535 HW 1

Gustavo Esparza

February 2019

1.)

What is the relationship between the predictors and the response? We will begin by fitting a linear model with all of our given predictors.

```
library(readr)
library(corrplot)
HW2 <- read_csv("Desktop/Math 535/4/HW2.csv")

SRP = HW2$SuggestedRetailPrice
EngSize =HW2$EngineSize
Cylinders = HW2$Cylinders
HP = HW2$Horsepower
HMPG = HW2$HighwayMPG
Weight = HW2$Weight
WB = HW2$WheelBase
Hybrid = HW2$Hybrid

model1 = lm(SRP~EngSize+Cylinders+HP+HMPG+Weight+WB+Hybrid)
summary(model1)
```

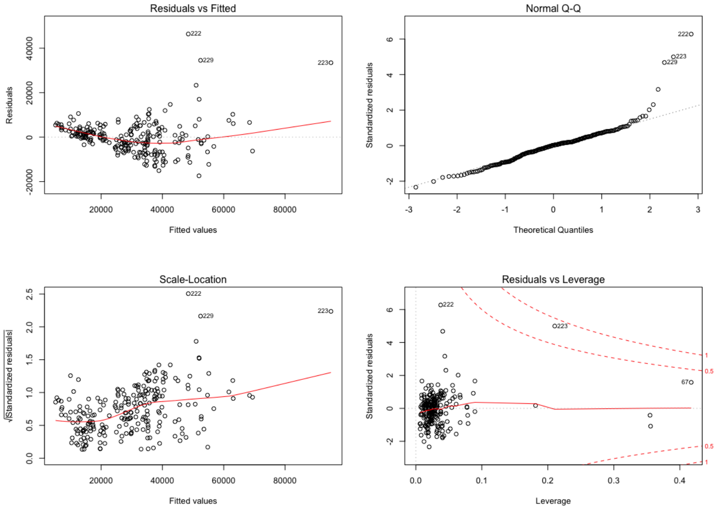
OUTPUT:

```
    Call:
lm(formula = SRP ~ EngSize + Cylinders + HP + HMPG + Weight +
    WB + Hybrid)

Residuals:
   Min     1Q Median     3Q    Max
-17436  -4134    173   3561  46392

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -68965.793  16180.381  -4.262 2.97e-05 ***
EngSize      -6957.457   1600.137  -4.348 2.08e-05 ***
Cylinders     3564.755    969.633   3.676 0.000296 ***
HP             179.702     16.411  10.950  < 2e-16 ***
HMPG           637.939    202.724   3.147 0.001873 **
Weight          11.911      2.658   4.481 1.18e-05 ***
WB              47.607    178.070   0.267 0.789444
Hybrid         431.759   6092.087   0.071 0.943562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 7533 on 226 degrees of freedom
Multiple R-squared:  0.7819,Adjusted R-squared:  0.7751
F-statistic: 115.7 on 7 and 226 DF,  p-value: < 2.2e-16
```

From this simple model, we can see that the p-value corresponding to the F-statistic is close to zero, which implies that our model is valid (ie there is a relationship between the predictors and the response). Now, we should investigate further and review our diagnostic plots.

```
    par(mfrow=c(2,2))
    plot(model1)
```



These plots imply many violations of the assumptions for a linear regression model.

From the residuals vs. fitted plot, we can see that the residuals do not seem to be distributed in a random fashion, which indicates that our model violates the assumption of linearity.

From the qq-plot, we can see that the tail ends deviate from our theoretical quantile line, which implies non-normality of our model.

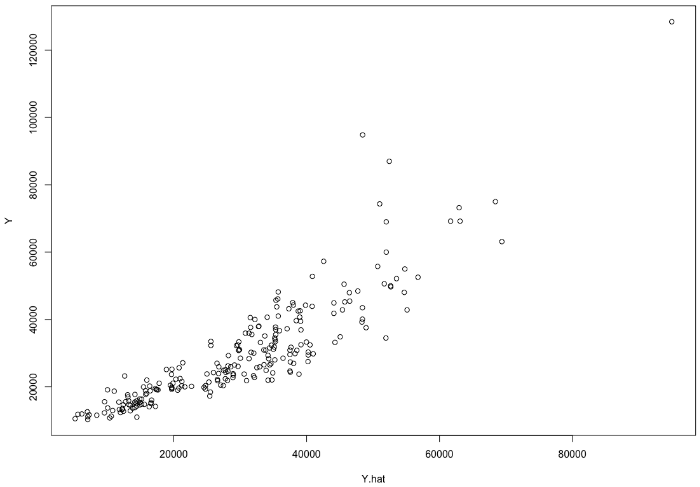
From the scale-location plot, we note that as the fitted values axis increases there is an increase in spread of our residuals, indicating non-constant variance.

Thus, there appears to be many issues with our model, despite our satisfactory p-value and R^2 value.

A transformation of the response may be necessary in order to resolve our issues regarding the assumptions for a valid model.

Taking a look at the plot of our predicted values versus our actual response may assist in deciding which transformation to perform.

```
X = data.frame(EngSize,Cylinders,HP,HMPG,Weight,WB,Hybrid)
Y = SRP
YX = data.frame(SRP,EngSize,Cylinders,HP,HMPG,Weight,WB,Hybrid)

Y.hat = lm(YX)$fit
plot(Y.hat,Y)
```



There is a very evident relationship between the fitted values from our simple model and the true values of our response variable. The pattern shown in the graph above implies that a log transform would be a viable option for resolving the violated assumptions and improving our model holistically.

We will make our desired transformation and run all appropriate diagnostics of our new model.

```
Y.trans=log(Y)
model2 = lm(Y.trans~EngSize+Cylinders+HP+HMPG+Weight+WB+Hybrid)
summary(model2)
```

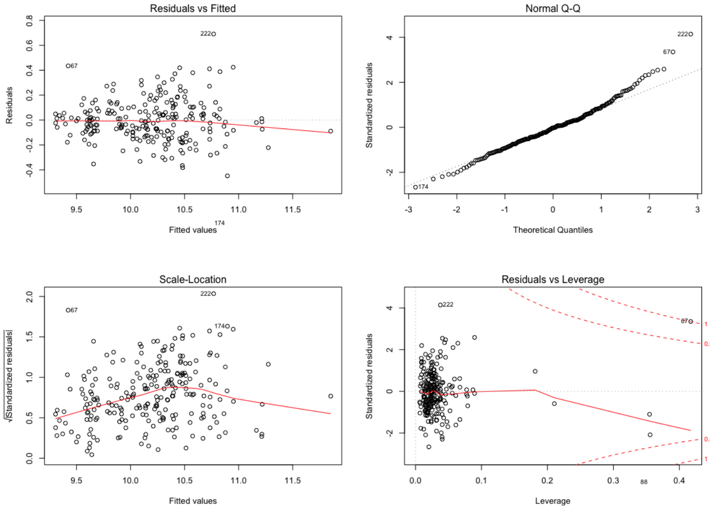
OUTPUT:

```
    Call:
lm(formula = Y.trans ~ EngSize + Cylinders + HP + HMPG + Weight +
    WB + Hybrid)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44808 -0.10082 -0.00456  0.09324  0.69074

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.4419088  0.3652375  20.376  < 2e-16 ***
EngSize     -0.1734769  0.0361197  -4.803 2.85e-06 ***
Cylinders    0.0315841  0.0218874   1.443   0.1504
HP           0.0044384  0.0003704  11.981  < 2e-16 ***
HMPG         0.0077142  0.0045761   1.686   0.0932 .
Weight       0.0006038  0.0000600  10.062  < 2e-16 ***
WB          -0.0004391  0.0040195  -0.109   0.9131
Hybrid       0.3250857  0.1375158   2.364   0.0189 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.17 on 226 degrees of freedom
Multiple R-squared:  0.8754,Adjusted R-squared:  0.8715
F-statistic: 226.8 on 7 and 226 DF,  p-value: < 2.2e-16
```

Our new model is still significant, according to our F-statistic and p-value. We can also note that a majority of our predictors also have a p-value that implies they are statistically significant to the model.

Now, let's check the assumptions with our residual plots.

```
par(mfrow=c(2,2))
plot(model2)
```



From our residual vs. fitted plot, we can clearly see that the violation of non-linearity has been resolved, as our residuals are now distributed in a random fashion.

Our qq-plot still has some deviation at the tail-ends, but it is much more aligned than it previously was in the first model.

The scale location plot now has a fairly constant spread across the entire x-axis, implying that the new model has constant variance.

So, our new model has essentially resolved all of our past issues of violated assumptions, while still retaining a statistically significant model that predicts our response variable.

Thus, the relationship between the predictors and the response can be explained by the following fit:

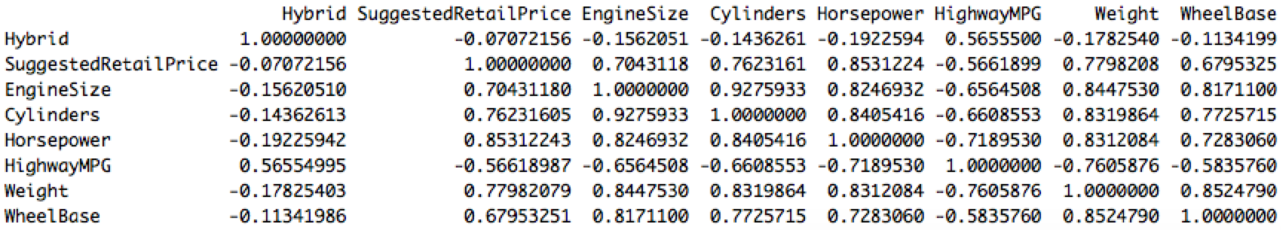
log(SRP) = 7.4419088 + -.1734769(Engine Size) + .0315841(Cylinders) + .0044384(HP) + .0077142(HMPG) + .0006038(Weight) + -.0004391 (WB) + .3250857(Hybrid)

2.) Are any of the predictors unimportant when it comes to predicting the response?

Earlier in our final prediction model, we stated that most of the predictors showed to have a small enough p-value to be considered statistically significant in the model. There were also the following predictors that showed to have larger p-value : Cylinders,WB and to a lesser extent HMPG.
The p-value does not tell the entire story, however. We will take a look at the correlation chart of our predictors to really see what is happening between our variables.

```
c=cor(HW2[, !(names(HW2)=="Vehicle Name")])
c
```

OUTPUT:



```
                          Hybrid SuggestedRetailPrice EngineSize  Cylinders Horsepower HighwayMPG     Weight  WheelBase
Hybrid                1.00000000          -0.07072156 -0.1562051 -0.1436261 -0.1922594  0.5655500 -0.1782540 -0.1134199
SuggestedRetailPrice -0.07072156           1.00000000  0.7043118  0.7623161  0.8531224 -0.5661899  0.7798208  0.6795325
EngineSize           -0.15620510           0.70431180  1.0000000  0.9275933  0.8246932 -0.6564508  0.8447530  0.8171100
Cylinders            -0.14362613           0.76231605  0.9275933  1.0000000  0.8405416 -0.6608553  0.8319864  0.7725715
Horsepower           -0.19225942           0.85312243  0.8246932  0.8405416  1.0000000 -0.7189530  0.8312084  0.7283060
HighwayMPG            0.56554995          -0.56618987 -0.6564508 -0.6608553 -0.7189530  1.0000000 -0.7605876 -0.5835760
Weight               -0.17825403           0.77982079  0.8447530  0.8319864  0.8312084 -0.7605876  1.0000000  0.8524790
WheelBase            -0.11341986           0.67953251  0.8171100  0.7725715  0.7283060 -0.5835760  0.8524790  1.0000000
```

Looking at the predictor variables that were shown to be possibly unimportant, we can see that they are in fact correlated well enough to still be considered for the final model. The reason why the fit shows that these predictors may not be useful is that they are also highly correlated with other variables in the model. This could indicate that more predictors are being used than necessary or that the high correlation is inconsequential if each predictor still remains highly correlated the response variable. Ultimately, more methodology is required to make this type of decision.

3.) A new vehicle is coming in that hasn't been priced for retail yet. The Vehicle is a 3.5 L engine size, has 6 Cylinders, 210 Horsepower, gets 29 HWY MPG, weighs 4210 lbs and has a wheelbase of 115. The vehicle is not a hybrid. Hank wants to know what the AVERAGE suggested retail would be for a vehicle like this.

```
predict(model2,newdata=data.frame(EngSize=3.5,Cylinders=6,HP=210,HMPG=29,
 Weight=4210,WB=115,Hybrid=0),interval="confidence")
```

OUTPUT:

```
     fit      lwr      upr
 10.67132 10.59057 10.75207
```

Using our prediction confidence interval with the provided information, we can see that the expected log(SRP) is equivalent to 10.67132. Thus, transforming our response back, we end up with the average suggested retail price for this scenario to be e10.67132 = 43101.80 dollars