Taylor Vandenberg

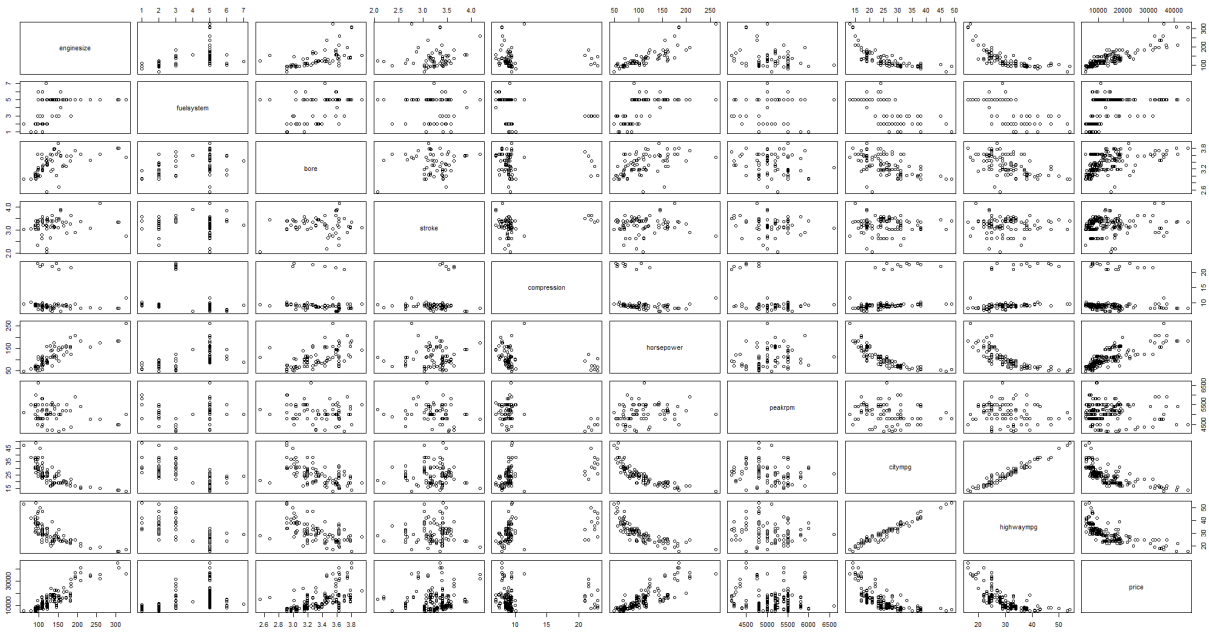
STA 2260

Kevin Bailey

5/21/2021

Extra Credit Project

The dataset I chose to work with was the cleaned UCR car file. After importing the csv file, I used the plot command in order to search for some possible relationships between different measurements in the file which could be used to predict price. When answering this type of question with data, the price is the result and things such as the weight of the car or horsepower are the predictors. I reduced the number of graphs shown as many of them were irrelevant for the purpose of this project and would make the picture below more difficult to read. To plot all graphs I would remove the range I added to the plot function.



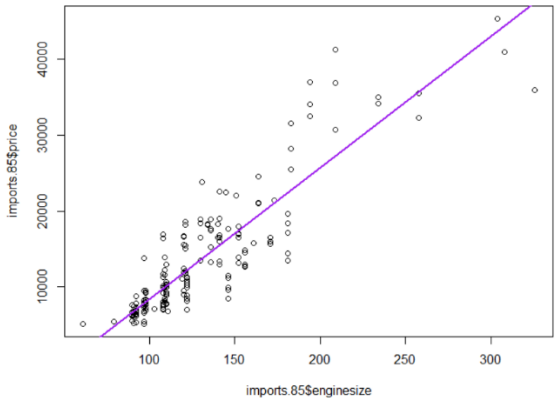
Based on these graphs, I predicted that the best one-predictor models to test for would be the engine size, highway mpg, horsepower, and bore. I then tested for adjusted r-squared scores by creating linear models, using the summary function, and found the following data.

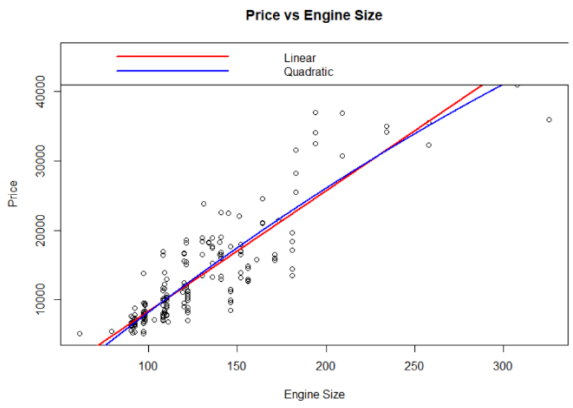
Engine Size	Highway MPG	Horsepower	Bore
.7888	.5147	.6583	.2948

The value of r-squared represents how close the data is to the line of best fit. In other words, the higher this score is, the more likely that the predictor and result are related to each other. The lower the score, the more likely it is that there is no correlation. Based on this, the engine size would be the best predictor from this group. To further verify this, it is also possible to compare the AIC values of the linear models. In this case, a lower value implies a greater correlation and a higher value means that there is no correlation. The AIC values for the linear models are found below.

Engine Size	Highway MPG	Horsepower	Bore
3724.902	3885.500	3817.784	3957.632

Based on these results, it can be concluded that the engine size of the car is the best predictor to find the price. I then plotted the data using the engine size on the X axis and the price on the Y axis and included the fitted line shown in purple by using the abline function. Model 1 (m1) is shown below. The model could also be made more accurate by using engine size squared and creating xhat and yhat to use for plotting the data. This also allows for the creation of a quadratic line. The second version of Model 1 is shown on the right and has an adjusted r-squared value of 0.7909.





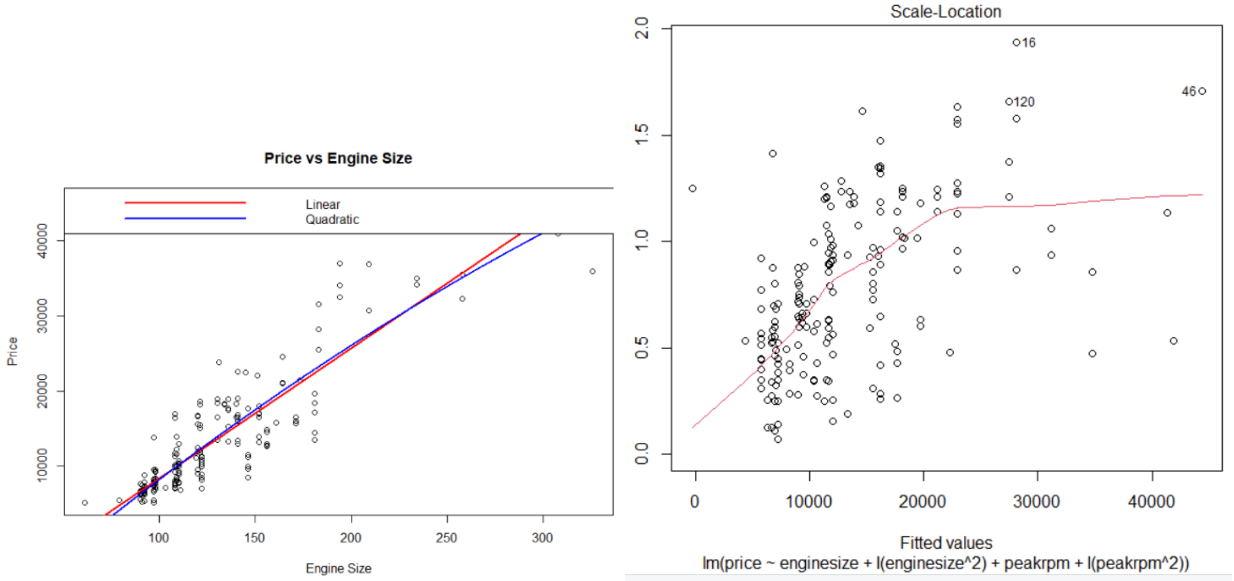
To decide how to create model 2, I made two new models. The first was created using another linear model using engine size and peak rpm as predictors with price as the result, and the second was created by multiplying the engine size and horsepower when creating the linear

model. Calculating the value of r-squared and AIC of the engine size – peak rpm model showed that the respective values were 0.8037 and 3713.728. The values for the engine size – horsepower model were 0.8015 and 3714.907. Both of these are extremely similar, but the engine size – peak rpm model is slightly more accurate as shown by the higher r-squared value and lower AIC value. Therefore, it will be used as Model 2.

Comparing Model 1 to Model 2 can be simplified using the following chart.

Comparison	R-Squared	AIC
Model 1	0.7909	3724.902
Model 2	0.8037	3713.728

Model 2 is more accurate than Model 1 as shown by the r-squared and AIC values because it includes more data in it, and as a result will ultimately be more accurate in predicting future data. The formulas for Model 1 (left) and Model 2 (right) are shown below as well as the graphs side by side.



```
Call:
lm(formula = price ~ enginesize + I(enginesize^2), data = imports.85)

Residuals:
    Min      1Q  Median      3Q     Max
-9402.6 -1980.5   -49.2  1462.8 13778.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.251e+04  2.312e+03  -5.410 1.88e-07 ***
enginesize       2.217e+02  2.943e+01   7.532 1.97e-12 ***
I(enginesize^2) -1.439e-01  8.469e-02  -1.700   0.0908 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3699 on 190 degrees of freedom
Multiple R-squared:  0.7931,    Adjusted R-squared:  0.7909
F-statistic: 364.1 on 2 and 190 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = price ~ enginesize + I(enginesize^2) + peakrpm +
    I(peakrpm^2), data = imports.85)

Residuals:
    Min      1Q  Median      3Q     Max
-9426.0 -1829.0  -261.6  1493.1 13201.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.555e+04  2.124e+04   1.203   0.2306
enginesize       2.302e+02  2.877e+01   8.002 1.23e-13 ***
I(enginesize^2) -1.582e-01  8.236e-02  -1.921   0.0562 .
peakrpm         -1.703e+01  8.196e+00  -2.078   0.0391 *
I(peakrpm^2)     1.829e-03  7.975e-04   2.293   0.0230 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3584 on 188 degrees of freedom
Multiple R-squared:  0.8078,    Adjusted R-squared:  0.8037
F-statistic: 197.6 on 4 and 188 DF,  p-value: < 2.2e-16
```