Math statistics 5120 Project

Gerard shu Fuhnwi and Godlove Jator

Course instructor: Dr. Mathew Jones

Introduction

The Ford Motor Company is an American multinational automaker headquartered in Dearborn, Michigan, a suburb of Detroit. It was founded by Henry Ford and incorporated on June 1903. The Company sells automobiles and commercial vehicles under the Ford brand (model) and most luxury cars under the Lincoln brand (model). Ford also has branches in Brazil, United Kingdom and Australia. We seek to build a model to predict the prices of used Ford Cars in certain locations in the United State of American.

Data description

Variable Name	Description
Color	Color of the Car
Year	The year in which the car was produced
Mileage	Number of miles a car covers
Location	Place where a car is in the United states
Price	Price of used car in $
Model	Brand of the used car
Age	Age of car

NB: This data was collected from 1990 to 2009

Data Link: https://assets.datacamp.com/production/course_1586/datasets/Fords.csv

Structure of the data

```
setwd("C:/Users/Gerard/Desktop/Gboy")
Ford_cars= read.csv("fords.csv")
names(Ford_cars)
[1] "X"        "Year"     "Mileage" "Price"    "Color"    "Location" "Model"   
"Age"
str(Ford_cars)
'data.frame':	635 obs. of  8 variables:
 $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Year    : int  1990 1994 1995 1995 1995 1996 1997 1998 1998 1999 ...
 $ Mileage : int  NA 94000 NA 68000 NA 115730 74564 143000 91000 88000 ...
 $ Price   : int  1600 1988 2288 2495 1995 2199 2995 1200 2488 3300 ...
 $ Color   : Factor w/ 10 levels "beige","black",..: NA 10 10 NA NA 1 8 3 9 1
0 ...
 $ Location: Factor w/ 6 levels "Cambridge","Dallas",..: 5 5 5 5 5 5 5 3 5 5
...
 $ Model   : Factor w/ 6 levels "GL","Limited",..: NA 1 NA NA 1 1 1 4 NA NA .
..
 $ Age     : int  19 15 14 14 14 13 12 11 11 10 ...

summary(Ford_cars)
       X               Year         Mileage            Price            Color    
 Min.   :  1.0   Min.   :1990   Min.   :     42   Min.   : 1200   gray    :191  
 1st Qu.:159.5   1st Qu.:2003   1st Qu.: 31773   1st Qu.: 5995   white   :101  
 Median :318.0   Median :2006   Median : 48898   Median : 8950   beige   : 63  
 Mean   :318.0   Mean   :2005   Mean   : 56016   Mean   : 9421   blue    : 59  
 3rd Qu.:476.5   3rd Qu.:2007   3rd Qu.: 74503   3rd Qu.:11665   black   : 55  
 Max.   :635.0   Max.   :2009   Max.   :181484   Max.   :21995   (Other):156  
                                NA's   :19       NA's   :6       NA's   : 10  
        Location       Model          Age       
 Cambridge   :141   GL     : 16   Min.   : 0.00  
 Dallas      :136   Limited: 32   1st Qu.: 2.00  
 Fresno      : 23   LX     : 12   Median : 3.00  
 Philadelphia:137   SE     :283   Mean   : 4.28  
 Phoenix     : 85   SEL    :208   3rd Qu.: 6.00  
 St Paul     :113   SES    : 76   Max.   :19.00  
                    NA's   :  8                  
head(Ford_cars,10)
    X Year Mileage Price Color Location Model Age
1   1 1990      NA  1600  <NA>  Phoenix  <NA>  19
2   2 1994   94000  1988 white  Phoenix    GL  15
3   3 1995      NA  2288 white  Phoenix  <NA>  14
4   4 1995   68000  2495  <NA>  Phoenix  <NA>  14
5   5 1995      NA  1995  <NA>  Phoenix    GL  14
6   6 1996  115730  2199 beige  Phoenix    GL  13
7   7 1997   74564  2995 green  Phoenix    GL  12
8   8 1998  143000  1200  blue   Fresno    SE  11
9   9 1998   91000  2488   red  Phoenix  <NA>  11
10 10 1999   88000  3300 white  Phoenix  <NA>  10
```

Data Cleaning

The dataset had some missing values for some variables like mileage(predictor) and price (response), so we had to replace them with the median since it is not affected by extreme values.

Code

```
Ford_cars1=Ford_cars

Ford_cars1$Mileage[which(is.na(Ford_cars$Mileage))]=median(Ford_cars1$Mileage,na.rm = T)

Ford_cars1

Ford_cars2=Ford_cars1

Ford_cars2$Price[which(is.na(Ford_cars1$Price))]=median(Ford_cars2$Price,na.rm = T)

head(Ford_cars2[-1], 10)
```

Output

```
head(Ford_cars2[-1],10)
   Year  Mileage Price Color Location Model Age
1  1990  48897.5  1600  <NA>  Phoenix  <NA>  19
2  1994  94000.0  1988 white  Phoenix    GL  15
3  1995  48897.5  2288 white  Phoenix  <NA>  14
4  1995  68000.0  2495  <NA>  Phoenix  <NA>  14
5  1995  48897.5  1995  <NA>  Phoenix    GL  14
6  1996 115730.0  2199 beige  Phoenix    GL  13
7  1997  74564.0  2995 green  Phoenix    GL  12
8  1998 143000.0  1200  blue   Fresno    SE  11
9  1998  91000.0  2488   red  Phoenix  <NA>  11
10 1999  88000.0  3300 white  Phoenix  <NA>  10
```

Checking for correlation of variables

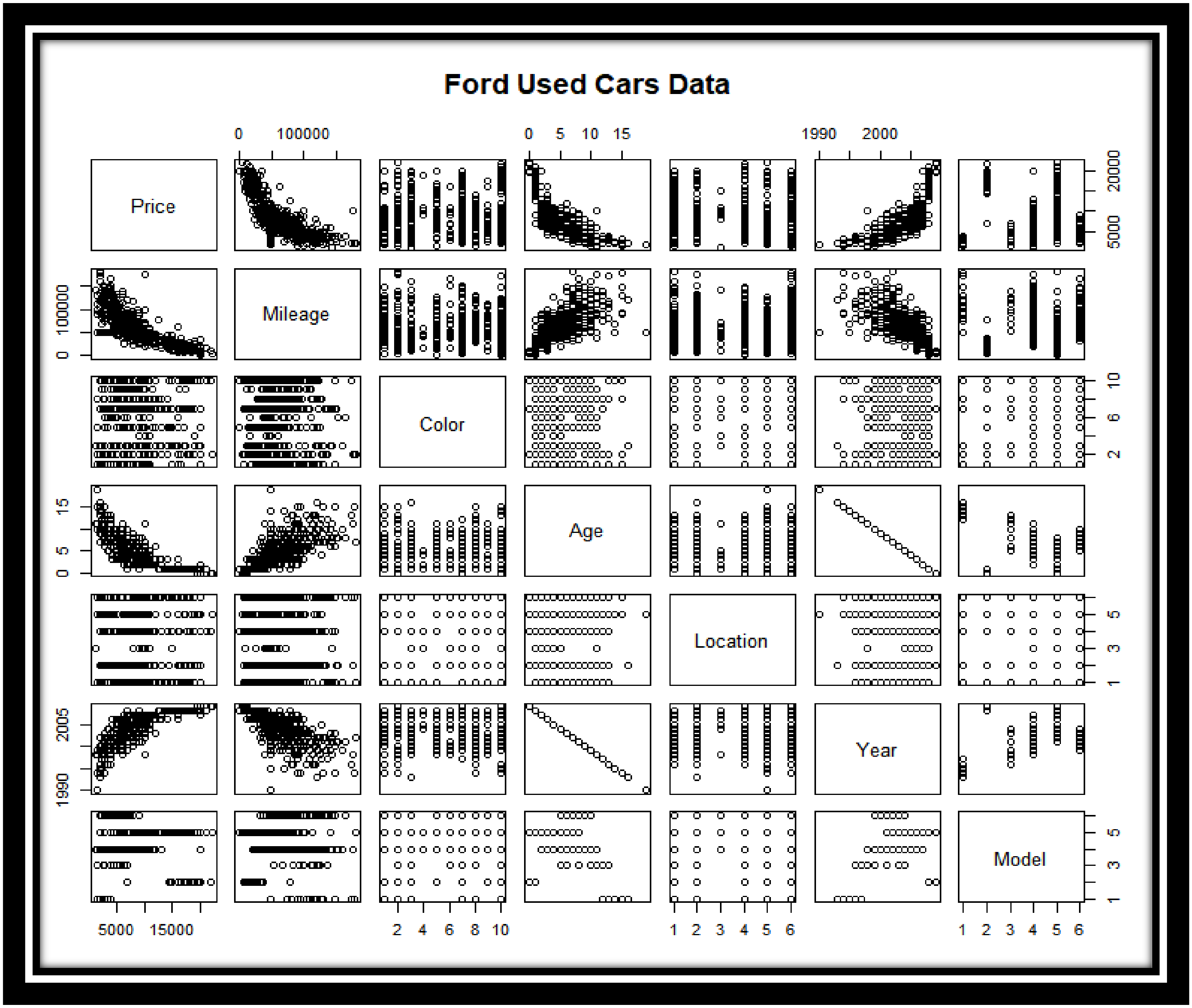
Code

cor(Ford_cars2[,-c(5:7)])

pairs(~ Price + Mileage + Color + Age + Location + Year + Model , data = Ford_cars2, main = "Ford Used Cars Data")

Output

```
cor(Ford_cars2[,-c(5:7)])
                  X        Year    Mileage       Price         Age
X        1.00000000 -0.05635155  0.1496194 -0.03270205  0.05635155
Year    -0.05635155  1.00000000 -0.7339419  0.78525840 -1.00000000
Mileage  0.14961940 -0.73394194  1.0000000 -0.78021127  0.73394194
Price   -0.03270205  0.78525840 -0.7802113  1.00000000 -0.78525840
Age      0.05635155 -1.00000000  0.7339419 -0.78525840  1.00000000
```



Observation: From the output above, we can observe that the response variable (Price) is highly correlated with the predictor variables (Year, Age and mileage).

NB: Since Year is highly correlated with age, we will remove year and used age in our model.

Regression model

Model 1

Code

fit1=lm(Price~., data = Ford_cars2)

summary(fit1)

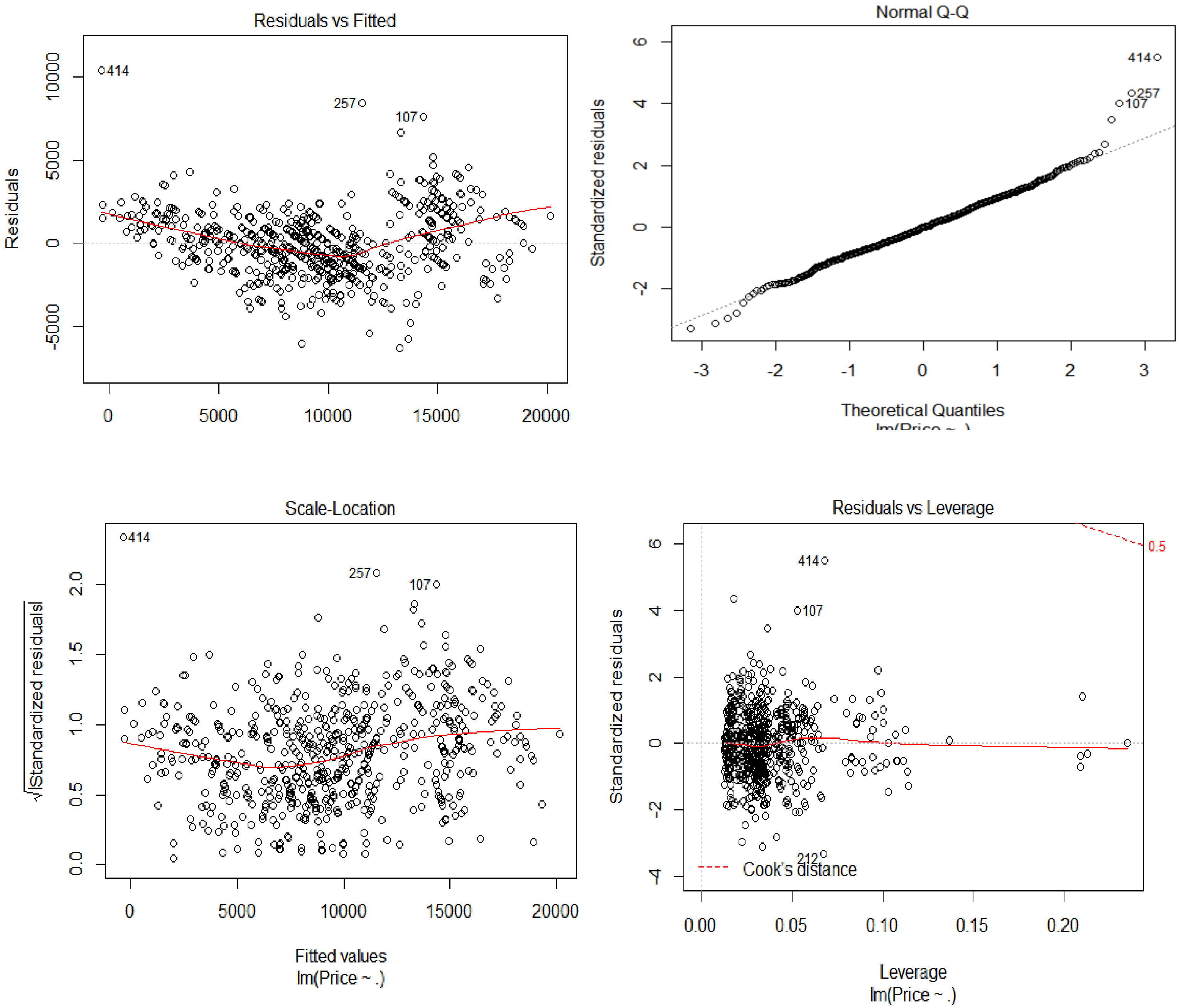
plot(fit1)

confint(fit1)

Output

```
summary(fit1)

Call:
lm(formula = Price ~ ., data = Ford_cars2)

Residuals:
    Min      1Q  Median      3Q     Max 
-6255.2 -1253.5   -31.7  1231.0 10372.6 

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -1.179e+06  9.525e+04 -12.384  < 2e-16 ***
X                    -8.549e+00  1.866e+00  -4.582 5.61e-06 ***
Year                  5.955e+02  4.758e+01  12.513  < 2e-16 ***
Mileage              -4.620e-02  3.822e-03 -12.088  < 2e-16 ***
Colorblack            1.355e+03  3.668e+02   3.693 0.000242 ***
Colorblue             1.375e+03  3.636e+02   3.781 0.000172 ***
Colorbrown            2.261e+03  9.154e+02   2.470 0.013789 *  
Colorburgundy         8.772e+02  3.784e+02   2.318 0.020779 *  
Colorgold             8.812e+02  5.139e+02   1.715 0.086930 .  
Colorgray             1.257e+03  2.980e+02   4.217 2.86e-05 ***
Colorgreen            6.331e+02  3.862e+02   1.639 0.101676    
Colorred              9.963e+02  4.585e+02   2.173 0.030165 *  
Colorwhite            2.047e+03  3.642e+02   5.621 2.92e-08 ***
LocationDallas        2.769e+03  6.568e+02   4.215 2.88e-05 ***
LocationFresno       -2.209e+03  5.134e+02  -4.302 1.98e-05 ***
LocationPhiladelphia  1.469e+03  3.435e+02   4.277 2.21e-05 ***
LocationPhoenix      -2.217e+03  3.761e+02  -5.894 6.30e-09 ***
LocationSt Paul       2.832e+03  6.431e+02   4.404 1.26e-05 ***
ModelLimited          2.487e+03  7.755e+02   3.207 0.001414 ** 
ModelLX              -1.582e+03  7.787e+02  -2.031 0.042686 *  
ModelSE              -3.017e+03  6.284e+02  -4.801 1.99e-06 ***
ModelSEL             -4.267e+02  6.754e+02  -0.632 0.527846    
ModelSES             -2.438e+03  6.192e+02  -3.938 9.19e-05 ***
Age                          NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1954 on 597 degrees of freedom
  (15 observations deleted due to missingness)
```

```
Multiple R-squared:  0.8302,   Adjusted R-squared:  0.8239
F-statistic: 132.7 on 22 and 597 DF,  p-value: < 2.2e-16
```



Observation: Most of the variable in the model above seem significant, but may be overfitted due to so many variables, and equally the predictor variable (age) seem to disappear in our model, which has a lot to do with the response variable (Price). So, we will fit another base mostly on the quantitative predictor variables (age and mileage). It can also be observed from the plots above that 257, 414 and 107 seem to be outliers.

This can also be confirmed by the outlierTest below:

```
outlierTest(fit1)
    rstudent unadjusted p-value Bonferonni p
414 5.639082         2.6421e-08   1.6381e-05
257 4.417334         1.1869e-05   7.3590e-03
107 4.065437         5.4375e-05   3.3713e-02
```

Training and Testing

- **Training**

Code

n=nrow(Ford_cars2)

trainindex = sample(1:n, size = round(0.7*n), replace = F)

train_Ford = Ford_cars2[trainindex,]

test_Ford = Ford_cars2[-trainindex,]

head(train_Ford)

head(test_Ford)

Output

```
head(train_Ford)
      X Year Mileage Price Color     Location Model Age
19   19 2002 48897.5  2788  gray      Phoenix   SEL   7
418 418 2004 79650.0  4950 black       Dallas    SE   5
423 423 2000 64600.0  5995 black      St Paul    SE   9
478 478 2006 65956.0  8999  gold       Dallas   SEL   3
322 322 2005 43675.0  7991 white Philadelphia    SE   4
555 555 2008 35508.0 13995 green      St Paul   SEL   1
> head(test_Ford)
    X Year  Mileage Price Color Location Model Age
5   5 1995  48897.5  1995  <NA>  Phoenix    GL  14
9   9 1998  91000.0  2488   red  Phoenix  <NA>  11
13 13 2000 115123.0  2995 white  Phoenix    SE   9
14 14 2000  99000.0  2988  gray  Phoenix   SES   9
17 17 2002  48897.5  2800  blue  Phoenix    SE   7
24 24 2003  80267.0  6491 white  Phoenix   SES   6
```

Simple Linear Model

Code

fit2=lm(Price~Age, data = train_Ford)

plot(Price~Age, data = train_Ford)

abline(fit2, lwd=2, col="red")

summary(fit2)

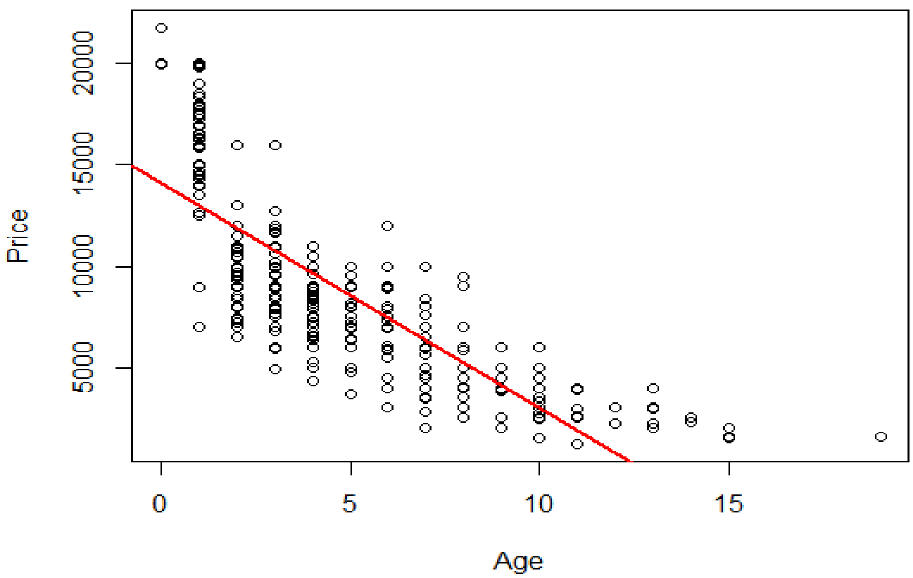
plot(fit2)

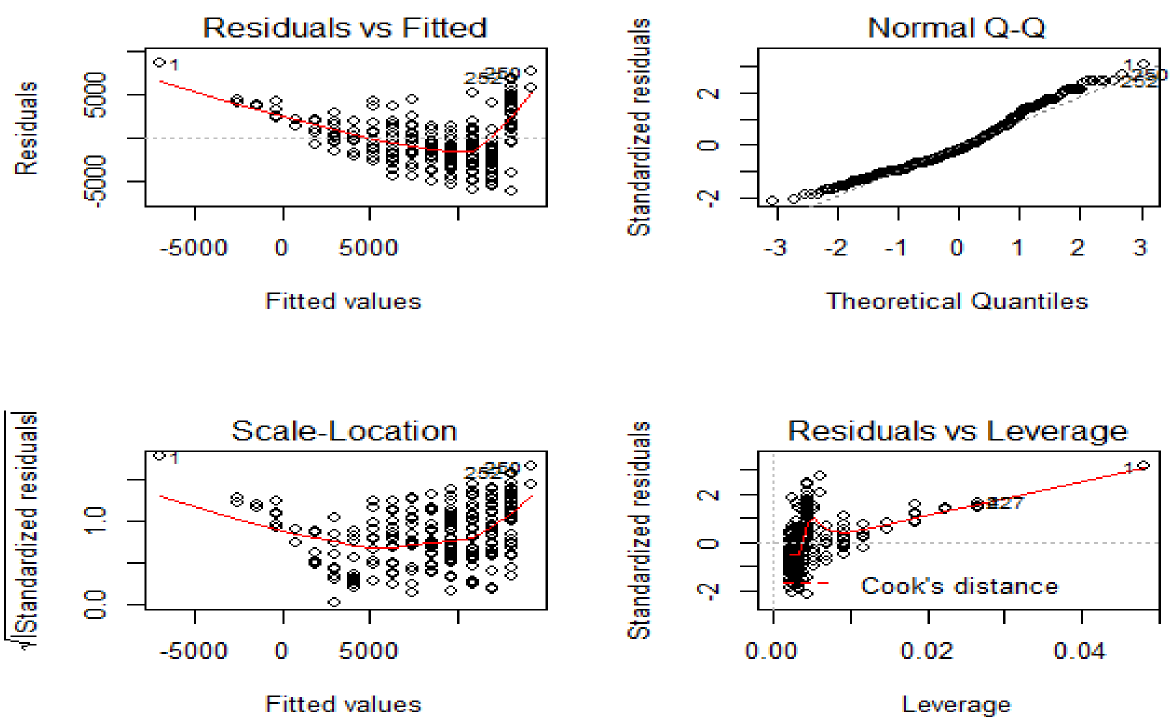
outlierTest(fit2)

confint(fit2)

Output

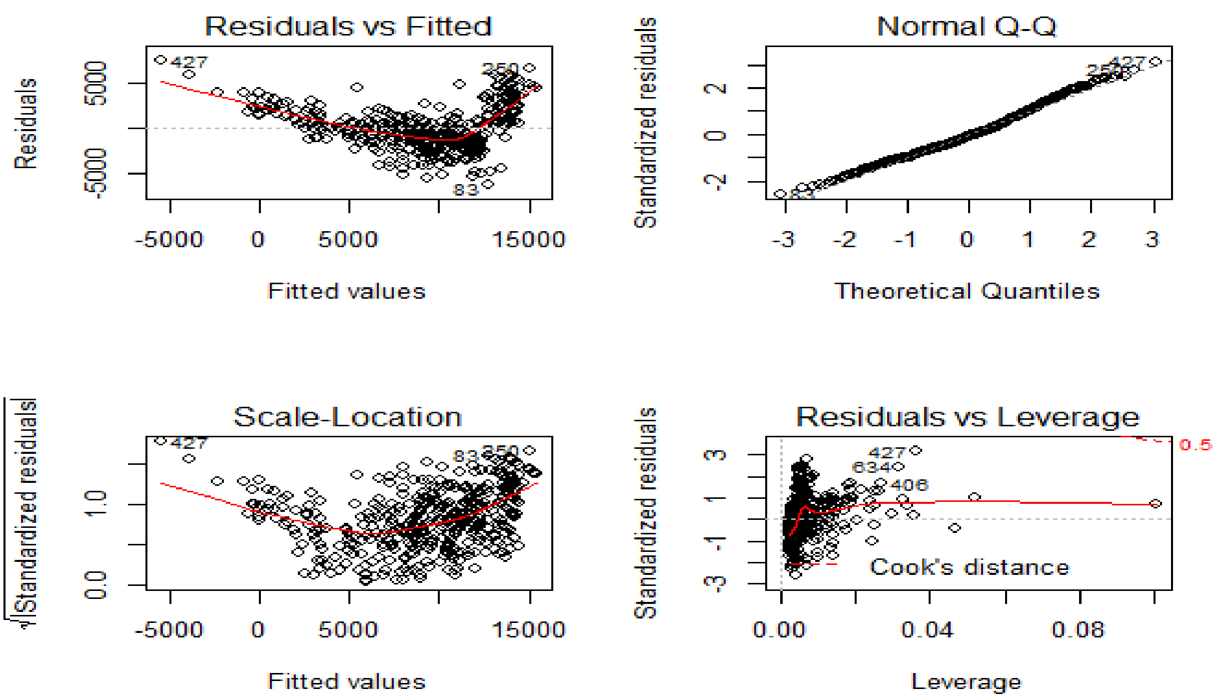
```
fit2=lm(Price~Age, data = train_Ford)
> summary(fit2)

Call:
lm(formula = Price ~ Age, data = train_Ford)

Residuals:
    Min      1Q  Median      3Q     Max 
-6019.0 -1950.8  -557.1  1670.4  8623.5 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14127.22     221.85   63.68   <2e-16 ***
Age         -1113.20      41.28  -26.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2838 on 442 degrees of freedom
Multiple R-squared:  0.622,	Adjusted R-squared:  0.6211 
F-statistic: 727.2 on 1 and 442 DF,  p-value: < 2.2e-16


outlierTest(fit2)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferonni p
1 3.145378          0.0017709       0.7863
> confint(fit2)
                2.5 %   97.5 %
(Intercept) 13691.215 14563.23
Age         -1194.327 -1032.07
```


Price
Age
20000
15000
10000
5000
0
5
10
15


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

Multiple Regression

```
fit3=lm(Price~Age + Mileage, data = train_Ford)
> summary(fit3)

Call:
lm(formula = Price ~ Age + Mileage, data = train_Ford)

Residuals:
   Min     1Q Median     3Q    Max 
 -6231  -1662   -205   1555   7525 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.570e+04  2.282e+02   68.80   <2e-16 ***
Age         -6.637e+02  5.042e+01  -13.16   <2e-16 ***
Mileage     -6.315e-02  5.034e-03  -12.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2439 on 441 degrees of freedom
Multiple R-squared:  0.7214,   Adjusted R-squared:  0.7201 
F-statistic:   571 on 2 and 441 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(fit)
```

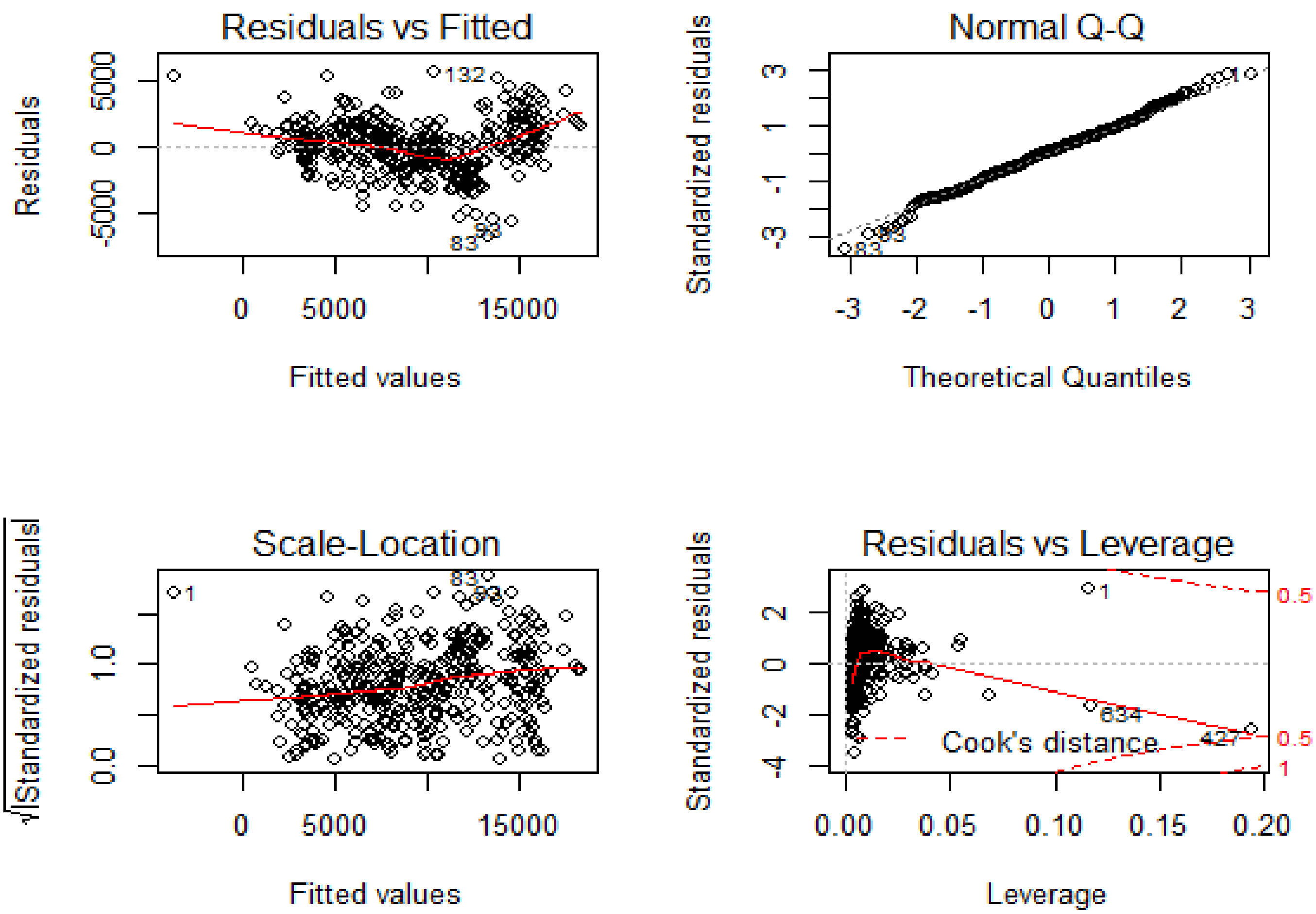


```
confint(fit3)
                    2.5 %        97.5 %
(Intercept)  1.525171e+04  1.614871e+04
Age         -7.628441e+02 -5.646556e+02
Mileage     -7.304136e-02 -5.325528e-02
> vif(fit3)
     Age  Mileage
2.019758 2.019758
> outlierTest(fit3)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
427 3.173831          0.0016099      0.71482
ncvTest(fit3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 15.8425    Df = 1     p = 6.883869e-05
```

Interactions

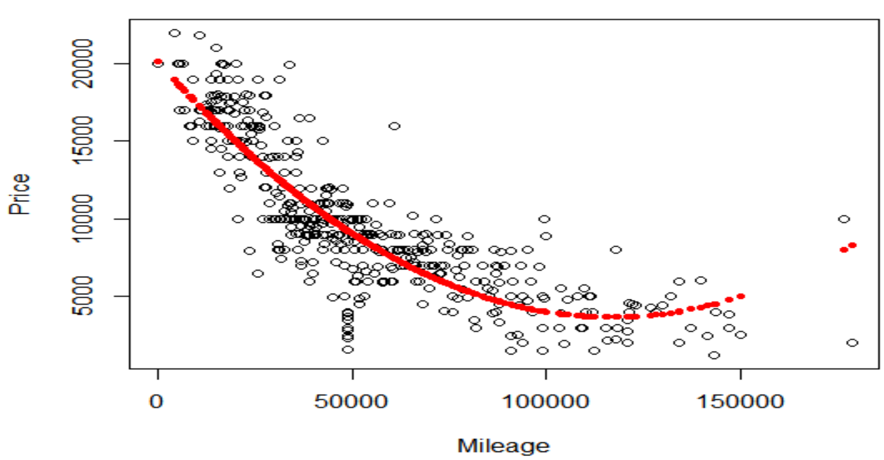
```
fit4 =lm(Price~Age*Mileage, data = train_Ford)
> summary(fit4)

Call:
lm(formula = Price ~ Age * Mileage, data = train_Ford)

Residuals:
    Min      1Q  Median      3Q     Max
-6801.9 -1276.1    79.4  1216.0  5639.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.893e+04  2.783e+02   68.03   <2e-16 ***
Age         -1.456e+03  6.537e+01  -22.28   <2e-16 ***
Mileage     -1.288e-01  5.870e-03  -21.95   <2e-16 ***
Age:Mileage  1.213e-02  7.838e-04   15.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1965 on 440 degrees of freedom
Multiple R-squared:  0.8196,   Adjusted R-squared:  0.8184
F-statistic: 666.3 on 3 and 440 DF,  p-value: < 2.2e-16


vif(fit4)
        Age     Mileage Age:Mileage
   5.230901    4.231971   10.211028
> ncvTest(fit4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 19.64682    Df = 1     p = 9.3158e-06
```

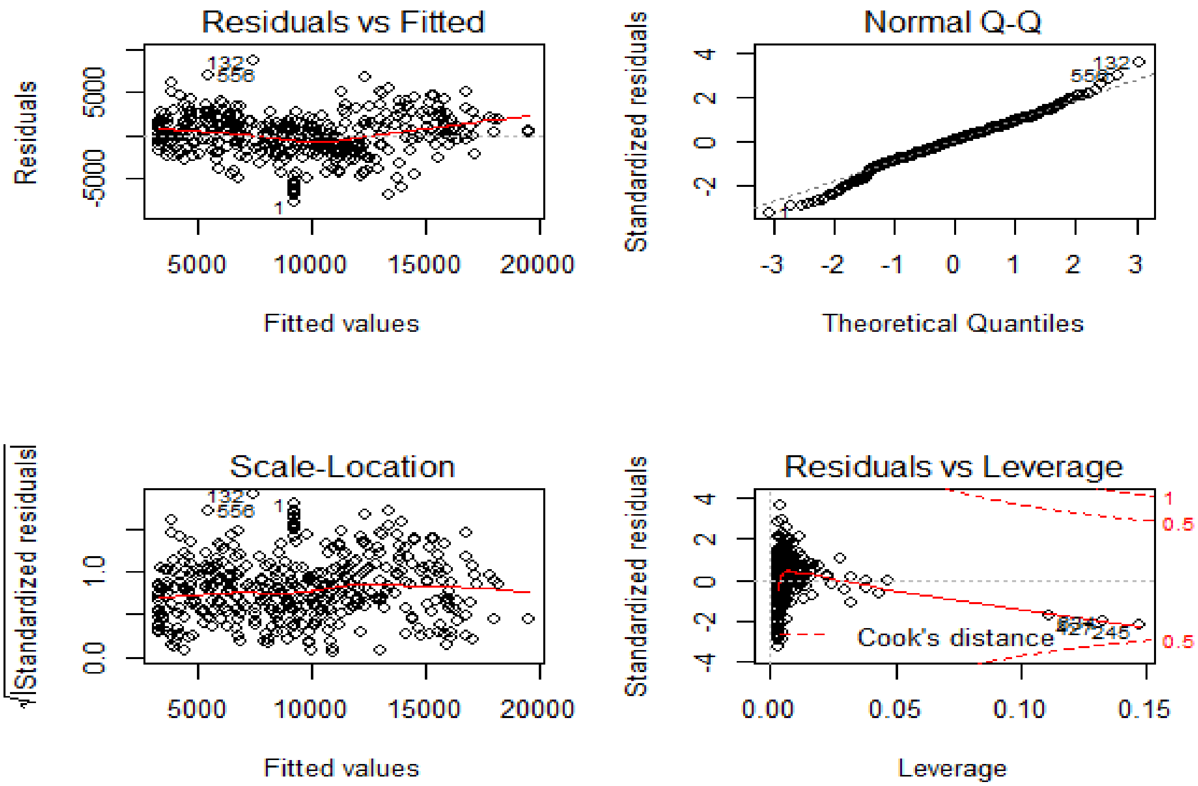
```
outlierTest(fit4)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
83 -3.513209          0.00048865      0.21696
```



Nonlinear terms

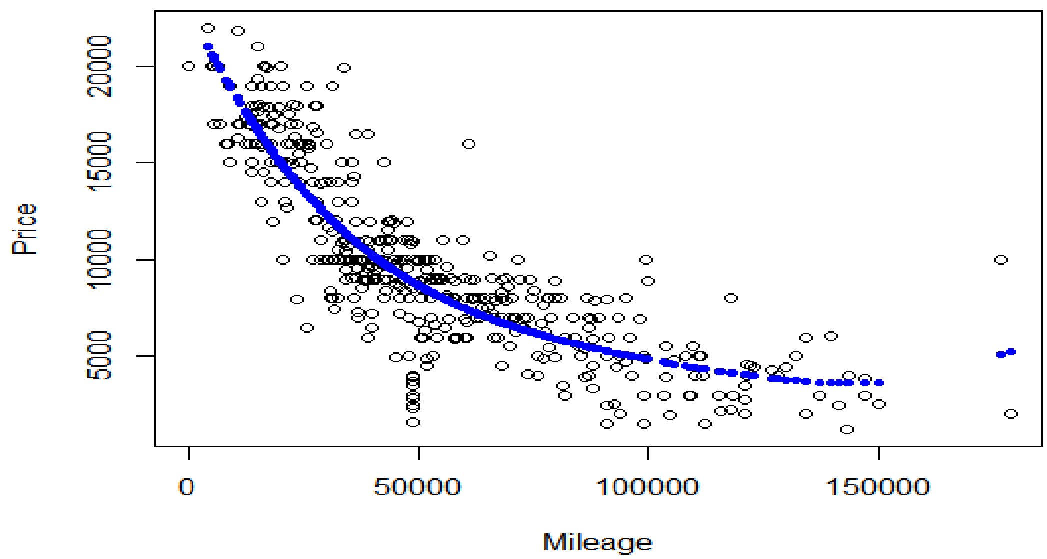
```
summary(fit5)

Call:
lm(formula = Price ~ Mileage + I(Mileage^2), data = train_Ford)

Residuals:
   Min     1Q Median     3Q    Max 
 -7612  -1398    151   1523   8555 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.953e+04  3.597e+02   54.28   <2e-16 ***
Mileage      -2.629e-01  1.121e-02  -23.46   <2e-16 ***
I(Mileage^2)  1.064e-06  7.426e-08   14.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2379 on 441 degrees of freedom
Multiple R-squared:  0.7351,	Adjusted R-squared:  0.7339 
F-statistic:   612 on 2 and 441 DF,  p-value: < 2.2e-16

outlierTest(fit5)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
132 3.653701         0.00028966      0.12861
vif(fit5)
     Mileage I(Mileage^2) 
    10.53042     10.53042 
> ncvTest(fit5)
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 3.974273    Df = 1     p = 0.0462004 
```

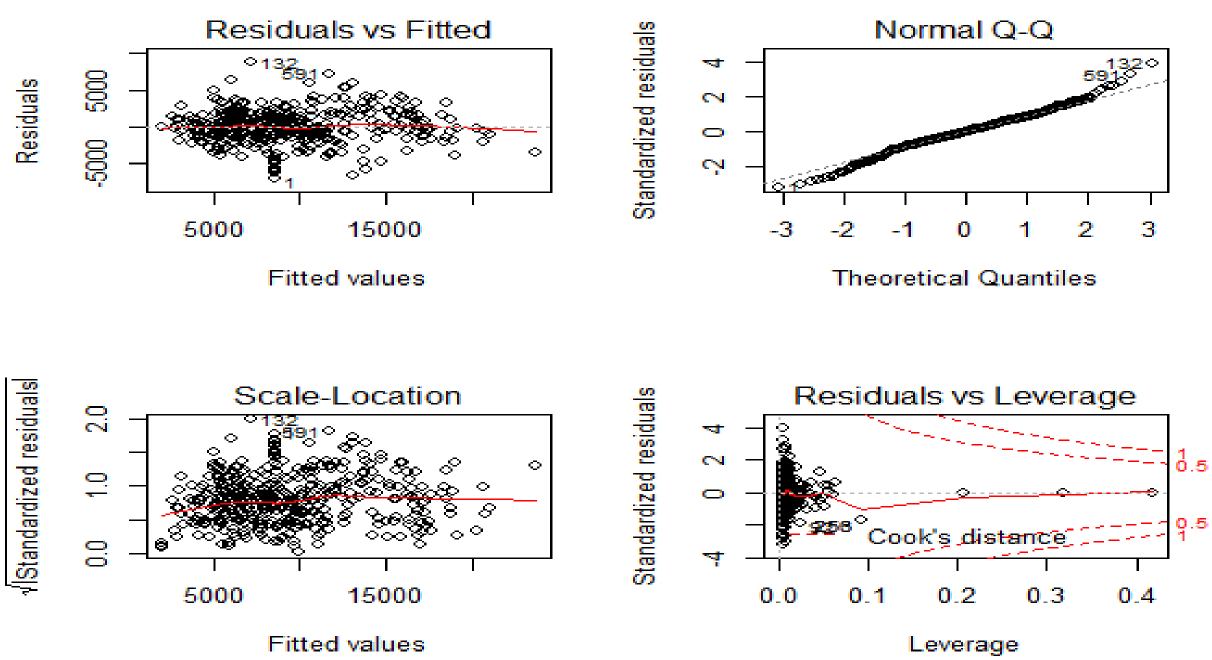





```
fit6=lm(Price~poly(Mileage,4), data = train_Ford)
> summary(fit6)

Call:
lm(formula = Price ~ poly(Mileage, 4), data = train_Ford)

Residuals:
    Min      1Q  Median      3Q     Max 
-6964.7 -1256.1    10.3  1431.5  8764.0 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         9373.6      105.3   89.05  < 2e-16 ***
poly(Mileage, 4)1 -75919.2     2218.1  -34.23  < 2e-16 ***
poly(Mileage, 4)2  34066.3     2218.1   15.36  < 2e-16 ***
poly(Mileage, 4)3 -17234.5     2218.1   -7.77 5.61e-14 ***
poly(Mileage, 4)4   6166.2     2218.1    2.78  0.00567 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2218 on 439 degrees of freedom
Multiple R-squared:  0.7707,   Adjusted R-squared:  0.7686 
F-statistic: 368.9 on 4 and 439 DF,  p-value: < 2.2e-16
```

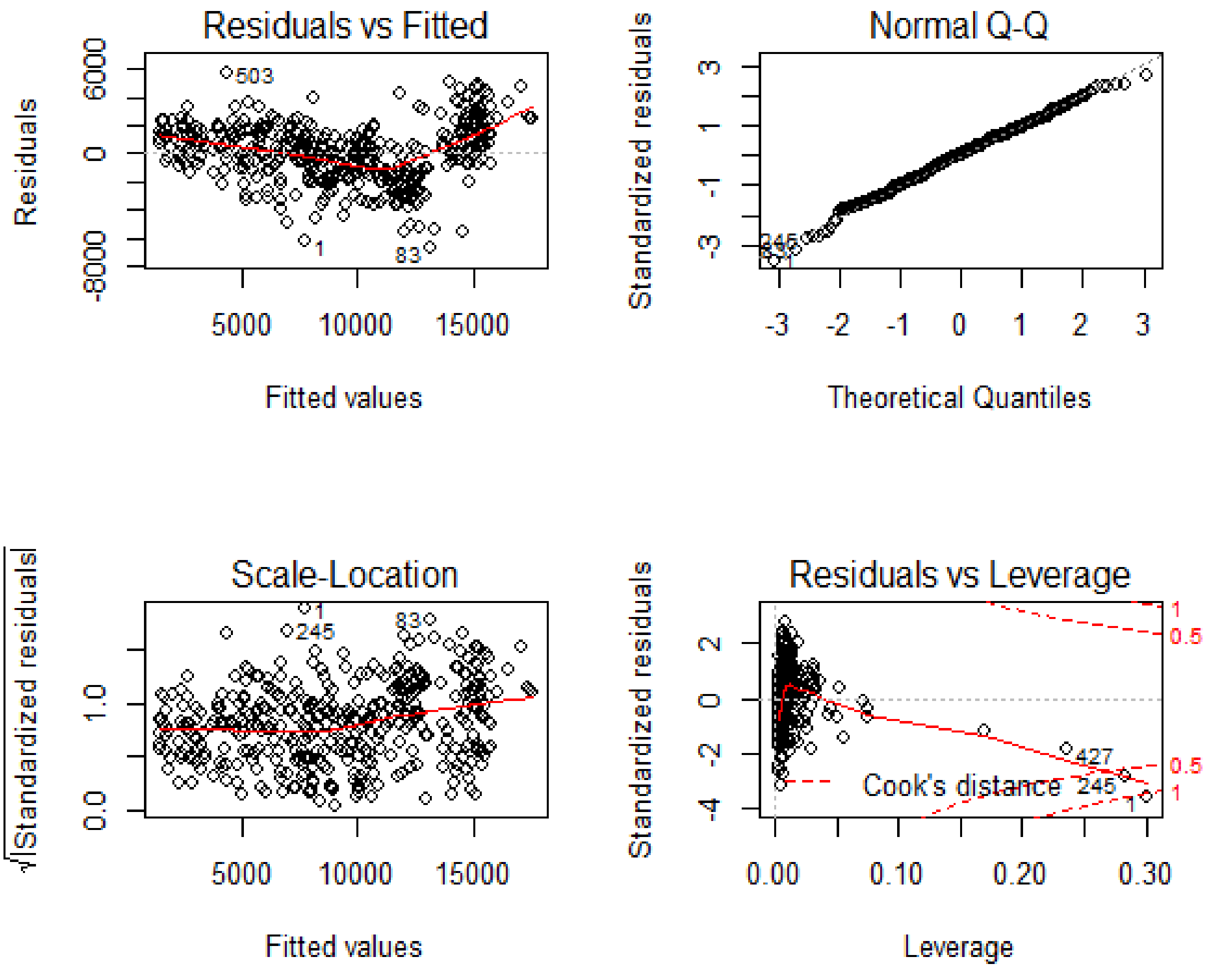
```
ncvTest(fit6)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.764245    Df = 1     p = 0.01635551
> outlierTest(fit6)
    rstudent unadjusted p-value Bonferonni p
132 4.028644         6.6134e-05     0.029363
```





Final Model

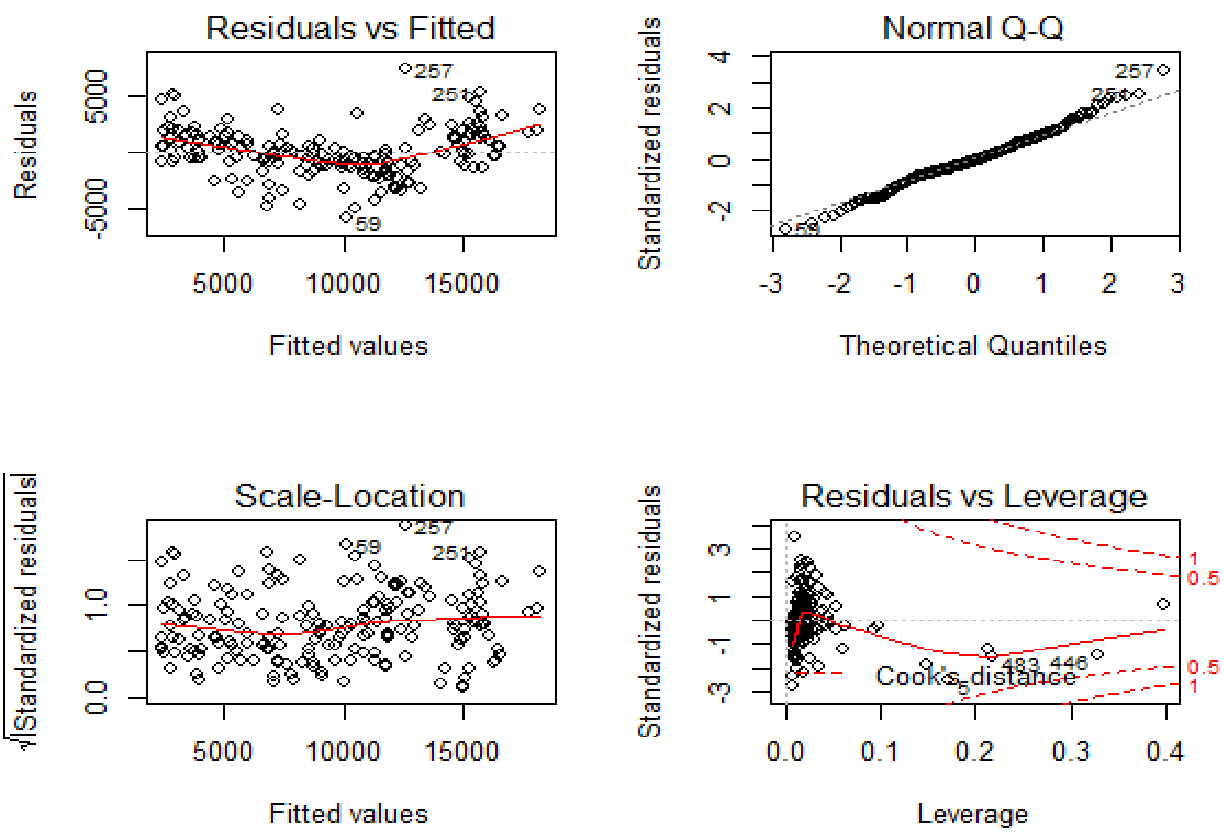
```
fit7=lm(Price~Mileage + Age + I(Age^2) + I(Mileage^4), data = train_Ford[-c(4
27,83,132,590)])
> summary(fit7)

Call:
lm(formula = Price ~ Mileage + Age + I(Age^2) + I(Mileage^4),
    data = train_Ford[-c(427, 83, 132, 590)])

Residuals:
    Min      1Q  Median      3Q     Max
-6617.5 -1478.5   111.5  1393.9  5635.5

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.785e+04  2.571e+02  69.436  < 2e-16 ***
Mileage      -7.840e-02  6.823e-03 -11.491  < 2e-16 ***
Age          -1.470e+03  1.238e+02 -11.876  < 2e-16 ***
I(Age^2)      5.982e+01  7.945e+00   7.530 2.91e-13 ***
I(Mileage^4)  9.854e-18  1.494e-18   6.596 1.22e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2084 on 439 degrees of freedom
Multiple R-squared:  0.7975,   Adjusted R-squared:  0.7957
F-statistic: 432.3 on 4 and 439 DF,  p-value: < 2.2e-16


vif(fit7)
     Mileage          Age     I(Age^2) I(Mileage^4)
    5.082076    16.683139    12.356260     2.826909
> outlierTest(fit7)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
1 -3.566284         0.00040198      0.17848
```



Residuals vs Fitted
Residuals
Fitted values
503
1
83
6000
0
-8000
5000
10000
15000
Normal Q-Q
Standardized residuals
Theoretical Quantiles
245
-3
-2
-1
0
1
2
3
Scale-Location
√|Standardized residuals|
Fitted values
1
245
83
0.0
1.0
5000
10000
15000
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
427
245
1
0.5
2
0
-2
-4
0.00
0.10
0.20
0.30

Testing the Model

```
fit9=lm(Price~Mileage + Age + I(Age^2) + I(Mileage^4), data = test_Ford)
> summary(fit9)

Call:
lm(formula = Price ~ Mileage + Age + I(Age^2) + I(Mileage^4),
    data = test_Ford)

Residuals:
    Min      1Q  Median      3Q     Max
-5784.2 -1176.5  -139.2  1341.9  7349.1

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.856e+04  3.891e+02  47.704  < 2e-16 ***
Mileage      -9.444e-02  1.136e-02  -8.315 1.89e-14 ***
Age          -1.537e+03  2.123e+02  -7.241 1.14e-11 ***
I(Age^2)      7.302e+01  1.390e+01   5.251 4.10e-07 ***
I(Mileage^4)  1.559e-17  2.285e-18   6.822 1.23e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2130 on 186 degrees of freedom
Multiple R-squared:  0.8117,   Adjusted R-squared:  0.8077
F-statistic: 200.5 on 4 and 186 DF,  p-value: < 2.2e-16


ncvTest(fit9)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.58147    Df = 1     p = 0.208549
> vif(fit9)
     Mileage          Age     I(Age^2) I(Mileage^4)
    6.915764    20.957953    14.464315     3.365227
> plot(fit9)
> outlierTest(fit9)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
257 3.571813         0.00045197     0.086327
```


Residuals vs Fitted
Residuals
Fitted values
257
251
59
Normal Q-Q
Standardized residuals
Theoretical Quantiles
257
Scale-Location
√|Standardized residuals|
Fitted values
257
59
251
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
483
446
1
0.5
0.5
1

Residual Analysis

Residual analysis is usually done graphically or using basic library in R.

- **Outlier**

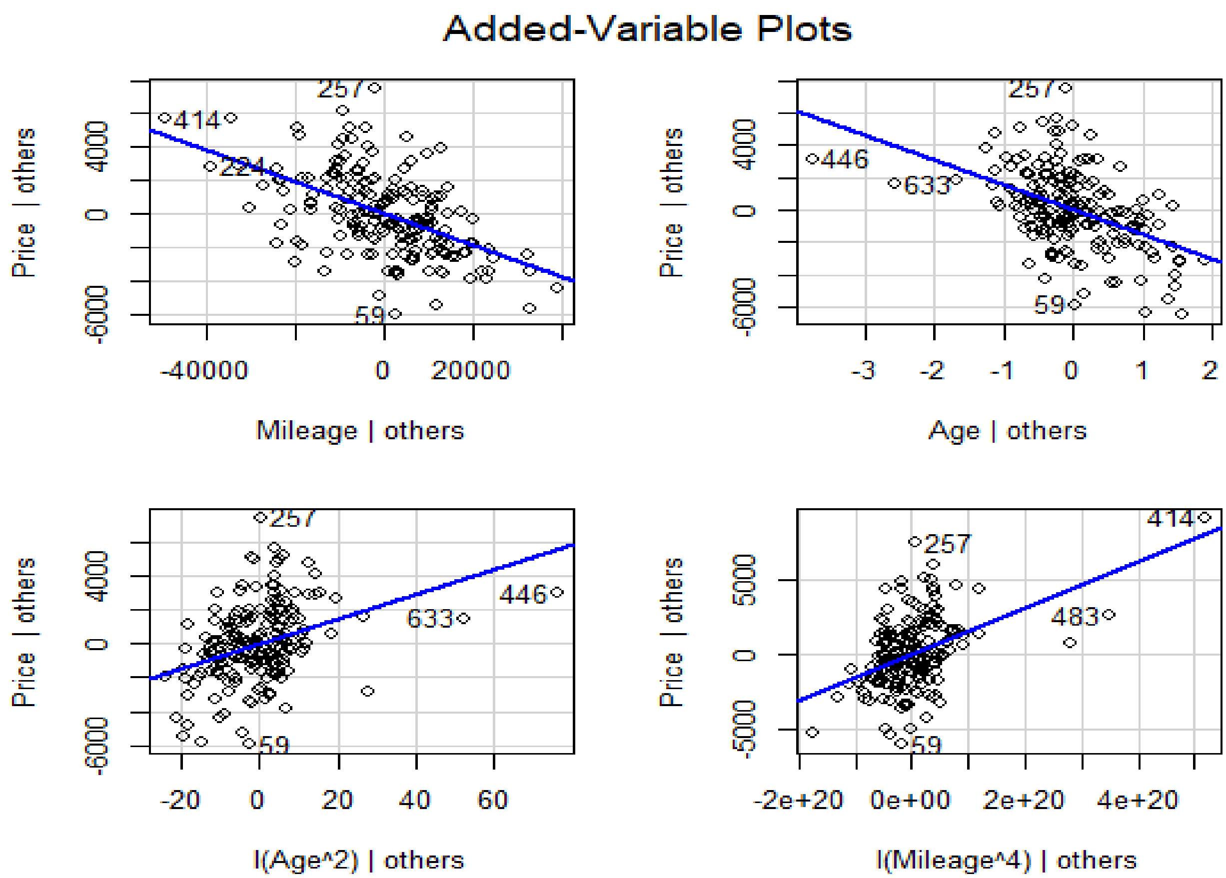
```
> outlierTest(fit9)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
257 3.571813         0.00045197      0.086327
```

From the outlier test, it shows that 257 is an outlier, so we can decide to remove it to improve on our model.

- **Influential observations**

This can be checked using added variable plots in R using the car package.

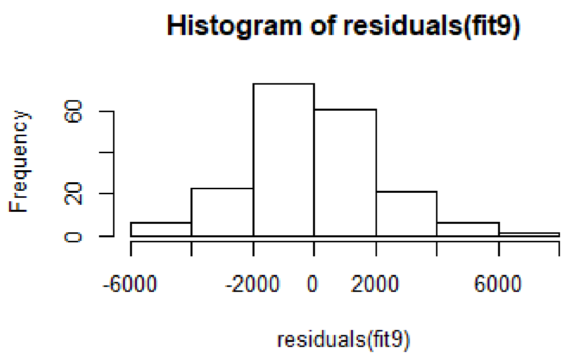
```
avPlots(fit9, id.n =2, id.cex=0.7)
```

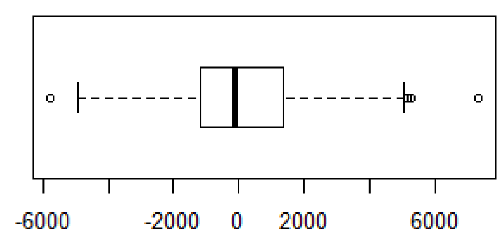


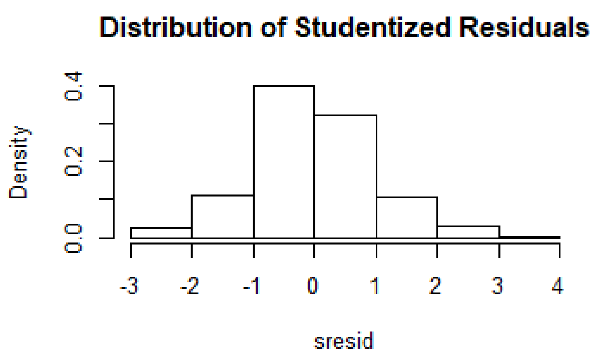
- **Non-Normality**

```
hist(residuals(fit9))
shapiro.test(residuals(fit9))
boxplot(residuals(fit9), horizontal = T)
sreid=studres(fit9)
hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals")
```

```
shapiro.test(residuals(fit9))

        Shapiro-Wilk normality test

data:  residuals(fit9)
W = 0.98979, p-value = 0.1913
```







- **Non-Constant Error Variance**

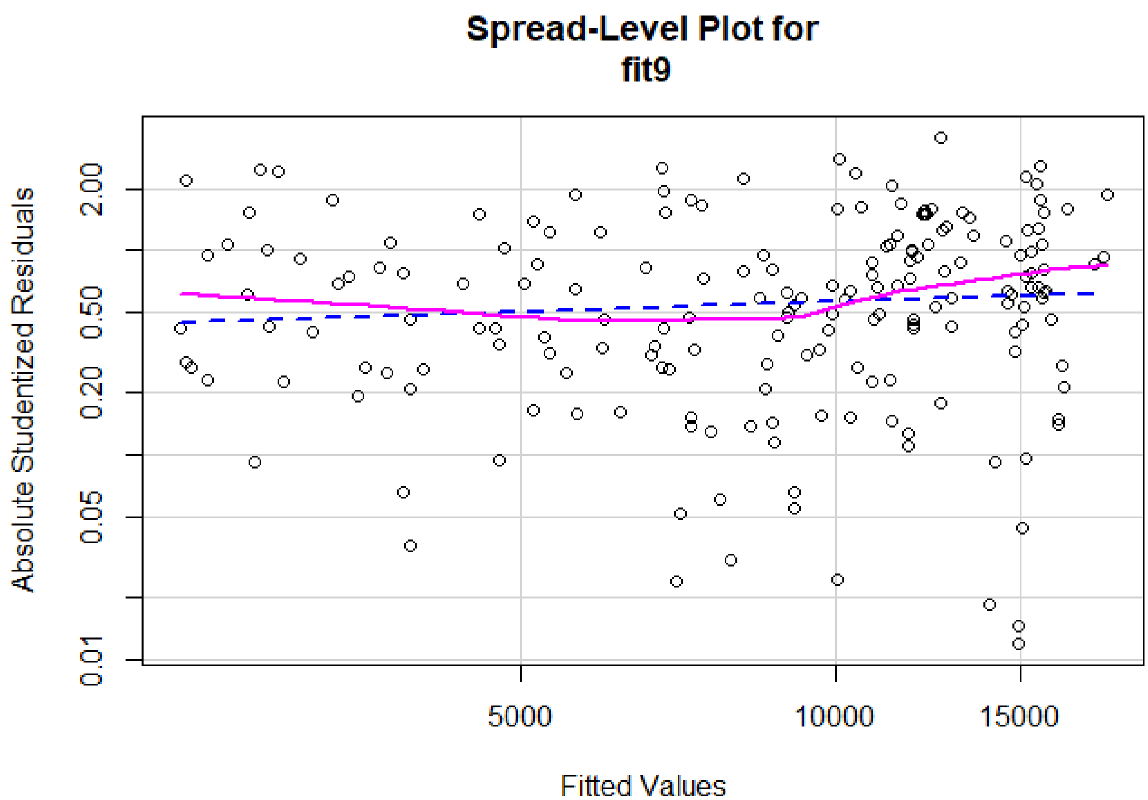
This can be done with the car library in R or graphically.

```
ncvTest(fit9)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.58147    Df = 1     p = 0.208549
```

```
spreadLevelPlot(fit9)
```

```
Suggested power transformation:  0.8355619
```



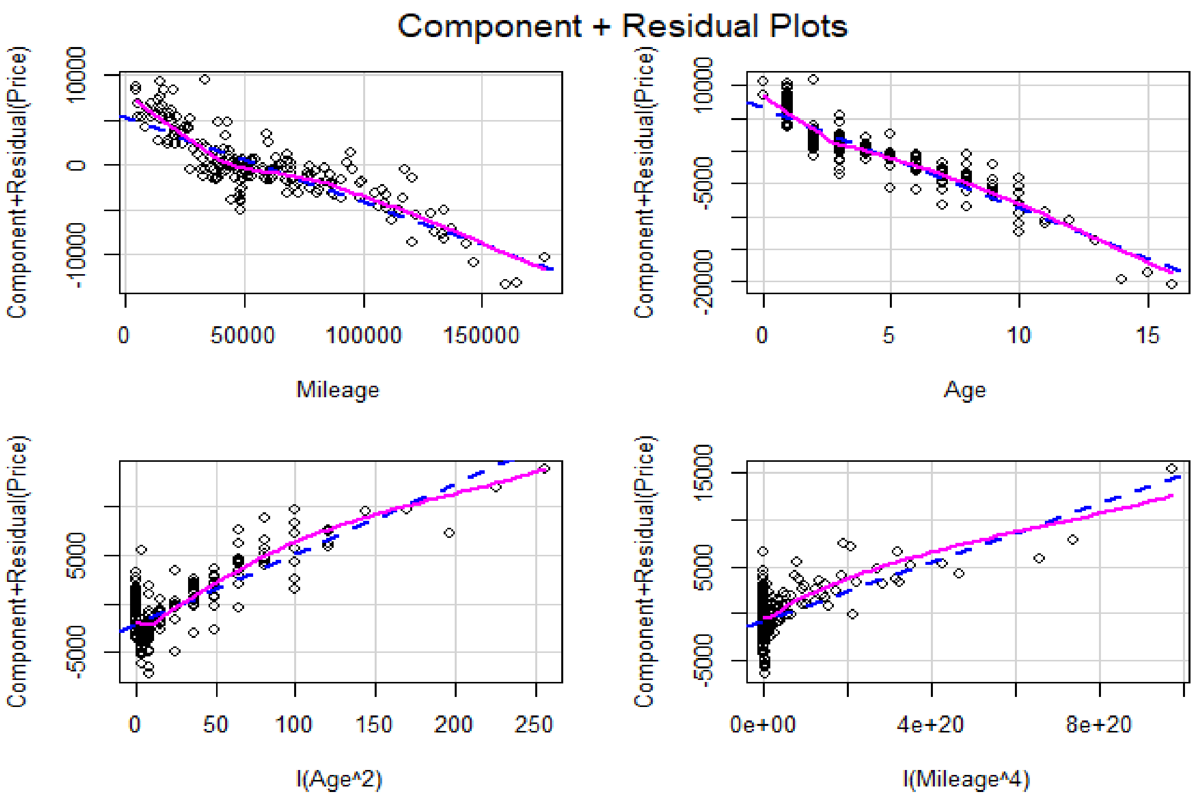
- **Multi-Collinearity**

```
vif(fit9)
     Mileage           Age      I(Age^2) I(Mileage^4)
    6.915764     20.957953     14.464315      3.365227
sqrt(vif(fit9))
 Mileage           Age      I(Age^2) I(Mileage^4)
2.629784      4.577986      3.803198      1.834455
```

- **Non-Linearity**

This can be determine using **crPlots** and **ceresplots** in R using the car package.

```
> crPlots(fit9)
```



- **Non-Independence of Errors**

```
durbinWatsonTest(fit9)
 lag Autocorrelation D-W Statistic p-value
   1        0.225358      1.515181       0
 Alternative hypothesis: rho != 0
```

- **Analysis of Variance**
- `anova(fit9)`
- `Analysis of Variance Table`
-
- `Response: Price`

```
              Df     Sum Sq    Mean Sq F value    Pr(>F)    
Mileage        1 2715016257 2715016257 598.364 < 2.2e-16 ***
Age            1  329821215  329821215  72.689 5.162e-15 ***
I(Age^2)       1  382938764  382938764  84.396 < 2.2e-16 ***
I(Mileage^4)   1  211138433  211138433  46.533 1.227e-10 ***
Residuals    186  843956112    4537398                      
---
```

Conclusion

Hence the best model for predicting the Price for used Ford cars in the United State is a polynomial Model of degree.