ML_sr55737

Sanjhana Rangaraj

2023-07-22

Libraries

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine

## Loading required package: lattice

## Loaded gbm 2.1.8.1

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## Loading required package: nnet

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##     cluster

## Loading required package: Matrix

## Loaded glmnet 4.1-7

##
## Attaching package: 'pls'

## The following object is masked from 'package:caret':
##
##     R2

## The following object is masked from 'package:stats':
##
##     loadings

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:ISLR2':
##
##     Boston

## corrplot 0.92 loaded

##
## Attaching package: 'corrplot'

## The following object is masked from 'package:pls':
##
##     corrplot
```

Chapter 2 Question 9

PART A

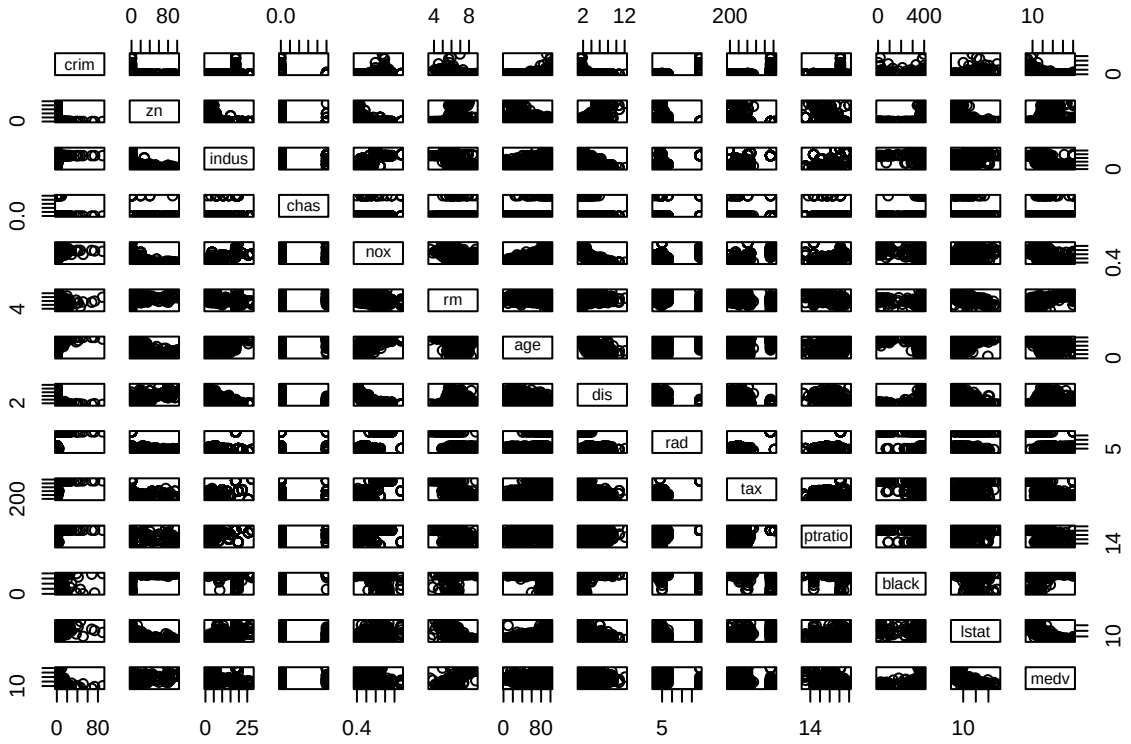
How many rows are in this data set? How many columns? What do the rows and columns represent?

Boston is a data set containing housing values in 506 suburbs of Boston with 13 feature variables.

506 Rows 13 Columns

PART B

Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



We observe the following correlations with these variables

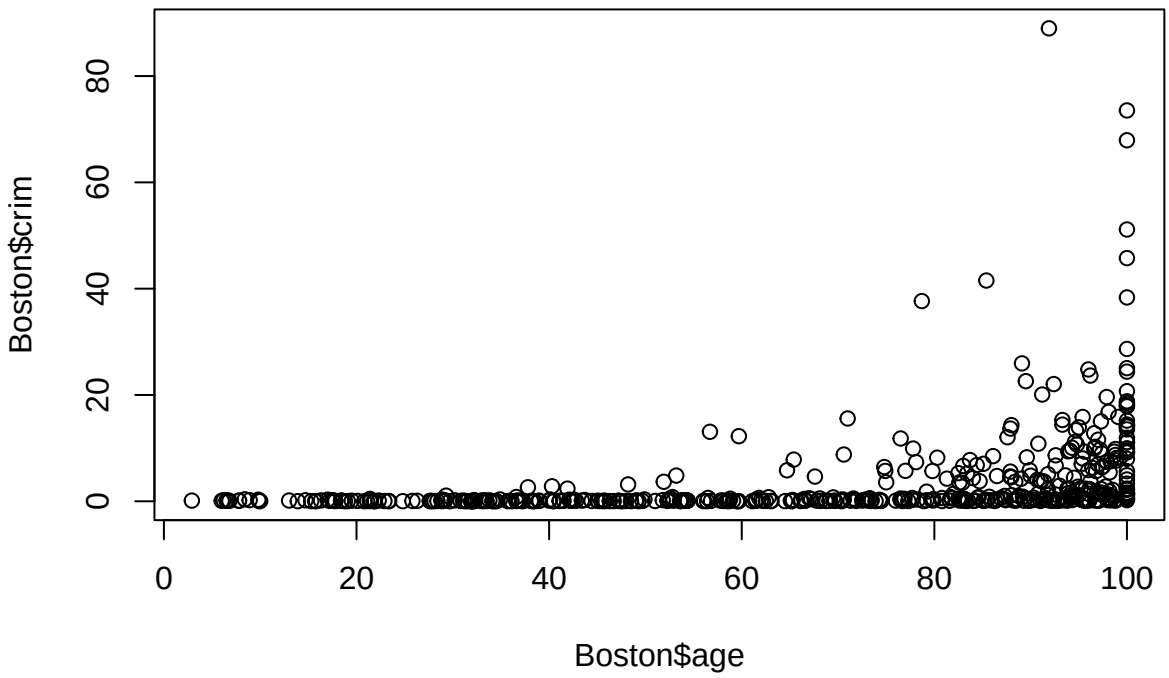
crim is related to age, dis, rad, tax, ptratio

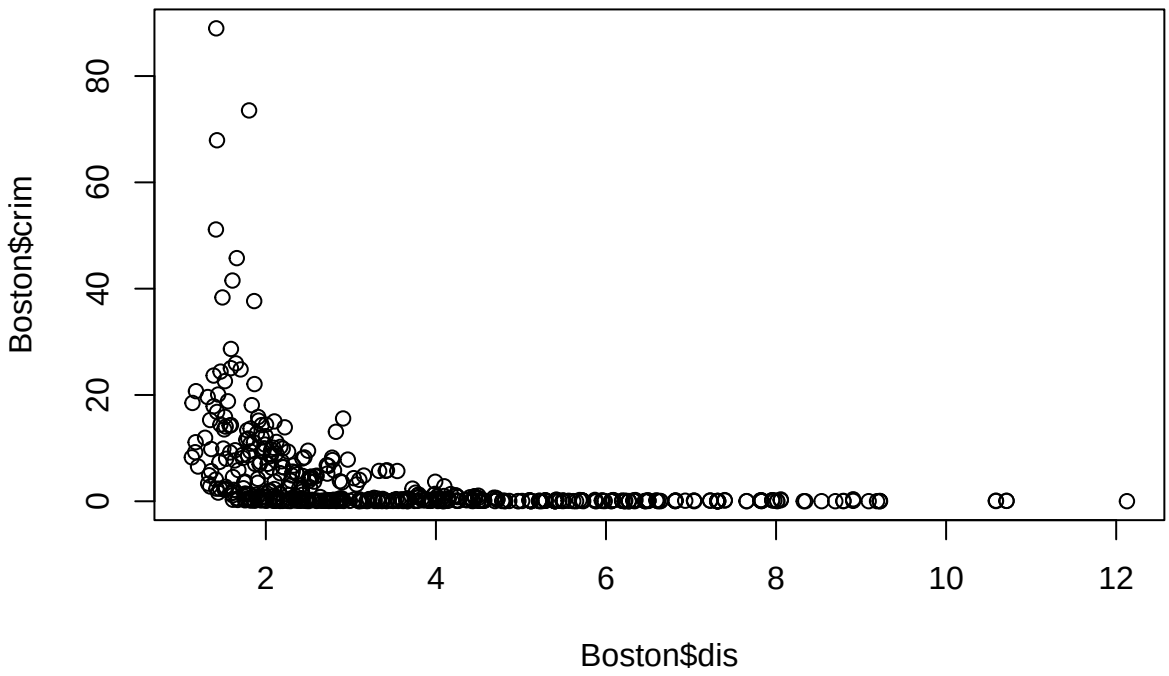
rm and medv are closely related

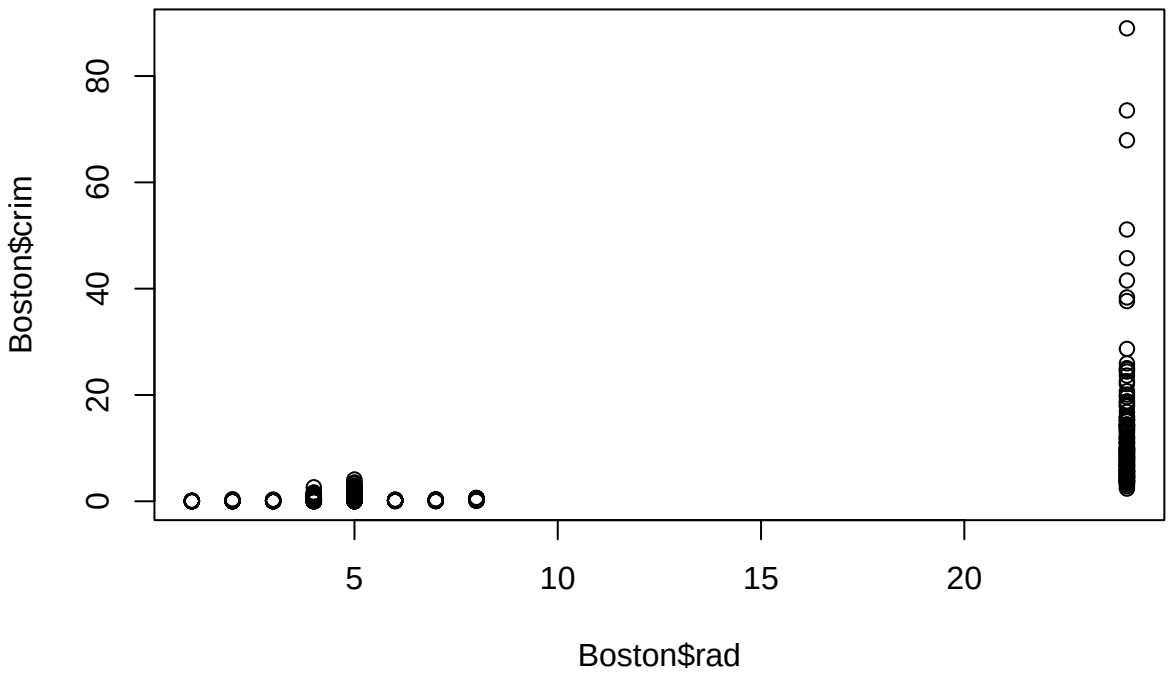
The overall observation is although some correlations can be spotted, to get a clear picture we need to do plot correlation matrix.

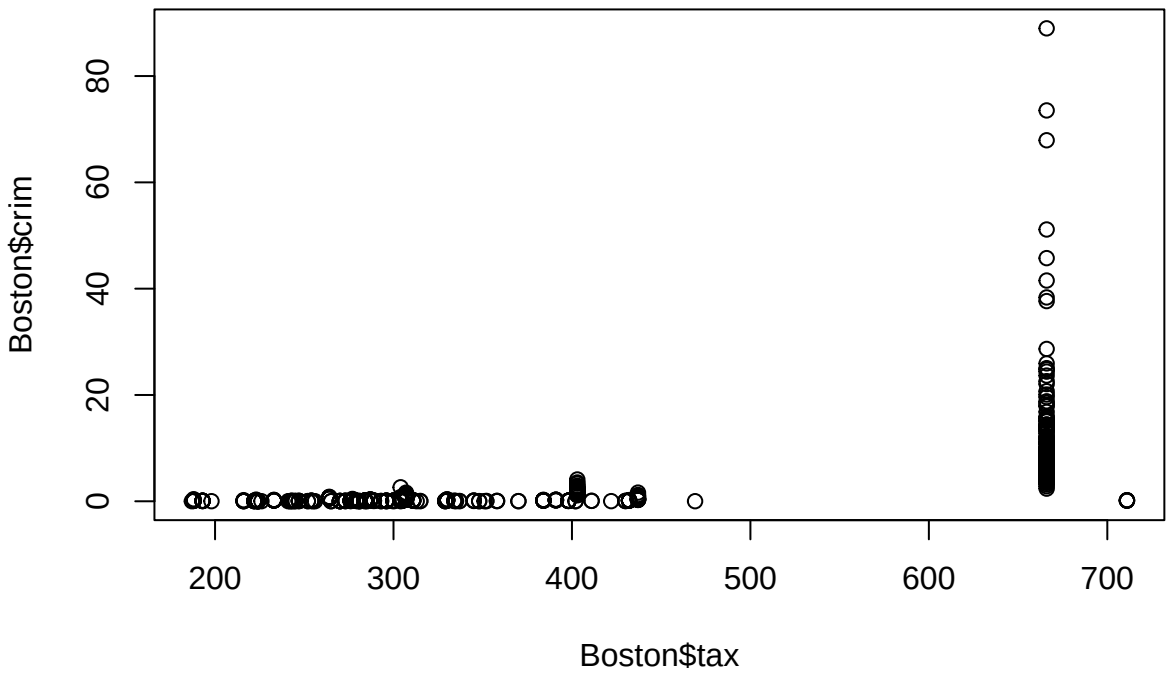
PART C

Are any of the predictors associated with per capita crime rate? If so, explain the relationship.


Boston$crim
0
20
40
60
80
0
20
40
60
80
100
Boston$age


Boston$crim
0
20
40
60
80
2
4
6
8
10
12
Boston$dis


Boston$crim
0
20
40
60
80
5
10
15
20
Boston$rad



We made the following observations:
Older homes, more crime

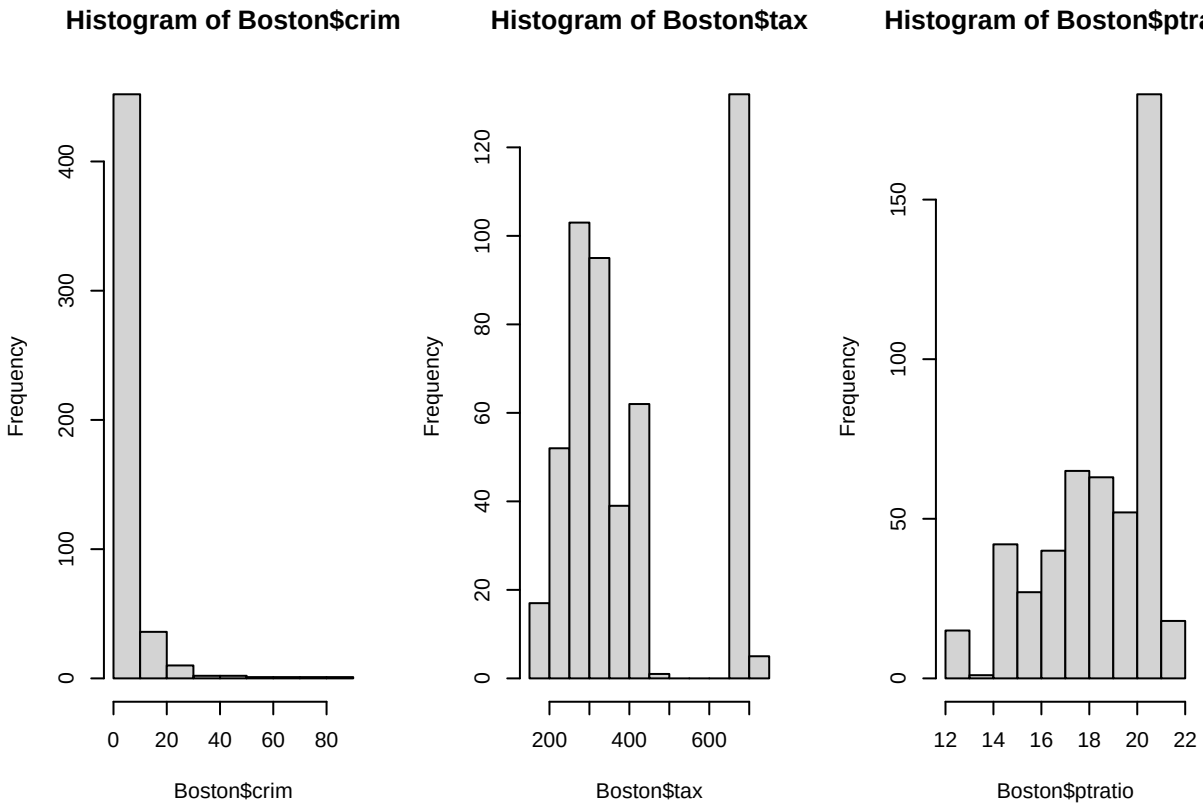
Closer to work-area, more crime

Higher index of accessibility to radial highways, more crime

Higher tax rate, more crime

PART D

Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



Observations:

Most suburbs have low crime rates, but around 18 suburbs have a very high crime rate

There is a large divide between suburbs with low tax rates and high tax rates

A skew towards high ratios, but others lie in the same range

PART E

How many of the census tracts in this data set bound the Charles river?

```
## [1] 35 14
```

35 Suburbs

PART F

What is the median pupil-teacher ratio among the towns in this data set?

```
## [1] 19.05
```

Median pupil-teacher ratio is 19.05

PART G

Which census tract of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
##             399      406
## crim     38.3518  67.9208
## zn        0.0000   0.0000
## indus    18.1000  18.1000
## chas      0.0000   0.0000
## nox       0.6930   0.6930
## rm        5.4530   5.6830
## age     100.0000 100.0000
## dis       1.4896   1.4254
## rad      24.0000  24.0000
## tax     666.0000 666.0000
## ptratio  20.2000  20.2000
## black   396.9000 384.9700
## lstat    30.5900  22.9800
## medv      5.0000   5.0000
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

Observations corresponding to each variable

crim: above 3rd quartile

zn: at min

indus: at 3rd quartile

chas: not bounded by river

nox: bove 3rd quartile

rm: below 1st quartile

age: at max

dis: below 1st quartile

rad: at max

tax: at 3rd quartile

ptratio: at 3rd quartile

black: 399 at max; 406 above 1st quartile

lstat: above 3rd quartile

medv: at min

PART H

In this dataset, how many of the census tract saver age more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

```
## [1] 64 14
## [1] 13 14
##       crim                zn             indus              chas        
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm              age              dis        
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black       
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv      
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0

##       crim                zn             indus              chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
```

```
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

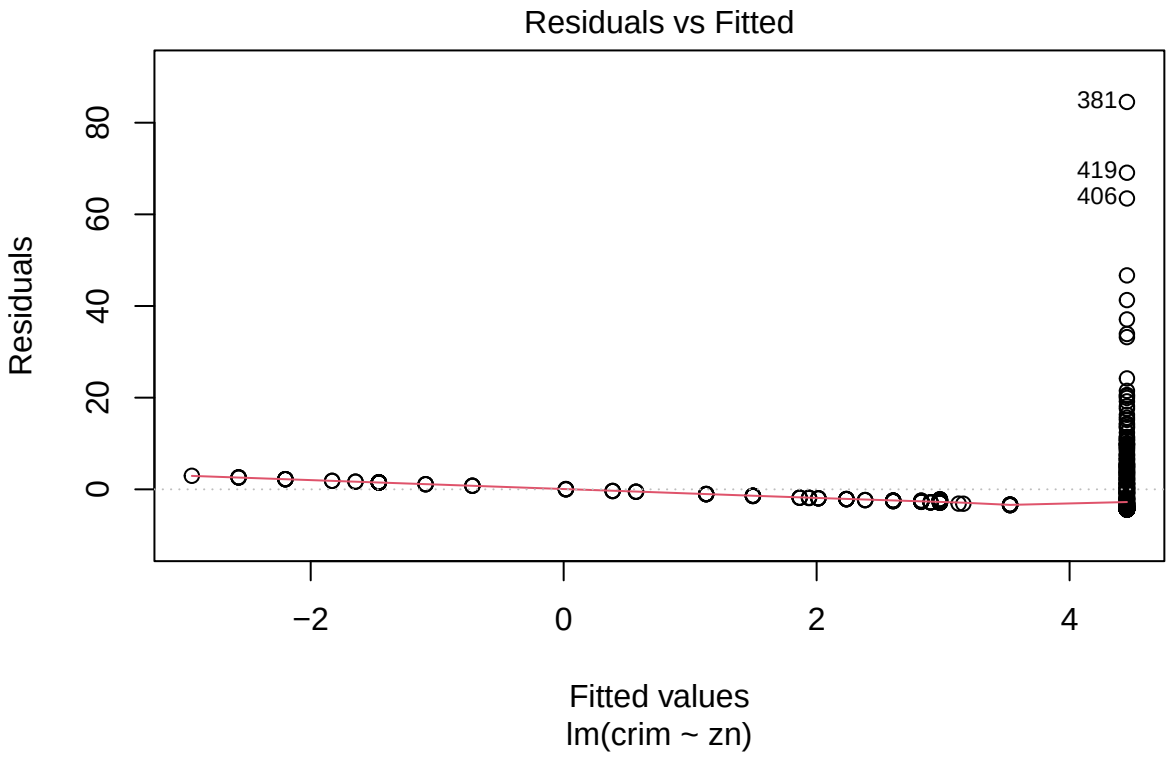
Observations

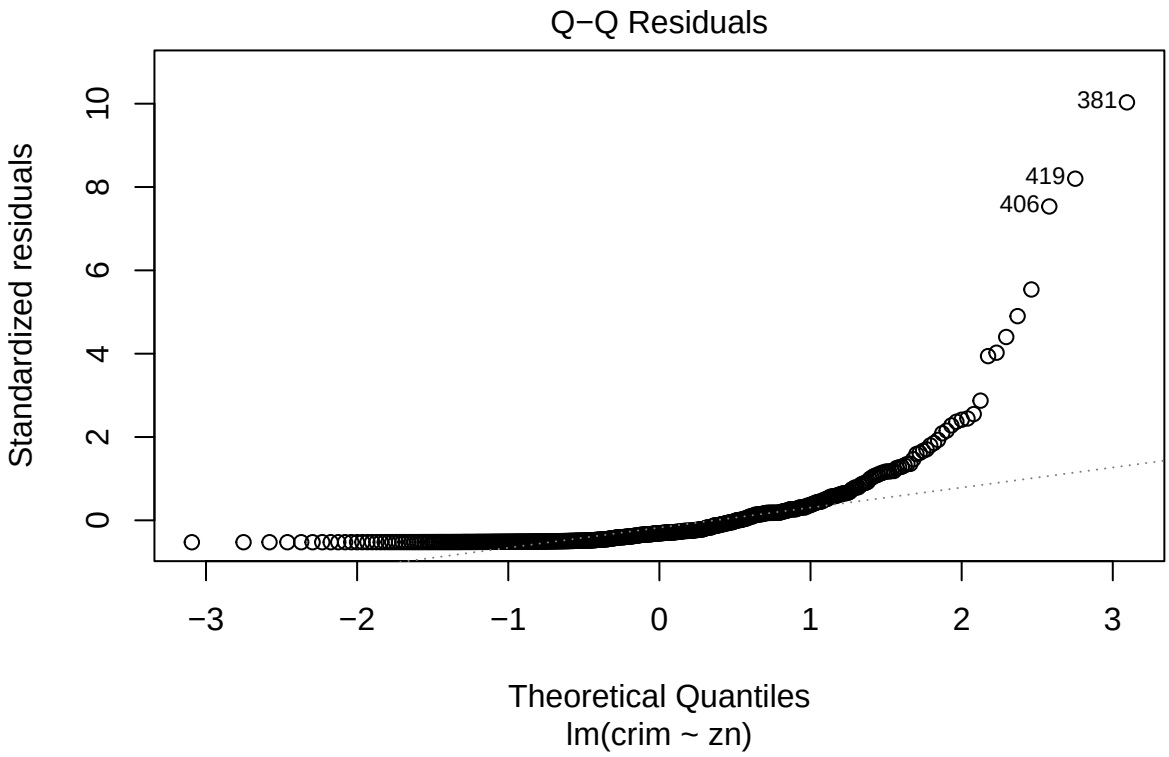
Relatively lower crime (comparing range), lower lstat (comparing range)

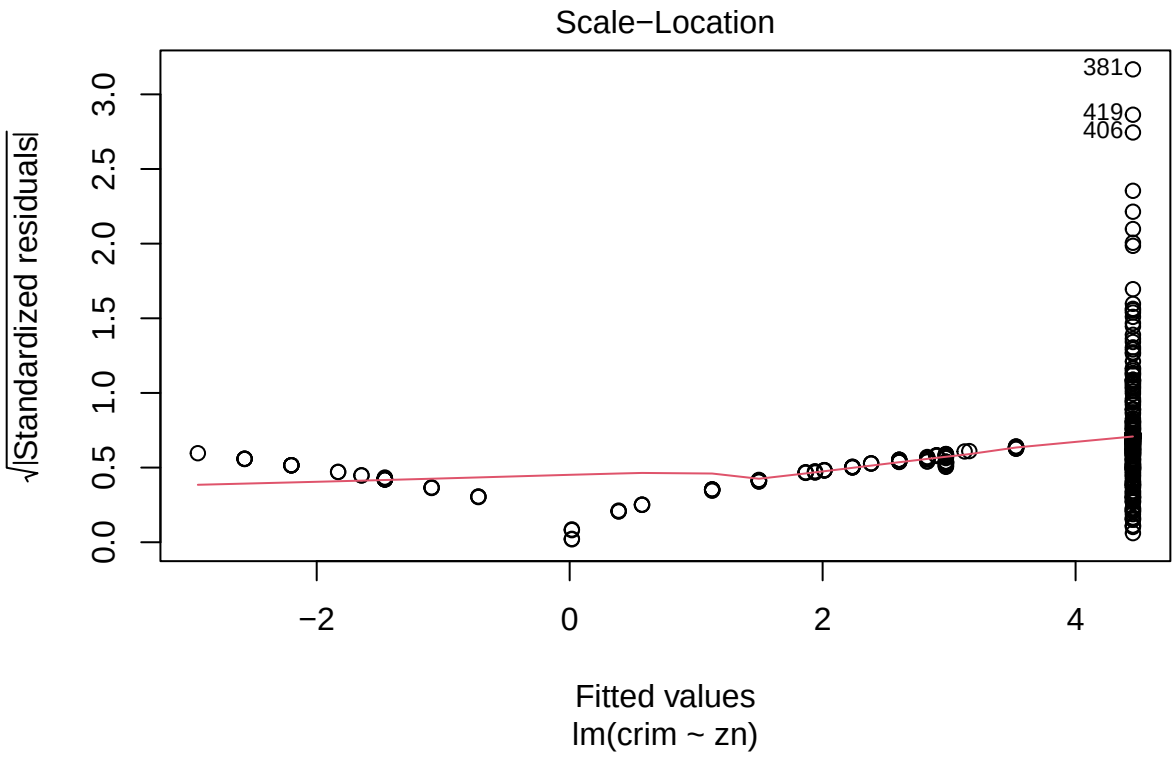
Chapter 3 Question 15

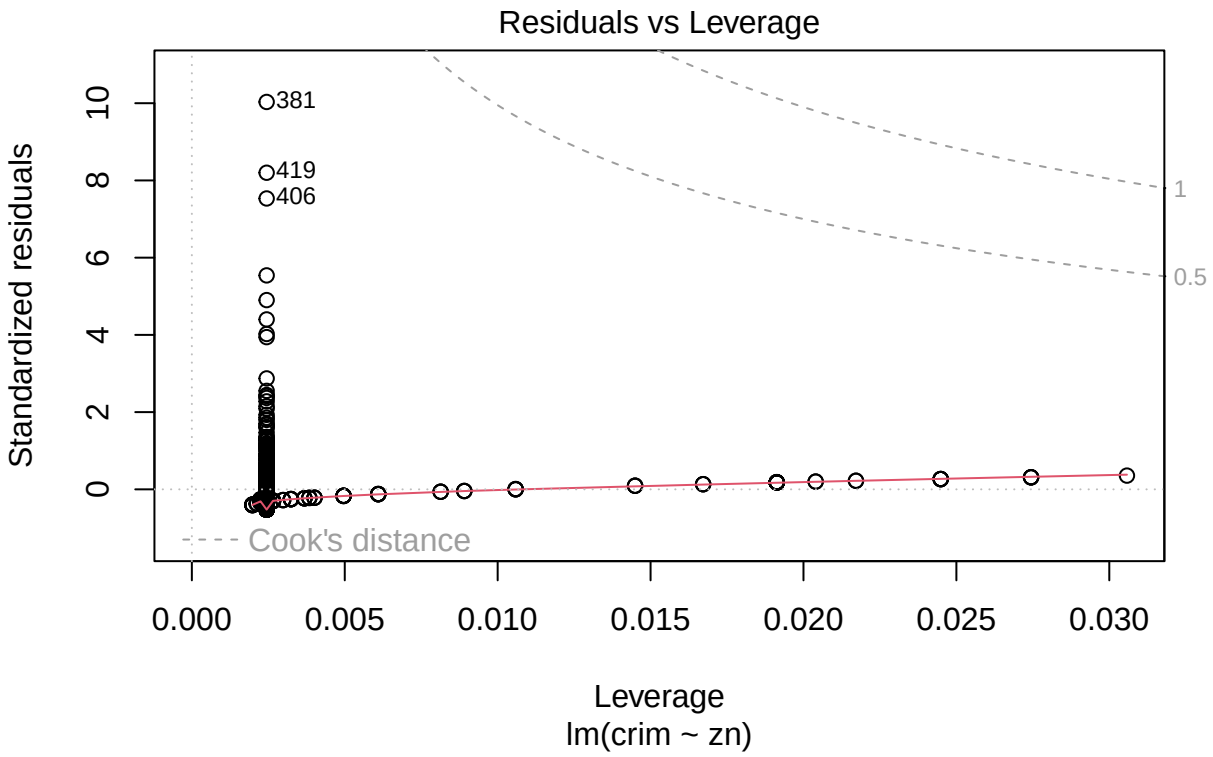
PART A

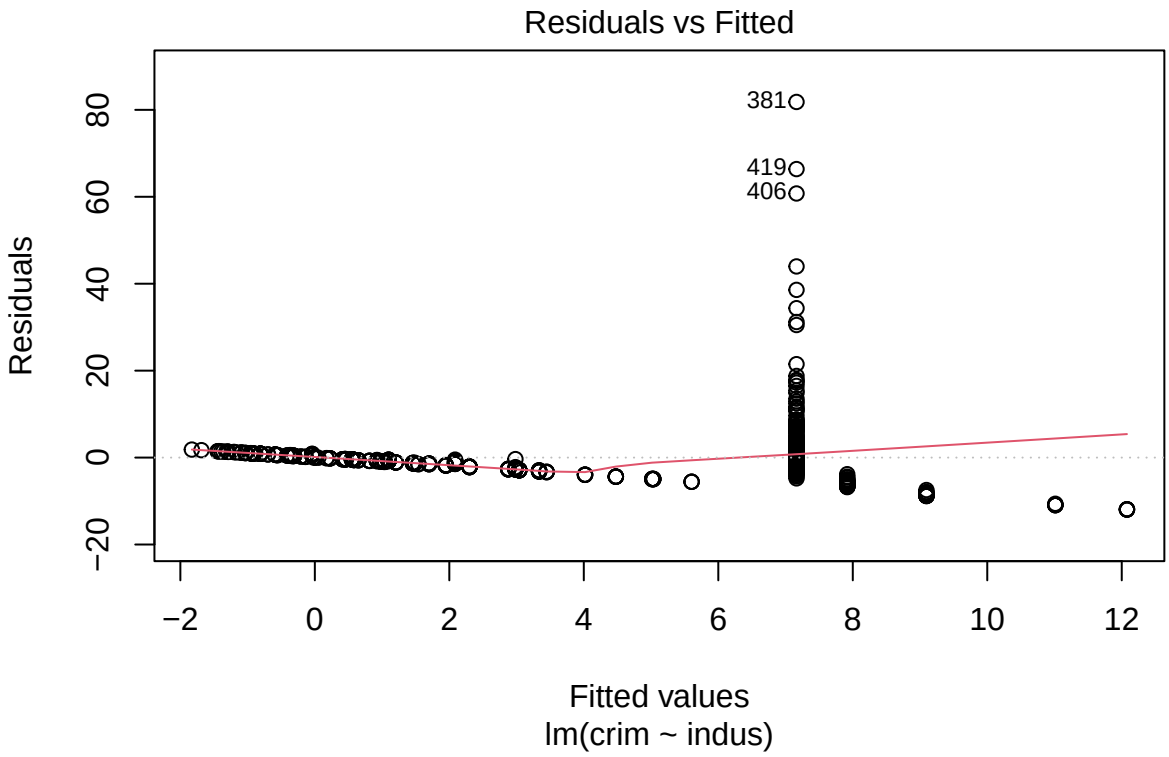
For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

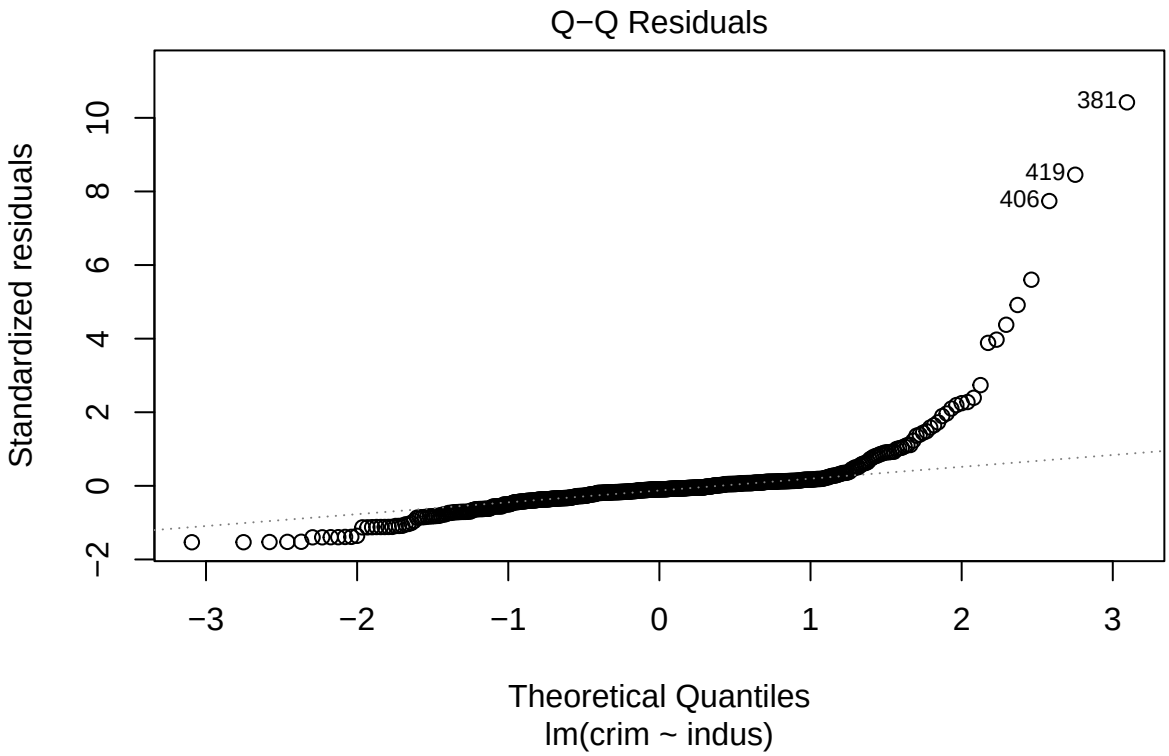

Residuals vs Fitted
Residuals
0
20
40
60
80
381
419
406
−2
0
2
4
Fitted values
lm(crim ~ zn)

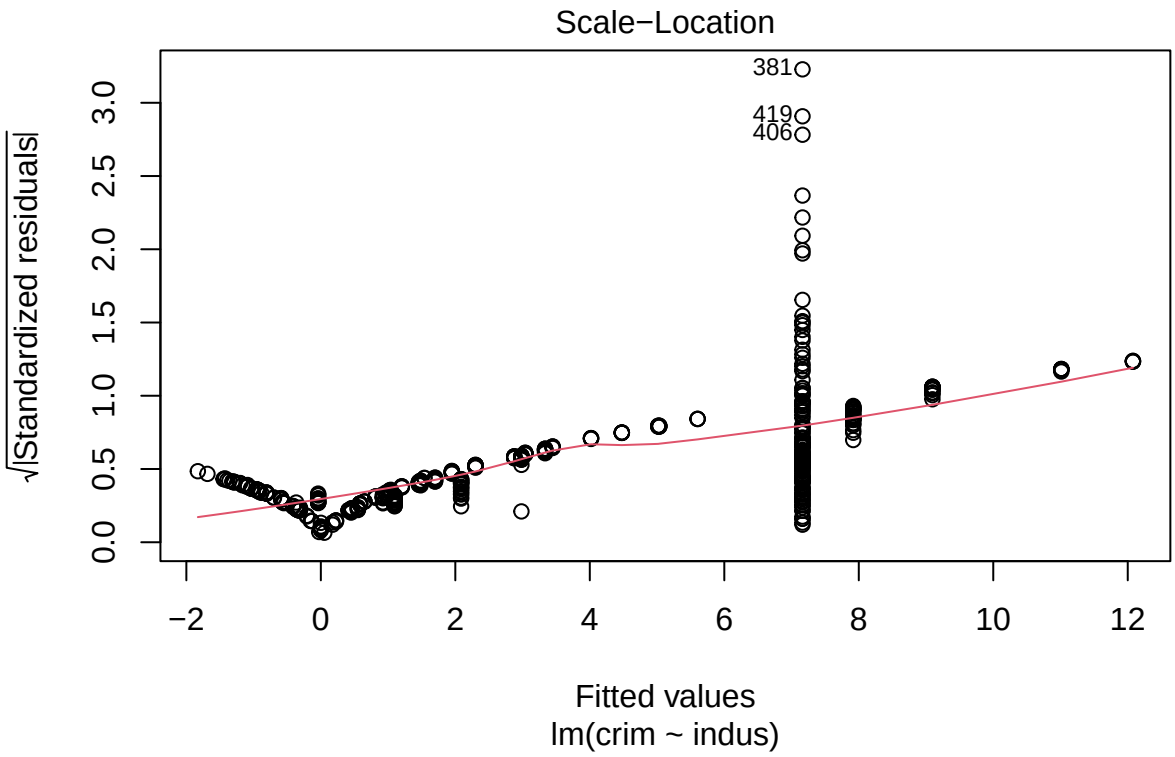

Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ zn)

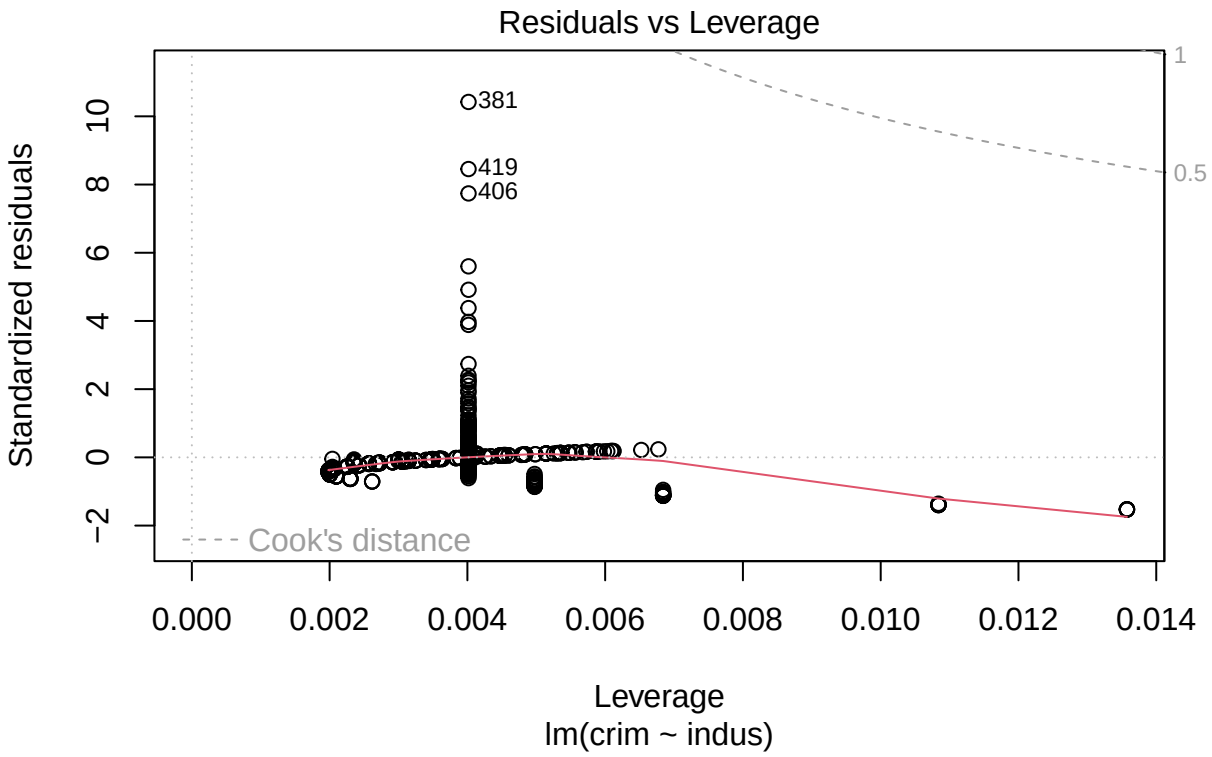

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
381
419
406
−2
0
2
4
Fitted values
lm(crim ~ zn)

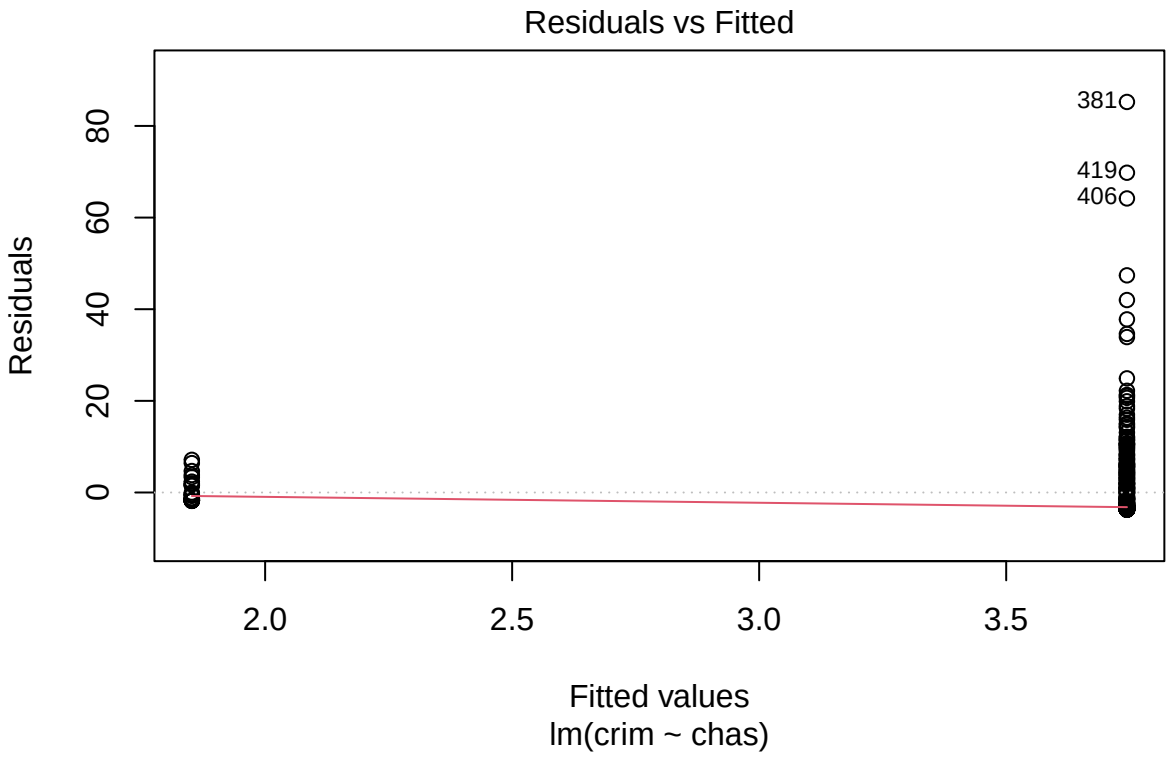

Residuals vs Leverage
Standardized residuals
Leverage
lm(crim ~ zn)
Cook's distance
1
0.5
381
419
406
0
2
4
6
8
10
0.000
0.005
0.010
0.015
0.020
0.025
0.030

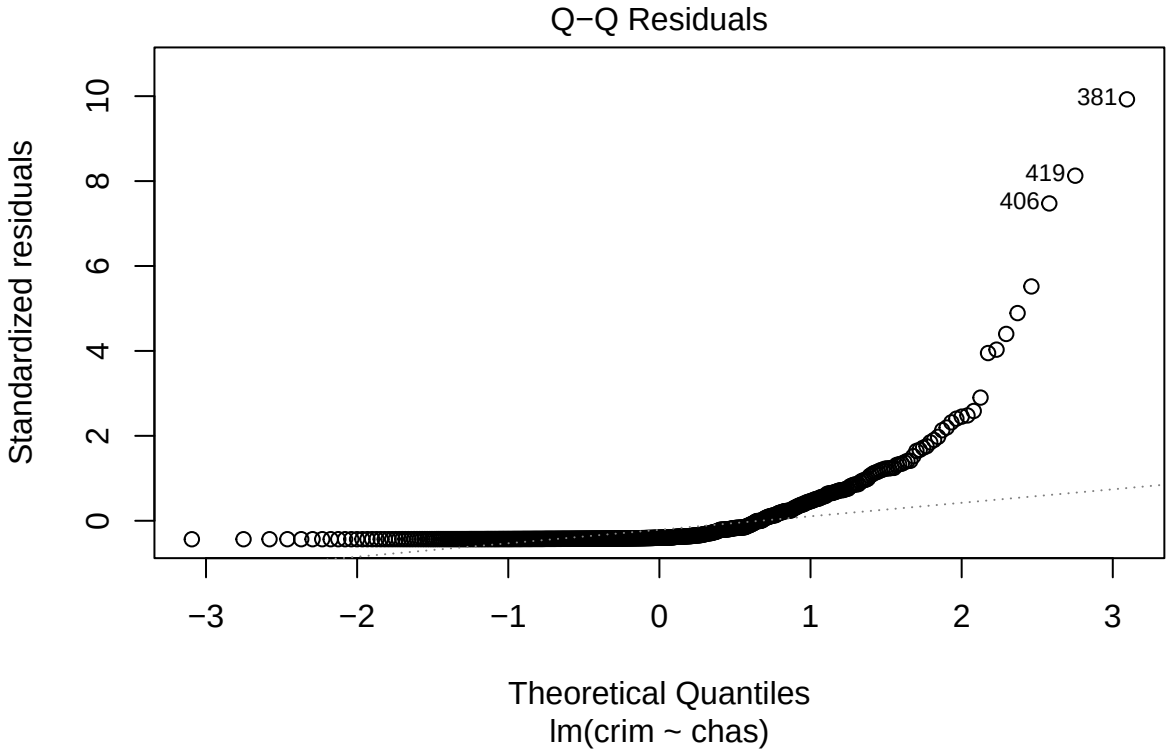

Residuals vs Fitted
Residuals
Fitted values
lm(crim ~ indus)
381
419
406
−20
0
20
40
60
80
−2
0
2
4
6
8
10
12

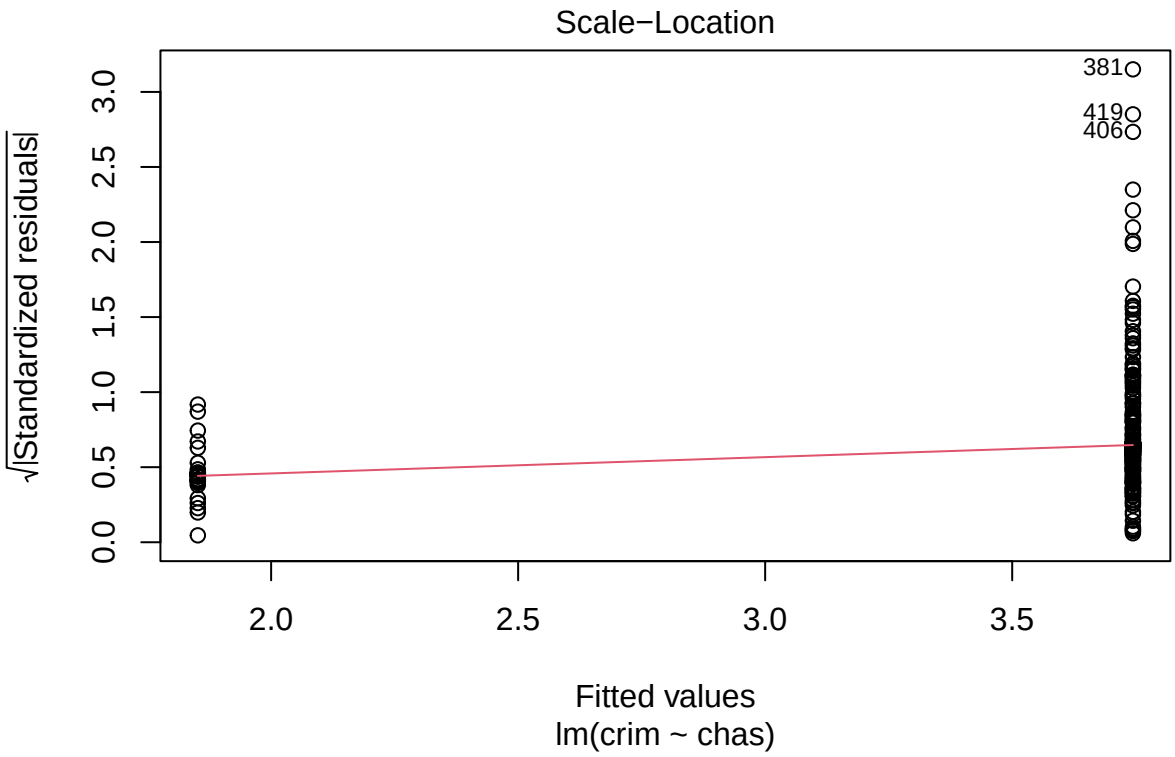

Q−Q Residuals
Standardized residuals
−2
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ indus)

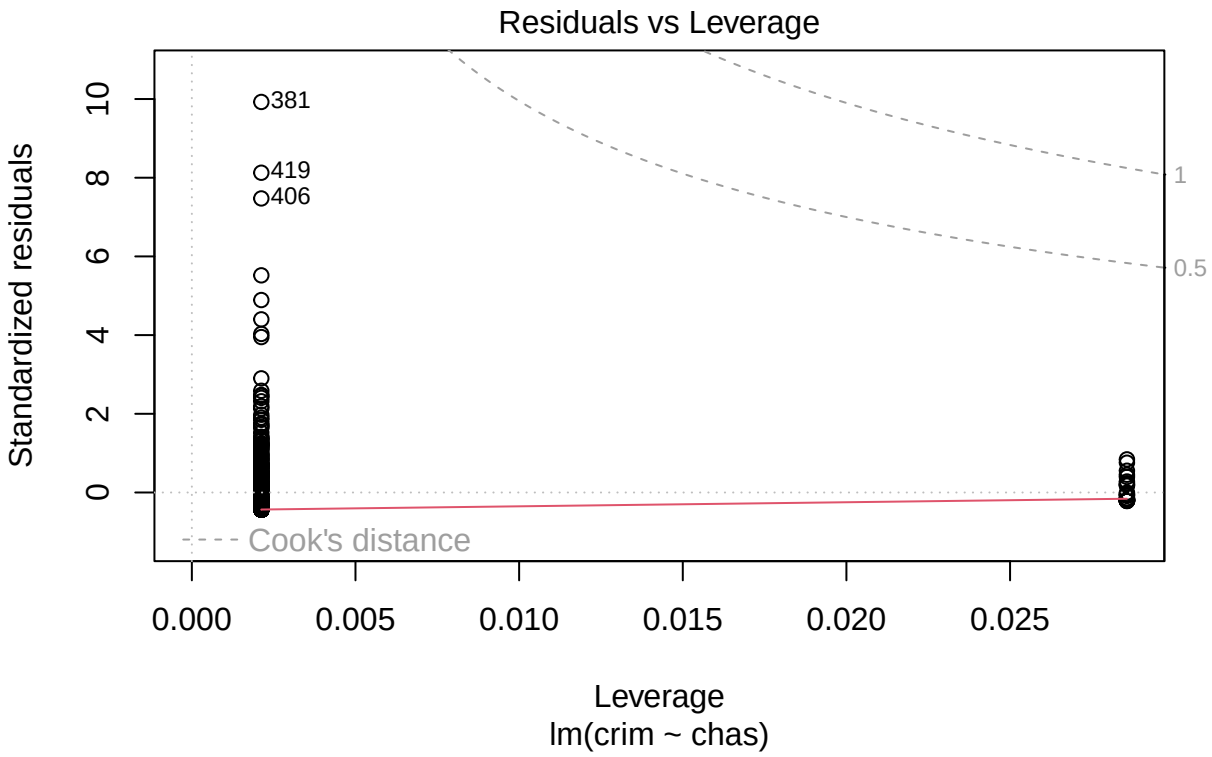

Scale−Location
381
419
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
0
2
4
6
8
10
12
Fitted values
lm(crim ~ indus)

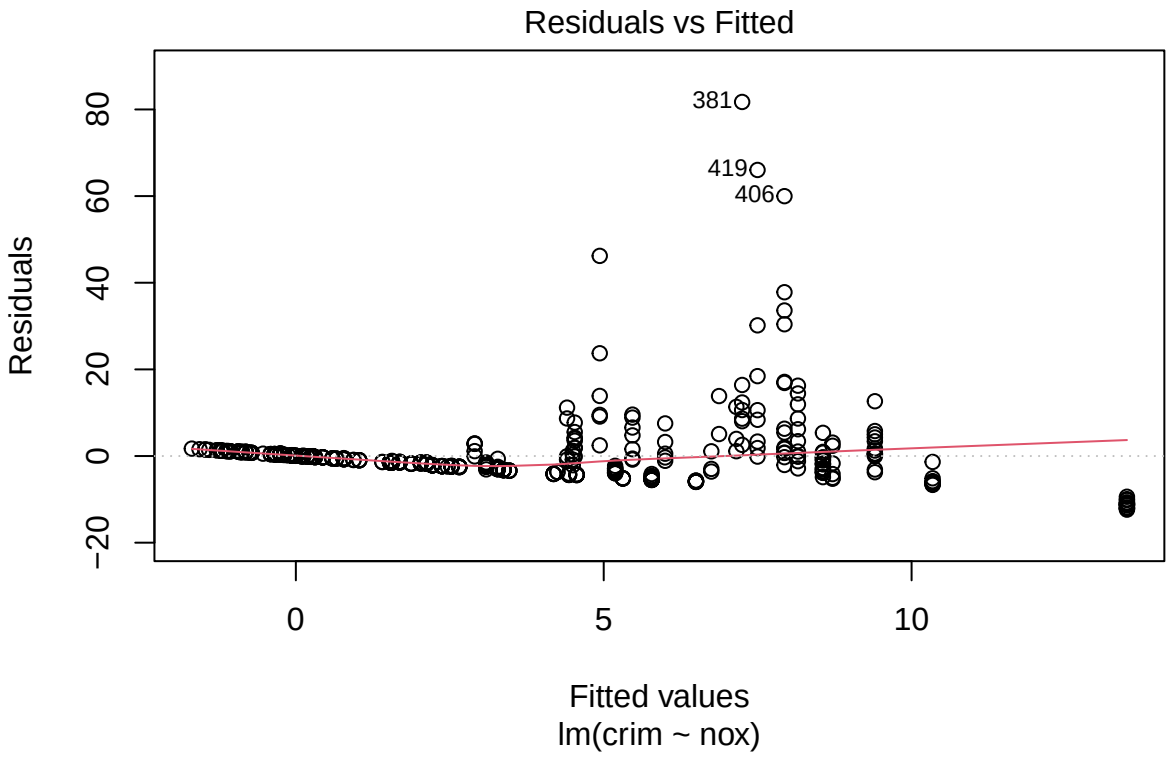

Residuals vs Leverage
Standardized residuals
Leverage
lm(crim ~ indus)
Cook's distance
1
0.5
381
419
406
−2
0
2
4
6
8
10
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014

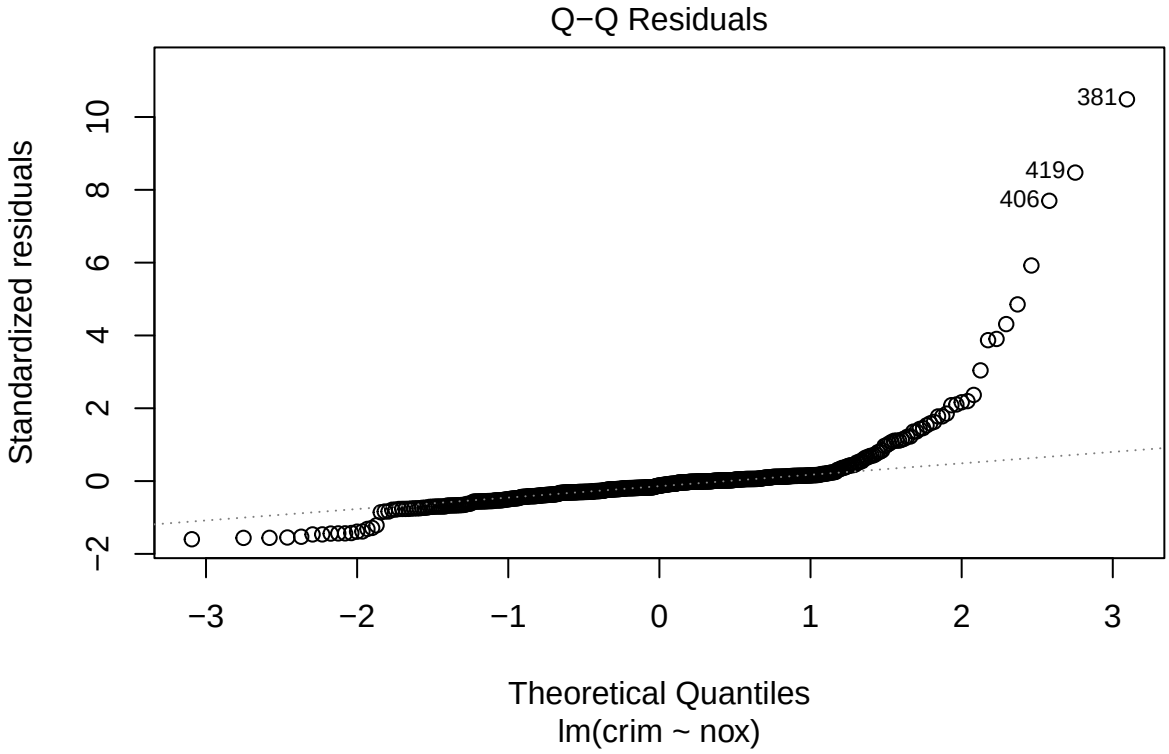

Residuals vs Fitted
Residuals
0
20
40
60
80
381
419
406
2.0
2.5
3.0
3.5
Fitted values
lm(crim ~ chas)

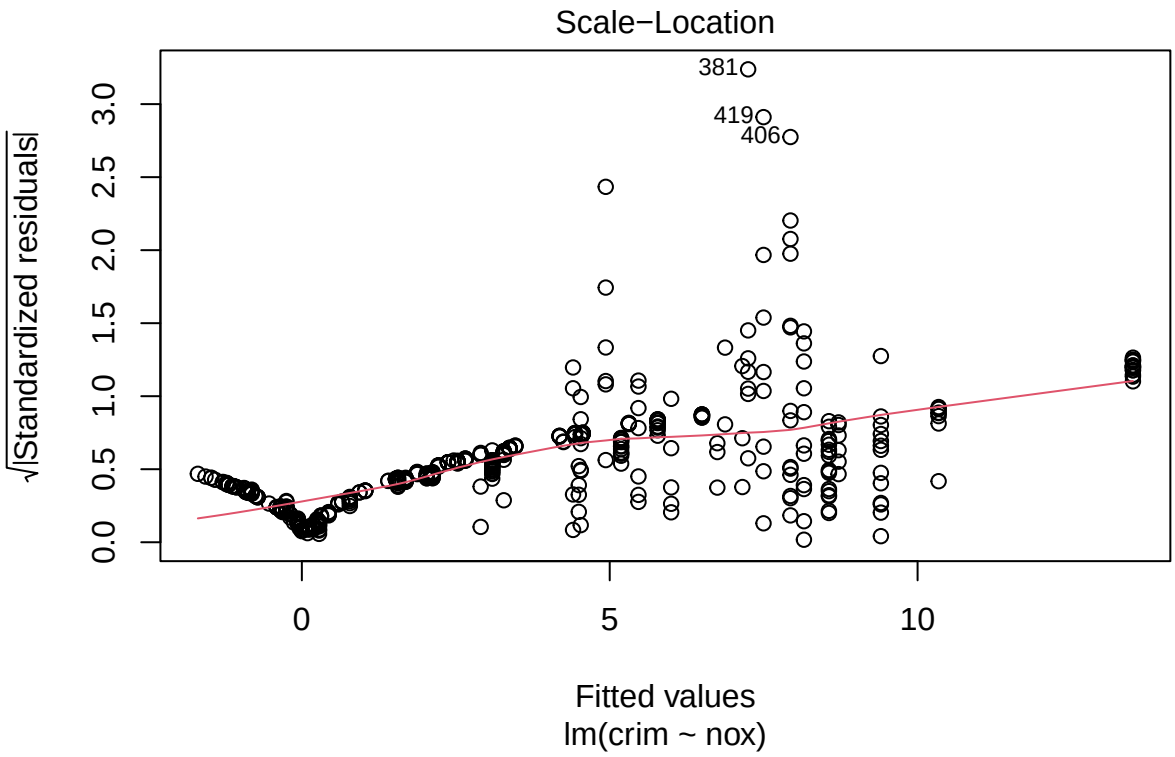

Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ chas)

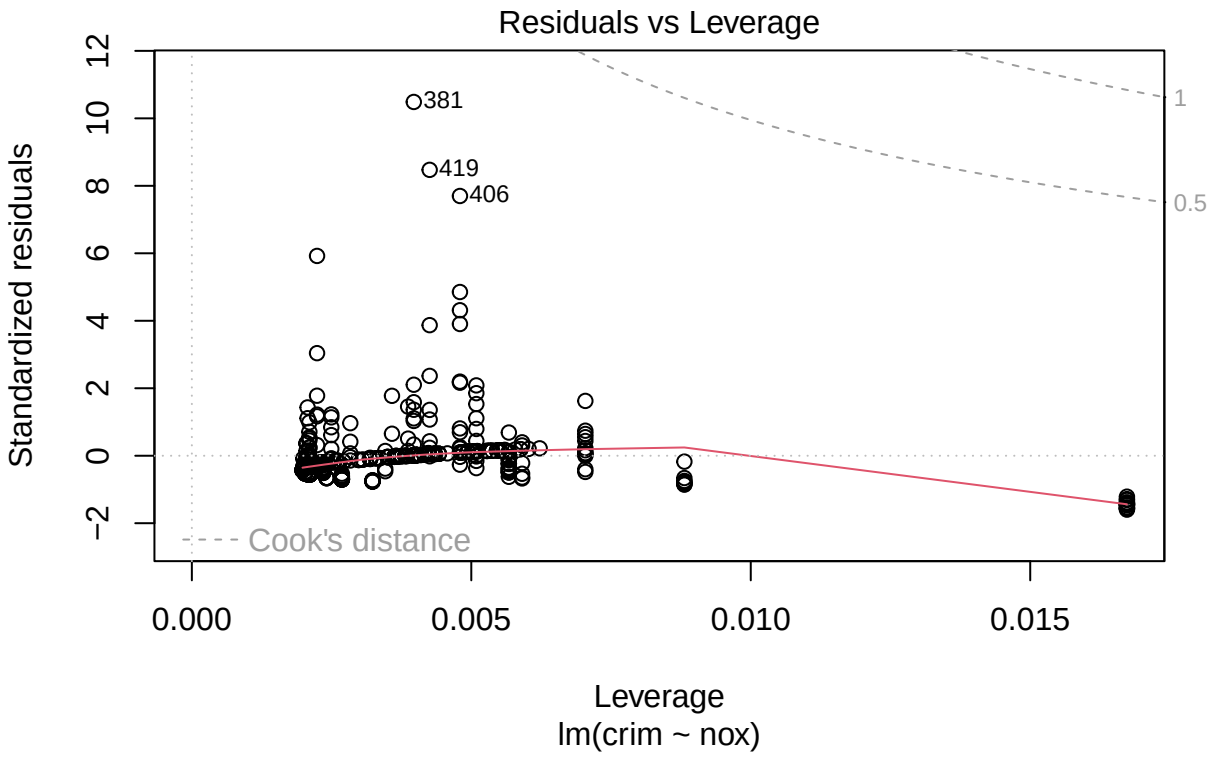

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
381
419
406
2.0
2.5
3.0
3.5
Fitted values
lm(crim ~ chas)

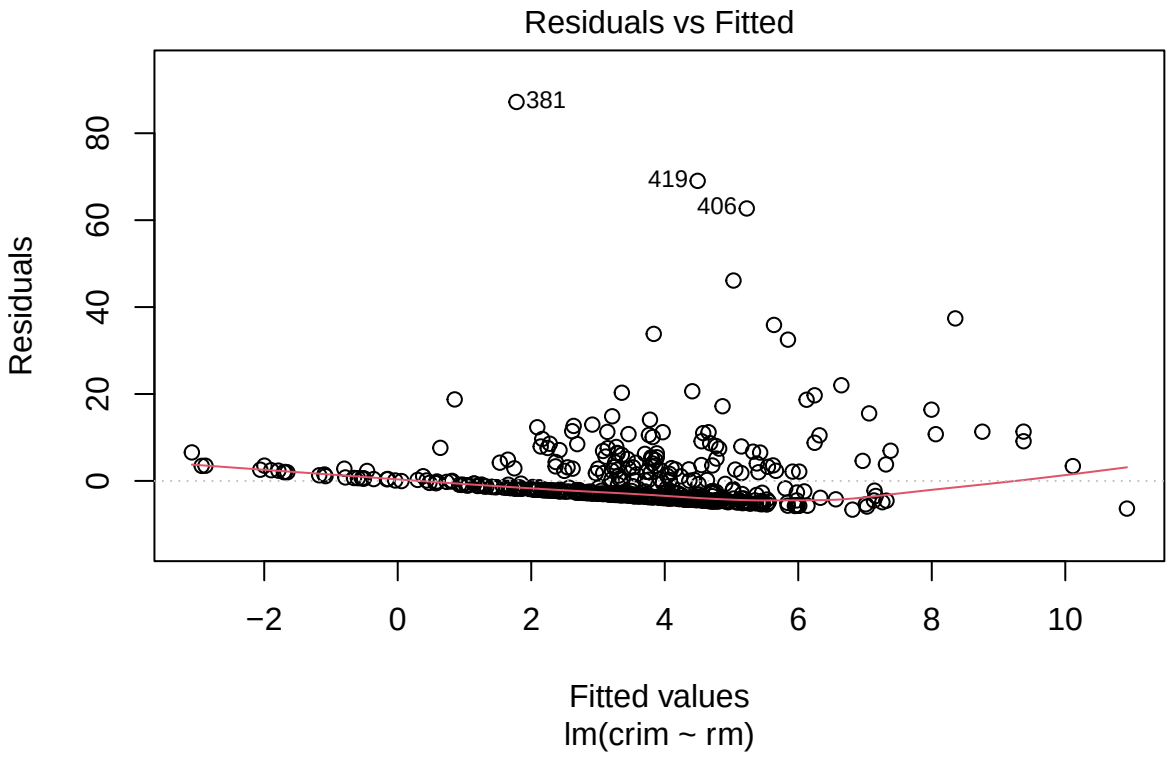

Residuals vs Leverage
Standardized residuals
0
2
4
6
8
10
381
419
406
1
0.5
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
lm(crim ~ chas)

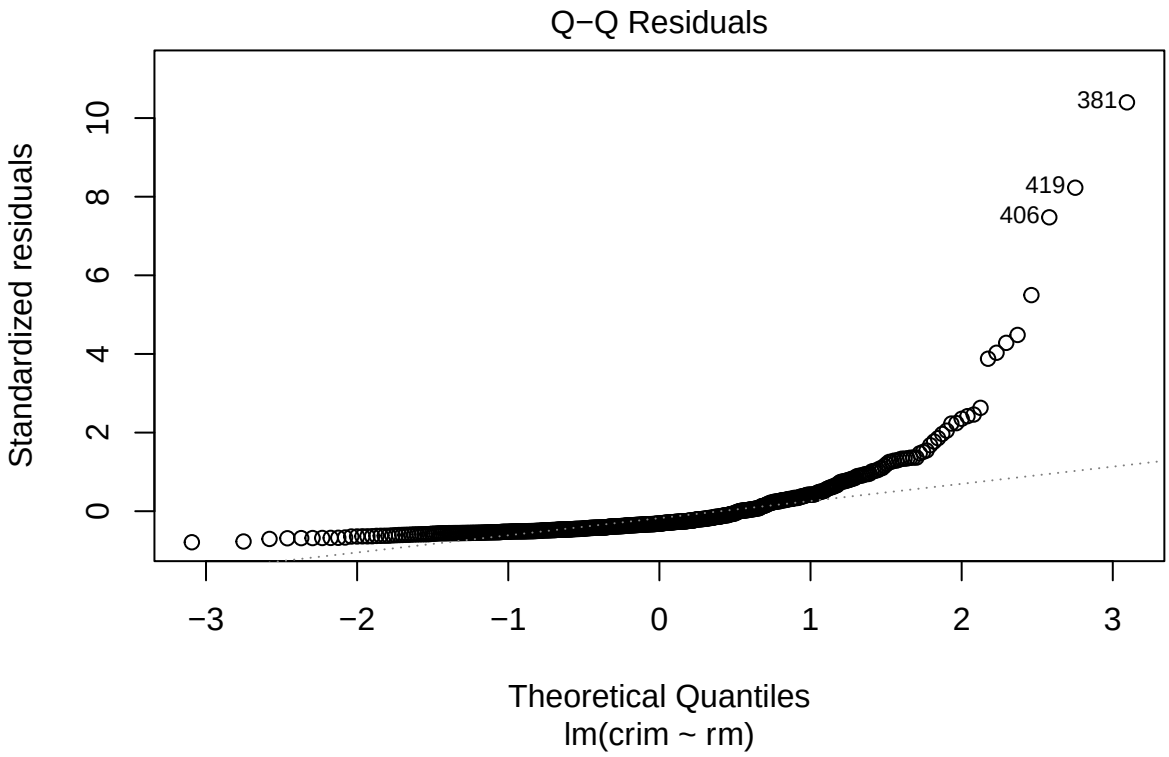

Residuals vs Fitted
381
419
406
Residuals
−20
0
20
40
60
80
0
5
10
Fitted values
lm(crim ~ nox)

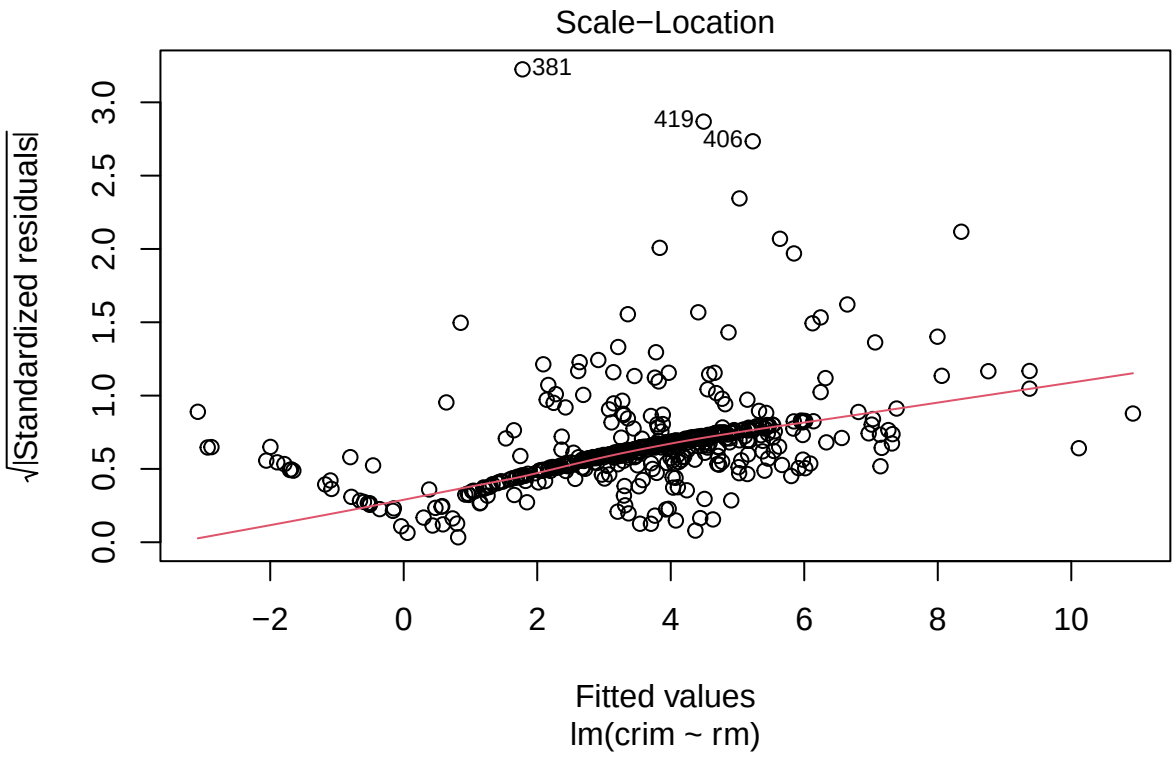

Q−Q Residuals
Standardized residuals
−2
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ nox)

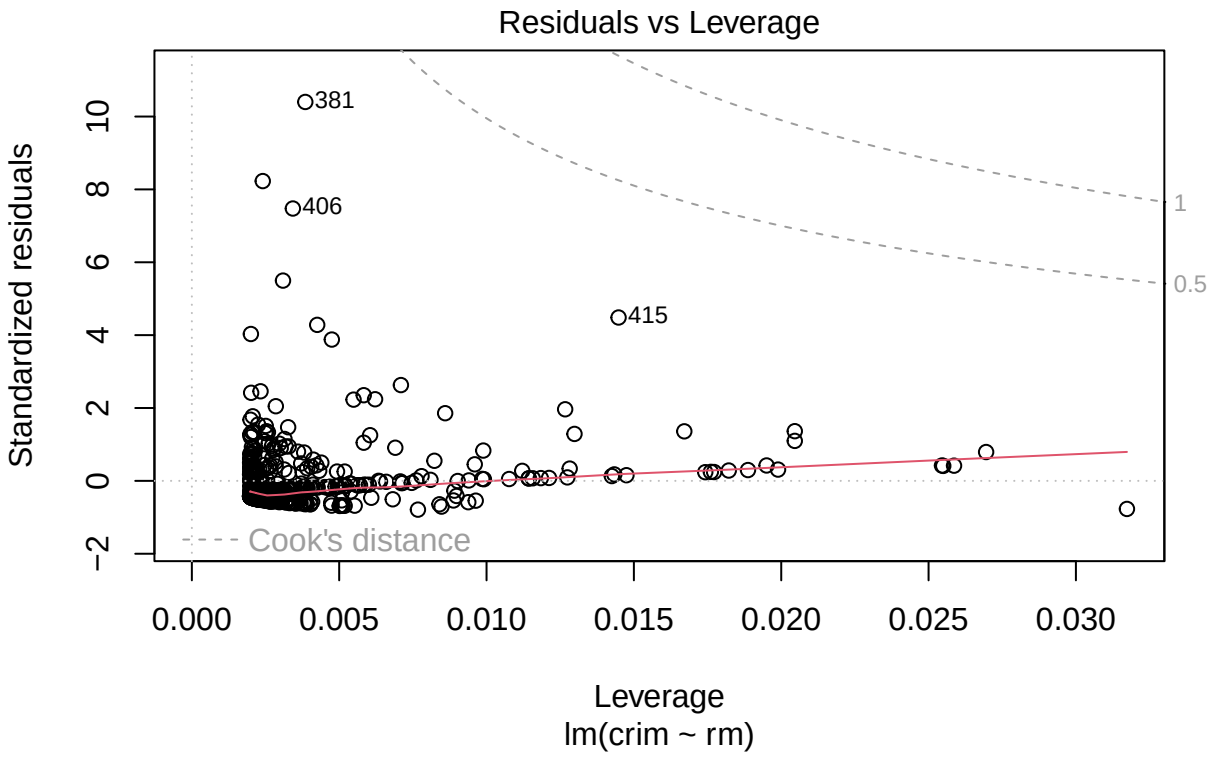

Scale−Location
381
419
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
5
10
Fitted values
lm(crim ~ nox)

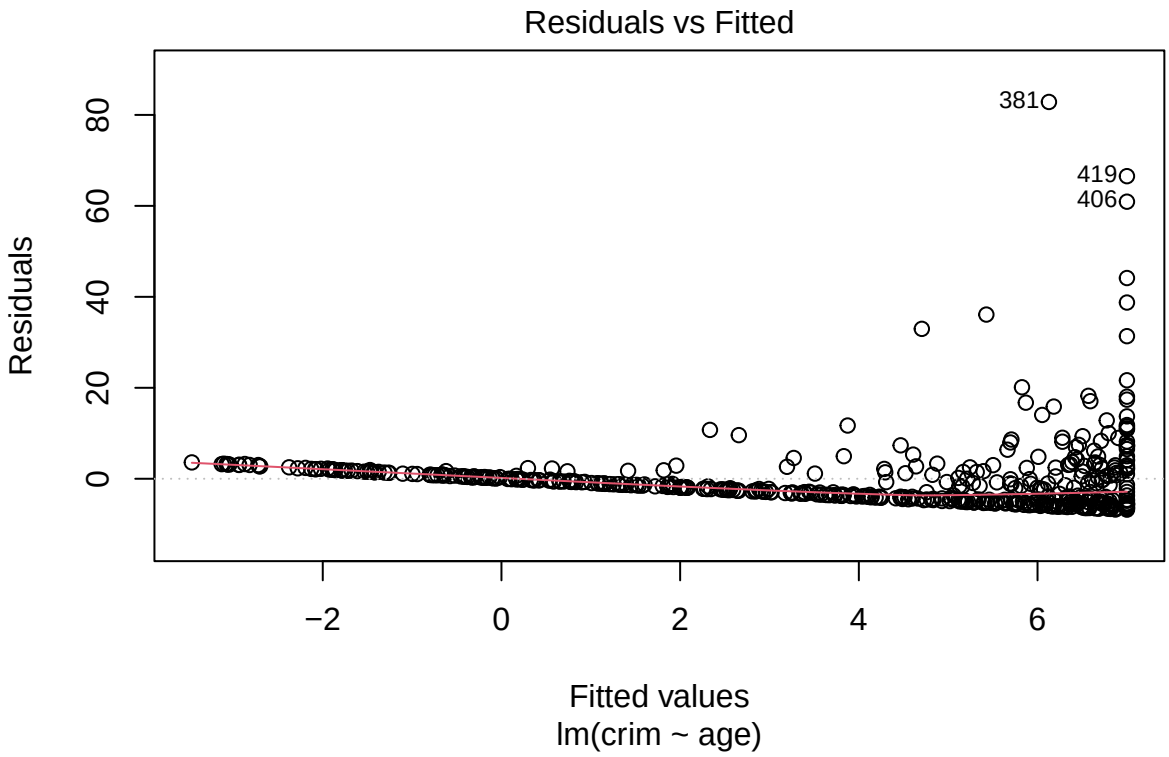

Residuals vs Leverage
Standardized residuals
12
10
8
6
4
2
0
−2
381
419
406
1
0.5
Cook's distance
0.000
0.005
0.010
0.015
Leverage
lm(crim ~ nox)

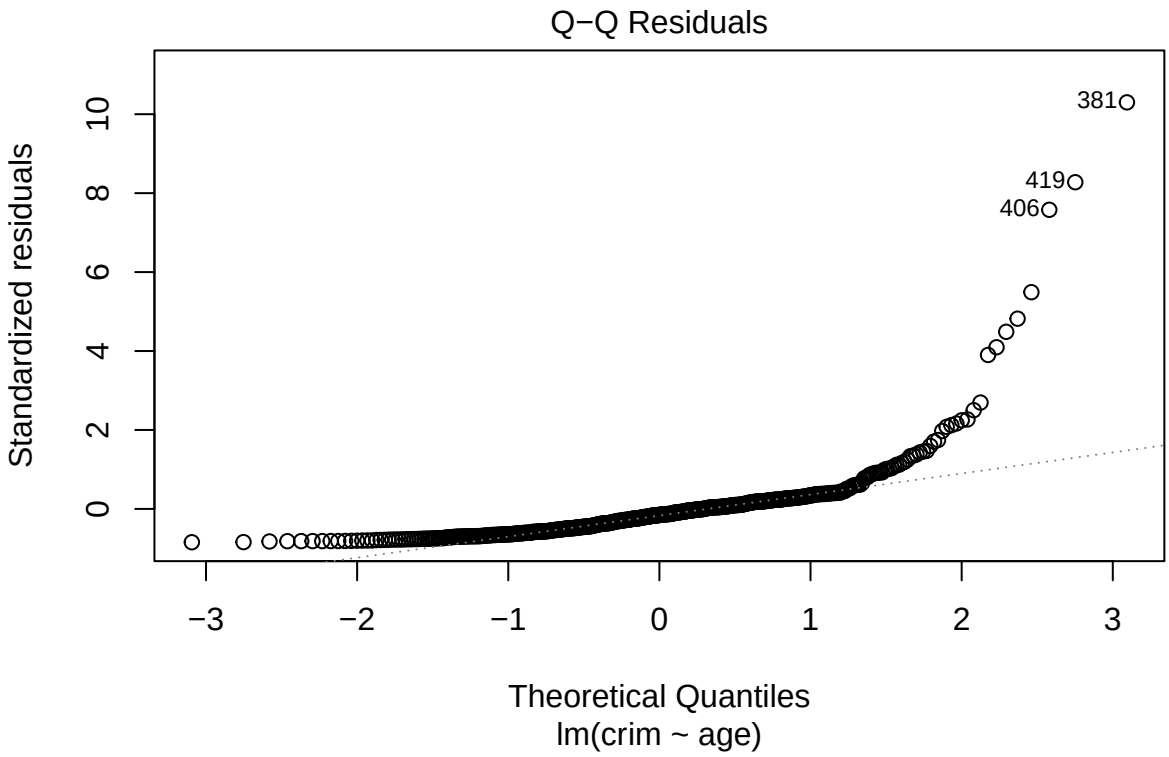

Residuals vs Fitted
381
419
406
Residuals
0
20
40
60
80
−2
0
2
4
6
8
10
Fitted values
lm(crim ~ rm)

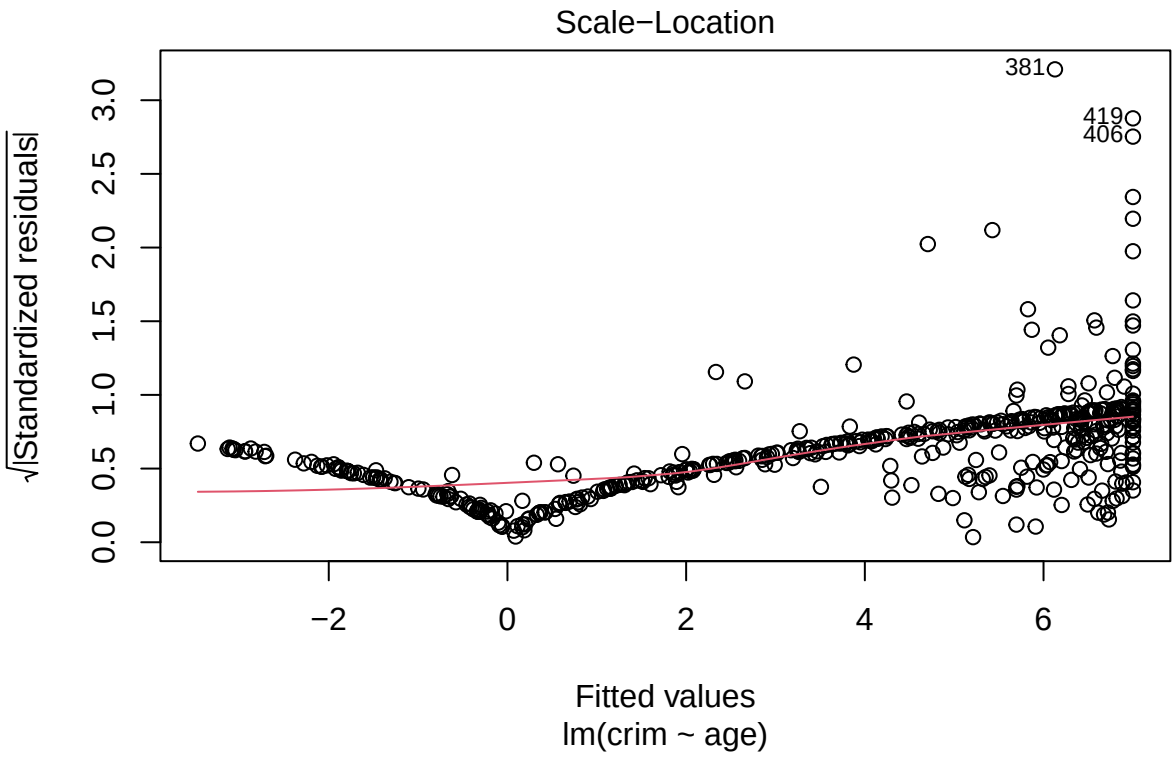

Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ rm)

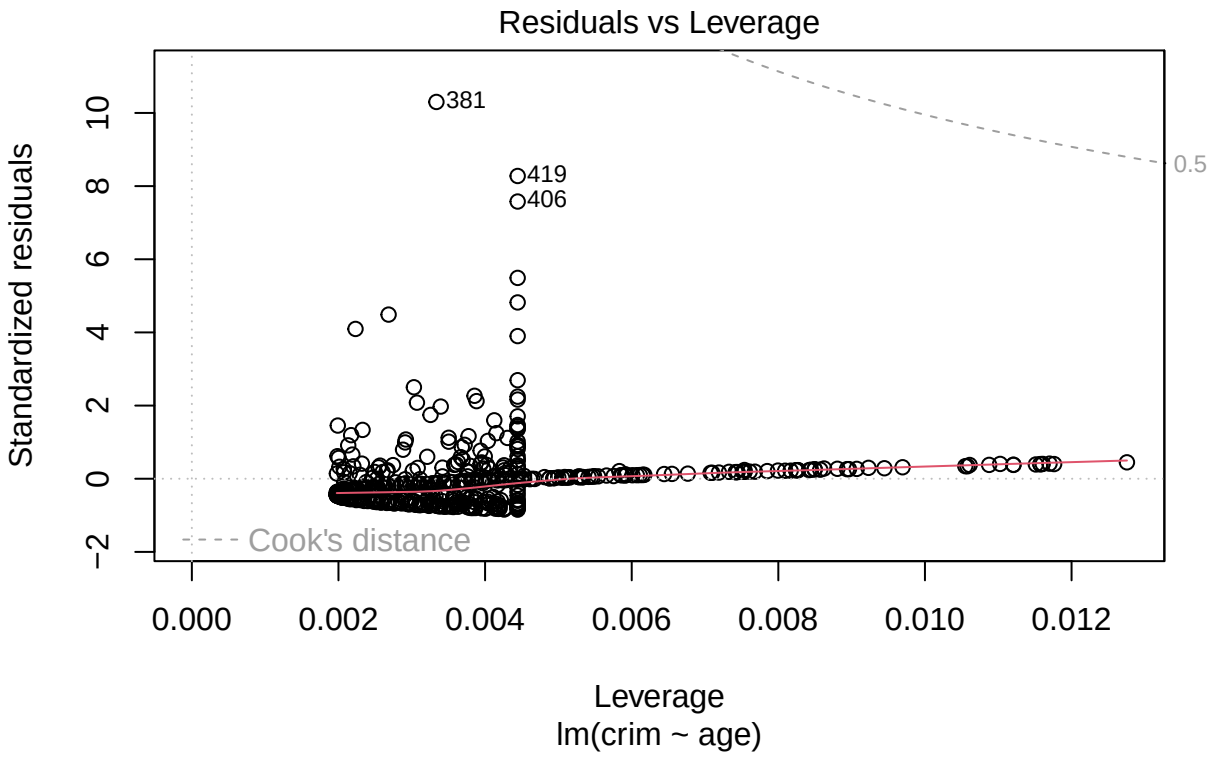

Scale−Location
381
419
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
0
2
4
6
8
10
Fitted values
lm(crim ~ rm)

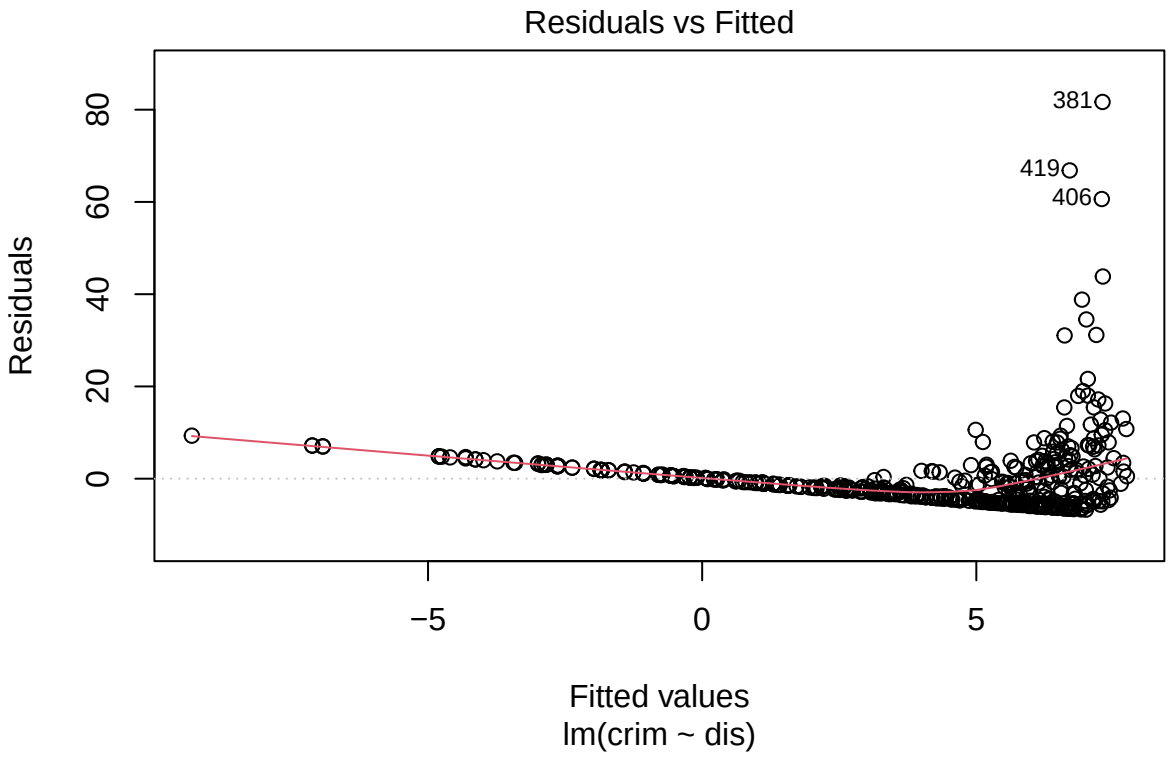

Residuals vs Leverage
Standardized residuals
10
8
6
4
2
0
−2
381
406
415
1
0.5
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(crim ~ rm)

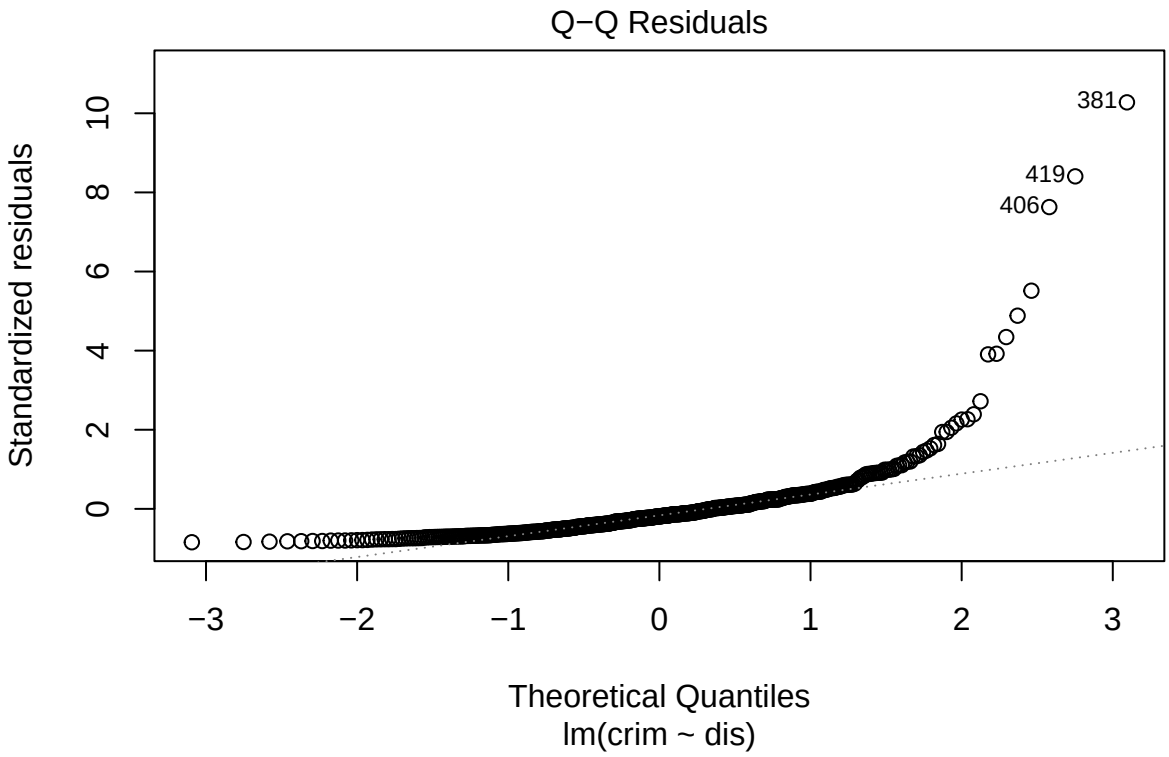

Residuals vs Fitted
Residuals
0
20
40
60
80
381
419
406
−2
0
2
4
6
Fitted values
lm(crim ~ age)

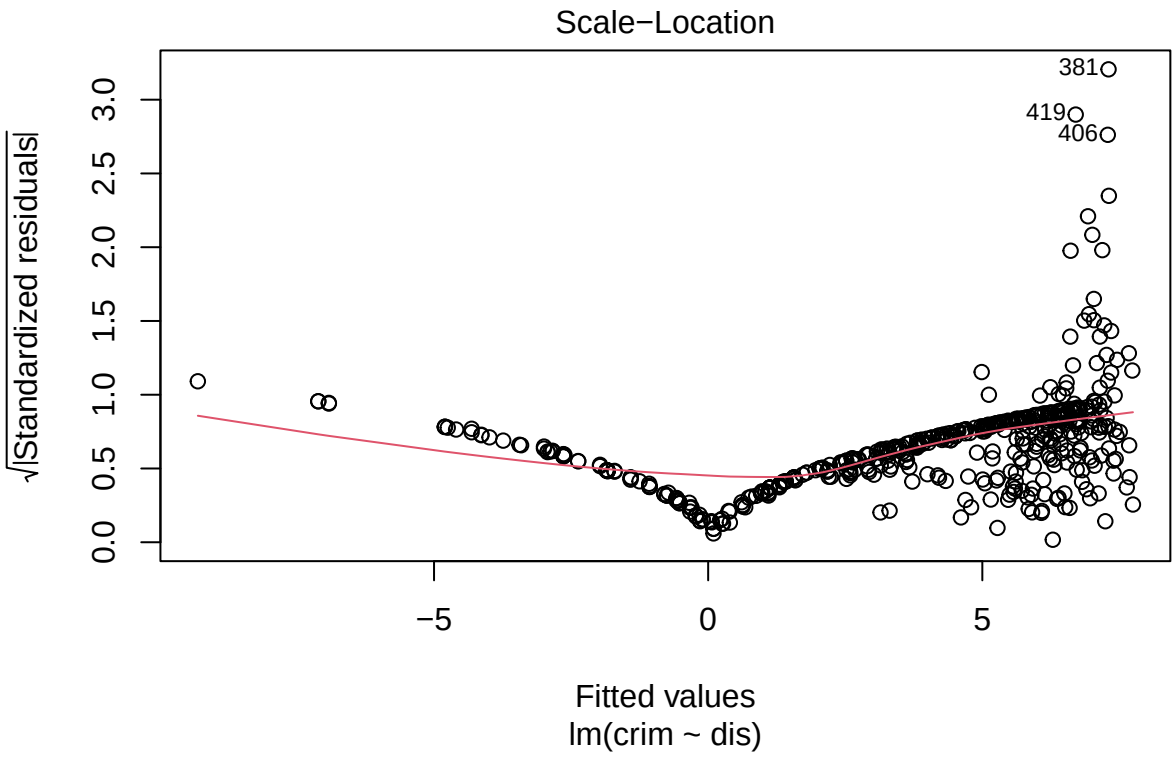

Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ age)

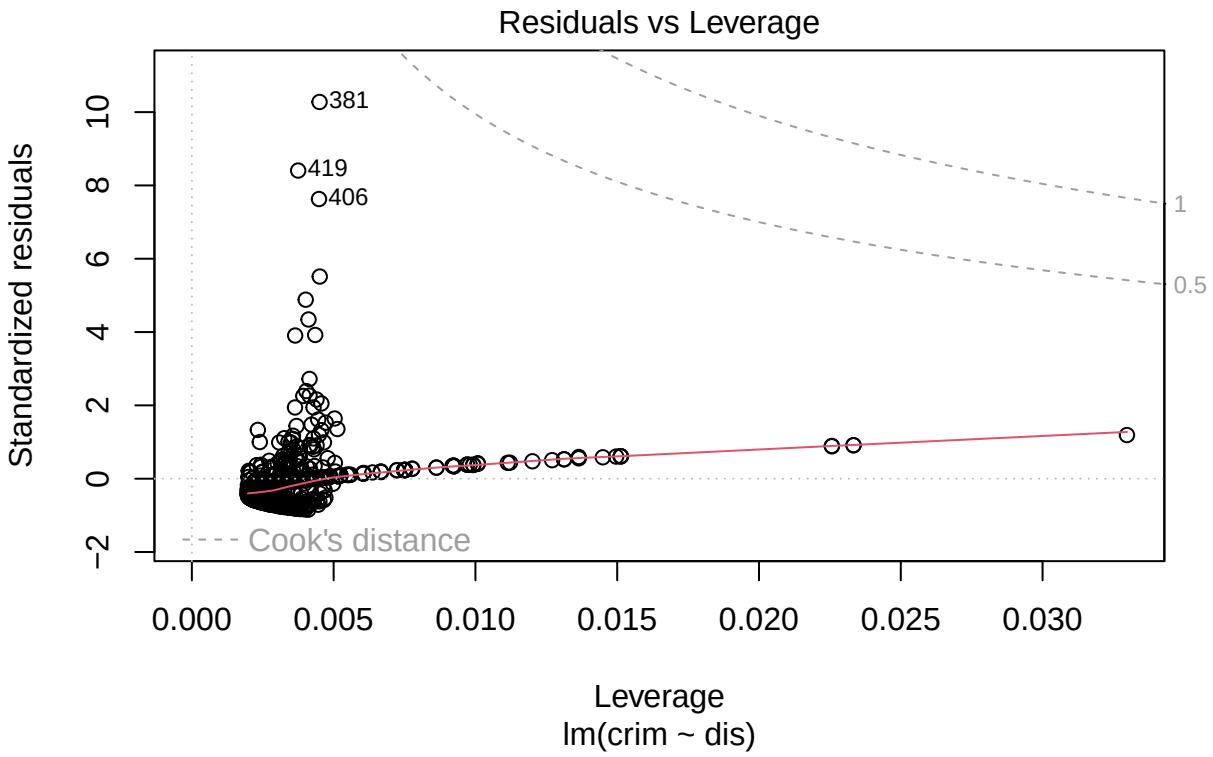

Scale−Location
381
419
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
0
2
4
6
Fitted values
lm(crim ~ age)

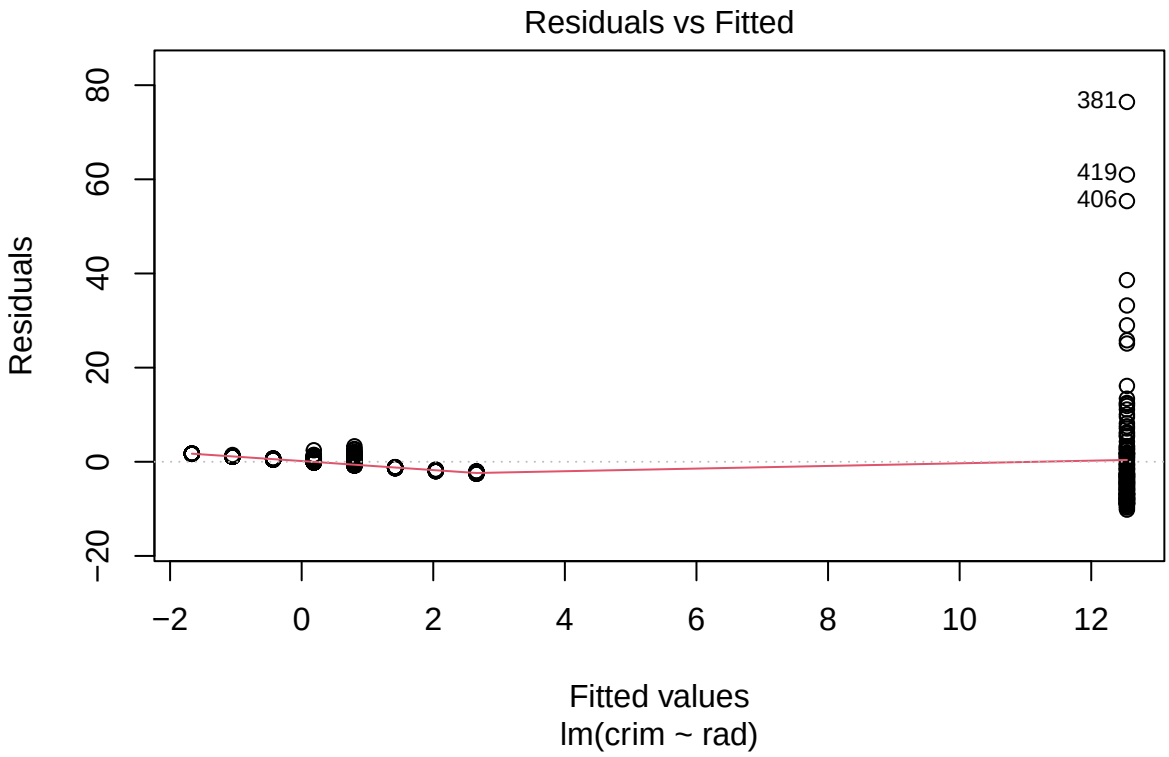

Residuals vs Leverage
Standardized residuals
Leverage
lm(crim ~ age)
Cook's distance
0.5
381
419
406
−2
0
2
4
6
8
10
0.000
0.002
0.004
0.006
0.008
0.010
0.012

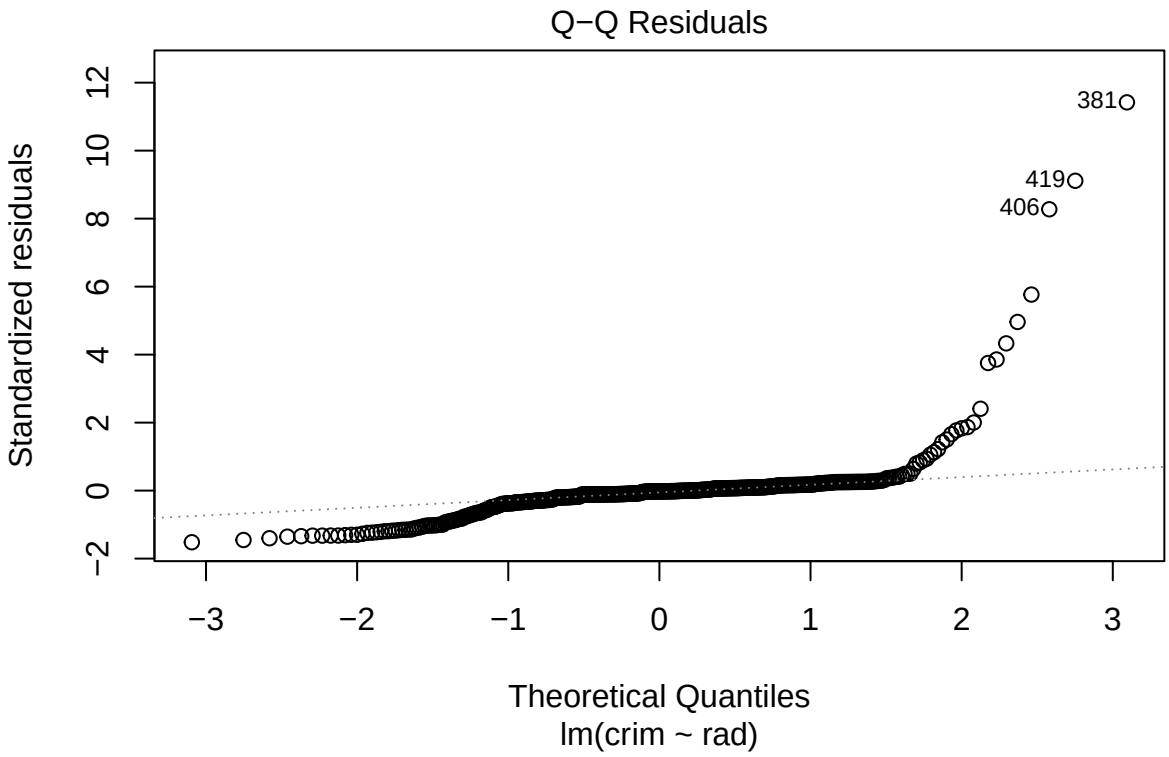

Residuals vs Fitted
Residuals
0
20
40
60
80
381
419
406
−5
0
5
Fitted values
lm(crim ~ dis)

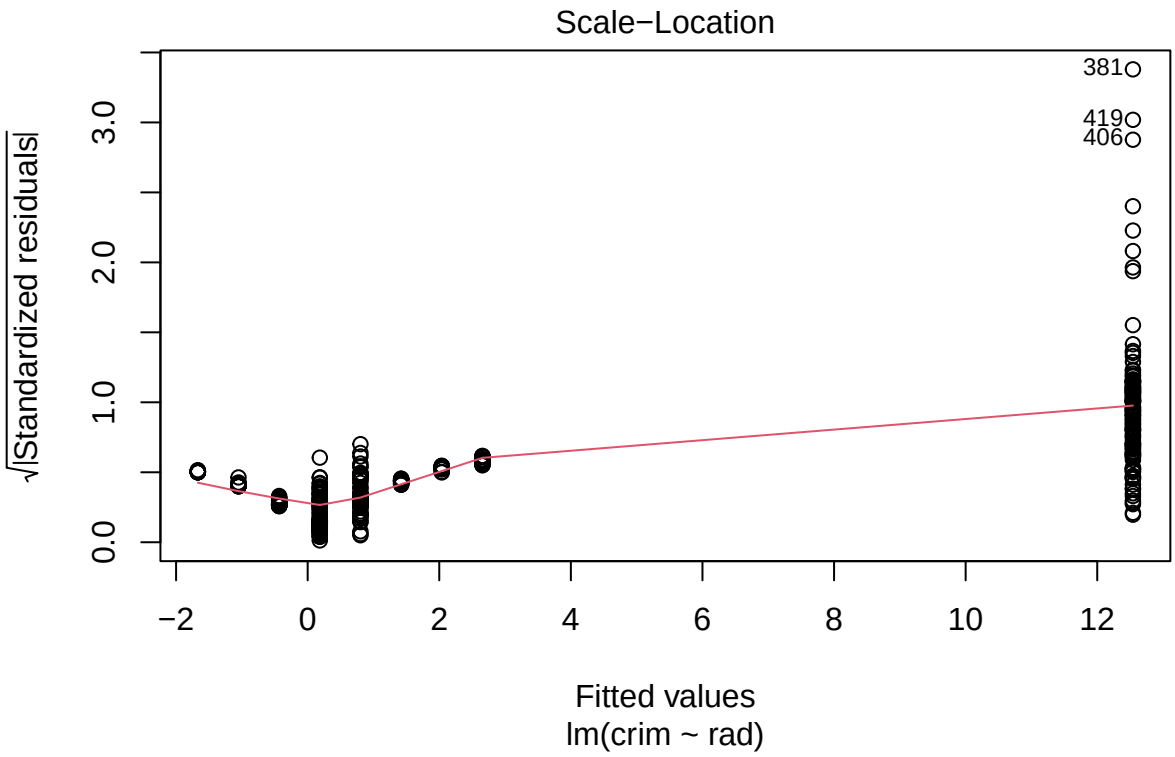

Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ dis)

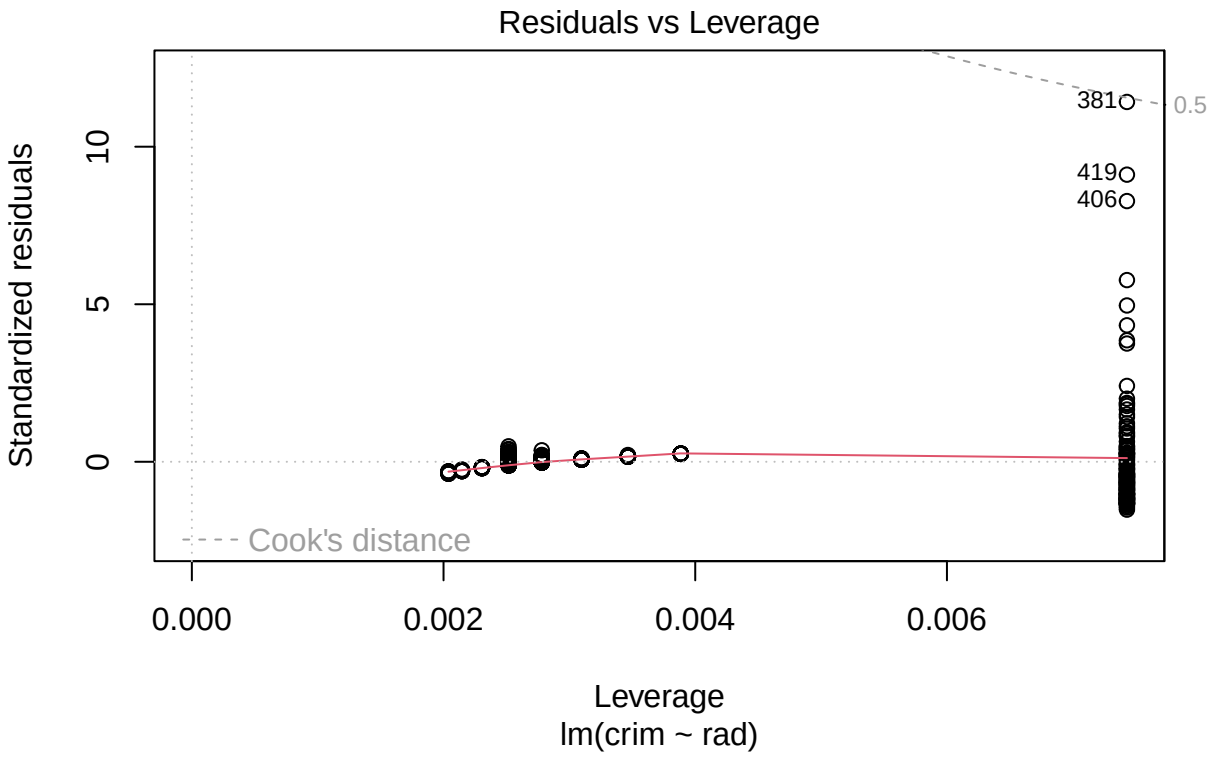

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
381
419
406
−5
0
5
Fitted values
lm(crim ~ dis)

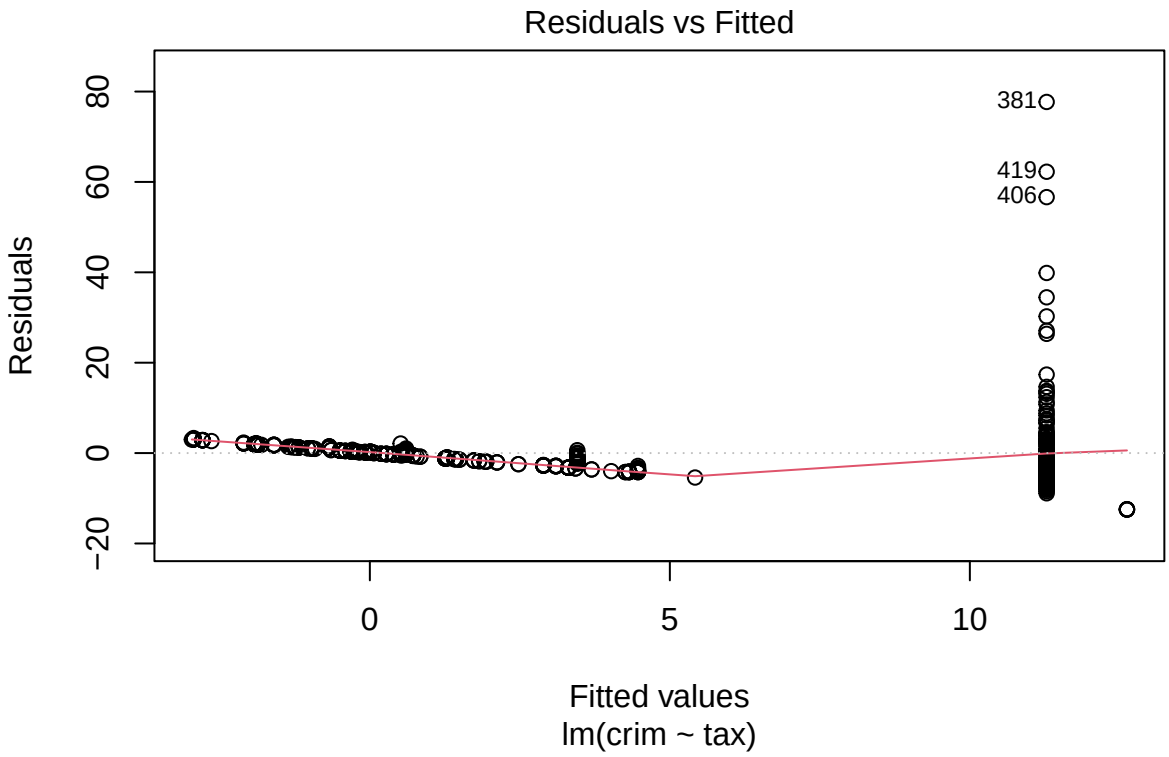

Residuals vs Leverage
Standardized residuals
10
8
6
4
2
0
−2
381
419
406
1
0.5
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(crim ~ dis)

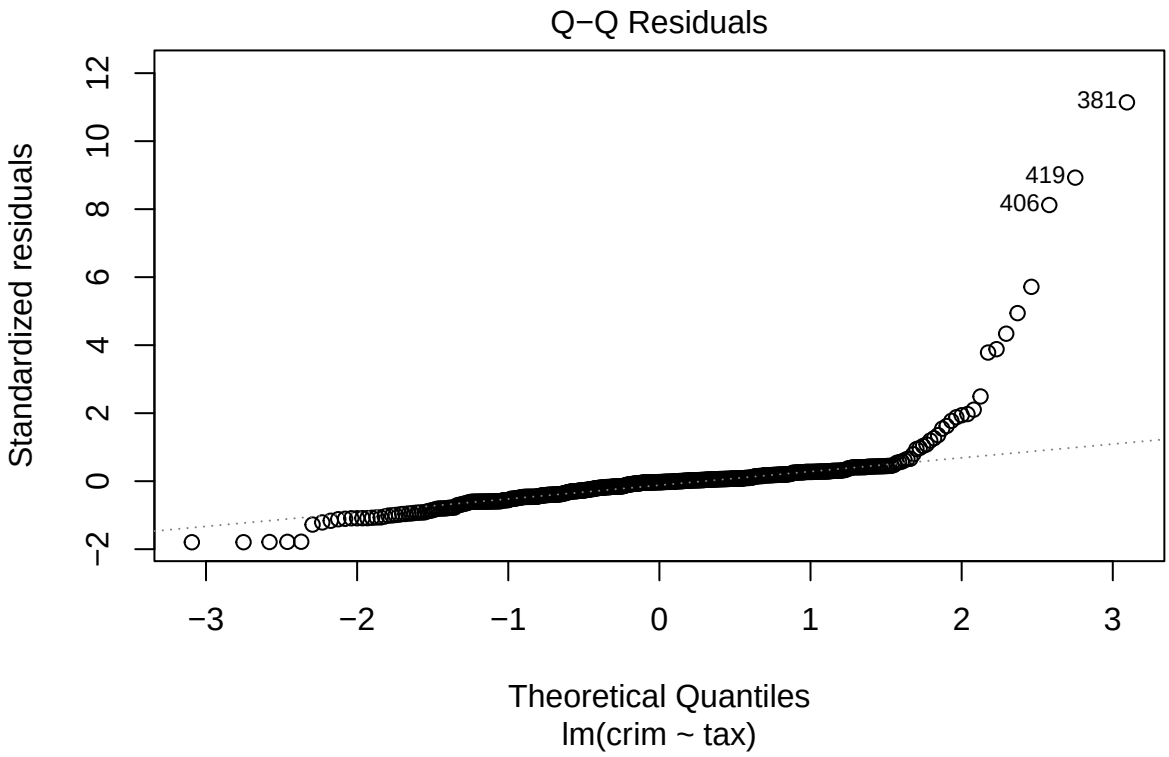

Residuals vs Fitted
Residuals
Fitted values
lm(crim ~ rad)
381
419
406
−20
0
20
40
60
80
−2
0
2
4
6
8
10
12

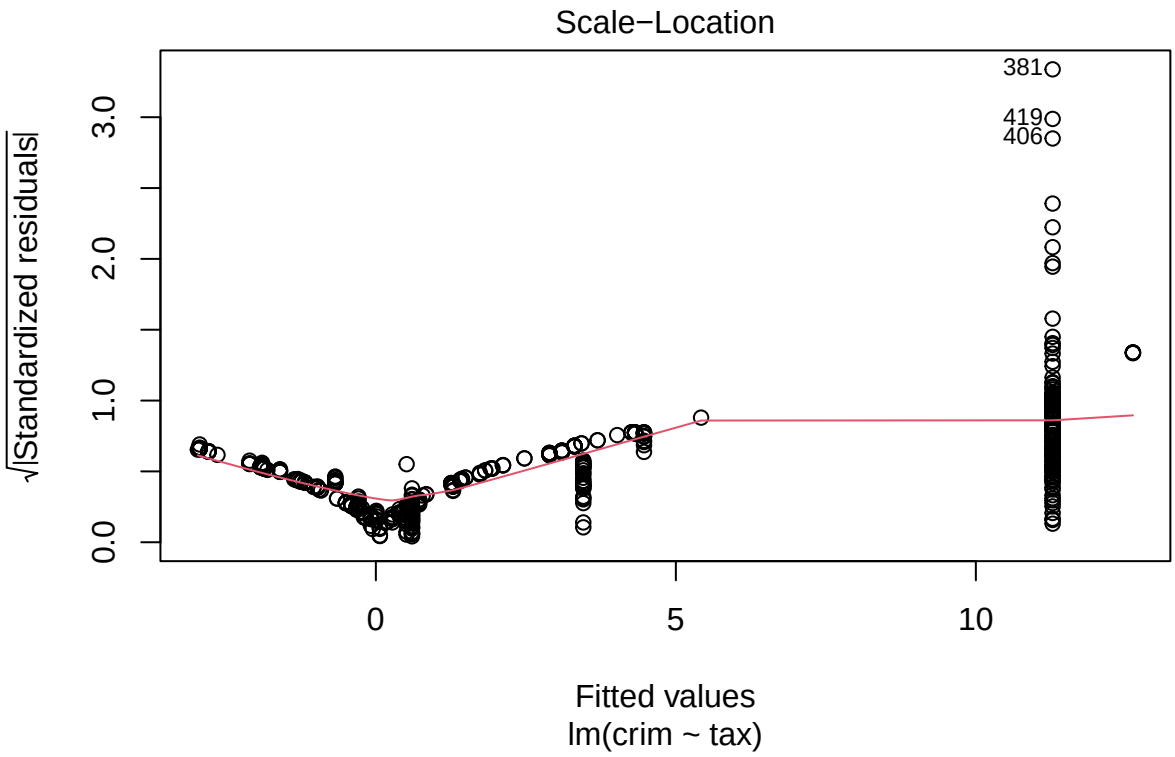

Q−Q Residuals
Standardized residuals
12
10
8
6
4
2
0
−2
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ rad)

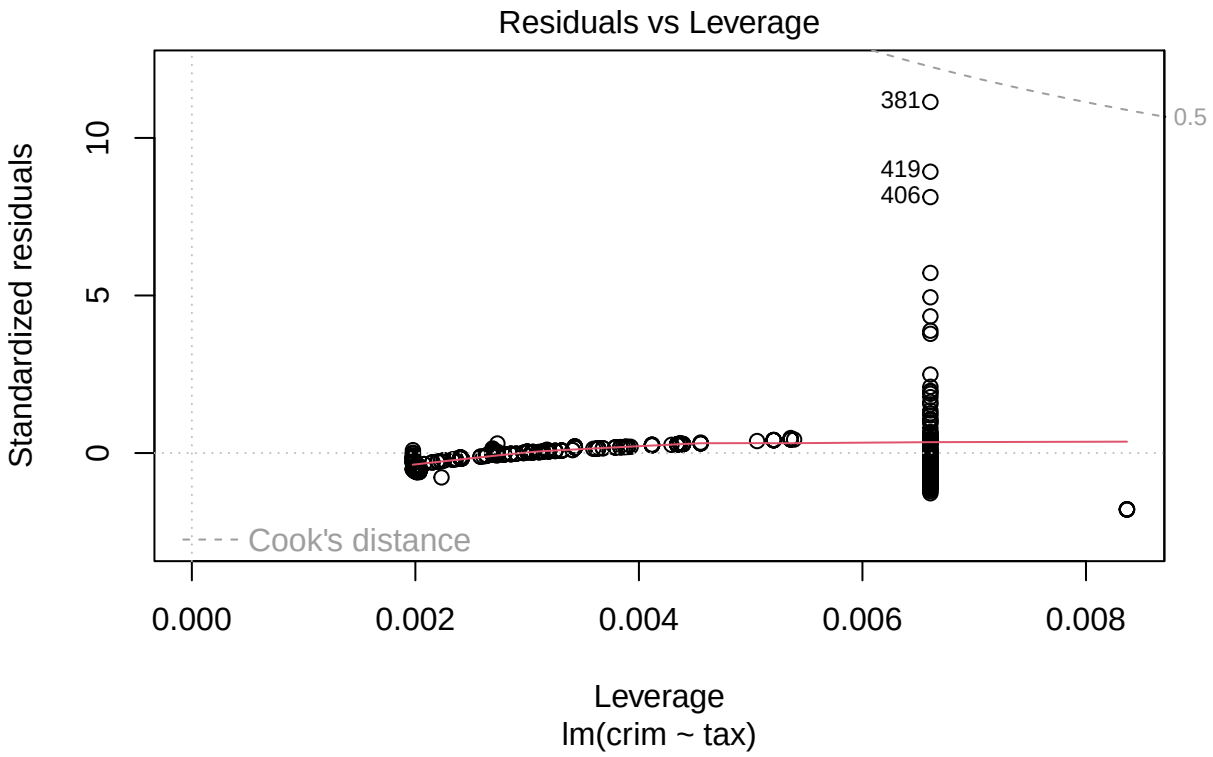

Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
381
419
406
−2
0
2
4
6
8
10
12
Fitted values
lm(crim ~ rad)

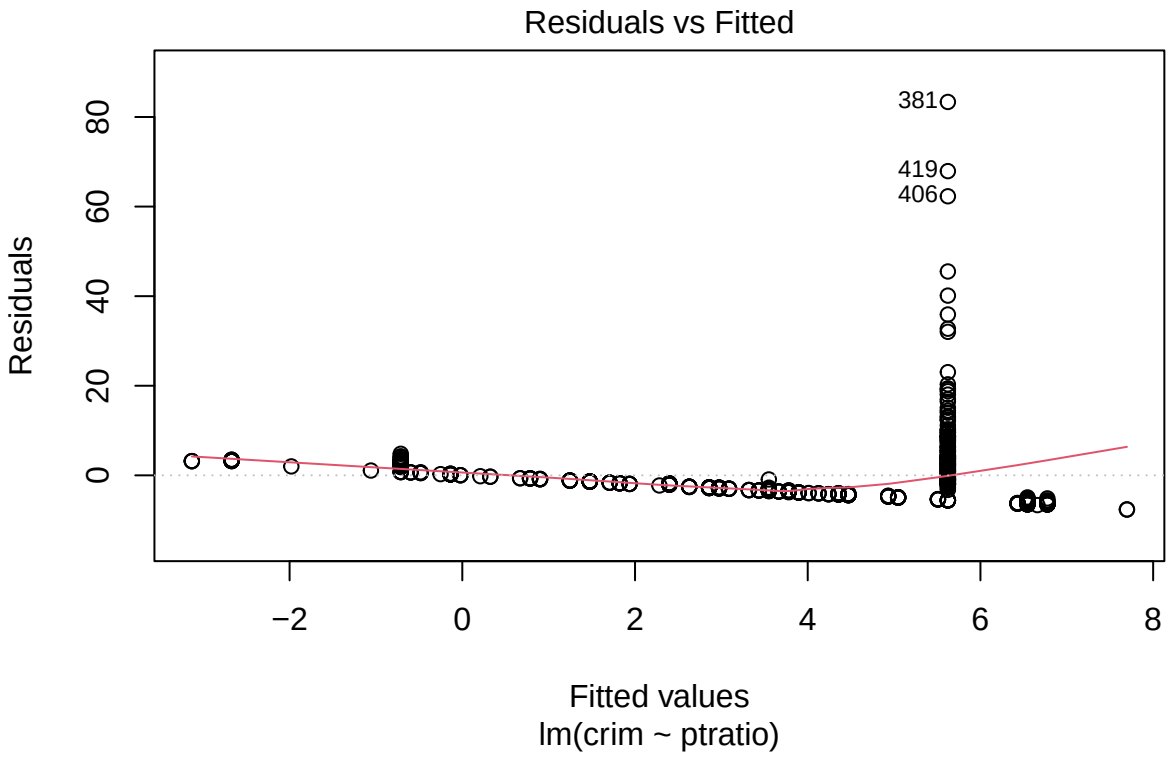

Residuals vs Leverage
Standardized residuals
10
5
0
381
419
406
0.5
Cook's distance
0.000
0.002
0.004
0.006
Leverage
lm(crim ~ rad)

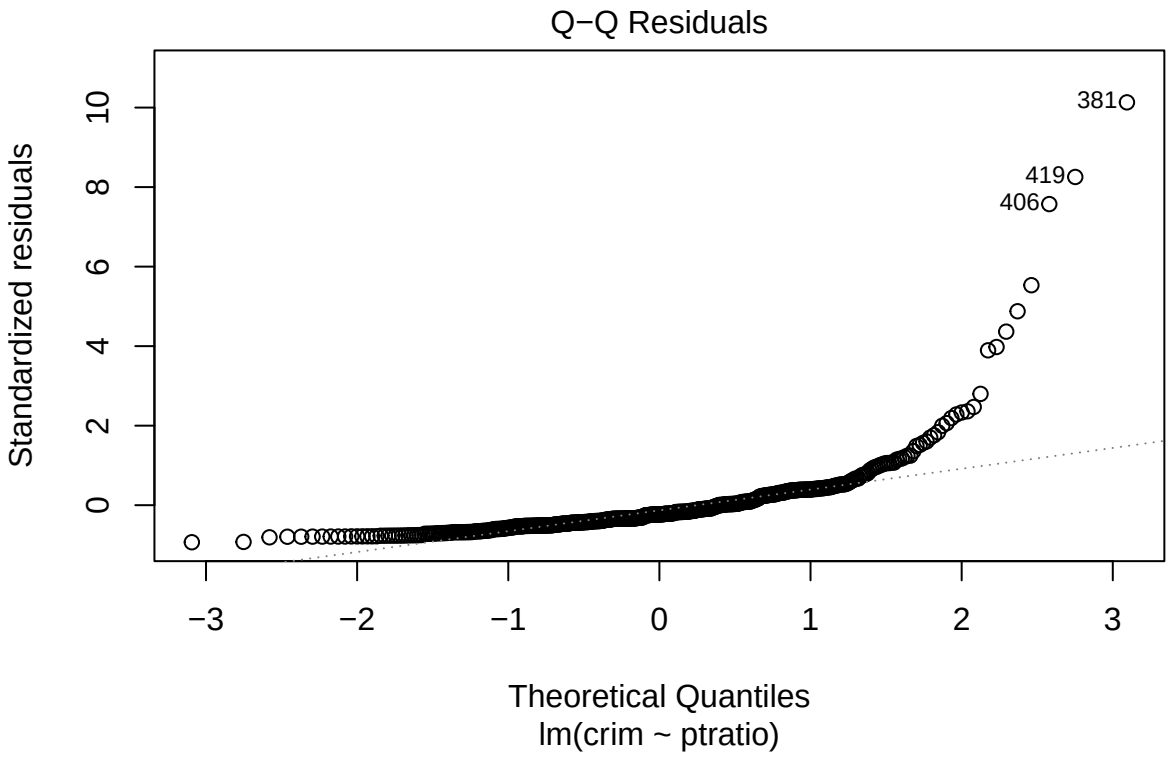

Residuals vs Fitted
Residuals
Fitted values
lm(crim ~ tax)
381
419
406
−20
0
20
40
60
80
0
5
10

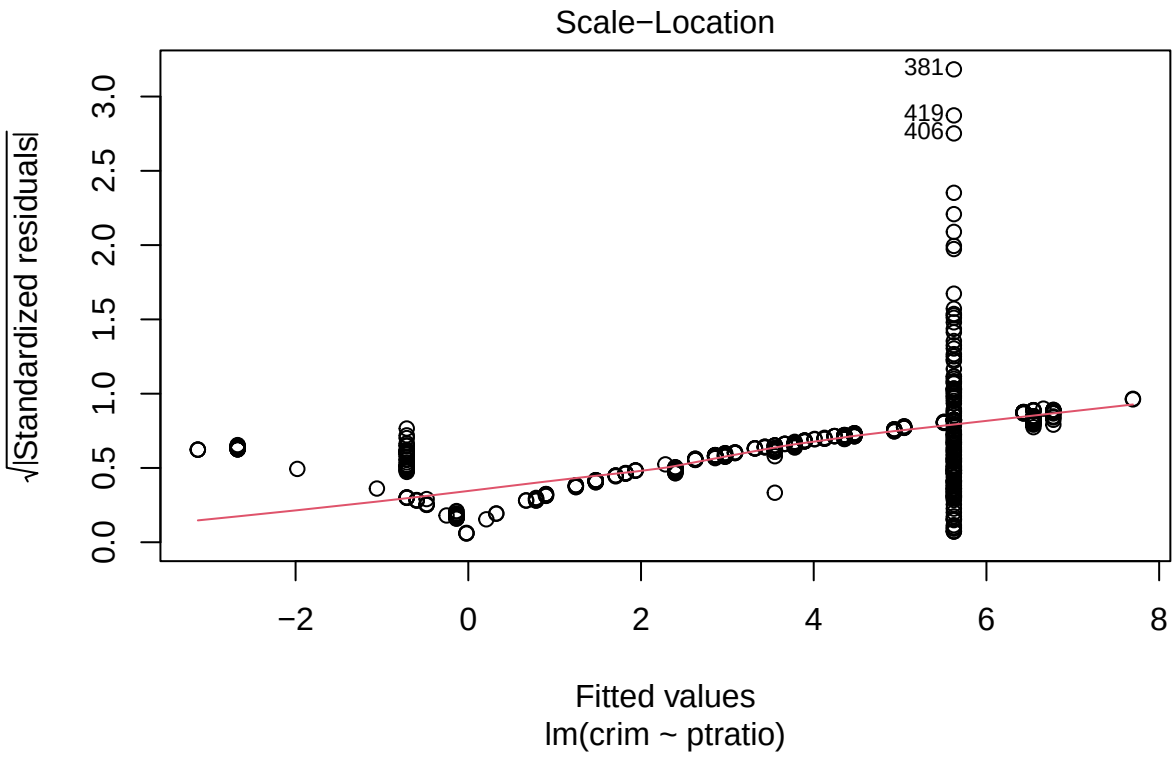

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(crim ~ tax)
381
419
406
−2
0
2
4
6
8
10
12
−3
−2
−1
0
1
2
3

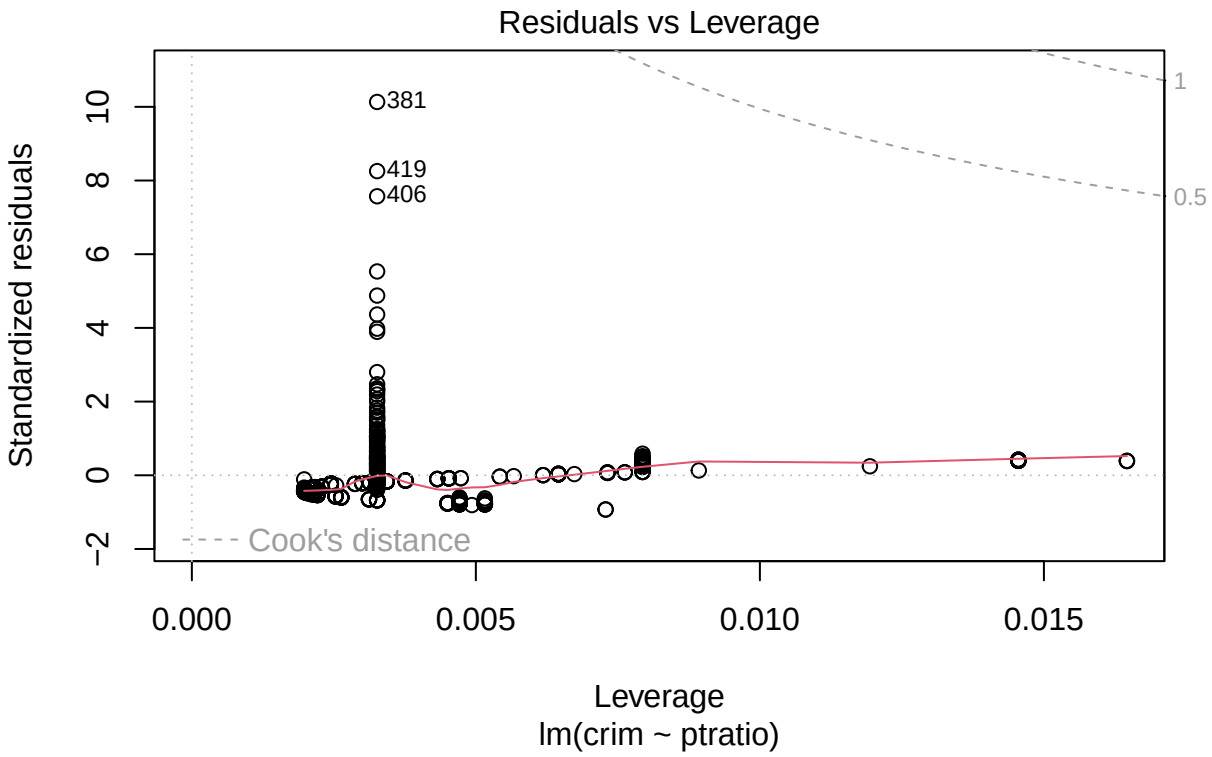

Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
381
419
406
0
5
10
Fitted values
lm(crim ~ tax)

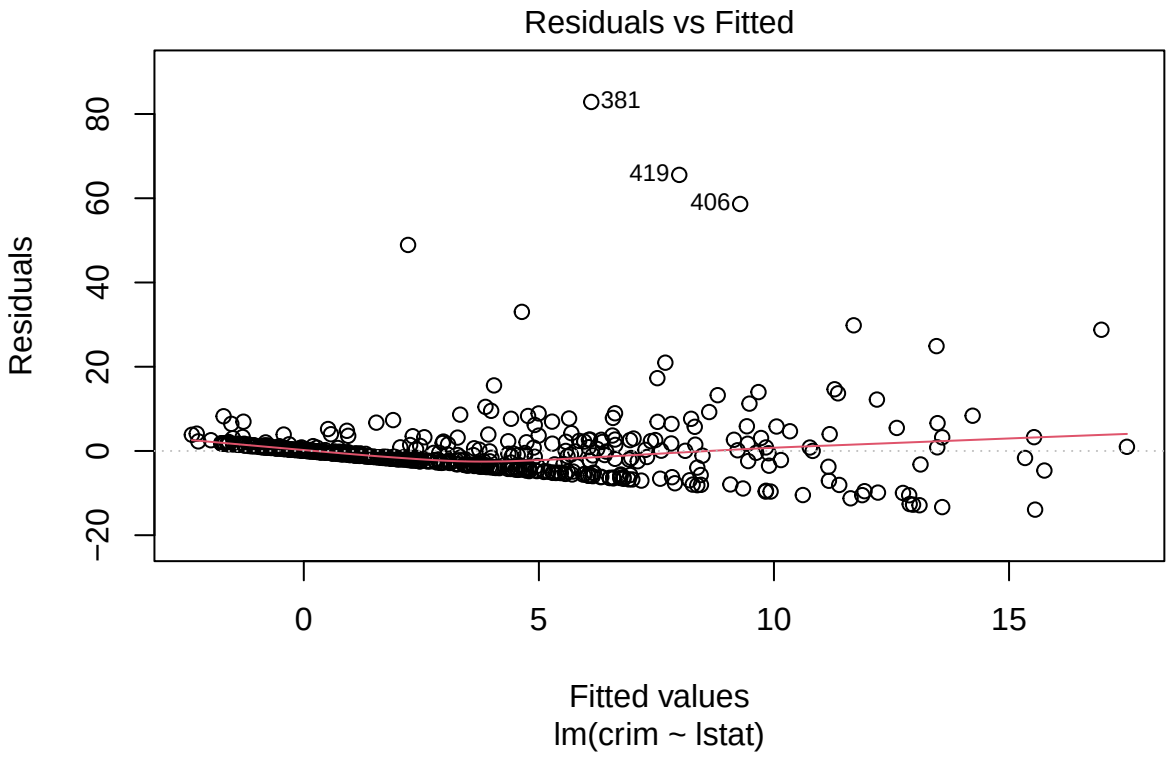

Residuals vs Leverage
Standardized residuals
10
5
0
381
419
406
0.5
Cook's distance
0.000
0.002
0.004
0.006
0.008
Leverage
lm(crim ~ tax)

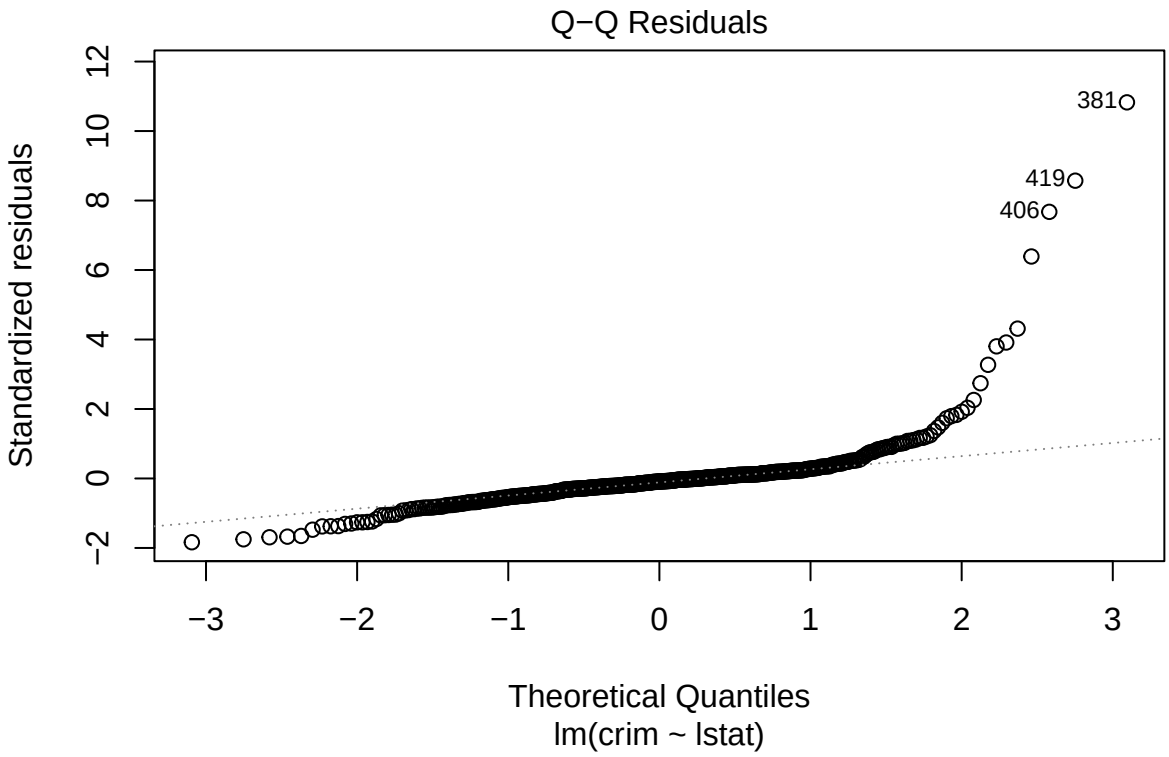

Residuals vs Fitted
Residuals
Fitted values
lm(crim ~ ptratio)
381
419
406
0
20
40
60
80
−2
0
2
4
6
8

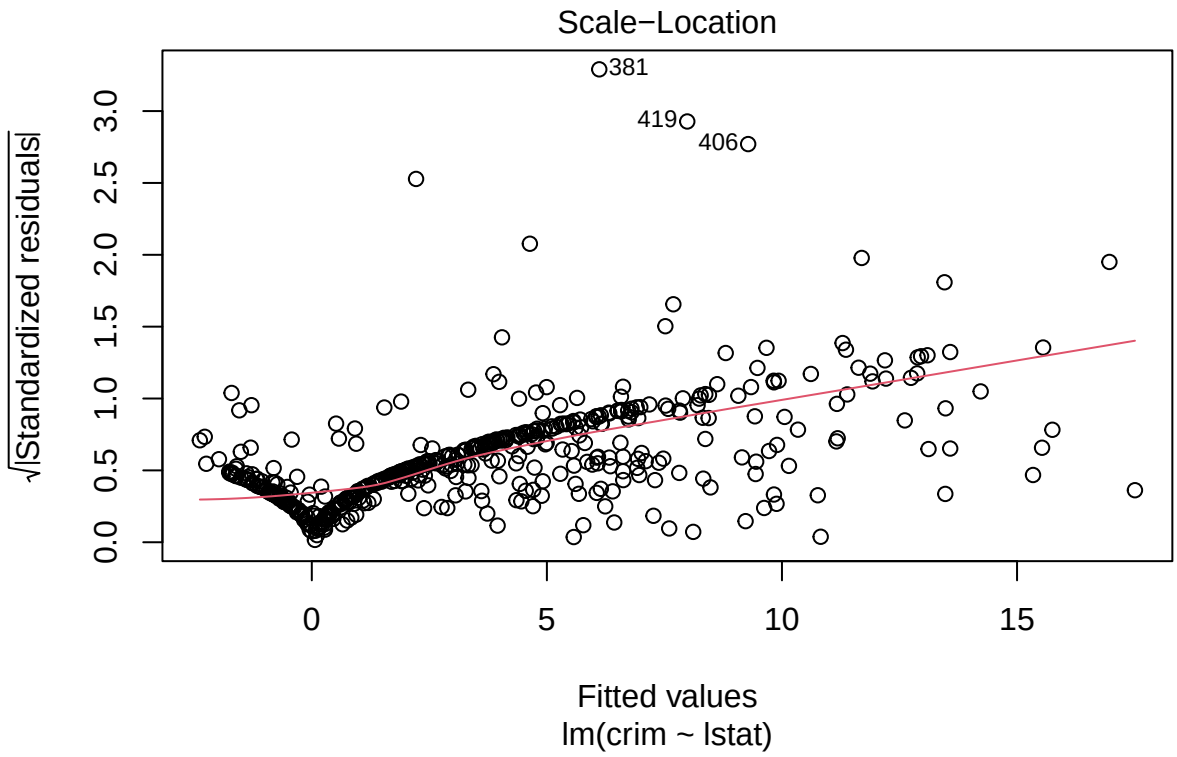

Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ ptratio)

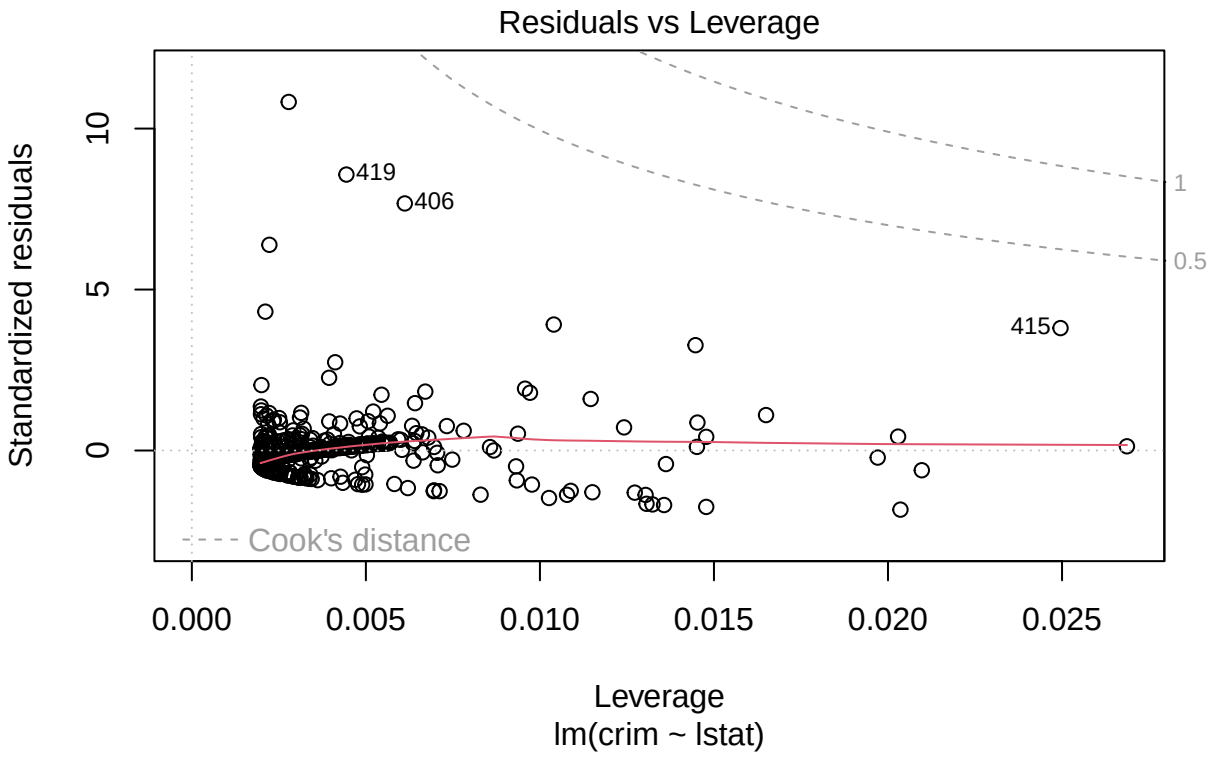

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
381
419
406
−2
0
2
4
6
8
Fitted values
lm(crim ~ ptratio)

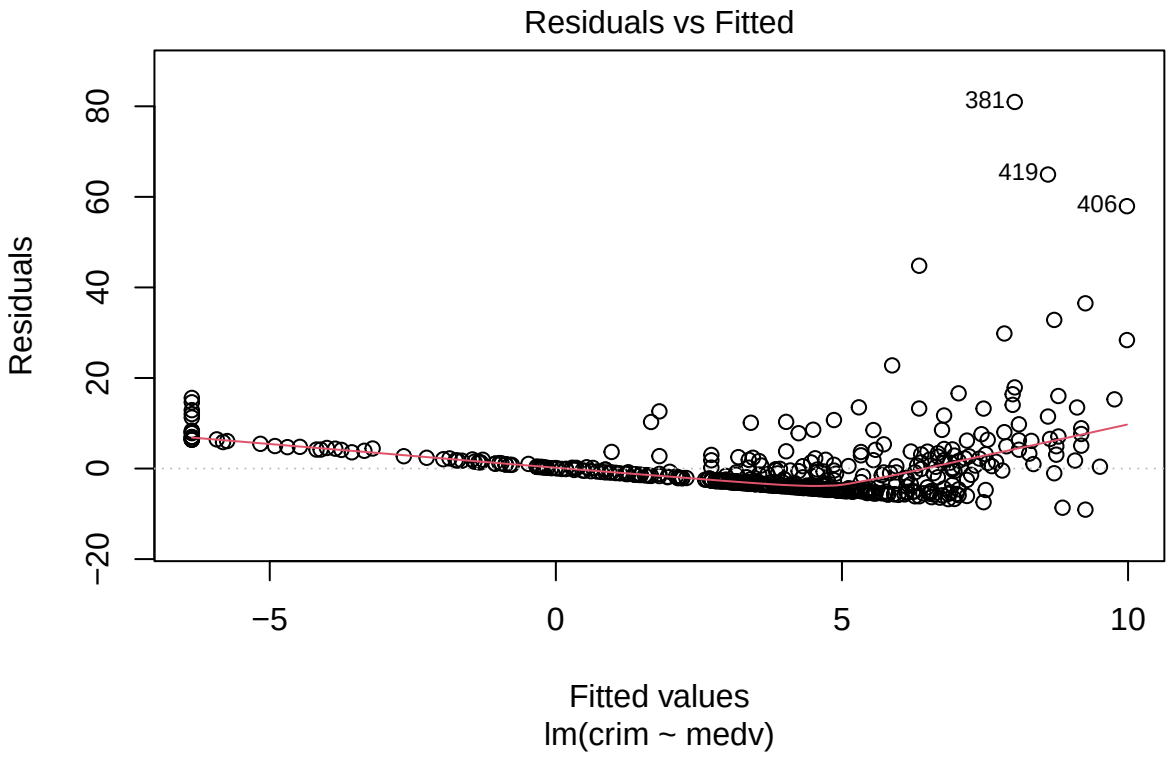

Residuals vs Leverage
Standardized residuals
Leverage
lm(crim ~ ptratio)
Cook's distance
1
0.5
381
419
406
−2
0
2
4
6
8
10
0.000
0.005
0.010
0.015

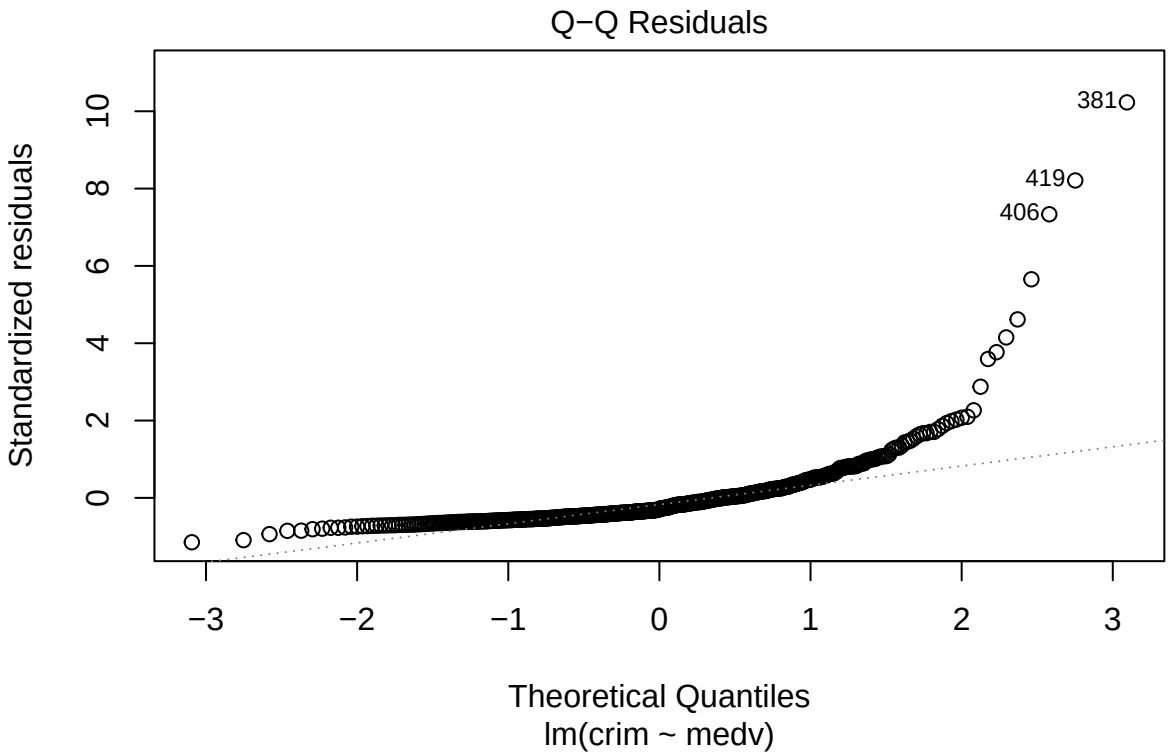

Residuals vs Fitted
381
419
406
Residuals
−20
0
20
40
60
80
0
5
10
15
Fitted values
lm(crim ~ lstat)

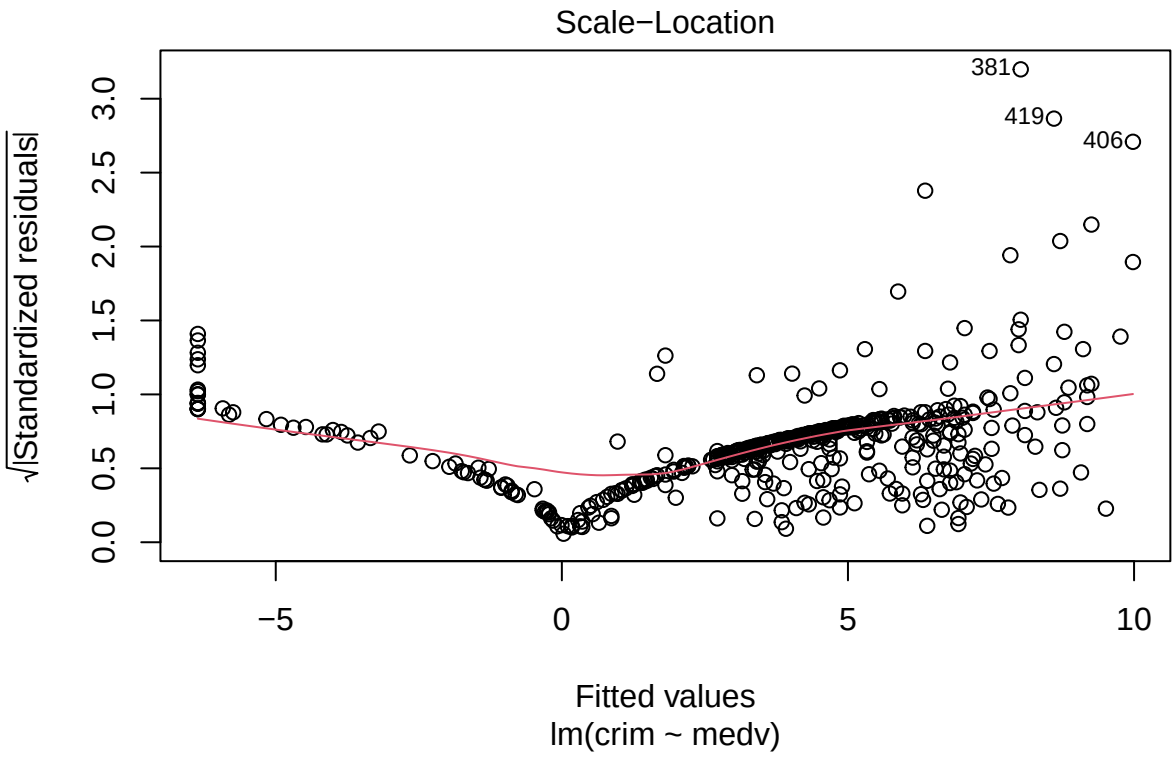

Q−Q Residuals
Standardized residuals
12
10
8
6
4
2
0
−2
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ lstat)

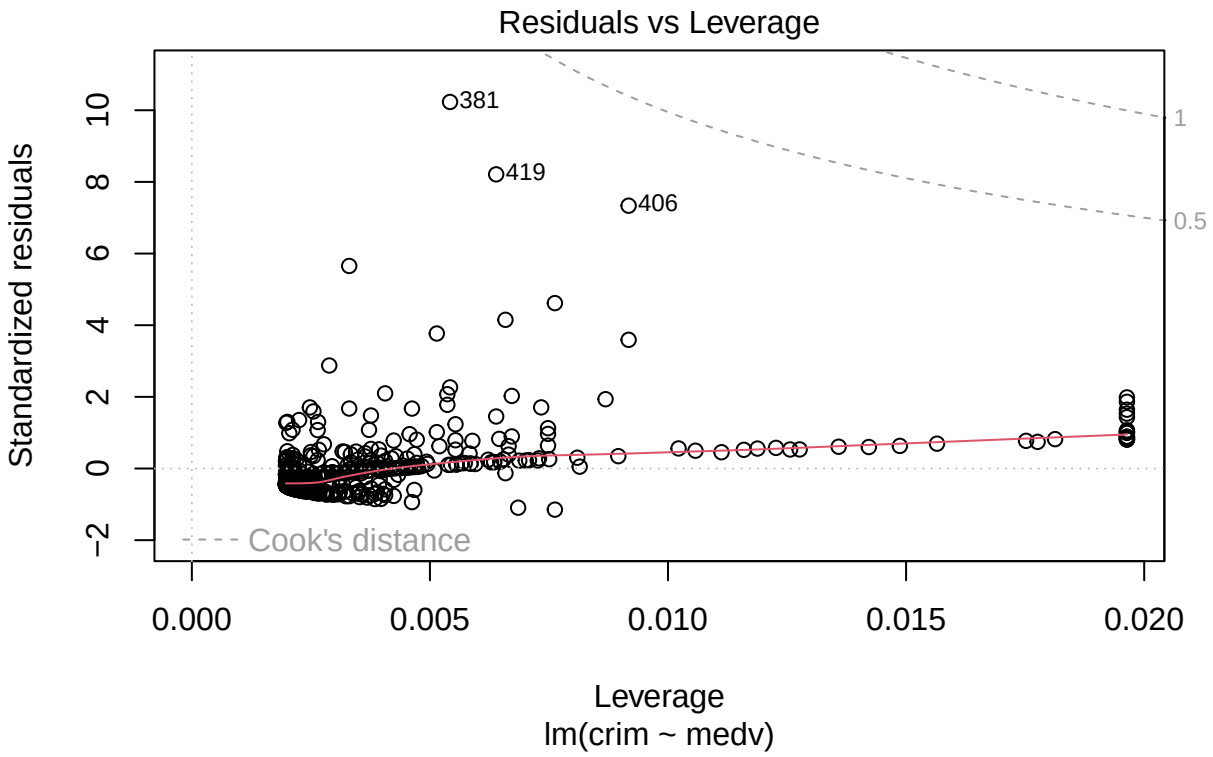

Scale−Location
381
419
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
5
10
15
Fitted values
lm(crim ~ lstat)


Residuals vs Leverage
Standardized residuals
10
5
0
419
406
415
1
0.5
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
lm(crim ~ lstat)


Residuals vs Fitted
Residuals
Fitted values
lm(crim ~ medv)
381
419
406
−20
0
20
40
60
80
−5
0
5
10


Q−Q Residuals
Standardized residuals
0
2
4
6
8
10
381
419
406
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(crim ~ medv)


Scale−Location
381
419
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−5
0
5
10
Fitted values
lm(crim ~ medv)



All, except chas there is a statistically significant association between the predictor and the response.

PART B

Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : Betaj = 0?

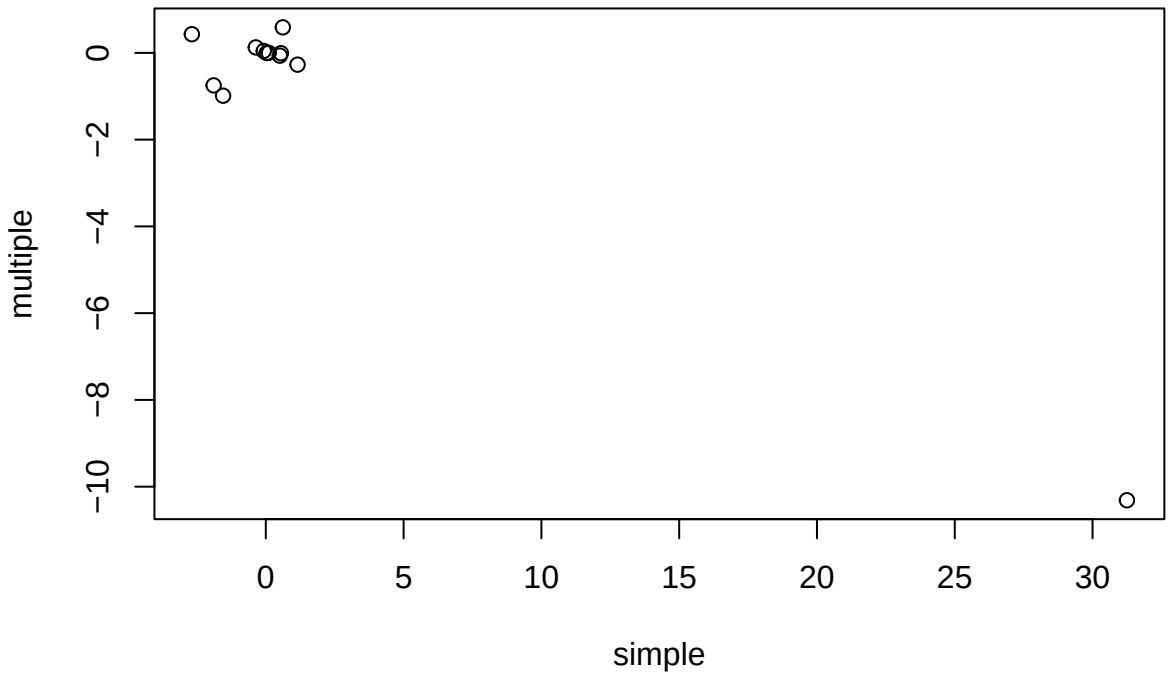
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chasY        -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
```

```
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

For the following predictors we can reject the null hypothesis:
zn, dis, rad, black, medv

PART C

How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regres- sion model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. #Compared to Part 1, fewer predictors in Part 2 had p-values that were low enough to provide strong evidence to reject the null hypothesis



Observations:

rm and chas increased the values and nox decreased

PART D

Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

```
##
## Call:
## lm(formula = crim ~ poly(zn, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.821 -4.614 -1.294  0.473 84.130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.6135     0.3722   9.709  < 2e-16 ***
## poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
## poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 **
## poly(zn, 3)3 -10.0719     8.3722  -1.203  0.22954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.372 on 502 degrees of freedom
## Multiple R-squared:  0.05824,    Adjusted R-squared:  0.05261
## F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
##
##
## Call:
## lm(formula = crim ~ poly(indus, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.278 -2.514  0.054  0.764 79.713
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.614      0.330  10.950  < 2e-16 ***
## poly(indus, 3)1   78.591      7.423  10.587  < 2e-16 ***
## poly(indus, 3)2  -24.395      7.423  -3.286  0.00109 **
## poly(indus, 3)3  -54.130      7.423  -7.292  1.2e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.423 on 502 degrees of freedom
## Multiple R-squared:  0.2597, Adjusted R-squared:  0.2552
## F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = crim ~ poly(nox, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.110 -2.068 -0.255  0.739 78.302
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3216  11.237  < 2e-16 ***
## poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
## poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
## poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.234 on 502 degrees of freedom
## Multiple R-squared:  0.297,  Adjusted R-squared:  0.2928
## F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = crim ~ poly(rm, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.485  -3.468  -2.221  -0.015  87.219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.6135     0.3703   9.758  < 2e-16 ***
## poly(rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
## poly(rm, 3)2  26.5768     8.3297   3.191  0.00151 **
## poly(rm, 3)3  -5.5103     8.3297  -0.662  0.50858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.33 on 502 degrees of freedom
## Multiple R-squared:  0.06779,    Adjusted R-squared:  0.06222
## F-statistic: 12.17 on 3 and 502 DF,  p-value: 1.067e-07
##
##
## Call:
## lm(formula = crim ~ poly(age, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.762 -2.673 -0.516  0.019 82.842
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3485  10.368  < 2e-16 ***
## poly(age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
## poly(age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
## poly(age, 3)3  21.3532     7.8397   2.724  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.84 on 502 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
## F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16
##
```

```
##
## Call:
## lm(formula = crim ~ poly(dis, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.757  -2.588   0.031   1.267  76.378
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3259  11.087  < 2e-16 ***
## poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
## poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
## poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.331 on 502 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2735
## F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = crim ~ poly(rad, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.381  -0.412  -0.269   0.179  76.217
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.2971  12.164  < 2e-16 ***
## poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
## poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 **
## poly(rad, 3)3   4.6985     6.6824   0.703  0.48231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.682 on 502 degrees of freedom
## Multiple R-squared:    0.4,  Adjusted R-squared:  0.3965
## F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = crim ~ poly(tax, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.273  -1.389   0.046   0.536  76.950
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.6135     0.3047  11.860  < 2e-16 ***
## poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
```

```
## poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
## poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.854 on 502 degrees of freedom
## Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
## F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = crim ~ poly(ptratio, 3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.833 -4.146 -1.655  1.408 82.697
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.614      0.361  10.008  < 2e-16 ***
## poly(ptratio, 3)1   56.045      8.122   6.901 1.57e-11 ***
## poly(ptratio, 3)2   24.775      8.122   3.050  0.00241 **
## poly(ptratio, 3)3  -22.280      8.122  -2.743  0.00630 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.122 on 502 degrees of freedom
## Multiple R-squared:  0.1138, Adjusted R-squared:  0.1085
## F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13
##
##
## Call:
## lm(formula = crim ~ poly(lstat, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.234  -2.151  -0.486   0.066  83.353
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.6135     0.3392  10.654   <2e-16 ***
## poly(lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
## poly(lstat, 3)2  15.8882     7.6294   2.082   0.0378 *
## poly(lstat, 3)3 -11.5740     7.6294  -1.517   0.1299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.629 on 502 degrees of freedom
## Multiple R-squared:  0.2179, Adjusted R-squared:  0.2133
## F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = crim ~ poly(medv, 3))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.427  -1.976  -0.437   0.439  73.655
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.614      0.292  12.374  < 2e-16 ***
## poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
## poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
## poly(medv, 3)3  -48.033      6.569  -7.312 1.05e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.569 on 502 degrees of freedom
## Multiple R-squared:  0.4202, Adjusted R-squared:  0.4167
## F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

In many variables there is evidence of non-linear relationship. Specifically indus, nox, age exhibit significant cubic terms which can be inferred from p-value.

Chapter 6 Question 9

In this exercise, we will predict the number of applications received using the other variables in the College data set.

PART A

Split the data set into a training set and a test set.

PART B
Fit a linear model using least squares on the training set, and report the test error obtained.

```
## [1] 1255496
```

PART C

Fit a ridge regression model on the training set, with chosen by cross-validation. Report the test error obtained.

```
## [1] 433.5803
## [1] 1094590
```

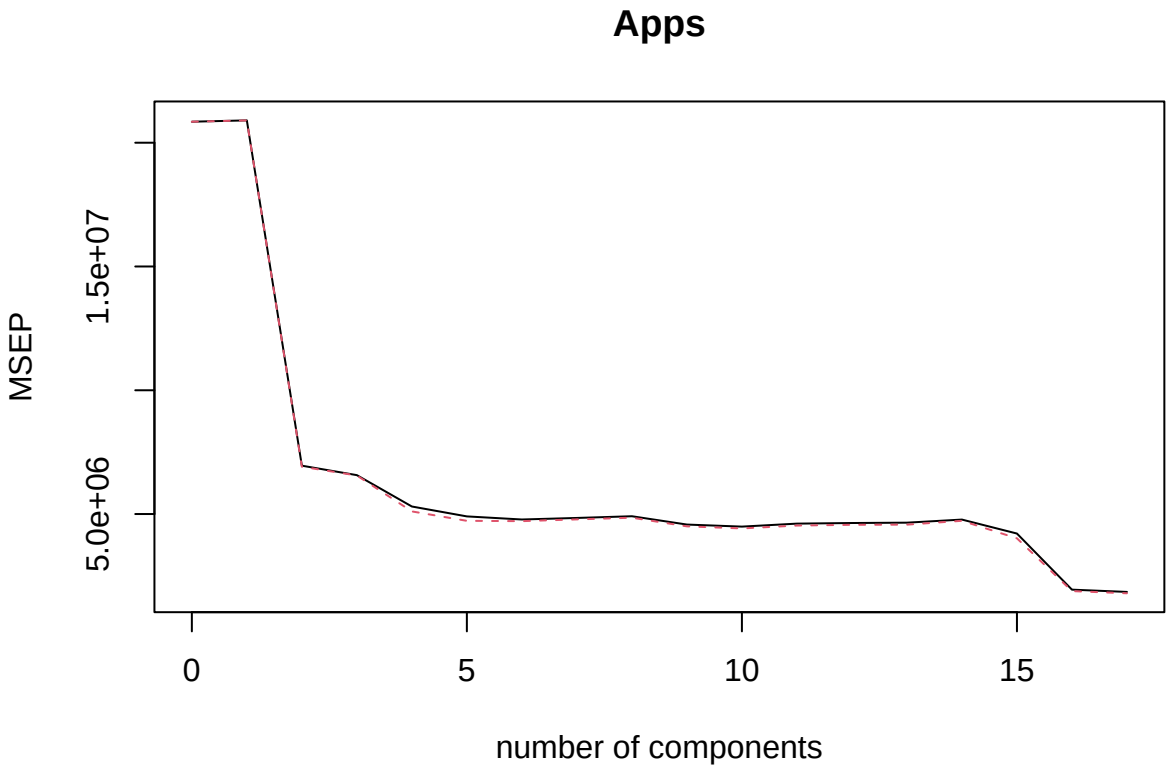
PART D

it a lasso model on the training set, with chosen by cross- validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

```
## [1] 2.313888
## [1] 1238666
## [1] "Number of Non-Zero Coefficients: 18"
```

PART E

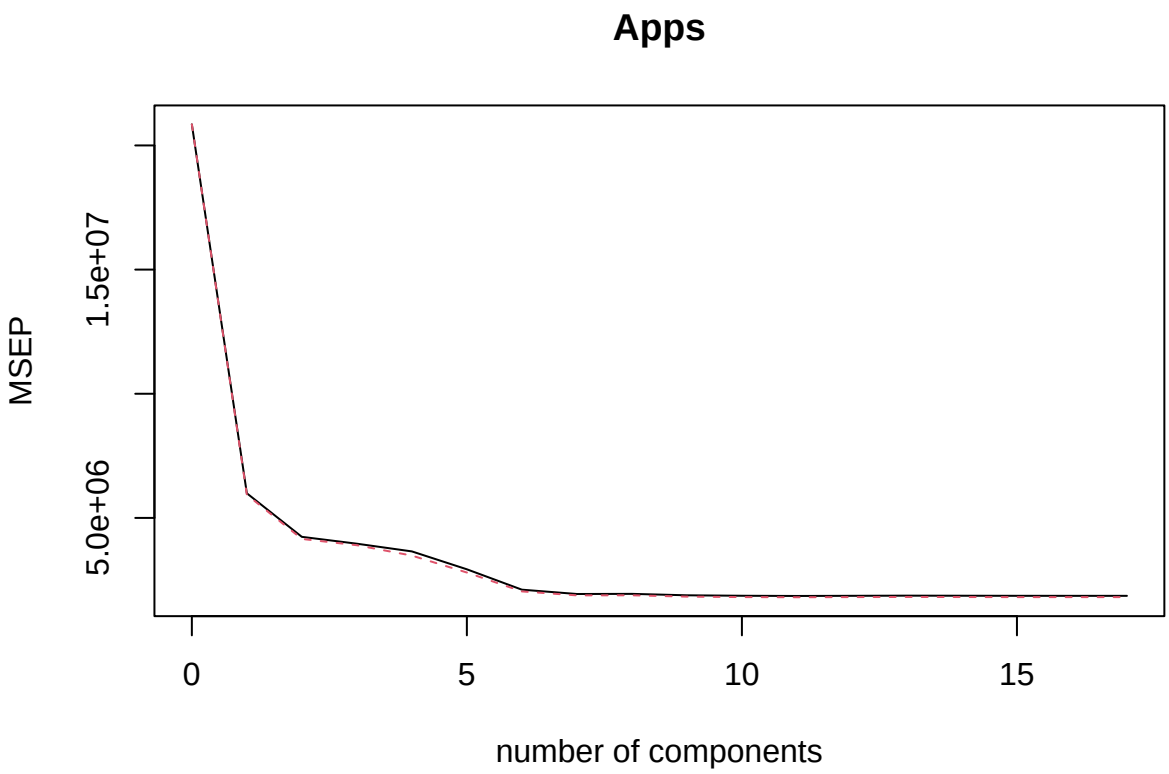
Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.




```
## [1] 1977556
```

PART F

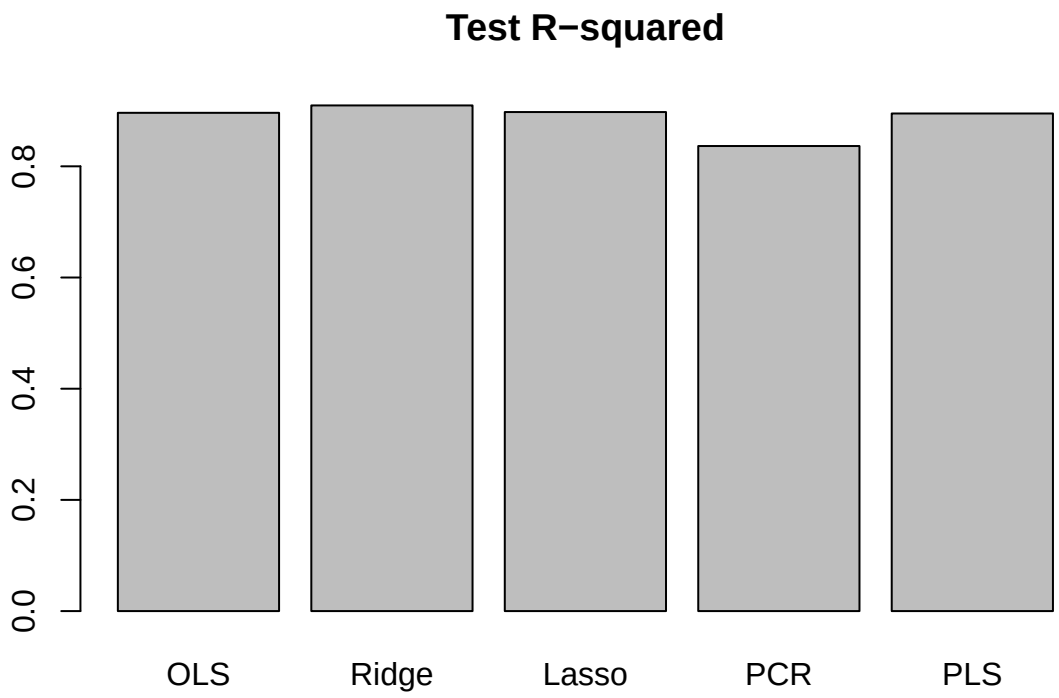
Fit a PLS model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.




```
## [1] 1270097
```

PART G

Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?



All models except PCR predicts college applications with high accuracy

Chapter 6 Question 11

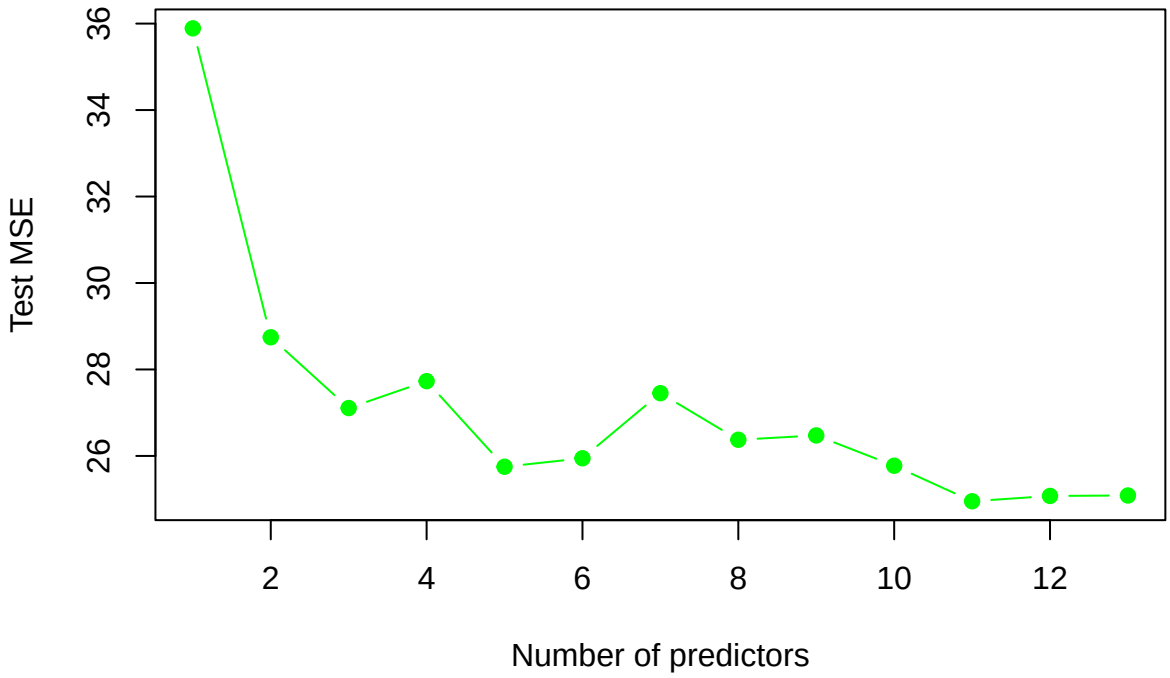
We will now try to predict per capita crime rate in the Boston data set

PART A

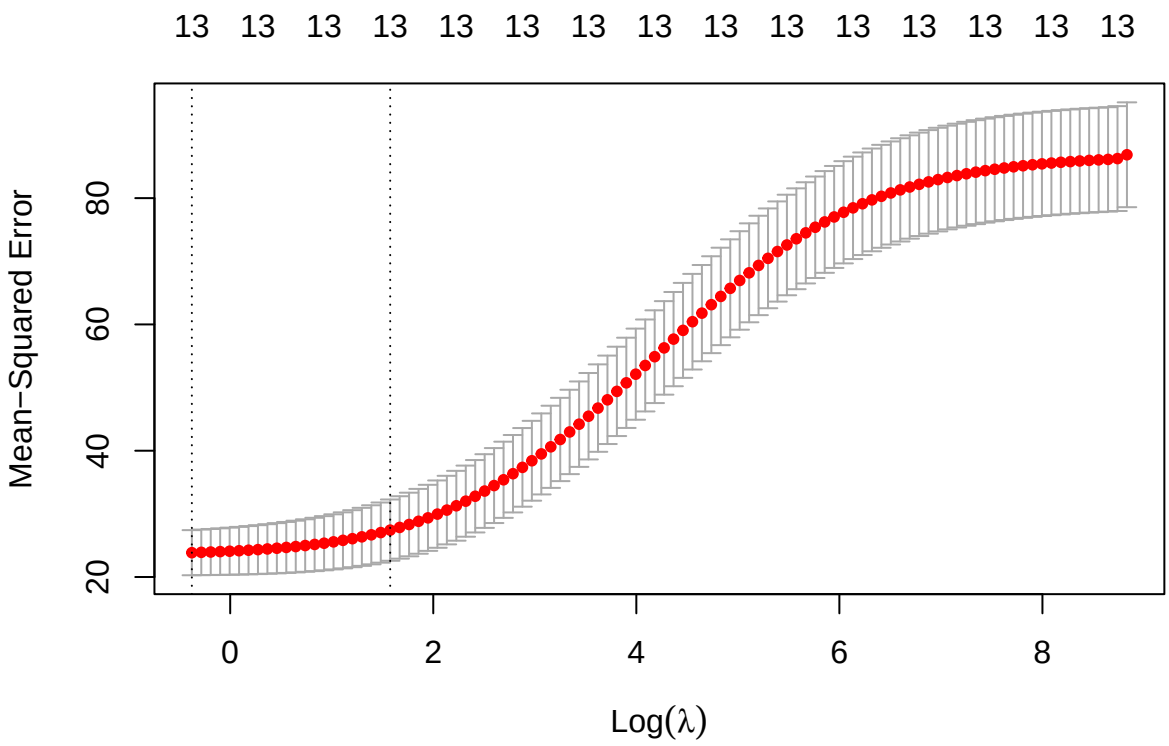
Try out some of the regression methods explored in this chapter, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
```

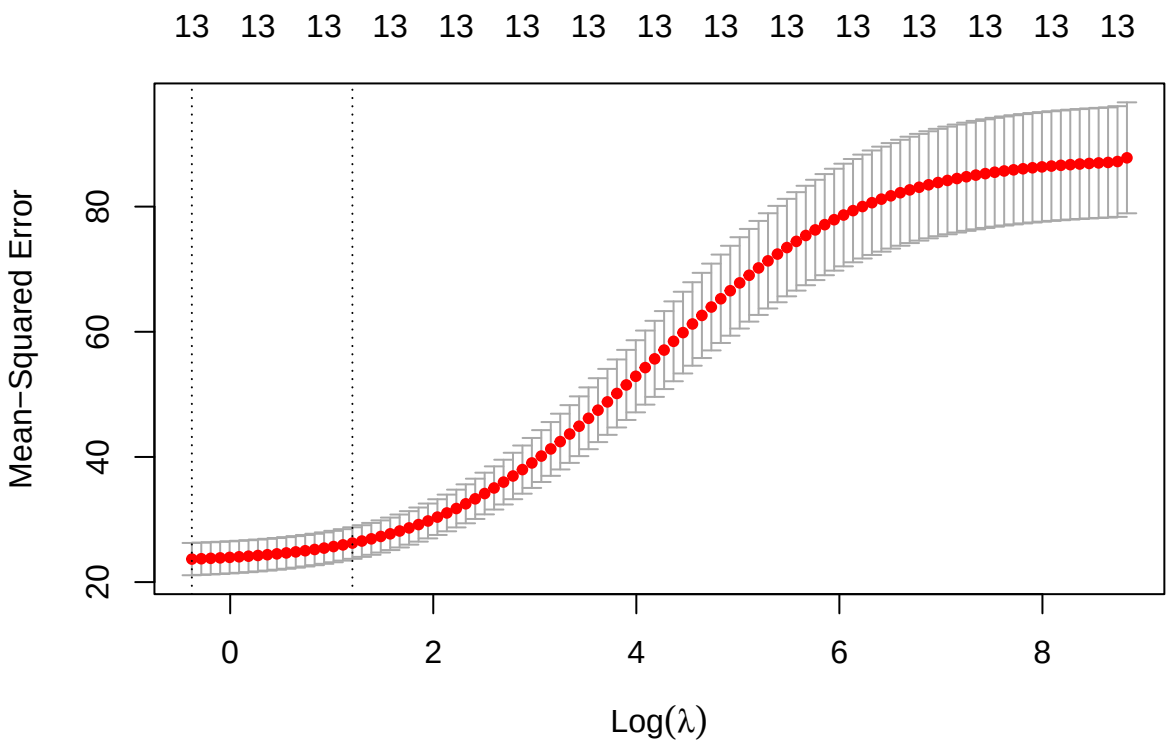
```
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
## Subset selection object
## Call: regsubsets.formula(medv ~ ., data = boston.train, nbest = 1,
##     nvmax = 13)
## 13 Variables  (and intercept)
##         Forced in Forced out
## crim        FALSE      FALSE
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##           crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 )  " "  " " " " " "   " "  " " " " " " " " " " " "     " "   "*"
## 2  ( 1 )  " "  " " " " " "   " "  " " "*" " " " " " " " " "     " "   "*"
## 3  ( 1 )  " "  " " " " " "   " "  " " "*" " " " " " " " " "*"     " "   "*"
## 4  ( 1 )  " "  " " " " " "   " "  " " "*" " " " " " " " " "*"     "*"   "*"
## 5  ( 1 )  " "  " " " " " "   " "  "*" "*" " " "*" " " " " " "*"     " "   "*"
## 6  ( 1 )  " "  " " " " " "   " "  "*" "*" " " "*" " " " " " "*"     "*"   "*"
## 7  ( 1 )  " "  "*" " " "   " "  "*" "*" " " "*" " " " " " "*"     "*"   "*"
## 8  ( 1 )  " "  "*" " " "   "*"  "*" "*" " " "*" " " " " " "*"     "*"   "*"
## 9  ( 1 )  " "  "*" " " "   " "  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 10  ( 1 ) "*"  "*" " " "   " "  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 11  ( 1 ) "*"  "*" " " "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 12  ( 1 ) "*"  "*" "*" "   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 13  ( 1 ) "*"  "*" "*" "   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```



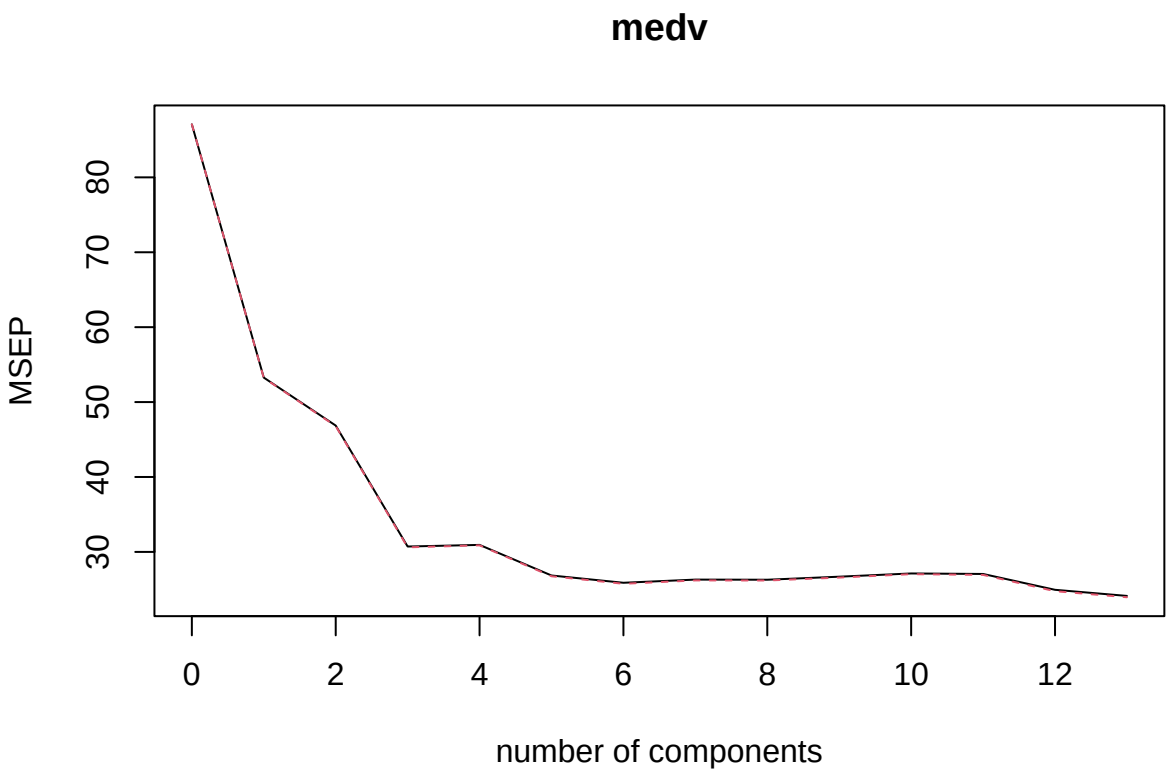

```
## [1] 24.95231
```




```
## [1] 0.6853196
## [1] 25.31414
```




```
## [1] 0.6853196
## [1] 25.31414
## Data:    X dimension: 354 13
##  Y dimension: 354 1
## Fit method: svdpc
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.331    7.298    6.846    5.542    5.562    5.182    5.087
## adjCV        9.331    7.296    6.841    5.533    5.557    5.172    5.074
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       5.129    5.127    5.166     5.208     5.201     4.994     4.912
## adjCV    5.118    5.117    5.155     5.199     5.189     4.975     4.894
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       47.46    58.83    68.45    75.23    81.33    86.17    90.28    93.30
## medv    39.21    49.00    66.01    66.50    70.76    71.72    71.72    71.85
##       9 comps  10 comps  11 comps  12 comps  13 comps
## X       95.26     96.90     98.33     99.56     100.0
## medv    71.93     71.95     73.00     75.03      75.8
```




```
## Data:    X dimension: 354 13
##  Y dimension: 354 1
## Fit method: svdpc
## Number of components considered: 5
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps
## X       47.46    58.83    68.45    75.23    81.33
## medv    39.21    49.00    66.01    66.50    70.76
## [1] 26.98119
```

PART B

Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross- validation, or some other reasonable alternative, as opposed to using training error.

BSM MSE 17.81448

Lasso MSE 17.53636

Ridge MSE 17.53636

PCR MSE 15.96318

PCR seems to be the best model for this dataset.

PART C

Does your chosen model involve all of the features in the data set? Why or why not?

My chosen model, PCR involves all the features in the data set.

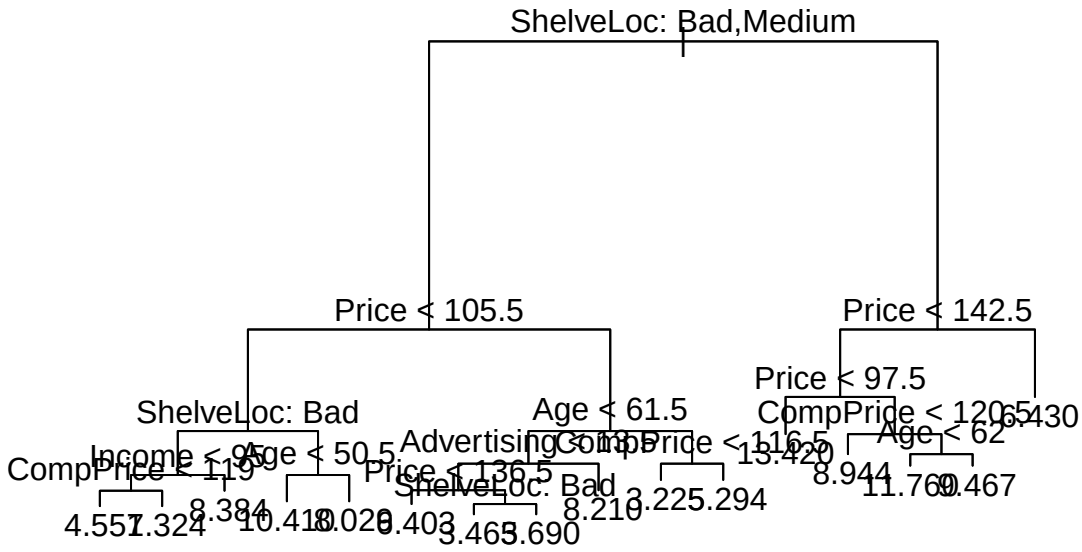
Chapter 8 Question 8

PART A

Split the data set into a training set and a test set.

PART B

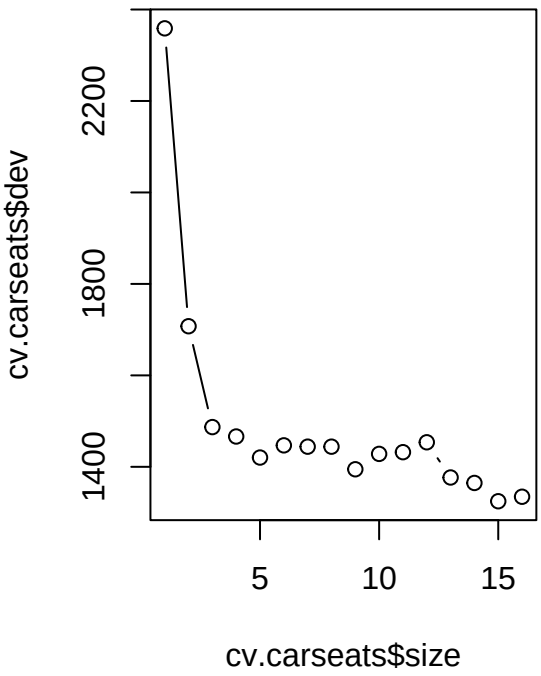
Fit a regression tree to the training set. Plot the tree, and inter- pret the results. What test MSE do you obtain?

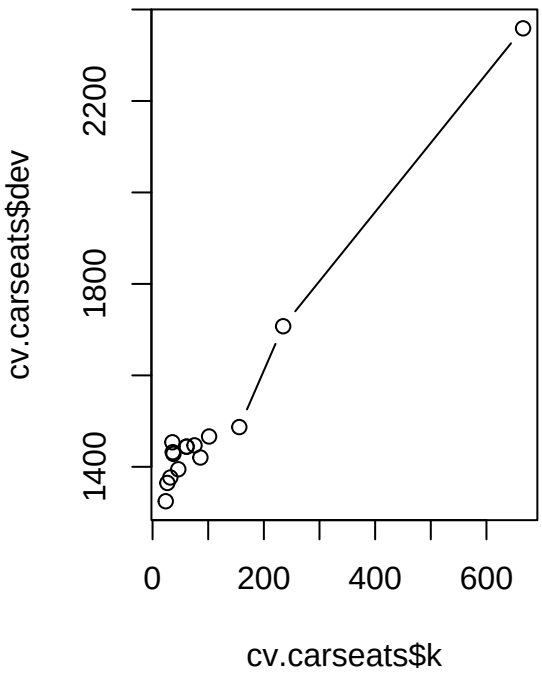


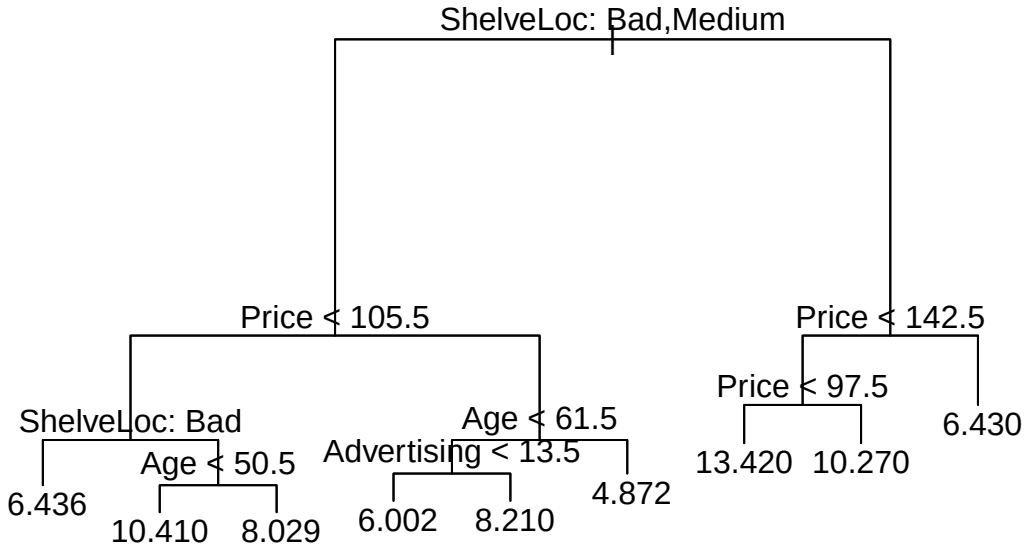

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = cartrain)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Income"      "CompPrice"   "Age"
## [6] "Advertising"
## Number of terminal nodes:  16
## Residual mean deviance:  2.511 = 662.8 / 264
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.86900 -1.09600 -0.05722  0.00000  1.16500  4.77600
## [1] 5.134134
```

PART C

Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?


cv.carseats$dev
1400
1800
2200
5
10
15
cv.carseats$size


cv.carseats$dev
1400
1800
2200
0
200
400
600
cv.carseats$k




```
## [1] 5.584126
```

No, pruning does not improve the test MSE.

PART D

Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

```
## [1] 2.698446
##               %IncMSE IncNodePurity
## CompPrice   26.194326    206.399166
## Income       9.407570    105.997185
## Advertising 18.059201    133.582805
## Population   1.349397     78.879498
## Price       68.541228    657.574271
## ShelveLoc   76.165372    773.725961
## Age         26.370140    253.705211
## Education    2.777495     60.119691
## Urban       -1.383561      6.468684
## US           1.265584      8.139500
```

ShelveLoc and Price variables are the most important according to the results.

PART E

Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

```
## [1] 2.568116
##                %IncMSE IncNodePurity
## CompPrice   17.7122647     193.24329
## Income       5.6138384     138.70920
## Advertising 16.2479259     157.15856
## Population   2.5601957     122.44596
## Price       52.7521338     585.45785
## ShelveLoc   61.4262507     668.64699
## Age         23.7194373     278.22551
## Education   -0.4780998      71.45863
## Urban        0.3304407      10.26137
## US           2.5160355      17.11510
```

ShelveLoc and Price variables are the most important according to the results.

PART F

Now analyze the data using BART, and report your results.

```
## *****Calling gbart: type=1
## *****Data:
## data:n,p,np: 280, 15, 120
## y1,yn: 0.088000, 4.528000
## x1,x[n*p]: 7.520000, 1.000000
## xp1,xp[np*p]: 10.810000, 1.000000
## *****Number of Trees: 200
## *****Number of Cut Points: 100 ... 1
## *****burn,nd,thin: 100,1000,1
## *****Prior:beta,alpha,tau,nu,lambda,offset: 2,0.95,0.284787,3,1.41657e-30,7.432
## *****sigma: 0.000000
## *****w (weights): 1.000000 ... 1.000000
## *****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,15,0
## *****printevery: 100
##
## MCMC
## done 0 (out of 1100)
## done 100 (out of 1100)
## done 200 (out of 1100)
## done 300 (out of 1100)
## done 400 (out of 1100)
## done 500 (out of 1100)
## done 600 (out of 1100)
## done 700 (out of 1100)
## done 800 (out of 1100)
## done 900 (out of 1100)
## done 1000 (out of 1100)
## time: 5s
## trcnt,tecnt: 1000,1000
## [1] 0.05883333
```

BART is the best model as it gives the lowest MSE.

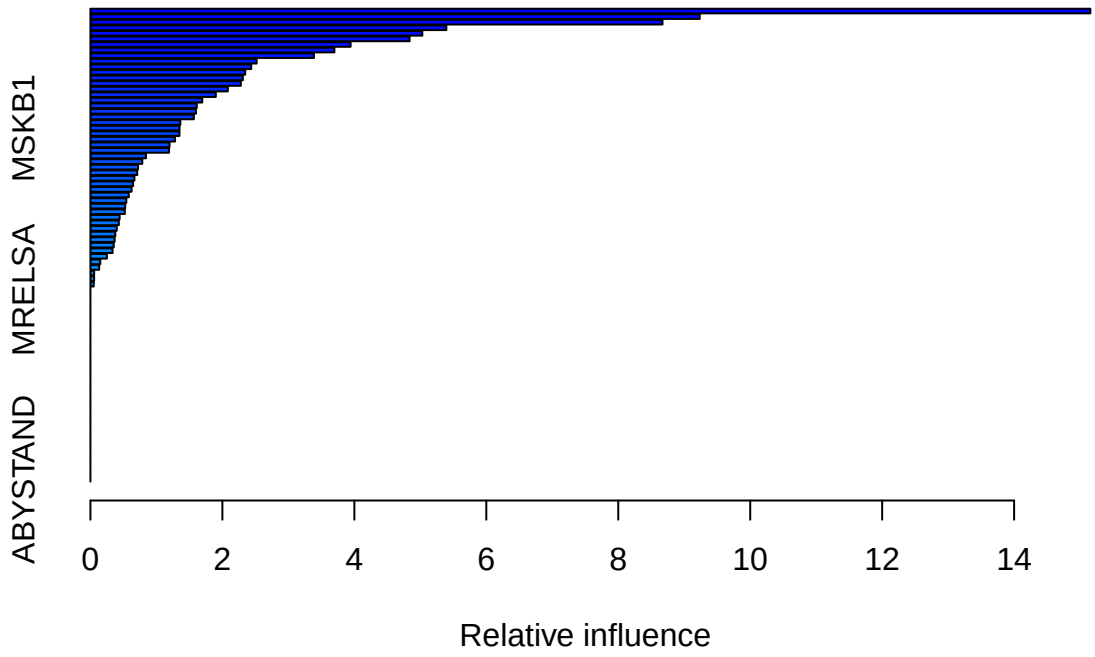
Chapter 8 Question 11

PART A

Create a training set consisting of the first 1,000 observations,and a test set consisting of the remaining observations.

PART B

Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?




```
##                var     rel.inf
## PPERSAUT PPERSAUT 15.15534009
## MKOOPKLA MKOOPKLA  9.23499526
## MOPLHOOG MOPLHOOG  8.67017024
## MBERMIDD MBERMIDD  5.39403655
## MGODGE     MGODGE  5.03047673
## PBRAND     PBRAND  4.83740038
## MINK3045 MINK3045  3.94305387
## ABRAND     ABRAND  3.69692919
## MOSTYPE   MOSTYPE  3.38768960
## PWAPART   PWAPART  2.51970169
## MGODPR     MGODPR  2.43689096
## MSKC         MSKC  2.34594774
## MAUT2       MAUT2  2.30973409
## MFWEKIND MFWEKIND  2.27959503
```

```
## MBERARBG MBERARBG  2.08245286
## MSKA         MSKA  1.90020973
## PBYSTAND PBYSTAND  1.69481877
## MGODOV     MGODOV  1.61147668
## MAUT1       MAUT1  1.59879109
## MBERHOOG MBERHOOG  1.56791308
## MINK7512 MINK7512  1.36255296
## MSKB1       MSKB1  1.35071475
## MINKGEM   MINKGEM  1.34913011
## MRELGE     MRELGE  1.28204167
## MAUT0       MAUT0  1.19929798
## MHHUUR     MHHUUR  1.19158719
## MFGEKIND MFGEKIND  0.84203310
## MRELOV     MRELOV  0.78554535
## MZPART     MZPART  0.72191139
## MINK4575 MINK4575  0.70935967
## MSKB2       MSKB2  0.66694112
## APERSAUT APERSAUT  0.64644681
## MGODRK     MGODRK  0.62380797
## MSKD         MSKD  0.58168337
## MINKM30   MINKM30  0.54392696
## PMOTSCO   PMOTSCO  0.52708603
## MOPLMIDD MOPLMIDD  0.52091706
## MGEMOMV   MGEMOMV  0.44231264
## MZFONDS   MZFONDS  0.43037800
## PLEVEN     PLEVEN  0.39901552
## MHKOOP     MHKOOP  0.37672230
## MBERARBO MBERARBO  0.36653424
## MBERBOER MBERBOER  0.35290257
## MINK123M MINK123M  0.33559225
## MGEMLEEF MGEMLEEF  0.24937634
## MFALLEEN MFALLEEN  0.14898856
## MOSHOOFD MOSHOOFD  0.13265308
## MOPLLAAG MOPLLAAG  0.05654615
## MBERZELF MBERZELF  0.05589282
## MAANTHUI MAANTHUI  0.05047841
## MRELSA     MRELSA  0.00000000
## PWABEDR   PWABEDR  0.00000000
## PWALAND   PWALAND  0.00000000
## PBESAUT   PBESAUT  0.00000000
## PVRAAUT   PVRAAUT  0.00000000
## PAANHANG PAANHANG  0.00000000
## PTRACTOR PTRACTOR  0.00000000
## PWERKT     PWERKT  0.00000000
## PBROM       PBROM  0.00000000
## PPERSONG PPERSONG  0.00000000
## PGEZONG   PGEZONG  0.00000000
## PWAOREG   PWAOREG  0.00000000
## PZEILPL   PZEILPL  0.00000000
## PPLEZIER PPLEZIER  0.00000000
## PFIETS     PFIETS  0.00000000
## PINBOED   PINBOED  0.00000000
## AWAPART   AWAPART  0.00000000
## AWABEDR   AWABEDR  0.00000000
```

```
## AWALAND   AWALAND  0.00000000
## ABESAUT   ABESAUT  0.00000000
## AMOTSCO   AMOTSCO  0.00000000
## AVRAAUT   AVRAAUT  0.00000000
## AAANHANG AAANHANG  0.00000000
## ATRACTOR ATRACTOR  0.00000000
## AWERKT     AWERKT  0.00000000
## ABROM       ABROM  0.00000000
## ALEVEN     ALEVEN  0.00000000
## APERSONG APERSONG  0.00000000
## AGEZONG   AGEZONG  0.00000000
## AWAOREG   AWAOREG  0.00000000
## AZEILPL   AZEILPL  0.00000000
## APLEZIER APLEZIER  0.00000000
## AFIETS     AFIETS  0.00000000
## AINBOED   AINBOED  0.00000000
## ABYSTAND ABYSTAND  0.00000000
```

PPERSAUT is the most important variable.

PART C

Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated prob- ability of purchase is greater than 20 %. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one? How does this compare with the results obtained from applying KNN or logistic regression to this data set?

```
##    pred.test
##        0    1
##   0 4396  137
##   1  255   34
## [1] 0.1988304
##    pred.test2
##        0    1
##   0 4183  350
##   1  231   58
## [1] 0.1421569
```

Boosting has a better true-positive rate than Logistic regression.

Chapter 10 Question 7

Fit a neural network to the Default data. Use a single hidden layer with 10 units, and dropout regularization. Have a look at Labs 10.9.1– 10.9.2 for guidance. Compare the classification performance of your model with that of linear logistic regression.

```
##
## Attaching package: 'tensorflow'
```

```
## The following object is masked from 'package:caret':
##
##     train
```

Problem 1 Beauty

PART 1

Using the data, estimate the effect of "beauty" into course ratings. Make sure to think about the potential many "other determinants". Describe your analysis and your conclusions.

```
##
## Call:
## lm(formula = CourseEvals ~ BeautyScore + female + lower + nonenglish +
##     tenuretrack, data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31385 -0.30202  0.01011  0.29815  1.04929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.06542    0.05145  79.020  < 2e-16 ***
## BeautyScore  0.30415    0.02543  11.959  < 2e-16 ***
## female      -0.33199    0.04075  -8.146 3.62e-15 ***
## lower       -0.34255    0.04282  -7.999 1.04e-14 ***
## nonenglish  -0.25808    0.08478  -3.044  0.00247 **
## tenuretrack -0.09945    0.04888  -2.035  0.04245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4273 on 457 degrees of freedom
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3399
## F-statistic: 48.58 on 5 and 457 DF,  p-value: < 2.2e-16
```

Result

We observe the following:

Increase in BeautyScore is associated with higher course ratings (coeff. 0.30415, $p < 2e\text{-}16$)

Other factors like being female, teaching lower-level courses, non-English speaking, and tenure track status, are linked to lower course ratings.

The model explains approximately 34.71% of the variance in course ratings (R-squared = 0.3471), suggesting that multiple factors influence course ratings.

PART 2

In his paper, Dr. Hamermesh has the following sentence: "Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible". Using the concepts we have talked about so far, what does he mean by that?

Dr. Hamermesh discusses the difficulty of identifying the true cause of a specific outcome, such as course ratings.

The outcome may be related to both productivity and discrimination. Disentangling these factors is complex, as various variables and interactions make it challenging to pinpoint the sole cause.

As a result, determining whether the outcome is dependent on productivity or discrimination is likely to be extremely difficult.

Problem 2: Housing Price Structure

PART 1

Is there a premium for brick houses everything else being equal?

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -9848.16257 9818.681830 -1.003003 3.178612e-01
## Nbhd         9790.37832 1804.111637  5.426703 2.997626e-07
## Offers      -8396.98324 1244.696367 -6.746210 5.540382e-10
## SqFt           50.14427    6.560109  7.643816 5.551626e-12
## BrickYes    15603.19192 2252.948428  6.925676 2.245953e-10
## Bedrooms     5618.29659 1816.568308  3.092808 2.462214e-03
## Bathrooms    8294.95035 2428.404807  3.415802 8.663942e-04
```

Yes, there is a premium for brick houses when everything else is equal.

The coefficient for the "BrickYes" variable is positive (15603.19192). This indicates that, after accounting for the effects of other predictors (neighborhood, number of offers, square footage, number of bedrooms, and number of bathrooms), houses made of brick tend to have higher selling prices compared to houses that are not made of brick.

PART 2

Is there a premium for houses in neighborhood 3?

The coefficient for the "Nbhd" variable is 9790.37832 , indicating houses in Neighborhood 3 have higher selling prices than the reference neighborhood. The statistical significance suggests this difference is not by chance.

We can conclude that houses in Neighborhood 3 has a premium in selling prices, accounting for other predictors in the model.

PART 3

Is there an extra premium for brick houses in neighborhood 3?

```
##                  Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)   -3560.91844 9651.823868 -0.3689374 7.128246e-01
## Nbhd           7253.37224 1905.577685  3.8063902 2.235271e-04
## BrickYes      -4086.69091 6453.284022 -0.6332731 5.277604e-01
## Offers        -8479.51754 1198.826239 -7.0731831 1.092157e-10
## SqFt             50.69028    6.319174  8.0216623 7.875770e-13
## Bedrooms       6249.68073 1760.051534  3.5508510 5.494925e-04
## Bathrooms      6455.46812 2406.331403  2.6827012 8.333289e-03
## Nbhd:BrickYes  9438.71570 2913.459951  3.2396930 1.548328e-03
```

The interaction term "Nbhd:BrickYes" has a positive and statistically significant coefficient (9438.71570).

This indicates an extra premium for brick houses in Neighborhood 3, where the price difference between brick and non-brick houses is more than in other neighborhoods.

PART 4

For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

Problem 3: What causes what??

PART A

Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)

Using a simple regression of "Crime" on "Police" to understand how more cops affect crime is inadequate due to bias by omitted variables.

Other factors influence crime rates (e.g., socioeconomic conditions), and these unaccounted variables lead to biased estimates.

To determine causation, researchers use advanced methods like randomized trials and instrumental variable regression.

PART B

How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below.

The researchers from UPenn aimed to isolate the effect of police presence on daily crime rates in Washington, D.C.

The table reports the total daily crime decreases on "High-Alert Days" along with additional control variables.

The R-squared values (.14 and .17) suggest that approximately 14% and 17% of the variation in daily crime can be explained by the variables in the respective models.

The researchers' findings support the hypothesis that increased police presence can lead to a decrease in crime.

PART C

Why did they have to control for METRO ridership? What was that trying to capture?

The researchers controlled for METRO ridership to account for a potential confounding variable that might have influenced the relationship between police presence and crime rates.

Controlling for METRO ridership helps isolate the specific impact of police presence on crime rates, minimizing the influence of other factors that could affect crime independently of police presence. It allows researchers to determine the unique effect of police on crime.

PART D

In the next page, I am showing you "Table 4" from the research paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

The model being estimated in "Table 4" aims to examine the reduction in crime on "High-Alert Days" with a focus on the National Mall area. The researchers used a regression analysis with several independent variables.

The results show a decrease in crime in the National Mall district during "High-Alert Days," while other districts show a smaller reduction. Also, higher METRO ridership is associated with reduced crime rates.

Problem 4: Final Project

Describe your contribution to the final group project

In our team project, we analyzed a Telco churn dataset from IBM with 10,000 rows and a 26% monthly churn rate. My main contributions were in two areas. First, I conducted Exploratory Data Analysis (EDA)

to understand patterns, trends, and relationships in the data, gaining insights for decision-making. Second, I built a Boosting model, calculating evaluation metrics like Out-of-Bag Loss (OLIB), In-Bag Loss (ILIB), and Mean Squared Error (MSE). Additionally, I handled data encoding for categorical features, ensuring our model can process all relevant information effectively.