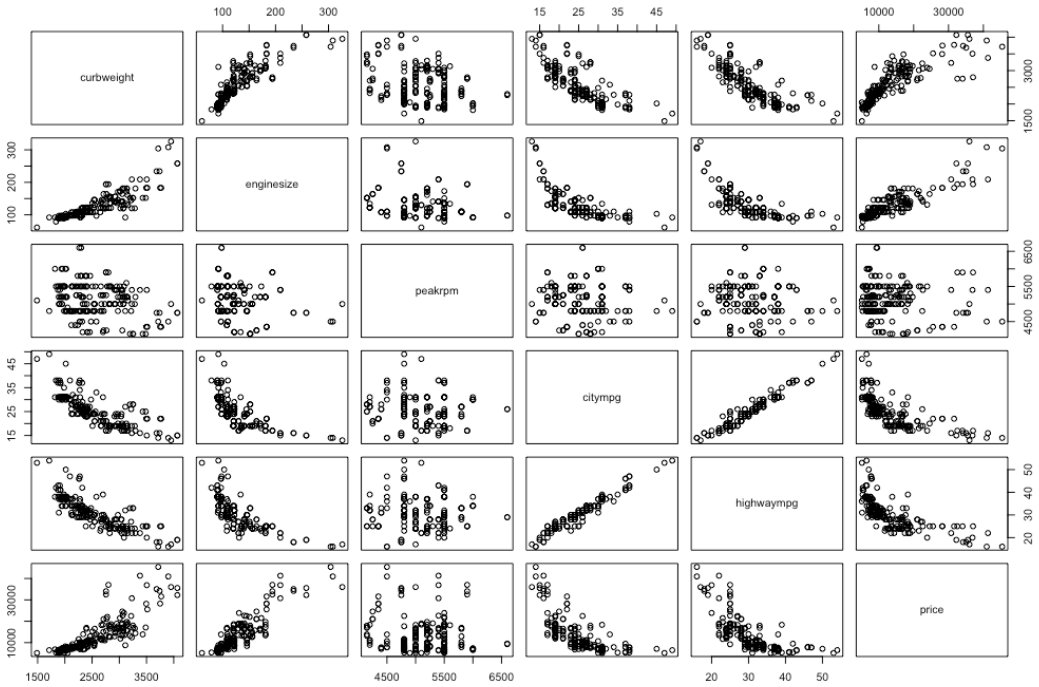
Apply Linear Regression for Predict Models

Paul (Dai) Vuong

Cal Poly Pomona

Apply Linear Regression for Predict Models

I chose Automobile Data Set from https://archive.ics.uci.edu/ml/datasets/Automobile. The data set was collected from many auto company (Audi, BMW, Chevrolet, Dodge, Honda, ISUZU, Jaguar, Mazda, …) However, the data was modified a bit from the UCI website, so there were 25 columns and 193 rows. I picked out some variables from data set for my project, include price, curb-weight, engine-size, peak-rpm, city-mpg, highway-mpg.



Model 1 – *one-predictor model*

I choose price as the dependent variable (the response) and curb-weight, engine-size, city-mpg, highway-mpg as the independent variables (the predictors). I apply the simple linear regression model for each independent variable:

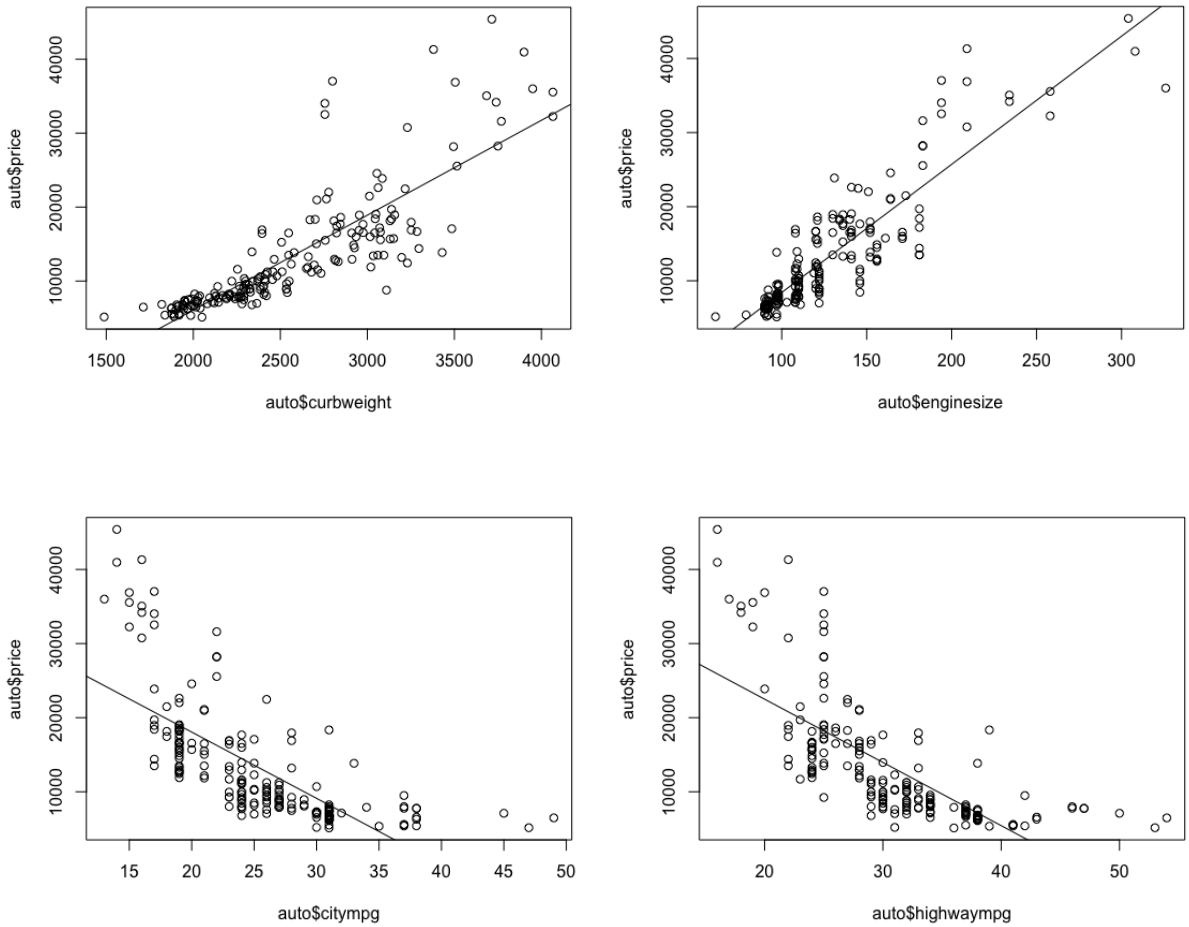
$$y = \widehat{B_0} + \widehat{B_1}x$$

And then, I get R^2 and AIC as this table:

Independent variables	R^2	AIC
curb-weight	0.6963	3795.058
engine-size	0.7888	3724.902
city-mpg	0.4967	3892.529
highway-mpg	0.5147	3885.500

We can easily see that engine-size has the lowest AIC and the highest adjusted R^2. Therefore, engine-size is the best one among those independent variables.

We can look at the scatter plots of each independent variable to the dependent variable.



The engine-size scatter plot looks more linear than the other three.
Because the engine-size is the best model so far, I use it for my one-predictor model.
Now, I try engine-size with quadratic and inverse
Quadratic formula:

$$y = \widehat{B_0} + \widehat{B_1}x + \widehat{B_2}x^2$$
$$price = \widehat{B_0} + \widehat{B_1}enginesize + \widehat{B_2}enginesize^2$$

Inverse formula:

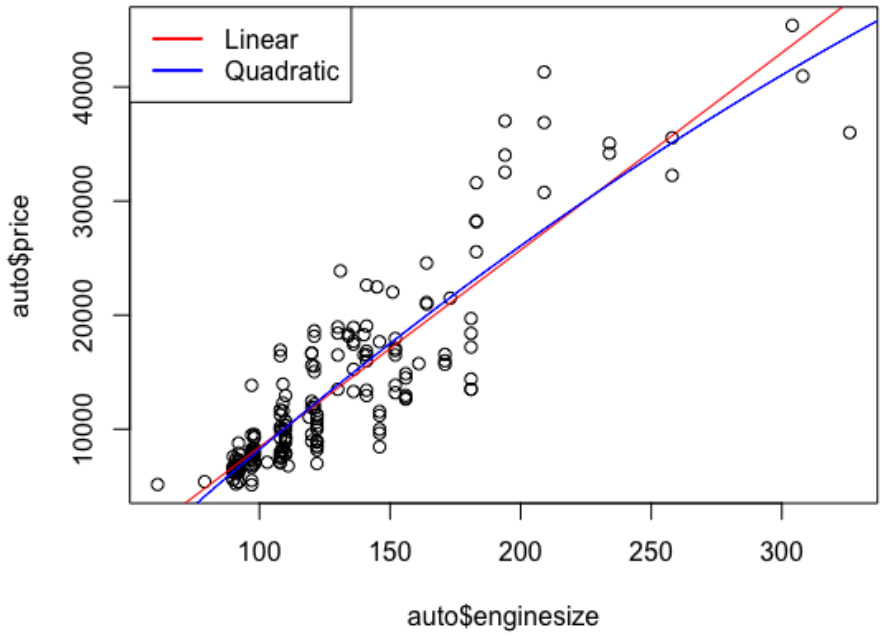
$$y = \widehat{B_0} + \widehat{B_1}x + \widehat{B_2}\frac{1}{x}$$
$$price = \widehat{B_0} + \widehat{B_1}enginesize + \widehat{B_2}\frac{1}{enginesize}$$

We get the R^2 and AIC as this table:

Independent variables	R^2	AIC
engine-size	0.7888	3724.902
engine-size quadratic	0.7909	3723.989
engine-size inverse	0.7880	3726.603

We can see that the quadratic has the lowest AIC and the highest adjusted R^2, so it is the best model for our one-predictor model, and it is the model for model 1.

Now I plot the linear line and quadratic line for our engine-size model. The two lines are slightly different.



Model 2 - *multiple-predictor model*

In this part, I try four models:

1. Engine-size and curb-weight

```
m2.engine.curb = lm(price ~ enginesize + curbweight, data=auto)
```

2. Engine-size and curb-weight square

```
m2.engine.curb.sq = lm(price ~ enginesize + I(enginesize^2) +
curbweight + I(curbweight^2), data=auto)
```

3. Engine-size and curb-weight inverse

```
m2.engine.curb.inv = lm(price ~ enginesize + I(1/enginesize) +
curbweight + I(1/curbweight), data=auto)
```

4. Engine-size, curb-weight, city-mpg and highway-mpg

```
m2.engine.curb.city.hwy = lm(price ~ enginesize + curbweight + citympg
+ highwaympg, data=auto)
```

We have the table of R^2 and AIC:

Independent variables	R^2	AIC
engine-size	0.7888	3724.902
engine-size quadratic	0.7909	3723.989
engine-size curb-weight	0.8083	3707.209
engine-size curb-weight quadratic	0.8073	3710.229
engine-size curb-weight inverse	0.8105	3706.990
engine-size curb-weight city-mpg hwy-mpg	0.8096	3707.863

According to the calculation result from the table, the engine-size and curb-weight inverse has highest AIC and lowest R^2. Therefore, it is the best model here, and it is the model for model 2.

Now, we compare the model 1 and model 2:

Independent variables	R^2	AIC
Model 1	0.7909	3723.989
Model 2	0.8105	3706.990

As the result, model 2 explains more variability in this data set (81.05% vs 79.09%), and also has better predictive power on the basis of AIC.