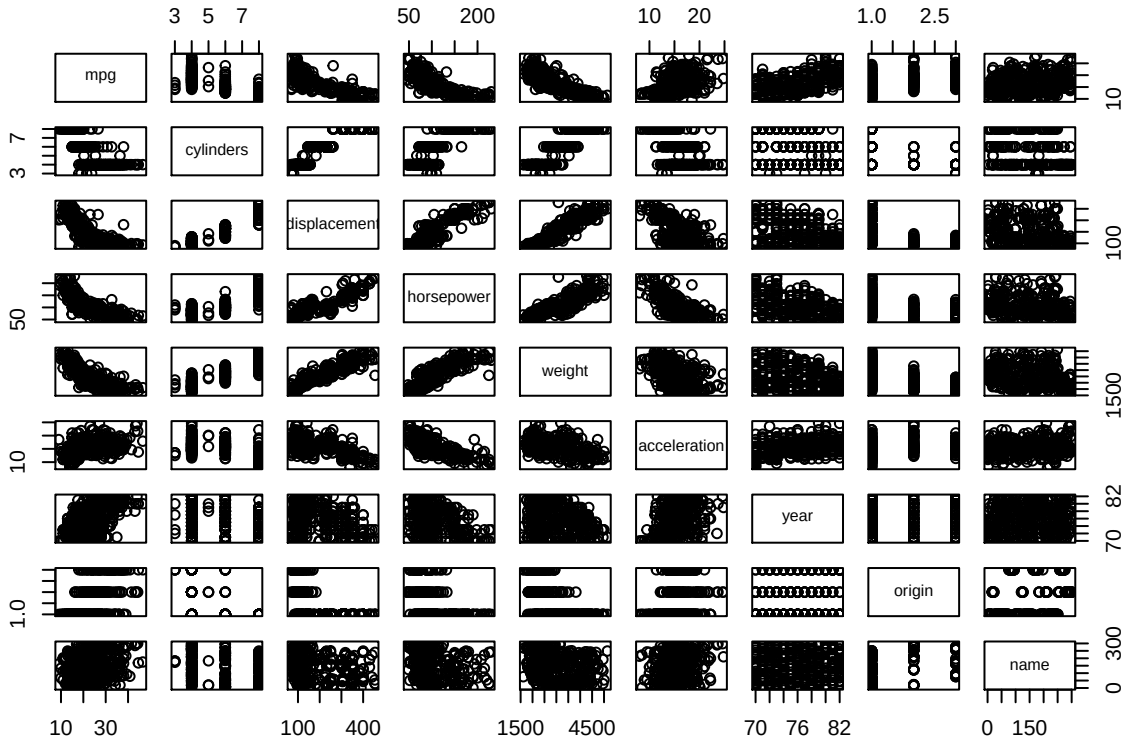
Question 9:

A:

```
Auto <- read.csv("/Users/matthewbradley/Downloads/Auto.csv")
Auto$name <- factor(Auto$name)
pairs(Auto)
```



B:

```
cor(Auto[,1:8])
```

```
##                     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
```

C:

```
linearModel <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration +
                    year + as.factor(origin), data = Auto)

summary(linearModel)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + as.factor(origin), data = Auto)
##
## 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 ***
## as.factor(origin)2  2.630e+00  5.664e-01   4.643 4.72e-06 ***
## as.factor(origin)3  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
```

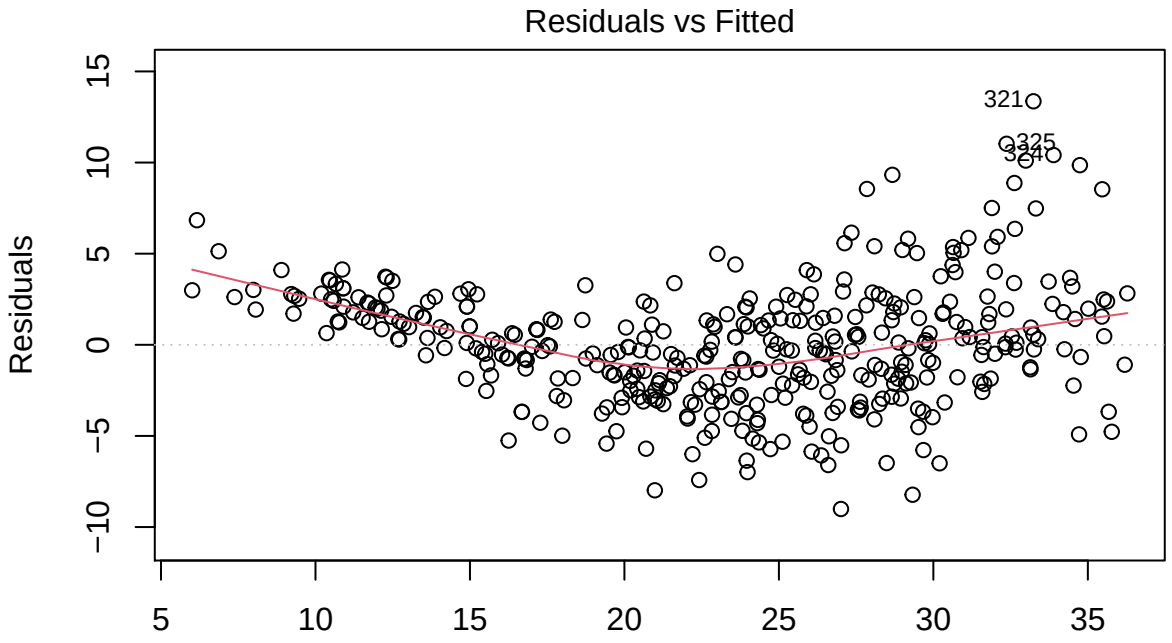
i. There is a relationship between predictor and response. Our R-squared is .8205, meaning we can predict about 82% of the variation of the data using our predictors. The F statistic is also extremely small (<2.2e-16), which allows us to reject the null hypothesis that there is no relationship.

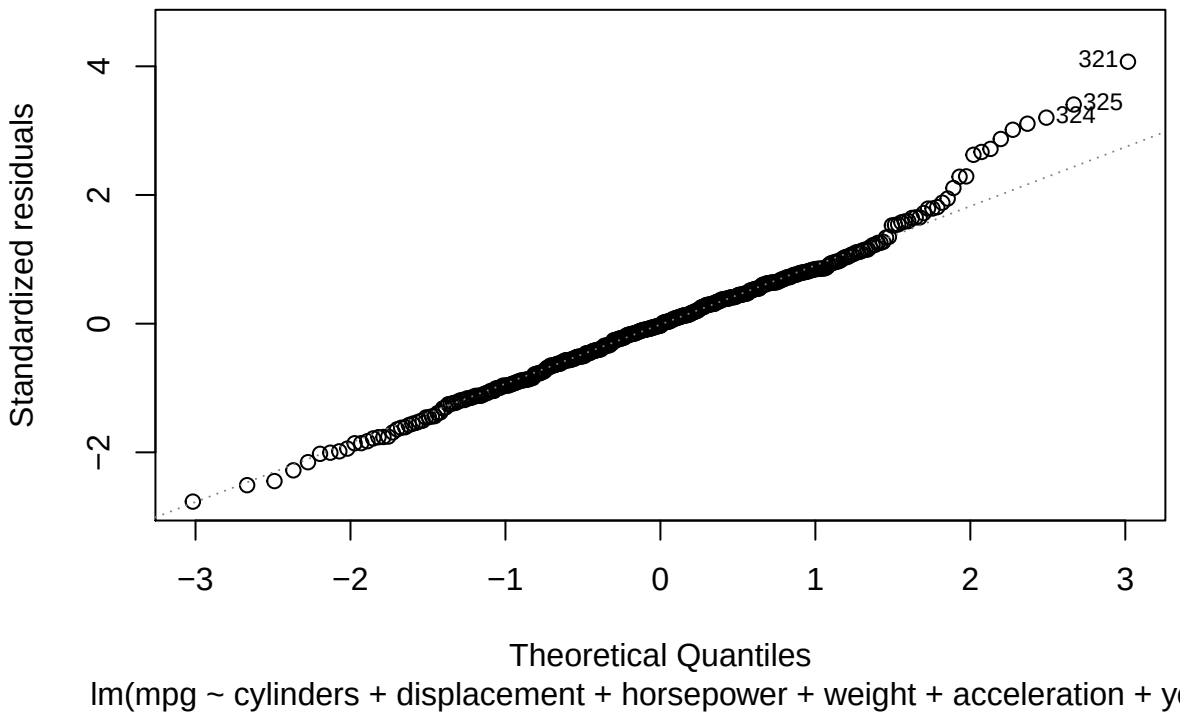
ii. Displacement, weight, year, and origin all appear to have statistically significant relationships with the response ($p<0.05$ for these variables)

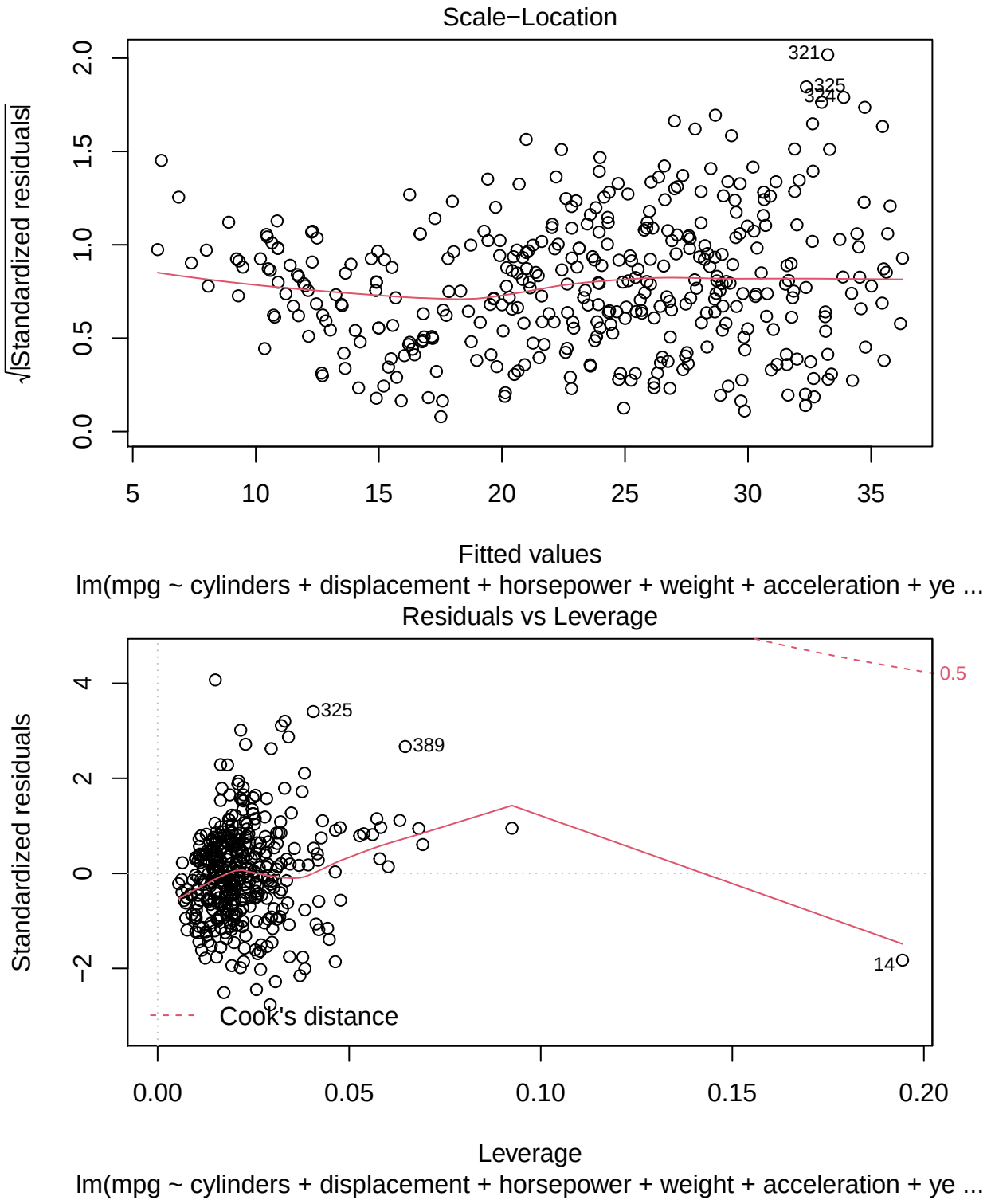
iii. The year coefficient (0.777) suggests that for every increase of 1 year, the response variable (mpg) increases by 0.777.

D:

```
plot(linearModel)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...
321
325
324


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...
321
325
324



The residual plot has a "U" shape, suggesting there could be non linearity in the data. There also do appear to be large outliers on the right hand side of the plot (marked 323, 326, and 327).

The Normal Q-Q plot also shows outliers on the right side, marked with the same numbers. This suggests these data points may be skewing the normality of the data.

The leverage plot suggests that point 14 has high leverage, suggesting it may be a particularly influential

point for our model.

E:

```
linearModel <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year +
                    as.factor(origin) + cylinders*displacement + cylinders:horsepower +
                    cylinders:weight +cylinders*acceleration + cylinders*year +
                    cylinders:as.factor(origin), data = Auto)

summary(linearModel)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + as.factor(origin) + cylinders * displacement +
##     cylinders:horsepower + cylinders:weight + cylinders * acceleration +
##     cylinders * year + cylinders:as.factor(origin), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4976 -1.7194  0.0678  1.3838 12.0082
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -4.084e+01  1.515e+01  -2.695  0.00736 **
## cylinders                     5.057e+00  2.709e+00   1.867  0.06271 .
## displacement                  2.243e-02  2.524e-02   0.889  0.37483
## horsepower                   -1.526e-01  5.552e-02  -2.748  0.00628 **
## weight                       -1.185e-02  2.593e-03  -4.569 6.67e-06 ***
## acceleration                  5.202e-02  2.988e-01   0.174  0.86188
## year                          1.441e+00  1.700e-01   8.477 5.33e-16 ***
## as.factor(origin)2            1.451e+00  3.277e+00   0.443  0.65832
## as.factor(origin)3           -3.900e+00  2.837e+00  -1.375  0.16998
## cylinders:displacement       -1.752e-03  3.727e-03  -0.470  0.63859
## cylinders:horsepower          1.592e-02  8.105e-03   1.964  0.05027 .
## cylinders:weight              1.089e-03  3.765e-04   2.893  0.00404 **
## cylinders:acceleration       -5.631e-03  5.417e-02  -0.104  0.91725
## cylinders:year               -1.305e-01  3.180e-02  -4.105 4.96e-05 ***
## cylinders:as.factor(origin)2  1.177e-01  7.636e-01   0.154  0.87755
## cylinders:as.factor(origin)3  1.347e+00  6.563e-01   2.053  0.04080 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.85 on 376 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8667
## F-statistic: 170.4 on 15 and 376 DF,  p-value: < 2.2e-16
```

I examined the interaction effects of the "cylinder" variable with all other variables. There do appear to be several significant interaction effects (with weight, year, and origin).

F:

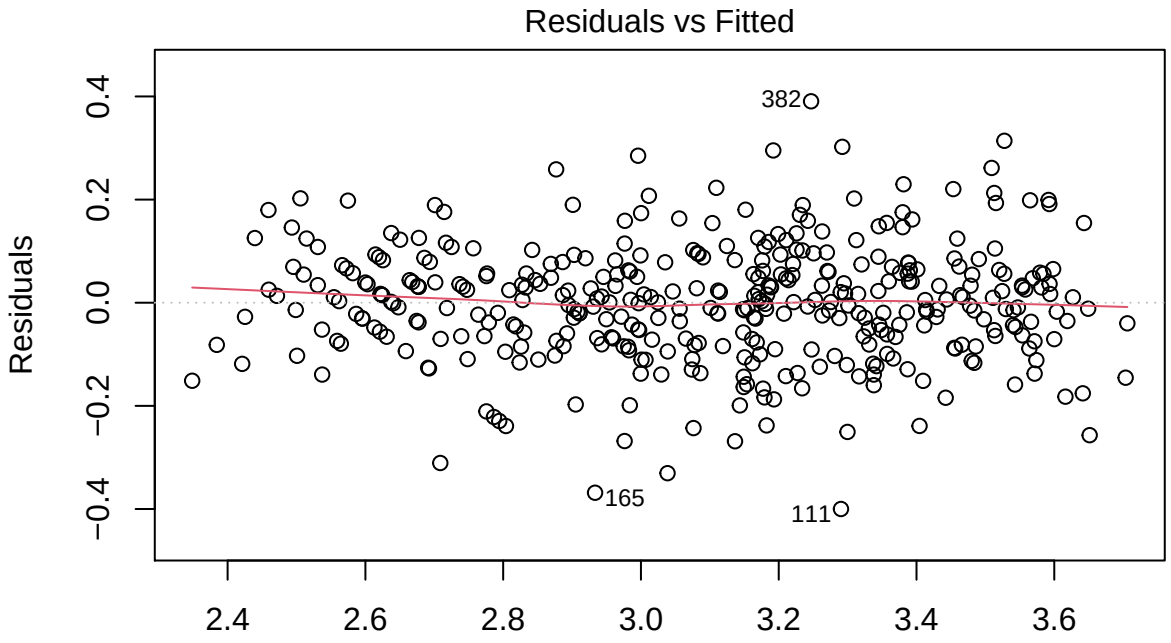
```
loggedData <- log(Auto[1:7])
origin = Auto[,8]
```

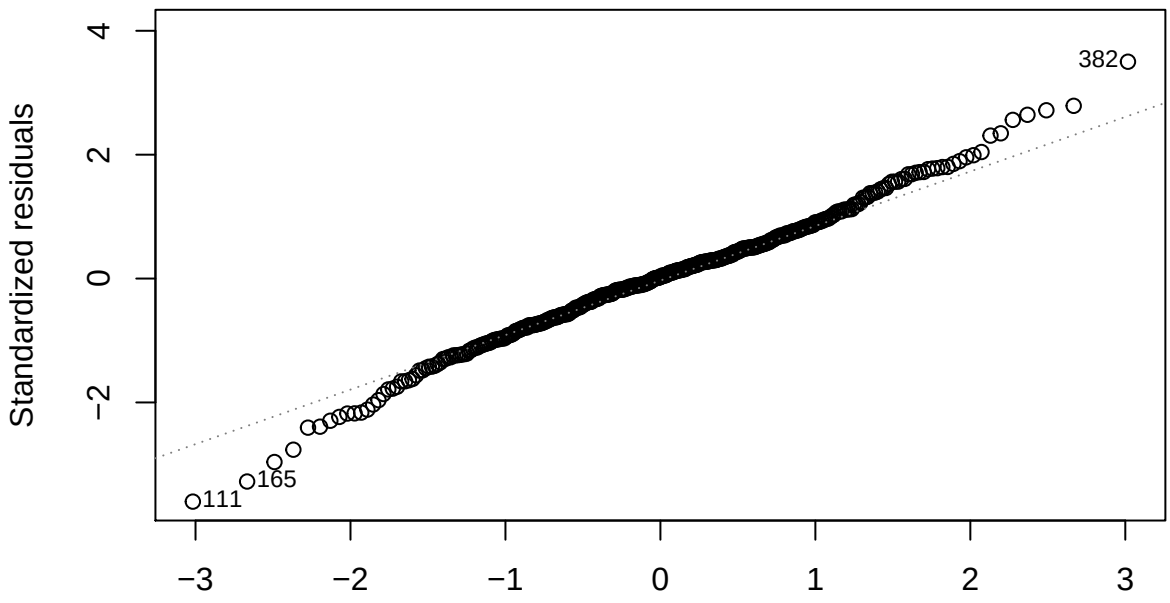
```r
loggedData <- cbind(loggedData, origin)
logModel <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year +
                 as.factor(origin), data = loggedData)
summary(logModel)
```

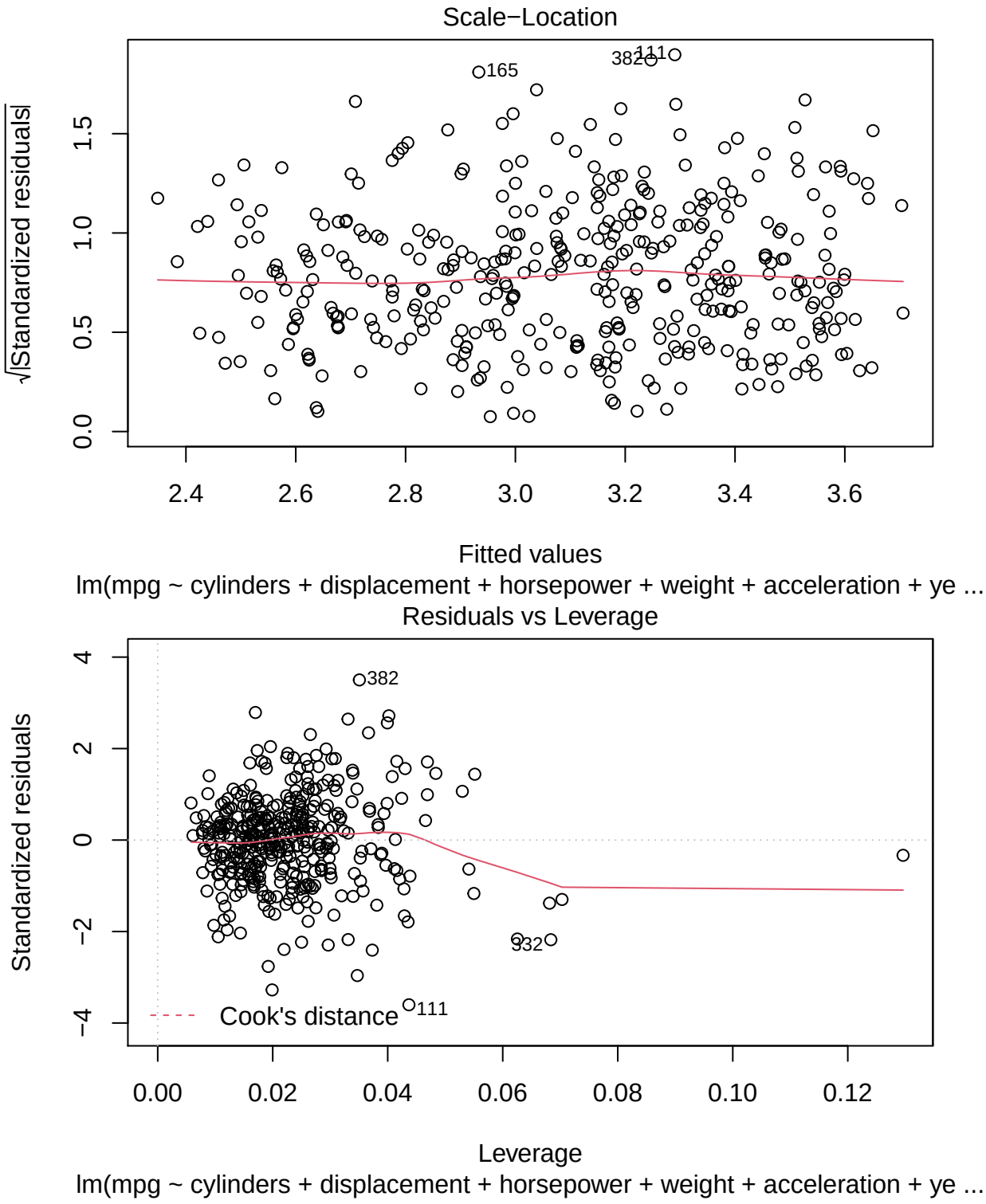
Log transformation:

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + as.factor(origin), data = loggedData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39999 -0.06970  0.00294  0.06304  0.39059
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.24514    0.65266  -0.376  0.70741
## cylinders          -0.08546    0.06145  -1.391  0.16510
## displacement        0.02303    0.05871   0.392  0.69503
## horsepower         -0.28422    0.05830  -4.875 1.60e-06 ***
## weight             -0.59696    0.08572  -6.964 1.45e-11 ***
## acceleration       -0.17066    0.05998  -2.845  0.00468 **
## year                2.27962    0.13521  16.859  < 2e-16 ***
## as.factor(origin)2  0.05004    0.02103   2.379  0.01785 *
## as.factor(origin)3  0.04736    0.02074   2.284  0.02293 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1136 on 383 degrees of freedom
## Multiple R-squared:  0.8907, Adjusted R-squared:  0.8884
## F-statistic: 390.2 on 8 and 383 DF,  p-value: < 2.2e-16
```

```r
plot(logModel)
```


Residuals vs Fitted
Residuals
382
165
111
0.4
0.2
0.0
−0.2
−0.4
2.4
2.6
2.8
3.0
3.2
3.4
3.6
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...


Normal Q−Q
Standardized residuals
382
165
111
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...



We have a slightly higher R-squared for the model using log values, suggesting that it may have helped limit the effects of non linearity and outliers. However, the Q-Q plot shows new outliers as well, so we would need to look deeper to determine if this model is truly better.

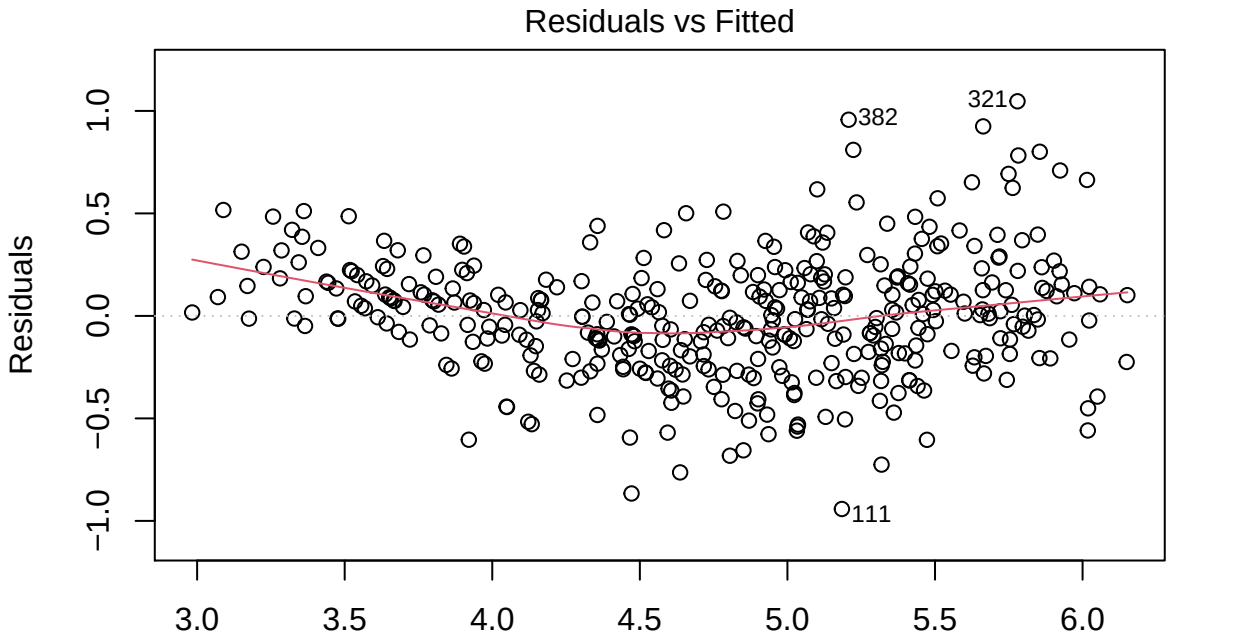
```
squareRootData <- sqrt(Auto[1:7])
```

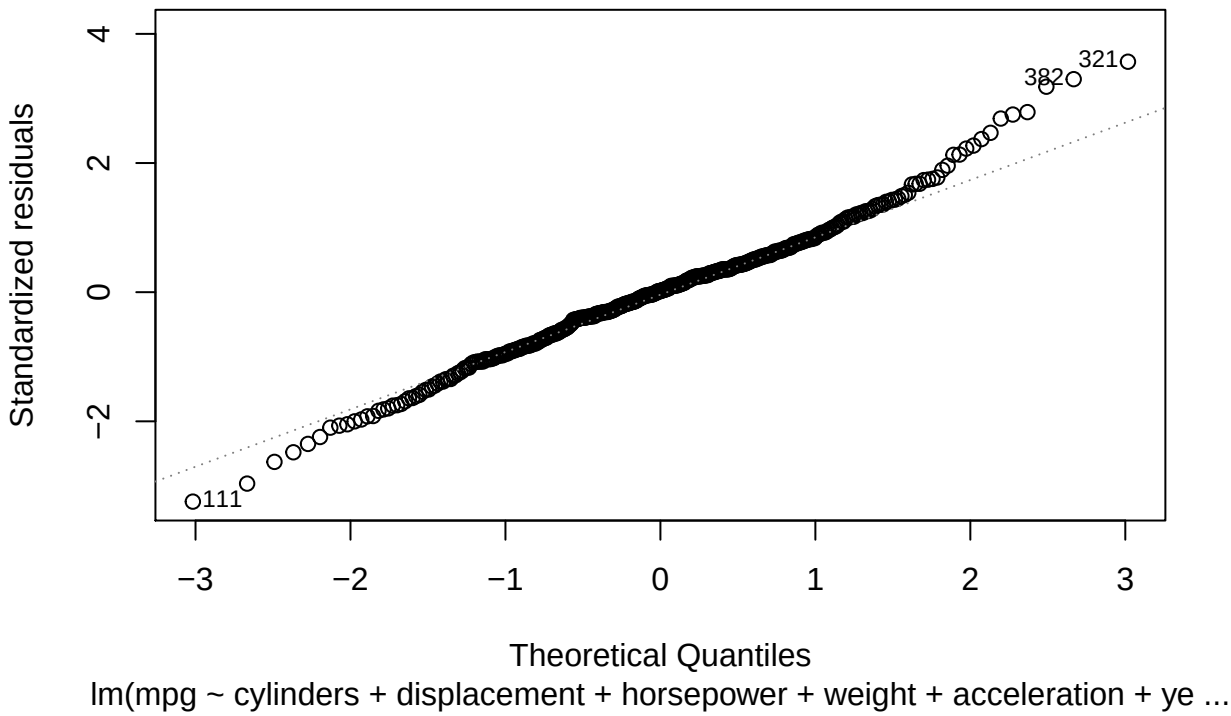
```r
squareRootData <- cbind(squareRootData, origin)
sqrtModel <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year +
                  as.factor(origin), data = squareRootData)
summary(sqrtModel)
```

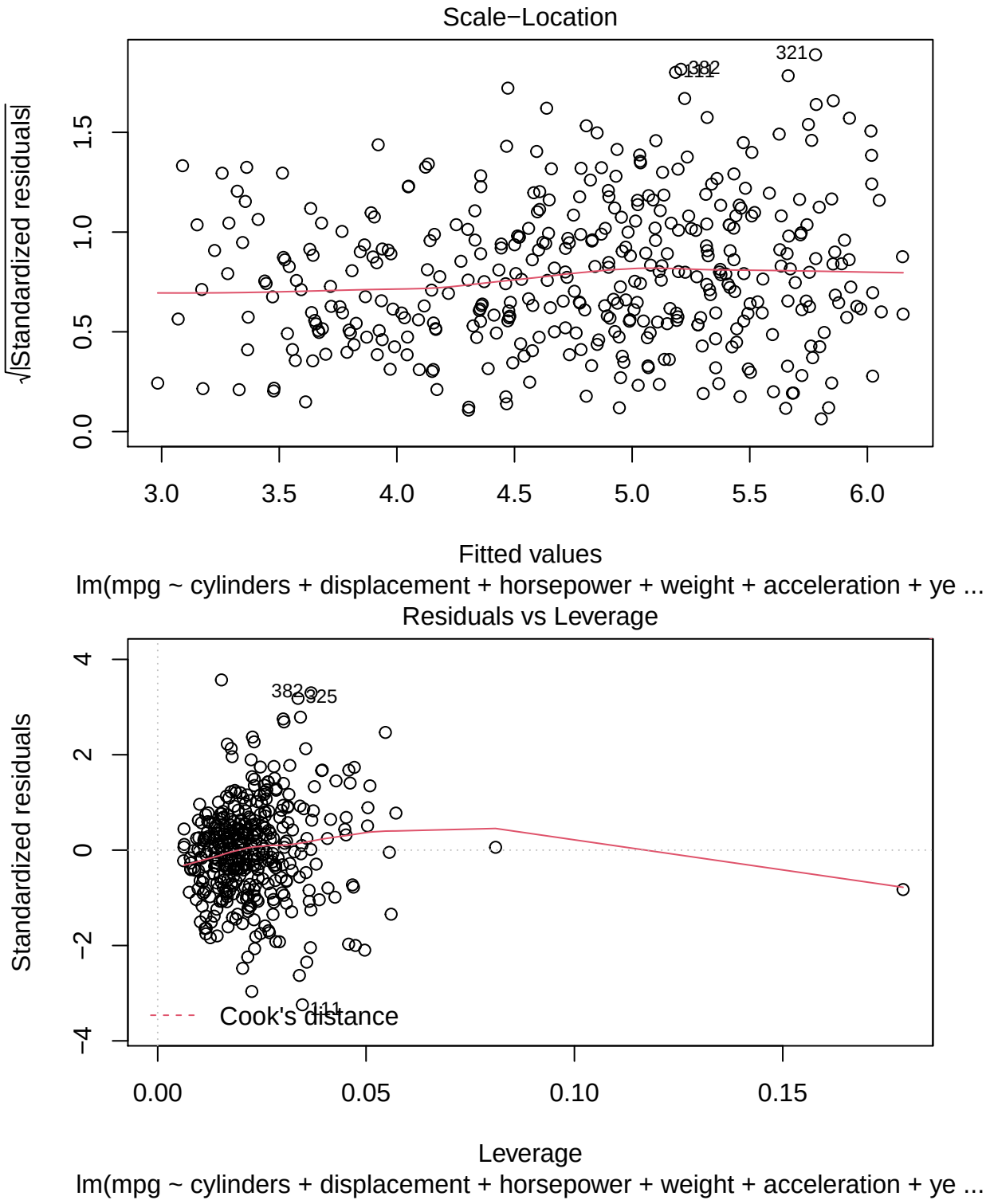
Square root transformation:

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + as.factor(origin), data = squareRootData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94193 -0.18566  0.00493  0.16375  1.04694
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.983035   0.857373  -2.313 0.021256 *
## cylinders          -0.118364   0.141690  -0.835 0.404029
## displacement        0.029209   0.021802   1.340 0.181127
## horsepower         -0.087893   0.028402  -3.095 0.002116 **
## weight             -0.064363   0.007463  -8.625  < 2e-16 ***
## acceleration       -0.097928   0.077020  -1.271 0.204336
## year                1.298207   0.081175  15.993  < 2e-16 ***
## as.factor(origin)2  0.190208   0.052807   3.602 0.000357 ***
## as.factor(origin)3  0.196275   0.051778   3.791 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2956 on 383 degrees of freedom
## Multiple R-squared:  0.8673, Adjusted R-squared:  0.8645
## F-statistic: 312.8 on 8 and 383 DF,  p-value: < 2.2e-16
```

```r
plot(sqrtModel)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
382
321
111
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...


Normal Q−Q
Standardized residuals
4
2
0
−2
321
382
111
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...



This model (square root values) has a similar R-squared to our original model, and this transformation did not seem to help much with outliers. Therefore, it would probably be unhelpful to use a square root transformation on the variables.