Chapter 3 exercises (ISLR)

Chapter 3: linear regression

Applied exercises

Exercise 9

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.6.3
```

```
?Auto
```

```
## starting httpd help server ... done
```

```
head(Auto)
```

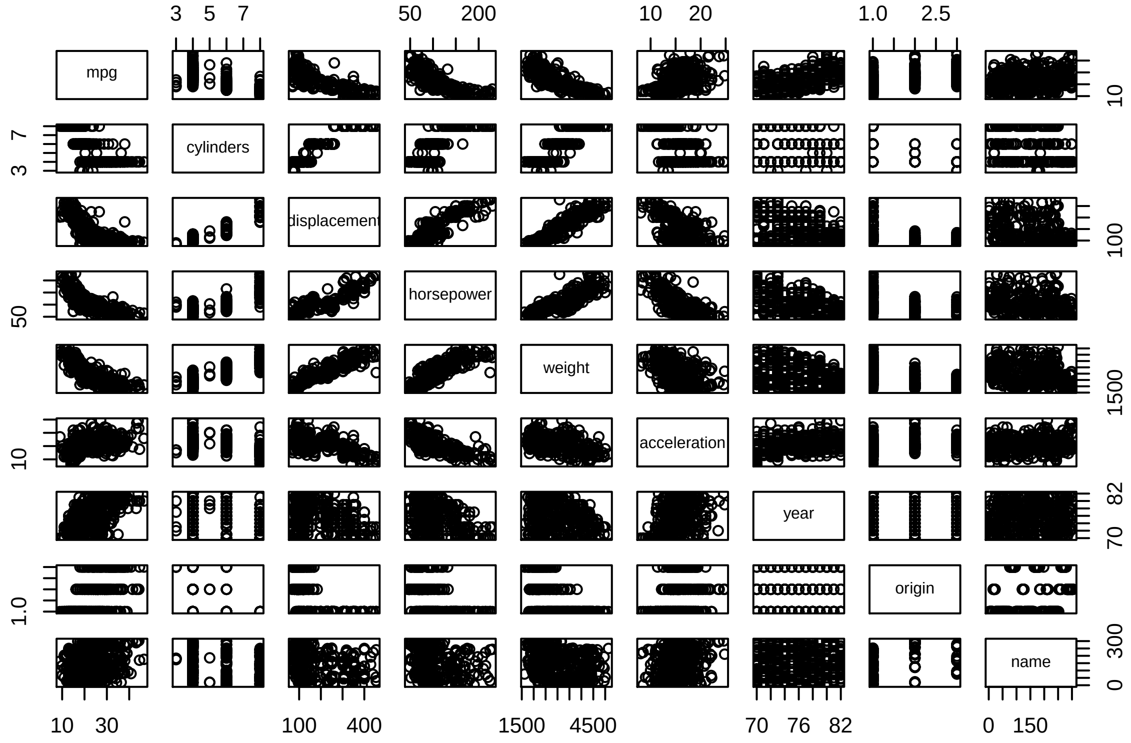
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
summary(Auto)
```

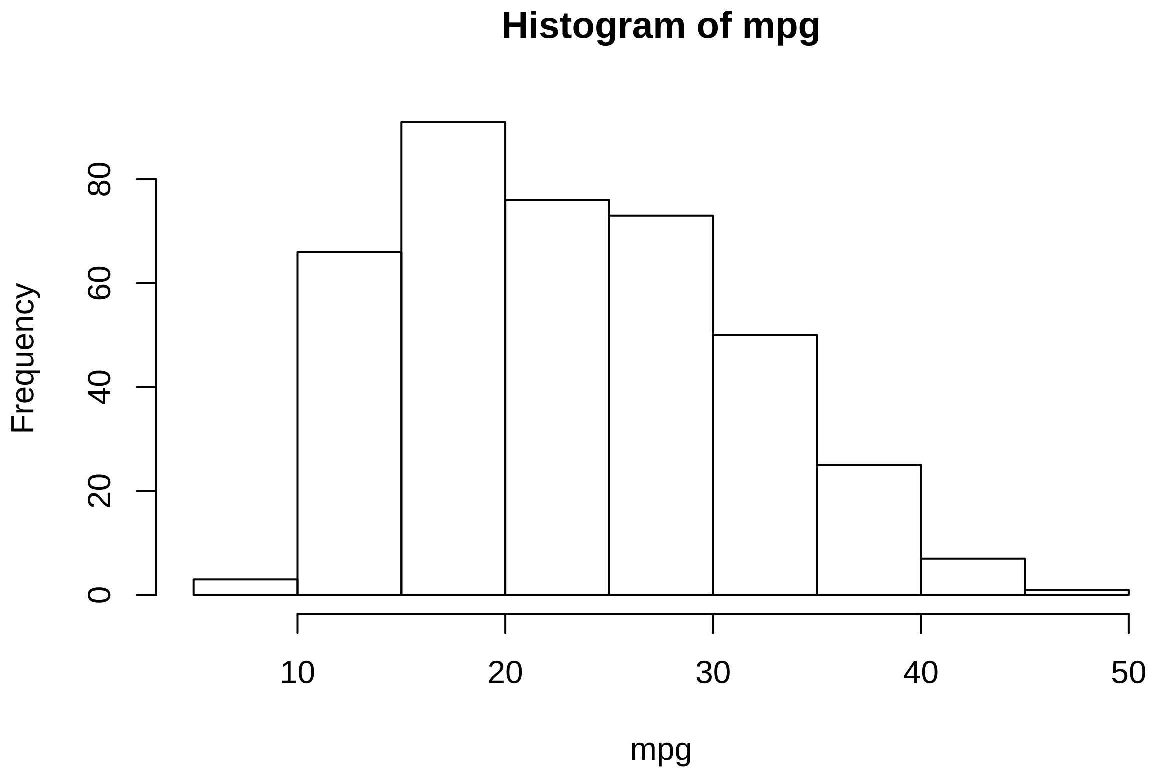
```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140  
##                                                                                
##   acceleration        year           origin                      name    
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5  
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5  
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5  
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4  
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4  
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4  
##                                                  (Other)           :365  
```

```r
attach(Auto)
```

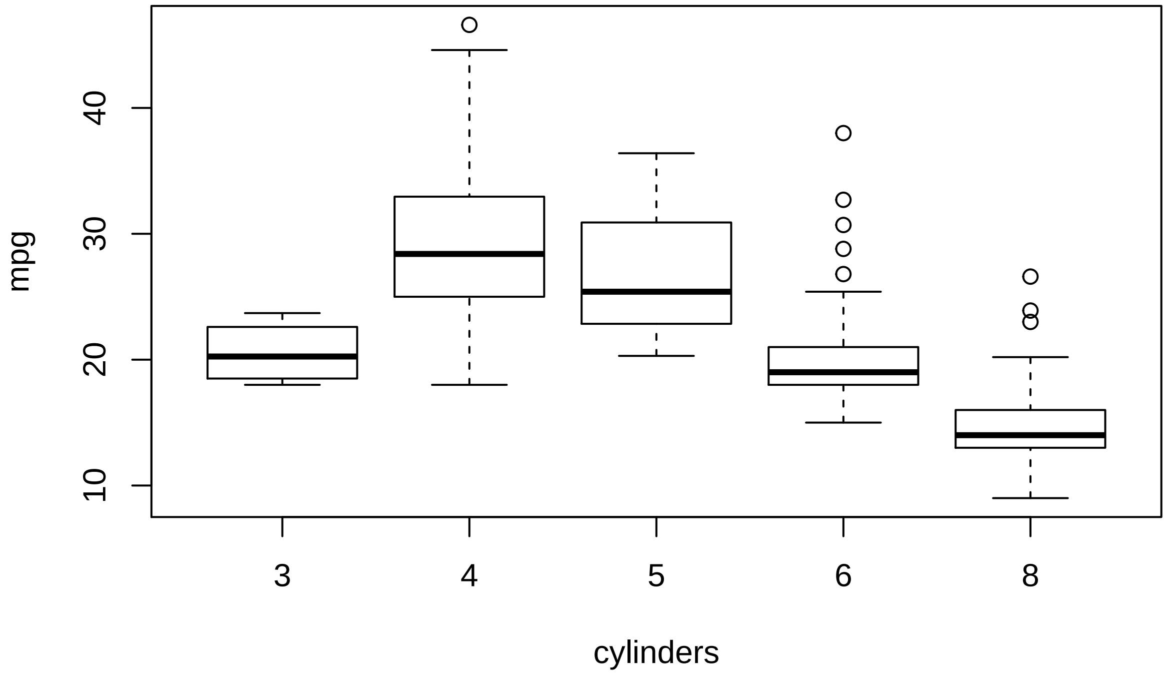
```r
pairs(Auto)
```



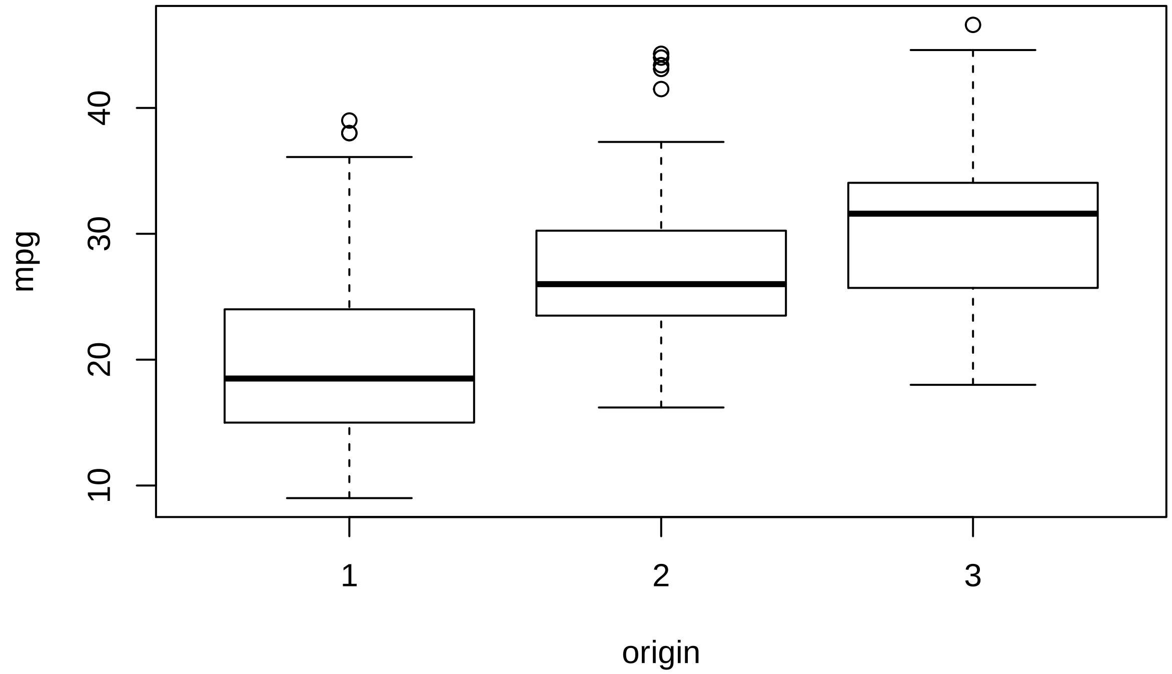
```r
hist(mpg)
```



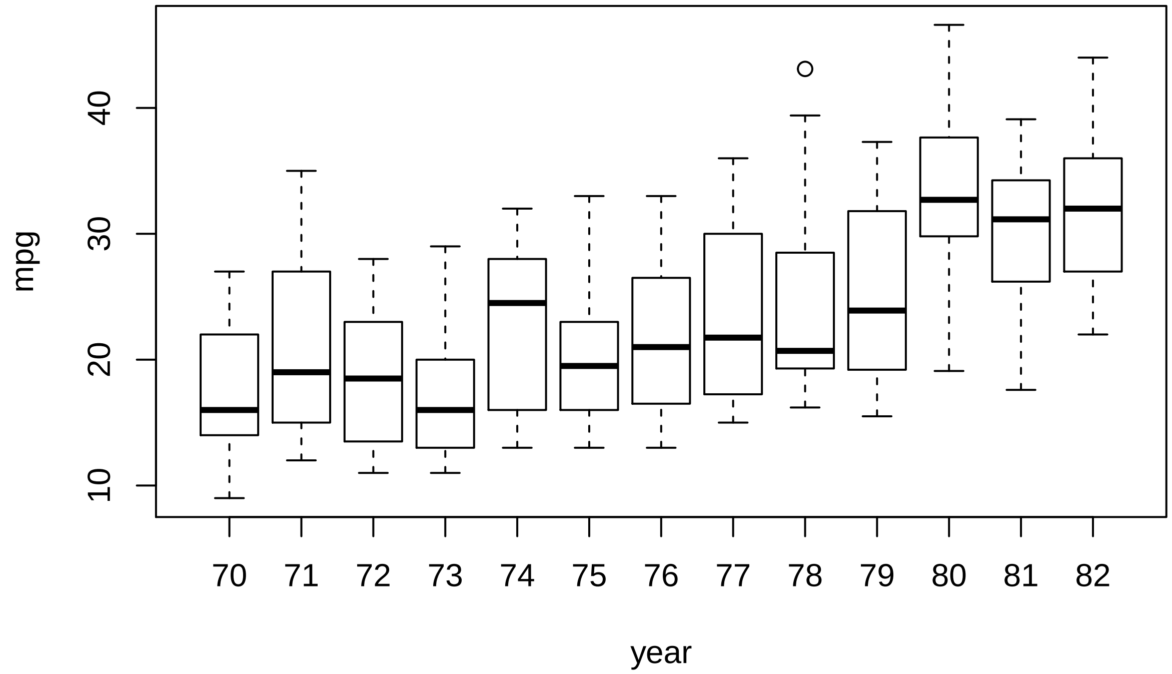
```
boxplot(mpg~cylinders)
```



```
boxplot(mpg~origin)
```



```
boxplot(mpg~year)
```



```r
# Matrix of correlation
cor(subset(Auto, select = -name))
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

There's a high correlation among the first 4 variables (mpg excluded).

Let's now build a linear model.

```r
# first transforming origin into a categorical
Auto$origin <- factor(Auto$origin, labels = c("American", "European", "Japanese"))
```

```r
fit <- lm(mpg~., data=subset(Auto, select = -name))
summary(fit)
```

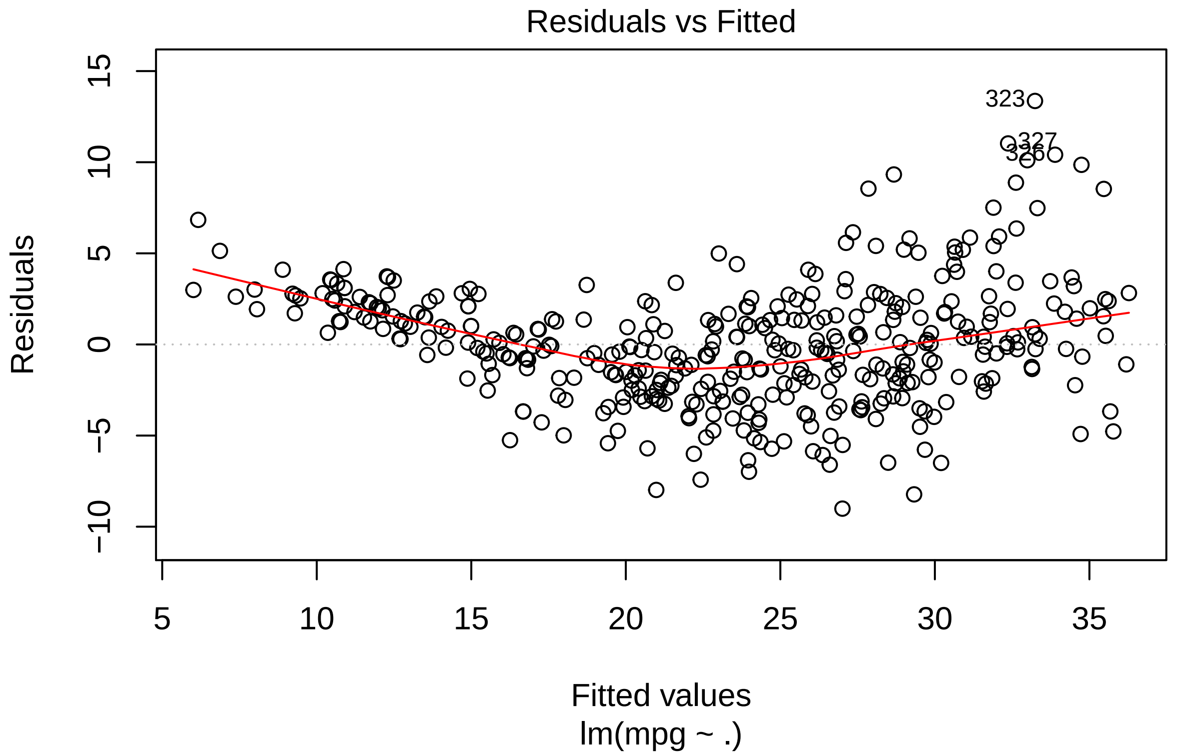
```
##
## Call:
## lm(formula = mpg ~ ., data = subset(Auto, select = -name))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders      -4.897e-01  3.212e-01  -1.524 0.128215
## displacement    2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower     -1.818e-02  1.371e-02  -1.326 0.185488
## weight         -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration    7.910e-02  9.822e-02   0.805 0.421101
## year            7.770e-01  5.178e-02  15.005  < 2e-16 ***
## originEuropean  2.630e+00  5.664e-01   4.643 4.72e-06 ***
## originJapanese  2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

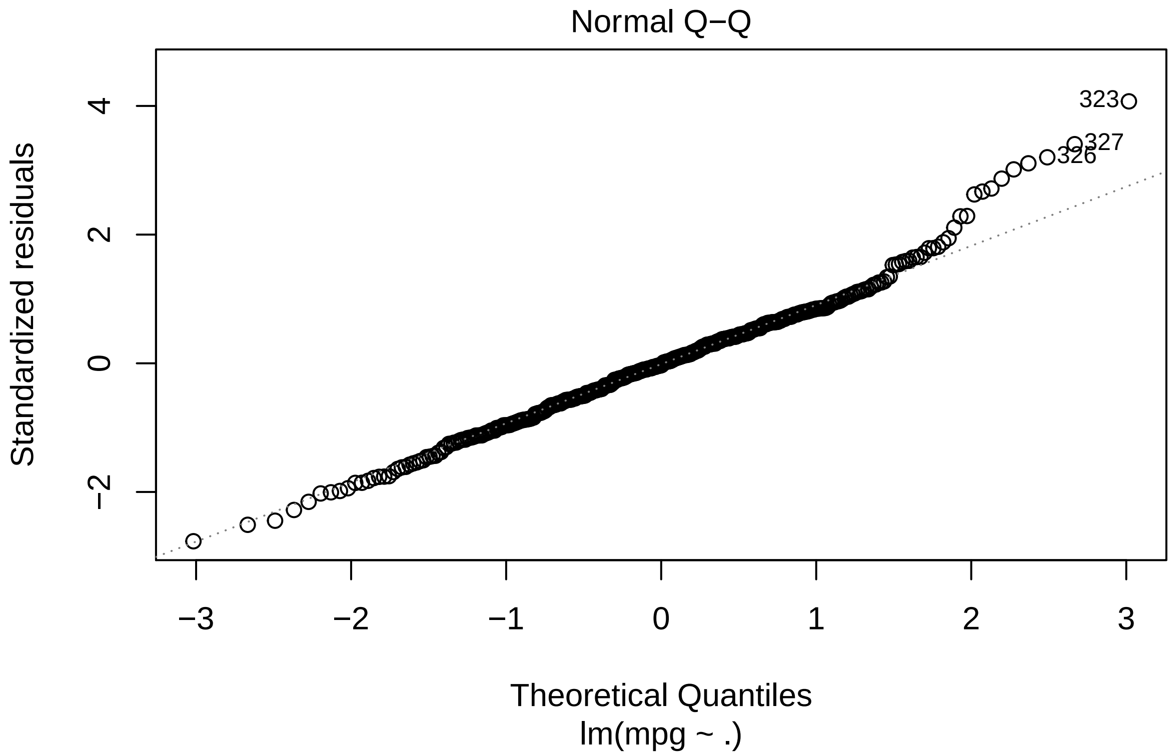
i. Is there a relationship between the predictors and the response? Answer: yes, the F-statistic is telling us that at least one variable in the model has a relationship with the response with an almost 100% confidence.

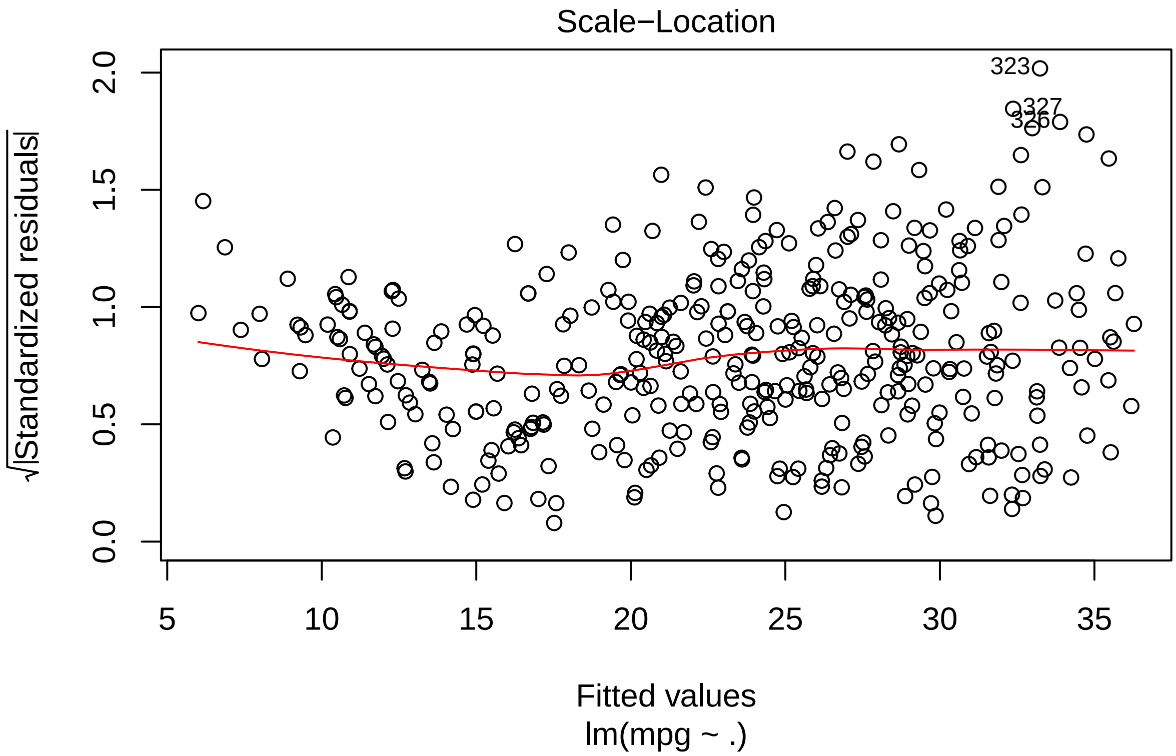
ii. Which predictors appear to have a statistically significant relationship to the response? Answer: Displacement, weight, year and origin. However, we have to take into consideration that there might be multi-collinearity in the model.

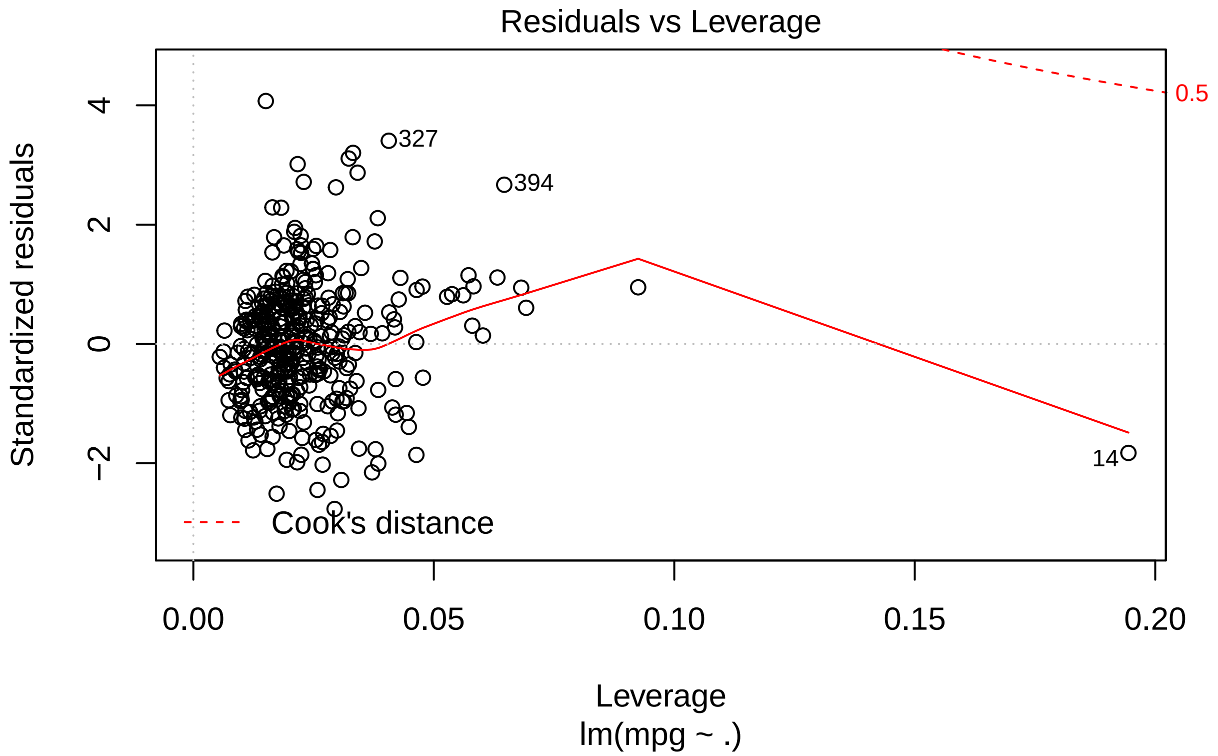
iii. What does the coefficient for the year variable suggest? Answer: that on average, leaving the rest fixed, each year the mpg increases by 0.75.

```r
# Diagnostic plots!
plot(fit)
```


Residuals vs Fitted
Residuals
15
10
5
0
−5
−10
5
10
15
20
25
30
35
Fitted values
lm(mpg ~ .)
323
327
326


Normal Q−Q
Standardized residuals
4
2
0
−2
323
327
326
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ .)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
30
35
Fitted values
lm(mpg ~ .)
323
327
326



The Tukey-Anscombe plot seem to be suggesting a missing quadratic term in the model.

The Q-Q plot seems to confirm the Normality hypothesis for the errors.

The Standardized residuals vs fitted values and leverage seem to suggest the presence of a few outliers (323,327,…) and a single high leverage point, which is not an outlier (14).

Let's model the interactions now.

```
fit.2 <- lm(mpg~.*., data=subset(Auto, select = -name))
summary(fit.2)
```

```
##
## Call:
## lm(formula = mpg ~ . * ., data = subset(Auto, select = -name))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6008 -1.2863  0.0813  1.2082 12.0382
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.401e+01  5.147e+01   0.855 0.393048
## cylinders                  3.302e+00  8.187e+00   0.403 0.686976
## displacement              -3.529e-01  1.974e-01  -1.788 0.074638 .
## horsepower                 5.312e-01  3.390e-01   1.567 0.117970
## weight                    -3.259e-03  1.820e-02  -0.179 0.857980
## acceleration              -6.048e+00  2.147e+00  -2.818 0.005109 **
## year                       4.833e-01  5.923e-01   0.816 0.415119
## originEuropean              -3.517e+01  1.260e+01  -2.790 0.005547 **
## originJapanese              -3.765e+01  1.426e+01  -2.640 0.008661 **
## cylinders:displacement      -6.316e-03  7.106e-03  -0.889 0.374707
## cylinders:horsepower         1.452e-02  2.457e-02   0.591 0.555109
## cylinders:weight             5.703e-04  9.044e-04   0.631 0.528709
## cylinders:acceleration       3.658e-01  1.671e-01   2.189 0.029261 *
## cylinders:year              -1.447e-01  9.652e-02  -1.499 0.134846
## cylinders:originEuropean    -7.210e-01  1.088e+00  -0.662 0.508100
## cylinders:originJapanese     1.226e+00  1.007e+00   1.217 0.224379
## displacement:horsepower     -5.407e-05  2.861e-04  -0.189 0.850212
## displacement:weight          2.659e-05  1.455e-05   1.828 0.068435 .
## displacement:acceleration   -2.547e-03  3.356e-03  -0.759 0.448415
## displacement:year            4.547e-03  2.446e-03   1.859 0.063842 .
## displacement:originEuropean -3.364e-02  4.220e-02  -0.797 0.425902
## displacement:originJapanese  5.375e-02  4.145e-02   1.297 0.195527
## horsepower:weight           -3.407e-05  2.955e-05  -1.153 0.249743
## horsepower:acceleration     -3.445e-03  3.937e-03  -0.875 0.382122
## horsepower:year             -6.427e-03  3.891e-03  -1.652 0.099487 .
## horsepower:originEuropean   -4.869e-03  5.061e-02  -0.096 0.923408
## horsepower:originJapanese    2.289e-02  6.252e-02   0.366 0.714533
## weight:acceleration         -6.851e-05  2.385e-04  -0.287 0.774061
## weight:year                 -8.065e-05  2.184e-04  -0.369 0.712223
## weight:originEuropean        2.277e-03  2.685e-03   0.848 0.397037
## weight:originJapanese       -4.498e-03  3.481e-03  -1.292 0.197101
## acceleration:year            6.141e-02  2.547e-02   2.412 0.016390 *
## acceleration:originEuropean  9.234e-01  2.641e-01   3.496 0.000531 ***
## acceleration:originJapanese  7.159e-01  3.258e-01   2.198 0.028614 *
## year:originEuropean          2.932e-01  1.444e-01   2.031 0.043005 *
## year:originJapanese          3.139e-01  1.483e-01   2.116 0.035034 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.628 on 356 degrees of freedom
## Multiple R-squared:  0.8967, Adjusted R-squared:  0.8866
## F-statistic: 88.34 on 35 and 356 DF,  p-value: < 2.2e-16
```

Note that the interactions have taken over most of the statistical significance of the model. The R-squared has increased by 5 points, hence we conclude that the model with the interactions is significantly better than the model without.

Now let's try omitting some possibly collinear variables.

```r
fit.3 <- lm(mpg~ origin + weight*horsepower + year*acceleration, data=subset(Auto, select = -name))
summary(fit.3)
```

```
##
## Call:
## lm(formula = mpg ~ origin + weight * horsepower + year * acceleration,
##     data = subset(Auto, select = -name))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1081 -1.5461 -0.1396  1.2778 11.5168
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.097e+02  1.761e+01   6.231 1.23e-09 ***
## originEuropean     1.267e+00  4.380e-01   2.893  0.00403 **
## originJapanese     1.393e+00  4.453e-01   3.128  0.00189 **
## weight            -1.049e-02  6.193e-04 -16.932  < 2e-16 ***
## horsepower        -2.354e-01  2.246e-02 -10.483  < 2e-16 ***
## year              -6.167e-01  2.271e-01  -2.716  0.00691 **
## acceleration      -6.868e+00  1.083e+00  -6.340 6.45e-10 ***
## weight:horsepower  5.373e-05  4.887e-06  10.994  < 2e-16 ***
## year:acceleration  8.819e-02  1.411e-02   6.251 1.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.788 on 383 degrees of freedom
## Multiple R-squared:  0.875,  Adjusted R-squared:  0.8724
## F-statistic: 335.2 on 8 and 383 DF,  p-value: < 2.2e-16
```

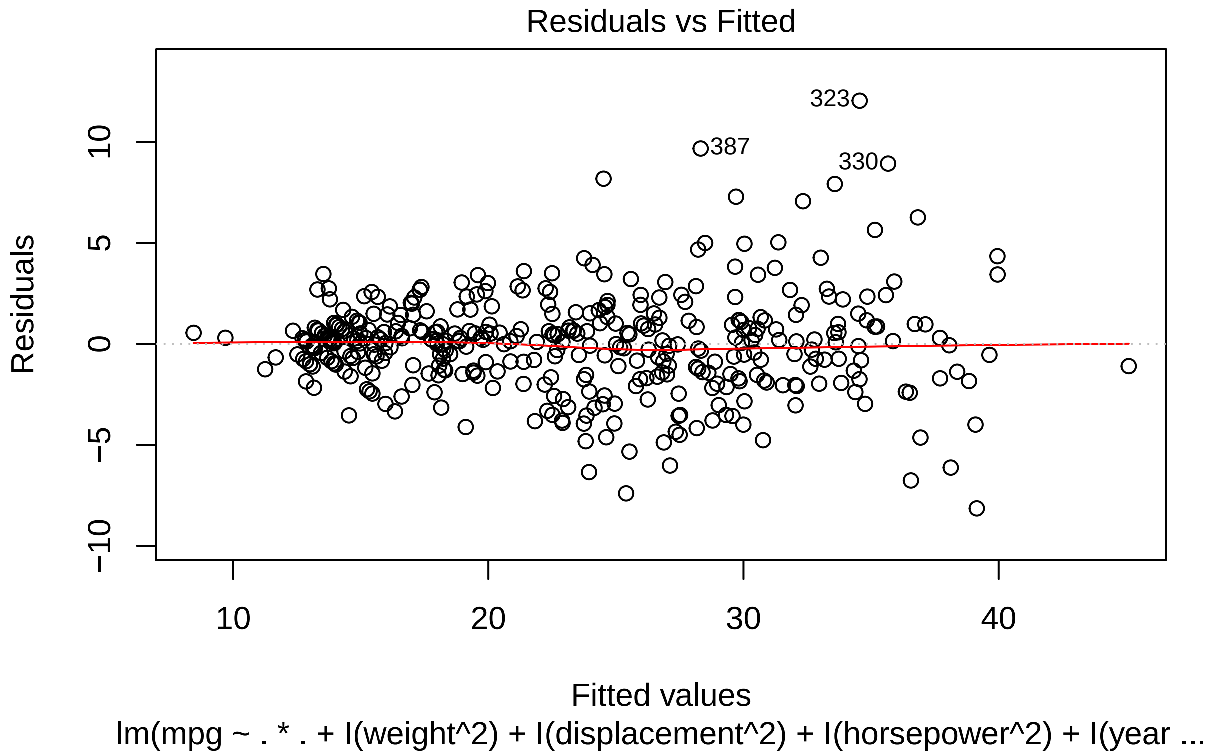
Although all the variables are statistically significant now, we don't get an increase in the adjusted R-squared.

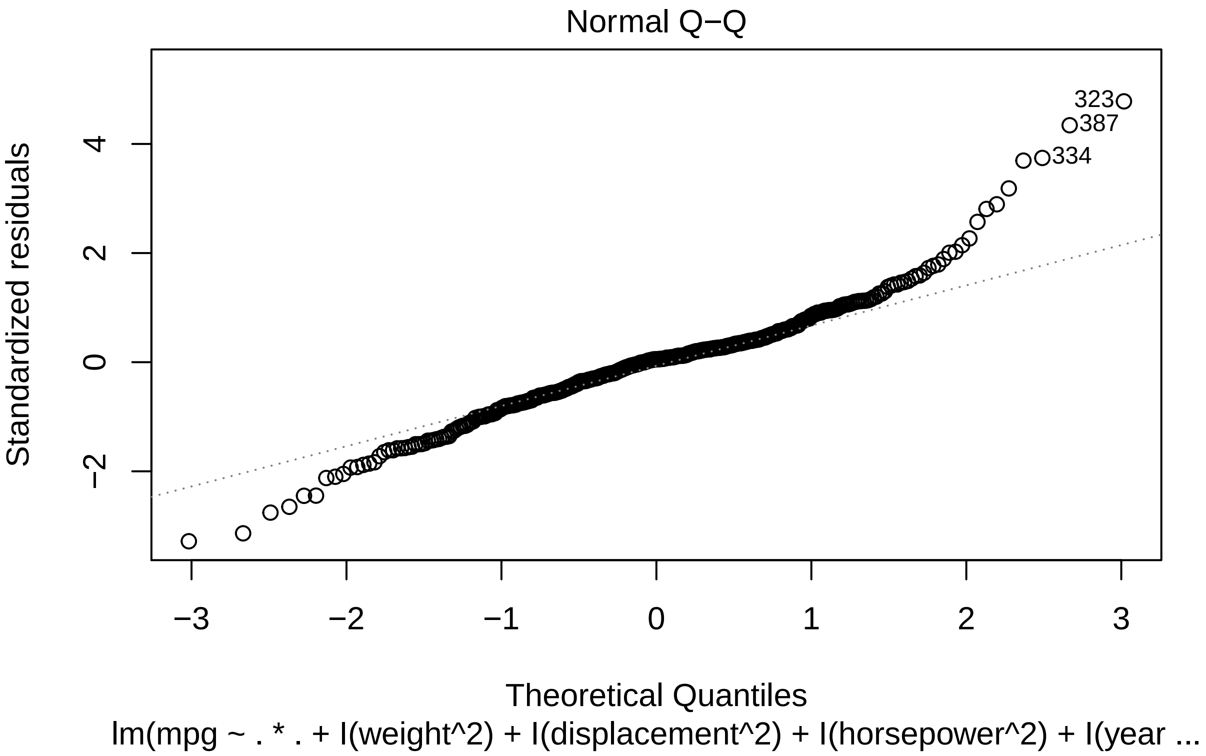
Let's try to include some polynomial terms as well. Which factor to transform is determined looking at the pairs plot.

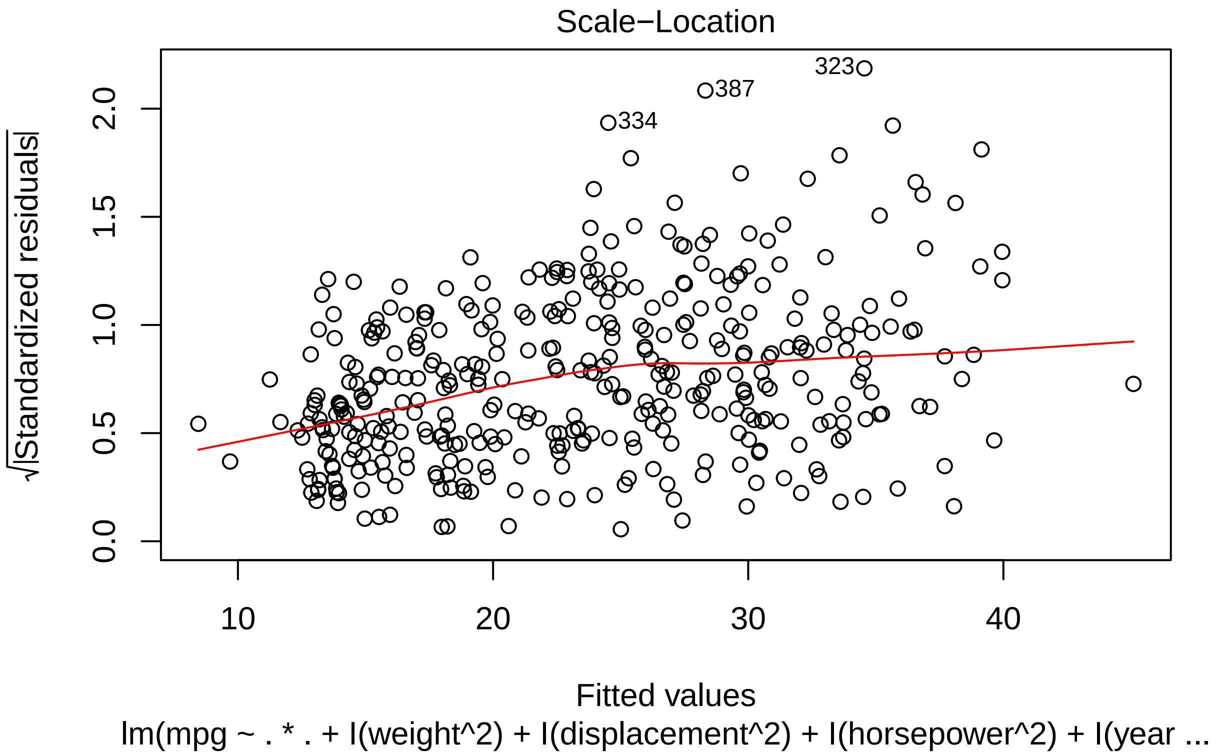
```
fit.4 <- lm(mpg~.*. + I(weight^2) + I(displacement^2) + I(horsepower^2) + I(year^2) , data=Auto[,-9])
summary(fit.4)
```

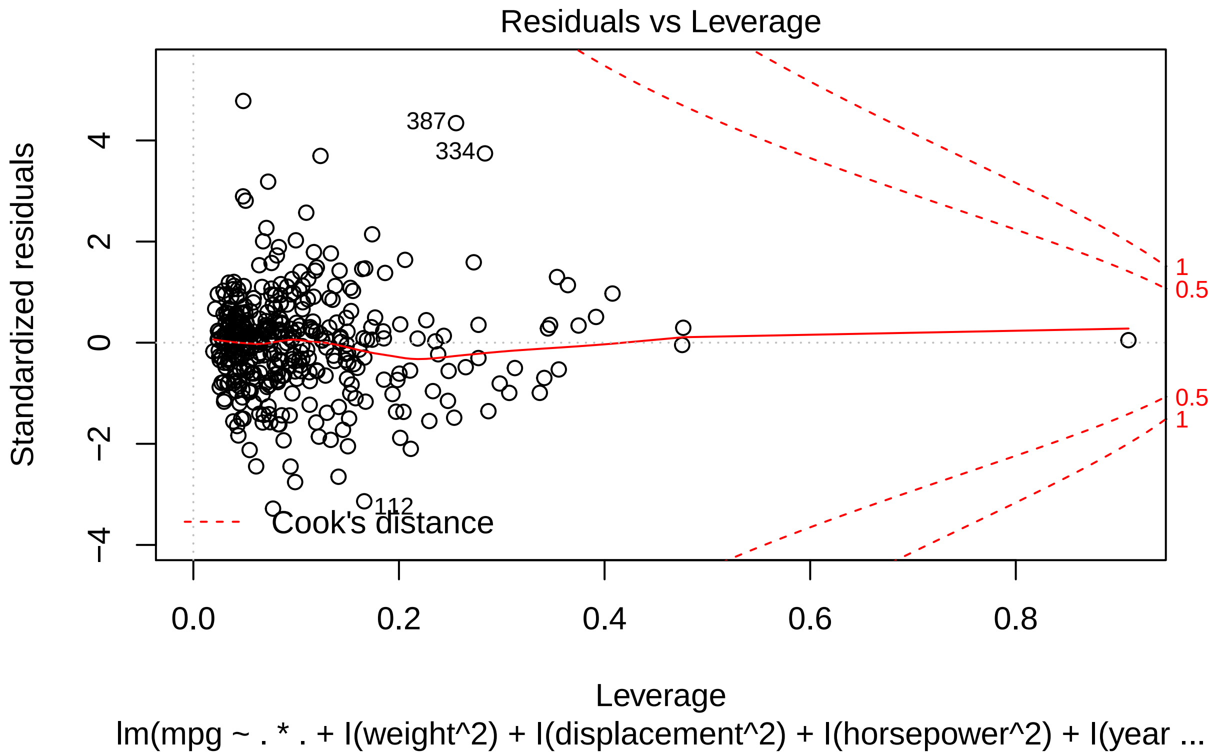
```
##
## Call:
## lm(formula = mpg ~ . * . + I(weight^2) + I(displacement^2) +
##     I(horsepower^2) + I(year^2), data = Auto[, -9])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1428 -1.3803  0.1319  1.0436 12.0466
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.849e+02  1.019e+02   3.777 0.000186 ***
## cylinders                   2.233e+00  8.169e+00   0.273 0.784728
## displacement               -4.449e-01  1.960e-01  -2.271 0.023777 *
## horsepower                  2.853e-01  3.765e-01   0.758 0.449188
## weight                      8.363e-03  1.862e-02   0.449 0.653548
## acceleration               -5.760e+00  2.176e+00  -2.647 0.008485 **
## year                       -8.353e+00  2.328e+00  -3.589 0.000379 ***
## originEuropean             -4.505e+01  1.274e+01  -3.535 0.000462 ***
## originJapanese             -3.723e+01  1.409e+01  -2.642 0.008603 **
## I(weight^2)                -3.550e-08  1.089e-06  -0.033 0.974007
## I(displacement^2)           1.530e-04  1.994e-04   0.767 0.443369
## I(horsepower^2)            -3.461e-04  5.568e-04  -0.622 0.534649
## I(year^2)                   5.664e-02  1.434e-02   3.950 9.43e-05 ***
## cylinders:displacement     -1.251e-02  1.112e-02  -1.125 0.261264
## cylinders:horsepower        1.719e-02  2.468e-02   0.696 0.486620
## cylinders:weight            1.059e-03  1.162e-03   0.912 0.362615
## cylinders:acceleration      2.884e-01  1.712e-01   1.684 0.093063 .
## cylinders:year             -1.194e-01  9.655e-02  -1.236 0.217129
## cylinders:originEuropean   -1.146e+00  1.185e+00  -0.967 0.334236
## cylinders:originJapanese    7.613e-01  1.086e+00   0.701 0.483658
## displacement:horsepower     5.998e-05  3.800e-04   0.158 0.874652
## displacement:weight          7.091e-06  2.437e-05   0.291 0.771268    
## displacement:acceleration    3.357e-04  3.720e-03   0.090 0.928159    
## displacement:year            5.303e-03  2.443e-03   2.171 0.030607 *  
## displacement:originEuropean -3.377e-02  4.393e-02  -0.769 0.442662    
## displacement:originJapanese  6.703e-02  4.412e-02   1.519 0.129578    
## horsepower:weight           -1.537e-05  4.133e-05  -0.372 0.710286    
## horsepower:acceleration     -8.399e-03  5.459e-03  -1.538 0.124834    
## horsepower:year             -2.568e-03  4.172e-03  -0.616 0.538619    
## horsepower:originEuropean    4.323e-03  5.122e-02   0.084 0.932785    
## horsepower:originJapanese    8.132e-03  6.224e-02   0.131 0.896120    
## weight:acceleration         -1.494e-05  2.477e-04  -0.060 0.951925    
## weight:year                 -2.364e-04  2.226e-04  -1.062 0.288901    
## weight:originEuropean        1.781e-03  2.697e-03   0.660 0.509491    
## weight:originJapanese       -4.420e-03  3.551e-03  -1.245 0.214121    
## acceleration:year            5.982e-02  2.582e-02   2.317 0.021092 *  
## acceleration:originEuropean  8.716e-01  2.603e-01   3.349 0.000899 ***
## acceleration:originJapanese  8.670e-01  3.267e-01   2.654 0.008321 ** 
## year:originEuropean          4.611e-01  1.480e-01   3.115 0.001993 ** 
## year:originJapanese          2.968e-01  1.463e-01   2.029 0.043246 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.583 on 352 degrees of freedom
## Multiple R-squared:  0.9014, Adjusted R-squared:  0.8905 
## F-statistic:  82.5 on 39 and 352 DF,  p-value: < 2.2e-16
```

```r
# Diagnostic plots
plot(fit.4)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ . * . + I(weight^2) + I(displacement^2) + I(horsepower^2) + I(year ...
10
5
0
−5
−10
10
20
30
40
323
387
330


Normal Q−Q
Standardized residuals
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
323
387
334
lm(mpg ~ . * . + I(weight^2) + I(displacement^2) + I(horsepower^2) + I(year ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10
20
30
40
Fitted values
lm(mpg ~ . * . + I(weight^2) + I(displacement^2) + I(horsepower^2) + I(year ...
323
387
334



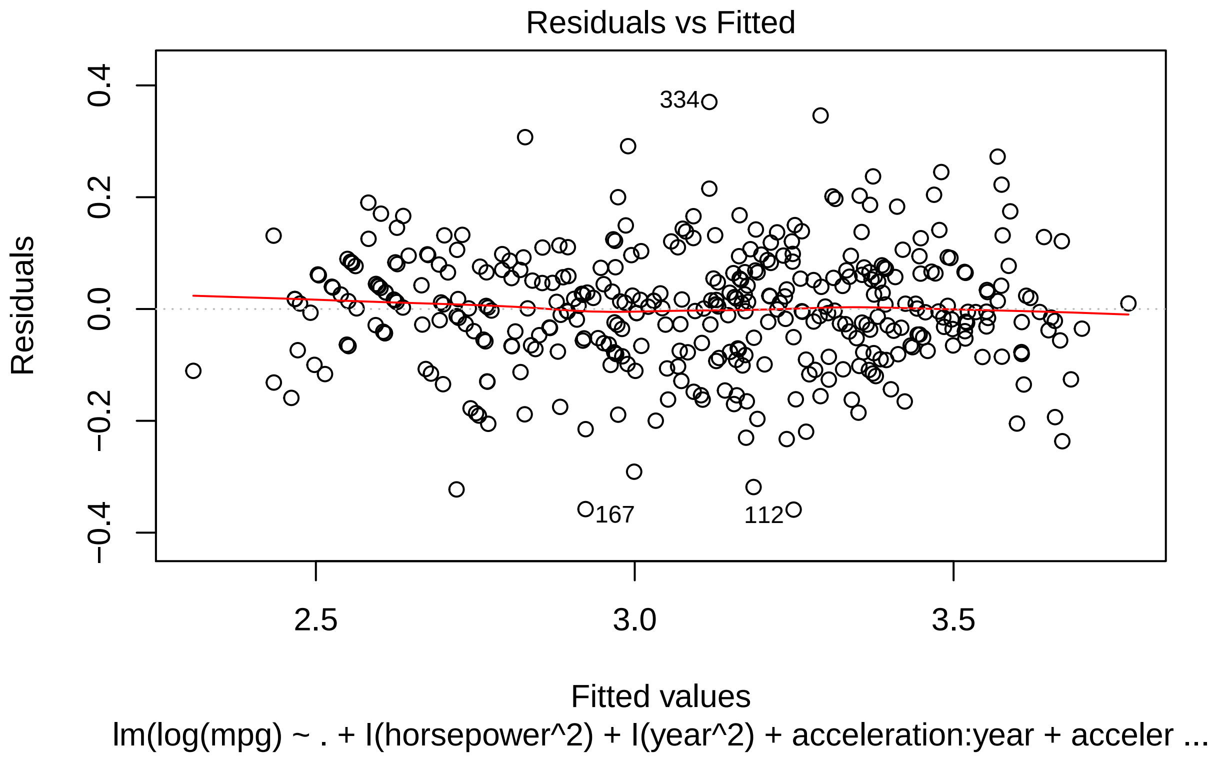
The Tukey-Anscombe plot seems to have a funnel shape. Let's transform the response to stabilize the predictions.

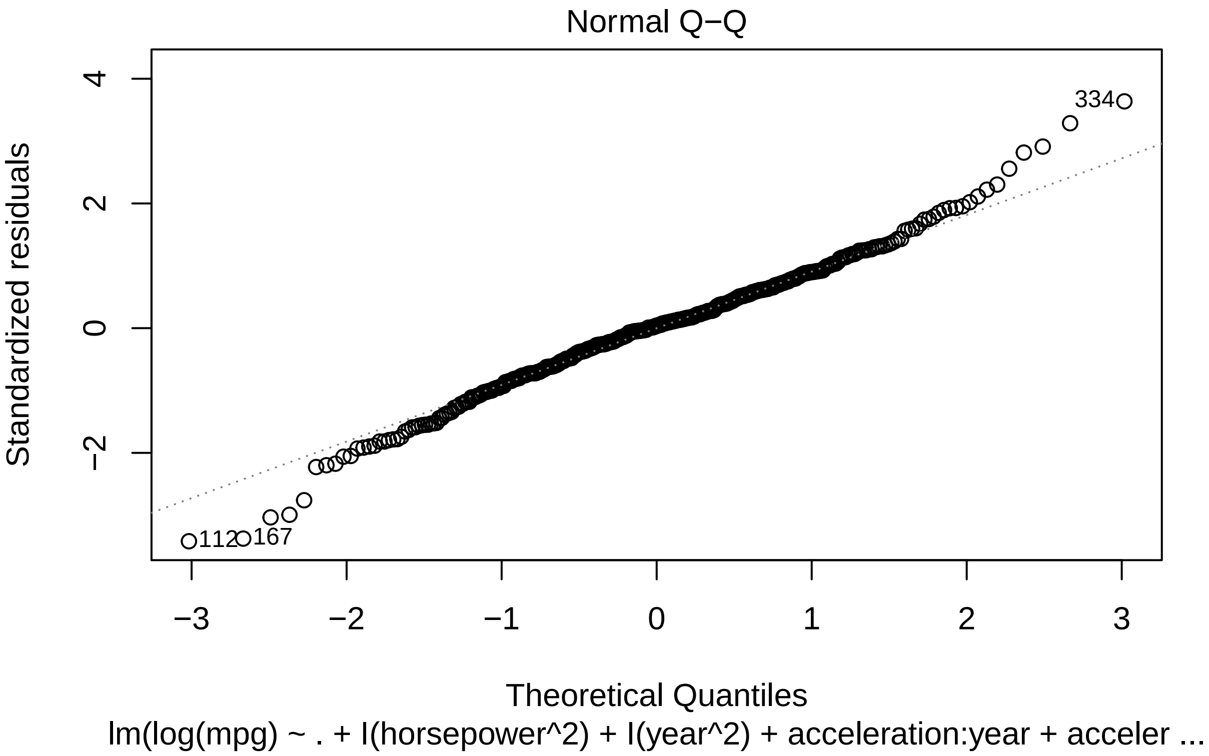
```
fit.5 <- lm(log(mpg)~.+ I(horsepower^2) + I(year^2) + acceleration:year + acceleration:origin, data=Aut
summary(fit.5)
```

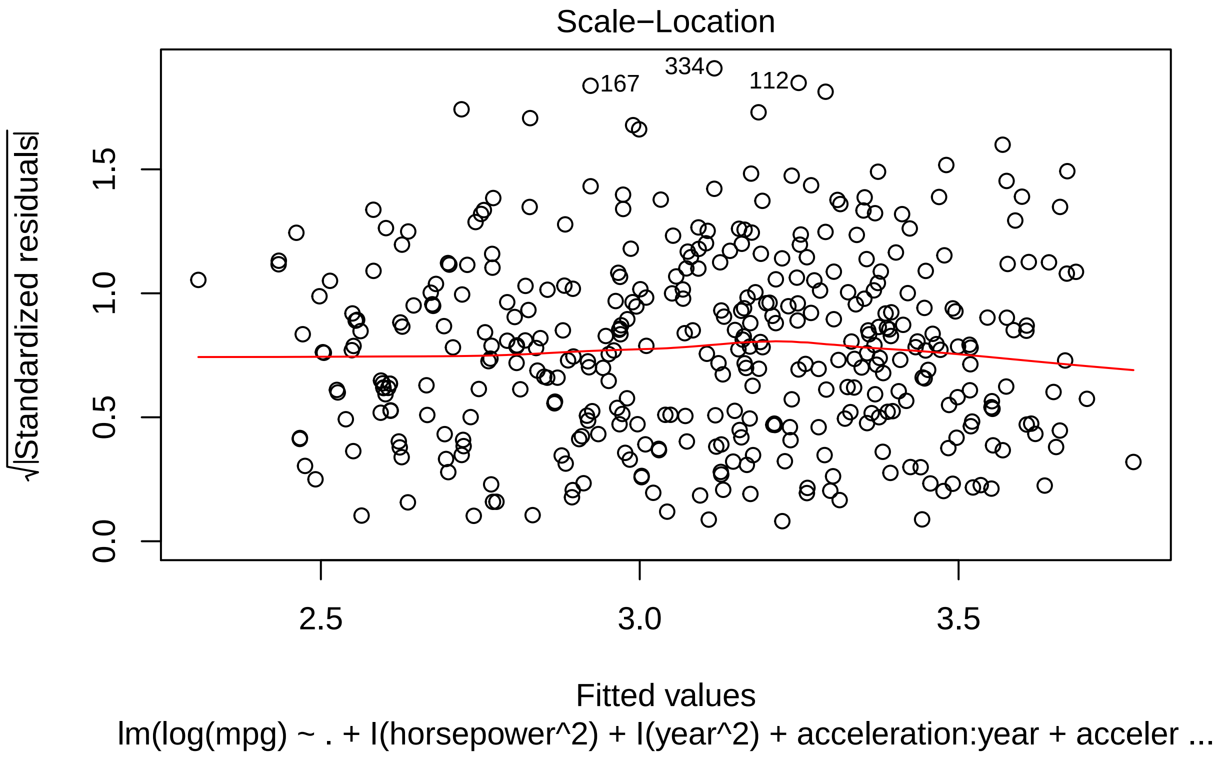
```
##
## Call:
## lm(formula = log(mpg) ~ . + I(horsepower^2) + I(year^2) + acceleration:year +
##     acceleration:origin, data = Auto[, -9])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35880 -0.06524  0.00426  0.06543  0.37036
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.337e+01  2.747e+00   4.867 1.67e-06 ***
## cylinders                  -7.042e-03  1.120e-02  -0.629 0.529851
## displacement               -2.693e-04  2.825e-04  -0.953 0.341067
## horsepower                 -6.849e-03  1.300e-03  -5.267 2.33e-07 ***
## weight                     -1.581e-04  2.569e-05  -6.155 1.92e-09 ***
## acceleration               -1.354e-01  4.645e-02  -2.915 0.003767 **
## year                       -2.402e-01  7.362e-02  -3.263 0.001201 **
## originEuropean             -2.782e-01  9.812e-02  -2.835 0.004829 **
## originJapanese             -2.694e-01  1.233e-01  -2.185 0.029527 *
## I(horsepower^2)             1.572e-05  4.123e-06   3.813 0.000161 ***
## I(year^2)                       1.617e-03  4.986e-04   3.243 0.001288 **
## acceleration:year               1.514e-03  6.100e-04   2.482 0.013496 *
## acceleration:originEuropean  2.004e-02  5.721e-03   3.503 0.000514 ***
## acceleration:originJapanese  2.013e-02  7.551e-03   2.666 0.008003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.108 on 378 degrees of freedom
## Multiple R-squared:  0.9025, Adjusted R-squared:  0.8991
## F-statistic: 269.1 on 13 and 378 DF,  p-value: < 2.2e-16
```

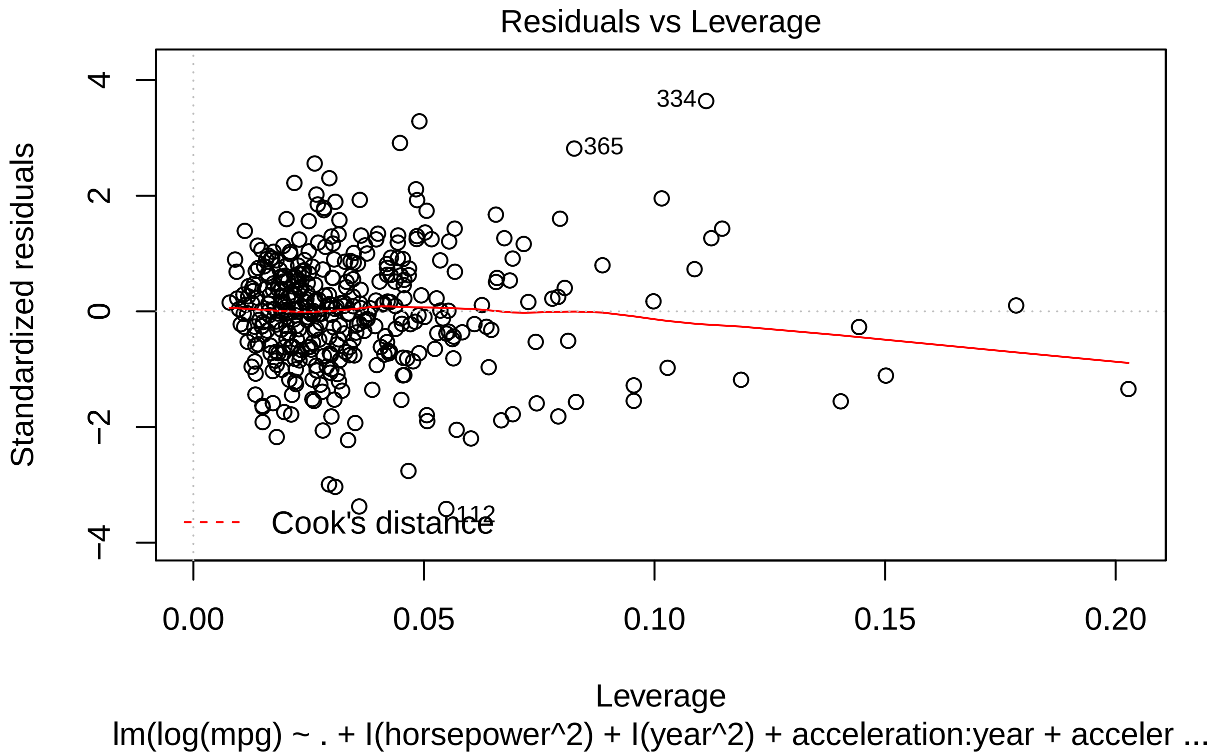
Diagnostic plot again.

```
plot(fit.5)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
334
112
167
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(mpg) ~ . + I(horsepower^2) + I(year^2) + acceleration:year + acceler ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
167
334
112
2.5
3.0
3.5
Fitted values
lm(log(mpg) ~ . + I(horsepower^2) + I(year^2) + acceleration:year + acceler ...



The log-transformation of the response variable seems to have had the desired impact on the residuals.

The Q-Q plot seems to suggest a heavier tails distribution for the residuals.

Exercise 10

Exercise 14

Exercise 15