IMT 573: Problem Set 7

Regression

Ali Qazi

Due: August 01, 2021

Collaborators: AKeel Qazi Anthony Mercado

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Download the `07_ps_regression.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `07_ps_regression.Rmd`.
2. Replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps7_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup: In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
```

```
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Problem 1: Housing Values in Suburbs of Boston

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. This data is much used in data science and statistics to demonstrate regression problems; and while it has a lot of advantages it will comes with concerns. Load this data and use it to answer the following questions.

```
data(Boston)
?Boston
```

(a) Briefly describe where these data come from and why they were collected. Be sure to mention any concerns you have about these data. 506 rows and 14 columns make up the Boston data frame. It includes information on housing costs in Boston's suburbs. I do not have any concerns but am curious why there is a column called black.

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102.

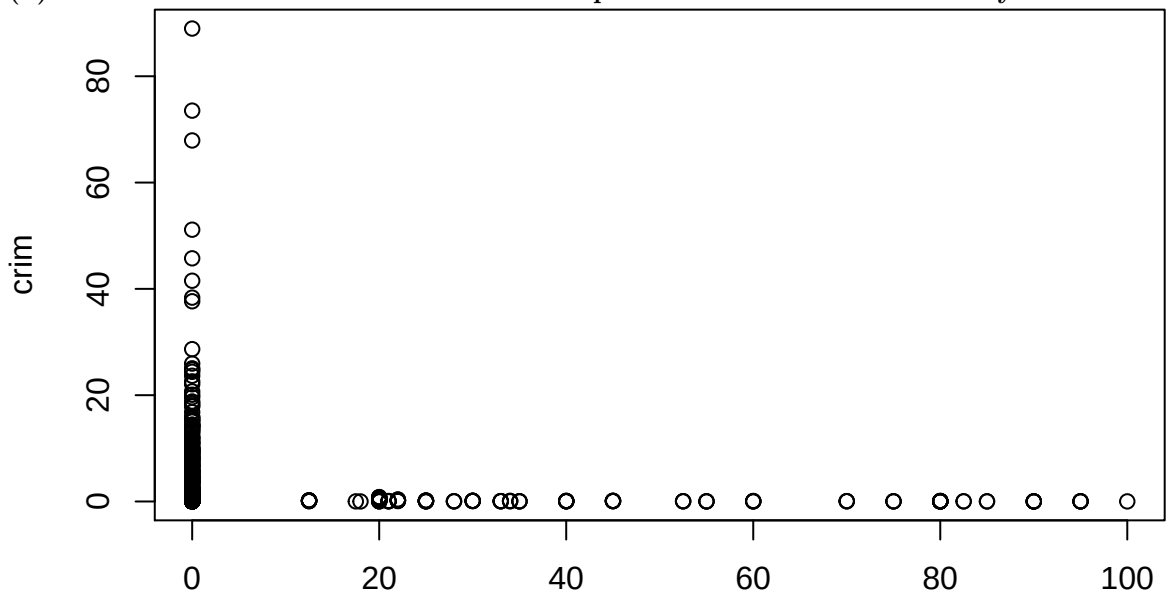
Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

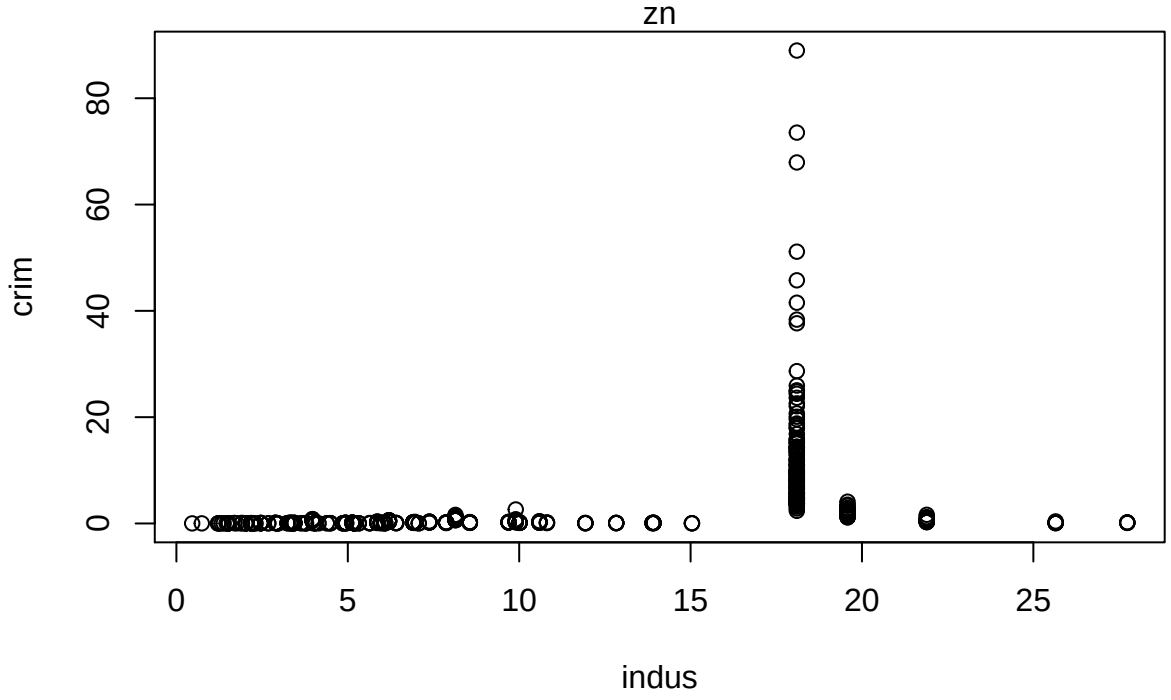
```
head(Boston)
```

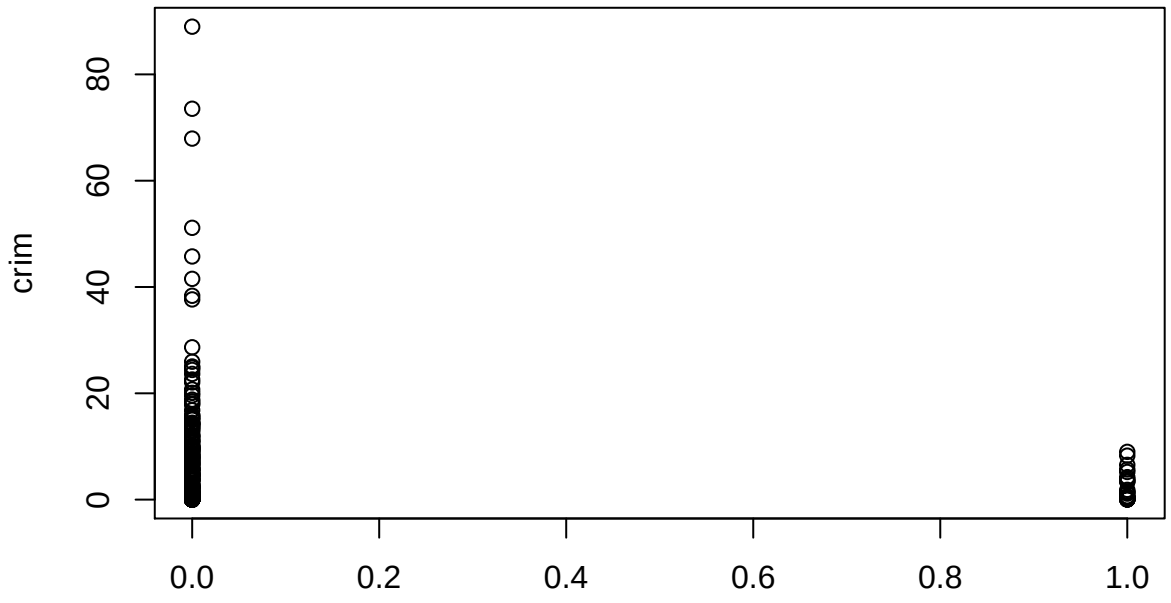
```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

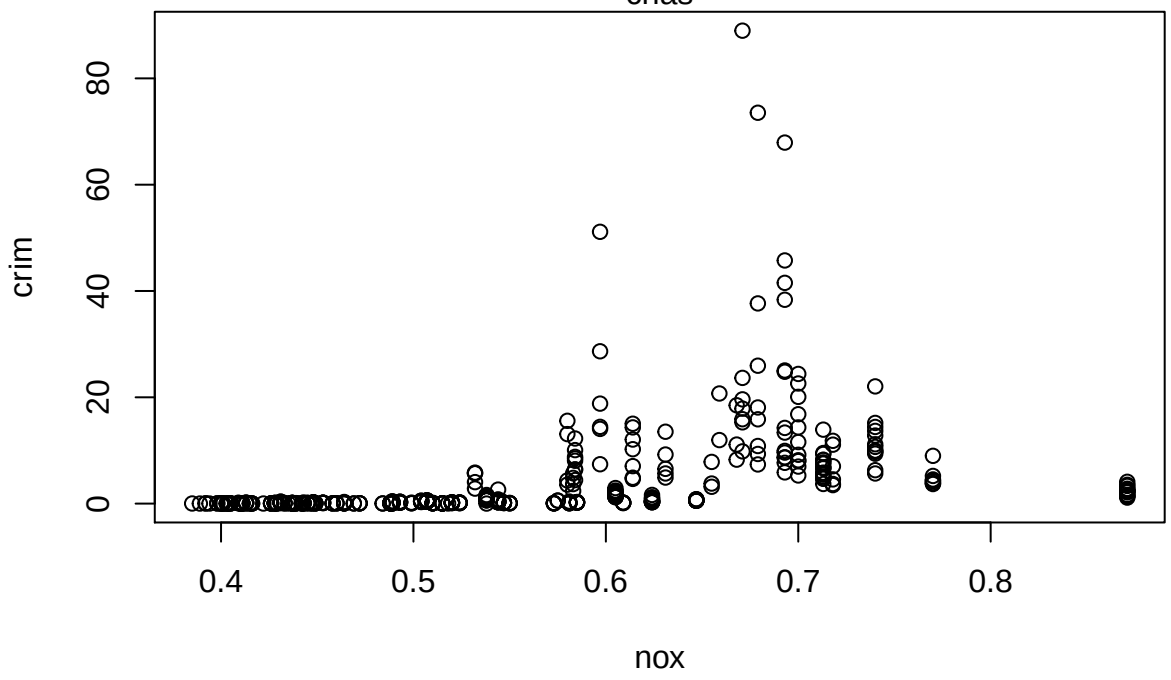
```
plot(crim ~ . - crim, data = Boston)
```

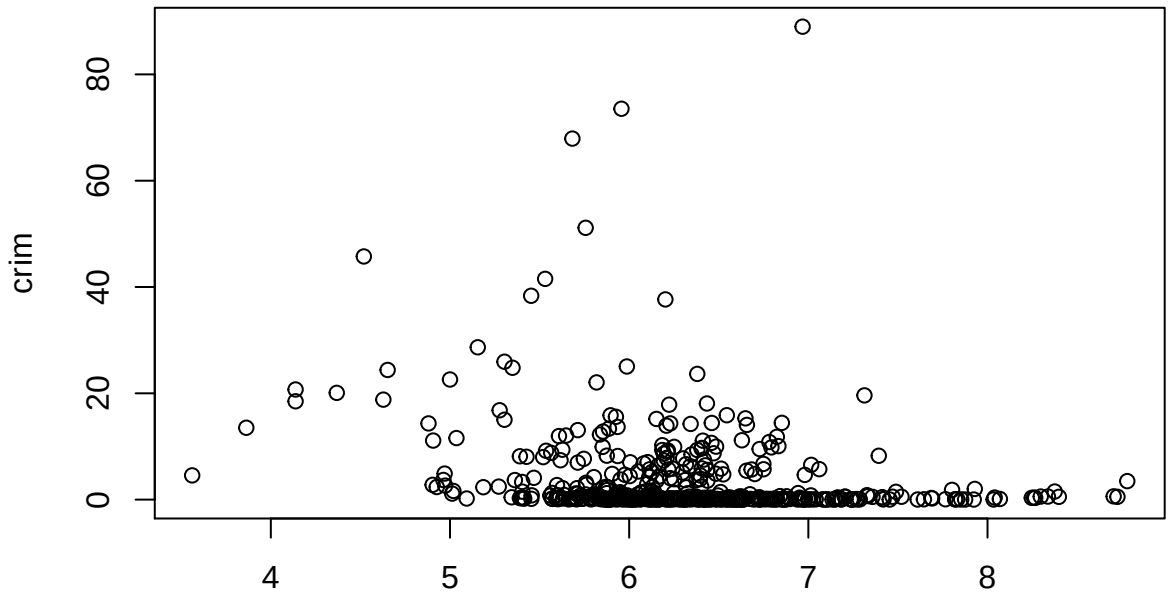
(b) Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.

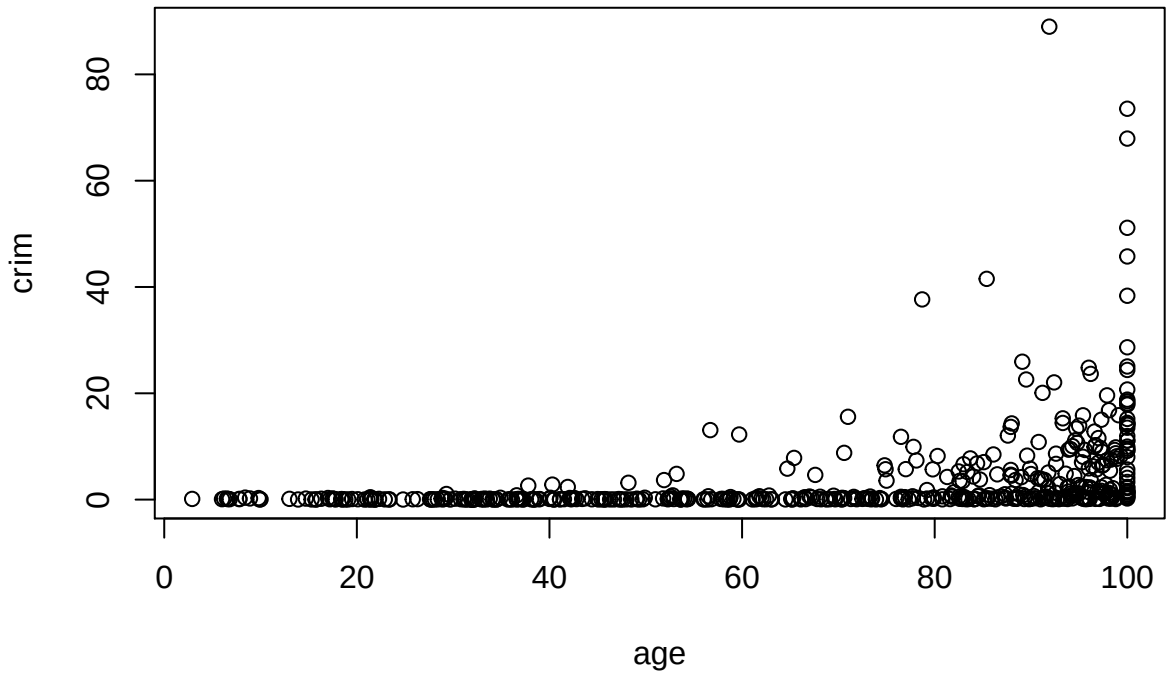


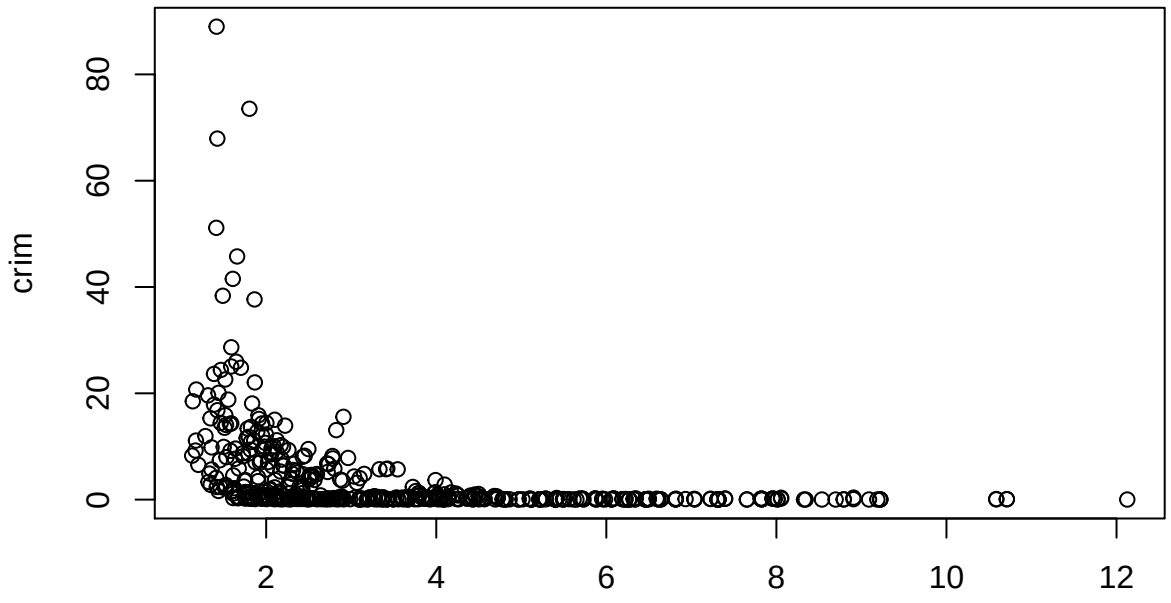


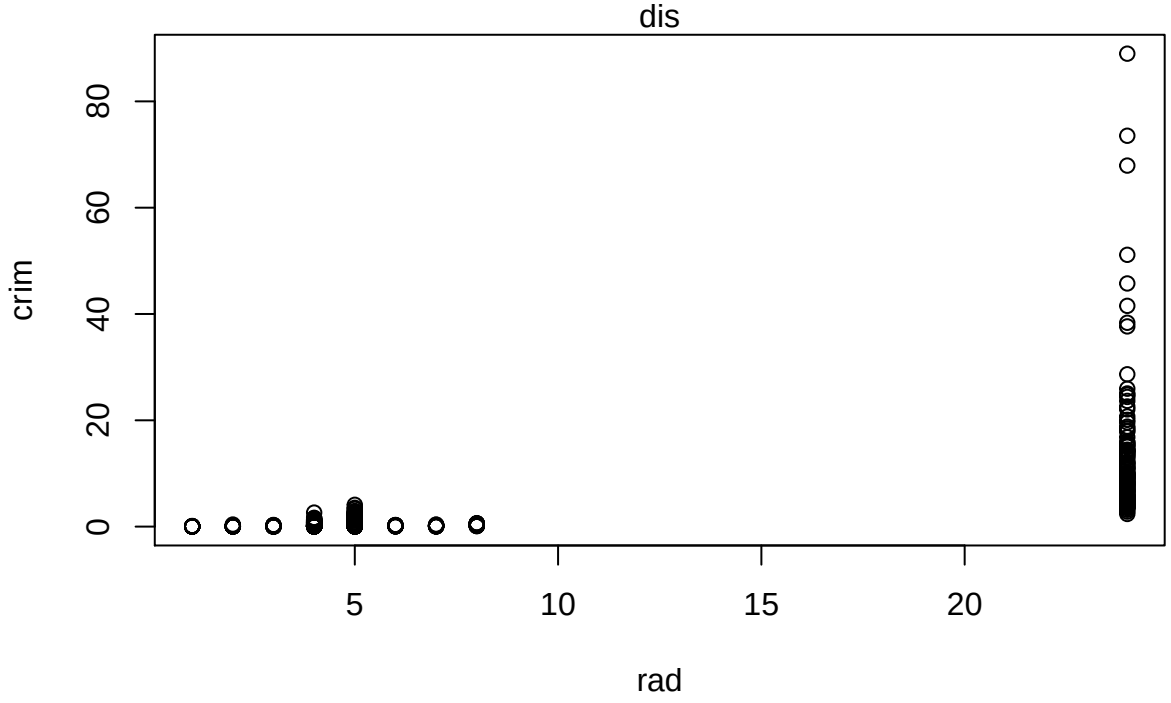

crim
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
chas

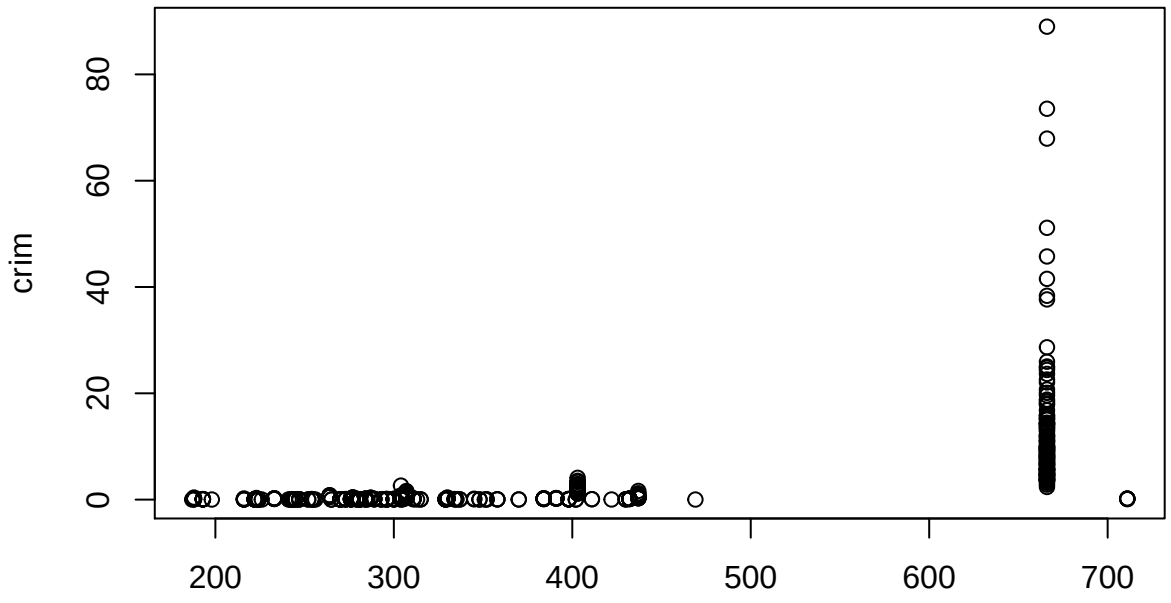

crim
0
20
40
60
80
0.4
0.5
0.6
0.7
0.8
nox

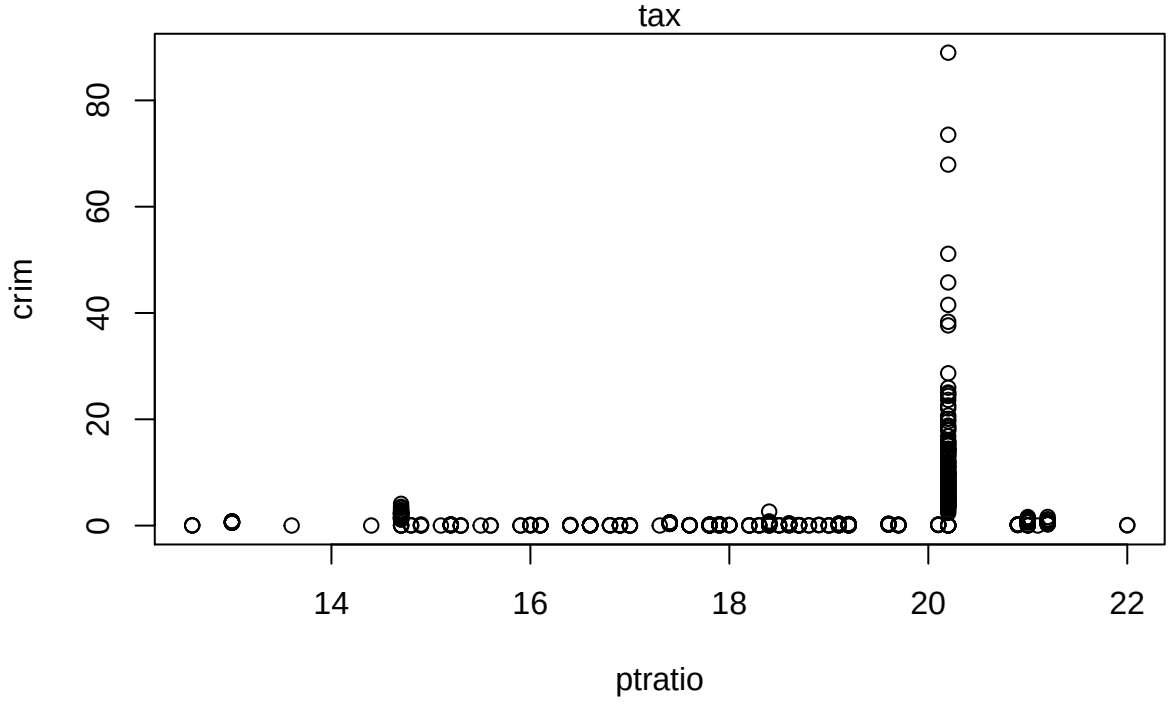

crim
0
20
40
60
80
4
5
6
7
8
rm

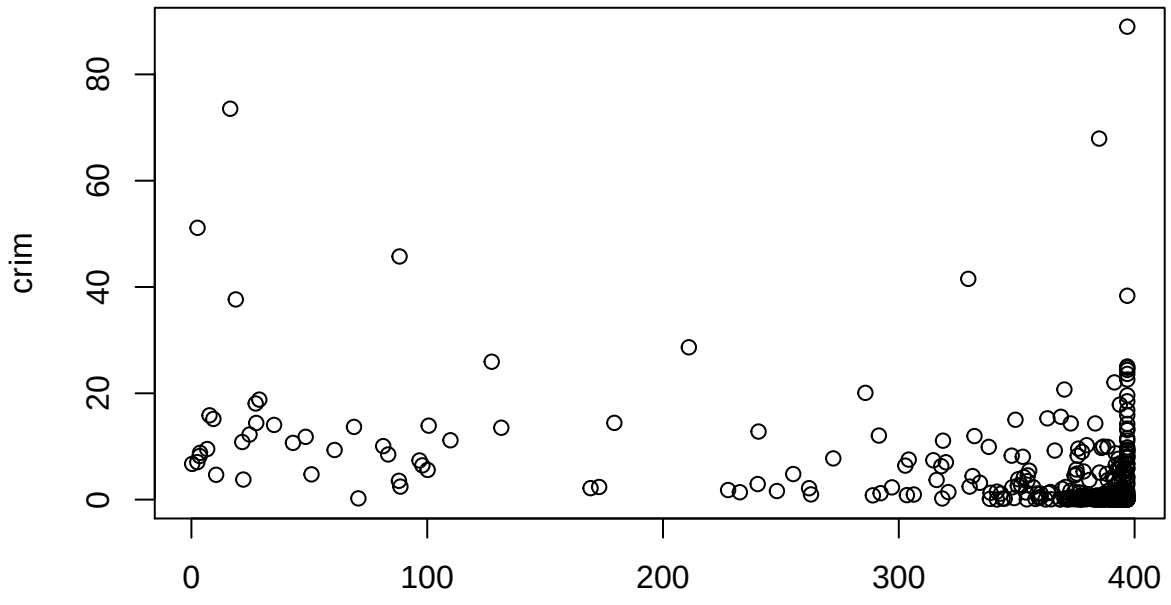

crim
0
20
40
60
80
0
20
40
60
80
100
age

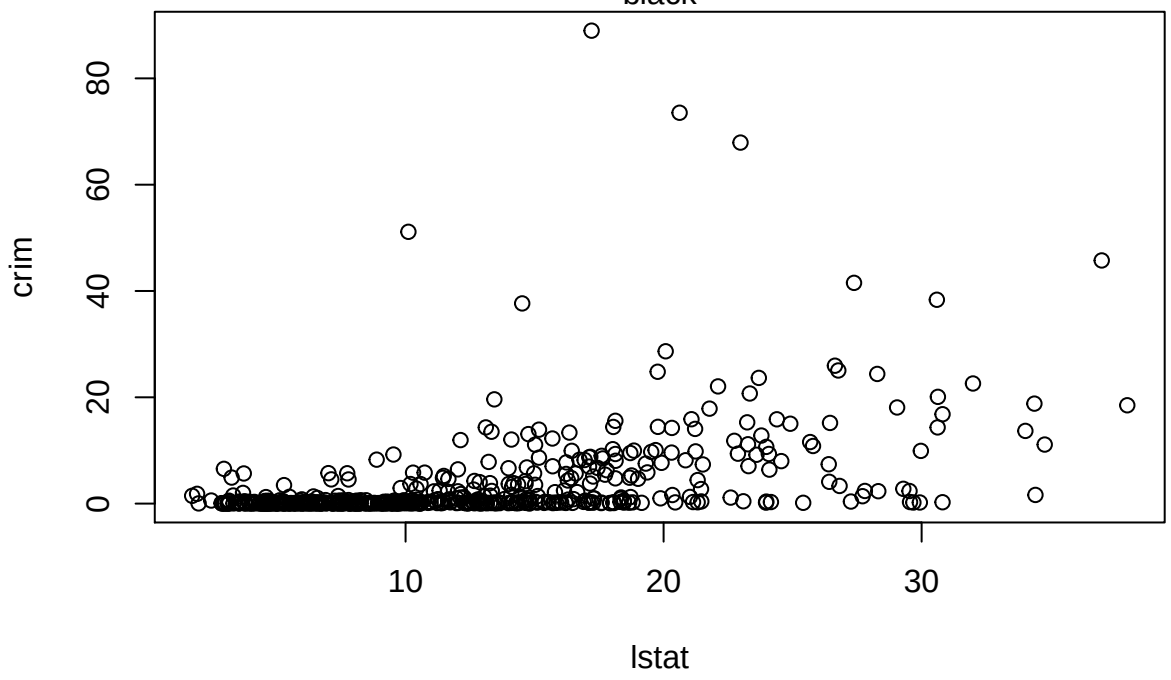

crim
0
20
40
60
80
2
4
6
8
10
12
dis

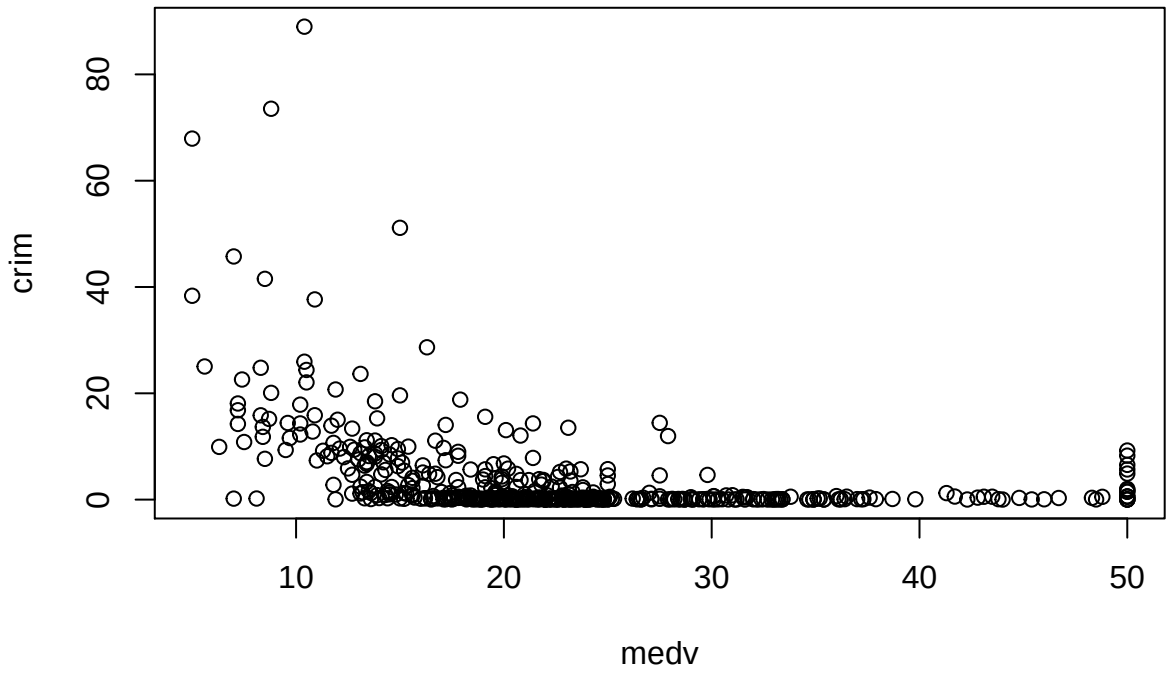

crim
0
20
40
60
80
5
10
15
20
rad


crim
0
20
40
60
80
200
300
400
500
600
700
tax


crim
0
20
40
60
80
14
16
18
20
22
ptratio


crim
0
20
40
60
80
0
100
200
300
400
black


crim
0
20
40
60
80
10
20
30
lstat



crim - per capita crime rate by town. zn - proportion of residential land zoned for lots over 25,000 sq.ft. indus - proportion of non-retail business acres per town. chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). nox - nitrogen oxides concentration (parts per 10 million). rm - average number of rooms per dwelling. age - proportion of owner-occupied units built prior to 1940. dis - weighted mean of distances to five Boston employment centres. rad - index of accessibility to radial highways. 2 tax - full-value property-tax rate per $10,000. ptratio - pupil-teacher ratio by town. black - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. lstat - lower status of the population (percent). medv - median value of owner-occupied homes in $1000s.

```
summary(lm(crim ~ zn, data = Boston))
```

(d) Consider this data in context, what is the response variable of interest?

```
##
## Call:
## lm(formula = crim ~ zn, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## zn          -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

Crime is the response variable.

```
summary(lm(crim ~ indus, data = Boston))
```

(e) For each predictor, fit a simple linear regression model to predict the response. 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.

```
##
## Call:
## lm(formula = crim ~ indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06374    0.66723  -3.093  0.00209 **
## indus        0.50978    0.05102   9.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ chas, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ chas, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7444     0.3961   9.453   <2e-16 ***
## chas         -1.8928     1.5061  -1.257    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

```
summary(lm(crim ~ nox, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ nox, data = Boston)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
## nox           31.249      2.999  10.419  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ nox, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ nox, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
## nox           31.249      2.999  10.419  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ rm, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ rm, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## rm            -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```
summary(lm(crim ~ age, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ age, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.789 -4.257 -1.230  1.527 82.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## age          0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```
summary(lm(crim ~ dis, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ dis, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.708 -4.134 -1.527  1.516 81.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4993     0.7304  13.006   <2e-16 ***
## dis          -1.5509     0.1683  -9.213   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ rad, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -10.164  -1.381  -0.141   0.660  76.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
## rad          0.61791    0.03433  17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.718 on 504 degrees of freedom
## Multiple R-squared:  0.3913, Adjusted R-squared:   0.39
## F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ tax, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.513  -2.738  -0.194   1.065  77.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
## tax          0.029742   0.001847   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ ptratio, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ ptratio, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
## ptratio       1.1520     0.1694   6.801 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```
summary(lm(crim ~ black, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ black, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.756  -2.299  -2.095  -1.296  86.822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.553529   1.425903  11.609   <2e-16 ***
## black       -0.036280   0.003873  -9.367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.946 on 504 degrees of freedom
## Multiple R-squared:  0.1483, Adjusted R-squared:  0.1466
## F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ lstat, data = Boston))
```

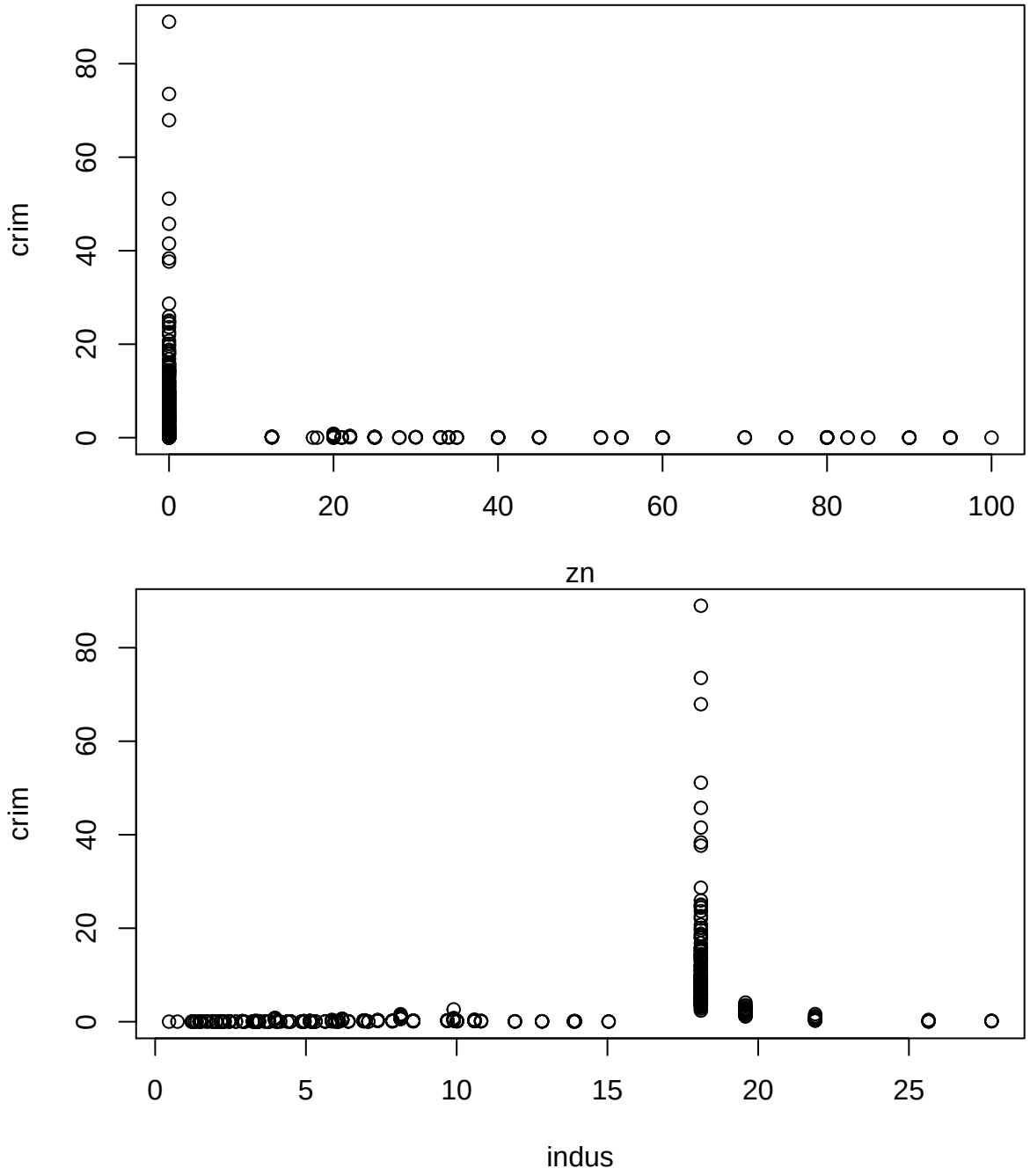
```
##
## Call:
## lm(formula = crim ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

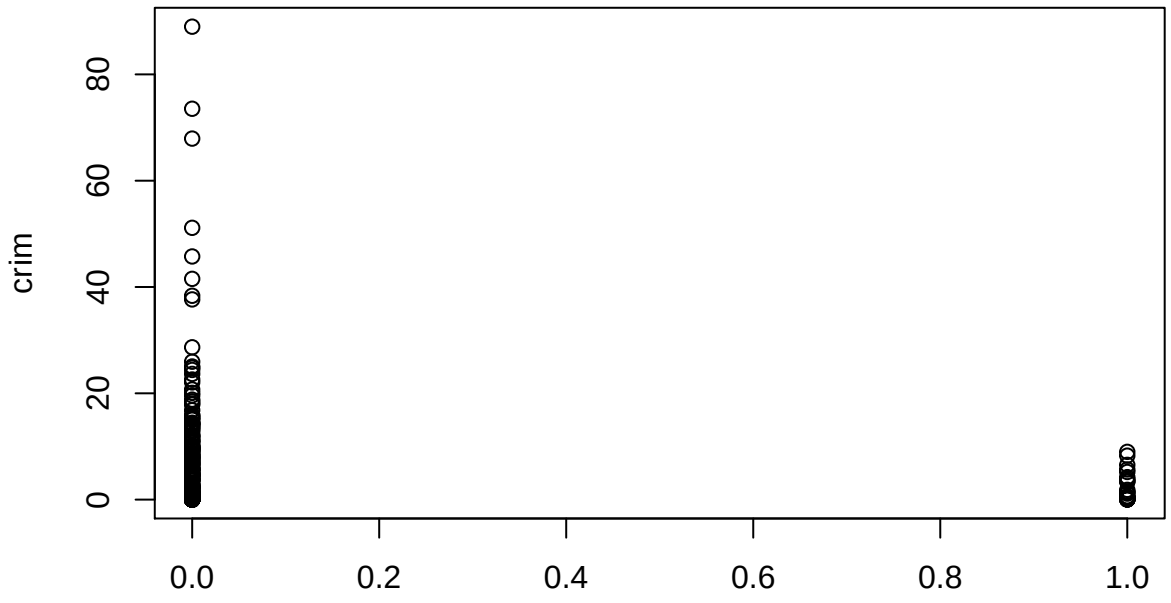
```
summary(lm(crim ~ medv, data = Boston))
```

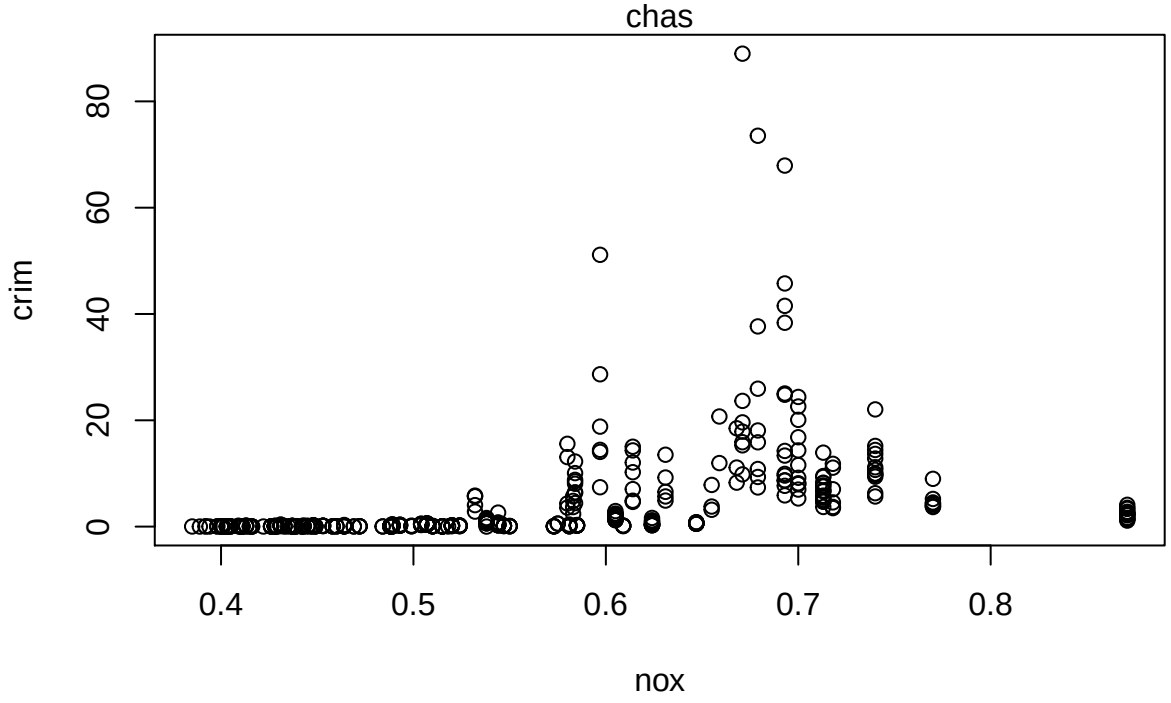
```
##
## Call:
## lm(formula = crim ~ medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
```

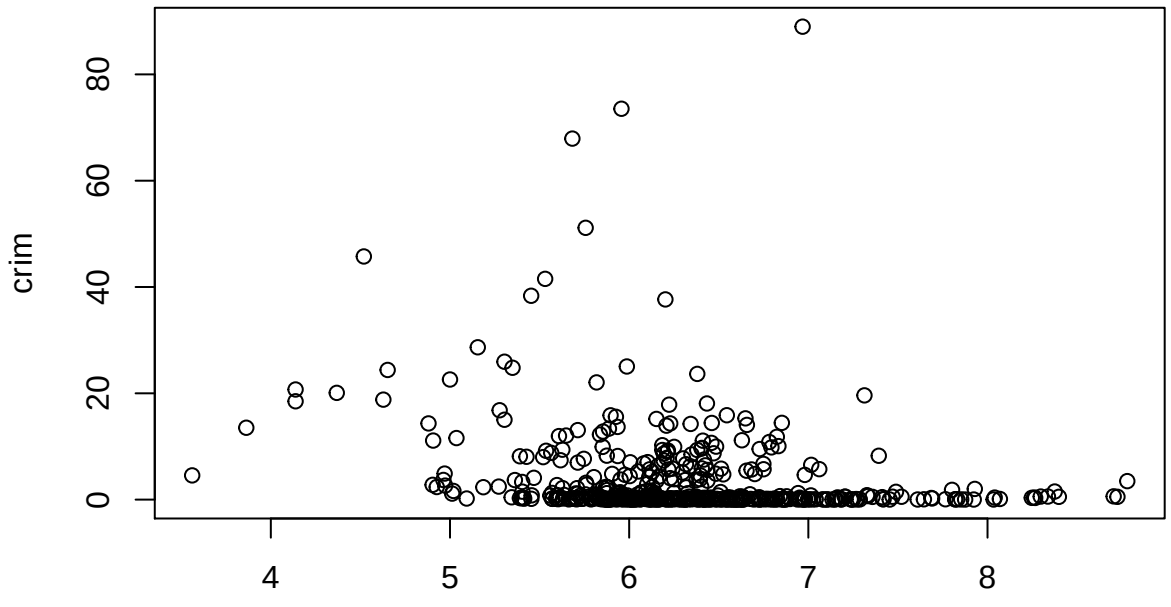
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

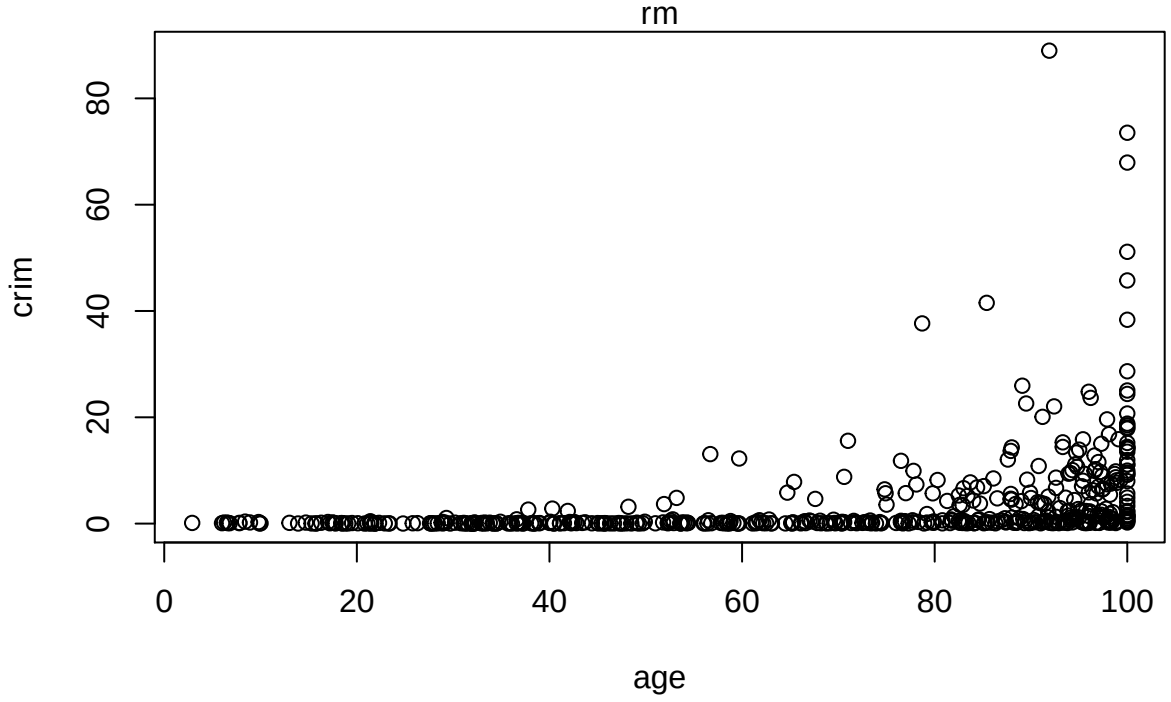
```
plot(crim ~ . - crim, data = Boston)
```

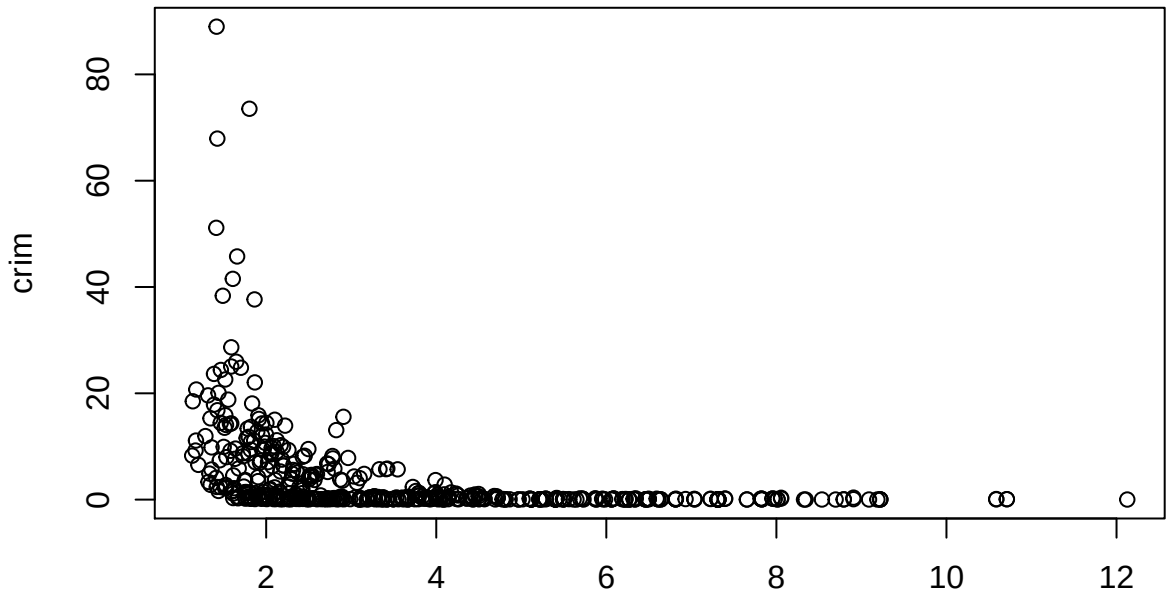


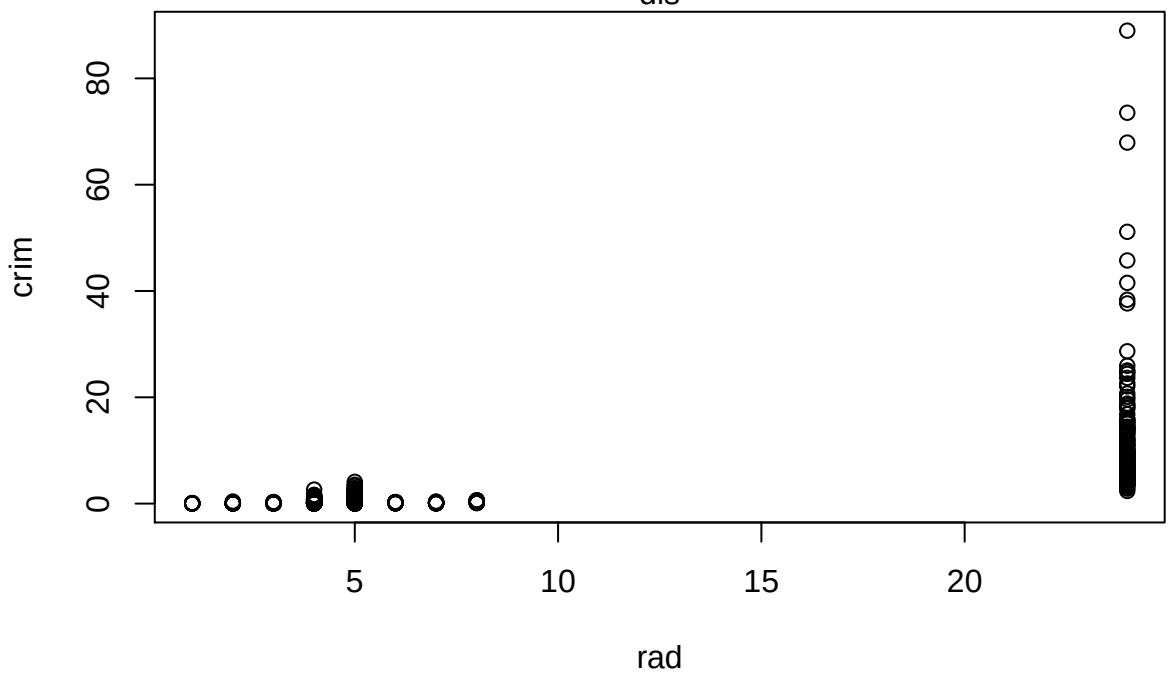

crim
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
chas

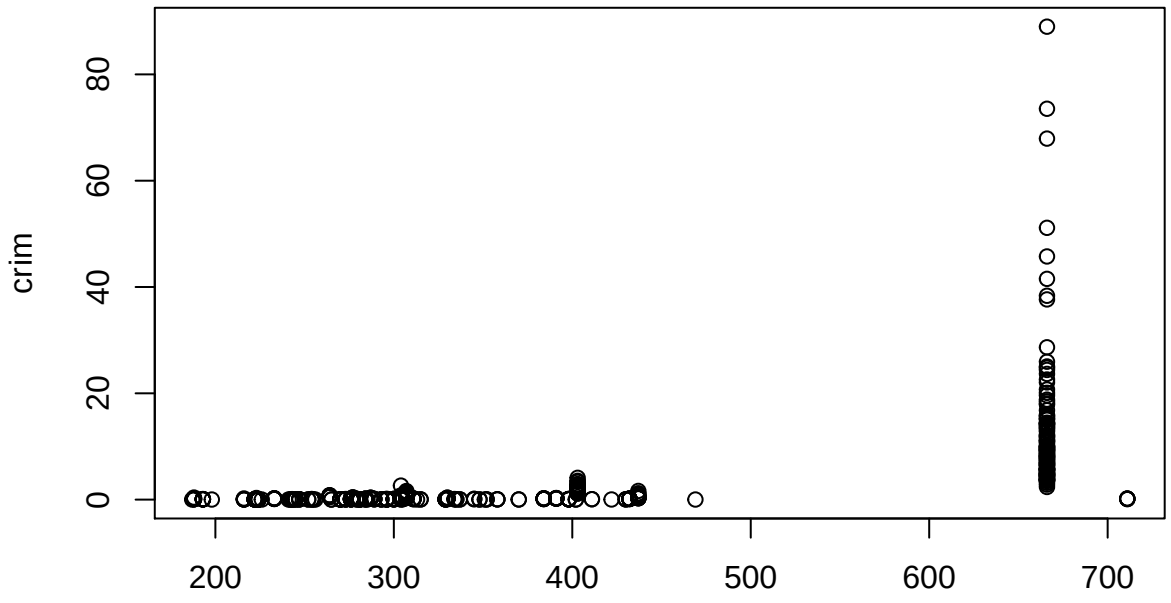

crim
0
20
40
60
80
0.4
0.5
0.6
0.7
0.8
nox

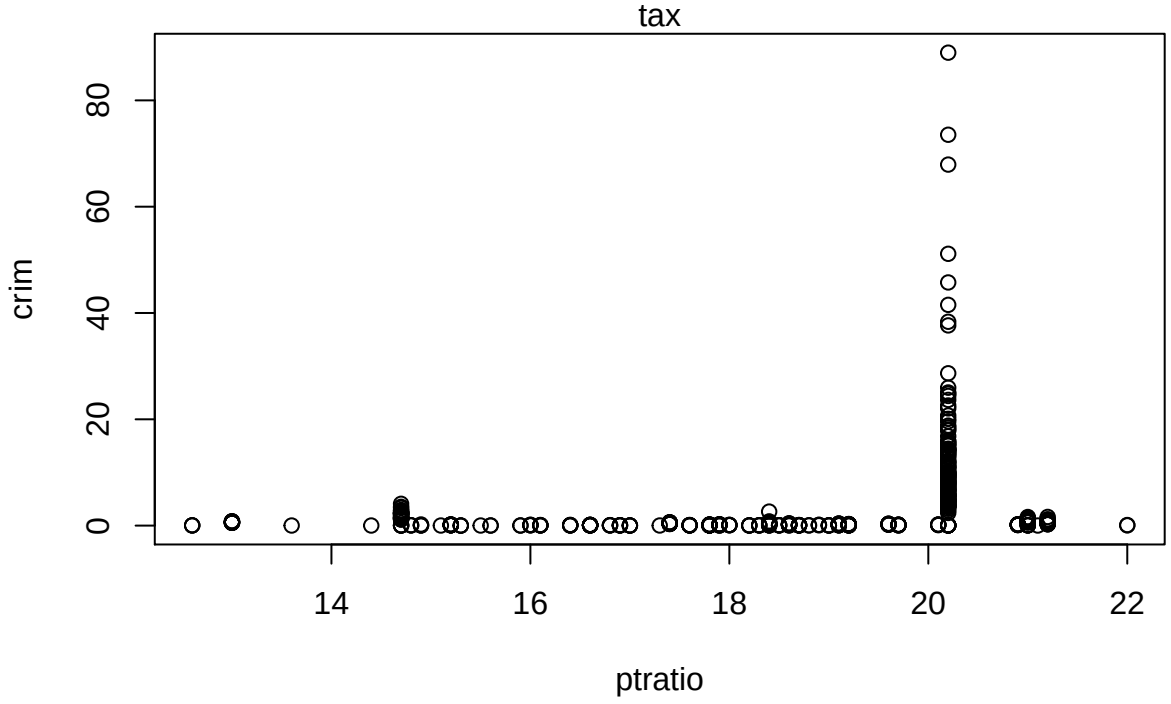

crim
0
20
40
60
80
4
5
6
7
8
rm

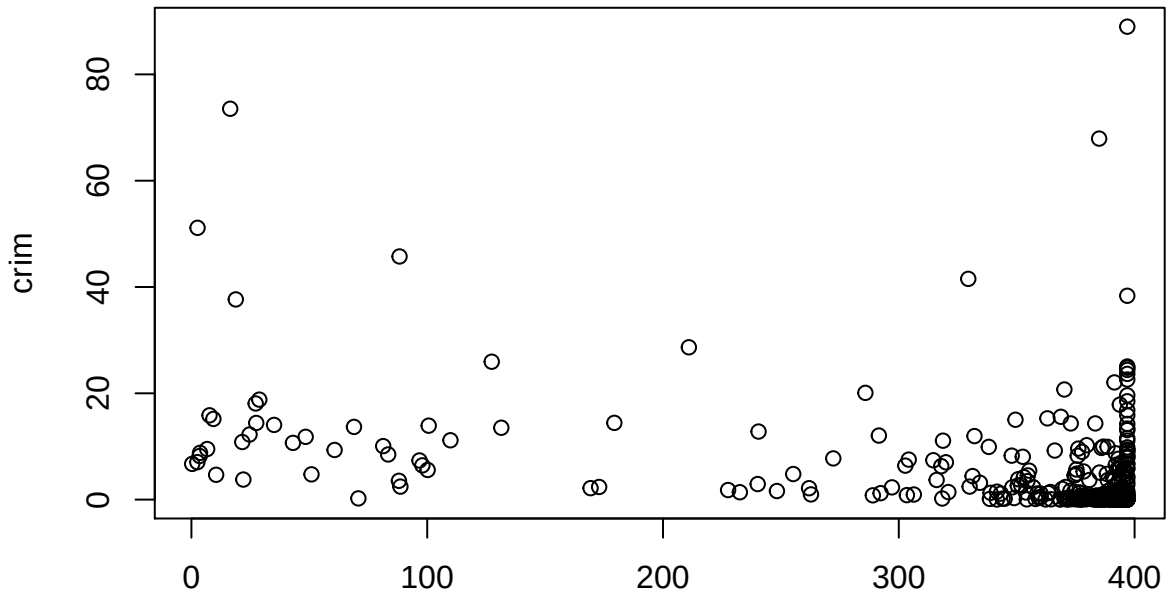

crim
0
20
40
60
80
0
20
40
60
80
100
age

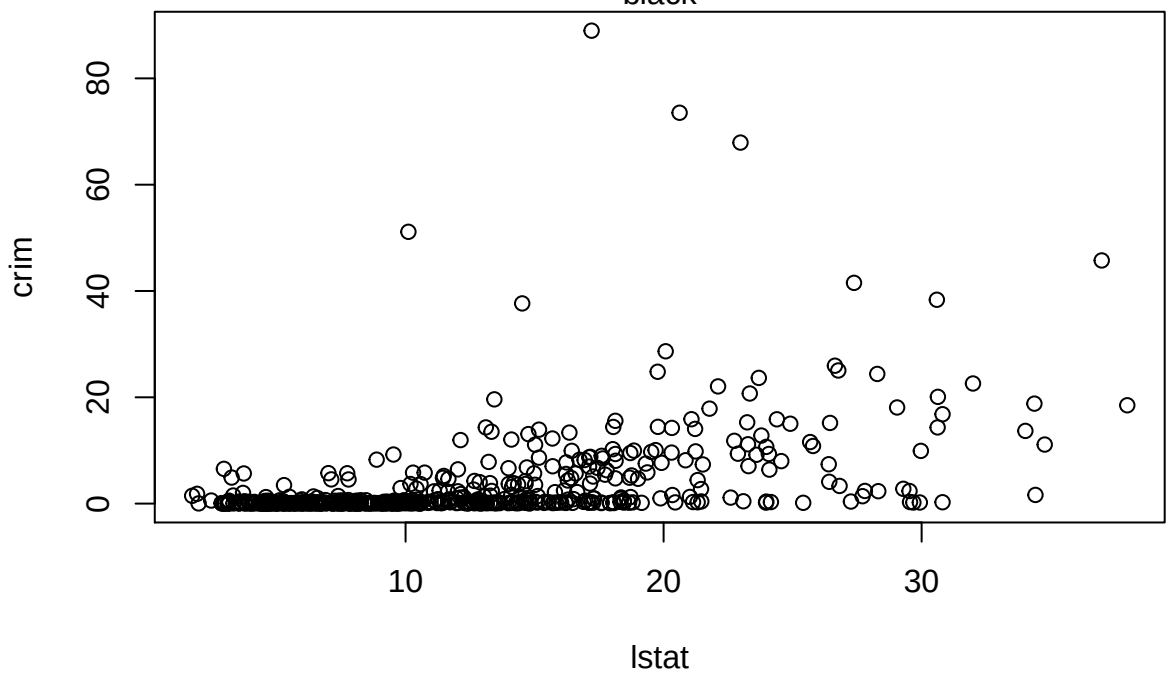

crim
0
20
40
60
80
2
4
6
8
10
12
dis

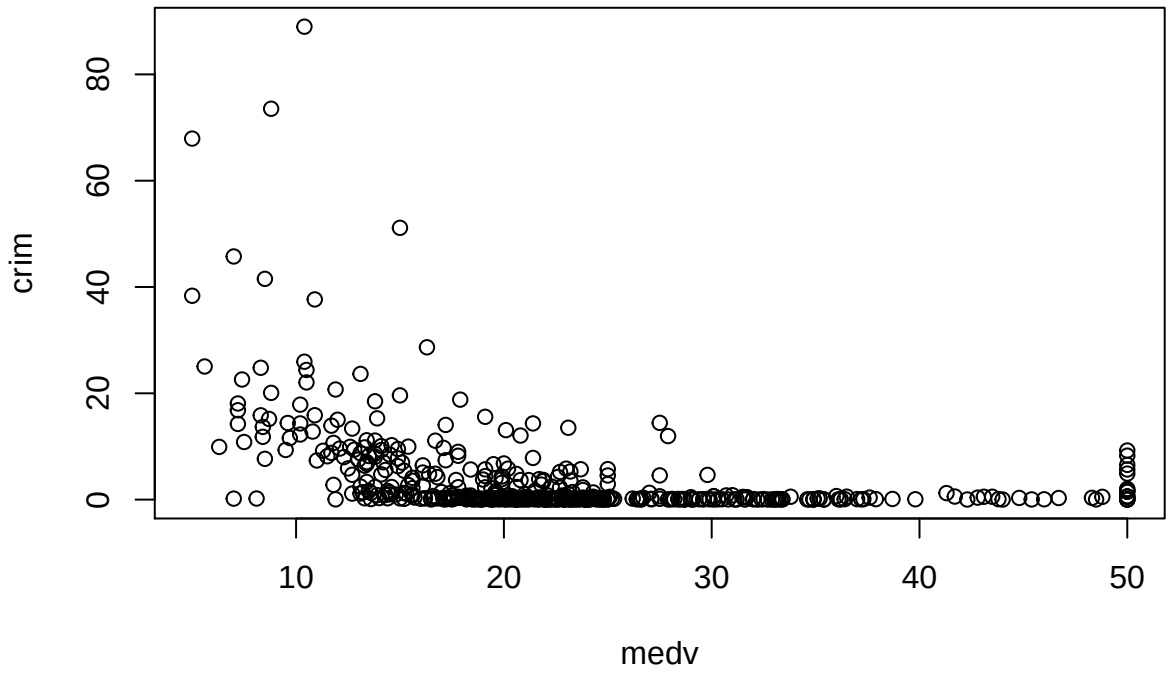

crim
0
20
40
60
80
5
10
15
20
rad


crim
0
20
40
60
80
200
300
400
500
600
700
tax


crim
0
20
40
60
80
14
16
18
20
22
ptratio


crim
0
20
40
60
80
0
100
200
300
400
black


crim
0
20
40
60
80
10
20
30
lstat




```
summary(lm(crim ~ . - crim, data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ . - 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
## chas         -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
```

Except for chas, there is a statistically significant relationship between the predictor and the response. R squared is small. In other words, the predictors describe a modest amount of variation.

```
univariate_coeff <- lm(crim ~ zn, data = Boston)$coefficients[2]
univariate_coeff <- append(univariate_coeff, lm(crim ~ indus, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ chas, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ nox, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ rm, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ age, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ dis, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ rad, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ tax, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ ptratio, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ black, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ lstat, data = Boston)$coefficients[2])
univariate_coeff <- append(univariate_coeff, lm(crim ~ medv, data = Boston)$coefficients[2])
booston <- (lm(crim ~ . - crim, data = Boston))
booston$coefficients[2:14]
```

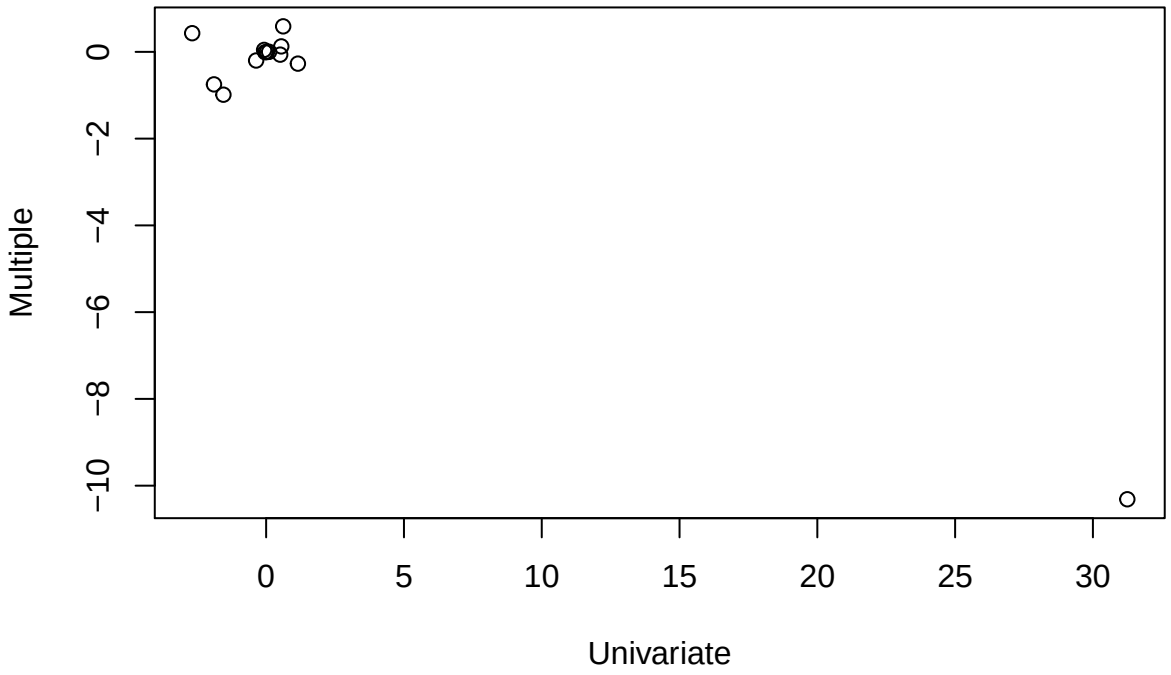
(f) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```
##            zn         indus          chas           nox            rm
##   0.044855215  -0.063854824  -0.749133611 -10.313534912   0.430130506
##           age           dis           rad           tax       ptratio
##   0.001451643  -0.987175726   0.588208591  -0.003780016  -0.271080558
##         black         lstat          medv
##  -0.007537505   0.126211376  -0.198886821
```

Only a small number of variables, such as dis, rad, medv, zn, and black, are significant in multiple regression models. The null hypothesis cannot be ruled out by the remaining variables. Because we are utilizing several regression models rather than the predictors alone, R squared is also significantly greater.

```
plot(univariate_coeff, booston$coefficients[2:14],
main = "Univariate Regression Coefficients vs. Multiple Regression Coefficients",
xlab = "Univariate", ylab = "Multiple")
```

(g) How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response.




```
summary(lm(crim ~ zn + I(zn^2) + I(zn^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ zn + I(zn^2) + I(zn^3), data = Boston)
##
## 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)  4.846e+00  4.330e-01  11.192  < 2e-16 ***
## zn          -3.322e-01  1.098e-01  -3.025  0.00261 **
## I(zn^2)      6.483e-03  3.861e-03   1.679  0.09375 .
## I(zn^3)     -3.776e-05  3.139e-05  -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
```

(h) Is there evidence of a 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:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

```
summary(lm(crim ~ chas + I(chas^2) + I(chas^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ chas + I(chas^2) + I(chas^3), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7444     0.3961   9.453   <2e-16 ***
## chas         -1.8928     1.5061  -1.257    0.209
## I(chas^2)         NA         NA      NA       NA
## I(chas^3)         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

```
summary(lm(crim ~ indus + I(indus^2) + I(indus^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ indus + I(indus^2) + I(indus^3), data = Boston)
##
## 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)  3.6625683  1.5739833   2.327   0.0204 *
## indus       -1.9652129  0.4819901  -4.077 5.30e-05 ***
## I(indus^2)   0.2519373  0.0393221   6.407 3.42e-10 ***
## I(indus^3)  -0.0069760  0.0009567  -7.292 1.20e-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
```

```
summary(lm(crim ~ nox + I(nox^2) + I(nox^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ nox + I(nox^2) + I(nox^3), data = Boston)
##
## 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)   233.09      33.64   6.928 1.31e-11 ***
## nox         -1279.37     170.40  -7.508 2.76e-13 ***
## I(nox^2)     2248.54     279.90   8.033 6.81e-15 ***
## I(nox^3)    -1245.70     149.28  -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
```

```
summary(lm(crim ~ rm + I(rm^2) + I(rm^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ rm + I(rm^2) + I(rm^3), data = Boston)
##
## 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) 112.6246    64.5172   1.746   0.0815 .
## rm          -39.1501    31.3115  -1.250   0.2118
## I(rm^2)       4.5509     5.0099   0.908   0.3641
## I(rm^3)      -0.1745     0.2637  -0.662   0.5086
## ---
## 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
```

```
summary(lm(crim ~ age + I(age^2) + I(age^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ age + I(age^2) + I(age^3), data = Boston)
##
## 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) -2.549e+00  2.769e+00  -0.920  0.35780
## age          2.737e-01  1.864e-01   1.468  0.14266
## I(age^2)    -7.230e-03  3.637e-03  -1.988  0.04738 *
```

```
## I(age^3)     5.745e-05  2.109e-05   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
```

```
summary(lm(crim ~ dis + I(dis^2) + I(dis^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ dis + I(dis^2) + I(dis^3), data = Boston)
##
## 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)  30.0476     2.4459  12.285  < 2e-16 ***
## dis         -15.5543     1.7360  -8.960  < 2e-16 ***
## I(dis^2)      2.4521     0.3464   7.078 4.94e-12 ***
## I(dis^3)     -0.1186     0.0204  -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
```

```
summary(lm(crim ~ rad + I(rad^2) + I(rad^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ rad + I(rad^2) + I(rad^3), data = Boston)
##
## 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) -0.605545   2.050108  -0.295    0.768
## rad          0.512736   1.043597   0.491    0.623
## I(rad^2)    -0.075177   0.148543  -0.506    0.613
## I(rad^3)     0.003209   0.004564   0.703    0.482
##
## 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
```

```
summary(lm(crim ~ tax + I(tax^2) + I(tax^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ tax + I(tax^2) + I(tax^3), data = Boston)
##
## 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)  1.918e+01  1.180e+01   1.626    0.105
## tax         -1.533e-01  9.568e-02  -1.602    0.110
## I(tax^2)     3.608e-04  2.425e-04   1.488    0.137
## I(tax^3)    -2.204e-07  1.889e-07  -1.167    0.244
##
## 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
```

```
summary(lm(crim ~ ptratio + I(ptratio^2) + I(ptratio^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ ptratio + I(ptratio^2) + I(ptratio^3), data = Boston)
##
## 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)  477.18405  156.79498   3.043  0.00246 **
## ptratio      -82.36054   27.64394  -2.979  0.00303 **
## I(ptratio^2)   4.63535    1.60832   2.882  0.00412 **
## I(ptratio^3)  -0.08476    0.03090  -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
```

```
summary(lm(crim ~ black + I(black^2) + I(black^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ black + I(black^2) + I(black^3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.096  -2.343  -2.128  -1.439  86.790
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.826e+01  2.305e+00   7.924  1.5e-14 ***
## black       -8.356e-02  5.633e-02  -1.483    0.139
## I(black^2)   2.137e-04  2.984e-04   0.716    0.474
## I(black^3)  -2.652e-07  4.364e-07  -0.608    0.544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.955 on 502 degrees of freedom
## Multiple R-squared:  0.1498, Adjusted R-squared:  0.1448
## F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
summary(lm(crim ~ lstat + I(lstat^2) + I(lstat^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ lstat + I(lstat^2) + I(lstat^3), data = Boston)
##
## 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)  1.2009656  2.0286452   0.592   0.5541
## lstat       -0.4490656  0.4648911  -0.966   0.3345
## I(lstat^2)   0.0557794  0.0301156   1.852   0.0646 .
## I(lstat^3)  -0.0008574  0.0005652  -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
```

```
summary(lm(crim ~ medv + I(medv^2) + I(medv^3), data = Boston))
```

```
##
## Call:
## lm(formula = crim ~ medv + I(medv^2) + I(medv^3), data = Boston)
##
## 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) 53.1655381  3.3563105  15.840  < 2e-16 ***
## medv        -5.0948305  0.4338321 -11.744  < 2e-16 ***
## I(medv^2)    0.1554965  0.0171904   9.046  < 2e-16 ***
## I(medv^3)   -0.0014901  0.0002038  -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
```

```
lmMod <- lm(crim ~ . , data = Boston)
selectedMod <- step(lmMod)
```

```
## Start:  AIC=1898.56
## crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv
##
##           Df Sum of Sq   RSS    AIC
## - age      1      0.27 20400 1896.6
## - chas     1     16.71 20417 1897.0
## - rm       1     20.43 20420 1897.1
## - tax      1     22.29 20422 1897.1
## - indus    1     24.30 20424 1897.2
## <none>                 20400 1898.6
## - ptratio  1     87.65 20488 1898.7
## - lstat    1    115.18 20515 1899.4
## - nox      1    158.47 20558 1900.5
## - black    1    174.58 20574 1900.9
## - zn       1    237.70 20638 1902.4
## - medv     1    447.85 20848 1907.5
## - dis      1    508.77 20909 1909.0
## - rad      1   1850.44 22250 1940.5
##
## Step:  AIC=1896.56
## crim ~ zn + indus + chas + nox + rm + dis + rad + tax + ptratio +
##     black + lstat + medv
##
##           Df Sum of Sq   RSS    AIC
## - chas     1     16.54 20417 1895.0
## - rm       1     22.14 20422 1895.1
## - tax      1     22.16 20422 1895.1
## - indus    1     24.30 20424 1895.2
## <none>                 20400 1896.6
## - ptratio  1     87.41 20488 1896.7
## - lstat    1    131.43 20532 1897.8
## - nox      1    166.37 20567 1898.7
## - black    1    174.40 20575 1898.9
## - zn       1    239.21 20639 1900.5
## - medv     1    447.81 20848 1905.5
## - dis      1    559.06 20959 1908.2
## - rad      1   1857.98 22258 1938.7
##
## Step:  AIC=1894.97
## crim ~ zn + indus + nox + rm + dis + rad + tax + ptratio + black +
##     lstat + medv
##
##           Df Sum of Sq   RSS    AIC
## - tax      1     18.81 20436 1893.4
```

```
## - rm       1     22.76 20440 1893.5
## - indus    1     28.82 20446 1893.7
## <none>                 20417 1895.0
## - ptratio  1     84.57 20501 1895.1
## - lstat    1    129.63 20546 1896.2
## - nox      1    175.96 20593 1897.3
## - black    1    178.37 20595 1897.4
## - zn       1    241.26 20658 1898.9
## - medv     1    483.38 20900 1904.8
## - dis      1    563.37 20980 1906.8
## - rad      1   1842.82 22260 1936.7
##
## Step:  AIC=1893.44
## crim ~ zn + indus + nox + rm + dis + rad + ptratio + black +
##     lstat + medv
##
##           Df Sum of Sq   RSS    AIC
## - rm       1      23.0 20459 1892.0
## - indus    1      64.4 20500 1893.0
## <none>                 20436 1893.4
## - ptratio  1      87.4 20523 1893.6
## - lstat    1     137.9 20574 1894.8
## - black    1     178.1 20614 1895.8
## - nox      1     181.9 20617 1895.9
## - zn       1     222.9 20658 1896.9
## - medv     1     465.3 20901 1902.8
## - dis      1     556.9 20992 1905.0
## - rad      1    4693.4 25129 1996.0
##
## Step:  AIC=1892.01
## crim ~ zn + indus + nox + dis + rad + ptratio + black + lstat +
##     medv
##
##           Df Sum of Sq   RSS    AIC
## - indus    1      74.0 20533 1891.8
## <none>                 20459 1892.0
## - ptratio  1      88.2 20547 1892.2
## - lstat    1     118.9 20577 1892.9
## - nox      1     176.9 20636 1894.4
## - black    1     202.4 20661 1895.0
## - zn       1     233.9 20692 1895.8
## - medv     1     458.7 20917 1901.2
## - dis      1     572.2 21031 1904.0
## - rad      1    4811.3 25270 1996.9
##
## Step:  AIC=1891.83
## crim ~ zn + nox + dis + rad + ptratio + black + lstat + medv
##
##           Df Sum of Sq   RSS    AIC
## <none>                 20533 1891.8
## - lstat    1     104.7 20637 1892.4
## - ptratio  1     119.0 20652 1892.8
## - black    1     198.4 20731 1894.7
## - zn       1     239.6 20772 1895.7
```

```
## - nox      1     296.6 20829 1897.1
## - medv     1     430.2 20963 1900.3
## - dis      1     507.8 21040 1902.2
## - rad      1    4739.5 25272 1994.9
```

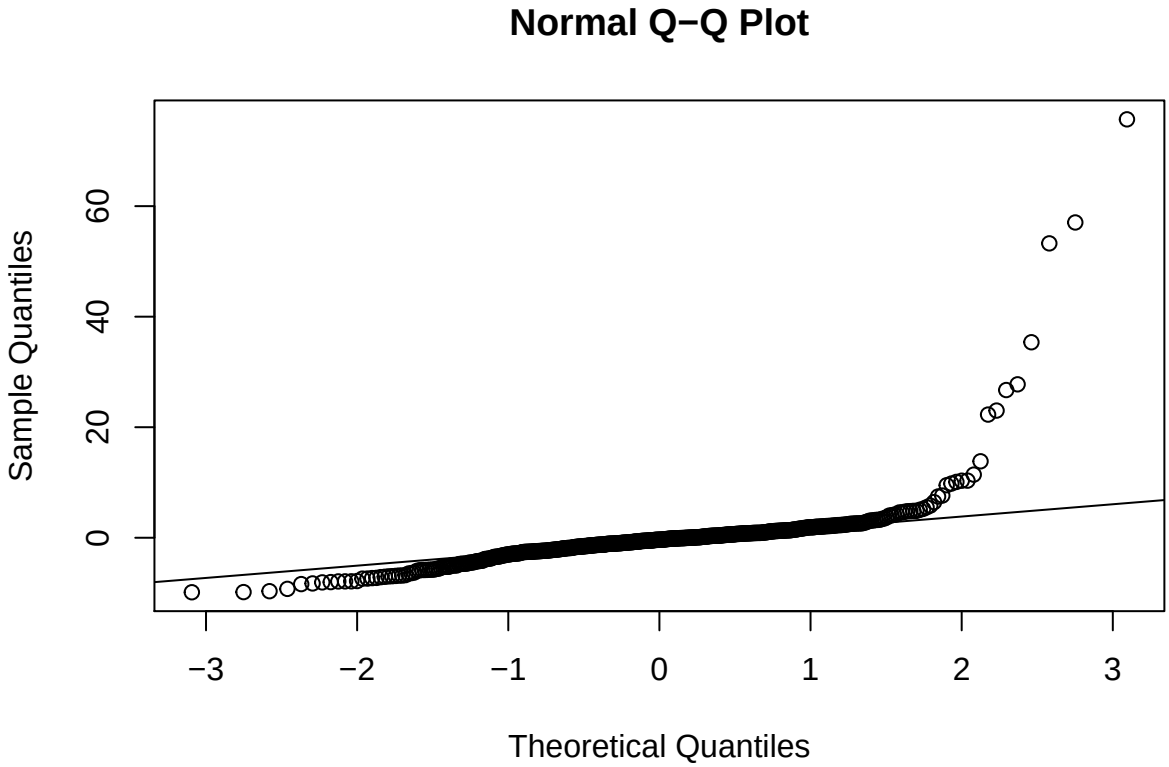
With indus, nox, dis, ptratio, age, and medv, there is proof of a non-linear connection between any of the predictors and the result. These variables are squared and rounded, making them statistically significant.

```
summary(selectedMod)
```

(i) Consider performing a stepwise model selection procedure to determine the bets fit model. Discuss your results. How is this model different from the model in (4)?

```
##
## Call:
## lm(formula = crim ~ zn + nox + dis + rad + ptratio + black +
##     lstat + medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.860 -2.102 -0.363  0.895 75.702
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.683128   6.086010   3.234 0.001301 **
## zn            0.043293   0.017977   2.408 0.016394 *
## nox         -12.753708   4.760157  -2.679 0.007623 **
## dis          -0.918318   0.261932  -3.506 0.000496 ***
## rad           0.532617   0.049727  10.711  < 2e-16 ***
## ptratio      -0.310541   0.182941  -1.697 0.090229 .
## black        -0.007922   0.003615  -2.191 0.028897 *
## lstat         0.110173   0.069219   1.592 0.112097
## medv         -0.174207   0.053988  -3.227 0.001334 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.428 on 497 degrees of freedom
## Multiple R-squared:  0.4505, Adjusted R-squared:  0.4416
## F-statistic: 50.92 on 8 and 497 DF,  p-value: < 2.2e-16
```

```
x <- resid(selectedMod)
qqnorm(x)
qqline(x)
```



(j) Evaluate the statistical assumptions in your regression analysis from (7) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

I know that a residual is the difference between the value that was seen and the value that the regression model predicted. When looking at the qq plot for the residuals, I seee that the data, with the exception of the dots at the top, follows the trend line.

Problem 2: A Critical Perspective to the Boston Housing Data

(a) When were these data collected? Did you note this in you descriptive above? Did the date surprise you? Data collection took place between 1978 and 1980. When I cited my source in my description, I did make note of this. Although the date surprised me because it was more than 40 years ago, it made more sense when I saw the dark column. I now wonder how the data was gathered and what techniques were employed.

(b) Amidst data features like number of rooms and access to highways are features like crime rate, and percentage Black per town. Whether intentional or not, someone looking at this data might infer a link between crime and race just due to the variables present; or even worse might use the data to support harsh policing policies based on race. Suppose for a moment we have a modern version of this dataset; the "Seattle Housing Data." Discuss, in a few paragraphs, how this hypothetical dataset could be used (1) in a harmful way, and (2) in a beneficial way for society. If the data and columns are comparable to the Boston housing data set, this hypothetical dataset might be misused. The information might be utilized to promote strict policing measures, as was mentioned in the question above. This would disadvantage some localities and property owners. By identifying elements like color and black individuals and connecting them to local criminality, this hypothetical dataset may likewise be used as a weapon.

By revealing additional elements that are influencing the value of residential property, the dataset may be useful to society. Building more educational facilities would be the solution, for instance, if the lack of educational facilities in the region had a negative impact on the value of residential property.Renters in Seattle, the 14th most expensive city to live in, would greatly benefit from rentals being priced at a price

that is fair and can match the cities average income.