THE AMERICAN
UNIVERSITY IN CAIRO

Department of Mathematics and Actuarial Science

MACT 4231: Applied Regression Methods

Professor Ali S. Hadi

Final Project

Cars Price Prediction using Multiple Linear Regression

Farida Simaika - 900201753

Katia Gabriel - 900202272

Acknowledgement

We would like to thank and express our sincere gratitude for our professor Dr. Ali Hadi for the constant encouragement and support throughout this course. We would also like to thank him for enriching our knowledge up until the completion of this final project.

Table of Contents:

Introduction

Cars have a wide variety of brands, models, prices, features and types. One might think that the price of a car depends on all of the aforementioned attributes and much more. The price of a car also gives an indication about its capabilities and power.

Accordingly, the goal of this project is to try and determine the most influential attributes that affect the price of a car and create a linear regression model based on the least squares estimation method to predict the price of any given car. This information can help both the buyer and the manufacturer to determine the most adequate price for a given car and gain a sense of the car's capabilities based on its price.

Data Description

The dataset contains 26 variables : 1 response variable and 25 predictor variables and encompasses 205 observations. A linear regression model will be developed to determine the price (response variable) of a car based on the following predictor variables:

1. car_ID: unique car observation number
2. symboling: assigned car insurance risk rating with 3 being the most risky and -3 the safest.
3. carName: categorical variable indicating the brand of the car
4. fueltype: categorical variable revealing the type of fuel used by car (gasoline or diesel)
5. aspiration: categorical variable indicating aspiration type of the car engine (standard (std) or turbo)
6. doornumber: number of doors in the car (categorical)
7. carbody: categorical variable indicating the type of car frame (sedan, wagon, hatchback and convertible)
8. drivewheel: categorical variable that explains type of drive wheel (all wheel drive (awd), front wheel drive (fwd) and four wheel drive (4wd).
9. enginelocation: categorical variable specifying the location of the car's engine (front).
10. wheelbase: quantitative variable measuring the wheelbase of a car in inches(distance between center of front wheel and center of rear wheels)
11. carlength: quantitative variable that gives the length of a car in inches
12. carwidth: quantitative variable measuring width of a car in inches
13. carheight: a quantitative variable, measuring the height of each car in inches.
14. curb weight: a quantitative variable, measuring the weight of each vehicle with a full tank

and all available equipment in Ib.

15. enginetype: a categorical variable that differentiates between two different engine types: Overhead Camshaft Engine (ohc) and Overhead Camshaft and Valve F engine (ohcf).
16. cylindernumber: a quantitative variable that gives the number of cylinders inside the engine.
17. Enginesize: a quantitative variable that measures the engine size in terms of the amount of fuel and air that can be pushed through the cylinders in the engine in cc.
18. Fuelsystem: a categorical variable that differentiates between 8 different systems that deliver the fuel from the tank to the car engine. The categories are (1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi).
19. Boreratio: a quantitative variable, that measure the ratio between the engine cylinder bore diameter and the stroke length.
20. Stroke: a quantitative variable that measures the distance that the piston travels during one cycle.
21. Compressionratio: a quantitative variable that measures the compression ratio inside the cylinder of the engines of each vehicle.
22. Horsepower: a quantitative variable that measures the power that an engine is producing and is measured in hp.
23. Peakrpm: a quantitative variable that measures the revolutions per minute of an engine at its highest horsepower and is measured in RPM.
24. Citympg: a quantitative variable that measures the mpg of each vehicle while driving in city conditions and is measured in mpg.

25. Highwaympg: a quantitative variable that measures the mpg of each vehicle while driving on a highway and is measured in mpg.
26. Price: quantitative response variable indicating the price of the car

Source of the dataset

The dataset can be found here: https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

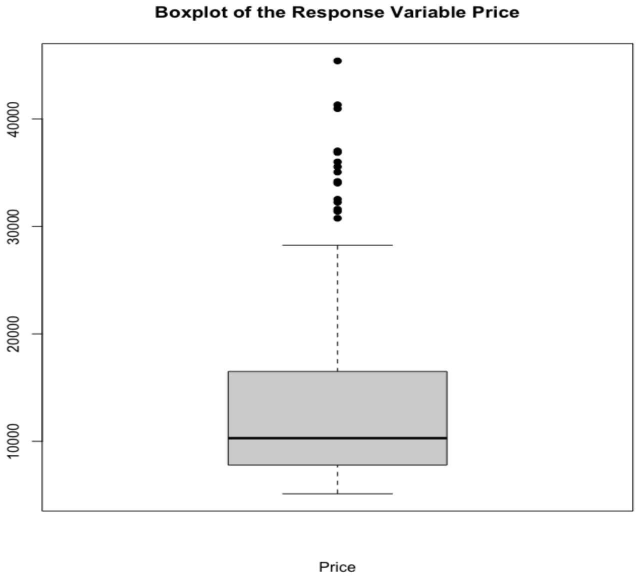
Multiple Linear Regression Procedure

Data preparation

As a first step before fitting any regression model, it is necessary to have an overall look at the dataset to determine whether the data is suitable for regression or not. It is evident that the data contains a number of categorical variables that have to be represented by indicator variables to be able to proceed with the regression. The categorical variables in the dataset are: car name, fuel type, aspiration, car body, drive wheel, engine location and fuel system. For each categorical variable new columns have been defined and the categories were converted into binary variables.

Graphs before fitting

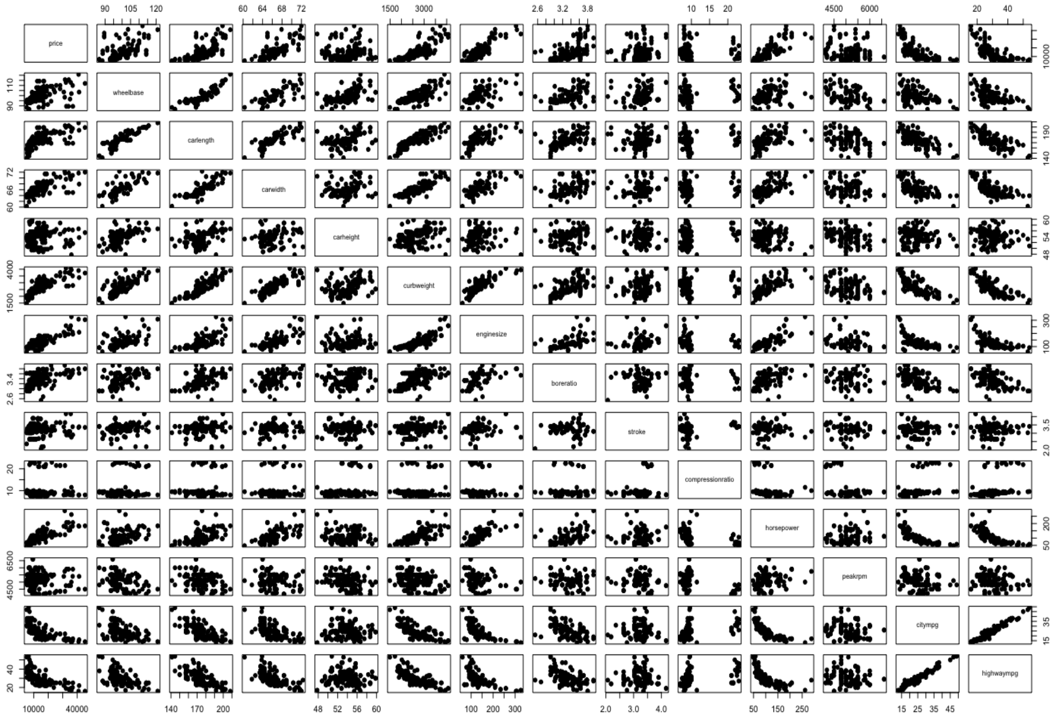
Examining the response variable



From the above graph, it can be seen that the response variable contains a number of outliers that raise a red flag and have to be further investigated. Moreover, the distribution of the response variable is right skewed with a median of 1000.

Examining the predictor variables:

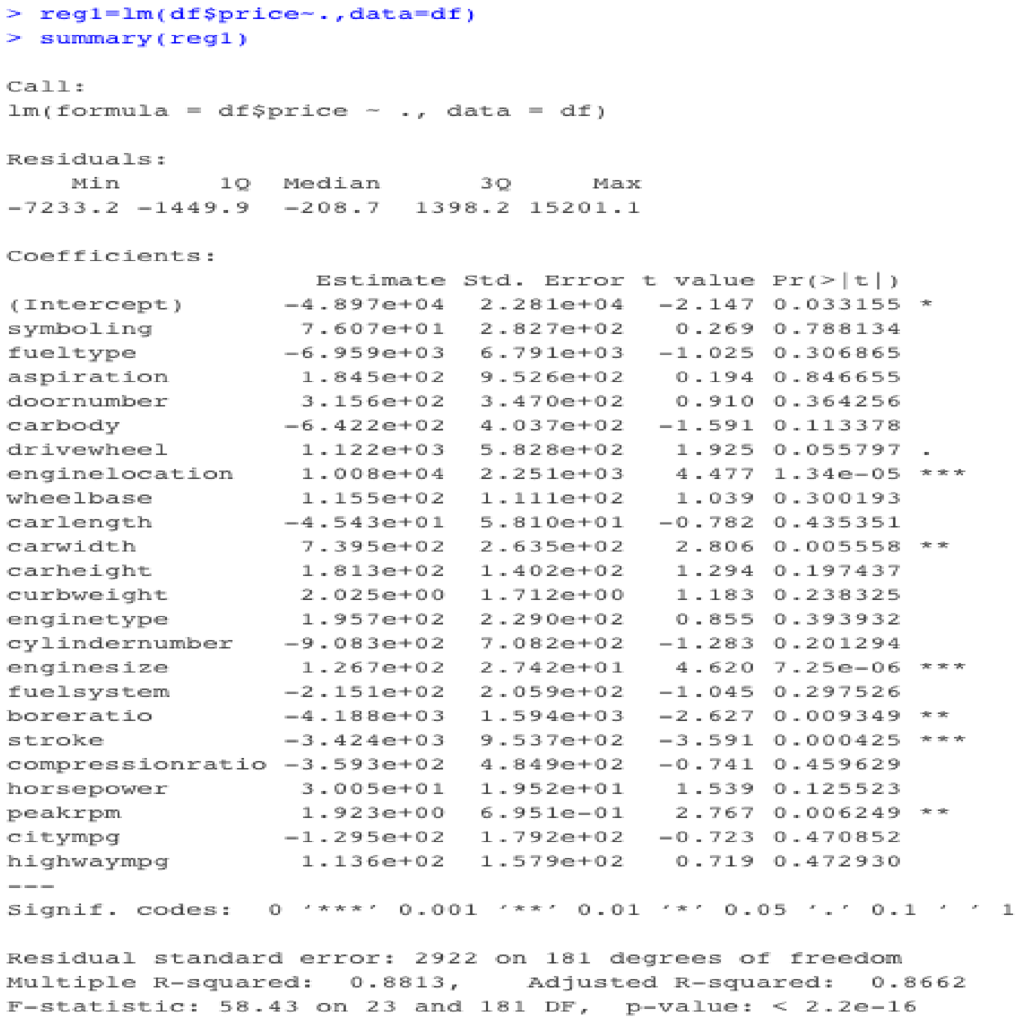
Since the data contains around 26 variables that have increased due to the indicator variable coding, it is nearly impossible to plot the pairs of all the predictors against each other. However, to be able to get a sense of the relationship between the predictor variables, a subset of the data with all the numerical variables has been created and the scatter plot matrix has been graphed as seen below.



It is evident that some of the predictors have a strong linear relationship with each other. For example the variable city mpg and highway mpg are almost perfectly correlated. Another example would be the variable curb weight and engine size. This suggests a multicollinearity problem among the predictors which violates the independence of the predictor variables assumption.

Iteration 1 :

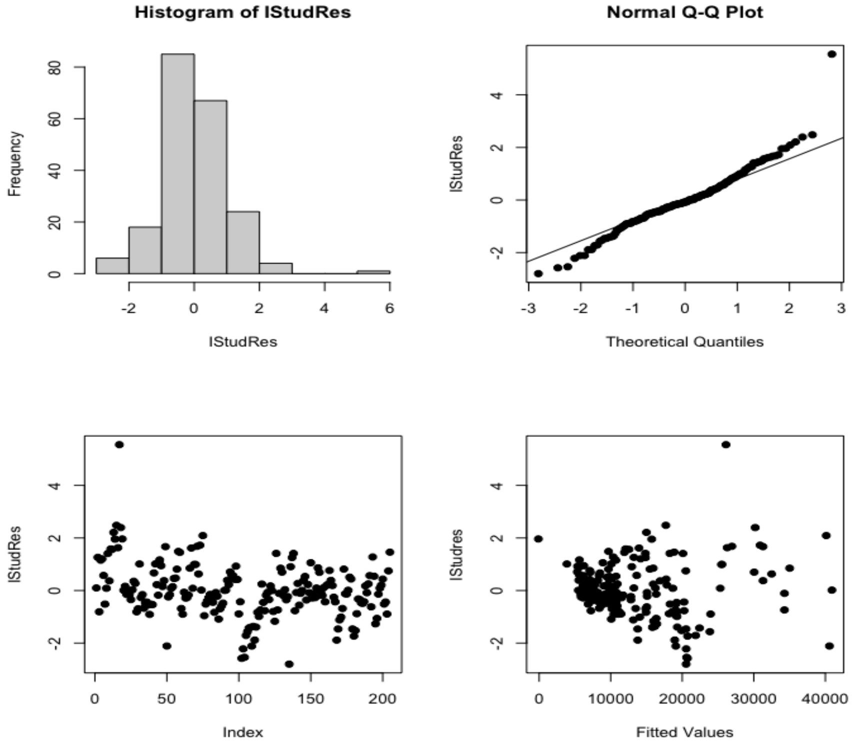
The response variable price was regressed on all the predictor variables after indicator variable coding: The result of the regression model are as follows:



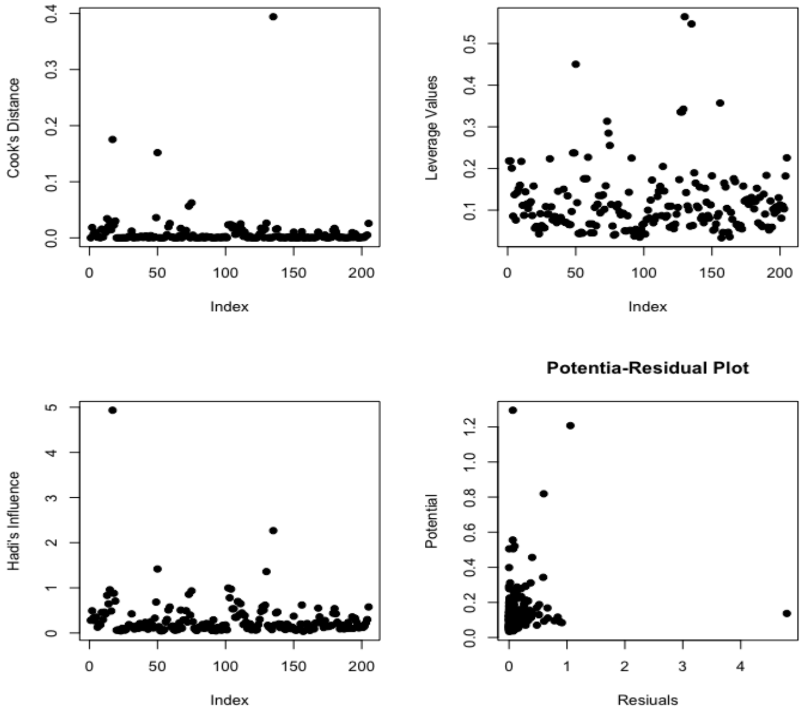
```
> reg1=lm(df$price~.,data=df)
> summary(reg1)

Call:
lm(formula = df$price ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-7233.2 -1449.9  -208.7  1398.2 15201.1 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -4.897e+04  2.281e+04  -2.147 0.033155 *  
symboling         7.607e+01  2.827e+02   0.269 0.788134    
fueltype         -6.959e+03  6.791e+03  -1.025 0.306865    
aspiration        1.845e+02  9.526e+02   0.194 0.846655    
doornumber        3.156e+02  3.470e+02   0.910 0.364256    
carbody          -6.422e+02  4.037e+02  -1.591 0.113378    
drivewheel        1.122e+03  5.828e+02   1.925 0.055797 .  
enginelocation    1.008e+04  2.251e+03   4.477 1.34e-05 ***
wheelbase         1.155e+02  1.111e+02   1.039 0.300193    
carlength        -4.543e+01  5.810e+01  -0.782 0.435351    
carwidth          7.395e+02  2.635e+02   2.806 0.005558 ** 
carheight         1.813e+02  1.402e+02   1.294 0.197437    
curbweight        2.025e+00  1.712e+00   1.183 0.238325    
enginetype        1.957e+02  2.290e+02   0.855 0.393932    
cylindernumber   -9.083e+02  7.082e+02  -1.283 0.201294    
enginesize        1.267e+02  2.742e+01   4.620 7.25e-06 ***
fuelsystem       -2.151e+02  2.059e+02  -1.045 0.297526    
boreratio        -4.188e+03  1.594e+03  -2.627 0.009349 ** 
stroke           -3.424e+03  9.537e+02  -3.591 0.000425 ***
compressionratio -3.593e+02  4.849e+02  -0.741 0.459629    
horsepower        3.005e+01  1.952e+01   1.539 0.125523    
peakrpm           1.923e+00  6.951e-01   2.767 0.006249 ** 
citympg          -1.295e+02  1.792e+02  -0.723 0.470852    
highwaympg        1.136e+02  1.579e+02   0.719 0.472930    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2922 on 181 degrees of freedom
Multiple R-squared:  0.8813,	Adjusted R-squared:  0.8662 
F-statistic: 58.43 on 23 and 181 DF,  p-value: < 2.2e-16
```

The preliminary results show that 88.13% of the variability in price can be explained by the predictor variables and the F-statistic is 58.43. However we observe that most of the variables are insignificant. No statistical inference can be concluded by this model as long as the 10 assumptions are not validated. Accordingly, as a next step the assumptions will be validated.

Graphs after fitting:



From the above graphs it can be seen that the residuals are not normally distributed as the Normal Q-Q Plot shows that some of the residuals deviate from the expected nature of the straight line. The normality assumption of the residuals is violated. Moreover, a discernable pattern and clustering can be seen in both the index plot of the standardized residuals and the plot of the standardized residuals versus the fitted values, which shows the need for possible transformation. The homoscedasticity and normality and independence of residuals assumptions are violated.



From the above graphs it is clear that the implicit assumptions are violated. The dataset contains both high leverage and influential points that might affect the regression outcomes and need further investigation.

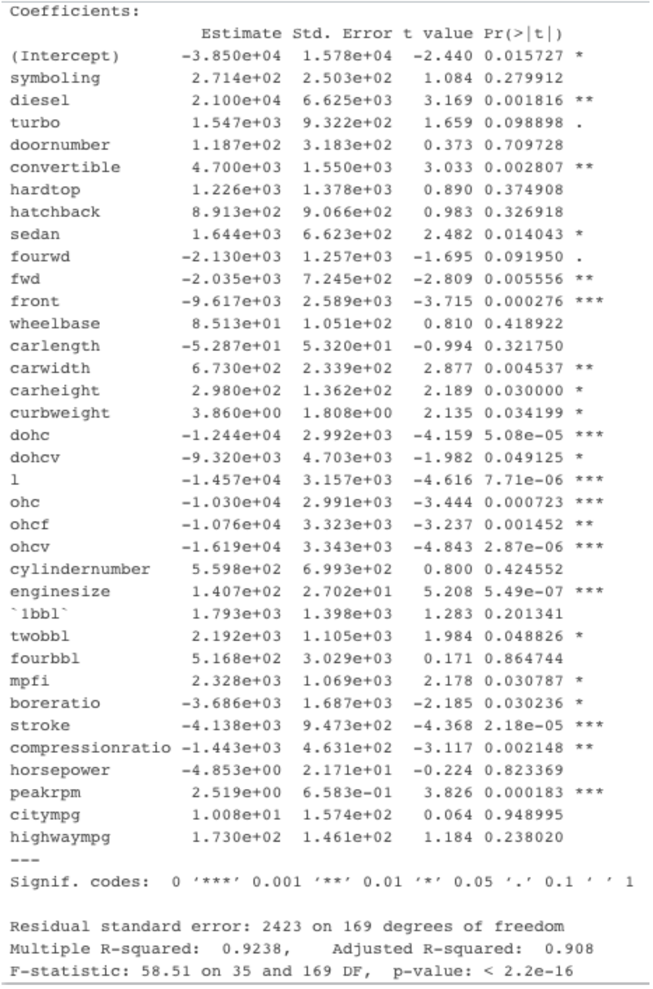
Steps taken to satisfy the assumptions:

1. Coding indicator variables into binary variables
2. Interaction Variables
3. Transformation of the variables
4. Multicollinearity Problem
5. Detection of outliers

1. **Coding indicator variables into binary variables.**

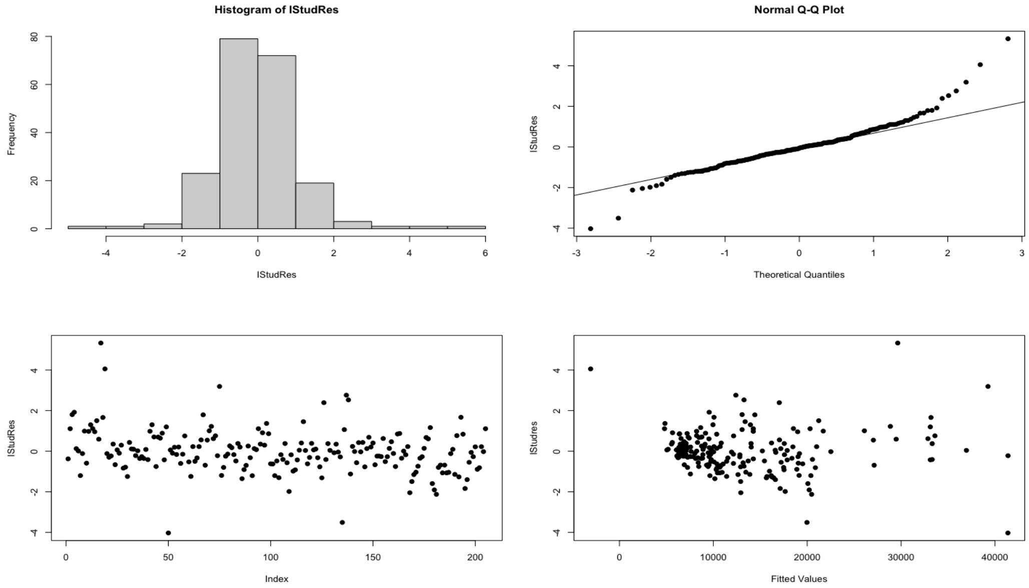
Since both the plot of the studentized residuals versus fitted values and the index plot of standardized residuals show a pattern. The categorical variables were converted into binary variables. One variable in each category was dropped to avoid collinearity. The dropped variables are the base and the regression coefficients will be interpreted with respect to the base variable. After implementing the above suggestion a second model was fitted.

Iteration 2:



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.850e+04  1.578e+04  -2.440 0.015727 *
symboling          2.714e+02  2.503e+02   1.084 0.279912
diesel             2.100e+04  6.625e+03   3.169 0.001816 **
turbo              1.547e+03  9.322e+02   1.659 0.098898 .
doornumber         1.187e+02  3.183e+02   0.373 0.709728
convertible        4.700e+03  1.550e+03   3.033 0.002807 **
hardtop            1.226e+03  1.378e+03   0.890 0.374908
hatchback          8.913e+02  9.066e+02   0.983 0.326918
sedan              1.644e+03  6.623e+02   2.482 0.014043 *
fourwd            -2.130e+03  1.257e+03  -1.695 0.091950 .
fwd               -2.035e+03  7.245e+02  -2.809 0.005556 **
front             -9.617e+03  2.589e+03  -3.715 0.000276 ***
wheelbase          8.513e+01  1.051e+02   0.810 0.418922
carlength         -5.287e+01  5.320e+01  -0.994 0.321750
carwidth           6.730e+02  2.339e+02   2.877 0.004537 **
carheight          2.980e+02  1.362e+02   2.189 0.030000 *
curbweight         3.860e+00  1.808e+00   2.135 0.034199 *
dohc              -1.244e+04  2.992e+03  -4.159 5.08e-05 ***
dohcv             -9.320e+03  4.703e+03  -1.982 0.049125 *
l                 -1.457e+04  3.157e+03  -4.616 7.71e-06 ***
ohc               -1.030e+04  2.991e+03  -3.444 0.000723 ***
ohcf              -1.076e+04  3.323e+03  -3.237 0.001452 **
ohcv              -1.619e+04  3.343e+03  -4.843 2.87e-06 ***
cylindernumber     5.598e+02  6.993e+02   0.800 0.424552
enginesize         1.407e+02  2.702e+01   5.208 5.49e-07 ***
`1bbl`             1.793e+03  1.398e+03   1.283 0.201341
twobbl             2.192e+03  1.105e+03   1.984 0.048826 *
fourbbl            5.168e+02  3.029e+03   0.171 0.864744
mpfi               2.328e+03  1.069e+03   2.178 0.030787 *
boreratio         -3.686e+03  1.687e+03  -2.185 0.030236 *
stroke            -4.138e+03  9.473e+02  -4.368 2.18e-05 ***
compressionratio -1.443e+03  4.631e+02  -3.117 0.002148 **
horsepower        -4.853e+00  2.171e+01  -0.224 0.823369
peakrpm            2.519e+00  6.583e-01   3.826 0.000183 ***
citympg            1.008e+01  1.574e+02   0.064 0.948995
highwaympg         1.730e+02  1.461e+02   1.184 0.238020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2423 on 169 degrees of freedom
Multiple R-squared:  0.9238,	Adjusted R-squared:  0.908
F-statistic: 58.51 on 35 and 169 DF,  p-value: < 2.2e-16
```

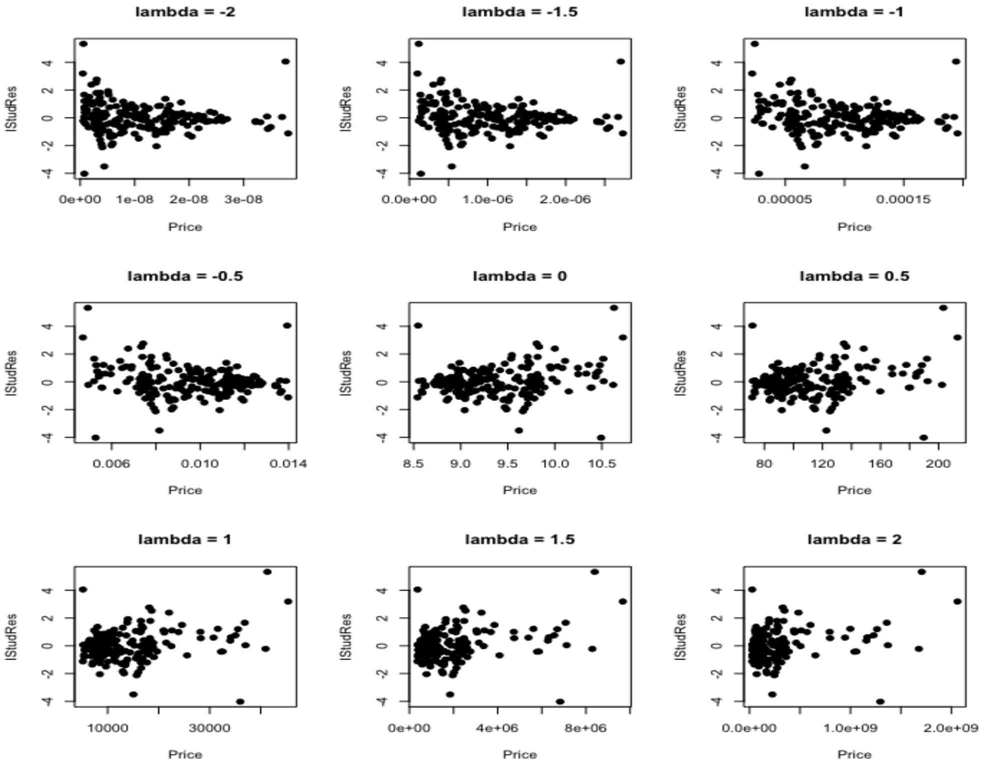
The multiple R-squared has significantly improved, almost 92.38% of the variability in price is accounted for by the predictor variables. Some of the predictor variables became more significant. The index plot of residuals and the plot of studentized residuals vs fitted values were plotted to validate the assumptions.



From the above plots, it can be seen that the residual plots have significantly improved showing no discernable pattern or clustering. However, the normality assumption is still not satisfied and needs further investigation.

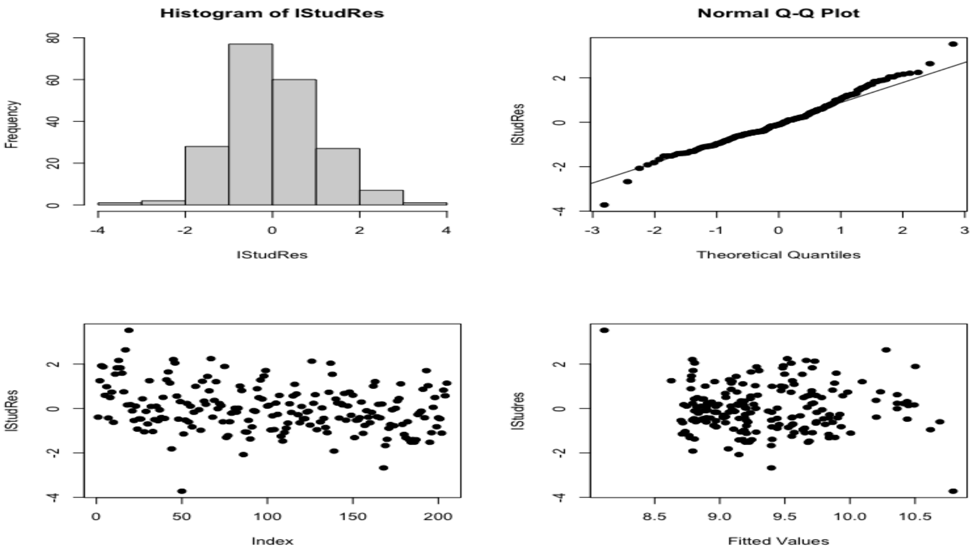
2. Power Transformation

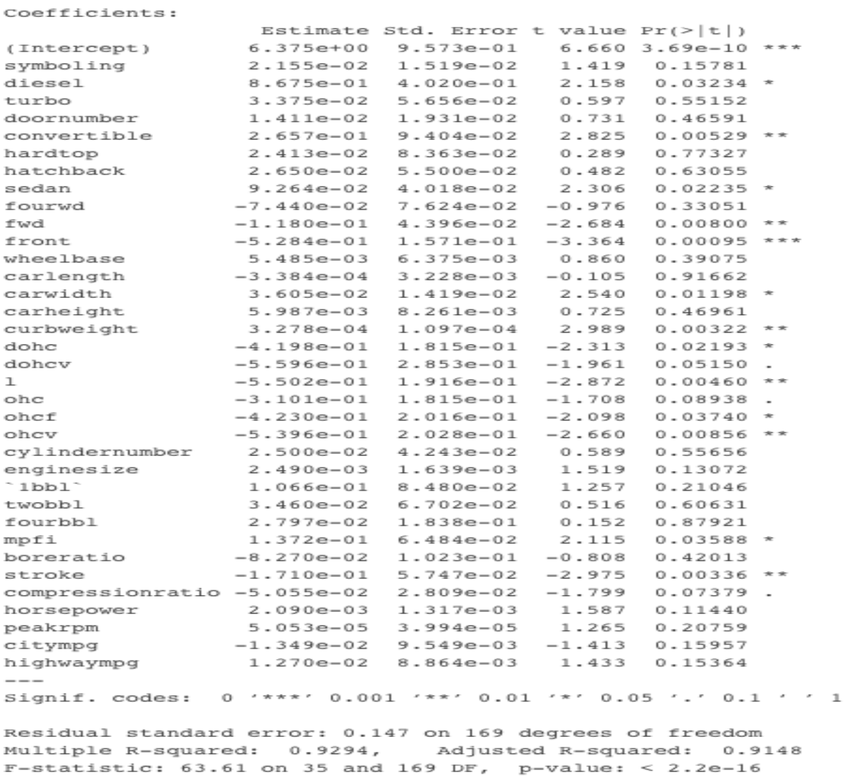
The need for a power transformation was obvious from the start. To be able to choose the most appropriate transformation, different lambda values were selected and the plot of the studentized residuals versus the response variable was plotted to indicate the best lambda variable. The process can be seen below:



The plots above show that the most appropriate transformation is the log transformation since the best graph among all is that of lambda = 0. Accordingly the response variable is transformed and the model is fitted again.

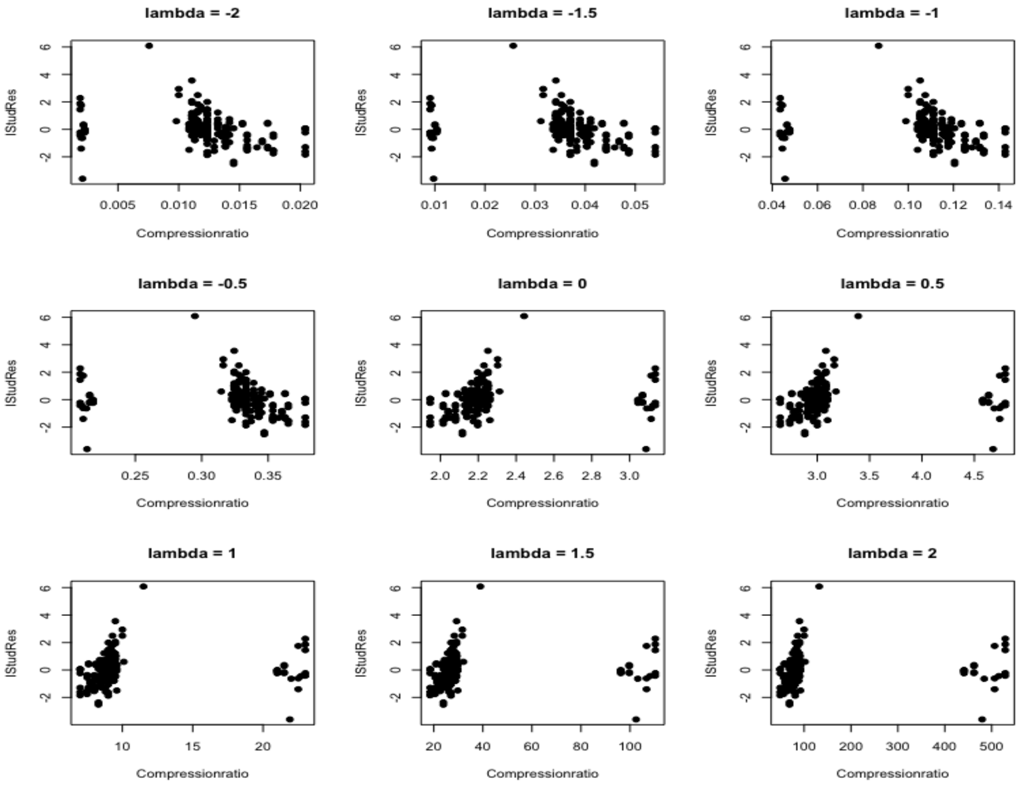
Iteration 3:




```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       6.375e+00  9.573e-01   6.660 3.69e-10 ***
symboling         2.155e-02  1.519e-02   1.419  0.15781    
diesel            8.675e-01  4.020e-01   2.158  0.03234 *  
turbo             3.375e-02  5.656e-02   0.597  0.55152    
doornumber        1.411e-02  1.931e-02   0.731  0.46591    
convertible       2.657e-01  9.404e-02   2.825  0.00529 ** 
hardtop           2.413e-02  8.363e-02   0.289  0.77327    
hatchback         2.650e-02  5.500e-02   0.482  0.63055    
sedan             9.264e-02  4.018e-02   2.306  0.02235 *  
fourwd           -7.440e-02  7.624e-02  -0.976  0.33051    
fwd              -1.180e-01  4.396e-02  -2.684  0.00800 ** 
front            -5.284e-01  1.571e-01  -3.364  0.00095 ***
wheelbase         5.485e-03  6.375e-03   0.860  0.39075    
carlength        -3.384e-04  3.228e-03  -0.105  0.91662    
carwidth          3.605e-02  1.419e-02   2.540  0.01198 *  
carheight         5.987e-03  8.261e-03   0.725  0.46961    
curbweight        3.278e-04  1.097e-04   2.989  0.00322 ** 
dohc             -4.198e-01  1.815e-01  -2.313  0.02193 *  
dohcv            -5.596e-01  2.853e-01  -1.961  0.05150 .  
l                -5.502e-01  1.916e-01  -2.872  0.00460 ** 
ohc              -3.101e-01  1.815e-01  -1.708  0.08938 .  
ohcf             -4.230e-01  2.016e-01  -2.098  0.03740 *  
ohcv             -5.396e-01  2.028e-01  -2.660  0.00856 ** 
cylindernumber    2.500e-02  4.243e-02   0.589  0.55656    
enginesize        2.490e-03  1.639e-03   1.519  0.13072    
`1bbl`            1.066e-01  8.480e-02   1.257  0.21046    
twobbl            3.460e-02  6.702e-02   0.516  0.60631    
fourbbl           2.797e-02  1.838e-01   0.152  0.87921    
mpfi              1.372e-01  6.484e-02   2.115  0.03588 *  
boreratio        -8.270e-02  1.023e-01  -0.808  0.42013    
stroke           -1.710e-01  5.747e-02  -2.975  0.00336 ** 
compressionratio -5.055e-02  2.809e-02  -1.799  0.07379 .  
horsepower        2.090e-03  1.317e-03   1.587  0.11440    
peakrpm           5.053e-05  3.994e-05   1.265  0.20759    
citympg          -1.349e-02  9.549e-03  -1.413  0.15957    
highwaympg        1.270e-02  8.864e-03   1.433  0.15364    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.147 on 169 degrees of freedom
Multiple R-squared:  0.9294,	Adjusted R-squared:  0.9148 
F-statistic: 63.61 on 35 and 169 DF,  p-value: < 2.2e-16
```

The assumptions have been satisfied after both converting the categorical variables into binary variables and power transformation has been performed. Even the normality assumption of the residuals is satisfied as the residuals have almost taken the nature of a straight line. The multiple R-squared has improved significantly to be 92.8% and the F-statistic has increased significantly to 63.6. This is the best model we have reached so far. However, the multicollinearity problem is still present and still requires further iterations to be made.

The plots of the studentized residuals against all the predictor variables were plotted to observe the need for transformation of the predictor variables. The compression ratio plot showed an obvious case of clustering. Transformation for this predictor variable has been made. As shown below, the transformation of the compression ratio variable with a value of lambda equal to -2 has significantly improved the plot of the standardized residuals against the fitted values.

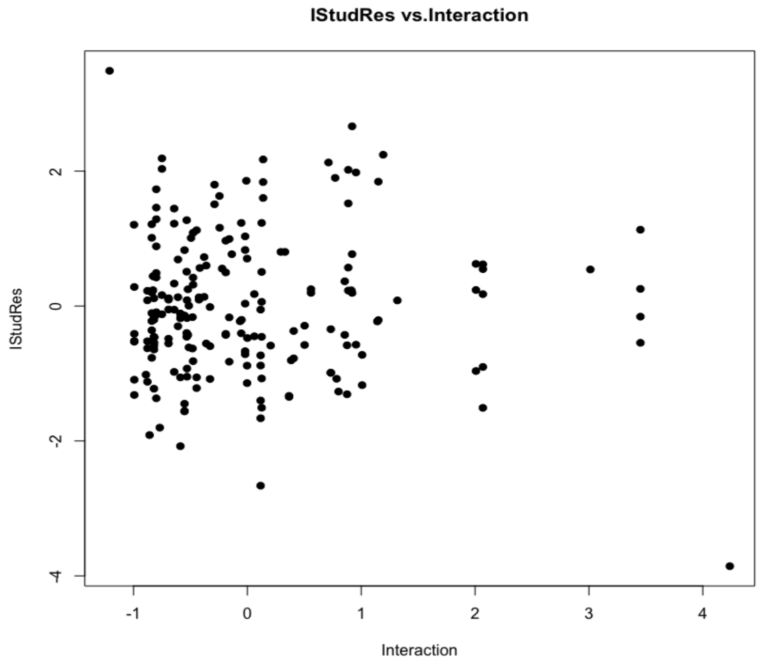


However, upon regressing the model with the transformed predictor variable, no change in the regression results was identified and hence transformation was not appropriate. This problem might be solved using interaction variables instead.

3. Interaction Variables

Since the transformation of the compression ratio variable did not yield to statistically significant results, the appropriate course of action would be to create an interaction variable using the compression ratio. The horsepower and compression ratio are closely related. The higher the compression ratio, the higher the horsepower. Accordingly, we created an interaction variable between the horsepower and compression ratio.

Iteration 4



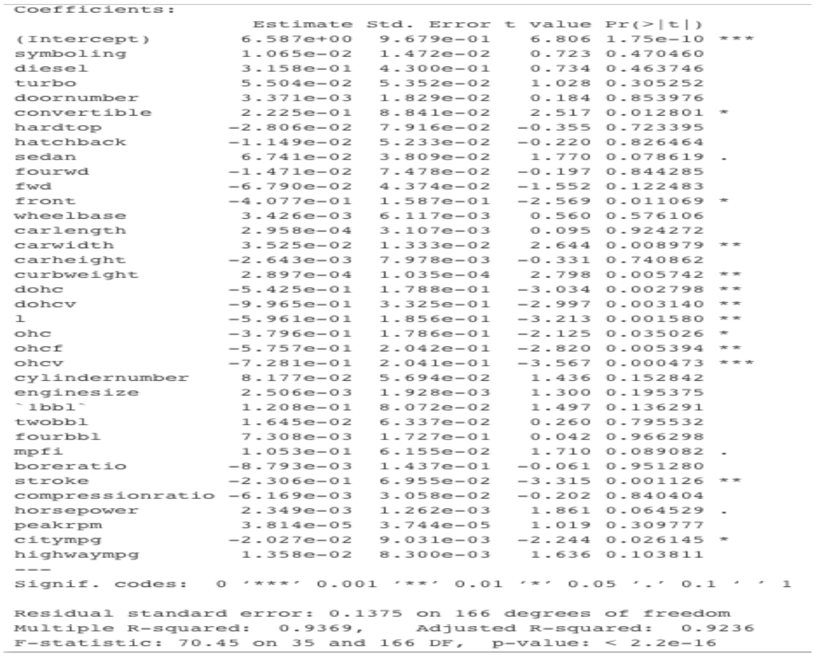
As shown above, the interaction between horsepower and compression ratio got rid of the clustering.

3. Detection of outliers

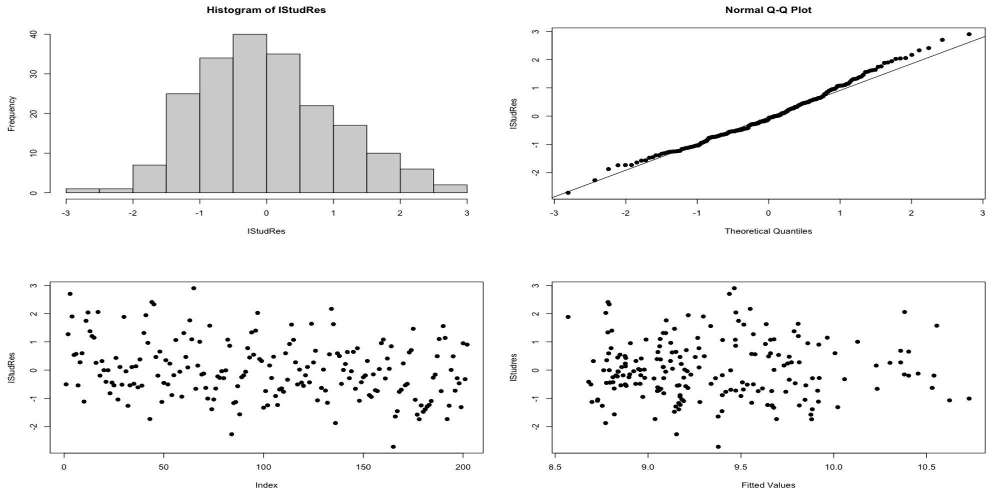
As seen from the previous regression outputs and plots, the data contains outliers and influential points and thus the implicit assumption is violated. We were able to successfully detect the outliers upon transforming the data and using the identify() function on R. We detected 3 outliers using Cook's Index Plot and Hadi's Influence Index Plot, namely: observation 19 (Chevrolet Impala), observation 50 (Jaguar XK) and observation 135 (Saab 99 Le).

After investigating the reasons behind the existence of outliers, we could not find a reason as to why they are considered outliers. One logical reason is that the brand name increases/ decreases a car's price regardless of its attributes. Accordingly, we have decided to drop the outliers and see whether they have a significant effect on the regression results or not.

Iteration 4:



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          6.587e+00  9.679e-01   6.806 1.75e-10 ***
symboling            1.065e-02  1.472e-02   0.723 0.470460    
diesel               3.158e-01  4.300e-01   0.734 0.463746    
turbo                5.504e-02  5.352e-02   1.028 0.305252    
doornumber           3.371e-03  1.829e-02   0.184 0.853976    
convertible          2.225e-01  8.841e-02   2.517 0.012801 *  
hardtop             -2.806e-02  7.916e-02  -0.355 0.723395    
hatchback           -1.149e-02  5.233e-02  -0.220 0.826464    
sedan                6.741e-02  3.809e-02   1.770 0.078619 .  
fourwd              -1.471e-02  7.478e-02  -0.197 0.844285    
fwd                 -6.790e-02  4.374e-02  -1.552 0.122483    
front               -4.077e-01  1.587e-01  -2.569 0.011069 *  
wheelbase            3.426e-03  6.117e-03   0.560 0.576106    
carlength            2.958e-04  3.107e-03   0.095 0.924272    
carwidth             3.525e-02  1.333e-02   2.644 0.008979 ** 
carheight           -2.643e-03  7.978e-03  -0.331 0.740862    
curbweight           2.897e-04  1.035e-04   2.798 0.005742 ** 
dohc                -5.425e-01  1.788e-01  -3.034 0.002798 ** 
dohcv               -9.965e-01  3.325e-01  -2.997 0.003140 ** 
l                   -5.961e-01  1.856e-01  -3.213 0.001580 ** 
ohc                 -3.796e-01  1.786e-01  -2.125 0.035026 *  
ohcf                -5.757e-01  2.042e-01  -2.820 0.005394 ** 
ohcv                -7.281e-01  2.041e-01  -3.567 0.000473 ***
cylindernumber       8.177e-02  5.694e-02   1.436 0.152842    
enginesize           2.506e-03  1.928e-03   1.300 0.195375    
`1bbl`               1.208e-01  8.072e-02   1.497 0.136291    
twobbl               1.645e-02  6.337e-02   0.260 0.795532    
fourbbl              7.308e-03  1.727e-01   0.042 0.966298    
mpfi                 1.053e-01  6.155e-02   1.710 0.089082 .  
boreratio           -8.793e-03  1.437e-01  -0.061 0.951280    
stroke              -2.306e-01  6.955e-02  -3.315 0.001126 ** 
compressionratio -6.169e-03  3.058e-02  -0.202 0.840404    
horsepower           2.349e-03  1.262e-03   1.861 0.064529 .  
peakrpm              3.814e-05  3.744e-05   1.019 0.309777    
citympg             -2.027e-02  9.031e-03  -2.244 0.026145 *  
highwaympg           1.358e-02  8.300e-03   1.636 0.103811    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1375 on 166 degrees of freedom
Multiple R-squared:  0.9369,	Adjusted R-squared:  0.9236 
F-statistic: 70.45 on 35 and 166 DF,  p-value: < 2.2e-16
```

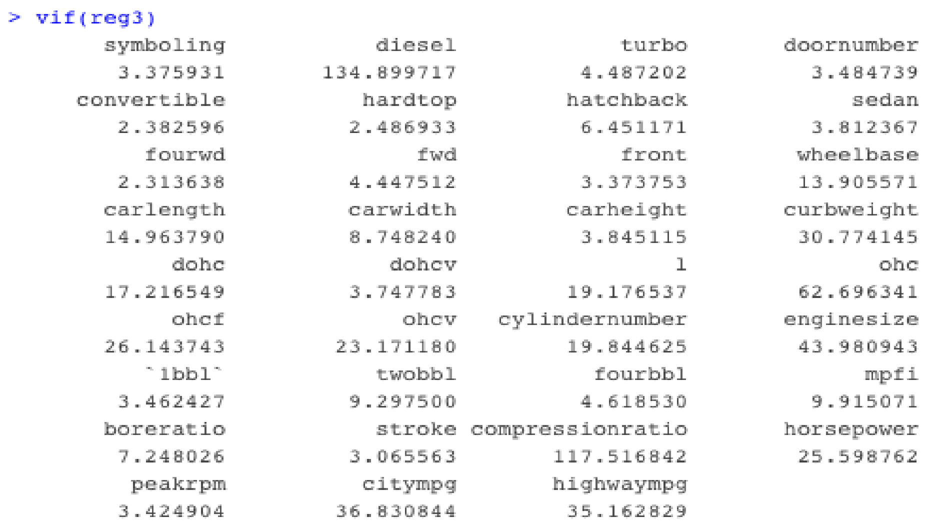


The omission of the outliers has significantly improved the model. The normality and homoscedasticity assumption hold. All other assumptions remain validated as well. We can observe a significant improvement in the model in terms of the multiple R-squared and F-statistic. 93.69% of the variability in price can be explained by the predictors.

4. Multicollinearity Problem:

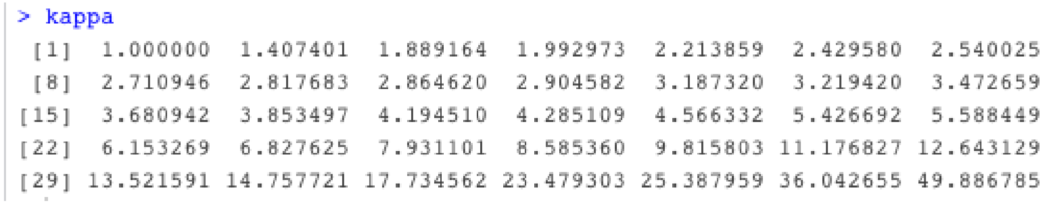
As seen from the scatter plot matrix above and all the regression results, it is obvious that there is a multicollinearity problem in this data set. The t-values are small and the p-values are insignificant. In addition, the regression coefficients have high standard errors. In order to substantiate this assumption, the VIF (Variance Inflation Factor) and the conditional indices were calculated.

The VIF shows an indication of collinearity among the following variables: diesel, wheel base, car length, curb weight, dohc, l, ohc,ohcf,ohc,cylinder number, engine size, compression ratio, horsepower, city mpg and highway mpg.



```
> vif(reg3)
        symboling            diesel             turbo        doornumber 
         3.375931        134.899717          4.487202          3.484739 
      convertible           hardtop         hatchback             sedan 
         2.382596          2.486933          6.451171          3.812367 
           fourwd               fwd             front         wheelbase 
         2.313638          4.447512          3.373753         13.905571 
        carlength          carwidth         carheight        curbweight 
        14.963790          8.748240          3.845115         30.774145 
             dohc            dohcv                 l               ohc 
        17.216549          3.747783         19.176537         62.696341 
             ohcf              ohcv    cylindernumber        enginesize 
        26.143743         23.171180         19.844625         43.980943 
           `1bbl`            twobbl          fourbbl              mpfi 
         3.462427          9.297500          4.618530          9.915071 
        boreratio            stroke compressionratio        horsepower 
         7.248026          3.065563        117.516842         25.598762 
          peakrpm           citympg        highwaympg 
         3.424904         36.830844         35.162829 
```

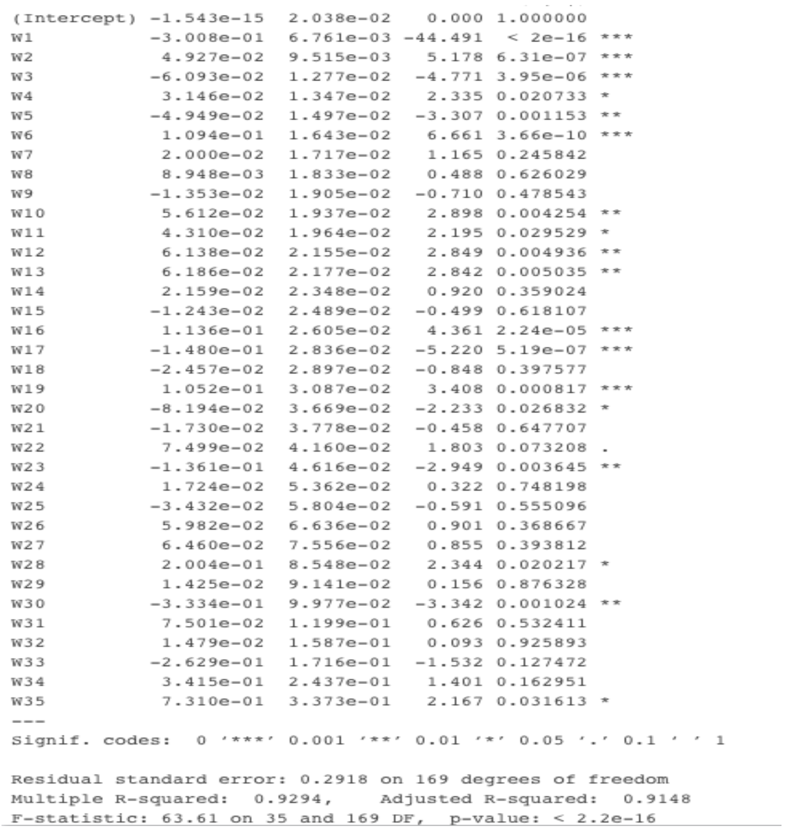
Since there are 9 large conditional indices that exceed the value of 10, there are 9 subsets of collinear variables.



```
> kappa
 [1]  1.000000  1.407401  1.889164  1.992973  2.213859  2.429580  2.540025
 [8]  2.710946  2.817683  2.864620  2.904582  3.187320  3.219420  3.472659
[15]  3.680942  3.853497  4.194510  4.285109  4.566332  5.426692  5.588449
[22]  6.153269  6.827625  7.931101  8.585360  9.815803 11.176827 12.643129
[29] 13.521591 14.757721 17.734562 23.479303 25.387959 36.042655 49.886785
```

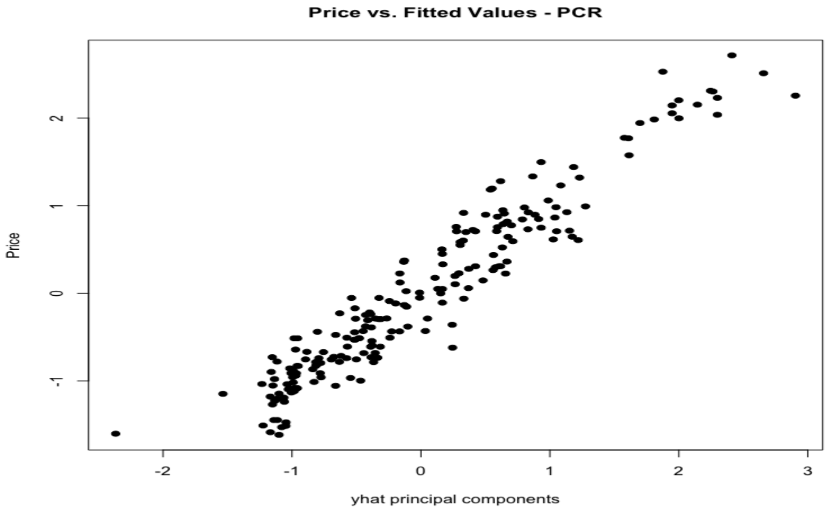
After proving the existence of collinearity that has harmful effects on the regression results, the principal components regression was implemented to get rid of collinearity and provide us with the best model of the data at hand.

Iteration 5:



```
(Intercept) -1.543e-15  2.038e-02   0.000 1.000000    
W1          -3.008e-01  6.761e-03 -44.491  < 2e-16 ***
W2           4.927e-02  9.515e-03   5.178 6.31e-07 ***
W3          -6.093e-02  1.277e-02  -4.771 3.95e-06 ***
W4           3.146e-02  1.347e-02   2.335 0.020733 *  
W5          -4.949e-02  1.497e-02  -3.307 0.001153 ** 
W6           1.094e-01  1.643e-02   6.661 3.66e-10 ***
W7           2.000e-02  1.717e-02   1.165 0.245842    
W8           8.948e-03  1.833e-02   0.488 0.626029    
W9          -1.353e-02  1.905e-02  -0.710 0.478543    
W10          5.612e-02  1.937e-02   2.898 0.004254 ** 
W11          4.310e-02  1.964e-02   2.195 0.029529 *  
W12          6.138e-02  2.155e-02   2.849 0.004936 ** 
W13          6.186e-02  2.177e-02   2.842 0.005035 ** 
W14          2.159e-02  2.348e-02   0.920 0.359024    
W15         -1.243e-02  2.489e-02  -0.499 0.618107    
W16          1.136e-01  2.605e-02   4.361 2.24e-05 ***
W17         -1.480e-01  2.836e-02  -5.220 5.19e-07 ***
W18         -2.457e-02  2.897e-02  -0.848 0.397577    
W19          1.052e-01  3.087e-02   3.408 0.000817 ***
W20         -8.194e-02  3.669e-02  -2.233 0.026832 *  
W21         -1.730e-02  3.778e-02  -0.458 0.647707    
W22          7.499e-02  4.160e-02   1.803 0.073208 .  
W23         -1.361e-01  4.616e-02  -2.949 0.003645 ** 
W24          1.724e-02  5.362e-02   0.322 0.748198    
W25         -3.432e-02  5.804e-02  -0.591 0.555096    
W26          5.982e-02  6.636e-02   0.901 0.368667    
W27          6.460e-02  7.556e-02   0.855 0.393812    
W28          2.004e-01  8.548e-02   2.344 0.020217 *  
W29          1.425e-02  9.141e-02   0.156 0.876328    
W30         -3.334e-01  9.977e-02  -3.342 0.001024 ** 
W31          7.501e-02  1.199e-01   0.626 0.532411    
W32          1.479e-02  1.587e-01   0.093 0.925893    
W33         -2.629e-01  1.716e-01  -1.532 0.127472    
W34          3.415e-01  2.437e-01   1.401 0.162951    
W35          7.310e-01  3.373e-01   2.167 0.031613 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2918 on 169 degrees of freedom
Multiple R-squared:  0.9294,	Adjusted R-squared:  0.9148 
F-statistic: 63.61 on 35 and 169 DF,  p-value: < 2.2e-16
```

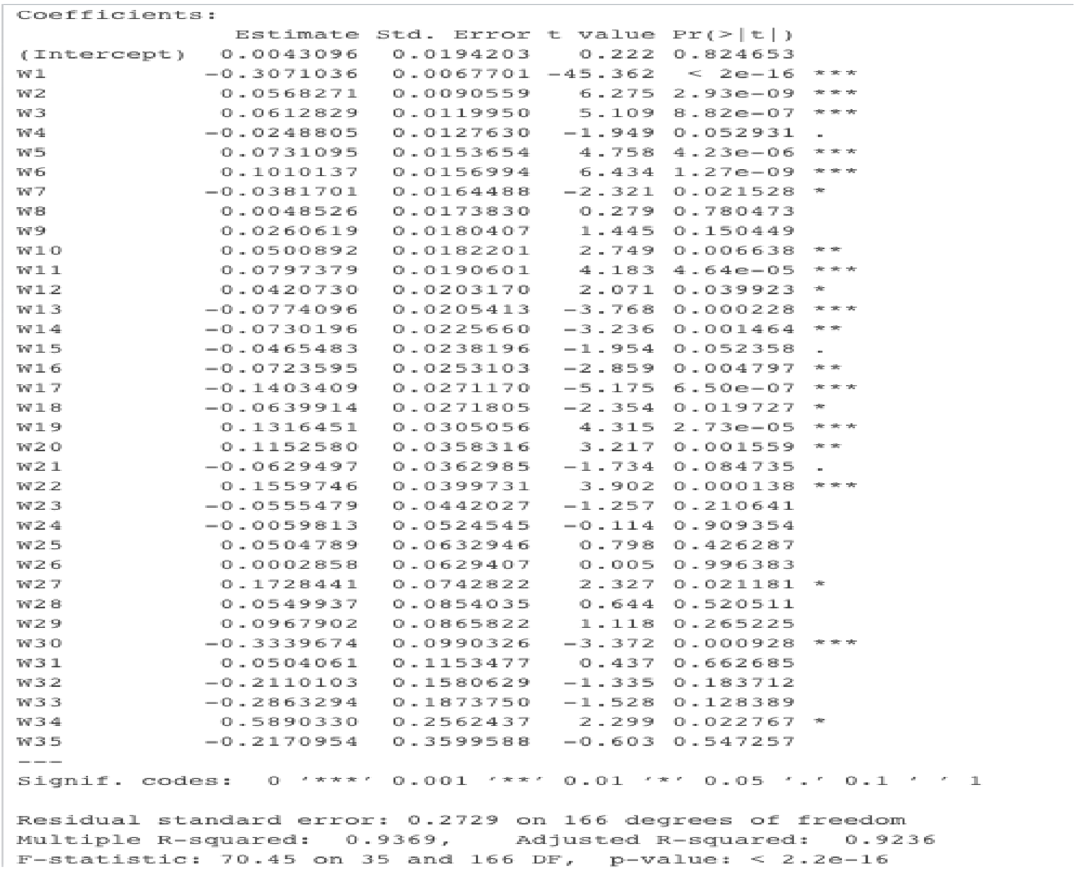
After implementing the principal components regression, the above regression results were obtained. The regression output shows a significant improvement in the model in terms of the multiple R-squared and F-statistic.



After implementing the principal components regression and removing insignificant components, we got rid of the multicollinearity problem and thus have reached the most adequate model yet. As seen from the plot above, the transformed response variable against the fitted values shows a straight line meaning that the model fitted using the principal components method accurately describes the data.

Iteration 6:

As shown below, upon omitting the outliers and implementing the principal components regression, the model has significantly improved in terms of the Multiple R-squared, F-statistic and predictors' significance. This model is the best model generated yet.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0043096  0.0194203   0.222 0.824653    
W1          -0.3071036  0.0067701 -45.362  < 2e-16 ***
W2           0.0568271  0.0090559   6.275 2.93e-09 ***
W3           0.0612829  0.0119950   5.109 8.82e-07 ***
W4          -0.0248805  0.0127630  -1.949 0.052931 .  
W5           0.0731095  0.0153654   4.758 4.23e-06 ***
W6           0.1010137  0.0156994   6.434 1.27e-09 ***
W7          -0.0381701  0.0164488  -2.321 0.021528 *  
W8           0.0048526  0.0173830   0.279 0.780473    
W9           0.0260619  0.0180407   1.445 0.150449    
W10          0.0500892  0.0182201   2.749 0.006638 ** 
W11          0.0797379  0.0190601   4.183 4.64e-05 ***
W12          0.0420730  0.0203170   2.071 0.039923 *  
W13         -0.0774096  0.0205413  -3.768 0.000228 ***
W14         -0.0730196  0.0225660  -3.236 0.001464 ** 
W15         -0.0465483  0.0238196  -1.954 0.052358 .  
W16         -0.0723595  0.0253103  -2.859 0.004797 ** 
W17         -0.1403409  0.0271170  -5.175 6.50e-07 ***
W18         -0.0639914  0.0271805  -2.354 0.019727 *  
W19          0.1316451  0.0305056   4.315 2.73e-05 ***
W20          0.1152580  0.0358316   3.217 0.001559 ** 
W21         -0.0629497  0.0362985  -1.734 0.084735 .  
W22          0.1559746  0.0399731   3.902 0.000138 ***
W23         -0.0555479  0.0442027  -1.257 0.210641    
W24         -0.0059813  0.0524545  -0.114 0.909354    
W25          0.0504789  0.0632946   0.798 0.426287    
W26          0.0002858  0.0629407   0.005 0.996383    
W27          0.1728441  0.0742822   2.327 0.021181 *  
W28          0.0549937  0.0854035   0.644 0.520511    
W29          0.0967902  0.0865822   1.118 0.265225    
W30         -0.3339674  0.0990326  -3.372 0.000928 ***
W31          0.0504061  0.1153477   0.437 0.662685    
W32         -0.2110103  0.1580629  -1.335 0.183712    
W33         -0.2863294  0.1873750  -1.528 0.128389    
W34          0.5890330  0.2562437   2.299 0.022767 *  
W35         -0.2170954  0.3599588  -0.603 0.547257    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2729 on 166 degrees of freedom
Multiple R-squared:  0.9369,	Adjusted R-squared:  0.9236 
F-statistic: 70.45 on 35 and 166 DF,  p-value: < 2.2e-16
```

Conclusion

After analyzing the data at hand and observing the different features that have an effect in determining the price of a car. It is evident that the linear regression model is suitable for explaining the relationship between the response variable and the predictor variable. However, one problem remains unresolved which is the problem of outliers. From only looking at the data there is no direct explanation as to why the observations detected are outliers. With that being said it is important to discuss the matter with a domain expert to get better insights as to why these observations were marked as outliers. Overall the project captured most of the methods that help in carrying out multiple regression, but it also revealed that regression is a very sensitive method that requires a ton of analysis to be able to draw the right conclusions and come up with the best model. The project captures 6 regression model iterations starting with a multiple R squared of 88% to a multiple R squared of 94%. With each iteration the assumptions were observed and validated to be able to draw conclusions. Lastly, the principal component regression was developed to yield the best model so far.