Assignment 5

Chiu-Yuan Wu

10/29/2020

```
library(tidyverse)
library(tidyr)
library(dplyr)
library(tsibble)
library(ggplot2)
library(ggpubr)
```

1) This question involves the use of multiple linear regression on the Auto data set from the course webpage (https://scads.eecs.wsu.edu/index.php/datasets/). Ensure that you remove missing values from the dataframe, and that values are represented in the appropriate types.

```
auto <- read.csv("Auto.csv", na.strings = "?")
auto <- na.omit(auto)
```

a. (5%) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Show a printout of the result (including coefficient, error and t values for each predictor). Comment on the output:

```
auto.fit = lm(mpg~.-name, data=auto)
summary(auto.fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
```

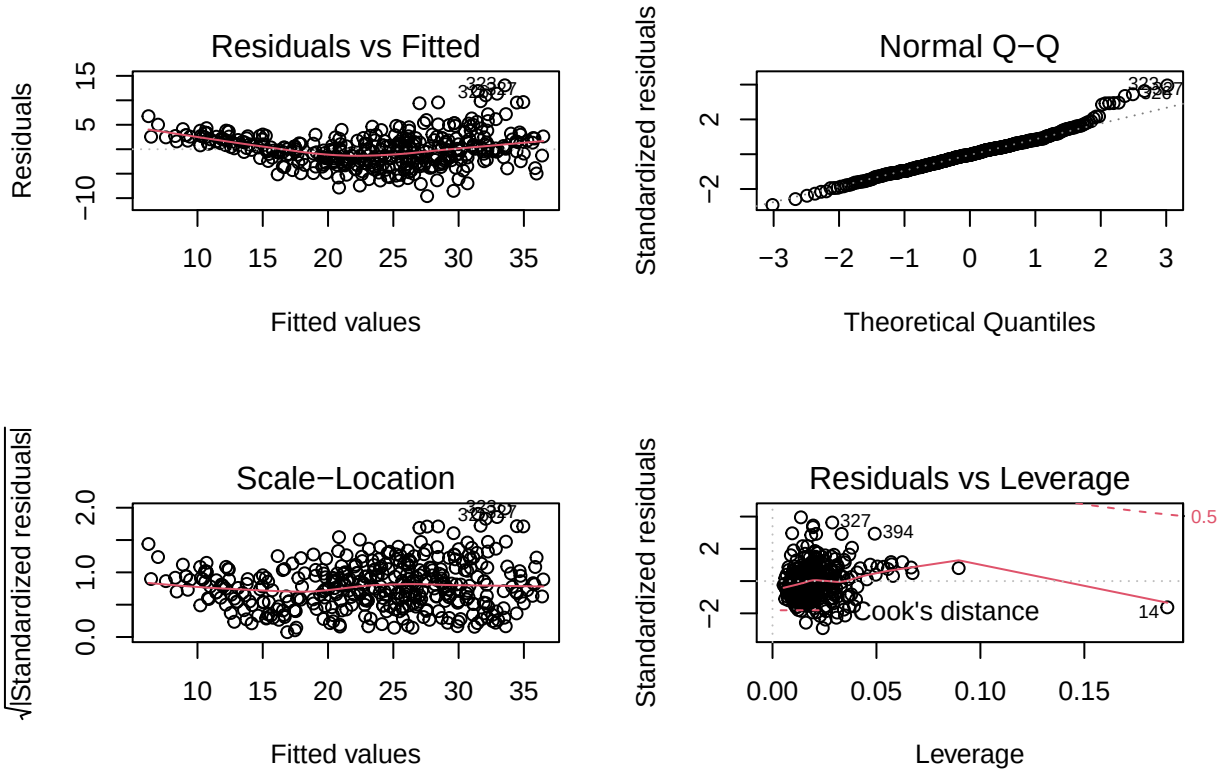
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i) Which predictors appear to have a statistically significant relationship to the response, and how do you determine this?
Displacement, weight, and year. They have low p-values.

ii) What does the coefficient for the displacement variable suggest, in simple terms?
The standard error is low so the confidence intervals are quite narrow. And p-value is also quite low, so there is only a tiny chance a value will go over |2.647|.

b. (5%) Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
par(mfrow=c(2,2))
plot(auto.fit)
```



In Residuals vs Leverage plot the fit doesn't work well, since most values are cluttered together.

c. (5%) Fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
summary(lm(mpg ~ weight * displacement + (weight * cylinders) + cylinders * displacement,
    data = auto))
```

```
##
## Call:
## lm(formula = mpg ~ weight * displacement + (weight * cylinders) +
##     cylinders * displacement, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1599  -2.5204  -0.3546   1.7851  17.8829
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.903e+01  6.743e+00   7.271 2.01e-12 ***
## weight                  -8.351e-03  3.026e-03  -2.759  0.00607 **
## displacement            -9.357e-02  3.919e-02  -2.387  0.01746 *
## cylinders                1.851e+00  2.075e+00   0.892  0.37289
## weight:displacement      2.499e-05  8.250e-06   3.029  0.00262 **
## weight:cylinders        -3.801e-04  6.720e-04  -0.566  0.57197
## displacement:cylinders -2.026e-03  3.826e-03  -0.529  0.59682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.106 on 385 degrees of freedom
## Multiple R-squared:  0.7275, Adjusted R-squared:  0.7232
## F-statistic: 171.3 on 6 and 385 DF,  p-value: < 2.2e-16
```

Weight and displacement are statistically significant.

2) This problem involves the Boston data set, which we saw in class. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
summary(Boston)
```

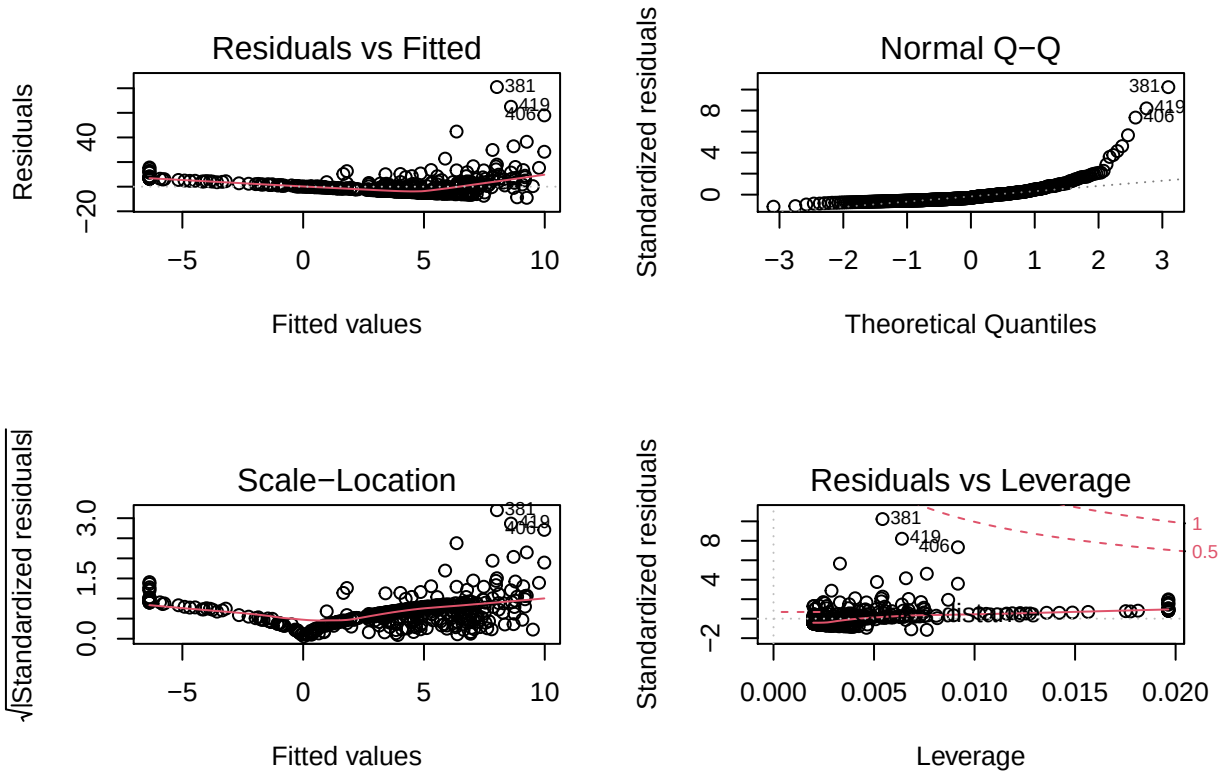
```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   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
```

a. (6%) For each predictor, fit a simple linear regression model to predict the response. Include the code, but not the output for all models in your solution.

```
lr_zn = lm(crim ~ zn , data = Boston)
lr_indus = lm(crim ~ indus , data = Boston)
lr_rm = lm(crim ~ rm , data = Boston)
lr_age = lm(crim ~ age , data = Boston)
lr_rad = lm(crim ~ rad , data = Boston)
lr_tax = lm(crim ~ tax , data = Boston)
lr_ptratio = lm(crim ~ ptratio , data = Boston)
lr_black = lm(crim ~ black , data = Boston)
lr_lstat = lm(crim ~ lstat , data = Boston)
lr_medv = lm(crim ~ medv , data = Boston)
lr_chas = lm(crim ~ chas , data = Boston)
lr_nox = lm(crim ~ nox , data = Boston)
lr_dis = lm(crim ~ dis , data = Boston)
```

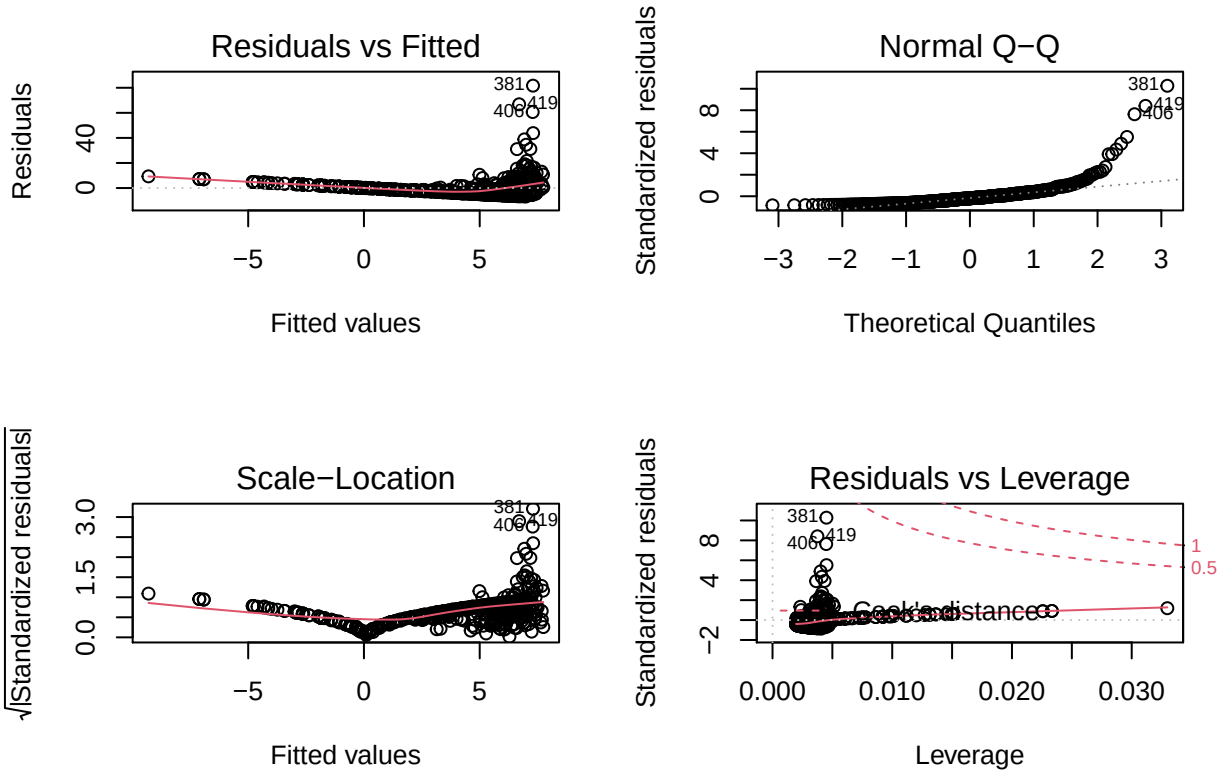
b. (6%) In which of the models is there a statistically significant association between the predictor and the response? Considering the meaning of each variable, discuss the relationship between crim and nox, chas, medv and dis in particular. How do these relationships differ?

```
lr_medv = lm(crim ~ medv , data = Boston)
par(mfrow = c(2,2))
plot(lr_medv)
```



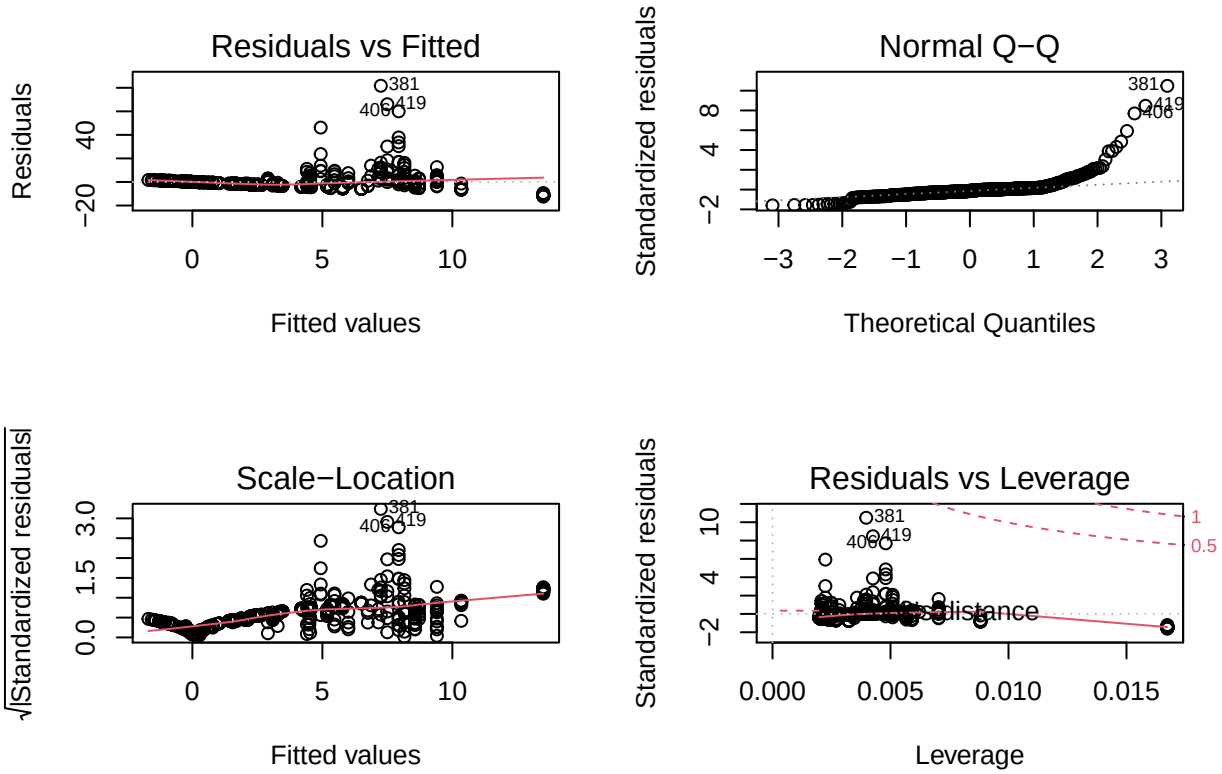
Two residuals plots fit better, they are mostly in a linear relation. Normal Q-Q would probably fit better in a polynomial model.

```
lr_dis = lm(crim ~ dis , data = Boston)
par(mfrow = c(2,2))
plot(lr_dis)
```



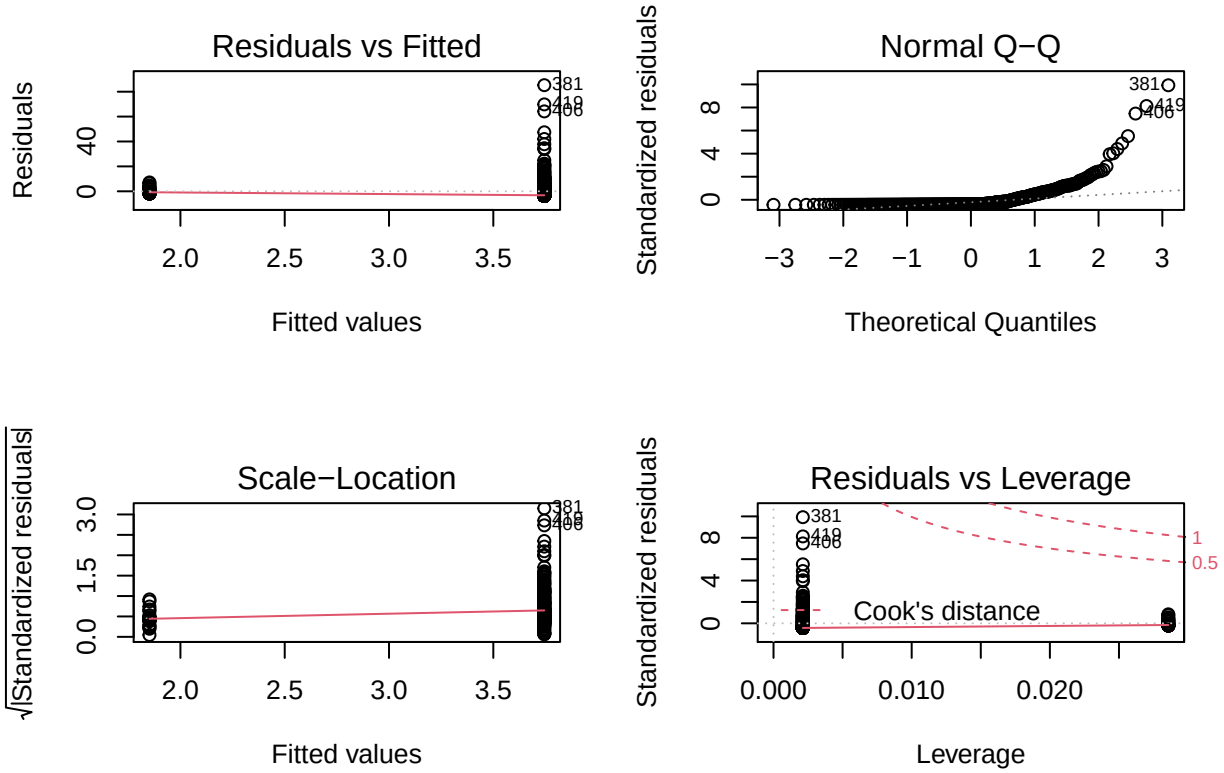
The plots for dis look pretty similar to crim.

```
par(mfrow = c(2,2))
plot(lr_nox)
```



The plots for nox have more value that are spread out on y axis.

```
par(mfrow = c(2,2))
plot(lr_chas)
```



The values for chas are catogorized, there is no certain relation.

c. (6%) Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : Bj = 0?

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

```
##
## 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
## 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 ***
```

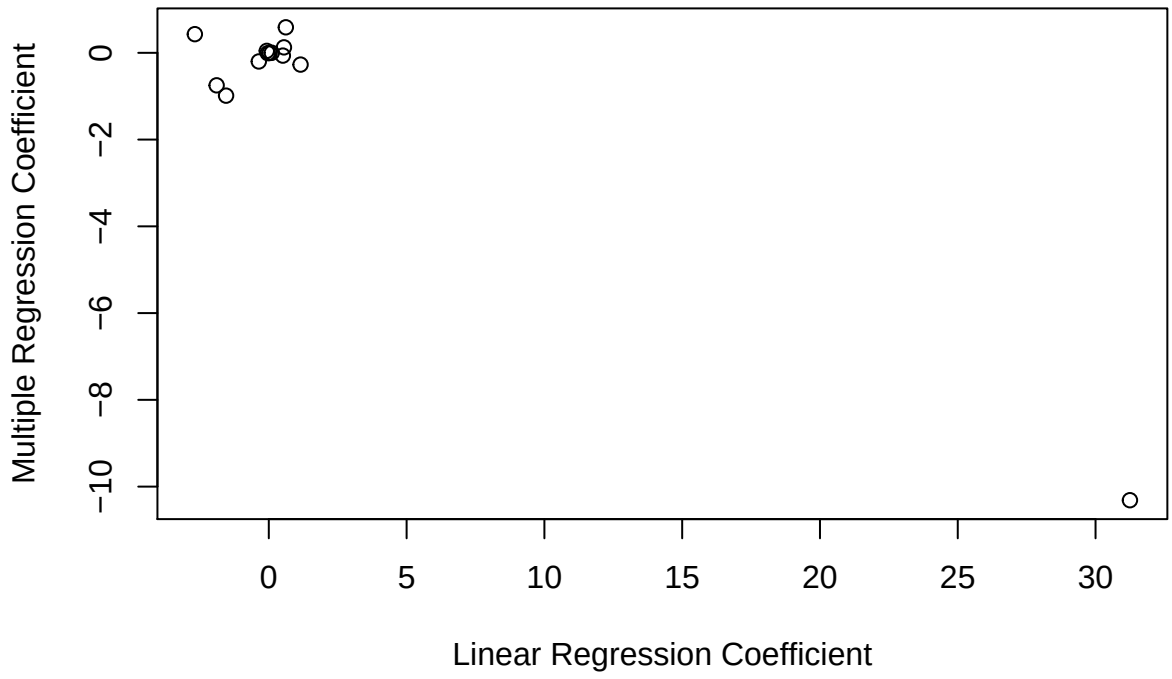
```
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   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
```

zn, dis, rad, black, and medv are rejected because they have statistically significant p-values.

d. (6%) How do your results from (a) compare to your results from (c)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (c) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. What does this plot tell you about the various predictors?

```r
linear_co <- list()
for (n in names(Boston[-1])) {
    linear_co[[n]] <- lm(crim ~ get(n), data = Boston)$coefficients[2]
}

multi_co <- lm(crim ~ ., data = Boston)$coefficients[-1]
plot(linear_co, multi_co, xlab = "Linear Regression Coefficient", ylab = "Multiple Regression Coefficie
```



e. (6%) 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 Y = B0 + B1X + B2X2 + B3X3+

e Hint: use the poly() function in R. Again, include the code, but not the output for each model in your solution, and instead describe any non-linear trends you uncover.

```
poly_zn = lm(crim ~ poly(zn, 3), data = Boston)
summary(poly_zn)
```

```
##
## Call:
## lm(formula = crim ~ poly(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)    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
```

```
poly_indus = lm(crim ~ poly(indus, 3), data = Boston)
poly_chas = lm(crim ~ chas + I(chas^2) + I(chas^3), data = Boston)
poly_nox = lm(crim ~ poly(nox, 3), data = Boston)
poly_rm = lm(crim ~ poly(rm, 3), data = Boston)
poly_age = lm(crim ~ poly(age, 3), data = Boston)
poly_dis = lm(crim ~ poly(dis, 3), data = Boston)
poly_rad = lm(crim ~ poly(rad, 3), data = Boston)
poly_tax = lm(crim ~ poly(tax, 3), data = Boston)
poly_ptratio = lm(crim ~ poly(ptratio, 3), data = Boston)
poly_black = lm(crim ~ poly(black, 3), data = Boston)
poly_lstat = lm(crim ~ poly(lstat, 3), data = Boston)
poly_medv = lm(crim ~ poly(medv, 3), data = Boston)
```

nox, age, dis, tax, and medv have high statistical significance (0.001) for Quadratic terms. zn, indus, rm, rad, ptratio, and black are at 0.01 for Quadratic terms. The predictors mentioned above has non-linear association between the response crim.

3)

a. (5%) Estimate the probability that a student who studies for 32 h, has a PSQI score of 12 and has an undergrad GPA of 3.0 gets an A in the class. Show your work.

```
temp = -7 + 0.1*32 + 1*3 - 0.04*12
px = exp(temp)/(1 + exp(temp))
px
```

```
## [1] 0.2175502
```

b. (5%) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class? Show your work.

```
# log(0.5/(1-0.5)) = -7 + 0.1*hours + 1*3 - 0.04*12
hours = (log(0.5/(1-0.5)) + 7 - 1*3 + 0.04*12) * 10
hours
```

```
## [1] 44.8
```

c. (5%) How many hours would a student with a 3.0 GPA and a PSQI score of 3 need to study to have a 50 % chance of getting an A in the class? Show your work.

```
# log(0.5/(1-0.5)) = -7 + 0.1*hours + 1*3 - 0.04*3
hours = (log(0.5/(1-0.5)) + 7 - 1*3 + 0.04*3) * 10
hours
```

```
## [1] 41.2
```

4)

a. Tokenization (20%)

```
library(tokenizers)
library(stopwords)
articles <- read.csv("GuardianArticles.csv")

token <- tokenize_word_stems(articles$body, stopwords = stopwords::stopwords("en"))
```

b. Classification (20%)

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
#train <- createDataPartition(token, p = .8)
```