CS498 AML HW6

Pengyu Cheng

10/19/2018

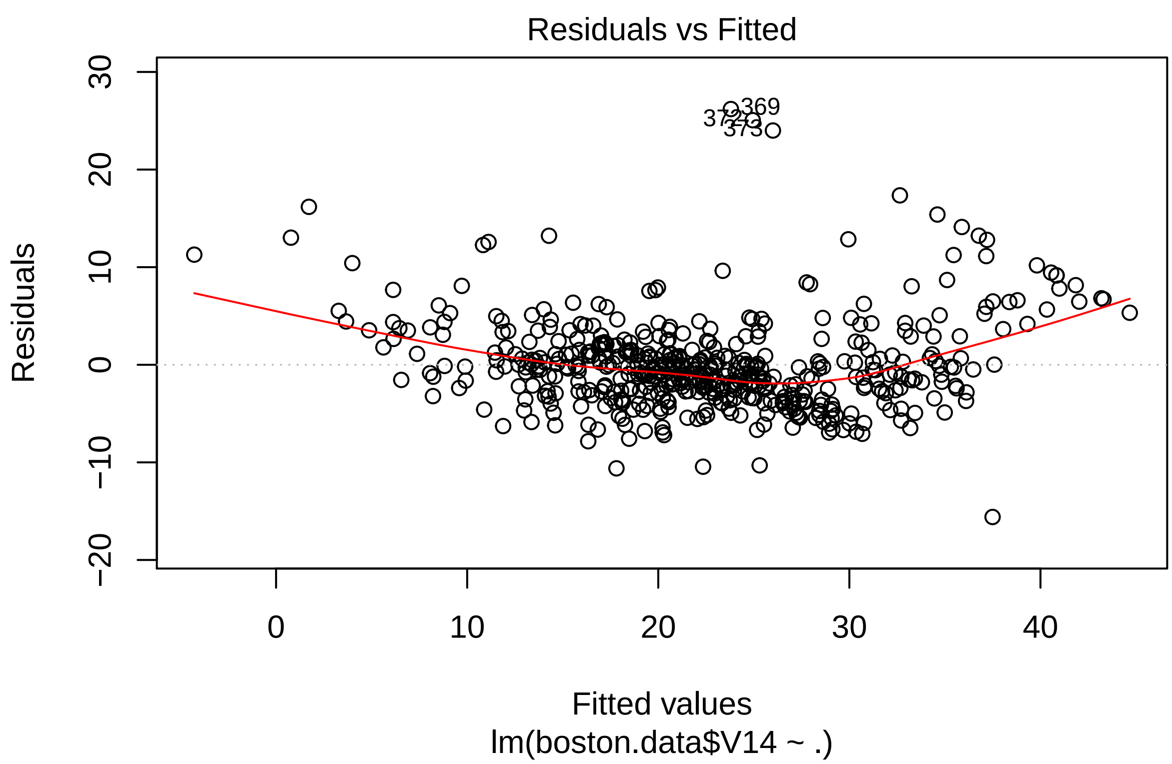
Question a): Regress house price (variable 14) against all others, and use leverage,

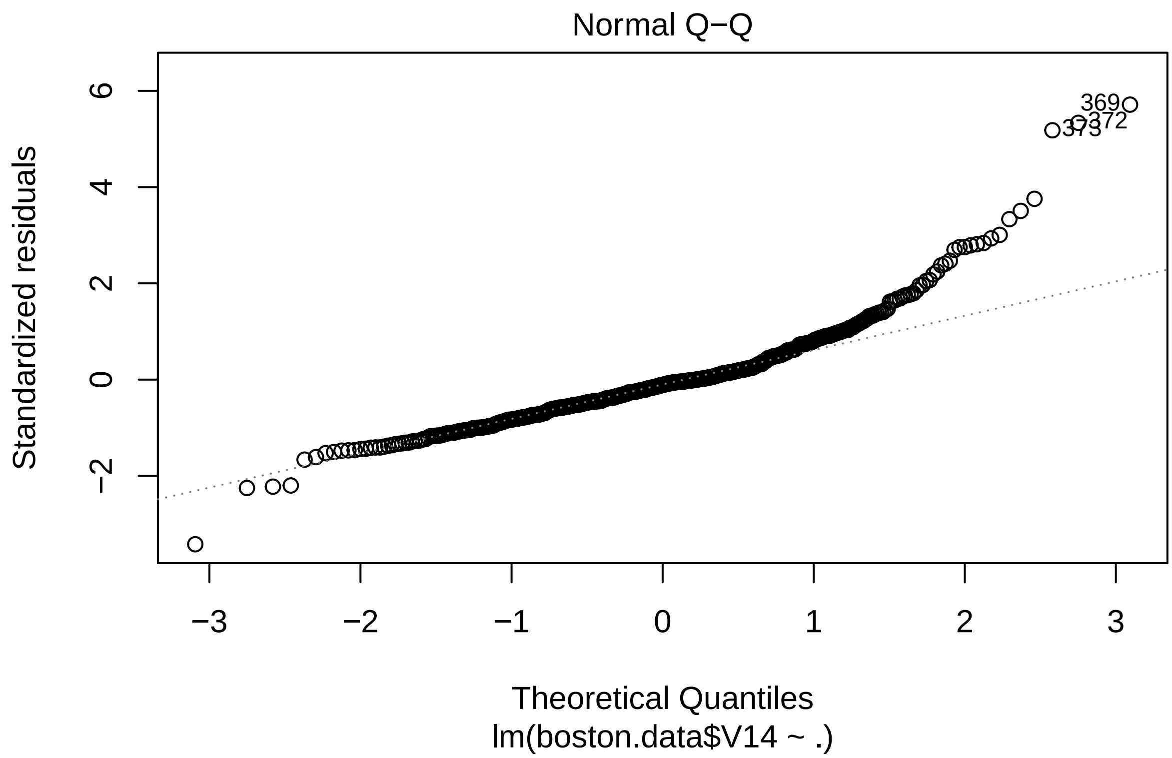
Cook's distance, and standardized residuals to find possible outliers. Produce a diagnostic plot that allows you to identify possible outliers

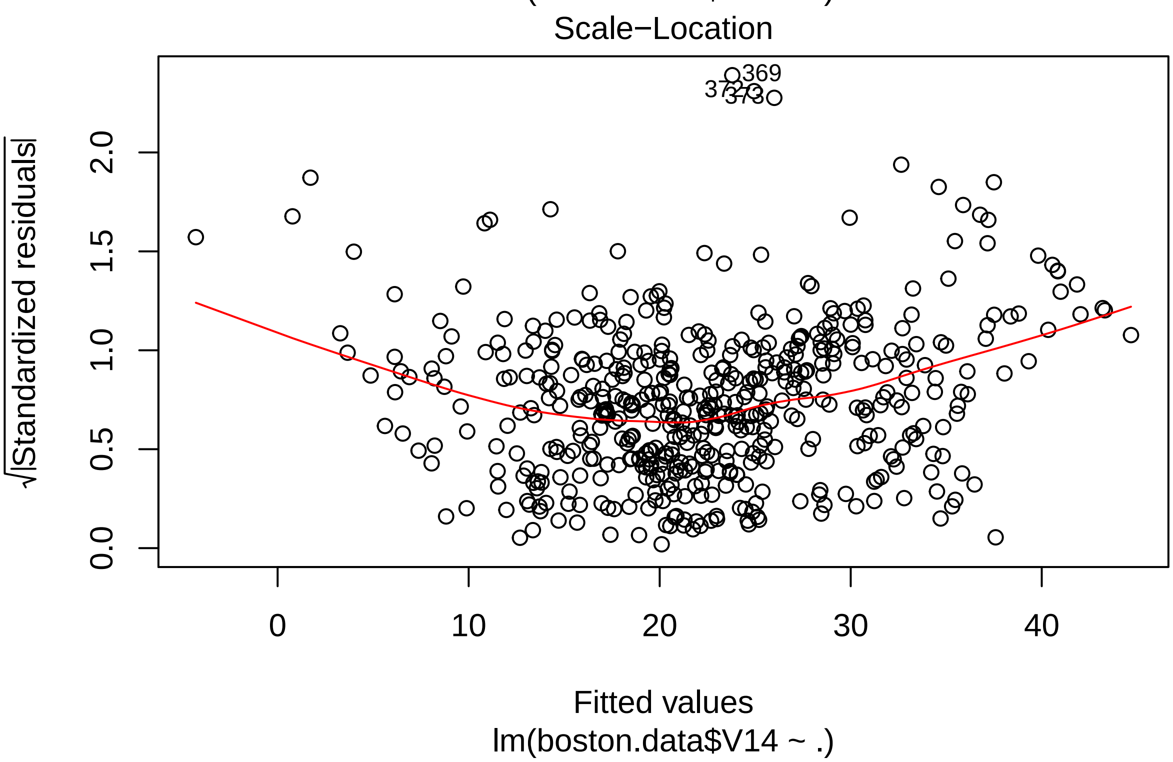
```r
library(data.table)
boston.data = fread('https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data')
head(boston.data)
```

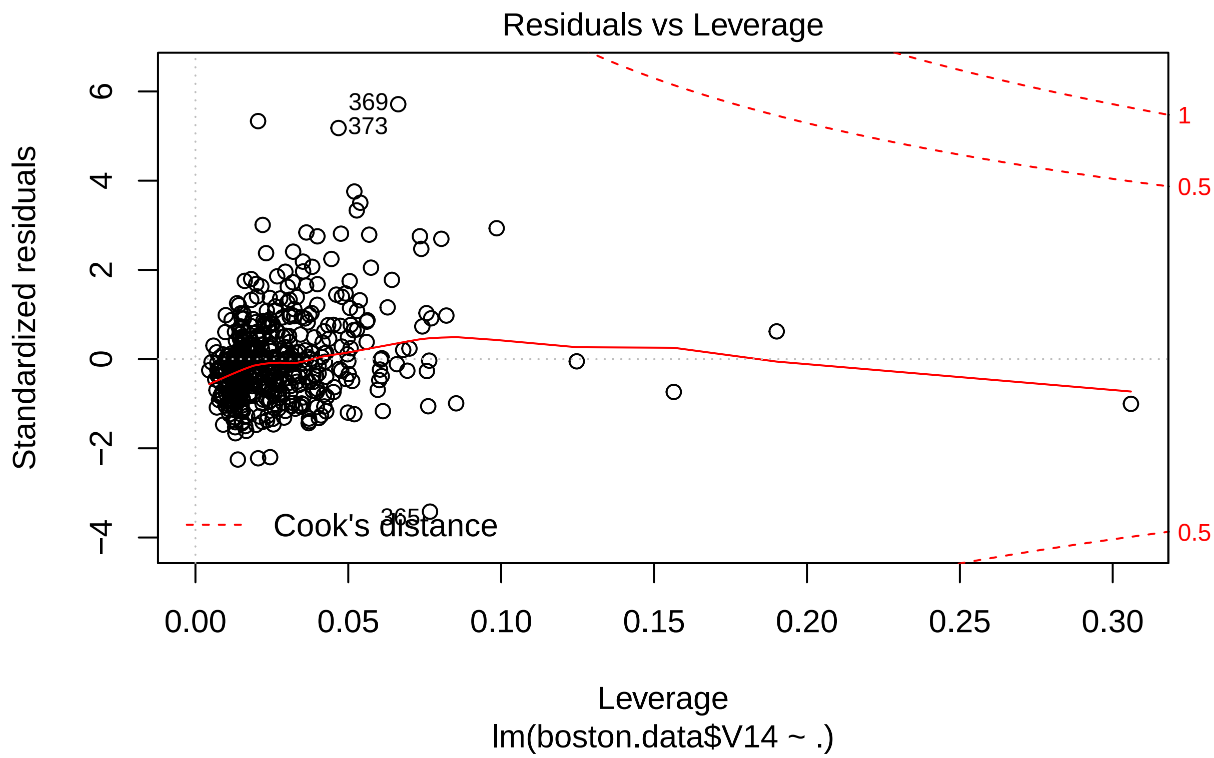
```
##         V1 V2   V3 V4    V5    V6   V7     V8 V9 V10  V11    V12  V13  V14
## 1: 0.00632 18 2.31  0 0.538 6.575 65.2 4.0900  1 296 15.3 396.90 4.98 24.0
## 2: 0.02731  0 7.07  0 0.469 6.421 78.9 4.9671  2 242 17.8 396.90 9.14 21.6
## 3: 0.02729  0 7.07  0 0.469 7.185 61.1 4.9671  2 242 17.8 392.83 4.03 34.7
## 4: 0.03237  0 2.18  0 0.458 6.998 45.8 6.0622  3 222 18.7 394.63 2.94 33.4
## 5: 0.06905  0 2.18  0 0.458 7.147 54.2 6.0622  3 222 18.7 396.90 5.33 36.2
## 6: 0.02985  0 2.18  0 0.458 6.430 58.7 6.0622  3 222 18.7 394.12 5.21 28.7
```

```r
model = lm(boston.data$V14~., data = boston.data)
plot(model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(boston.data$V14 ~ .)
369
372
373

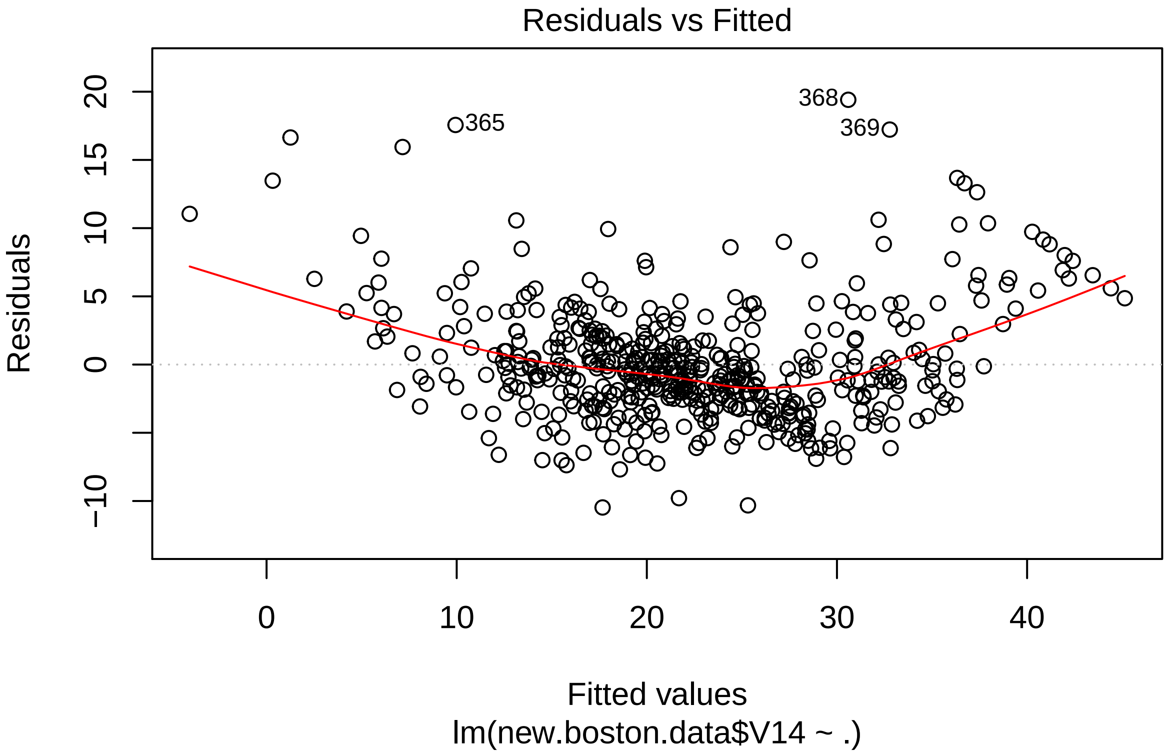

Scale−Location
√|Standardized residuals|
Fitted values
lm(boston.data$V14 ~ .)
369
372
373

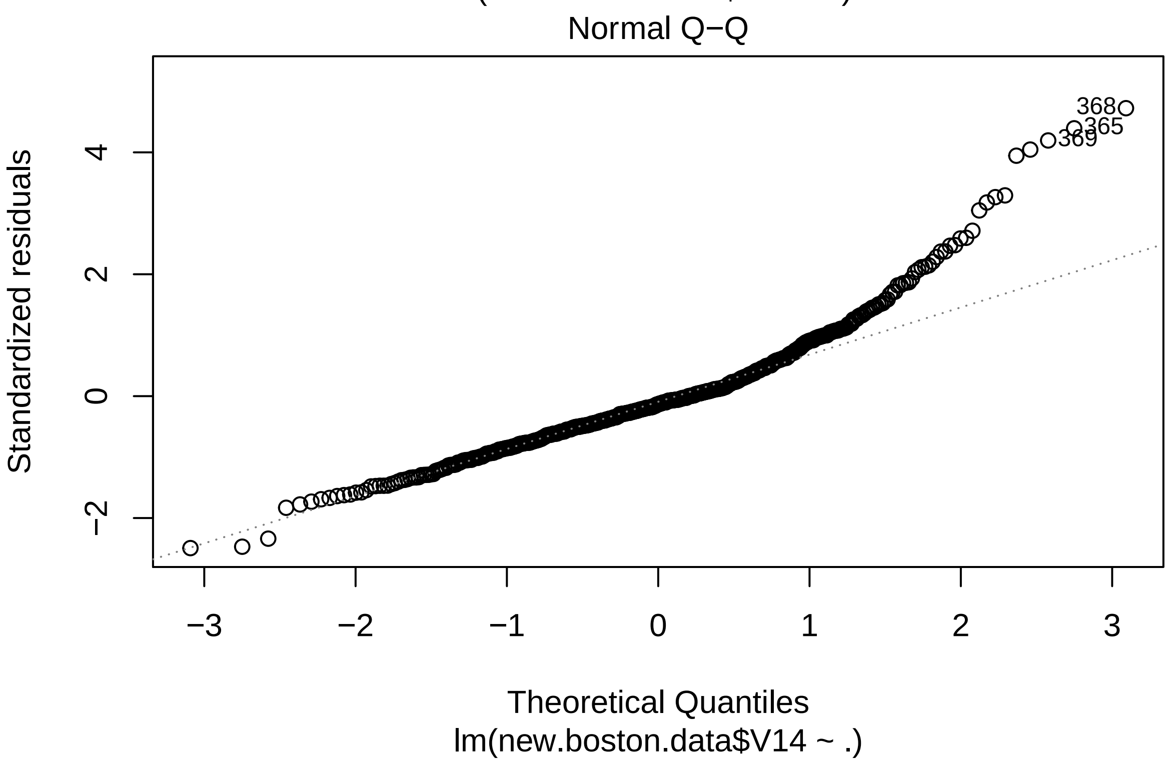


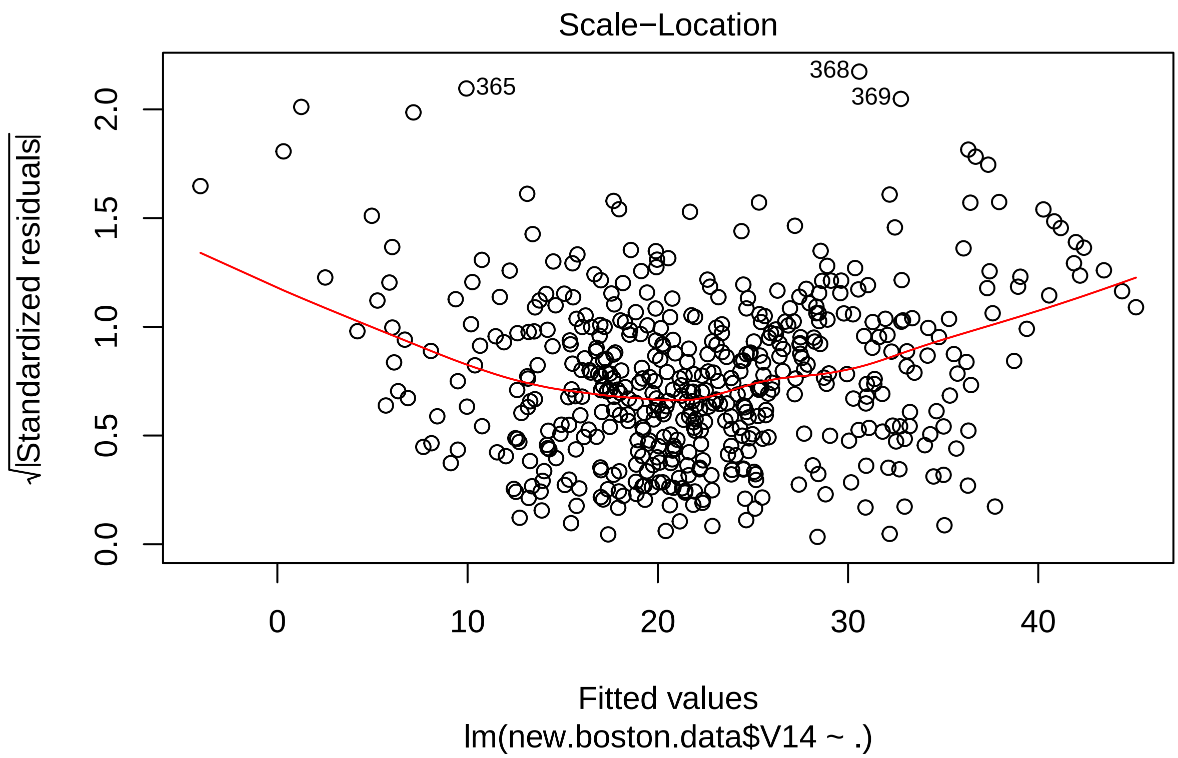
From the "residual vs fitted" plot we can see there exists some non-linearity replationship and point 369, 372 and 373 are identified as possible outliers. We refer to "scale-location" for further information. In this plot, we find that points 373, 372 and 369 are indeed beyond 2 standard deviations;thus we feel comfortable to flag them as possible outliers. From the "residuals vs leverage" plot, we identify another point as possible outlier: point 365 with suspiciously large Cook's distance. There are points that have high leverage but small residuals and we leave it as it is for now.

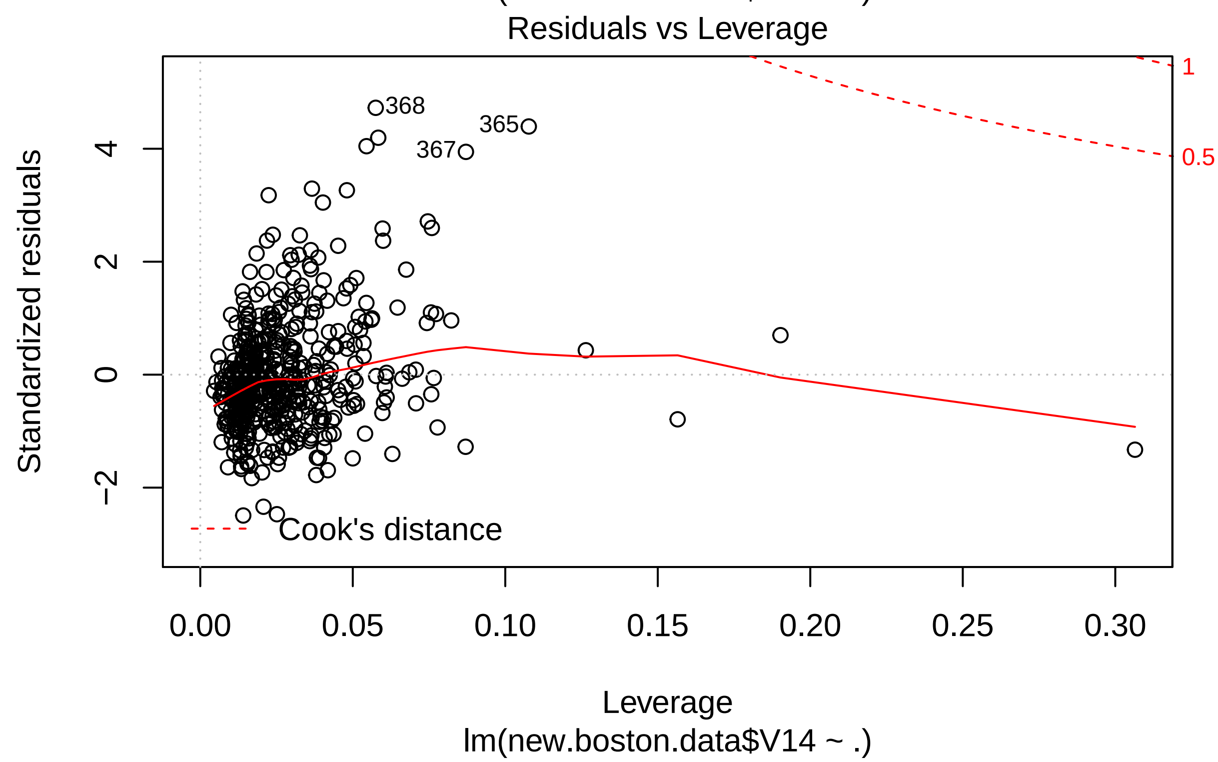
Question b): Remove all the points you suspect as outliers, and compute a new regression. Produce a diagnostic plot that allows you to identify possible outliers.

```
#remove data points 373, 369 365 and 372
new.boston.data = boston.data[-c(373, 372, 369, 365),]
new.model = lm(new.boston.data$V14~., data=new.boston.data)
plot(new.model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(new.boston.data$V14 ~ .)
368
369
365
−10
0
5
10
15
20
0
10
20
30
40


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(new.boston.data$V14 ~ .)
368
365
369
−2
0
2
4
−3
−2
−1
0
1
2
3

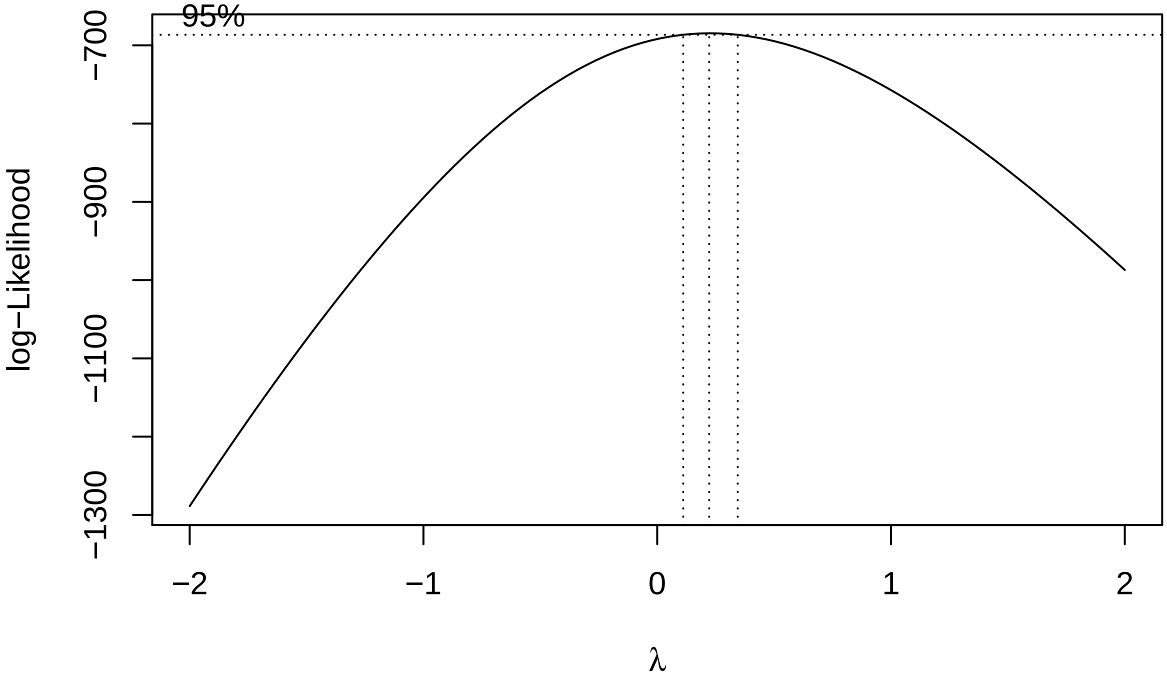




Question c): Apply a Box-Cox transformation to the dependent variable – what is the

best value of the parameter?

```r
library(MASS)
boxcox(new.model)
boxcox.transform = boxcox(new.model)
```




```r
best.lambda = boxcox.transform$x[which(boxcox.transform$y == max(boxcox.transform$y))]
```

The best parameter of λ is 0.2222222