MGSC 310 - Problem Set #3

Ananya Vittal

1.

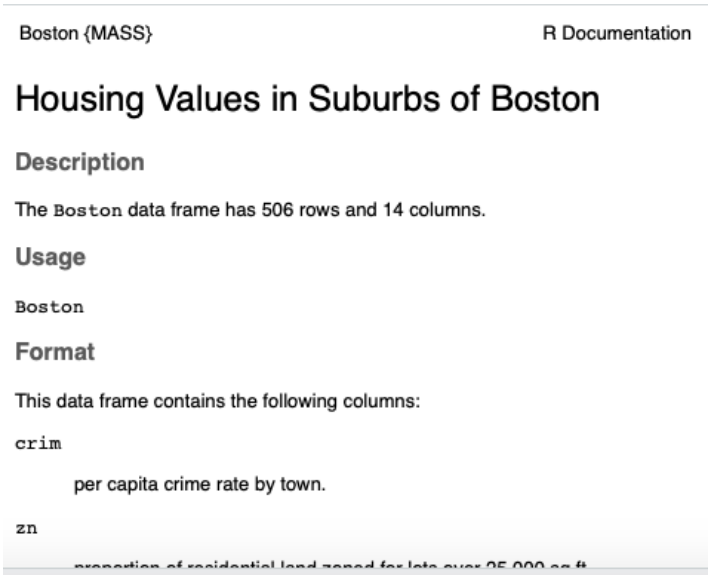
a) If the true relationship betwen X and Y is linear, then when comparing the training RSS for the linear regression and the training RSS for the cubic regression, I would expect the training RSS for the linear regression to have a lower value because the true relationship is linear so there would be less errors in the linear regression.

b) When comparing the test RSS for the linear regression and the test RSS for the cubic regression, I would still expect the test RSS for linear regression to have a lower value becuase the cubic regression on the test data might result in more errors due to overfitting.

2.

a)

```
help(Boston)

## No documentation for 'Boston' in specified packages and libraries:
## you could try '??Boston'
```



Boston {MASS} R Documentation

Housing Values in Suburbs of Boston

Description

The Boston data frame has 506 rows and 14 columns.

Usage

Boston

Format

This data frame contains the following columns:

crim

per capita crime rate by town.

zn

There are 506 observations and 14 variables in the dataset.

b)

```
library(MASS)
data(Boston)

correlations <- cor(Boston, use="complete.obs", method="pearson")
correlations
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                  rm         age         dis          rad         tax
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593
##             ptratio       black      lstat       medv
## crim     0.2899456 -0.38506394  0.4556215 -0.3883046
## zn      -0.3916785  0.17552032 -0.4129946  0.3604453
## indus    0.3832476 -0.35697654  0.6037997 -0.4837252
## chas    -0.1215152  0.04878848 -0.0539293  0.1752602
## nox      0.1889327 -0.38005064  0.5908789 -0.4273208
## rm      -0.3555015  0.12806864 -0.6138083  0.6953599
## age      0.2615150 -0.27353398  0.6023385 -0.3769546
## dis     -0.2324705  0.29151167 -0.4969958  0.2499287
## rad      0.4647412 -0.44441282  0.4886763 -0.3816262
## tax      0.4608530 -0.44180801  0.5439934 -0.4685359
## ptratio  1.0000000 -0.17738330  0.3740443 -0.5077867
```

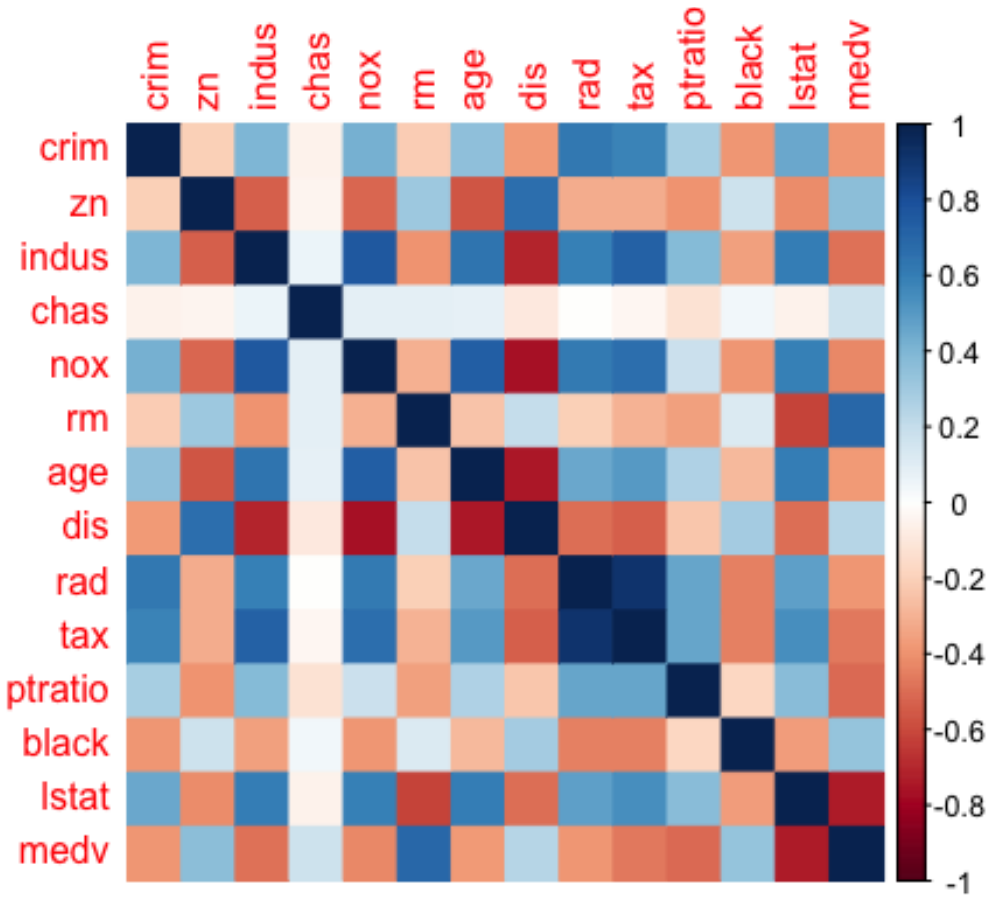
```
## black   -0.1773833  1.00000000 -0.3660869  0.3334608
## lstat    0.3740443 -0.36608690  1.0000000 -0.7376627
## medv    -0.5077867  0.33346082 -0.7376627  1.0000000
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.4.2
```

```
## corrplot 0.84 loaded
```

```
corrplot(correlations,method='color')
```



The four variables that are most strongly correlated with medv are: lstat, rm, ptratio, indus

c)

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.4.4
```

```r
# Code needed to replicate results
set.seed(99)

#Regression
regression1 <- lm(medv ~ lstat + rm + ptratio + indus, data = Boston)
summary(regression1)
```

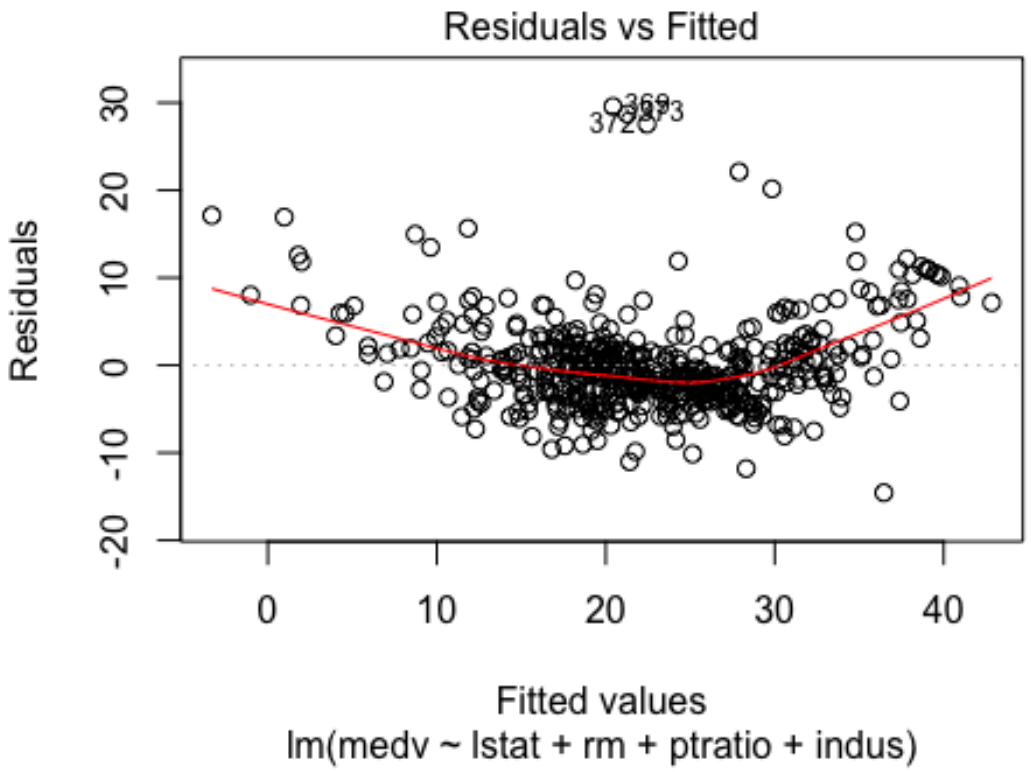
```
##
## Call:
## lm(formula = medv ~ lstat + rm + ptratio + indus, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5602  -3.1379  -0.7984   1.7783  29.5739
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.614970   3.926680   4.741 2.78e-06 ***
## lstat       -0.575711   0.047885 -12.023  < 2e-16 ***
## rm           4.515179   0.426286  10.592  < 2e-16 ***
## ptratio     -0.935122   0.120464  -7.763 4.71e-14 ***
## indus        0.007567   0.043594   0.174    0.862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.234 on 501 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6761
## F-statistic: 264.5 on 4 and 501 DF,  p-value: < 2.2e-16
```

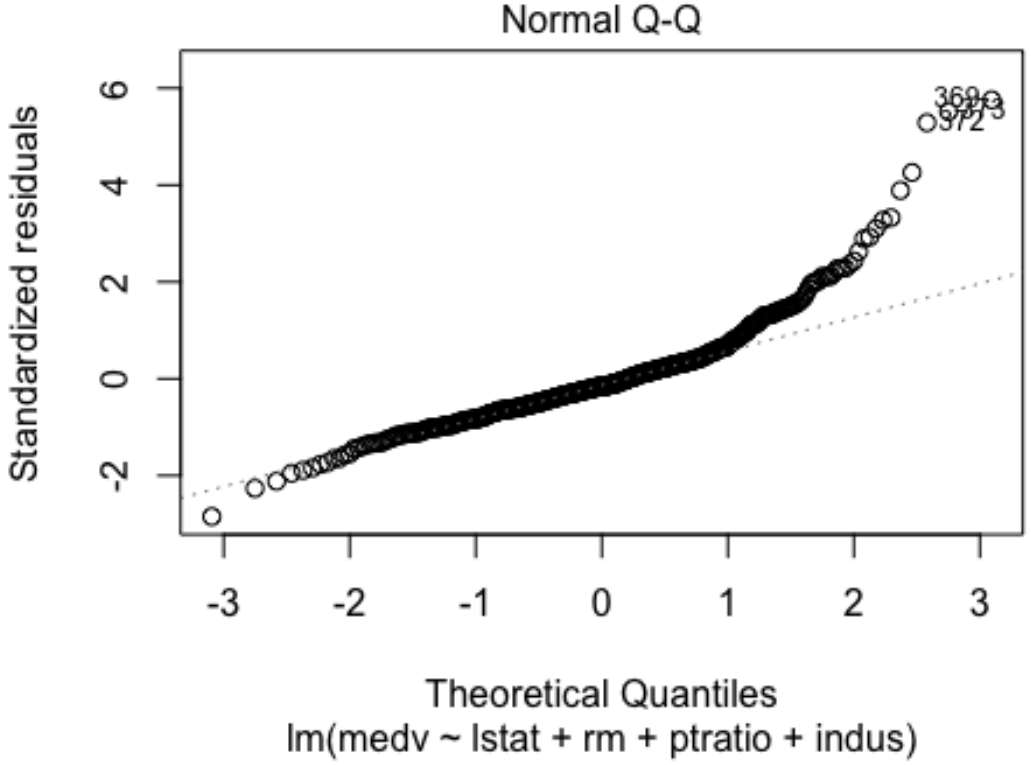
d)

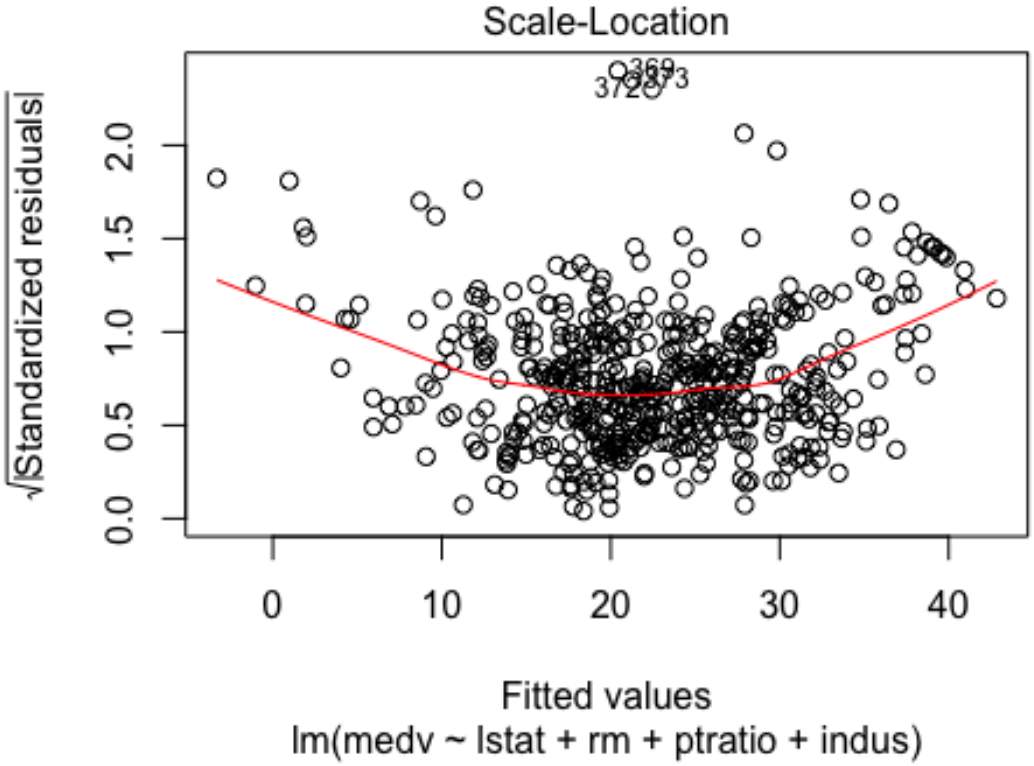
The coefficients for lstat, rm, and ptratio are less that 0.05 which means that they are statistically significant. However, the coefficient for indus is greater than 0.05 which means that it is not statistically significant.

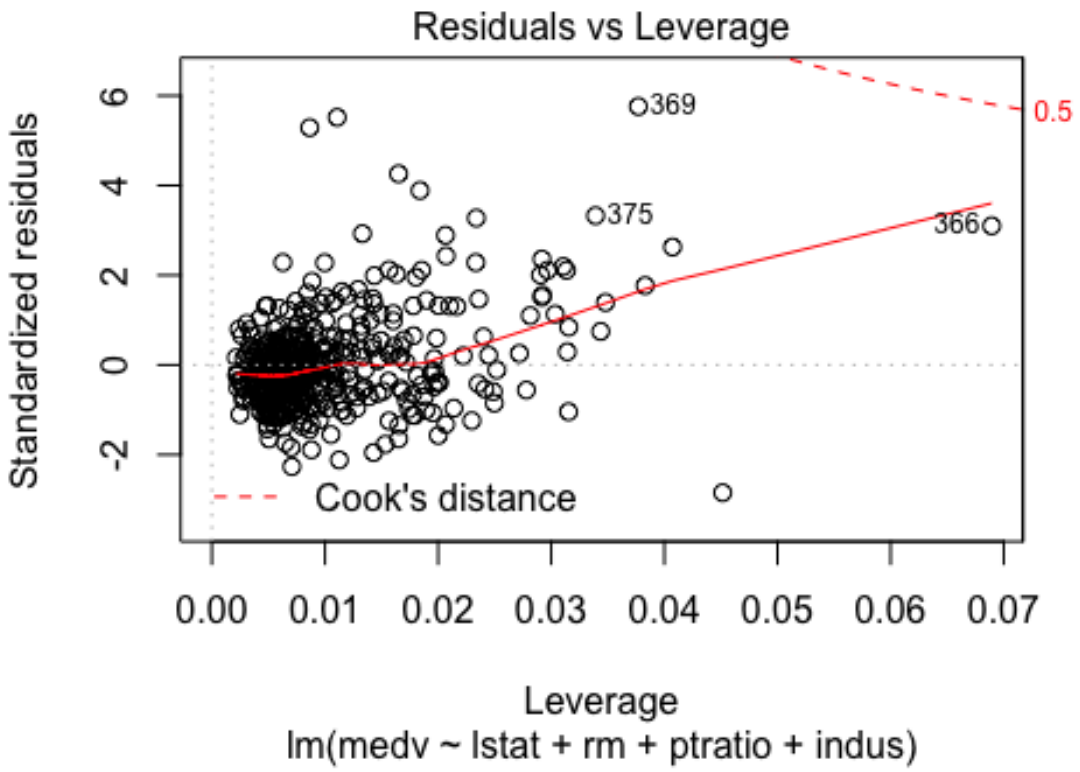
e)

```r
resid <- as.data.frame(residuals(regression1))
plot(regression1)
```


Residuals vs Fitted
Residuals
30
20
10
0
-10
-20
0
10
20
30
40
Fitted values
lm(medv ~ lstat + rm + ptratio + indus)


Normal Q-Q
Standardized residuals
-2
0
2
4
6
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(medv ~ lstat + rm + ptratio + indus)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Fitted values
lm(medv ~ lstat + rm + ptratio + indus)



f)

Looking at the residuals vs. fitted graph, there is some evidence of heteroscedacity because there is non-constant variance of errors.

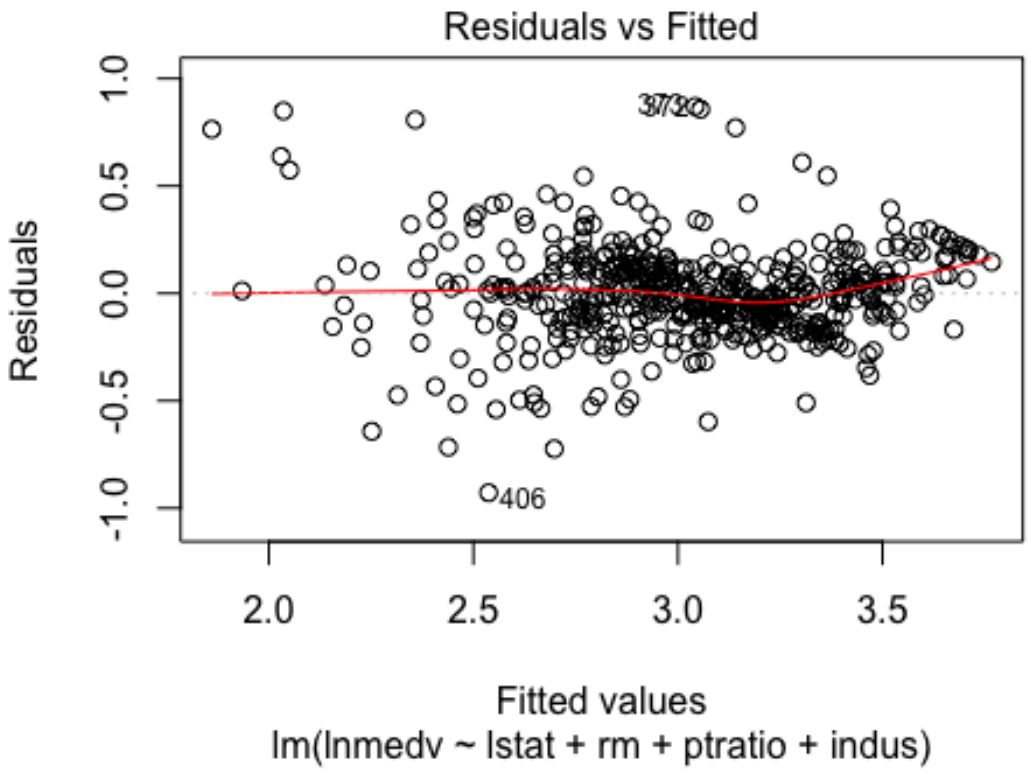
g)

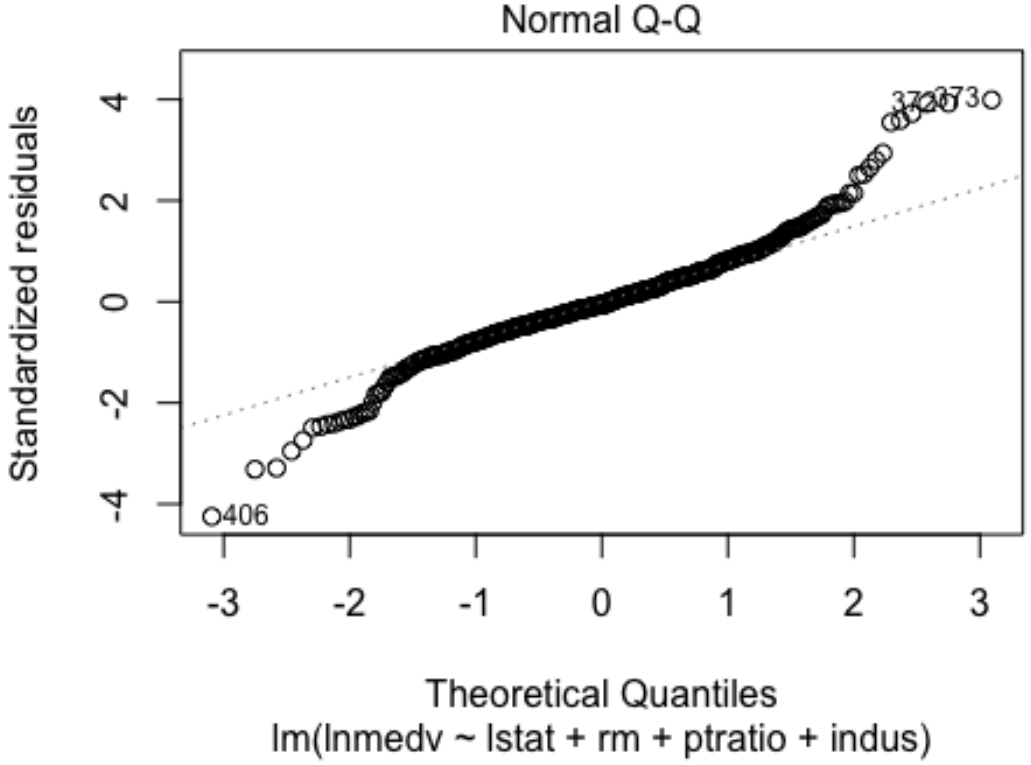
```
Boston$lnmedv <- log(Boston$medv)
regression2 <- lm(lnmedv ~ lstat + rm + ptratio + indus, data = Boston)
summary(regression2)

##
## Call:
## lm(formula = lnmedv ~ lstat + rm + ptratio + indus, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92790 -0.11001 -0.01274  0.10998  0.86993
##
## Coefficients:
```

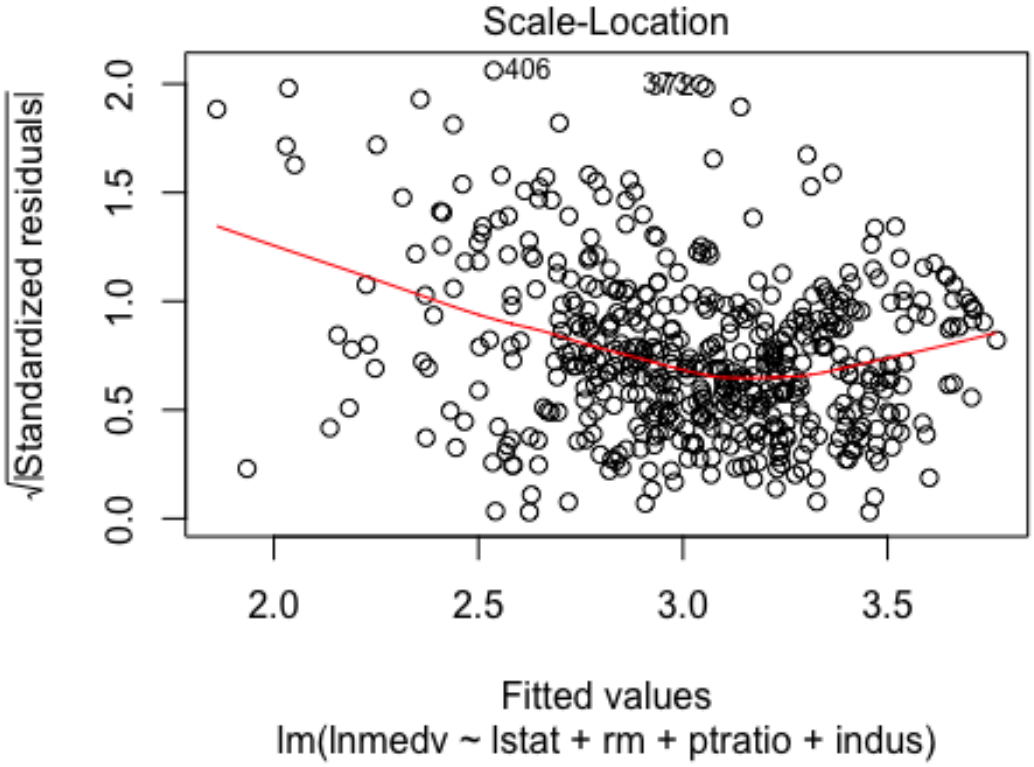
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.535070   0.164366  21.507  < 2e-16 ***
## lstat       -0.034376   0.002004 -17.150  < 2e-16 ***
## rm           0.104442   0.017844   5.853 8.73e-09 ***
## ptratio     -0.037987   0.005042  -7.533 2.33e-13 ***
## indus       -0.001878   0.001825  -1.029    0.304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2191 on 501 degrees of freedom
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.7127
## F-statistic: 314.1 on 4 and 501 DF,  p-value: < 2.2e-16
```

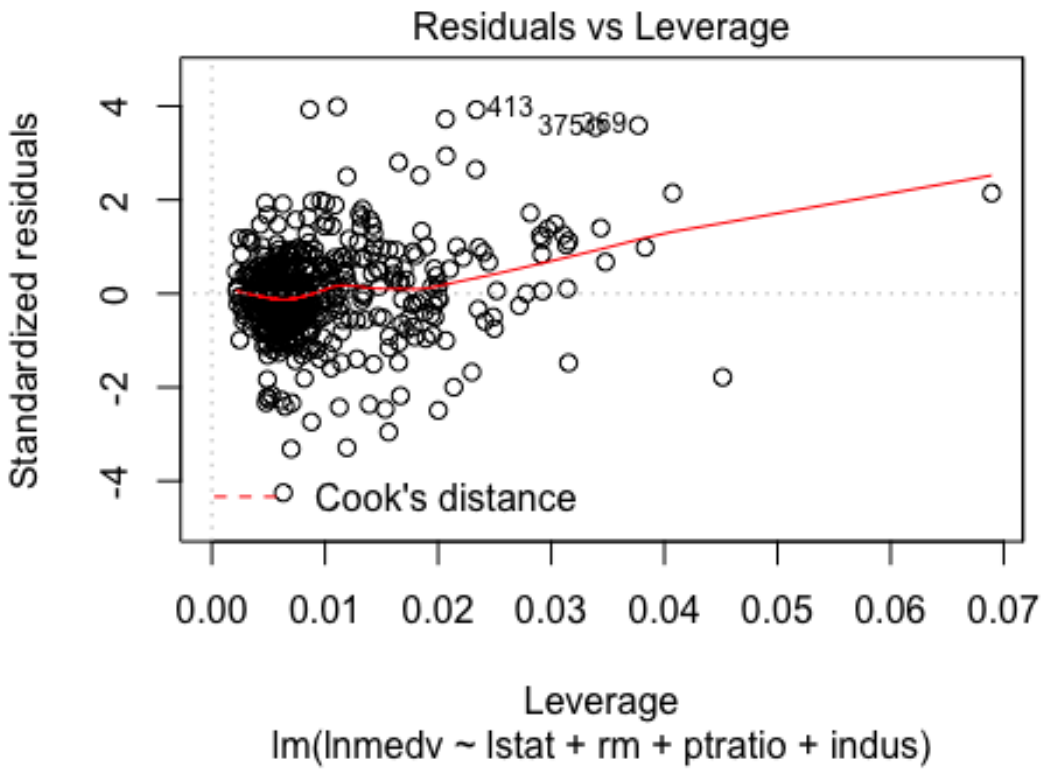
h)

```
resid2 <- as.data.frame(residuals(regression2))
plot(regression2)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
-0.5
-1.0
373
406
2.0
2.5
3.0
3.5
Fitted values
lm(lnmedv ~ lstat + rm + ptratio + indus)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
406
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(lnmedv ~ lstat + rm + ptratio + indus)


Scale-Location
406
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.0
2.5
3.0
3.5
Fitted values
lm(lnmedv ~ lstat + rm + ptratio + indus)



Yes, there is still evidence of heteroscedacity because of many residuals and non-constant variance of errors.

i)

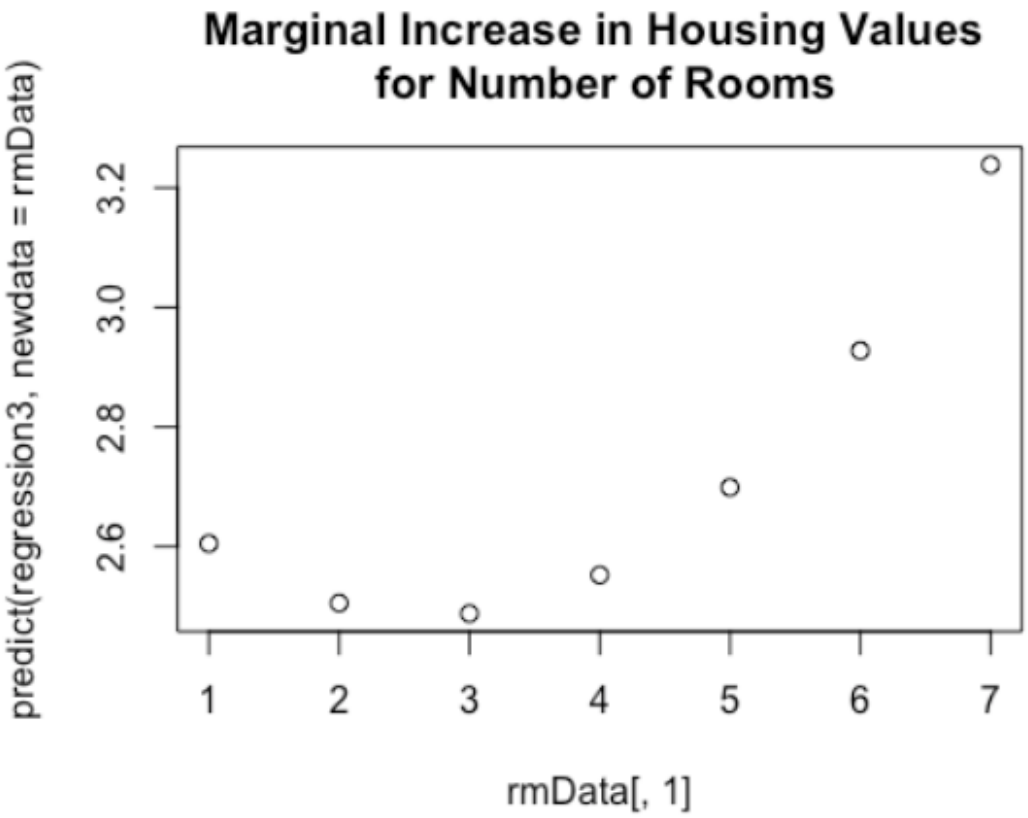
```
Boston$rmSq <- Boston$rm * Boston$rm
regression3 <- lm(lnmedv ~ rm + rmSq + ptratio, data=Boston)
summary(regression3)
```

```
##
## Call:
## lm(formula = lnmedv ~ rm + rmSq + ptratio, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1430 -0.1217  0.0590  0.1714  1.3125
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.855021   0.576644   6.685 6.15e-11 ***
```

```
## rm          -0.222718   0.176997  -1.258  0.20886
## rmSq         0.041036   0.013753   2.984  0.00299 **
## ptratio     -0.057533   0.006447  -8.924  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.291 on 502 degrees of freedom
## Multiple R-squared:  0.4962, Adjusted R-squared:  0.4932
## F-statistic: 164.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

j)

```
rmData <- data.frame(rm = 1:7, rmSq = 1:7 * 1:7, ptratio = rep(18.57,
 7))
plot(rmData[, 1], predict(regression3, newdata = rmData), main = "Marginal
Increase in Housing Values \n for Number of Rooms")
```



There is more of a steep marginal increase from 6 to 7 rooms than there is for 4 to 5 rooms. For example, for 6 to 7, the values increase from around 2.9 to 3.3. But for 4 to 5, the values only increase from around 2.5 to 2.7. Therefore the approximate impact of moving from 6 to 7 is 0.4 or about a 13.8% change and the approximate impact of moving from 4 to 5 is 0.2 or about a 8% change.