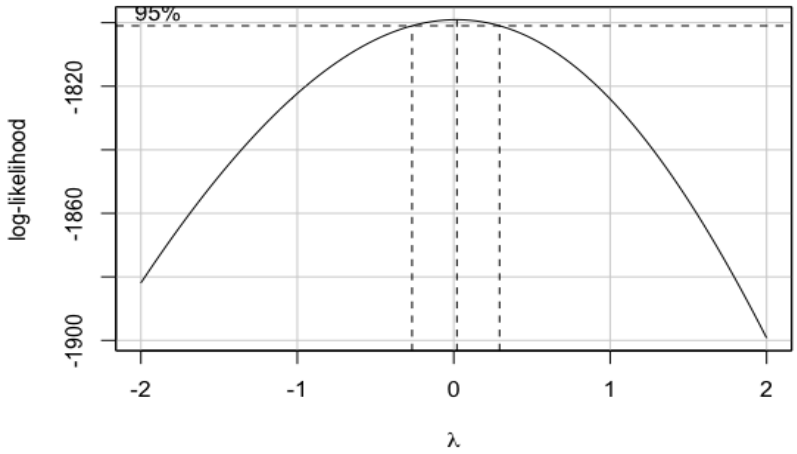
The stepwise selection returned the following model:

$$MSRP = accelrate + MPG + carclass_{id} + accelrate * MPG + accelrate * carclass_{id} + MPG * carclass_{id}$$

The AIC for this model is 2892.4 (compared to the original model's AIC of 3052.47. We then used similar techniques backward elimination and forward selection. Both resulted in the same model as the stepwise function.



To test our assumptions, we looked at the QQ, Residuals vs. Fitted, Scale-Location, and Residuals vs. Leverage plots for our model. After analyzing the outputs, we noticed there was a cause for concern in the QQ plot— the data might be right skewed, a violation of the assumption of normality. To fix this assumption, we used the Box-Cox method to find the best transformation.



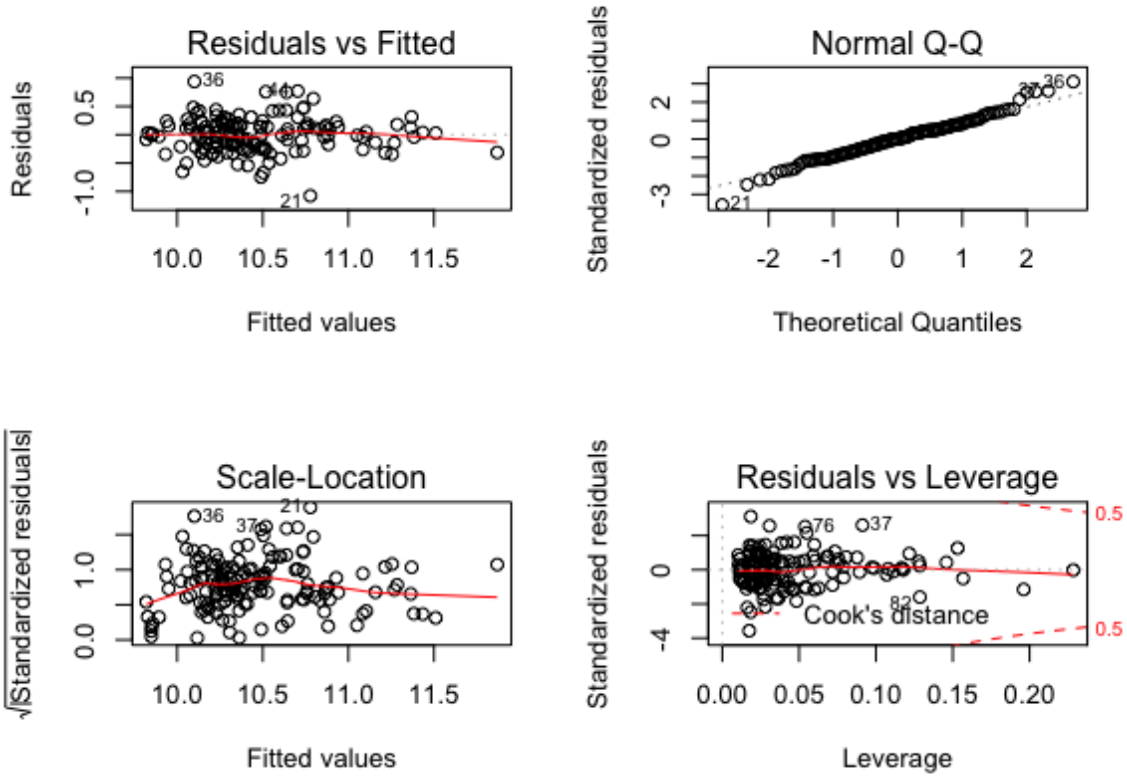
The Box-Cox function returned a lambda very close to zero, meaning that a log-transformation of the predictor would be beneficial for our model.

$$log(MSRP) = accelrate + MPG + carclass_{id} + accelerate * MPG + accelrate * carclass_id + MPG * carclass_{id}$$

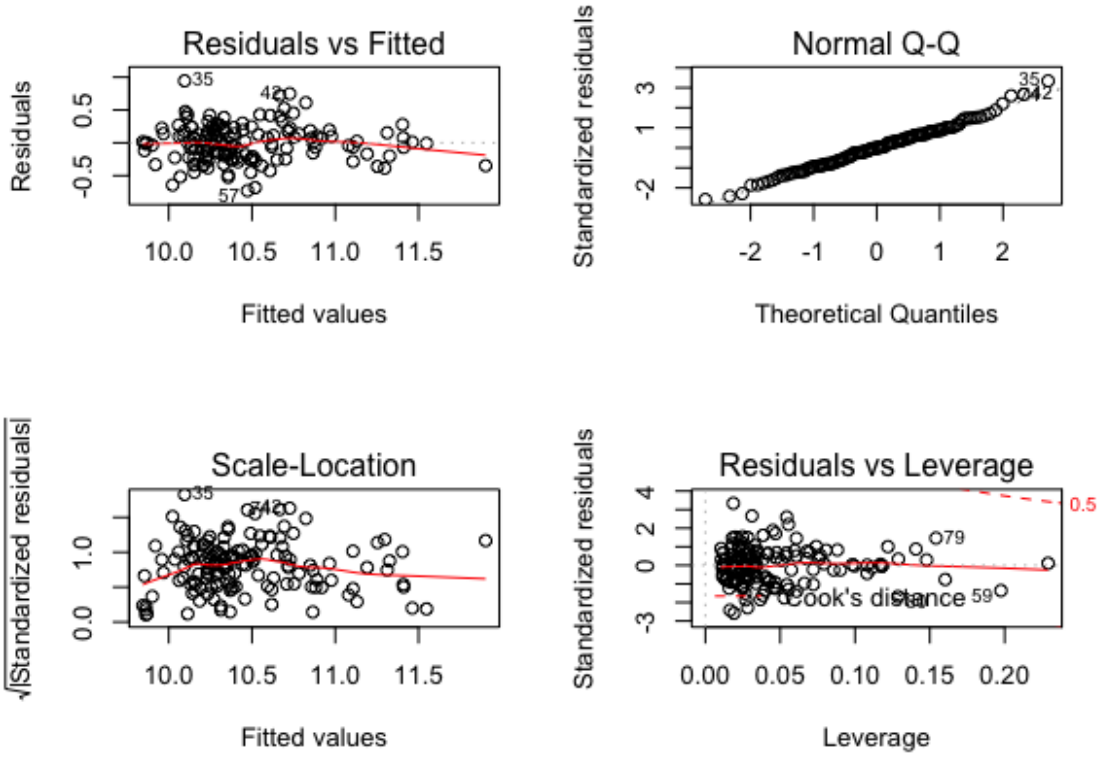
```
Call:
lm(formula = msrp ~ accelrate + mpg + carclass_id + accelrate:mpg +
    accelrate:carclass_id + mpg:carclass_id)

Residuals:
   Min     1Q Median     3Q    Max
-39043  -7680  -1121   5555  41701

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -162356.90   25366.42  -6.400 1.99e-09 ***
accelrate            17661.78    1785.67   9.891  < 2e-16 ***
mpg                   3590.69     542.78   6.615 6.56e-10 ***
carclass_id          18818.10    4676.09   4.024 9.14e-05 ***
accelrate:mpg         -325.51      42.88  -7.590 3.47e-12 ***
accelrate:carclass_id -1141.27     265.51  -4.298 3.13e-05 ***
mpg:carclass_id        -187.45      59.08  -3.173  0.00184 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12460 on 146 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.6619
F-statistic: 50.59 on 6 and 146 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log(msrp) ~ accelrate + mpg + carclass_id + accelrate:mpg +
    accelrate:carclass_id + mpg:carclass_id)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07646 -0.19859 -0.00024  0.16507  0.93883

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.253360   0.620131  10.084  < 2e-16 ***
accelrate             0.329087   0.043654   7.538 4.63e-12 ***
mpg                   0.067062   0.013269   5.054 1.28e-06 ***
carclass_id           0.499189   0.114316   4.367 2.38e-05 ***
accelrate:mpg        -0.005127   0.001048  -4.890 2.63e-06 ***
accelrate:carclass_id -0.024085   0.006491  -3.711 0.000293 ***
mpg:carclass_id       -0.006452   0.001444  -4.467 1.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3045 on 146 degrees of freedom
Multiple R-squared:  0.6215,    Adjusted R-squared:  0.6059
F-statistic: 39.95 on 6 and 146 DF,  p-value: < 2.2e-16
```

This newer model (with a log-transformed predictor) has a slightly lower R^2 value, 0.6215 compared to 0.6752, the R^2 value of the previous model. Despite its lower correlation coefficient, the log-transformed of the model helps adjust the normality of our distribution.



However, after taking another look at the diagnostic plots, observations 21 and 37 sway most of the assumptions. By removing these observations from our data, our R^2 increases to 0.6615 and our diagnostic plots hold normality assumptions. This signifies that 66.15% of the variation in MSRP can be explained by our model.



We conclude our final model:

$$log(MSRP) = accelrate + MPG + carclass_{id} + accelrate * MPG + accelrate * carclass_{id} + MPG * carclass_{id}$$

What does this mean? If I want to sell my 2002 SUV with an acceleration of 8 km per hour per second and 30 miles per gallon, our model predicts a price of $31,108.92, with a 95% confidence interval of $26,032.99 to $37,174.19. This confidence interval expands the lower and upper bounds of what you should pay for a vehicle of the described attributes—closer to $26,032.99 describing a much better deal for the buyer, and approaching $37,174.19 as a better deal for the seller.

Conclusions

Final Model:

$$log(MSRP) = (0.32909) * accelrate + (0.06706) * MPG + (0.49919) * carclass_{id} + (-0.00512) * accelrate * MPG$$
$$+(-0.02409) * accelrate * carclass_{id} + (0.00645) * MPG * carclass_{id}$$

Our final model contains Accelrate, MPG, Carclass_id in addition to the interaction between Accelrate and MPG, between Accelrate and Carclass_id, and between MPG and Carclass_id. From the dataset, these variables had the most significant impact on MSRP. With all other predictors remaining unchanged, one additional unit of Accelerate increased MSRP by 38.96%, one additional unit of MPG increased MSRP by 6.94%, and one additional unit of Carclass_id increased MSRP by 64.73%. Reflecting on our analysis and final model, we believe our results accurately interpret MSRP, and thus can effectively predict MSRP. However, there is no doubt there is room for improvement. Given only 153 observations in our data set, our model would benefit from a larger sample size. Additionally, the analysis could be improved by obtaining other potential predictor variables not recorded in the original data set. Variables such as horsepower or luxury features such as navigation systems, would only make the model stronger as they add variables that play a role in the pricing of vehicles.

Appendix 1: R code

```
library(MASS)
library(car)

#Read in the data
hybrid <- read.csv("C:/Users/Isabelle/Desktop/hybrid_reg.csv", header = TRUE)

#The Full Model with response MSRP
full.model = lm(msrp ~ year + accelrate + mpg + carclass_id, data = hybrid)
#We removed carid (indexing variable), vehicle (Character vehicle name), MPGMPGE (equal to MPG in almost every ca
se), and replaced carclass with class_id.

#fullmodel2 includes the same variables as full.model, but includes all second order interactions.
full.model2 = lm(msrp~(. - carid - vehicle - mpgmpge - carclass)^2, data = hybrid)
summary(full.model)
anova(full.model, full.model2)

#Based on the summary, we see the most significant predictor is accelrate due to its lowest p-value.
#We then use an anova table to test the null hypothesis that all of the interactions slopes are 0 (not useful to the model).
#Our p-value (8.656e-10) is significantly low, thus we reject our null hypothesis and conclude at least one interaction is i
mportant in modeling the data.

full.model3 = lm(msrp ~ (.- carid - vehicle - mpgmpge - carclass)^3, data = hybrid)
anova(full.model2, full.model3)
#We used an anova table to test the null hypothesis that all of the third order interactions slopes are equal to 0 (not useful
to the model).
#Our p-value (0.8796) indicates we fail to reject the null and conclude we will not include third order interactions in our
model.

fit.0 = lm(msrp ~ 1, data = hybrid)
fit.1 = step(fit.0, scope = list (upper = full.model2), direction = 'both')
#Using Step wise selection we are able to find the best model of msrp~accelrate+mpg+carclass_id+accelrate:mpg+accel
rate:carclass_id+mpg:carclass_id.
#The AIC for this model is 2894.37
#Full model of interactions+variables AIC = 3052.47

fit.2 = step(fit.0, scope = list( upper = full.model2), direction = 'forward')
#Using forward addition we find that the model is the same as found our step-wise selection.

fit.3 = step(full.model2, direction = 'backward')
#Backwards elimination gives us the same model as our previous techniques.
#Implies a strong model

summary(fit.1)
#All of our p-values are low-- indicating importance.
#R-squared = 0.6752

plot(fit.1)
#The Residuals vs. Fitted plot implies there may be something wrong with our assumptions; try transformations.
#There is also a problem with normality indicated by the right skew seen in the QQ plot.

accelrate.2 = 1/(hybrid$accelrate)
fit.4 = lm( msrp ~ accelrate.2 + mpg + carclass_id + accelrate.2:mpg + accelrate.2:carclass_id + mpg:carclass_id, data = h
ybrid)
```

```
summary(fit.4)
plot(fit.4)
#Attempted to transform accelrate: log as well as inverse.
#Neither resulted in a better r^2.
#However the residuals plot appeared slightly better.

#Attempting to transform the response variable. (
msrp.2 = log(hybrid$msrp)
fit.5 = lm( msrp.2 ~ accelrate + mpg + carclass_id + accelrate:mpg + accelrate:carclass_id + mpg:carclass_id, data = hybrid)
summary(fit.5)
plot(fit.5)

# Exploring Izzy's Final model
m3 = lm( msrp ~ accelrate + mpg + carclass_id + accelrate:mpg + accelrate:carclass_id + mpg:carclass_id, data = hybrid)
summary(m3)
par(mfrow=c(2,2))
#Checking Diagnostic Plots
plot(m3)
#To determine log-transformation
boxcox(m3)
#New model with log-transform
m4 = lm( log(msrp) ~ accelrate + mpg + carclass_id + accelrate:mpg + accelrate:carclass_id + mpg:carclass_id, data = hybrid)
summary(m4)
#Check Diagnostic Plots
plot(m4)

#Taking out observations 21 and 37
hybridnew2 <- hybrid[-21,]
hybridnew3 <- hybridnew2[-36,]
attach(hybridnew3)

#Using new Data set(
summary(m3)
plot(m3)
summary(m4)
plot(m4)
avPlots(m4)
#Confidence interval for 2002 SUV
newdata = data.frame( accelrate = 8, mpg = 30, carclass_id = 6)
predict(m4, newdata, interval="confidence")
```