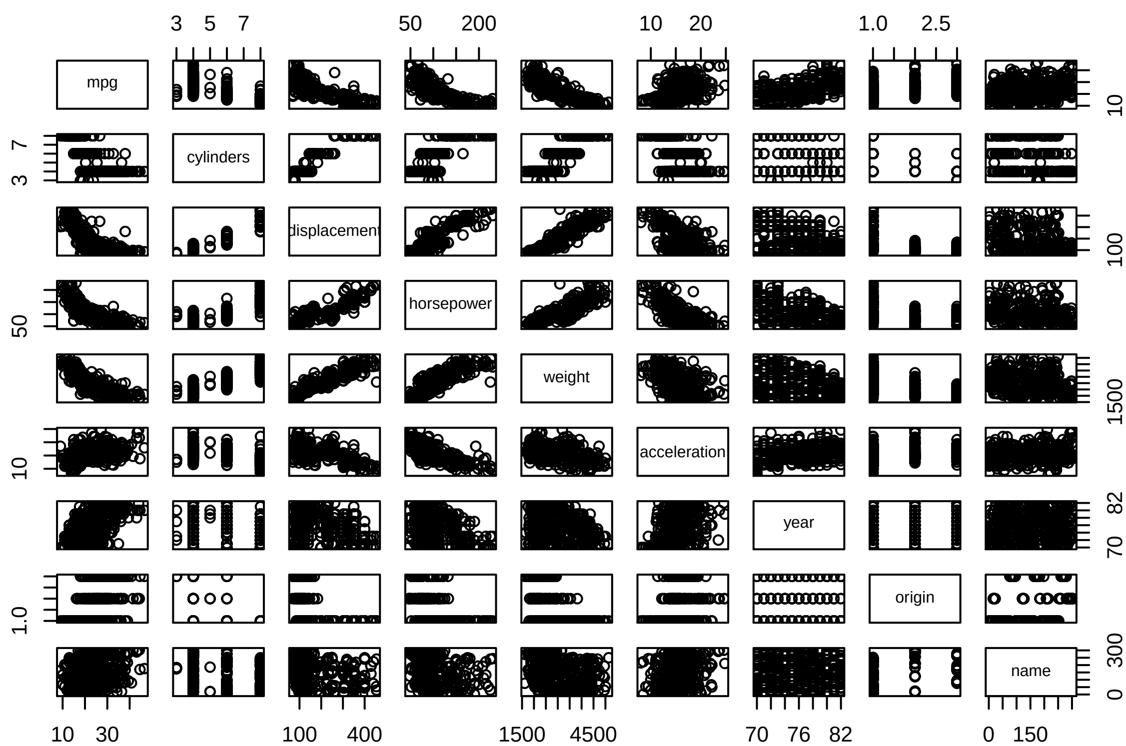
STAT 627 HW-3

1.

a.

```r
load("C:/Users/jumawidi/Desktop/AU R Studio/Auto.rda")
attach(Auto)
plot(Auto)
```



b.

```r
cor(Auto[ names(Auto) != "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
```

c.

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

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

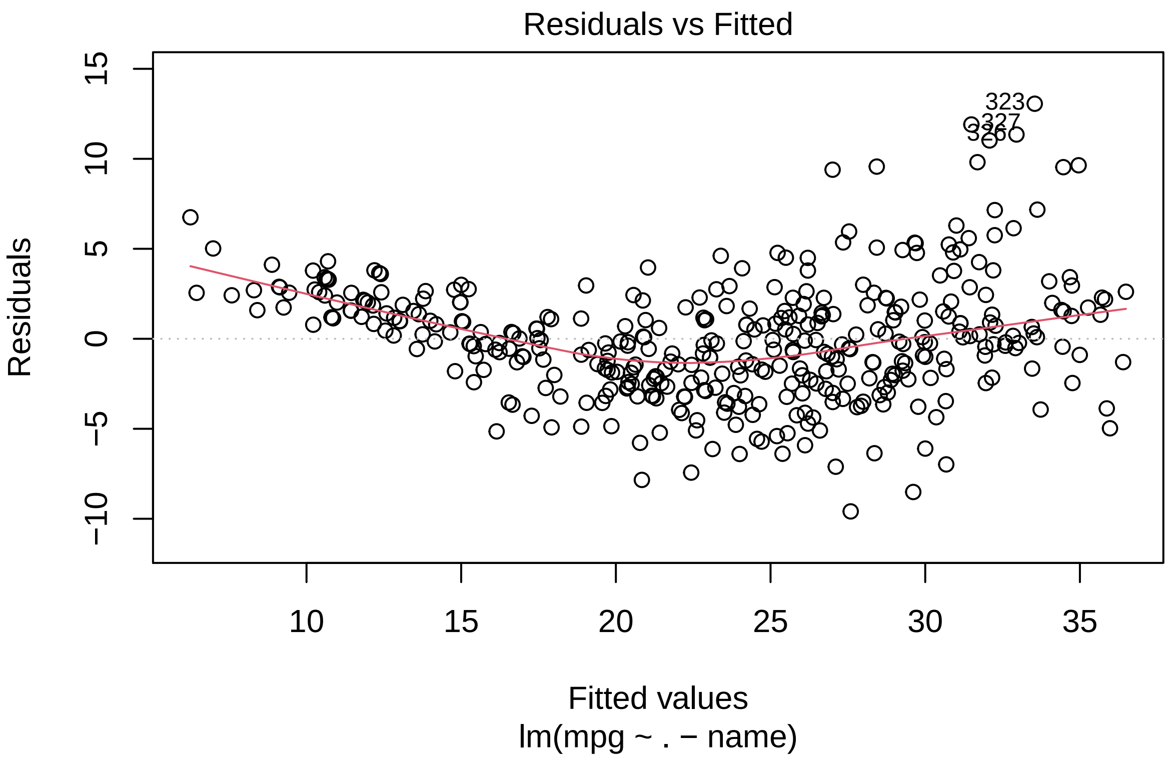
i. Yes there is a relationship. Many of the variables' estimates are statistically significant. We also see that the R2 value is .8182, indicating the line explains 81% of the vairance.

ii. displacement, weight, year, origin.

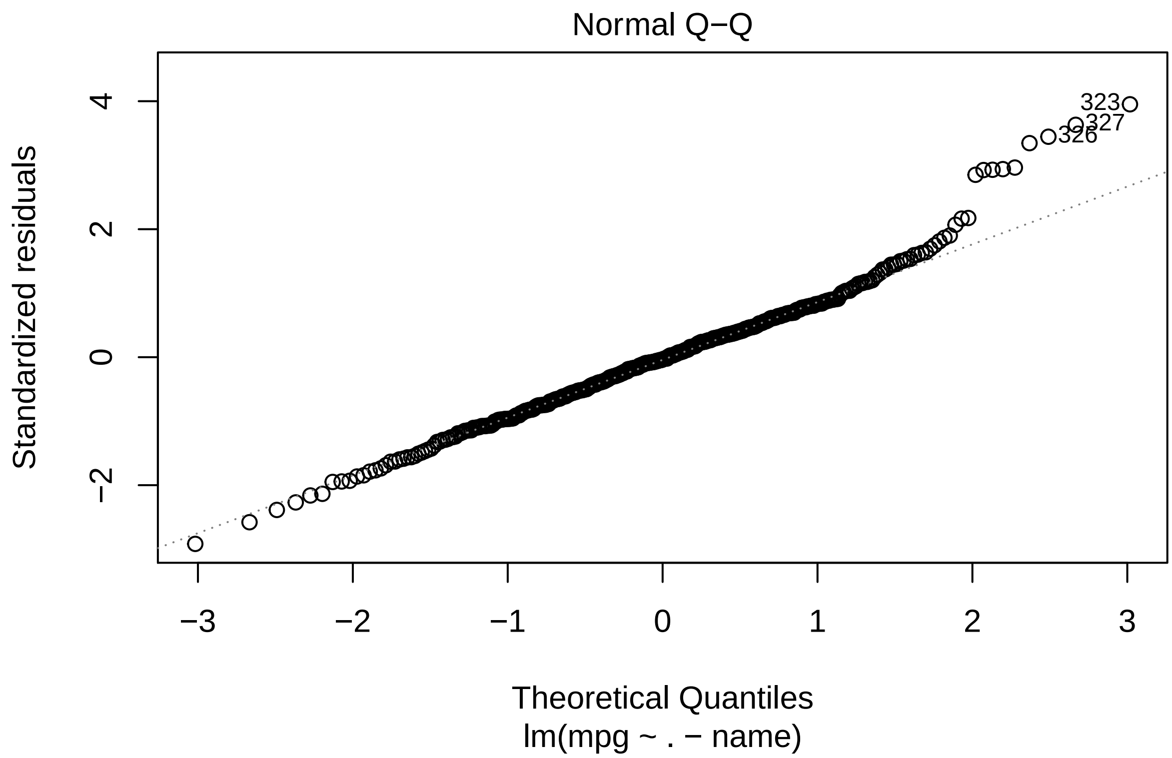
iii. The coefficient indicates that miles per gallon increases by 0.750773 every year.

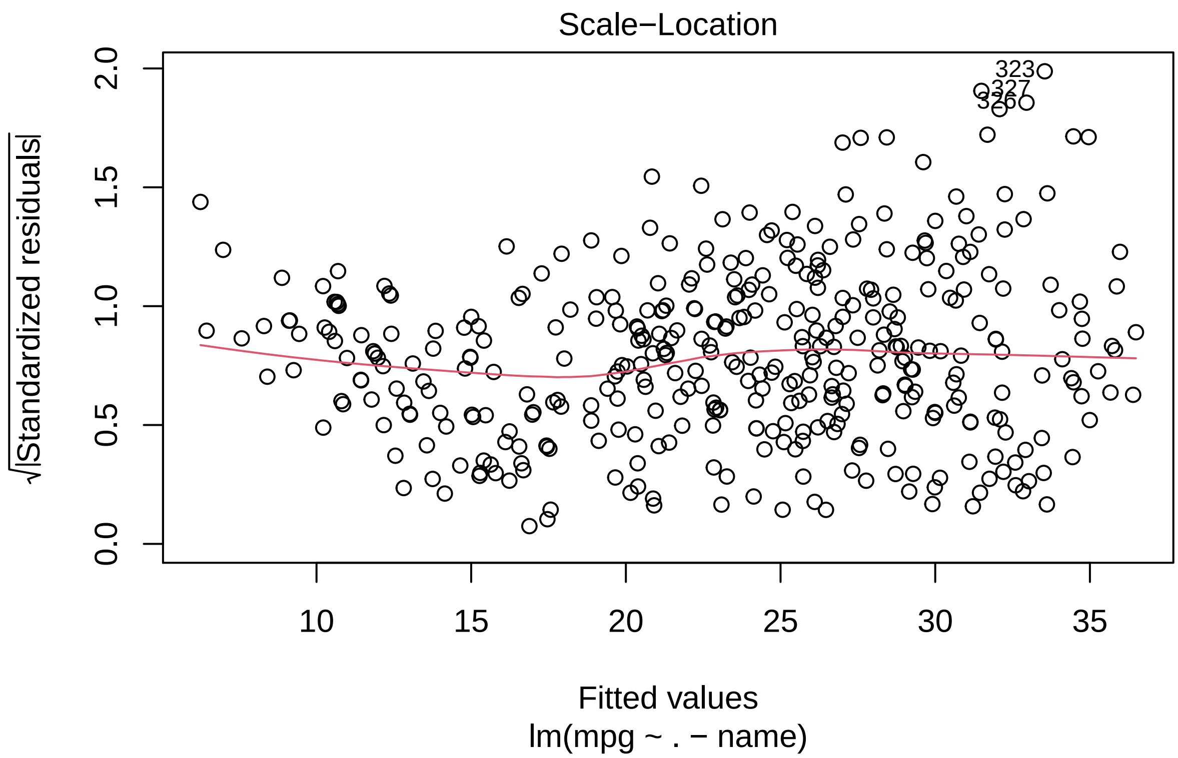
iv.

```r
plot(reg)
```

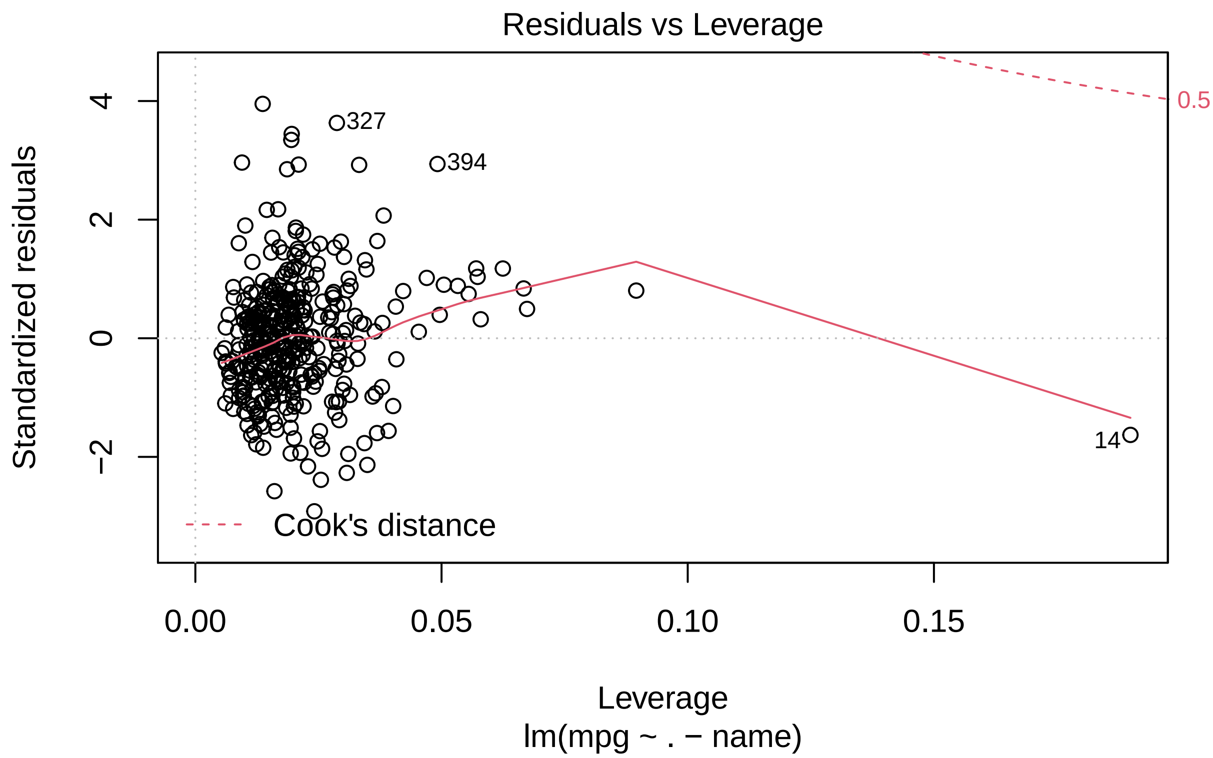


Residuals vs Fitted
Residuals
15
10
5
0
−5
−10
323
327
326
10
15
20
25
30
35
Fitted values
lm(mpg ~ . − name)


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



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



The first graph shows that the response and predictor's relationship is not linear since there is a clear pattern. The QQ plot shows that there is normal distribution except for the end values. The third shows a violation of the constant variance assumption, and the las tgraph shows that there is a potential outlier at 14.

e.

```r
reg <- lm(mpg~.-name+ displacement:cylinders + displacement:weight + acceleration:horsepower, data = Au
summary(reg)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + displacement:cylinders + displacement:weight +
##     acceleration:horsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3344 -1.6333  0.0188  1.4740 11.9723
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.725e+01  5.328e+00  -3.237  0.00131 **
## cylinders                6.354e-01  6.106e-01   1.041  0.29870
## displacement            -6.805e-02  1.337e-02  -5.088 5.68e-07 ***
## horsepower               6.026e-02  2.601e-02   2.317  0.02105 *
## weight                  -8.864e-03  1.097e-03  -8.084 8.43e-15 ***
```

```
## acceleration             6.257e-01  1.592e-01   3.931  0.00010 ***
## year                     7.845e-01  4.470e-02  17.549  < 2e-16 ***
## origin                   4.668e-01  2.595e-01   1.799  0.07284 .
## cylinders:displacement  -1.337e-03  2.726e-03  -0.490  0.62415
## displacement:weight      2.071e-05  3.638e-06   5.694 2.49e-08 ***
## horsepower:acceleration -7.467e-03  1.784e-03  -4.185 3.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.905 on 381 degrees of freedom
## Multiple R-squared:  0.865,  Adjusted R-squared:  0.8615
## F-statistic: 244.2 on 10 and 381 DF,  p-value: < 2.2e-16
```

I found that the interaction variables for displacement:weight and horsepower:acceleration are statistically significant.

f.

```
reg <- lm(mpg~.-name + sqrt(weight) + log(horsepower), data = Auto)
summary(reg)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + sqrt(weight) + log(horsepower),
##     data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9249 -1.4947 -0.1793  1.4600 12.1595
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      95.284814  11.099752   8.584 2.34e-16 ***
## cylinders        -0.175045   0.287395  -0.609 0.542836
## displacement      0.000156   0.007094   0.022 0.982468
## horsepower        0.109398   0.028358   3.858 0.000134 ***
## weight            0.010411   0.003658   2.846 0.004667 **
## acceleration     -0.208539   0.099998  -2.085 0.037693 *
## year              0.770153   0.045152  17.057  < 2e-16 ***
## origin            0.700298   0.253702   2.760 0.006053 **
## sqrt(weight)     -1.614885   0.422111  -3.826 0.000152 ***
## log(horsepower) -17.975649   3.489375  -5.152 4.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.908 on 382 degrees of freedom
## Multiple R-squared:  0.8644, Adjusted R-squared:  0.8612
## F-statistic: 270.5 on 9 and 382 DF,  p-value: < 2.2e-16
```

Using the sqrt of weight and the log of horsepower (both statistically significant), I found a pretty decent regression line with a R2 value of .8644.

2.

a.

```r
load("C:/Users/jumawidi/Desktop/AU R Studio/Carseats.rda")

attach(Carseats)
reg2 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(reg2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b. US Stores in urban location sales = 14.222 - .0544 * Price

Non-US stores in an urban location sales = 13.022 - .0544 * Price

US stores in rural location sales = 14.244 - .0544 * Price

Non-US stores in rural location Sales = 13.043 - 0.0544 * Price

c.

```r
reg2
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Coefficients:
## (Intercept)        Price     UrbanYes        USYes
##    13.04347     -0.05446     -0.02192      1.20057
```

```r
reduced <- lm(Sales ~ Price)
anova(reg2, reduced)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    396 2420.8
## 2    398 2552.2 -2   -131.41 10.748 2.848e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

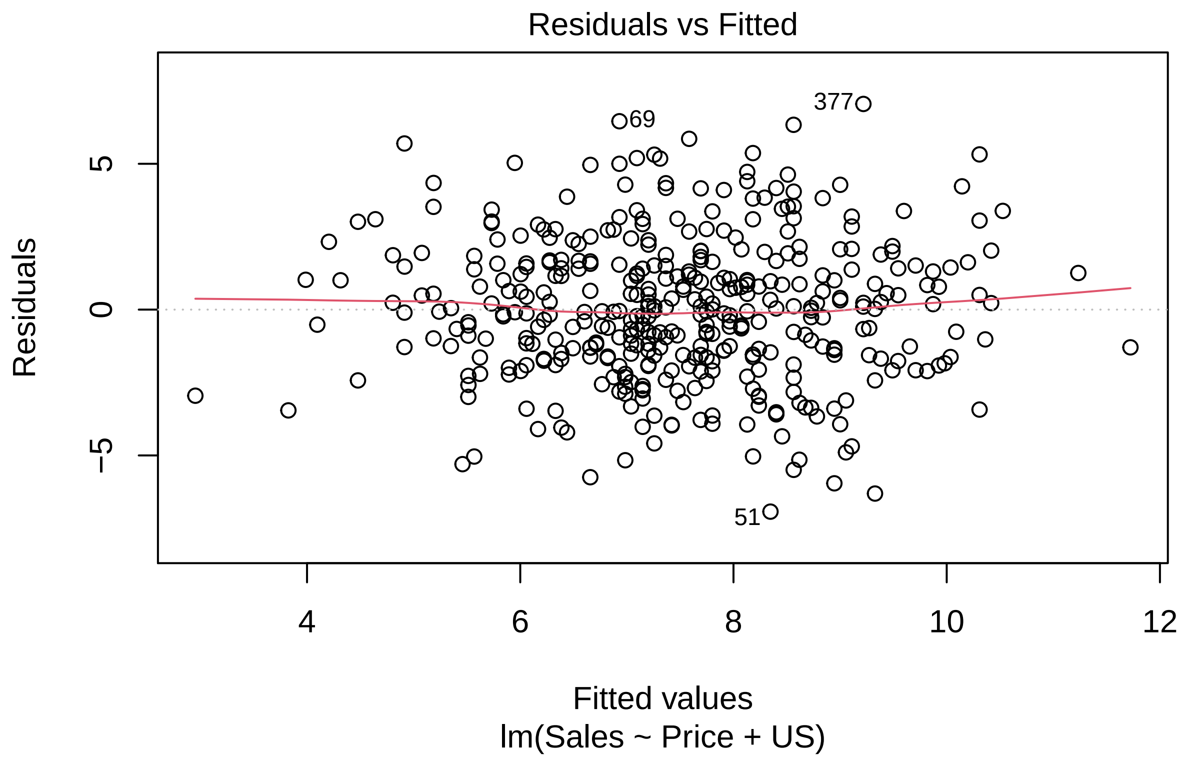
From the t-test (in the summary) we find that the UrbanYes variable is the only value where the null hypotheses is not rejected. From the anova test, we see that the full model with all variables included provides a better fit.

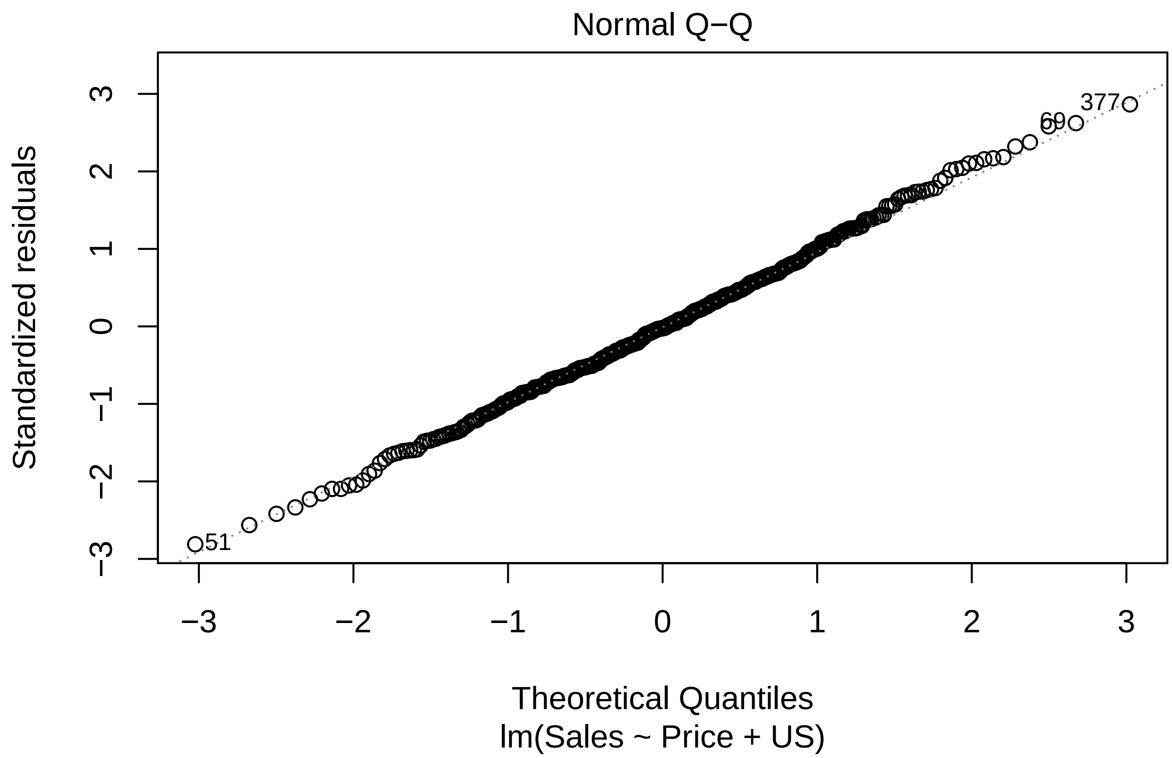
d.

```r
reg2 <- lm(Sales ~ Price + US, data = Carseats)
summary(reg2)
```

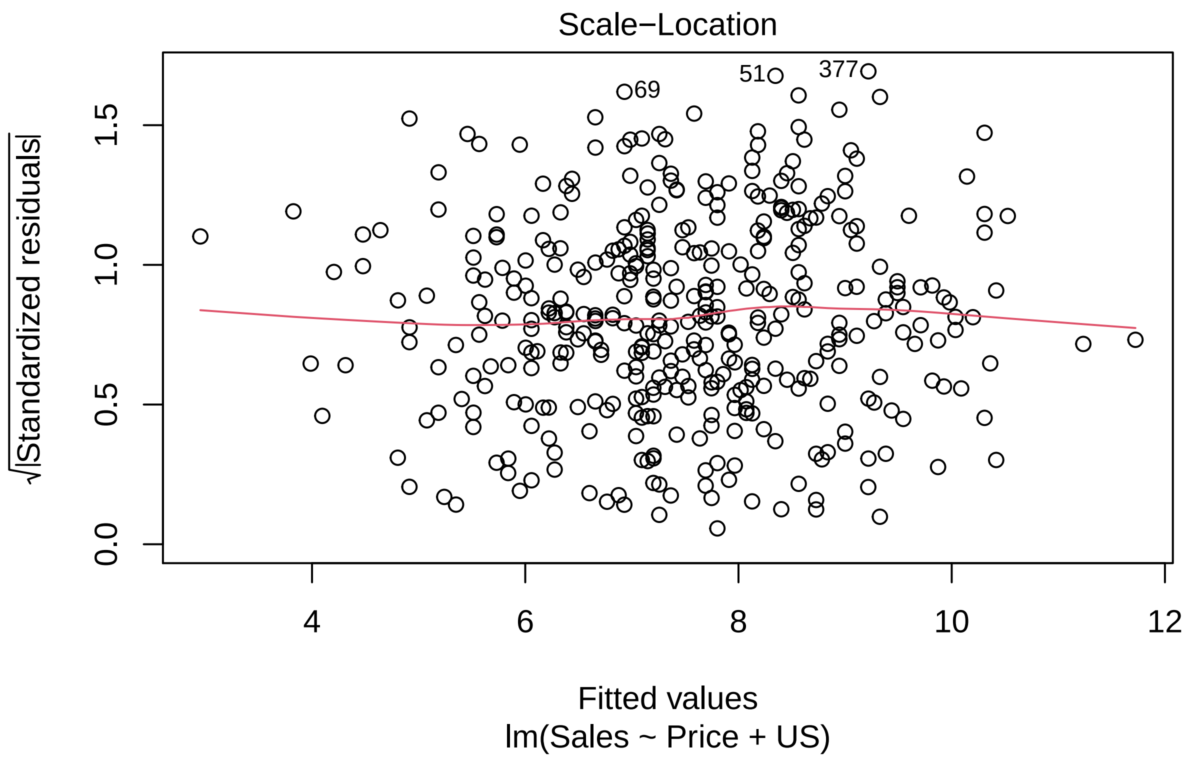
```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

```r
plot(reg2)
```

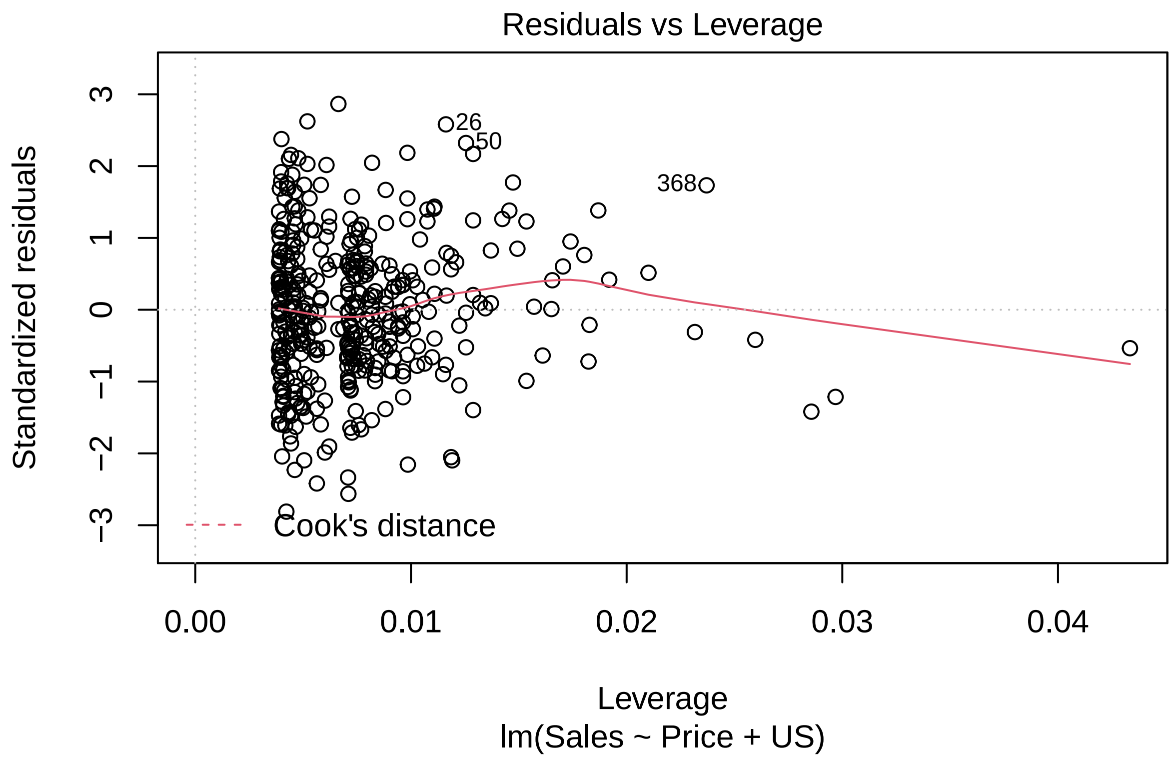

Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ Price + US)
5
0
−5
4
6
8
10
12
69
377
51



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Price + US)
377
69
51
−3
−2
−1
0
1
2
3



Scale−Location
69
51
377
0.0
0.5
1.0
1.5
√|Standardized residuals|
4
6
8
10
12
Fitted values
lm(Sales ~ Price + US)



According to the diagnostic charts, the data seems to be homoscedastic (from the third graph), linear (from the first graph) and normal (from the second graph). The fourth graph indicates that there is at least on point of high leverage that could be a potential outlier.

```r
#install.packages("car")
library(car)
```

```
## Loading required package: carData
```

```r
vif(reg2)
```

```
##    Price       US
## 1.003359 1.003359
```

Both of these variables are not multicolinear because the values are close to 1.

g.

```r
outlierTest(reg2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 377 2.891521          0.0040452           NA
```

```r
outlierTest(reduced)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 377 3.003268          0.0028401           NA
```

From the outlier tests on some of our models we found that there are no outliers in the data.

h.

```r
USyes = 1*(US == "Yes")

shapiro.test(reg2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg2$residuals
## W = 0.99799, p-value = 0.9199
```

```r
ncvTest(reg2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.566777, Df = 1, p = 0.21068
```

Because the p-values for both of the tests are high, we can conclude that the data is homoscedastic and normal

i.

```r
regh <- lm(Sales~ Price)
summary(regh)
```

```
##
## Call:
## lm(formula = Sales ~ Price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5224 -1.8442 -0.1459  1.6503  7.5108
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.641915   0.632812  21.558   <2e-16 ***
## Price       -0.053073   0.005354  -9.912   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.532 on 398 degrees of freedom
## Multiple R-squared:  0.198,  Adjusted R-squared:  0.196
## F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

This indicates a lack of fit since the R2 value is very low (.198)