INFX 573: Problem Set 6 - Regression

Meeta Pandit

Due: Tuesday, November 15, 2016

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset6.Rmd` file from Canvas. Open `problemset6.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset6.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.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. 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 R Markdown file to `YourLastName_YourFirstName_ps6.Rmd`, knit a 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)
```

```
## Warning: package 'tidyverse' was built under R version 3.2.5

## Warning: package 'ggplot2' was built under R version 3.2.5

## Warning: package 'tibble' was built under R version 3.2.5

## Warning: package 'tidyr' was built under R version 3.2.5

## Warning: package 'readr' was built under R version 3.2.5

## Warning: package 'purrr' was built under R version 3.2.5

## Warning: package 'dplyr' was built under R version 3.2.5
```

```
library(MASS) # Modern applied statistics functions
```

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.

```r
#data(package = "MASS")

boston_data <- Boston
#?Boston

#str(boston_data)
```

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

   The data set contains information about varios parameters about Boston's population and locality. For example, the per capita crime rate by town,proportion of residential land, proportion of non-retail business per town, if the tract occupies charles river or not (boolean variable), nitrogen oxide concetration,average numbe of rooms in an apartment, proportionof owner-occupied units, weighted mean of distances to five Boston employment centers, index of accessibility to radial highways, property-tax rate per $10,000, pupil-teacher ratio by town, proportion of blacks by town, lower status of the population and median value of owner-occupied homes.

2. Consider this data in context, what is the response variable of interest? Discuss how you think some of the possible predictor variables might be associated with this response.

   The response variable in the context of this data would be median value of owner-occupied homes in $1000s (medv). With the help of other variables like proportion of owner-occupied units, weighted mean of distances to 5 Boston employment centres, pupil-teacher ratio by town, proportion of blacks by town, median value of owner occupied homes, lower status of the population (percent), proportion of non-retail business acres per town, per capita crime rate by town. Here, we can find an association of how median value of owner-occupied homes increases or decreases if we consider proportion of owner-occupied units, weighted mean of distances to 5 Boston towns etc as predictor variables.

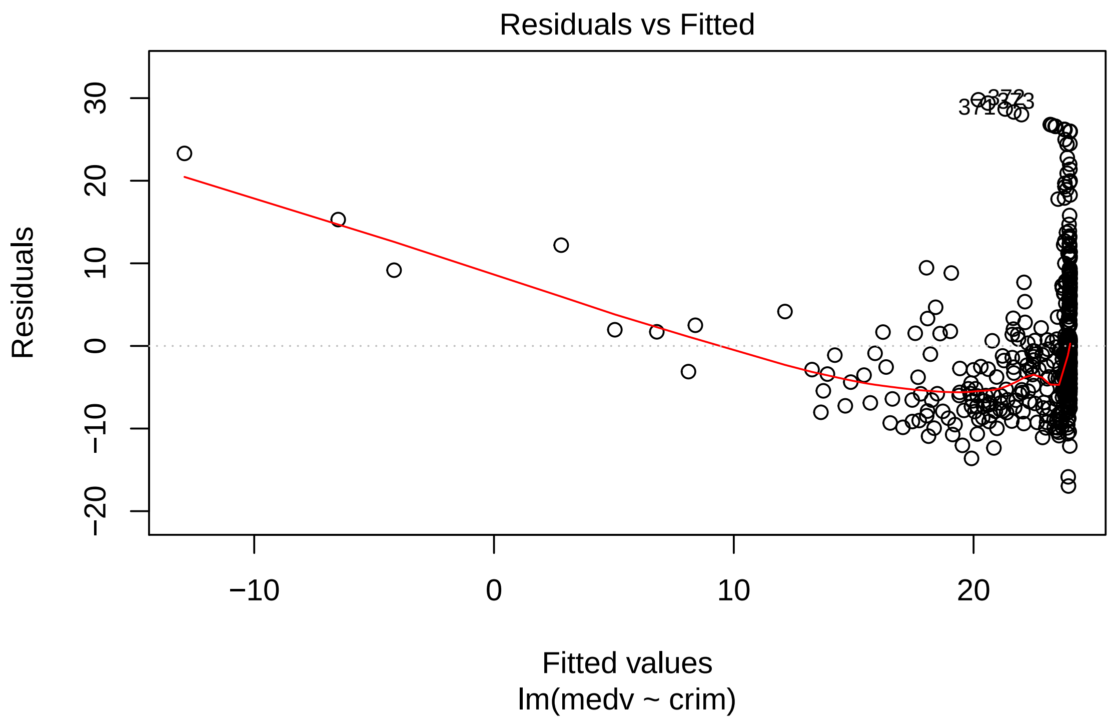
3. For each predictor, fit a simple linear regression model to predict the response. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

   Let's consider the association between median value of owner-occupied homes in $1000s as the response variable and crime rate by town as the predictor varaible

   ```r
   lm_crim <- lm(medv~crim, data = boston_data)
   summary(lm_crim)
   ```

   ```
   ## 
   ## Call:
   ## lm(formula = medv ~ crim, data = boston_data)
   ## 
   ## 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
   ```

```r
plot(lm_crim, which = c(1))
```



The visualization indicates that there is not a linear relationship between crim and medv. The data points are not bouncing randomly across the linear regression line as there is no sero line due to the uneven nature of the points. But, there is definitely a significant association between median value of homes and crime rate as the p-value is much less than 0.05.

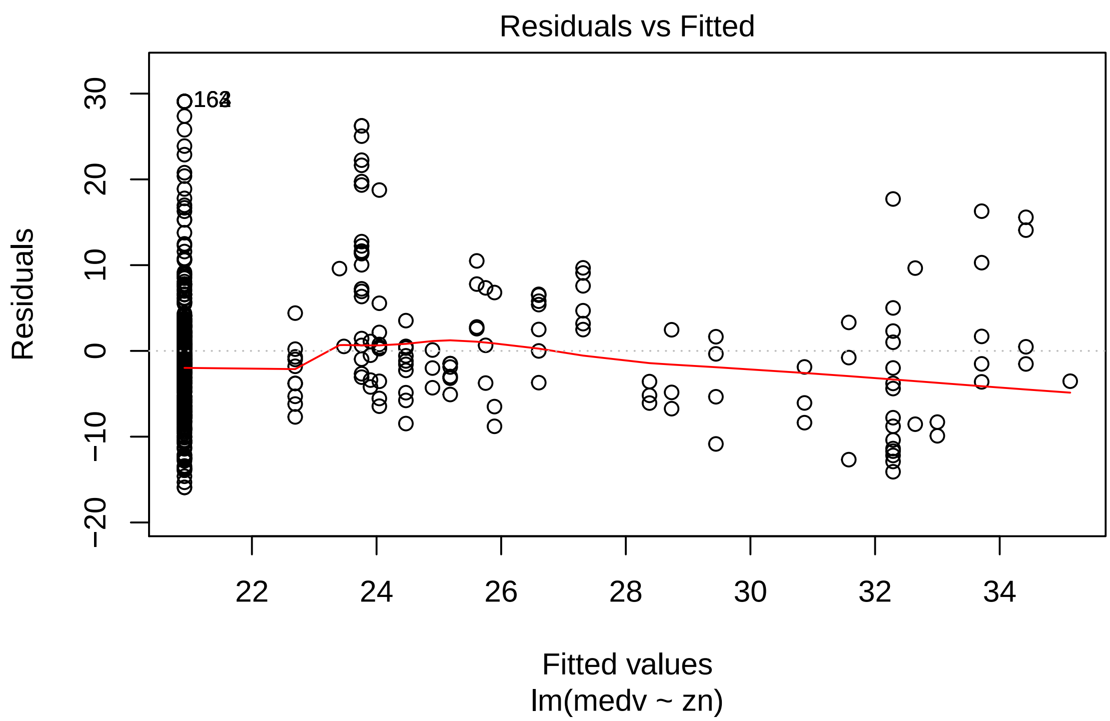
We can also analyze the relationship between median value of homes and proportion of residential land zoned for lots over 25000 sq.ft.

```r
lm_zn <- lm(medv~zn, data = boston_data)
summary(lm_zn)
```

```
##
## Call:
## lm(formula = medv ~ zn, data = boston_data)
##
## 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
```

```r
plot(lm_zn, which = c(1))
```



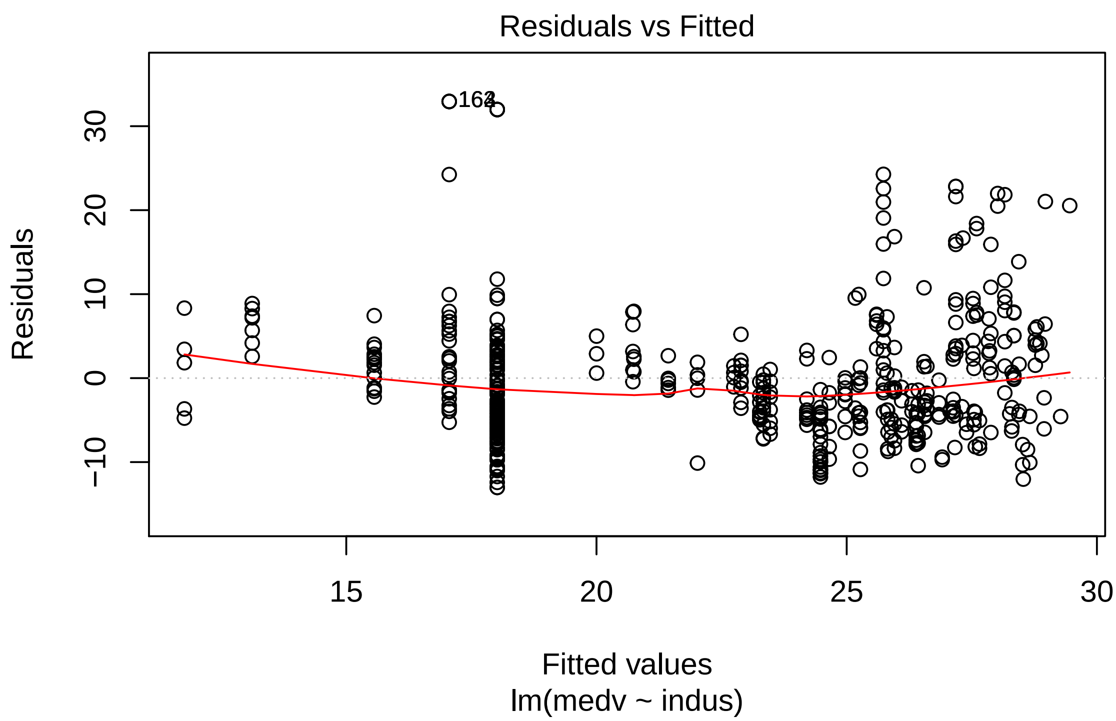
The reiduals vs fitted values graph indicates that the median value of homes and proportion of residential land zoned for lots are not linearly related as for every corresponding fitted value there is not a residual error to nullify the effect of that error. Also, the regression line is not linear. But, the the p-value is much less than 0.05 which indicates that there is a significant relationship between medv and zn but not linear.

```r
lm_indus <- lm(medv~indus, data = boston_data)
summary(lm_indus)
```

```
##
## Call:
## lm(formula = medv ~ indus, data = boston_data)
##
## 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
```

```r
plot(lm_indus, which = c(1))
```

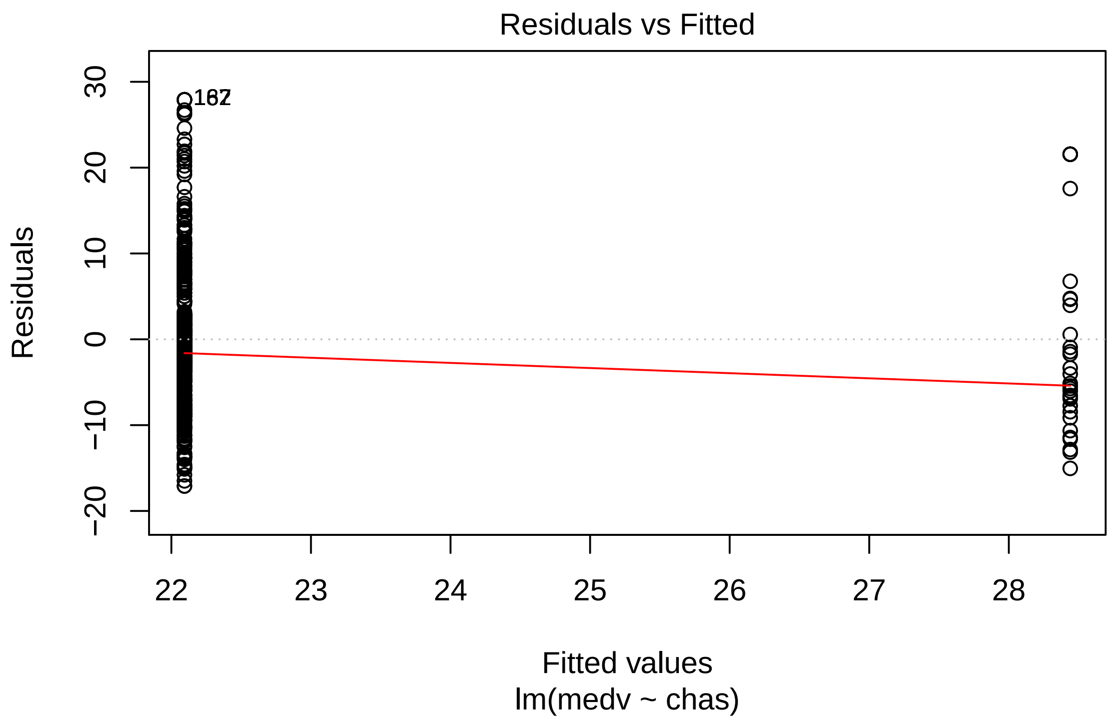


The shape of the regression line indicates that there is a non-linear relationship between median values of homes and proportion of non-retail business acres. Also, there are some outliers which do not have corresponding residuals below the trend line.

```r
lm_chas <- lm(medv~chas, data = boston_data)
summary(lm_chas)
```

```
##
## Call:
## lm(formula = medv ~ chas, data = boston_data)
##
## 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
plot(lm_chas, which = c(1))
```

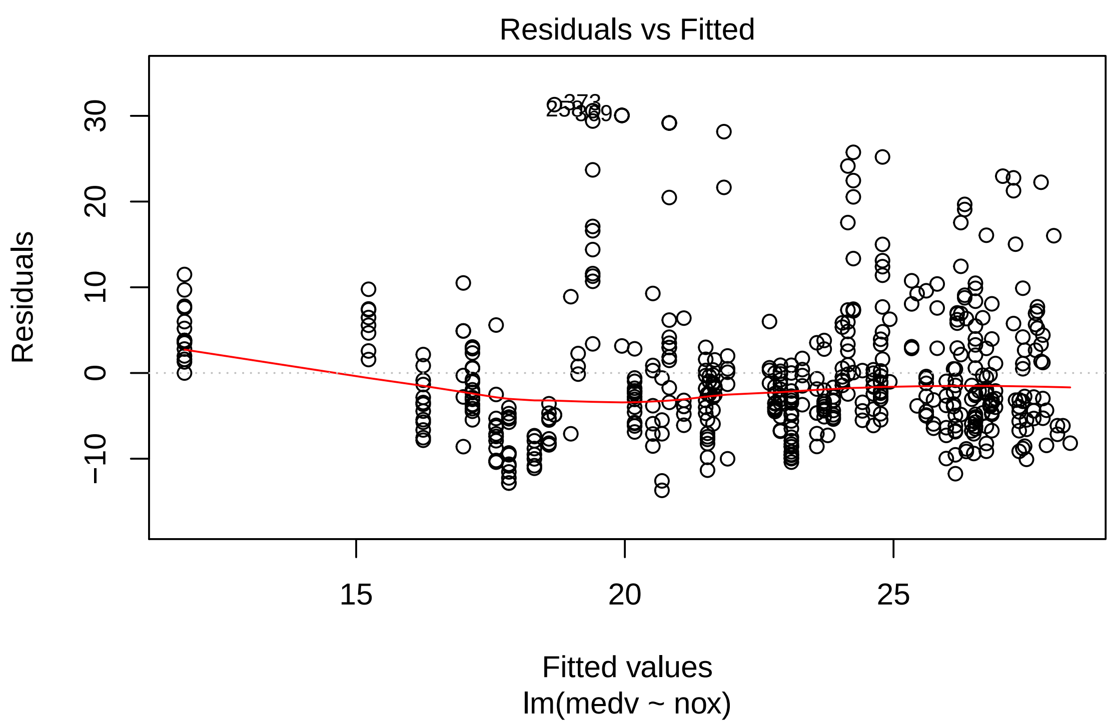


Fitting a linear model for variables with a Bernoulli distribution is viable. For example, one unit increase in a berboulli variable is either a 1 or 0 and there are no continuous values for such variables. So, it is not helpful to analyze such variables using linear regression models, logistic models fit best for Bernoulli distributed variables.

```r
lm_nox <- lm(medv~nox, data = boston_data)
summary(lm_nox)
```

```
##
## Call:
## lm(formula = medv ~ nox, data = boston_data)
##
## 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
plot(lm_nox, which = c(1))
```



The fitted values are scattered all over the graph and they are not evenly distributed across the trend line. Also, the trend line is not a zero line and hence does not satisfy linear relationship criterion. But, the summary statistics shows there is a significant relationship between median value of homes and nitrogen oxides concentration as the p-value is much less than 0.05 may not be linear.

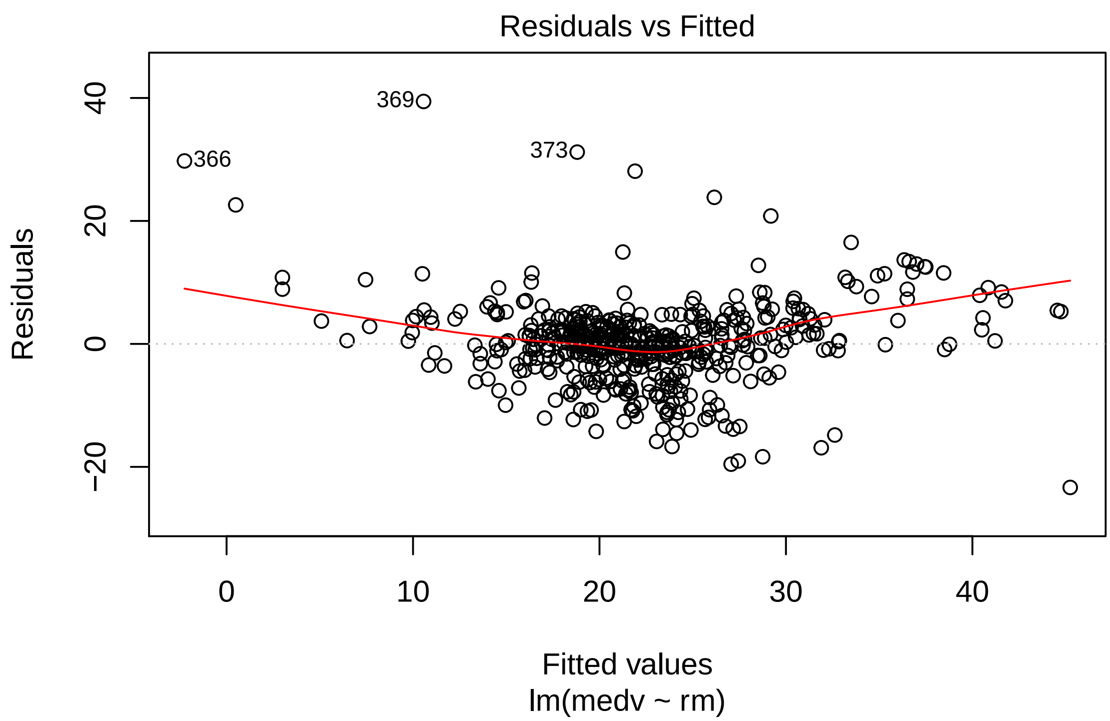
Let's also consider the relationship between median value of homes and average number of rooms per dwelling.

```r
lm_rm <- lm(medv~rm, data = boston_data)
summary(lm_rm)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = boston_data)
##
## 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
plot(lm_rm, which = c(1))
```



The points are spread unevenly across the trend line. Each point does not have a corresponding residual value to compensate for the error term. Hence, there is not a linear relationship between the median values of homes and average number of rooms per dwelling.

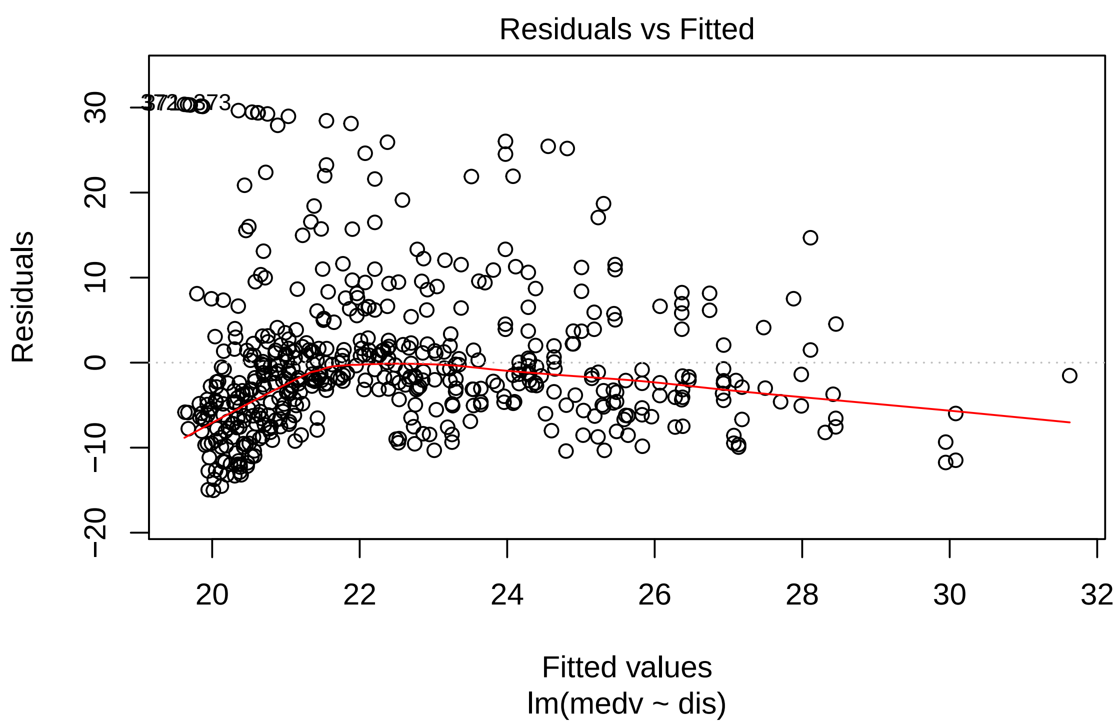
Let's consider another redictor variable dis which is the weighted mean of distances to 5 Boston employment centres.

```r
lm_dis <- lm(medv~dis, data = boston_data)
summary(lm_dis)
```

```
##
## Call:
## lm(formula = medv ~ dis, data = boston_data)
##
## 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
```

```r
plot(lm_dis, which = c(1))
```



The regression line is not linear and the points are scattered across the line unevenly. Hence, there is ni linear relationship between the median value of homes and weighted mean of distances to 5 Boston employment centres. Also, there are outliers marked by the row numbers in the graph.

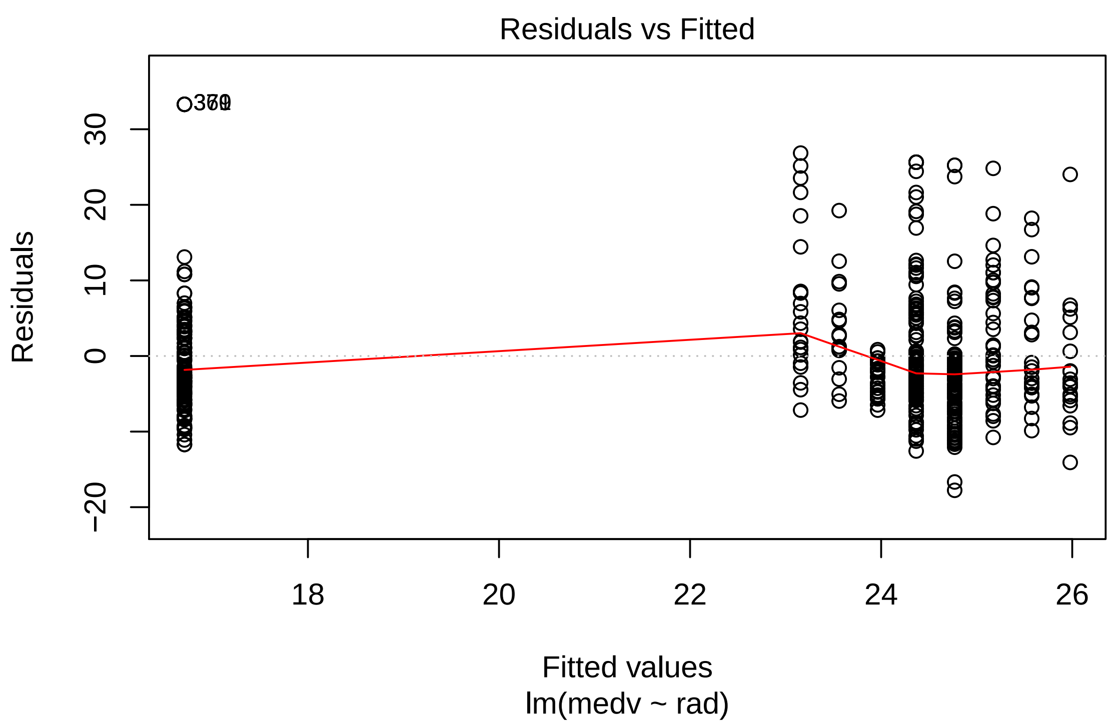
Let's also consider the median value of homes to accessibility to radial highways.

```r
lm_rad <- lm(medv~rad, data = boston_data)
summary(lm_rad)
```

```
##
## Call:
## lm(formula = medv ~ rad, data = boston_data)
##
## 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
```

```r
plot(lm_rad, which = c(1))
```



The curve indicates a non-linear relationship between the 2 variables and also the data points are scattered across the curve with no corresponding residuals on both sides of the curve.

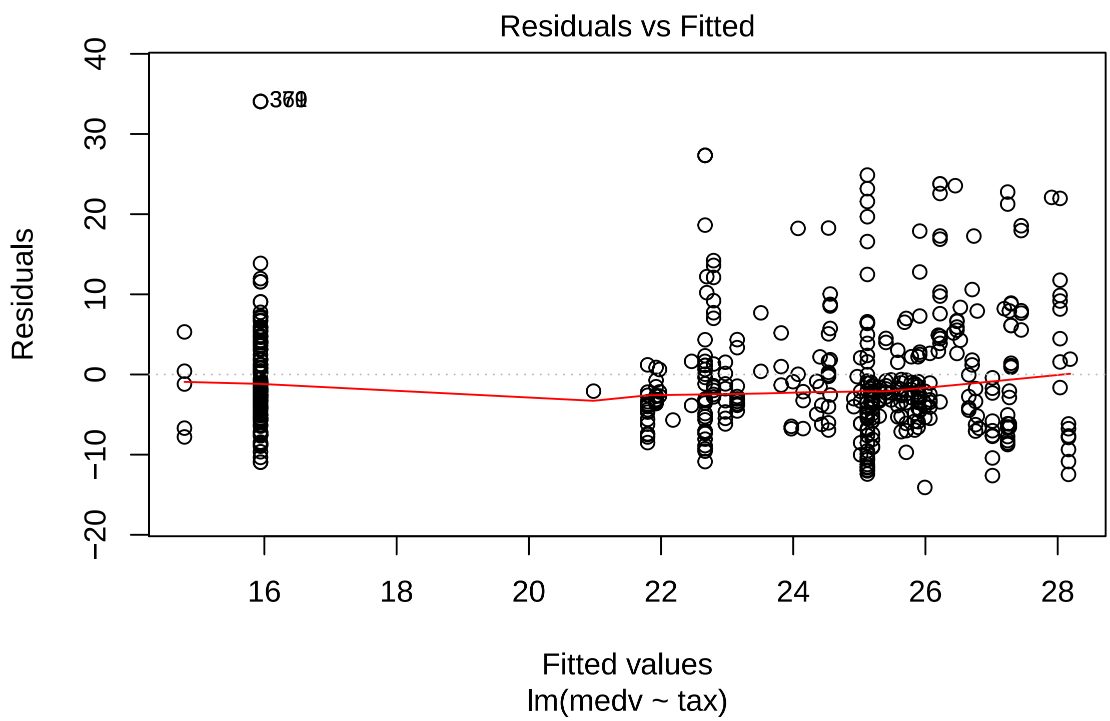
We will also consider the effect of full-value property tax on median value of homes.

```r
lm_tax <- lm(medv~tax, data = boston_data)
summary(lm_tax)
```

```
##
## Call:
## lm(formula = medv ~ tax, data = boston_data)
##
## 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
```

```r
plot(lm_tax, which = c(1))
```



The residuals vs fitted values graph shows that the regression line is almost linear but with some amount of non-linearity due to the uneven distribution of data points.

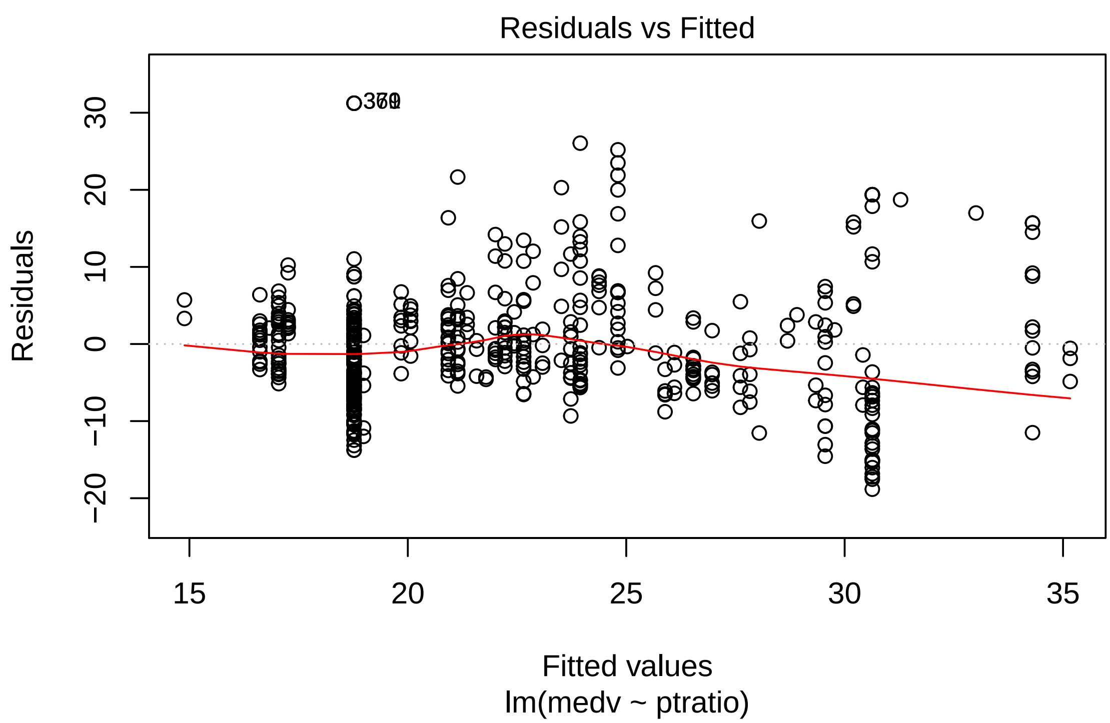
Relationship between pupil-teacher ratio and median value of homes

```r
lm_ptratio <- lm(medv~ptratio, data = boston_data)
summary(lm_ptratio)
```

```
##
## Call:
## lm(formula = medv ~ ptratio, data = boston_data)
##
## 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
plot(lm_ptratio, which = c(1))
```



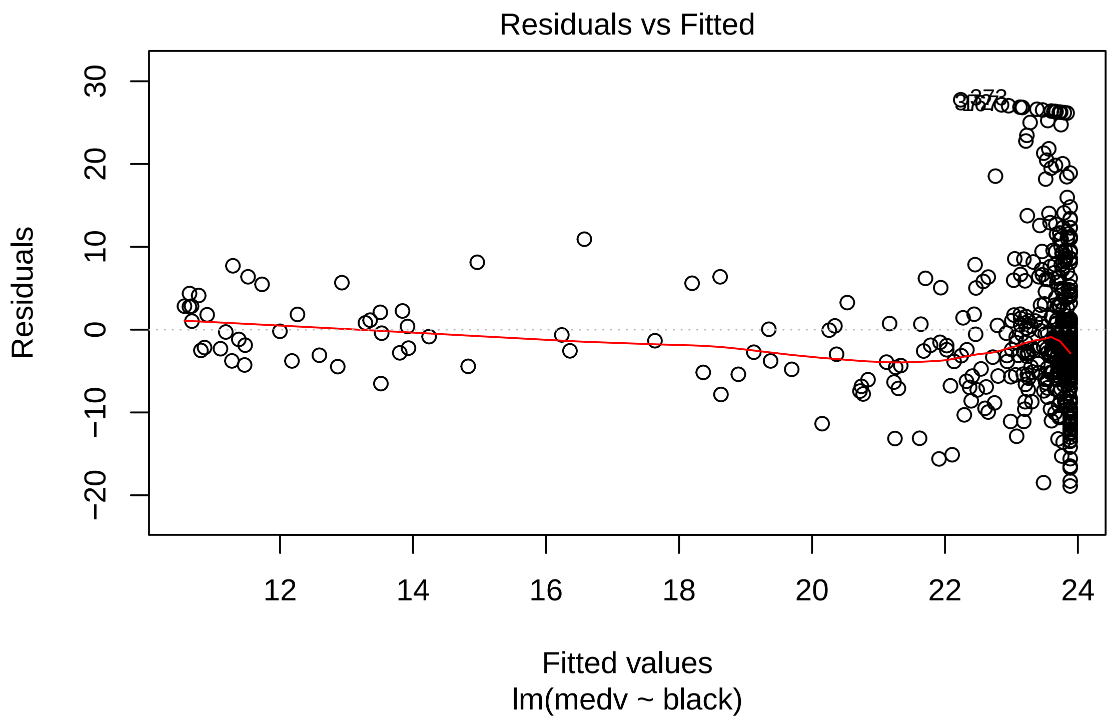
The relationship between median value of homes and pupil-teacher ration is clearly no-linear as the each ftted value does not have a residual value associated with it on the other side of the trend line.

Relationship between proportion of blacks by town and median value of homes

```r
lm_black <- lm(medv~black , data = boston_data)
summary(lm_black)
```

```
##
## Call:
## lm(formula = medv ~ black, data = boston_data)
##
## 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
plot(lm_black, which = c(1))
```



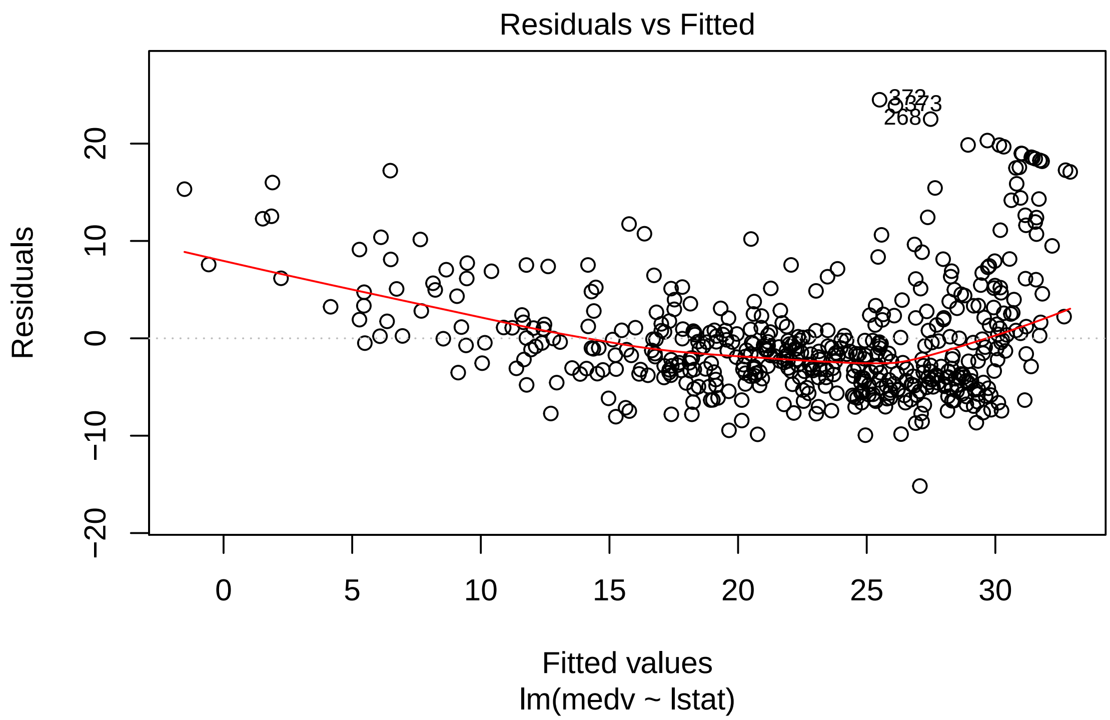
Most of the fitted values are clustered around one area and also they are not randomly bouncing across the trend line. Hence, there is no evidence of a linear relationship between the median value of homes and proportion of black population by town.

Lower status of population and median value of homes

```r
lm_lstat <- lm(medv~lstat, data = boston_data)
summary(lm_lstat)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = boston_data)
##
## 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
plot(lm_lstat, which = c(1))
```

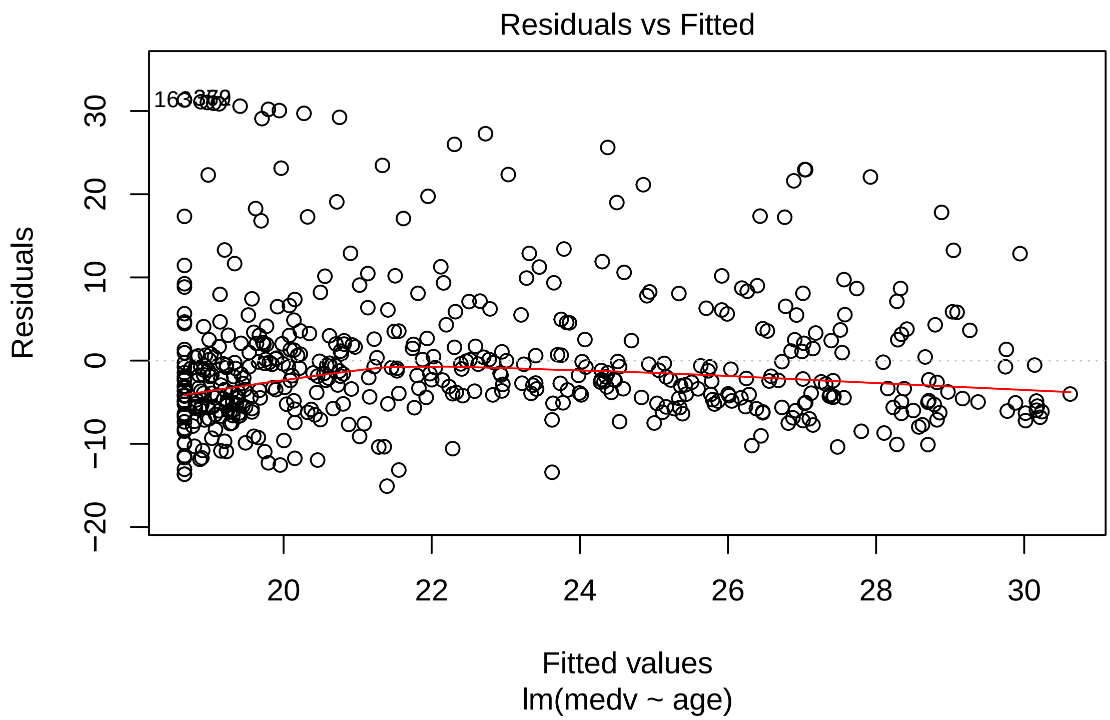


Again, the trend line indicates a non-linear relationship between the 2 variables and also there are some outliers with no corresponding residuals to nulligy the effect of this error.

```r
lm_age <- lm(medv~age , data = boston_data)
summary(lm_age)
```

```
##
## Call:
## lm(formula = medv ~ age, data = boston_data)
##
## 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
```

```r
plot(lm_age, which = c(1))
```



The points are randomly bouncing around the trend line but there are outliers which have no corresponding residual error value on the other side of the trend line. Thus, they do not staisfy the criterion for linear relationship.

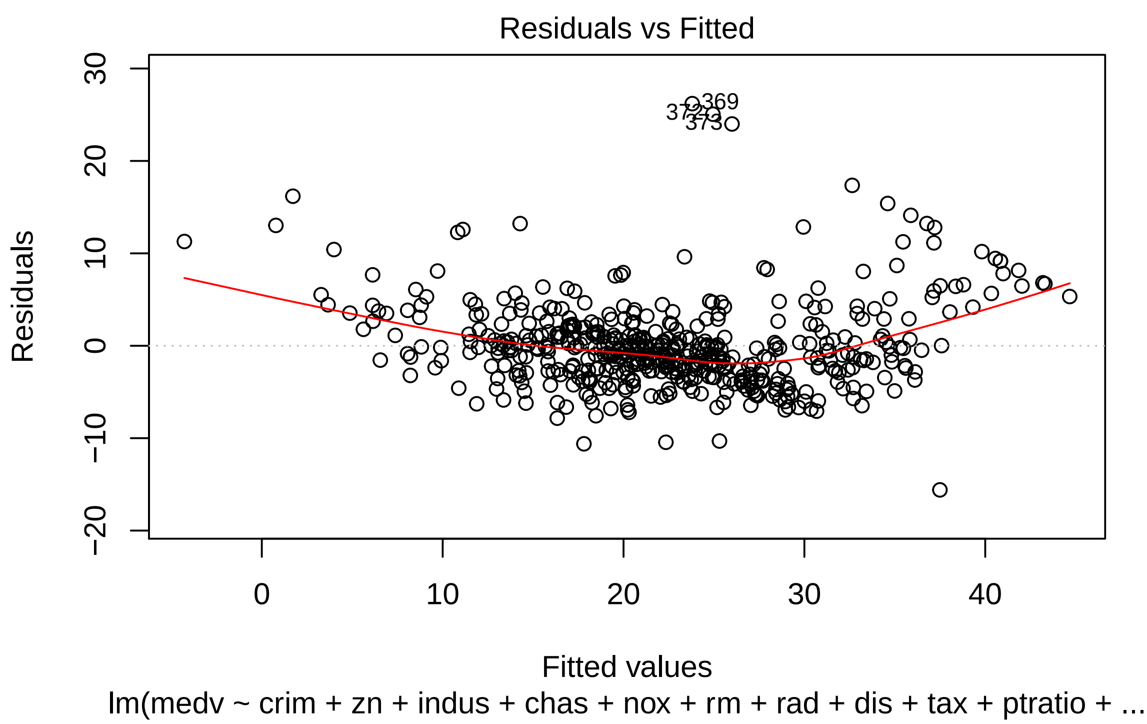
4. Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```r
lm_multiple <- lm(medv~crim+zn+indus+chas+nox+rm+rad+dis+tax+
                    ptratio+lstat+black+age, data = boston_data)
summary(lm_multiple)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + rad +
##     dis + tax + ptratio + lstat + black + age, data = boston_data)
##
## 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 ***
## rad           3.060e-01  6.635e-02   4.613 5.07e-06 ***
## dis          -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## tax          -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio      -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## lstat        -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## black         9.312e-03  2.686e-03   3.467 0.000573 ***
## age           6.922e-04  1.321e-02   0.052 0.958229
## ---
## 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
```

```
plot(lm_multiple, which = c(1))
```



The multiple regression model shows that we can reject the null hypothesis for all those predictors whose p-value is much less than 0.05 which means there is a significant relationship between the response and predictor. But, the residuals vs fitted values graph shows that there is no linear relationship between the response and all the predictor variables as the trend line is not linear which indicates that there are certain data points which do not have a corresponding match of residuals to nullify the elemnt of error. Also, there are some outliers which refrain the model from following the linearity priciples.

Thus, after analyzing the summary statistics of multiple regression model, we can safely reject the null hypothesis for the following variables: crim, zn, chas, nox, dis, ptratio, rad, rm, tax, lstat and black.

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.

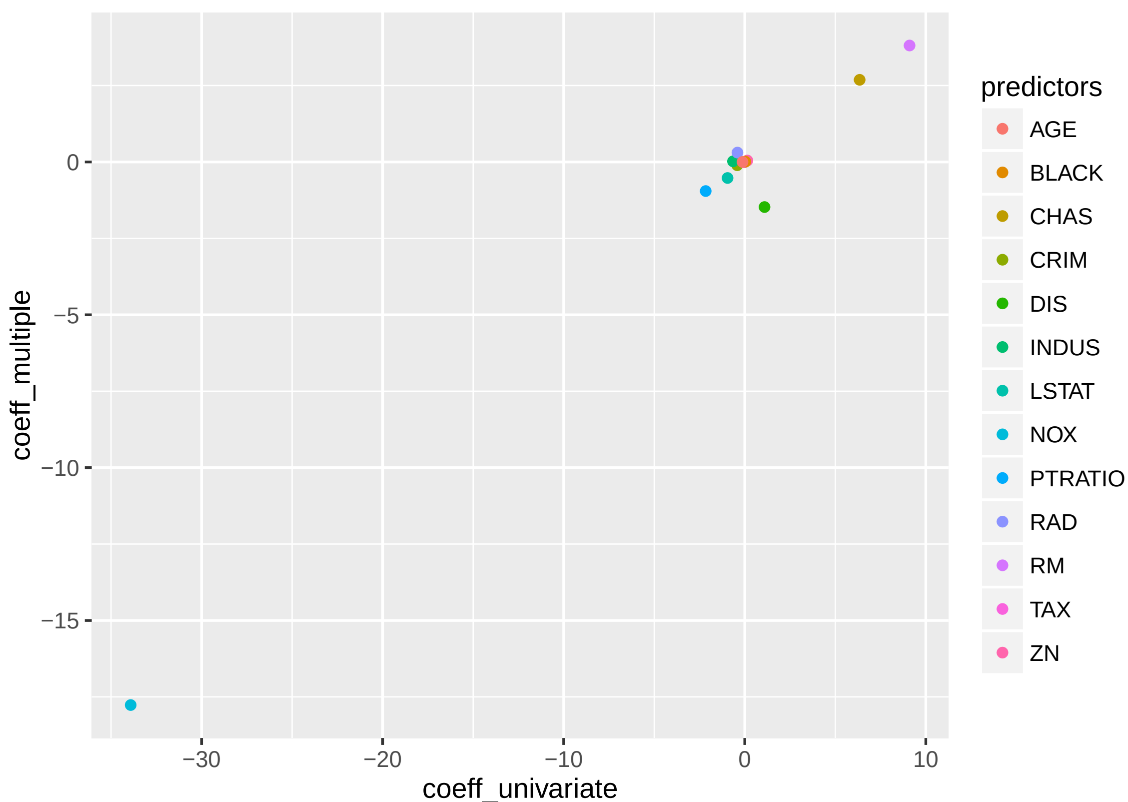
```
#Create data frame for the coefficients of linear and multiple regression
predictors <- c('CRIM', 'ZN', 'INDUS', 'CHAS','NOX', 'RM','RAD','DIS',
                'TAX','PTRATIO', 'LSTAT', 'BLACK', 'AGE')

coeff_univariate <- c(coefficients(lm_crim)[2], coefficients(lm_zn)[2],
                     coefficients(lm_indus)[2], coefficients(lm_chas)[2],
                     coefficients(lm_nox)[2],
                      coefficients(lm_rm)[2], coefficients(lm_rad)[2],
                     coefficients(lm_dis)[2],coefficients(lm_tax)[2],
                     coefficients(lm_ptratio)[2], coefficients(lm_lstat)[2],
                     coefficients(lm_black)[2],coefficients(lm_age)[2])

coeff_multiple <- c(coefficients(lm_multiple)[2:14])

coeff_df <- data.frame(predictors, coeff_univariate, coeff_multiple)

ggplot(coeff_df, aes(coeff_univariate, coeff_multiple, color = predictors))+
       geom_point()
```



When we observe the coefficients of univariate analysis and multivariate analysis in the intermediate table coeff_df, it is evident that in multiple regression model the coefficient values decrease as compared to univariate regression model. From the graph it is visible that all the coefficients that have higher values in univariate analysis shows reduced coefficient values in the graph.

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:

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

```r
#Non-linear regression models
nlm_crim <- lm(medv~crim+I(crim^2)+I(crim^3), data = boston_data)
summary(nlm_crim)
```

```
##
## Call:
## lm(formula = medv ~ crim + I(crim^2) + I(crim^3), data = boston_data)
##
## 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)  2.519e+01  4.355e-01  57.846  < 2e-16 ***
## crim        -1.136e+00  1.444e-01  -7.868 2.24e-14 ***
## I(crim^2)    2.378e-02  6.808e-03   3.494 0.000518 ***
## I(crim^3)   -1.489e-04  6.641e-05  -2.242 0.025411 *
## ---
## 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
```

```r
nlm_zn <- lm(medv~zn+I(zn^2)+I(zn^3), data = boston_data)
summary(nlm_zn)
```

```
##
## Call:
## lm(formula = medv ~ zn + I(zn^2) + I(zn^3), data = boston_data)
##
## 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) 20.4485972  0.4359536  46.905  < 2e-16 ***
## zn           0.6433652  0.1105611   5.819 1.06e-08 ***
## I(zn^2)     -0.0167646  0.0038872  -4.313 1.94e-05 ***
## I(zn^3)      0.0001257  0.0000316   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
```

```r
nlm_indus <- lm(medv~indus+I(indus^2)+I(indus^3), data = boston_data)
summary(nlm_indus)
```

```
##
```

```
## Call:
## lm(formula = medv ~ indus + I(indus^2) + I(indus^3), data = boston_data)
##
## 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) 37.080160   1.663326  22.293  < 2e-16 ***
## indus       -2.806994   0.509349  -5.511 5.71e-08 ***
## I(indus^2)   0.140462   0.041554   3.380 0.000781 ***
## I(indus^3)  -0.002399   0.001011  -2.373 0.018026 *
## ---
## 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
```

```
nlm_chas <- lm(medv~chas+I(chas^2)+I(chas^3), data = boston_data)
summary(nlm_chas)
```

```
##
## Call:
## lm(formula = medv ~ chas + I(chas^2) + I(chas^3), data = boston_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients: (2 not defined because of singularities)
##             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 ***
## 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: 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
```

```
nlm_nox <- lm(medv~nox+I(nox^2)+I(nox^3), data = boston_data)
summary(nlm_nox)
```

```
##
## Call:
## lm(formula = medv ~ nox + I(nox^2) + I(nox^3), data = boston_data)
##
## 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)   -22.49      38.52  -0.584   0.5596
## nox           315.10     195.10   1.615   0.1069
## I(nox^2)     -615.83     320.48  -1.922   0.0552 .
## I(nox^3)      350.19     170.92   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
```

```r
nlm_rm <- lm(medv~rm+I(rm^2)+I(rm^3), data = boston_data)
summary(nlm_rm)
```

```
##
## Call:
## lm(formula = medv ~ rm + I(rm^2) + I(rm^3), data = boston_data)
##
## 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)  241.3108    47.3275   5.099 4.85e-07 ***
## rm          -109.3906    22.9690  -4.763 2.51e-06 ***
## I(rm^2)       16.4910     3.6750   4.487 8.95e-06 ***
## I(rm^3)       -0.7404     0.1935  -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
```

```r
nlm_age <- lm(medv~age+I(age^2)+I(age^3), data = boston_data)
summary(nlm_age)
```

```
##
## Call:
## lm(formula = medv ~ age + I(age^2) + I(age^3), data = boston_data)
##
## 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)  2.893e+01  2.992e+00   9.668   <2e-16 ***
## age         -1.224e-01  2.014e-01  -0.608    0.544
## I(age^2)     2.355e-03  3.930e-03   0.599    0.549
## I(age^3)    -2.318e-05  2.279e-05  -1.017    0.310
## ---
```

```
## 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
```

```
nlm_dis <- lm(medv~dis+I(dis^2)+I(dis^3), data = boston_data)
summary(nlm_dis)
```

```
##
## Call:
## lm(formula = medv ~ dis + I(dis^2) + I(dis^3), data = boston_data)
##
## 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)  7.03789    2.91134   2.417  0.01599 *
## dis          8.59284    2.06633   4.158 3.77e-05 ***
## I(dis^2)    -1.24953    0.41235  -3.030  0.00257 **
## I(dis^3)     0.05602    0.02428   2.307  0.02146 *
## ---
## 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
```

```
nlm_rad <- lm(medv~rad+I(rad^2)+I(rad^3), data = boston_data)
summary(nlm_rad)
```

```
##
## Call:
## lm(formula = medv ~ rad + I(rad^2) + I(rad^3), data = boston_data)
##
## 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) 30.251303   2.567860  11.781  < 2e-16 ***
## rad         -3.799454   1.307156  -2.907 0.003815 **
## I(rad^2)     0.616347   0.186057   3.313 0.000991 ***
## I(rad^3)    -0.020086   0.005717  -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
```

```
nlm_tax <- lm(medv~tax+I(tax^2)+I(tax^3), data = boston_data)
summary(nlm_tax)
```

```
##
## Call:
## lm(formula = medv ~ tax + I(tax^2) + I(tax^3), data = boston_data)
##
## 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)  5.222e+01  1.397e+01   3.739 0.000206 ***
## tax         -1.635e-01  1.133e-01  -1.443 0.149646
## I(tax^2)     3.029e-04  2.872e-04   1.055 0.292004
## I(tax^3)    -2.079e-07  2.236e-07  -0.930 0.353061
## ---
## 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
nlm_ptratio <- lm(medv~ptratio+I(ptratio^2)+I(ptratio^3), data = boston_data)
summary(nlm_ptratio)
```

```
##
## Call:
## lm(formula = medv ~ ptratio + I(ptratio^2) + I(ptratio^3), data = boston_data)
##
## 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)  312.28642  152.48693   2.048   0.0411 *
## ptratio      -48.69114   26.88441  -1.811   0.0707 .
## I(ptratio^2)   2.83995    1.56413   1.816   0.0700 .
## I(ptratio^3)  -0.05686    0.03005  -1.892   0.0590 .
## ---
## 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
nlm_black <- lm(medv~black+I(black^2)+I(black^3), data = boston_data)
summary(nlm_black)
```

```
##
## Call:
## lm(formula = medv ~ black + I(black^2) + I(black^3), data = boston_data)
##
## 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)  1.260e+01  2.517e+00   5.006  7.7e-07 ***
## black       -1.703e-02  6.150e-02  -0.277    0.782
## I(black^2)   2.036e-04  3.258e-04   0.625    0.532
## I(black^3)  -2.224e-07  4.765e-07  -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
```

```
nlm_lstat <- lm(medv~lstat+I(lstat^2)+I(lstat^3), data = boston_data)
summary(nlm_lstat)
```

```
##
## Call:
## lm(formula = medv ~ lstat + I(lstat^2) + I(lstat^3), data = boston_data)
##
## 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) 48.6496253  1.4347240  33.909  < 2e-16 ***
## lstat       -3.8655928  0.3287861 -11.757  < 2e-16 ***
## I(lstat^2)   0.1487385  0.0212987   6.983 9.18e-12 ***
## I(lstat^3)  -0.0020039  0.0003997  -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
```

On analyzing the summary statistics, it is observed that most of the predictors have the cubical coefficients which have significant p-values. Hence, the model cannot ignore these coefficients. Thus, if we retain the cubical coefficients then we have to keep the lower exponential coefficients. Thus, this indictaes that there is non-linear reationship betwwen the reposne that is median values of homes and other predictos. Though there are some predictors for which the quadratic and cubic coefficients are not very significant as their p-value is much higher. Thus, we cannot consider predictors like age, tax, and black to have even a non-linear relationship with the response.

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)?

```
#Stepwise model selection
step_model <- stepAIC(lm_multiple, direction = "both")
```

```
## Start:  AIC=1589.64
## medv ~ crim + zn + indus + chas + nox + rm + rad + dis + tax +
##     ptratio + lstat + black + age
##
```

```
##           Df Sum of Sq   RSS    AIC
## - age      1      0.06 11079 1587.7
## - indus    1      2.52 11081 1587.8
## <none>                 11079 1589.6
## - chas     1    218.97 11298 1597.5
## - tax      1    242.26 11321 1598.6
## - crim     1    243.22 11322 1598.6
## - zn       1    257.49 11336 1599.3
## - black    1    270.63 11349 1599.8
## - rad      1    479.15 11558 1609.1
## - nox      1    487.16 11566 1609.4
## - ptratio  1   1194.23 12273 1639.4
## - dis      1   1232.41 12311 1641.0
## - rm       1   1871.32 12950 1666.6
## - lstat    1   2410.84 13490 1687.3
##
## Step:  AIC=1587.65
## medv ~ crim + zn + indus + chas + nox + rm + rad + dis + tax +
##     ptratio + lstat + black
##
##           Df Sum of Sq   RSS    AIC
## - indus    1      2.52 11081 1585.8
## <none>                 11079 1587.7
## + age      1      0.06 11079 1589.6
## - chas     1    219.91 11299 1595.6
## - tax      1    242.24 11321 1596.6
## - crim     1    243.20 11322 1596.6
## - zn       1    260.32 11339 1597.4
## - black    1    272.26 11351 1597.9
## - rad      1    481.09 11560 1607.2
## - nox      1    520.87 11600 1608.9
## - ptratio  1   1200.23 12279 1637.7
## - dis      1   1352.26 12431 1643.9
## - rm       1   1959.55 13038 1668.0
## - lstat    1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## medv ~ crim + zn + chas + nox + rm + rad + dis + tax + ptratio +
##     lstat + black
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## + indus    1      2.52 11079 1587.7
## + age      1      0.06 11081 1587.8
## - chas     1    227.21 11309 1594.0
## - crim     1    245.37 11327 1594.8
## - zn       1    257.82 11339 1595.4
## - black    1    270.82 11352 1596.0
## - tax      1    273.62 11355 1596.1
## - rad      1    500.92 11582 1606.1
## - nox      1    541.91 11623 1607.9
## - ptratio  1   1206.45 12288 1636.0
## - dis      1   1448.94 12530 1645.9
## - rm       1   1963.66 13045 1666.3
```

```
## - lstat    1   2723.48 13805 1695.0
```

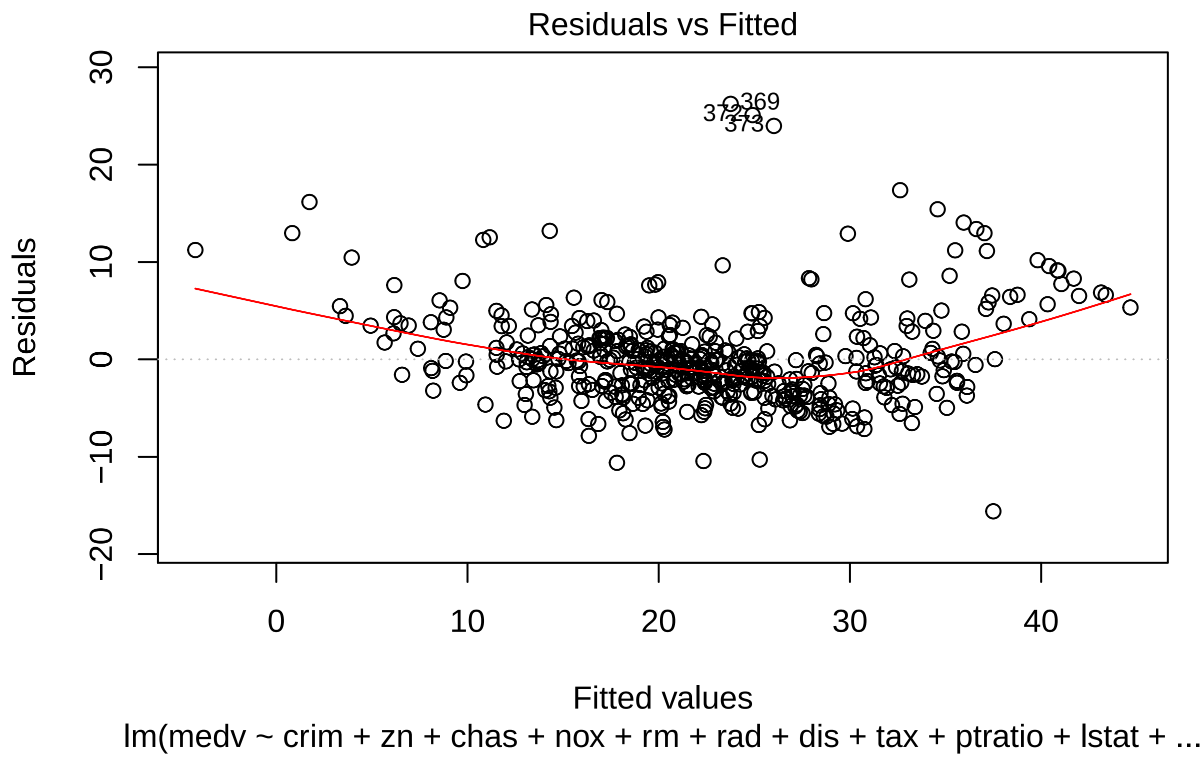
```
step_model$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## medv ~ crim + zn + indus + chas + nox + rm + rad + dis + tax +
##     ptratio + lstat + black + age
##
## Final Model:
## medv ~ crim + zn + chas + nox + rm + rad + dis + tax + ptratio +
##     lstat + black
##
##
##      Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                             492   11078.78 1589.643
## 2   - age  1 0.06183435       493   11078.85 1587.646
## 3 - indus  1 2.51754013       494   11081.36 1585.761
```

From the final step of stepwise model selection, does not consider the variables indus and age to generate best fir multiple regression model. This is because we start with all variables and the AIC value is 1589.64. Then we try removing age and the AIC value decreases to 1587. Then we try eliminating indus and the value increases a bit. Hence, in the next step we start with the best least AIC value got so far which is 1587.65 and eliminate age from this step. Now, when we eliminate indus, the value of AIC goes down further to 1585. Also, in the same step when we try to add indus it increases. So, in this step we got the least AIC of 1585. In the final step, we start with this AIC value obtained and try to add the removed values again. BUt, it is observed that AIC values goes on increasing. Hence, the best fit model has the least AIC when we eliminate age and indus.

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.

```r
#Plot of residuals vs. fitted for the stepwise model
plot(step_model, which = c(1))
```



The residuals vs. fitted values graph shows that there is a non-linear relationship between the response and the predictors. This is because the regression line is not coinciding with the 0 line in the above plot. This means the fitted values are not randomly bouncing arund the 0-line. Also, there are many outliers in the graph which do not have a corresponding residual value to nullify the effect of error introduced in the model.