IMT 573: Problem Set 7 - Regression

Divya Gaurav Tripathi

Due: Tuesday, November 19, 2019

Collaborators:

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 Server.

1. Download the `problemset7.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problemset7.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset7.Rmd`.
2. Replace the "Insert 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 not 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. (Note: I am also using the `include=FALSE` option here to not include this code in the PDF, but you need to remove this or change it to `TRUE` if you want to include the code chunk.)
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 PDF`, 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(ggplot2)
```

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. Load this data and use it to answer the following questions.

```
bostondata <- data("Boston")
str(bostondata)
```

```
##  chr "Boston"
```

```
str(Boston)
```

```
## '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 ...
```

1. Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.
2. Consider this data in context, what is the response variable of interest?
3. 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.
4. 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 $H_0 : \beta_j = 0$?
5. 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.
6. 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:

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

7. 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)?
8. 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.

Question 1: Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.

Answer: This data was collected by the U.S Census Servic and first published in 1978. The Boston data frame has 506 rows and 14 columns. It has various variables which may impact the price of houses in Boston Mass.

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. 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.

Question 2: Consider this data in context, what is the response variable of interest?

Answer: medv:median value of owner-occupied homes in $1000s. We are trying to predict the price of houses based on several factors.

Question 3: 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.

Answer

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

```
##
## Call:
## lm(formula = medv ~ crim, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## crim        -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 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
```

```
#we can tell from p value, crim is statistically significant

linearzn <-lm(medv ~ zn, data = Boston)
summary(linearzn)
```

```
##
## Call:
## lm(formula = medv ~ zn, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.918  -5.518  -1.006   2.757  29.082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 20.91758    0.42474  49.248   <2e-16 ***
## zn           0.14214    0.01638   8.675   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.587 on 504 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1282
## F-statistic: 75.26 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, zn is statistically significant

linearindus <- lm(medv ~ indus, data = Boston)
summary(linearindus)
```

```
##
## Call:
## lm(formula = medv ~ indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.017  -4.917  -1.457   3.180  32.943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.75490    0.68345   43.54   <2e-16 ***
## indus       -0.64849    0.05226  -12.41   <2e-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:  0.234,  Adjusted R-squared:  0.2325
## F-statistic:   154 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, indus is statistically significant

linearchas <- lm(medv ~ chas, data = Boston)
summary(linearchas)
```

```
##
## Call:
## lm(formula = medv ~ chas, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## chas          6.3462     1.5880   3.996 7.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

```r
#we can tell from p value, chas is statistically significant

linearnox <- lm(medv ~ nox, data = Boston)
summary(linearnox)
```

```
##
## Call:
## lm(formula = medv ~ nox, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.691  -5.121  -2.161   2.959  31.310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.346      1.811   22.83   <2e-16 ***
## nox          -33.916      3.196  -10.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.323 on 504 degrees of freedom
## Multiple R-squared:  0.1826, Adjusted R-squared:  0.181
## F-statistic: 112.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
#we can tell from p value, nox is statistically significant

linearrm <- lm(medv ~ rm, data = Boston)
summary(linearrm)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
#we can tell from p value, rm is statistically significant

linearage <- lm(medv ~ age, data = Boston)
summary(linearage)
```

```
##
## Call:
```

```
## lm(formula = medv ~ age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## age         -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, age is statistically significant

lineardis <- lm(medv ~ dis, data = Boston)
summary(lineardis)
```

```
##
## Call:
## lm(formula = medv ~ dis, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.016  -5.556  -1.865   2.288  30.377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3901     0.8174  22.499  < 2e-16 ***
## dis           1.0916     0.1884   5.795 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 504 degrees of freedom
## Multiple R-squared:  0.06246,    Adjusted R-squared:  0.0606
## F-statistic: 33.58 on 1 and 504 DF,  p-value: 1.207e-08
```

```
#we can tell from p value, dis is statistically significant

linearrad <- lm(medv ~ rad, data = Boston)
summary(linearrad)
```

```
##
## Call:
## lm(formula = medv ~ rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.770  -5.199  -1.967   3.321  33.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 26.38213    0.56176  46.964   <2e-16 ***
## rad         -0.40310    0.04349  -9.269   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.509 on 504 degrees of freedom
## Multiple R-squared:  0.1456, Adjusted R-squared:  0.1439
## F-statistic: 85.91 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, rad is statistically significant

lineartax <- lm(medv ~ tax, data = Boston)
summary(lineartax)
```

```
##
## Call:
## lm(formula = medv ~ tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.091  -5.173  -2.085   3.158  34.058
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.970654   0.948296   34.77   <2e-16 ***
## tax         -0.025568   0.002147  -11.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.133 on 504 degrees of freedom
## Multiple R-squared:  0.2195, Adjusted R-squared:  0.218
## F-statistic: 141.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, tax is statistically significant

linearptratio <- lm(medv ~ ptratio, data = Boston)
summary(linearptratio)
```

```
##
## Call:
## lm(formula = medv ~ ptratio, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.345      3.029   20.58   <2e-16 ***
## ptratio       -2.157      0.163  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.931 on 504 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2564
## F-statistic: 175.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
#we can tell from p value, ptratio is statistically significant

linearblack <- lm(medv ~ black, data = Boston)
summary(linearblack)
```

```
##
## Call:
## lm(formula = medv ~ black, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.884  -4.862  -1.684   2.932  27.763
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.551034   1.557463   6.775 3.49e-11 ***
## black        0.033593   0.004231   7.941 1.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.679 on 504 degrees of freedom
## Multiple R-squared:  0.1112, Adjusted R-squared:  0.1094
## F-statistic: 63.05 on 1 and 504 DF,  p-value: 1.318e-14
```

```r
#we can tell from p value, black is statistically significant

linearlstat <- lm(medv ~ lstat, data = Boston)
summary(linearlstat)
```

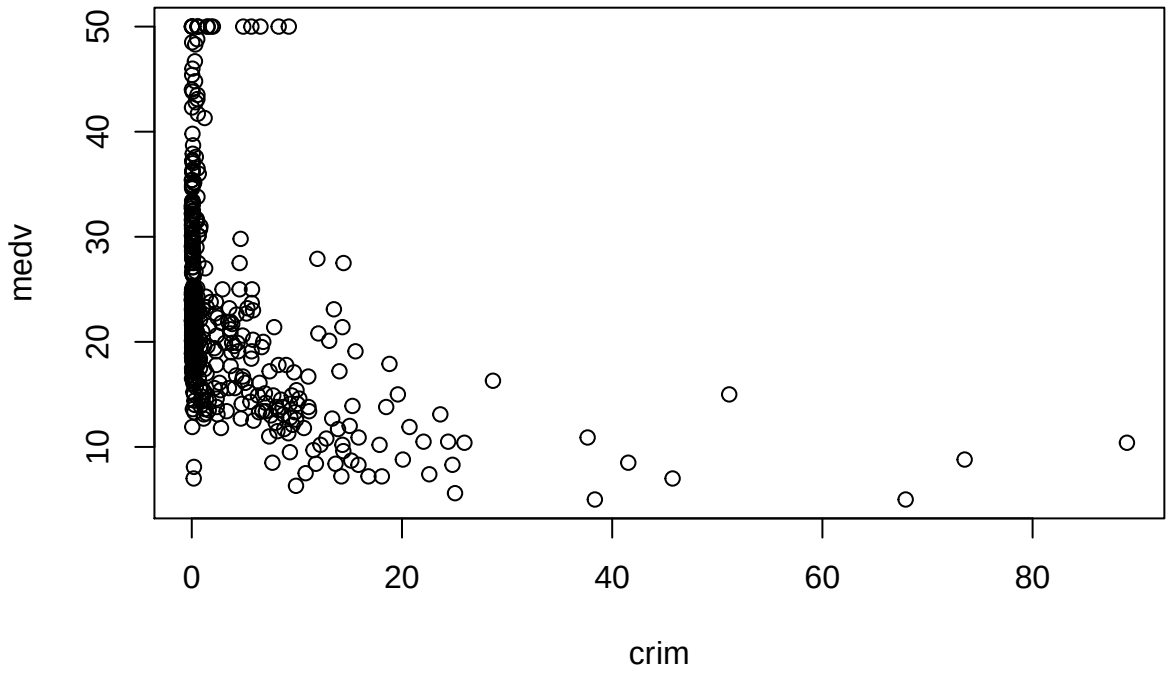
```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

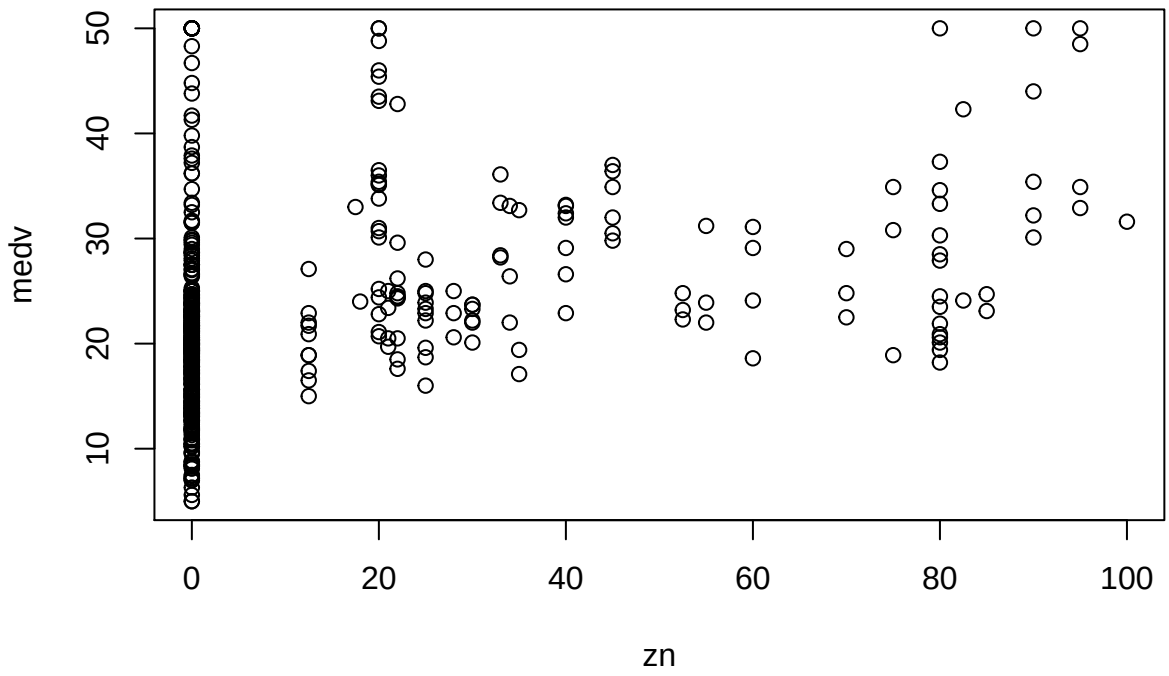
```r
#we can tell from p value, lstat is statistically significant


str(linearcrim$coefficients)
```

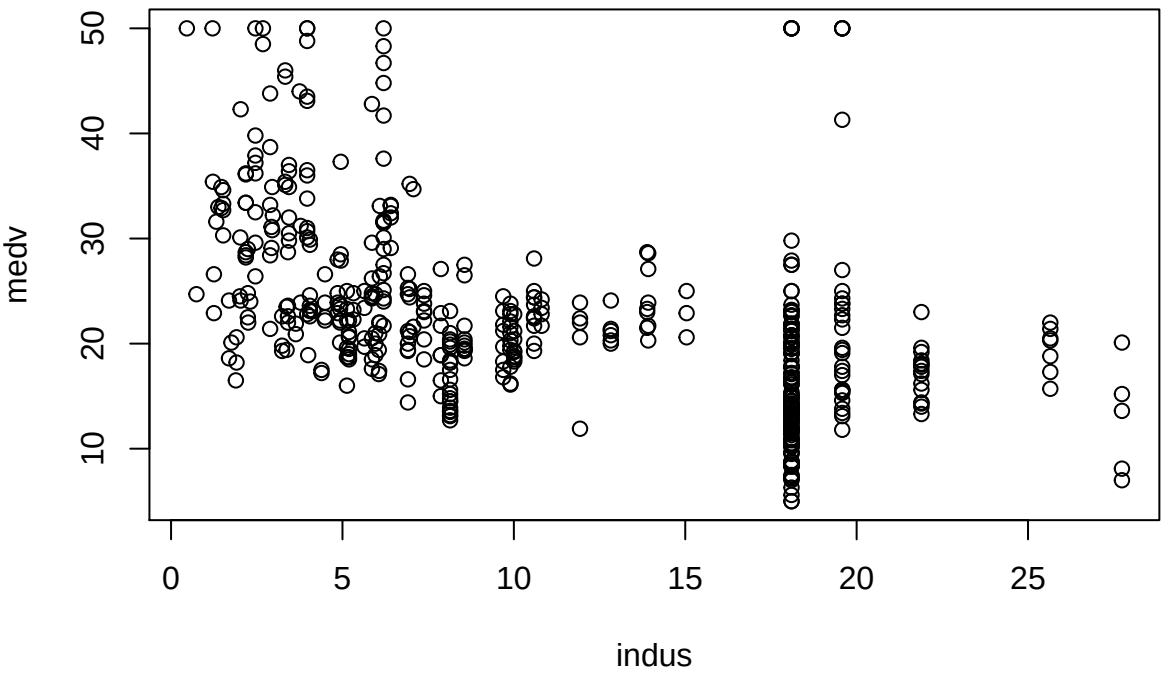
```
##  Named num [1:2] 24.033 -0.415
##  - attr(*, "names")= chr [1:2] "(Intercept)" "crim"
```

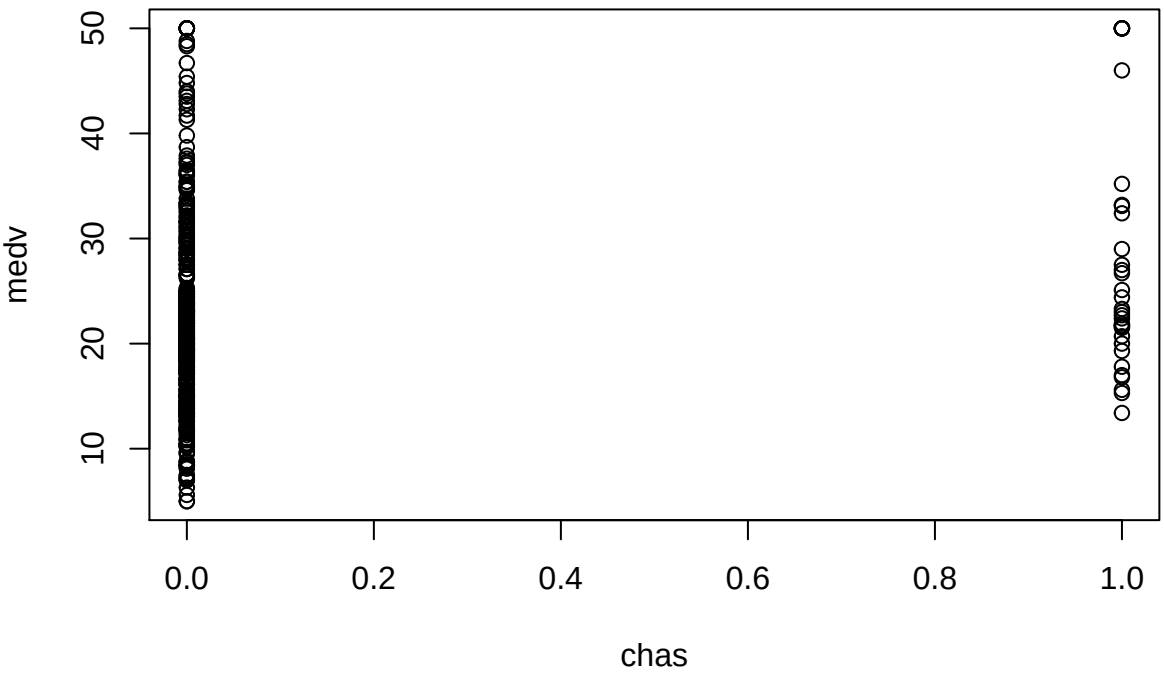
We fitted linear regression models seperately for each variables.

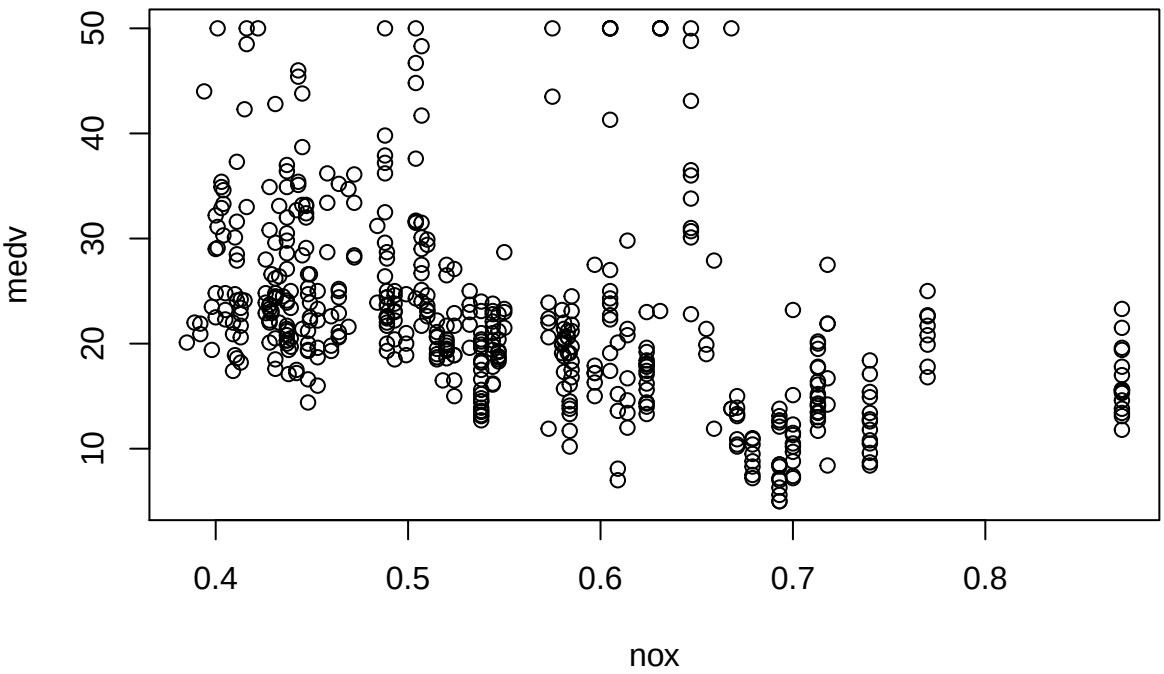
```
plot(medv ~ . - medv, data = Boston)
```

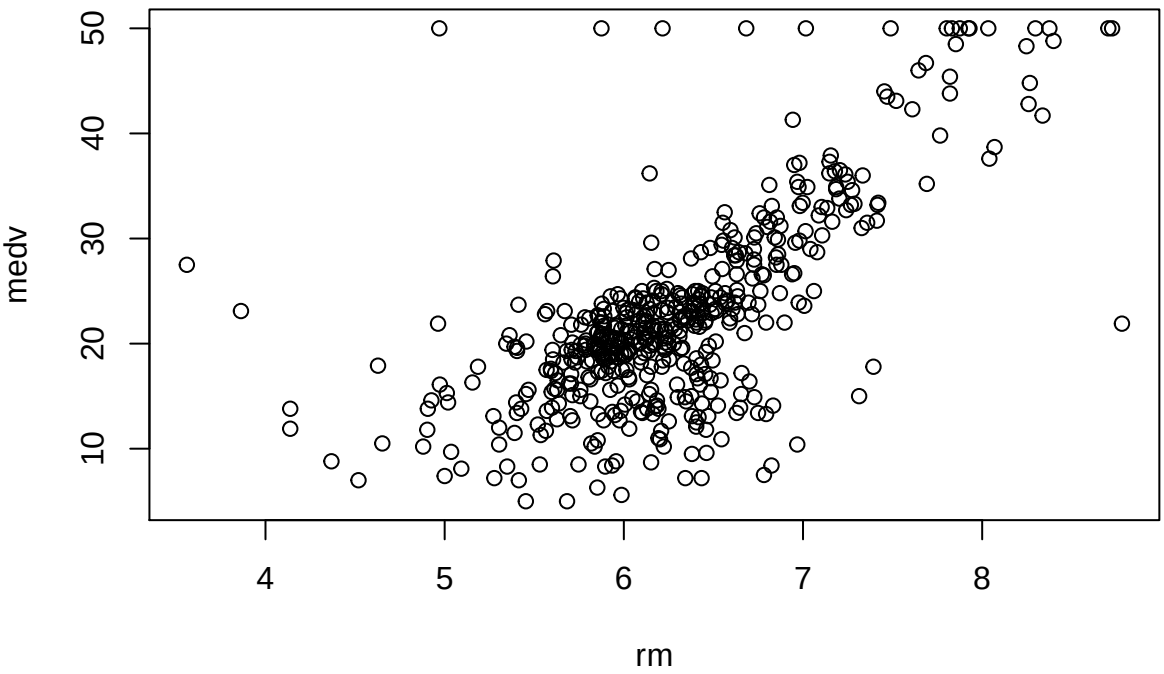


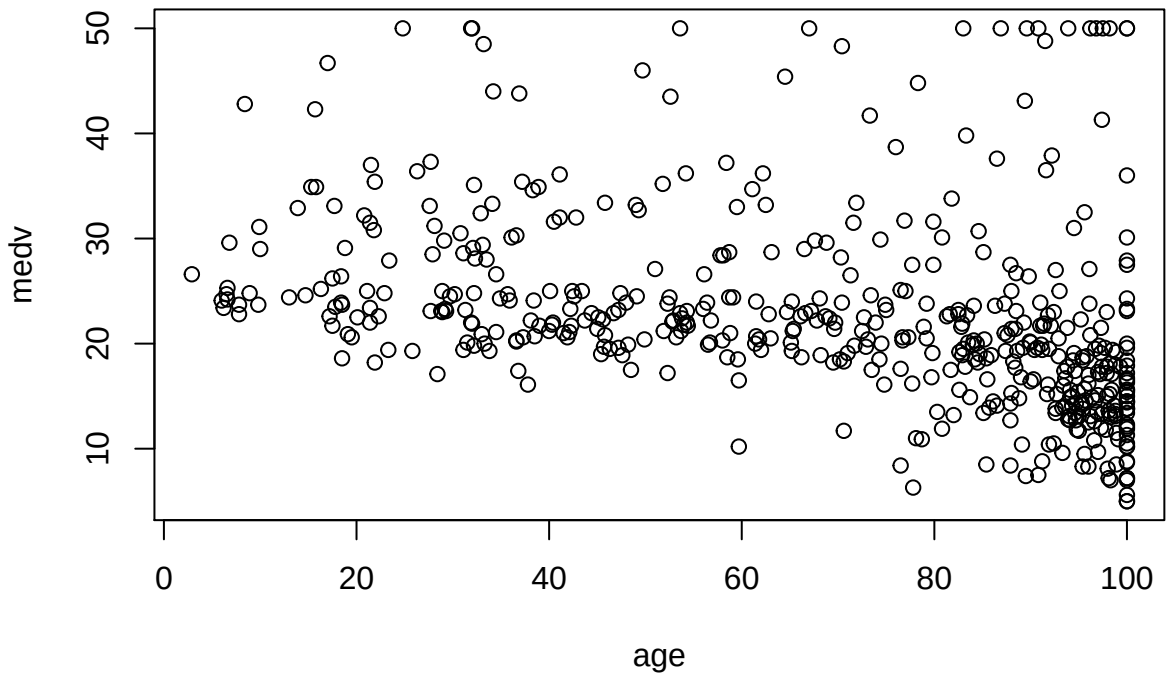

medv
10
20
30
40
50
0
20
40
60
80
100
zn

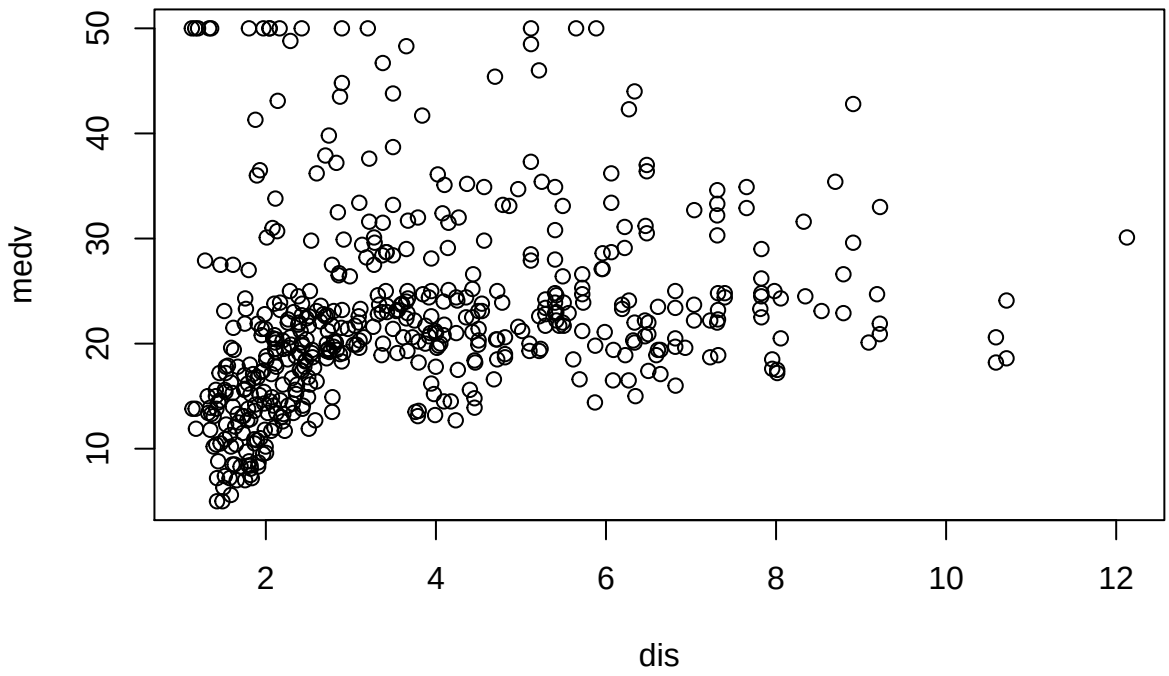

medv
indus
10
20
30
40
50
0
5
10
15
20
25

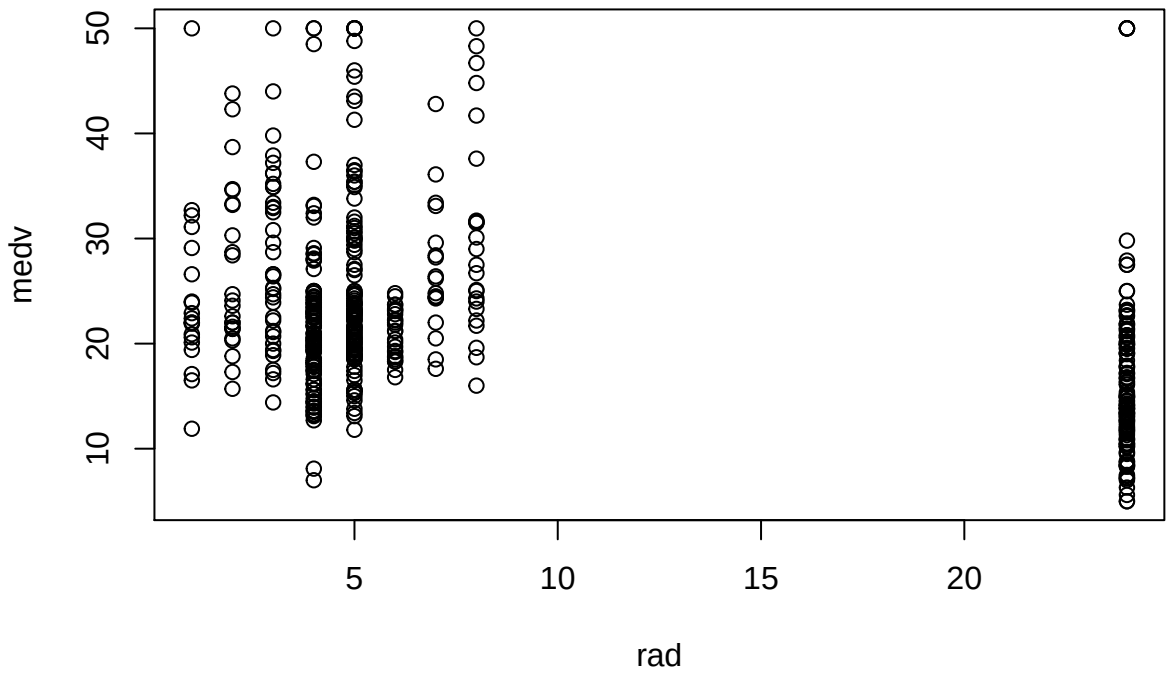

medv
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
chas

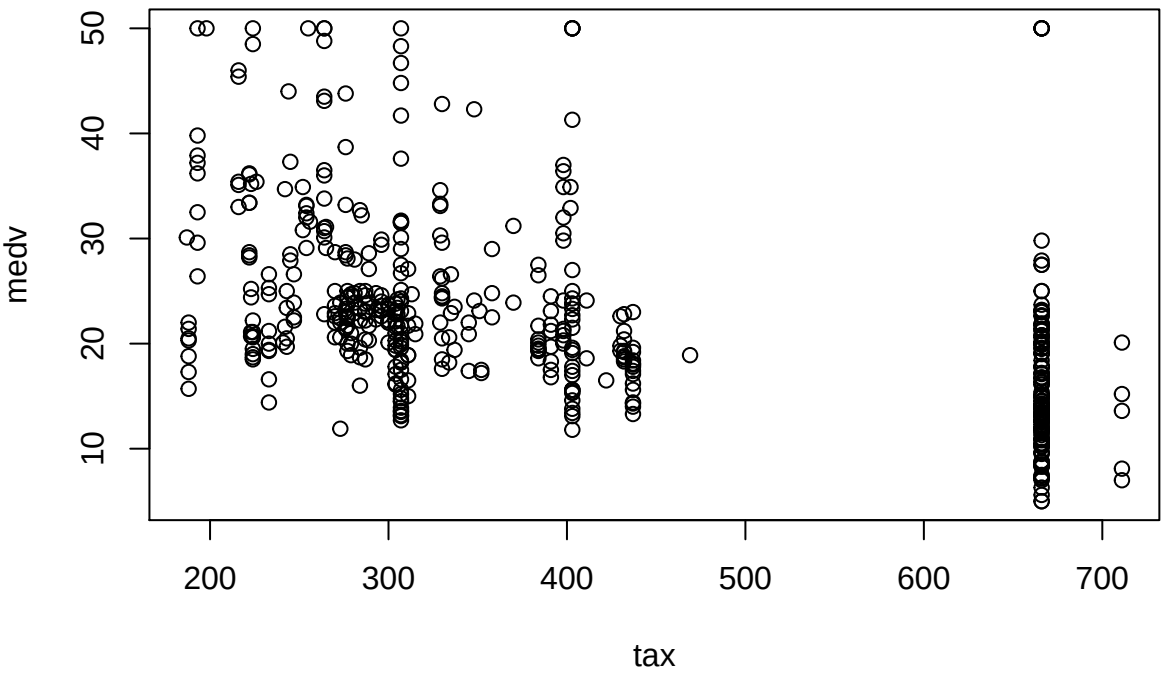

medv
10
20
30
40
50
0.4
0.5
0.6
0.7
0.8
nox

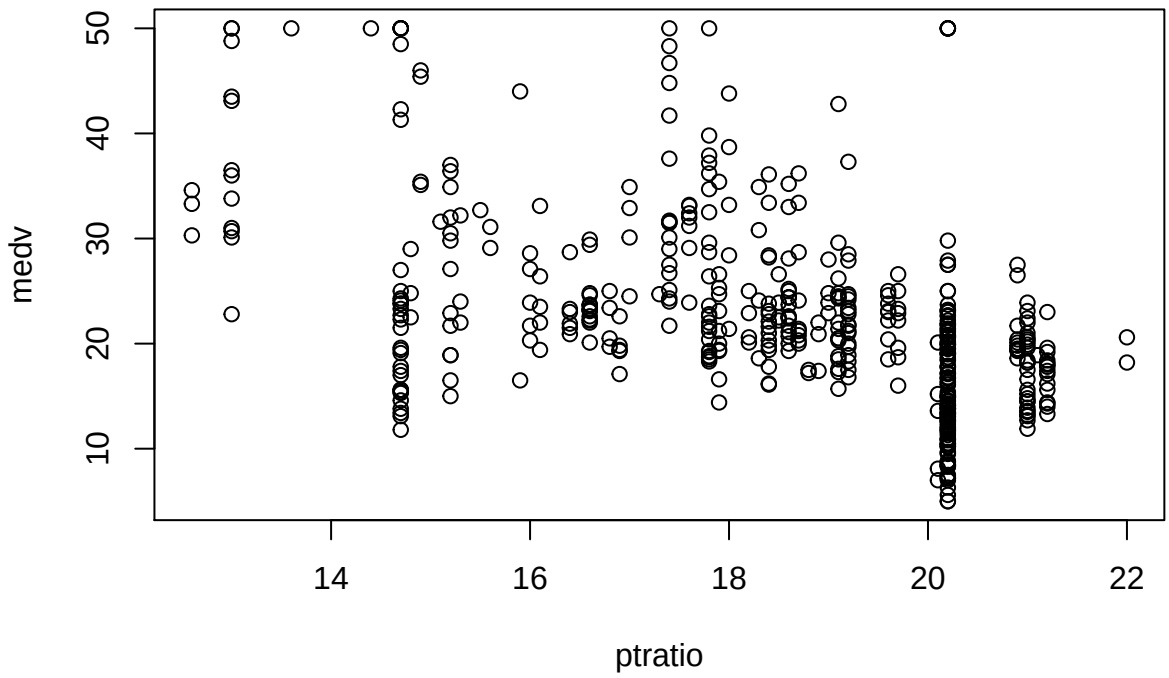

medv
10
20
30
40
50
4
5
6
7
8
rm

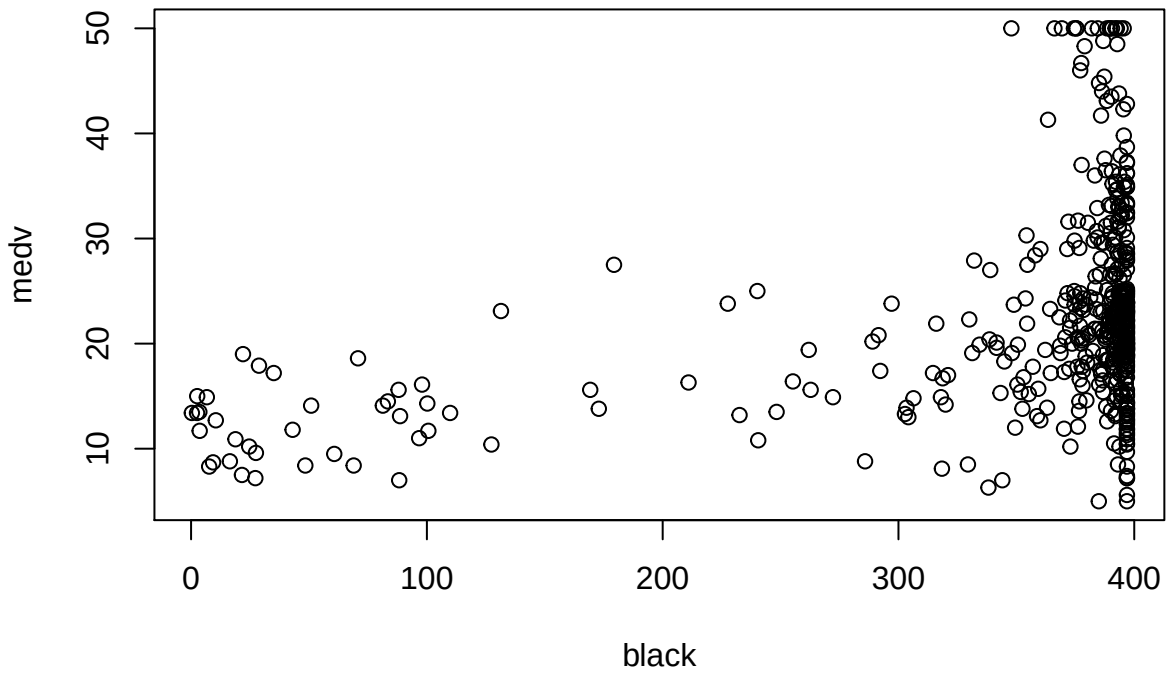

medv
10
20
30
40
50
0
20
40
60
80
100
age

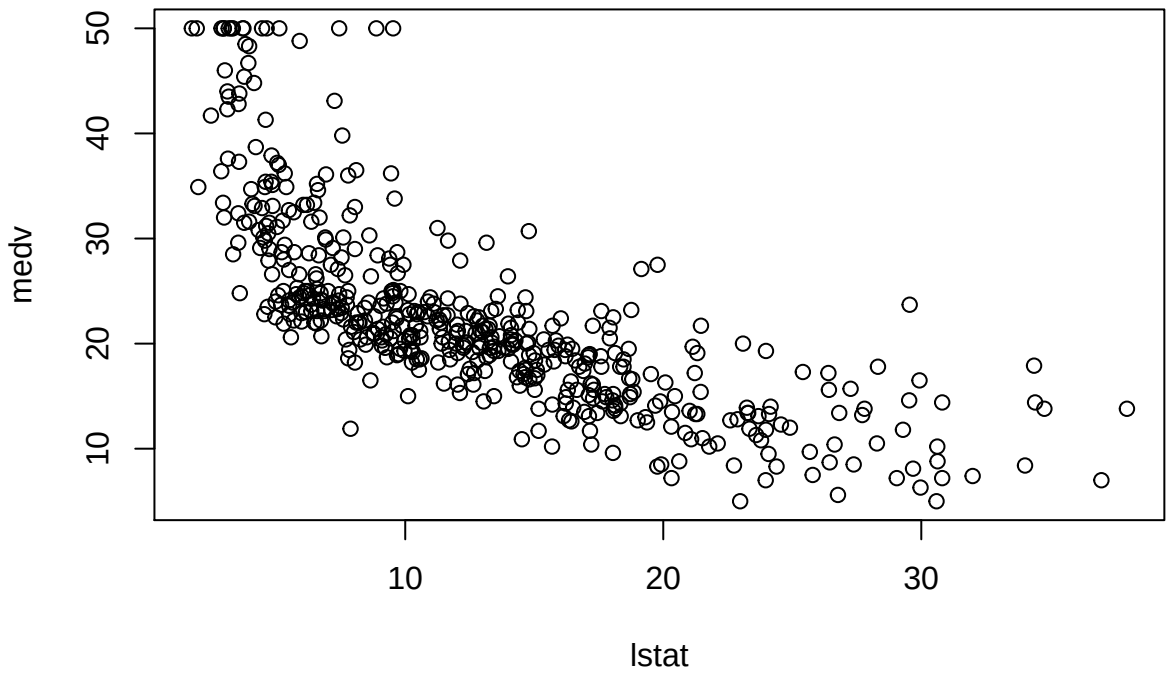

medv
10
20
30
40
50
2
4
6
8
10
12
dis


medv
10
20
30
40
50
5
10
15
20
rad


medv
10
20
30
40
50
200
300
400
500
600
700
tax


medv
10
20
30
40
50
14
16
18
20
22
ptratio


medv
10
20
30
40
50
0
100
200
300
400
black



When we fitted linear regression models for these predictors, the values told all of these are statistically significant. However I would like to say that when we check the plots, although the predictors are statistically significant, all of them do not appear to have strong linear relationship with medv. rm has a more linear relation with medv than tax. Some predictors like black also have strong skew.

Question 4: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 $H_0 : \beta_j = 0$?

```
multiple_regression <- lm(medv ~ .-medv, data = Boston)
#We created a multiole regression model, medv is response variable
#-medv says it is not a predictor variable
summary(multiple_regression)
```

```
##
## Call:
## lm(formula = medv ~ . - medv, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
```

```
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

We find that all predictors are statistically significant except indus and age.

We can reject null hypothesis H_0

$$\beta_j = 0$$

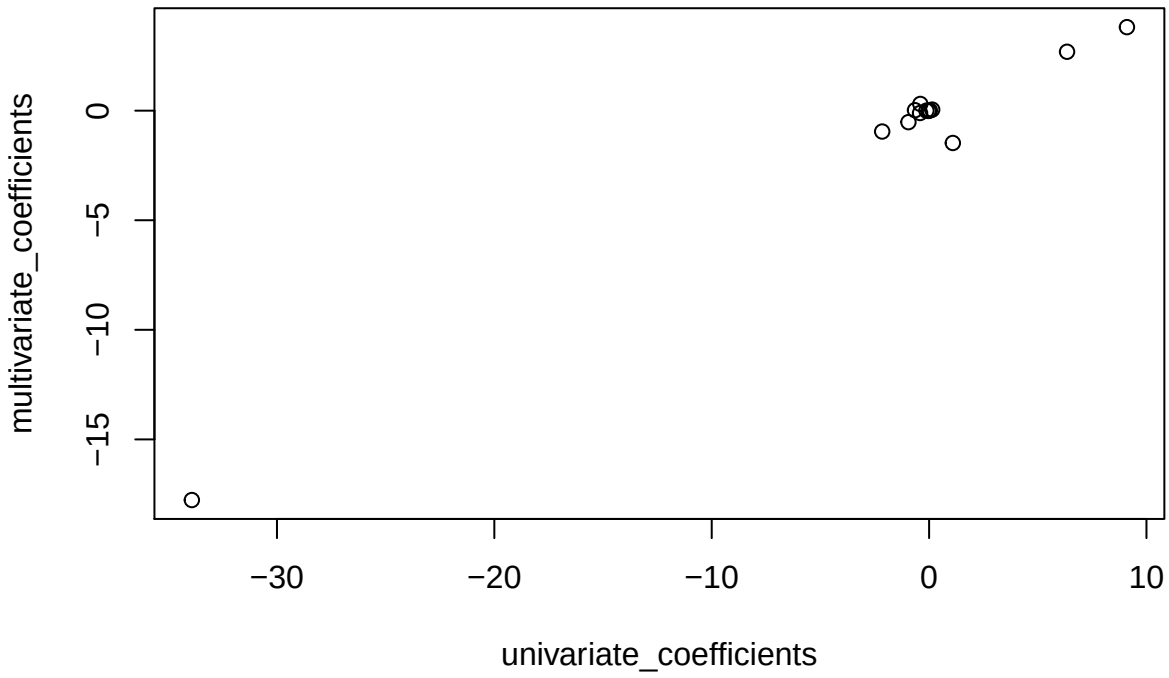
for all of the predictors except indus and age.

Question 5: 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.

In (3) we compared every predictor individually and fitted seperate linear regression models; we found that all of the predictors were statistically significant. In (4) we fitted a multiple regression model taking all predictors together and found indus and age are not statistically significant. Sometimes there are underlying relations and covariance between two or more variables which is taken into account only when all variables are considered together. Also the impact of a predictor of univariate linear regression on the response variable may be different from its impact on the response variable in a multivariate linear regression model.

```r
#univariatemultiple <- (lm(medv ~ . -medv, data = Boston))

univariate_coefficients <- c(linearcrim$coefficients[2],
              linearzn$coefficients[2],
              linearindus$coefficients[2],
              linearchas$coefficients[2],
              linearnox$coefficients[2],
              linearrm$coefficients[2],
              linearage$coefficients[2],
              lineardis$coefficients[2],
              linearrad$coefficients[2],
              lineartax$coefficients[2],
              linearptratio$coefficients[2],
              linearblack$coefficients[2],
              linearlstat$coefficients[2]
              )

#We made a list of all univariate linear coefficients without the intercept

multivariate_coefficients <- multiple_regression$coefficients[-1]

#We made a list of all multivariate linear coefficients without the intercept
```

```r
plot(univariate_coefficients,multivariate_coefficients)
```



We can find from this graph that the values of coefficients are different in univariate and multivariate linear regressions.

Question 6: 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:

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

Answer: We saw from the above graphs that every variable did not have total linear relationship with medv, and had very strong skews. We are going to fit polynomial regression for all the predictors.

chas onky takes a binary value(0 or 1), so we cannot do polynomial regression for that.

```r
polycrim <- lm(medv ~ poly(crim,3), data = Boston)
summary(polycrim)
```

```
##
## Call:
## lm(formula = medv ~ poly(crim, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.983  -4.975  -1.940   2.881  33.391
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     22.5328     0.3627  62.124  < 2e-16 ***
## poly(crim, 3)1 -80.2545     8.1589  -9.836  < 2e-16 ***
## poly(crim, 3)2  50.2416     8.1589   6.158 1.51e-09 ***
## poly(crim, 3)3 -18.2905     8.1589  -2.242   0.0254 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.159 on 502 degrees of freedom
## Multiple R-squared:  0.2177, Adjusted R-squared:  0.213
## F-statistic: 46.57 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, crim is statistically significant

polyzn <-lm(medv ~ poly(zn,3), data = Boston)
summary(polyzn)
```

```
##
## Call:
## lm(formula = medv ~ poly(zn, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.449  -5.549  -1.049   3.225  29.551
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.5328     0.3747  60.129  < 2e-16 ***
## poly(zn, 3)1  74.4966     8.4296   8.837  < 2e-16 ***
## poly(zn, 3)2 -19.2591     8.4296  -2.285   0.0227 *
## poly(zn, 3)3  33.5309     8.4296   3.978 7.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.43 on 502 degrees of freedom
## Multiple R-squared:  0.1649, Adjusted R-squared:  0.1599
## F-statistic: 33.05 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, zn is statistically significant

polyindus <- lm(medv ~ poly(indus,3), data = Boston)
summary(polyindus)
```

```
##
## Call:
## lm(formula = medv ~ poly(indus, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.760  -4.725  -1.009   2.932  32.038
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.5328     0.3487  64.614  < 2e-16 ***
## poly(indus, 3)1 -99.9759     7.8445 -12.745  < 2e-16 ***
## poly(indus, 3)2  38.5184     7.8445   4.910 1.23e-06 ***
```

```
## poly(indus, 3)3 -18.6140     7.8445  -2.373    0.018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.844 on 502 degrees of freedom
## Multiple R-squared:  0.2768, Adjusted R-squared:  0.2725
## F-statistic: 64.06 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, indus is statistically significant

#polychas <- lm(medv ~ poly(chas,3), data = Boston)
#we can tell from p value, chas is statistically significant

polynox <- lm(medv ~ poly(nox,3), data = Boston)
summary(polynox)
```

```
##
## Call:
## lm(formula = medv ~ poly(nox, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.104  -5.020  -2.144   2.747  32.416
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.5328     0.3682  61.199   <2e-16 ***
## poly(nox, 3)1 -88.3183     8.2823 -10.664   <2e-16 ***
## poly(nox, 3)2  13.8989     8.2823   1.678   0.0939 .
## poly(nox, 3)3  16.9686     8.2823   2.049   0.0410 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.282 on 502 degrees of freedom
## Multiple R-squared:  0.1939, Adjusted R-squared:  0.189
## F-statistic: 40.24 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, nox is statistically significant

polyrm <- lm(medv ~ poly(rm,3), data = Boston)
summary(polyrm)
```

```
##
## Call:
## lm(formula = medv ~ poly(rm, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.102  -2.674   0.569   3.011  35.911
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.5328     0.2716  82.952  < 2e-16 ***
## poly(rm, 3)1 143.7164     6.1103  23.520  < 2e-16 ***
## poly(rm, 3)2  52.6526     6.1103   8.617  < 2e-16 ***
## poly(rm, 3)3 -23.3832     6.1103  -3.827 0.000146 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.11 on 502 degrees of freedom
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5586
## F-statistic:   214 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, rm is statistically significant

polyage <- lm(medv ~ poly(age,3), data = Boston)
summary(polyage)
```

```
##
## Call:
## lm(formula = medv ~ poly(age, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.443  -4.909  -2.234   2.185  32.944
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.5328     0.3766  59.830   <2e-16 ***
## poly(age, 3)1 -77.9087     8.4717  -9.196   <2e-16 ***
## poly(age, 3)2 -23.3290     8.4717  -2.754   0.0061 **
## poly(age, 3)3  -8.6148     8.4717  -1.017   0.3097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.472 on 502 degrees of freedom
## Multiple R-squared:  0.1566, Adjusted R-squared:  0.1515
## F-statistic: 31.06 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, age is statistically significant

polydis <- lm(medv ~ poly(dis,3), data = Boston)
summary(polydis)
```

```
##
## Call:
## lm(formula = medv ~ poly(dis, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.571  -5.242  -2.037   2.397  34.769
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.5328     0.3879  58.082  < 2e-16 ***
## poly(dis, 3)1  51.6551     8.7267   5.919 6.00e-09 ***
## poly(dis, 3)2 -37.5859     8.7267  -4.307 1.99e-05 ***
## poly(dis, 3)3  20.1322     8.7267   2.307   0.0215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.727 on 502 degrees of freedom
```

```
## Multiple R-squared:  0.105,  Adjusted R-squared:  0.09968
## F-statistic: 19.64 on 3 and 502 DF,  p-value: 4.736e-12
```

```
#we can tell from p value, dis is statistically significant

polyrad <- lm(medv ~ poly(rad,3), data = Boston)
summary(polyrad)
```

```
##
## Call:
## lm(formula = medv ~ poly(rad, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.630  -5.151  -2.017   3.169  33.594
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.5328     0.3721  60.557  < 2e-16 ***
## poly(rad, 3)1 -78.8742     8.3700  -9.423  < 2e-16 ***
## poly(rad, 3)2 -21.4799     8.3700  -2.566 0.010568 *
## poly(rad, 3)3 -29.4095     8.3700  -3.514 0.000482 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.37 on 502 degrees of freedom
## Multiple R-squared:  0.1767, Adjusted R-squared:  0.1718
## F-statistic: 35.91 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, rad is statistically significant

polytax <- lm(medv ~ poly(tax,3), data = Boston)
summary(polytax)
```

```
##
## Call:
## lm(formula = medv ~ poly(tax, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.109  -4.952  -1.878   2.957  33.694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.5328     0.3608  62.460   <2e-16 ***
## poly(tax, 3)1 -96.8366     8.1150 -11.933   <2e-16 ***
## poly(tax, 3)2  14.9703     8.1150   1.845   0.0657 .
## poly(tax, 3)3  -7.5431     8.1150  -0.930   0.3531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.115 on 502 degrees of freedom
## Multiple R-squared:  0.2261, Adjusted R-squared:  0.2215
## F-statistic: 48.89 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
#we can tell from p value, tax is statistically significant

polyptratio <- lm(medv ~ poly(ptratio,3), data = Boston)
summary(polyptratio)
```

```
##
## Call:
## lm(formula = medv ~ poly(ptratio, 3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7795  -5.0364  -0.9778   3.4766  31.1636
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          22.5328     0.3511  64.173   <2e-16 ***
## poly(ptratio, 3)1 -104.9490     7.8984 -13.287   <2e-16 ***
## poly(ptratio, 3)2  -12.6952     7.8984  -1.607    0.109
## poly(ptratio, 3)3  -14.9472     7.8984  -1.892    0.059 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.898 on 502 degrees of freedom
## Multiple R-squared:  0.2669, Adjusted R-squared:  0.2625
## F-statistic: 60.91 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
#we can tell from p value, ptratio is statistically significant

polyblack <- lm(medv ~ poly(black,3), data = Boston)
summary(polyblack)
```

```
##
## Call:
## lm(formula = medv ~ poly(black, 3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.005  -4.802  -1.613   2.852  28.051
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.5328     0.3861  58.360  < 2e-16 ***
## poly(black, 3)1  68.9194     8.6851   7.935 1.38e-14 ***
## poly(black, 3)2   9.1467     8.6851   1.053    0.293
## poly(black, 3)3  -4.0541     8.6851  -0.467    0.641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.685 on 502 degrees of freedom
## Multiple R-squared:  0.1135, Adjusted R-squared:  0.1082
## F-statistic: 21.43 on 3 and 502 DF,  p-value: 4.463e-13
```

```r
#we can tell from p value, black is statistically significant

polylstat <- lm(medv ~ poly(lstat,3), data = Boston)
```

```
summary(polylstat)
```

```
##
## Call:
## lm(formula = medv ~ poly(lstat, 3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5441  -3.7122  -0.5145   2.4846  26.4153
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       22.5328     0.2399  93.937  < 2e-16 ***
## poly(lstat, 3)1 -152.4595     5.3958 -28.255  < 2e-16 ***
## poly(lstat, 3)2   64.2272     5.3958  11.903  < 2e-16 ***
## poly(lstat, 3)3  -27.0511     5.3958  -5.013 7.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.396 on 502 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6558
## F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#we can tell from p value, lstat is statistically significant
```

We fitted a polynomial regression of the form

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

for each of the predictors.

tax,ptratio,black have linear relationship with medv.

The following have non linear relation: crim(X,X^{2,X}3 are statistically significant), zn(X,X^{2,X}3 are statistically significant), indus(X,X^{2,X}3 are statistically significant), nox(X,X^3 are statistically significant), rm(X,X^{2,X}3 are statistically significant), age(X,X^2 are statistically significant), dis(X,X^{2,X}3 are statistically significant), rad(X,X^{2,X}3 are statistically significant), lstat(X,X^{2,X}3 are statistically significant)

Question 7: 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)?

Answer: We would perform stepwise model selection through forward model selections.

```
model_withoutvariables <- lm(medv ~ 1, data = Boston)
#We created a model without any determinant

model_withvariables <- lm(medv ~ .-medv, data = Boston)
#We created a model with all determinants

model_forward_selection <- step(model_withoutvariables,

                    scope = list(lower = model_withoutvariables, upper = model_withvariables), directi
```

```
## Start:  AIC=2246.51
## medv ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + lstat    1   23243.9 19472 1851.0
```

```
## + rm       1   20654.4 22062 1914.2
## + ptratio  1   11014.3 31702 2097.6
## + indus    1    9995.2 32721 2113.6
## + tax      1    9377.3 33339 2123.1
## + nox      1    7800.1 34916 2146.5
## + crim     1    6440.8 36276 2165.8
## + rad      1    6221.1 36495 2168.9
## + age      1    6069.8 36647 2171.0
## + zn       1    5549.7 37167 2178.1
## + black    1    4749.9 37966 2188.9
## + dis      1    2668.2 40048 2215.9
## + chas     1    1312.1 41404 2232.7
## <none>                 42716 2246.5
##
## Step:  AIC=1851.01
## medv ~ lstat
##
##           Df Sum of Sq   RSS    AIC
## + rm       1    4033.1 15439 1735.6
## + ptratio  1    2670.1 16802 1778.4
## + chas     1     786.3 18686 1832.2
## + dis      1     772.4 18700 1832.5
## + age      1     304.3 19168 1845.0
## + tax      1     274.4 19198 1845.8
## + black    1     198.3 19274 1847.8
## + zn       1     160.3 19312 1848.8
## + crim     1     146.9 19325 1849.2
## + indus    1      98.7 19374 1850.4
## <none>                 19472 1851.0
## + rad      1      25.1 19447 1852.4
## + nox      1       4.8 19468 1852.9
##
## Step:  AIC=1735.58
## medv ~ lstat + rm
##
##           Df Sum of Sq   RSS    AIC
## + ptratio  1   1711.32 13728 1678.1
## + chas     1    548.53 14891 1719.3
## + black    1    512.31 14927 1720.5
## + tax      1    425.16 15014 1723.5
## + dis      1    351.15 15088 1725.9
## + crim     1    311.42 15128 1727.3
## + rad      1    180.45 15259 1731.6
## + indus    1     61.09 15378 1735.6
## <none>                 15439 1735.6
## + zn       1     56.56 15383 1735.7
## + age      1     20.18 15419 1736.9
## + nox      1     14.90 15424 1737.1
##
## Step:  AIC=1678.13
## medv ~ lstat + rm + ptratio
##
##         Df Sum of Sq   RSS    AIC
## + dis    1    499.08 13229 1661.4
```

```
## + black  1    389.68 13338 1665.6
## + chas   1    377.96 13350 1666.0
## + crim   1    122.52 13606 1675.6
## + age    1     66.24 13662 1677.7
## <none>               13728 1678.1
## + tax    1     44.36 13684 1678.5
## + nox    1     24.81 13703 1679.2
## + zn     1     14.96 13713 1679.6
## + rad    1      6.07 13722 1679.9
## + indus  1      0.83 13727 1680.1
##
## Step:  AIC=1661.39
## medv ~ lstat + rm + ptratio + dis
##
##         Df Sum of Sq   RSS    AIC
## + nox    1    759.56 12469 1633.5
## + black  1    502.64 12726 1643.8
## + chas   1    267.43 12962 1653.1
## + indus  1    242.65 12986 1654.0
## + tax    1    240.34 12989 1654.1
## + crim   1    233.54 12995 1654.4
## + zn     1    144.81 13084 1657.8
## + age    1     61.36 13168 1661.0
## <none>               13229 1661.4
## + rad    1     22.40 13206 1662.5
##
## Step:  AIC=1633.47
## medv ~ lstat + rm + ptratio + dis + nox
##
##         Df Sum of Sq   RSS    AIC
## + chas   1    328.27 12141 1622.0
## + black  1    311.83 12158 1622.7
## + zn     1    151.71 12318 1629.3
## + crim   1    141.43 12328 1629.7
## + rad    1     53.48 12416 1633.3
## <none>               12469 1633.5
## + indus  1     17.10 12452 1634.8
## + tax    1     10.50 12459 1635.0
## + age    1      0.25 12469 1635.5
##
## Step:  AIC=1621.97
## medv ~ lstat + rm + ptratio + dis + nox + chas
##
##         Df Sum of Sq   RSS    AIC
## + black  1   272.837 11868 1612.5
## + zn     1   164.406 11977 1617.1
## + crim   1   116.330 12025 1619.1
## + rad    1    58.556 12082 1621.5
## <none>               12141 1622.0
## + indus  1    26.274 12115 1622.9
## + tax    1     4.187 12137 1623.8
## + age    1     2.331 12139 1623.9
##
## Step:  AIC=1612.47
```

```
## medv ~ lstat + rm + ptratio + dis + nox + chas + black
##
##         Df Sum of Sq   RSS    AIC
## + zn     1   189.936 11678 1606.3
## + rad    1   144.320 11724 1608.3
## + crim   1    55.633 11813 1612.1
## <none>               11868 1612.5
## + indus  1    15.584 11853 1613.8
## + age    1     9.446 11859 1614.1
## + tax    1     2.703 11866 1614.4
##
## Step:  AIC=1606.31
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn
##
##         Df Sum of Sq   RSS    AIC
## + crim   1    94.712 11584 1604.2
## + rad    1    93.614 11585 1604.2
## <none>               11678 1606.3
## + indus  1    16.048 11662 1607.6
## + tax    1     3.952 11674 1608.1
## + age    1     1.491 11677 1608.2
##
## Step:  AIC=1604.19
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim
##
##         Df Sum of Sq   RSS    AIC
## + rad    1   228.604 11355 1596.1
## <none>               11584 1604.2
## + indus  1    15.773 11568 1605.5
## + age    1     2.470 11581 1606.1
## + tax    1     1.305 11582 1606.1
##
## Step:  AIC=1596.1
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim + rad
##
##         Df Sum of Sq   RSS    AIC
## + tax    1   273.619 11081 1585.8
## <none>               11355 1596.1
## + indus  1    33.894 11321 1596.6
## + age    1     0.096 11355 1598.1
##
## Step:  AIC=1585.76
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim + rad + tax
##
##         Df Sum of Sq   RSS    AIC
## <none>               11081 1585.8
## + indus  1   2.51754 11079 1587.7
## + age    1   0.06271 11081 1587.8
```

```
#We did a built in forward selection function
```

In the forward selection model also, all the predictors are statistically significant except indus and age. So the

stepwise selection model which we got from forward selection is not different from what we had got through multivariate linear regression in question 4.

So we can confirm from two approaches(1. fittingg a multivariate regression model 2. using forward selection method) that all the predictors except indus and age are statistically significant. Let us try to fit a multivariate linear regression model without indus and age.

```
modelselect <- lm(medv ~ . -indus -age, data = Boston)
summary(modelselect)
```

```
##
## Call:
## lm(formula = medv ~ . - indus - age, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## zn            0.045845   0.013523   3.390 0.000754 ***
## chas          2.718716   0.854240   3.183 0.001551 **
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## rad           0.299608   0.063402   4.726 3.00e-06 ***
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## black         0.009291   0.002674   3.475 0.000557 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

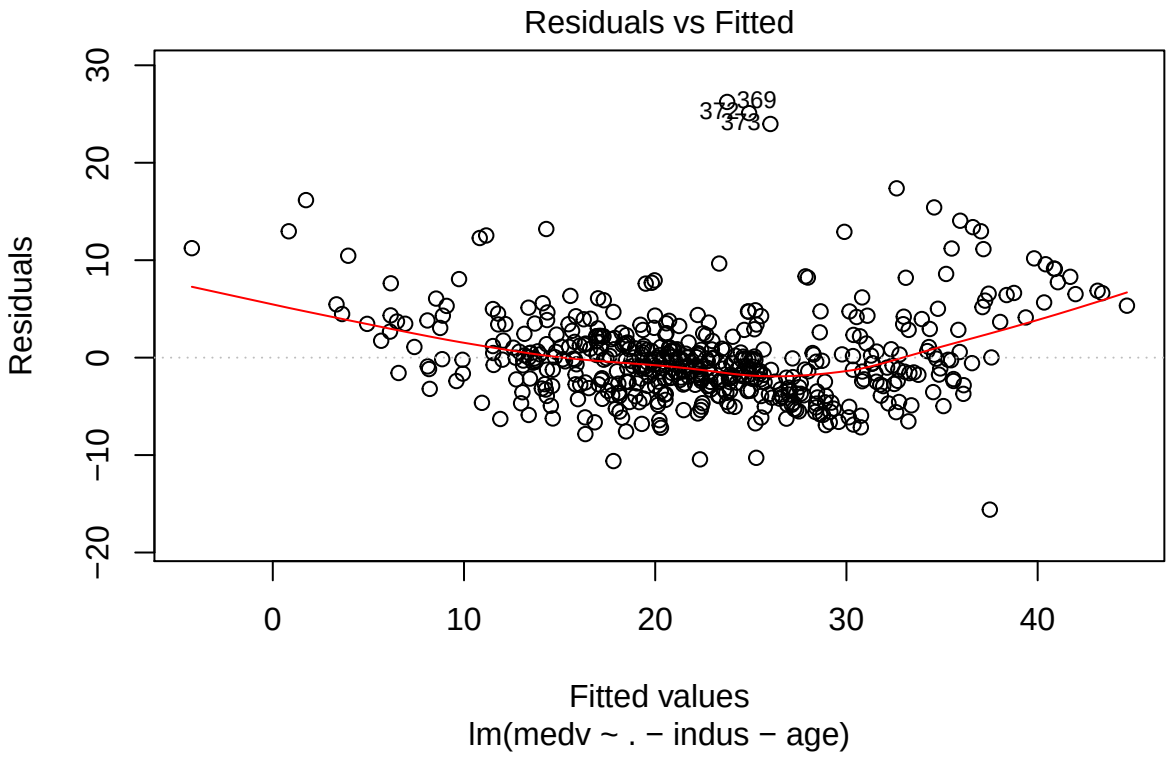
The Adjusted R-squared = 0.73, which means our model can explain 73% of the variability.

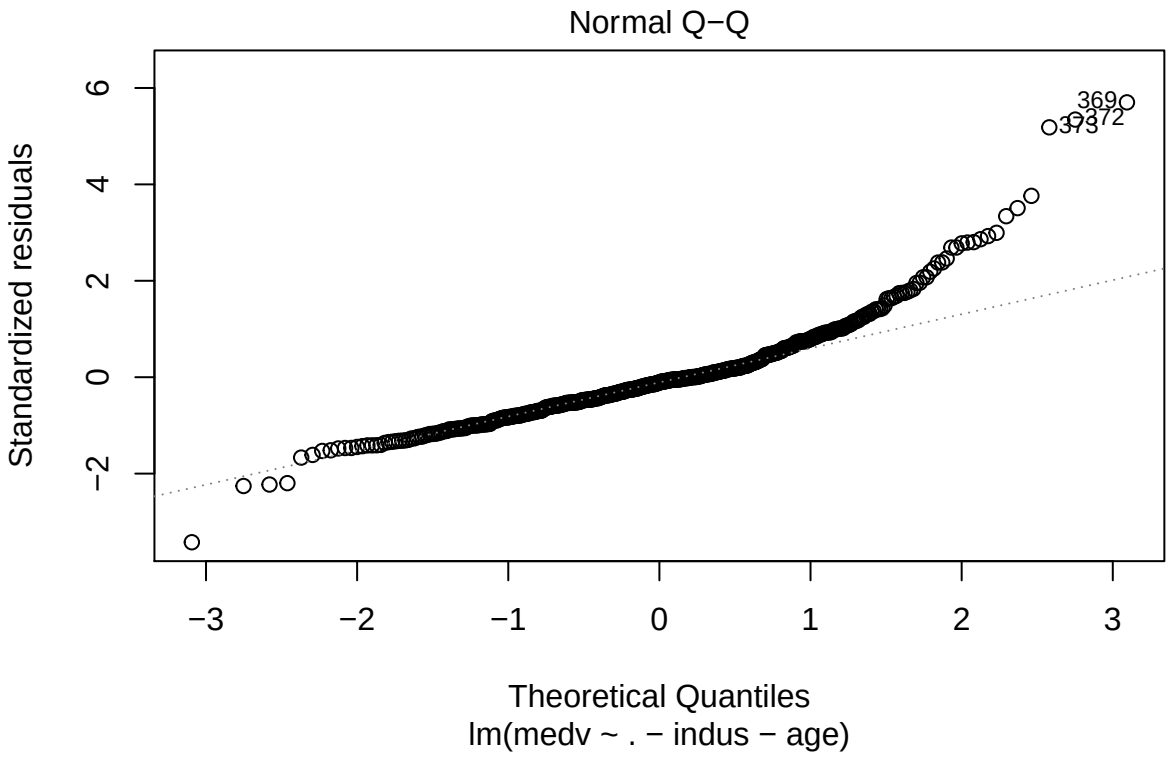
Question 8: Evaluate the statistical assumptions in your regression analysis from (7) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

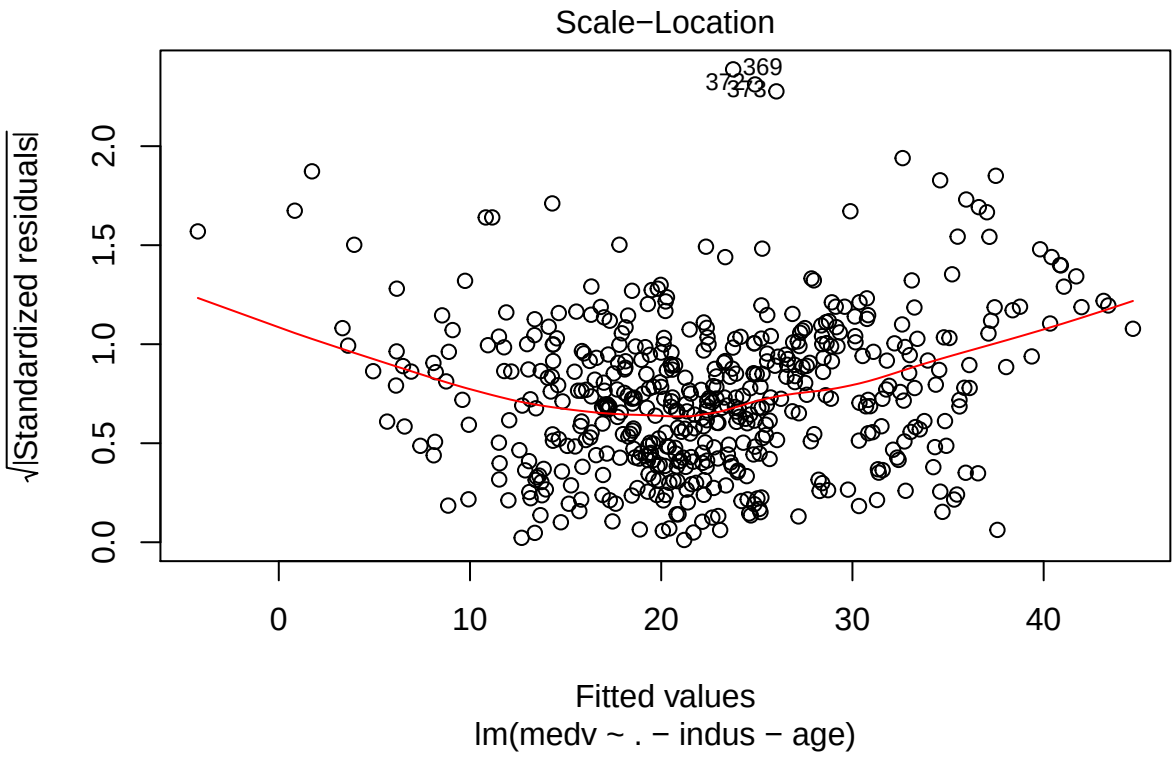
Answer: The following are the assumtions for multivariate regression analysis: 1. the residuals of the model are nearly normal, 2. the variability of the residuals is nearly constant, 3. the residuals are independent, and 4. each variable is linearly related to the outcome.

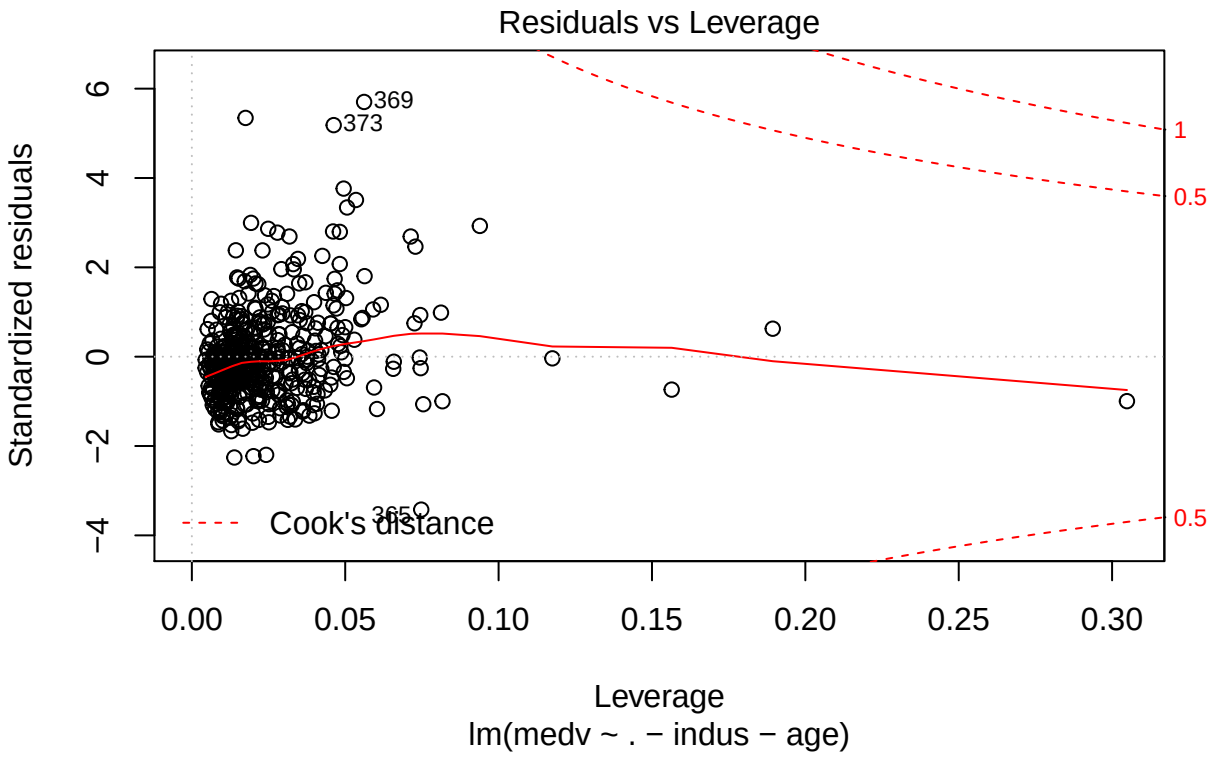
Let us try to make residual plots for our selected model(without indus and age).

```
plot(modelselect)
```


Residuals vs Fitted
Residuals
Fitted values
lm(medv ~ . − indus − age)
369
373
30
20
10
0
−10
−20
0
10
20
30
40


Normal Q−Q
Standardized residuals
6
4
2
0
−2
369
372
373
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(medv ~ . − indus − age)


Scale−Location
369
372
373
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Fitted values
lm(medv ~ . − indus − age)




```
#We get tge diagnostic plots using the built in plot function for our model
```

From the Residuals vs Fitted plot, we can find that the red line is depicting the pattern of the residuals. This proves that the condition of linearity is not fully met. From the normal Q-Q plot, we can see that the residuals are not totally lying on the line on which they should have been if they were normal, so our residuals are not fully normal and have strong skew. Our assumptions are not met. So our linear model may not be sufficient to predict the outcome variable which is medv, there might be some predictors which follow non linear relationship.