INFX 573: Problem Set 6 - Regression

Aakash Bang

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)
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.

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

   The Boston dataset contains data about the housing values in suburbs of Boston. Variables - 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. ltstat - lower status of the population (percent). medv - median value of owner-occupied homes in \$1000s

```r
#1

BostonData <- MASS::Boston

# Change column names
colnames(BostonData) <- c("Crime_Rate", "Zoned_Land", "Indus", "Tract_Bound", "NOX",
"Avg_Rooms", "Owner_Occupied", "Distance", "Rad",
"Tax", "PTRatio", "Blacks", "Lower_Status", "Median_Value")
```

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.

   Response variable - median value of owner-occupied homes

   Possible Predictor Variables -

```r
#2
fit = lm(log(Median_Value) ~ ., data = BostonData)
summary(fit)
```

```
##
## Call:
## lm(formula = log(Median_Value) ~ ., data = BostonData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73361 -0.09747 -0.01657  0.09629  0.86435
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.1020423  0.2042726  20.081  < 2e-16 ***
## Crime_Rate     -0.0102715  0.0013155  -7.808 3.52e-14 ***
## Zoned_Land      0.0011725  0.0005495   2.134 0.033349 *
## Indus           0.0024668  0.0024614   1.002 0.316755
## Tract_Bound     0.1008876  0.0344859   2.925 0.003598 **
## NOX            -0.7783993  0.1528902  -5.091 5.07e-07 ***
## Avg_Rooms       0.0908331  0.0167280   5.430 8.87e-08 ***
## Owner_Occupied  0.0002106  0.0005287   0.398 0.690567
## Distance       -0.0490873  0.0079834  -6.149 1.62e-09 ***
## Rad             0.0142673  0.0026556   5.373 1.20e-07 ***
## Tax            -0.0006258  0.0001505  -4.157 3.80e-05 ***
## PTRatio        -0.0382715  0.0052365  -7.309 1.10e-12 ***
## Blacks          0.0004136  0.0001075   3.847 0.000135 ***
## Lower_Status   -0.0290355  0.0020299 -14.304  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1899 on 492 degrees of freedom
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7841
## F-statistic: 142.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

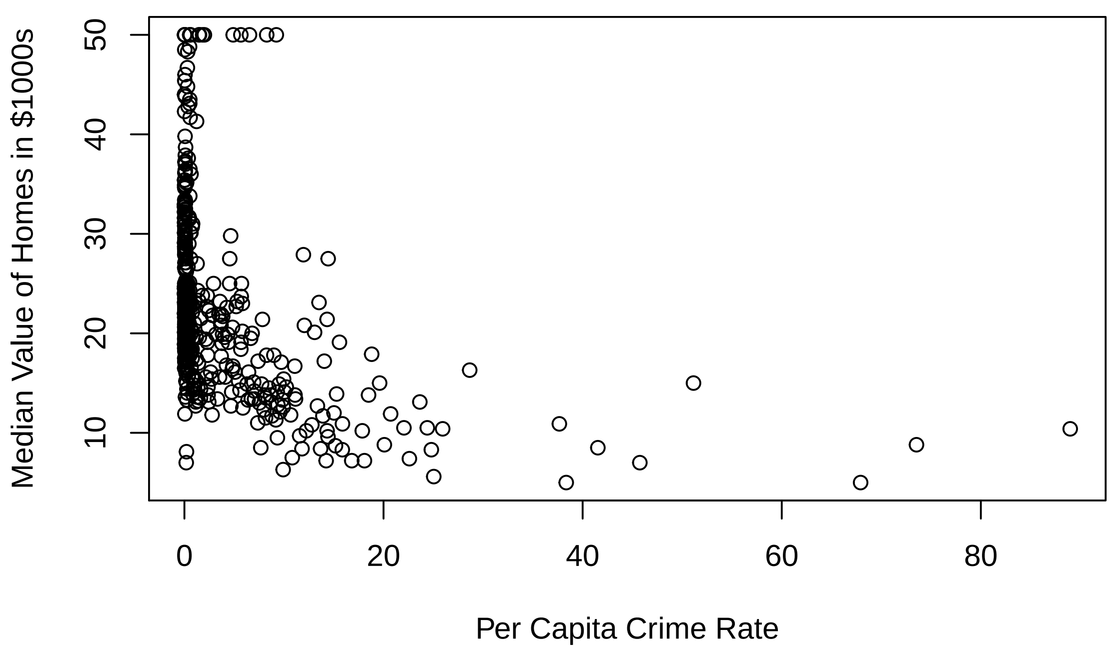
   We can observe that Indus, Owner_Occupied are not statistically significant and can be removed from the model. The remaining variables are statistically significant based on the summary.

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.

```r
#3
#Crime Rate Vs Median Value
fit_crim <- lm(Median_Value ~ Crime_Rate, data = BostonData)
summary(fit_crim)
```

```
##
## Call:
## lm(formula = Median_Value ~ Crime_Rate, data = BostonData)
##
## 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 ***
## Crime_Rate  -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
```

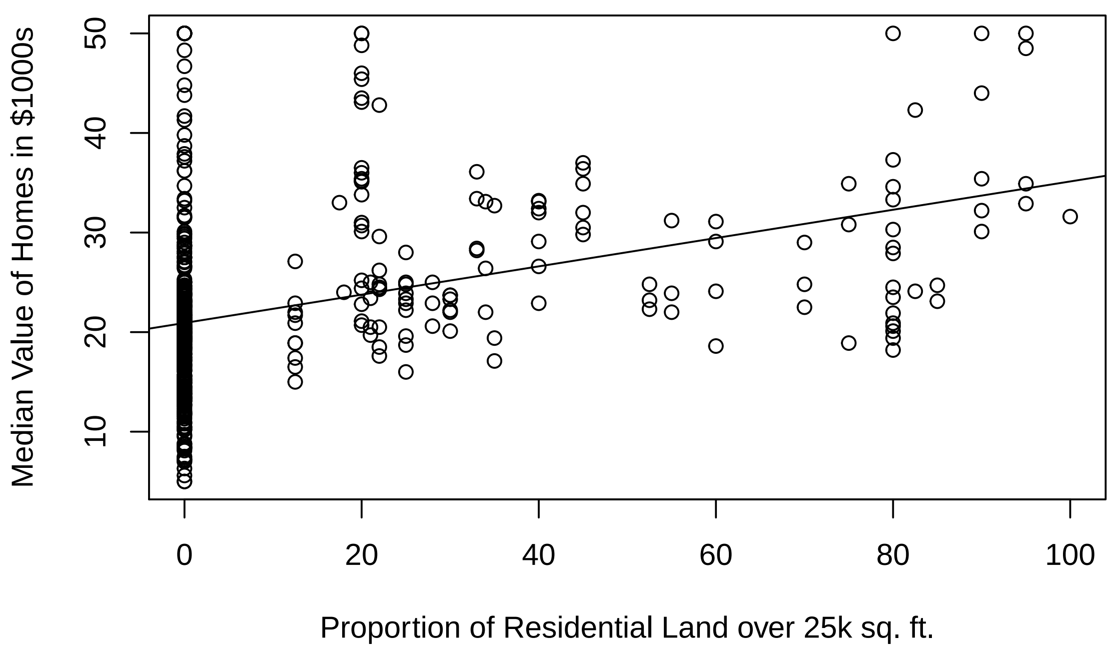
```r
plot(BostonData$Crime_Rate, BostonData$Median_Value,
xlab = "Per Capita Crime Rate", ylab = "Median Value of Homes in $1000s")
```



```r
#Zoned_Land vs Median_Value
fit_zn <- lm(Median_Value ~ Zoned_Land, data = BostonData)
summary(fit_zn)
```

```
##
## Call:
## lm(formula = Median_Value ~ Zoned_Land, data = BostonData)
##
## 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 ***
## Zoned_Land   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
```

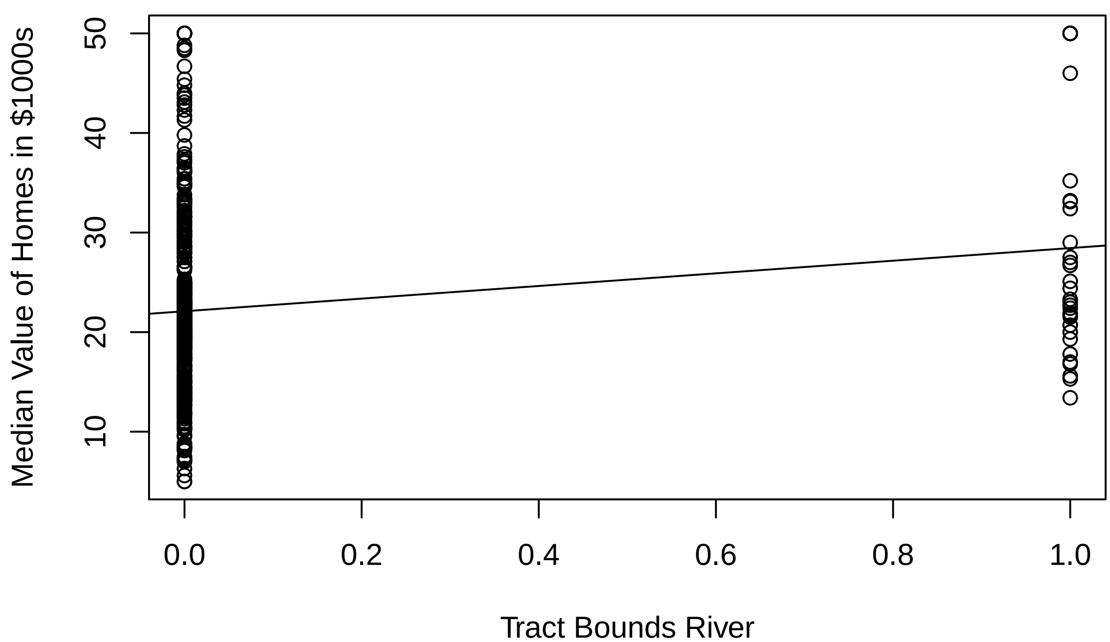
```r
plot(BostonData$Zoned_Land, BostonData$Median_Value,
xlab = "Proportion of Residential Land over 25k sq. ft.", ylab = "Median Value of Homes in $1000s")
abline(fit_zn)
```



```r
#Tract_Bound vs Median_Value
fit_tract <- lm(Median_Value ~ Tract_Bound, data = BostonData)
summary(fit_tract)
```

```
##
## Call:
## lm(formula = Median_Value ~ Tract_Bound, data = BostonData)
##
## 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 ***
## Tract_Bound   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
```

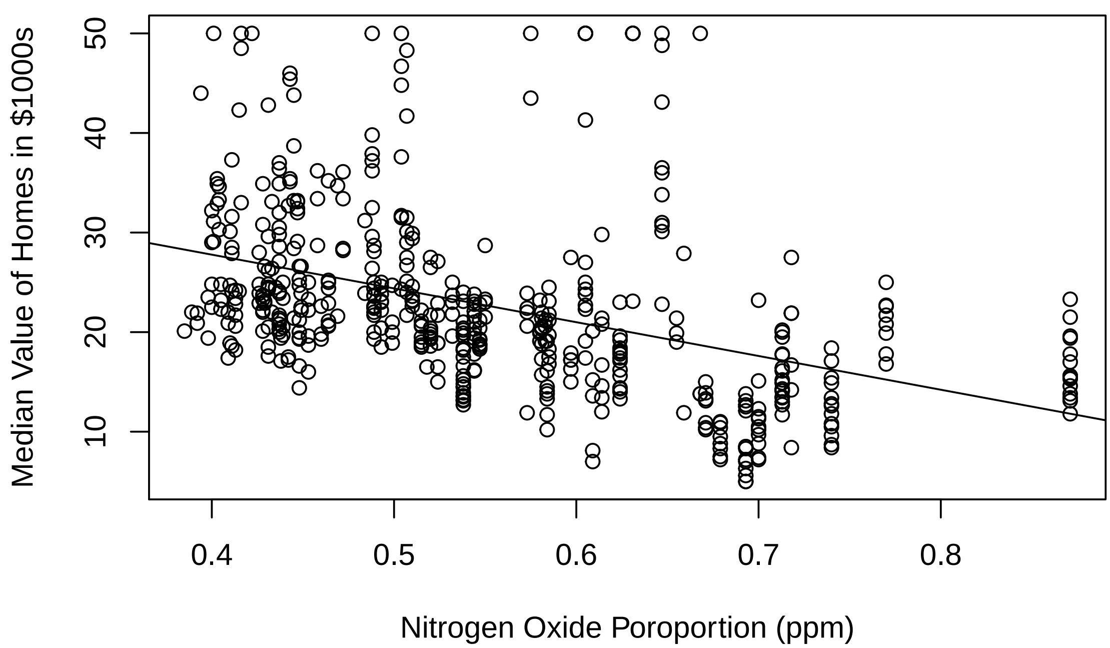
```r
plot(BostonData$Tract_Bound, BostonData$Median_Value,
xlab = "Tract Bounds River", ylab = "Median Value of Homes in $1000s")
abline(fit_tract)
```



```r
#NOX vs Median_Value
fit_nox <- lm(Median_Value ~ NOX, data = BostonData)
summary(fit_nox)
```

```
##
## Call:
## lm(formula = Median_Value ~ NOX, data = BostonData)
##
## 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
```

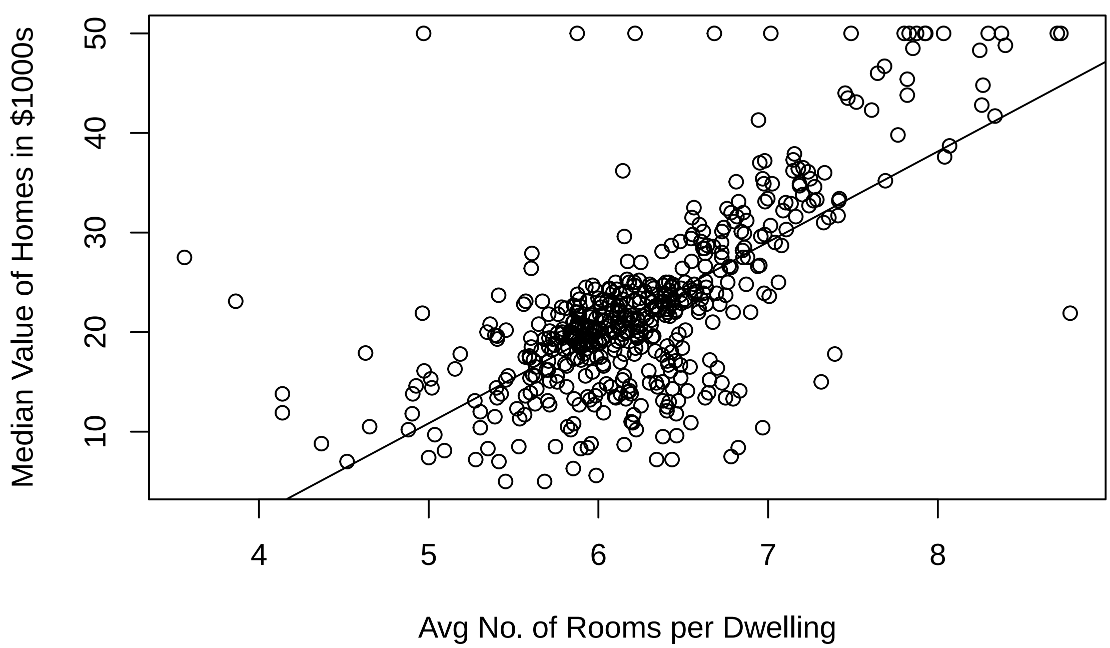
```r
plot(BostonData$NOX, BostonData$Median_Value,
xlab = "Nitrogen Oxide Poroportion (ppm)", ylab = "Median Value of Homes in $1000s")
abline(fit_nox)
```



```r
#Avg Rooms vs Median_Value
fit_rooms <- lm(Median_Value ~ Avg_Rooms, data = BostonData)
summary(fit_rooms)
```

```
##
## Call:
## lm(formula = Median_Value ~ Avg_Rooms, data = BostonData)
##
## 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 ***
## Avg_Rooms      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
```

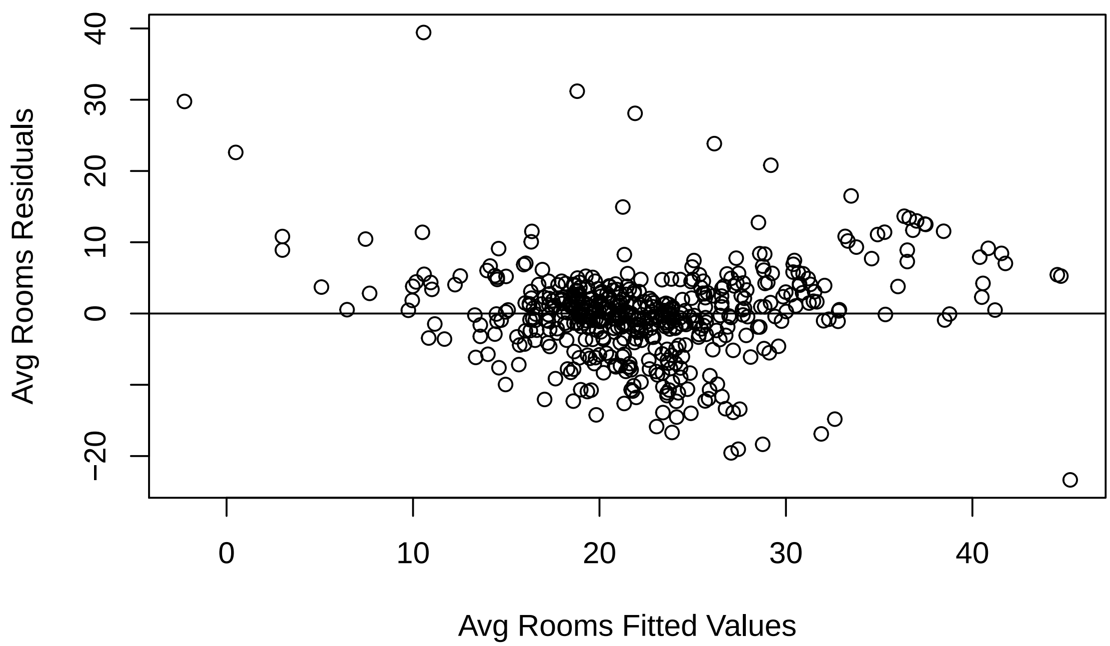
```r
plot(BostonData$Avg_Rooms, BostonData$Median_Value,
xlab = "Avg No. of Rooms per Dwelling", ylab = "Median Value of Homes in $1000s")
abline(fit_rooms)
```



The plot shows a strong association between average number of rooms and median value of the houses.

We will draw a plot of fitted values and residuals to back up this assertion.

```r
residual_rooms = lm(fit_rooms$residuals~ fit_rooms$fitted.values)
plot(fit_rooms$fitted.values, fit_rooms$residuals,
xlab = "Avg Rooms Fitted Values", ylab = "Avg Rooms Residuals")
abline(residual_rooms)
```



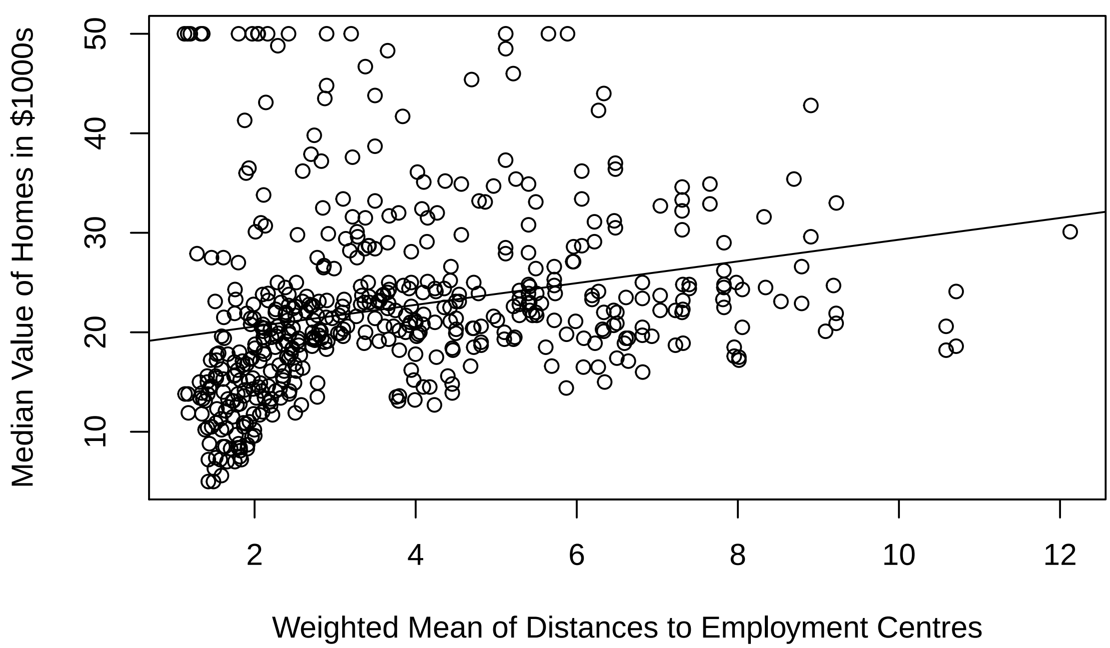
The plot of fitted values vs residuals shows a strong grouping of observations around the zero line and the plot looks similar to the average rooms vs median value plot. Thus, average room is a significant variable in our model.

```r
#Distance vs Median_Value
fit_distance <- lm(Median_Value ~ Distance, data = BostonData)
summary(fit_distance)
```

```
##
## Call:
## lm(formula = Median_Value ~ Distance, data = BostonData)
##
## 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 ***
## Distance      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(BostonData$Distance, BostonData$Median_Value,
xlab = "Weighted Mean of Distances to Employment Centres", ylab = "Median Value of Homes in $1000s")
abline(fit_distance)
```

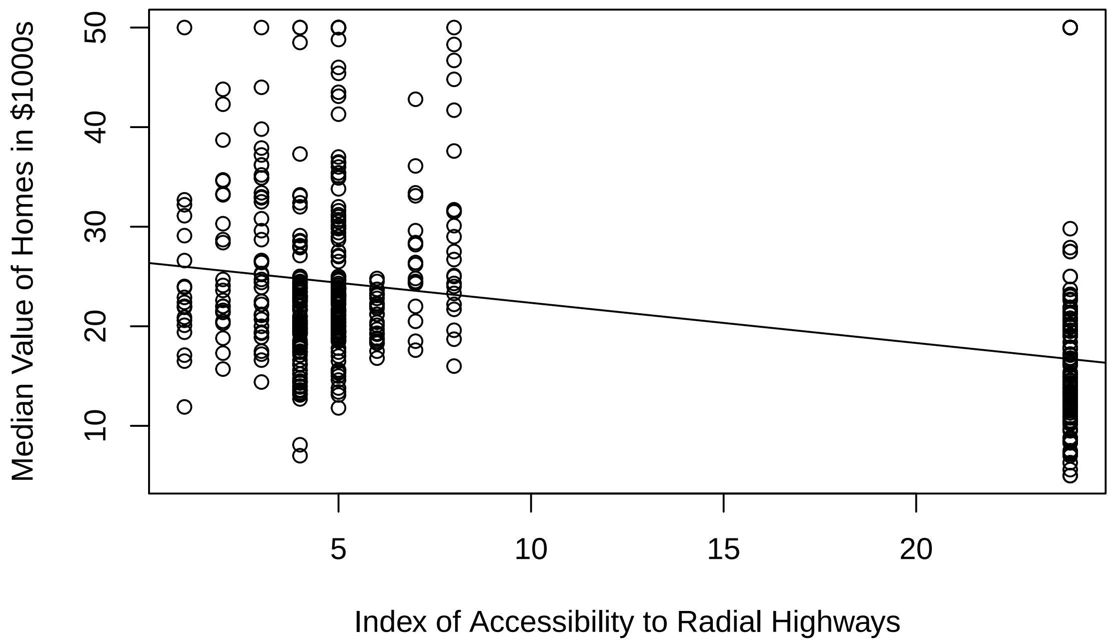



```r
#Rad vs Median_Value
fit_rad <- lm(Median_Value ~ Rad, data = BostonData)
summary(fit_rad)
```

```
##
## Call:
## lm(formula = Median_Value ~ Rad, data = BostonData)
##
## 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
```

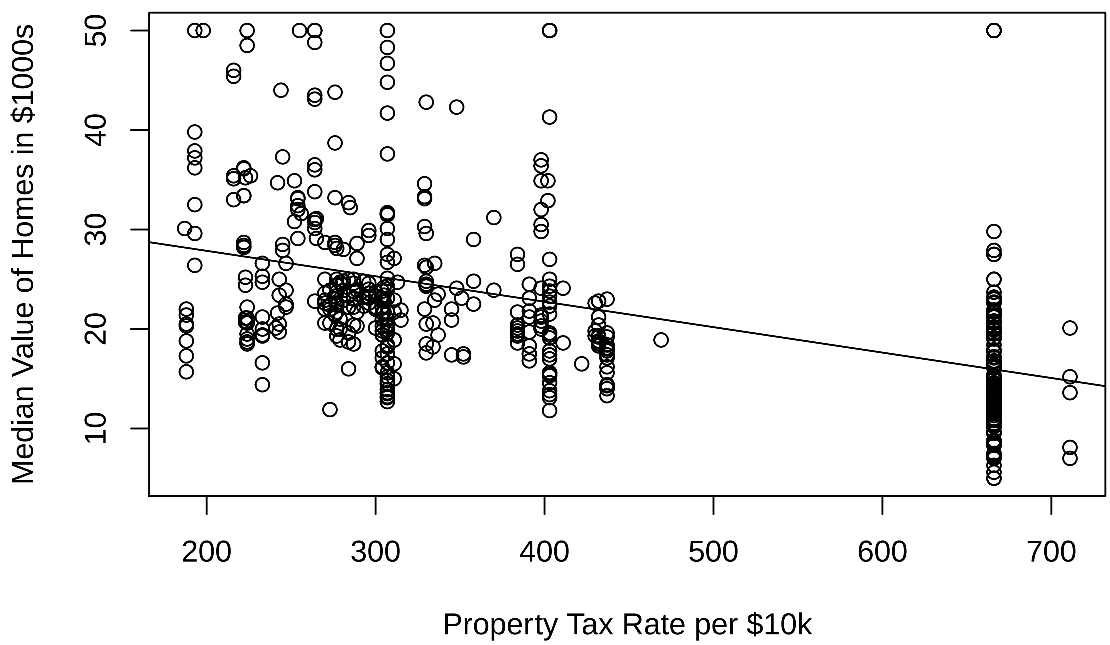
```r
plot(BostonData$Rad, BostonData$Median_Value,
xlab = "Index of Accessibility to Radial Highways", ylab = "Median Value of Homes in $1000s")
abline(fit_rad)
```




```r
#Tax vs Median_Value
fit_tax <- lm(Median_Value ~ Tax, data = BostonData)
summary(fit_tax)
```

```
##
## Call:
## lm(formula = Median_Value ~ Tax, data = BostonData)
##
## 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
```

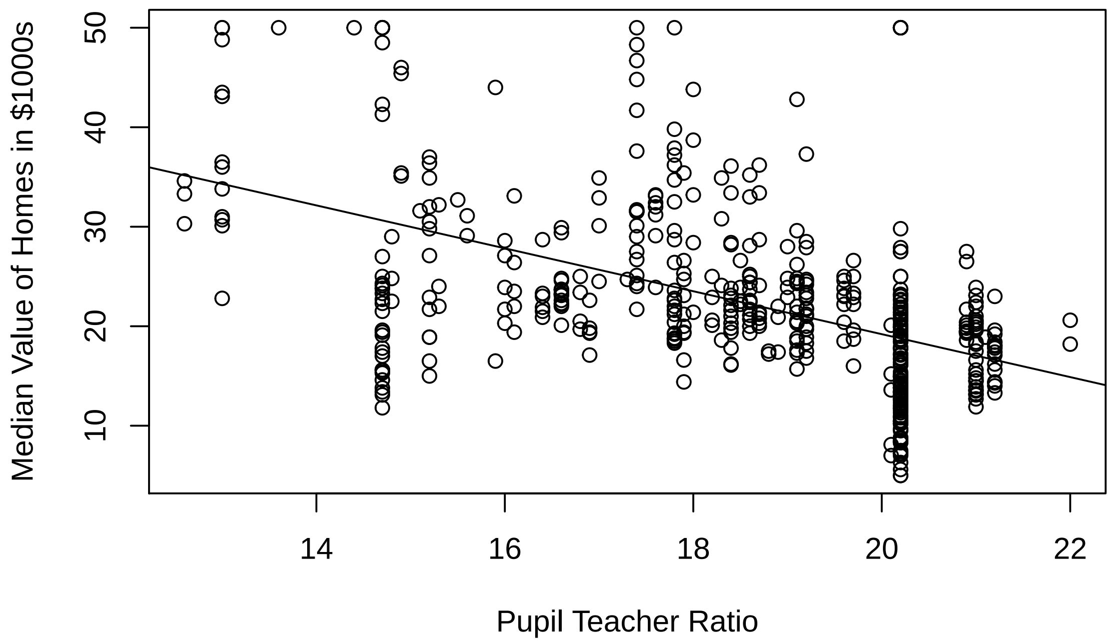
```r
plot(BostonData$Tax, BostonData$Median_Value,
xlab = "Property Tax Rate per $10k", ylab = "Median Value of Homes in $1000s")
abline(fit_tax)
```




```r
#PTRatio vs Median_Value
fit_ptr <- lm(Median_Value ~ PTRatio, data = BostonData)
summary(fit_ptr)
```

```
##
## Call:
## lm(formula = Median_Value ~ PTRatio, data = BostonData)
##
## 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
```

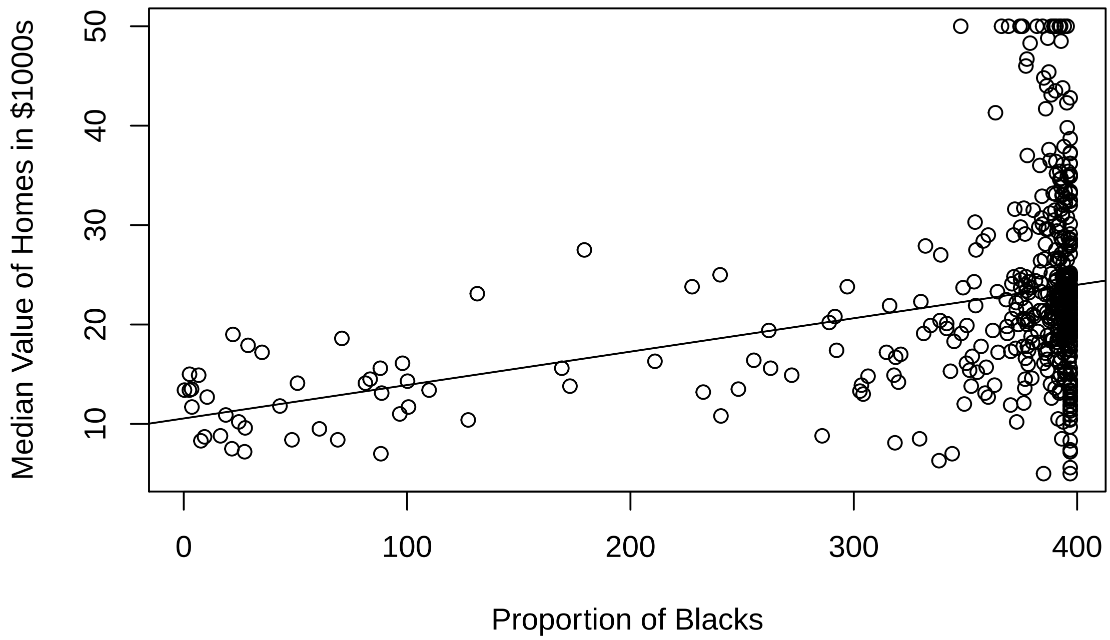
```r
plot(BostonData$PTRatio, BostonData$Median_Value,
xlab = "Pupil Teacher Ratio", ylab = "Median Value of Homes in $1000s")
abline(fit_ptr)
```




```r
#Blacks vs Median_Value
fit_blacks <- lm(Median_Value ~ Blacks, data = BostonData)
summary(fit_blacks)
```

```
##
## Call:
## lm(formula = Median_Value ~ Blacks, data = BostonData)
##
## 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 ***
## Blacks       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
```

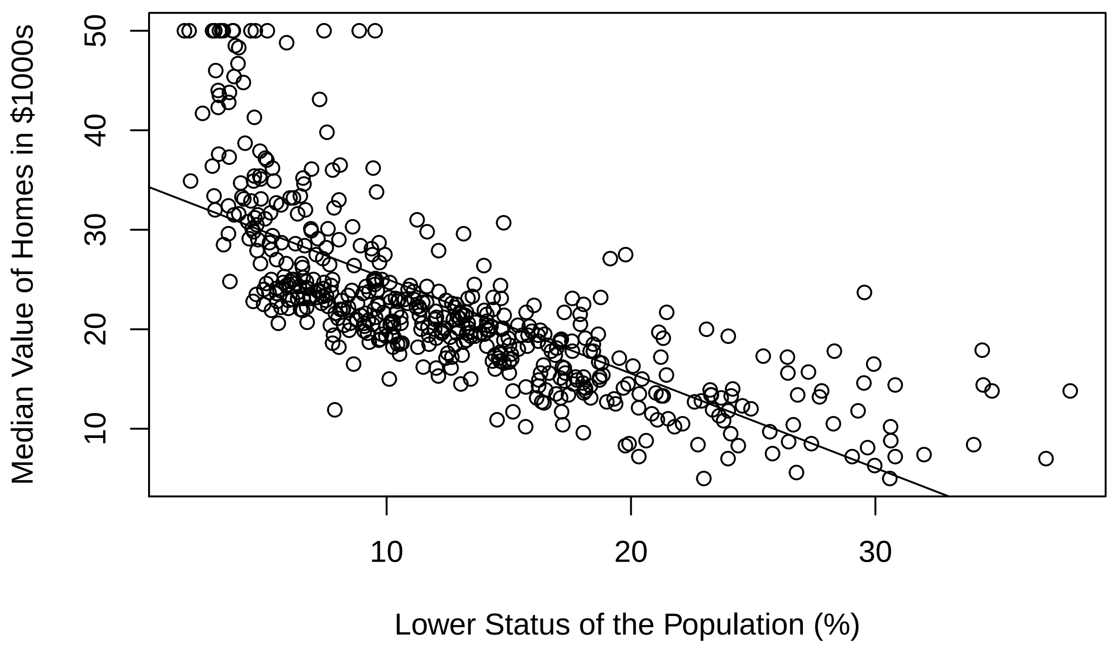
```r
plot(BostonData$Blacks, BostonData$Median_Value,
xlab = "Proportion of Blacks", ylab = "Median Value of Homes in $1000s")
abline(fit_blacks)
```




```r
#Lower_Status vs Median_Value
fit_status <- lm(Median_Value ~ Lower_Status, data = BostonData)
summary(fit_status)
```

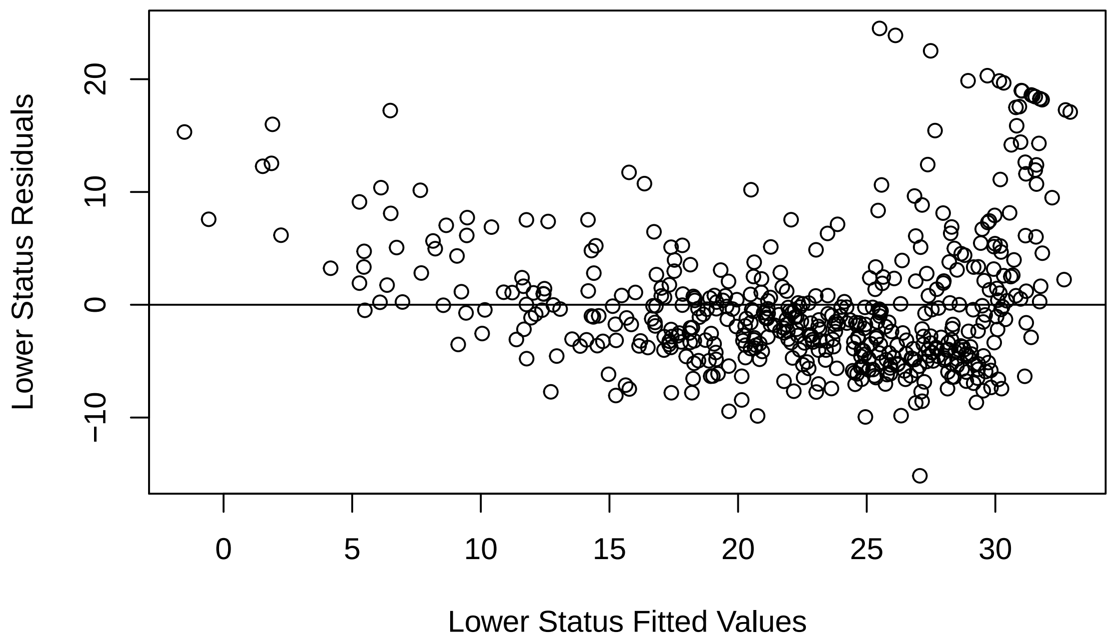
```
##
## Call:
## lm(formula = Median_Value ~ Lower_Status, data = BostonData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.55384    0.56263   61.41   <2e-16 ***
## Lower_Status -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(BostonData$Lower_Status, BostonData$Median_Value,
xlab = "Lower Status of the Population (%)", ylab = "Median Value of Homes in $1000s")
abline(fit_status)
```



The plot shows a strong association between lower status of the population and median value of the houses. We will draw a plot of fitted values and residuals to back up this assertion.

```r
residual_lower_status = lm(fit_status$residuals~ fit_status$fitted.values)
plot(fit_status$fitted.values, fit_status$residuals,
xlab = "Lower Status Fitted Values", ylab = "Lower Status Residuals")
abline(residual_lower_status)
```



The plot of fitted values vs residuals shows a fairly decent grouping of observations around the zero line and the plot looks similar to the lower status vs median value plot. Thus, lower status is a significant variable in our model.

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 $H_0 : \beta_j = 0$?

```
#4
fit_multiple <- lm(formula = Median_Value ~ Crime_Rate + Zoned_Land + Tract_Bound + NOX +
Avg_Rooms + Distance + Rad + Tax + PTRatio + Blacks +
Lower_Status, data = BostonData)

summary(fit_multiple)
```

```
##
## Call:
## lm(formula = Median_Value ~ Crime_Rate + Zoned_Land + Tract_Bound +
##     NOX + Avg_Rooms + Distance + Rad + Tax + PTRatio + Blacks +
##     Lower_Status, data = BostonData)
##
## 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 ***
## Crime_Rate    -0.108413   0.032779  -3.307 0.001010 **
## Zoned_Land     0.045845   0.013523   3.390 0.000754 ***
```

```
## Tract_Bound    2.718716   0.854240   3.183 0.001551 **
## NOX          -17.376023   3.535243  -4.915 1.21e-06 ***
## Avg_Rooms      3.801579   0.406316   9.356  < 2e-16 ***
## Distance      -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 ***
## Blacks         0.009291   0.002674   3.475 0.000557 ***
## Lower_Status  -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
```

Multiple Regression Model:

M1 = 36.34 - 0.11 * Crime_Rate + 0.04 * Zoned_Land + 2.71 * Tract_Bound - 17.37 * NOX + 3.80 * Avg_Rooms - 1.49 * Distance + 0.30 * Rad - 0.01 * Tax - 0.94 * PTRatio + 0.009 * Blacks - 0.52 * Lower_Status

We can reject null hypothesis for all the predictors considering the t and p values from the summary above which show all of the predictors are statistically significant. Also the F-statistic = 128.2 >> 1 suggests there is at least one predictor that is related to the median value of houses and thus we can reject null hypothesis.

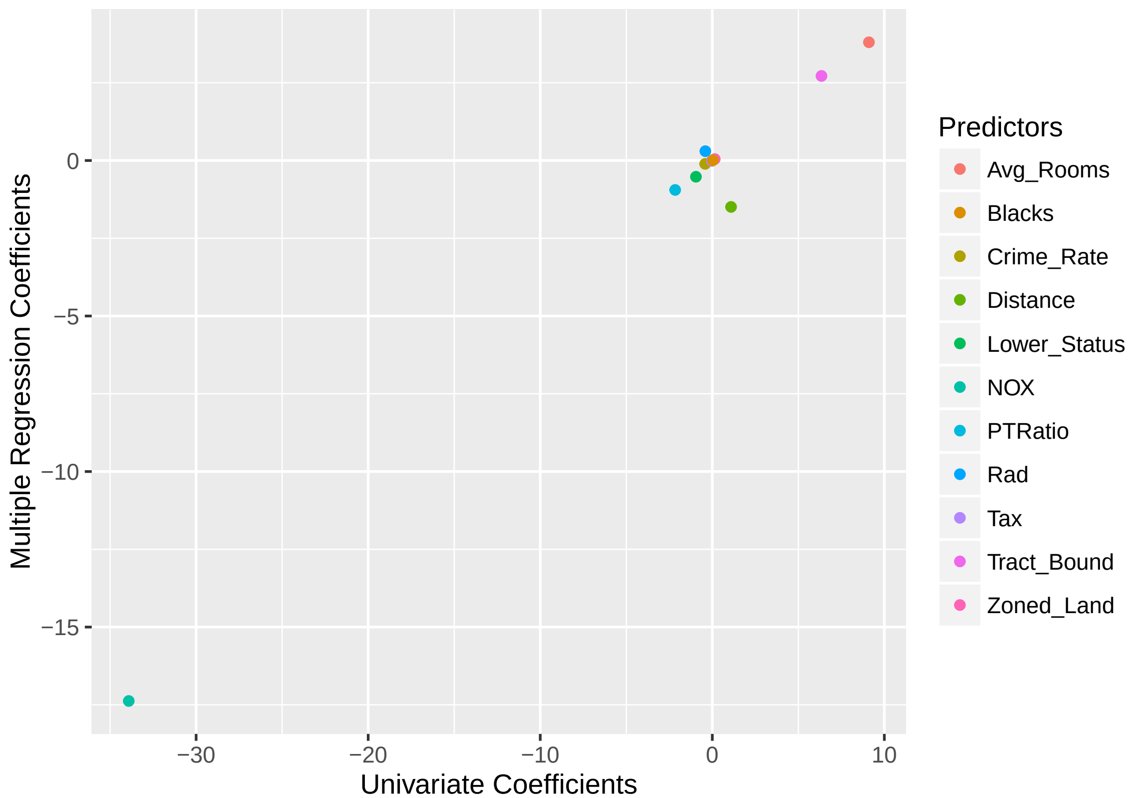
5. How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response.

```r
#5
Predictors = c("Crime_Rate", "Zoned_Land", "Tract_Bound" ,"NOX", "Avg_Rooms", "Distance",
"Rad", "Tax", "PTRatio", "Blacks", "Lower_Status")
univariate_coeffs = c(fit_crim$coefficients[2], fit_zn$coefficients[2],
fit_tract$coefficients[2], fit_nox$coefficients[2],
fit_rooms$coefficients[2], fit_distance$coefficients[2],
fit_rad$coefficients[2], fit_tax$coefficients[2],
fit_ptr$coefficients[2], fit_blacks$coefficients[2],
fit_status$coefficients[2])

multiple_coeffs = c(fit_multiple$coefficients[2],fit_multiple$coefficients[3],
fit_multiple$coefficients[4],fit_multiple$coefficients[5],
fit_multiple$coefficients[6],fit_multiple$coefficients[7],
fit_multiple$coefficients[8],fit_multiple$coefficients[9],
fit_multiple$coefficients[10], fit_multiple$coefficients[11],
fit_multiple$coefficients[12])

df = data.frame(Predictors, univariate_coeffs, multiple_coeffs)

ggplot(data = df, aes(x = univariate_coeffs, y = multiple_coeffs, color = Predictors)) +
geom_point() +
labs(x = "Univariate Coefficients", y = "Multiple Regression Coefficients")
```



The value of univariate regression coeffcients are more extreme compared to multiple regression coefficients. By more extreme I mean, if univariate coefficients are positive, the multiple regression coefficients are less in value while if the univariate regression coefficients are negative multiple regression coefficients take higher values.

Also most values are between -1 and +1 with a cluster around (0,0).

An interesting observation is for predictor Distance - the 2 coefficients have opposite signs and we may need to inspect this variable closely while devising a regression model.

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:

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

```
#6

lm.out = lm(Median_Value ~ Crime_Rate + I(Crime_Rate^2) + I(Crime_Rate^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Crime_Rate + I(Crime_Rate^2) + I(Crime_Rate^3),
##     data = BostonData)
##
## 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 ***
## Crime_Rate      -1.136e+00  1.444e-01  -7.868 2.24e-14 ***
## I(Crime_Rate^2)  2.378e-02  6.808e-03   3.494 0.000518 ***
## I(Crime_Rate^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
#Median_Value = 0.2519 - 1.136 * Crime_Rate + 0.02378 * Crime_Rate^2 - .0001489 * Crime_Rate^3
```

```r
lm.out = lm(Median_Value ~ Zoned_Land + I(Zoned_Land^2) + I(Zoned_Land^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Zoned_Land + I(Zoned_Land^2) + I(Zoned_Land^3),
##     data = BostonData)
##
## 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 ***
## Zoned_Land       0.6433652  0.1105611   5.819 1.06e-08 ***
## I(Zoned_Land^2) -0.0167646  0.0038872  -4.313 1.94e-05 ***
## I(Zoned_Land^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
# Median_Value = 20.45 + 0.64 * Zoned_Land - 0.0167 * Zoned_Land^2 + 0.000125 * Zoned_Land^3
```

```r
lm.out = lm(Median_Value ~ NOX + I(NOX^2) + I(NOX^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ NOX + I(NOX^2) + I(NOX^3), data = BostonData)
##
## 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
```

The summary shows the coefficients for the first and second power do not make a significant impact (from t and p values) and we can conclude there is absence of non-linear regression.

```r
lm.out = lm(Median_Value ~ Avg_Rooms + I(Avg_Rooms^2) + I(Avg_Rooms^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Avg_Rooms + I(Avg_Rooms^2) + I(Avg_Rooms^3),
##     data = BostonData)
##
## 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 ***
## Avg_Rooms      -109.3906    22.9690  -4.763 2.51e-06 ***
## I(Avg_Rooms^2)   16.4910     3.6750   4.487 8.95e-06 ***
## I(Avg_Rooms^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
# Median_Value = 241.31 - 109.39 * Avg_Rooms + 16.49 * Avg_Rooms^2 - 0.74 * Avg_Rooms^3
```

```r
lm.out = lm(Median_Value ~ Distance + I(Distance^2) + I(Distance^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Distance + I(Distance^2) + I(Distance^3),
##     data = BostonData)
##
## 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 *
## Distance        8.59284    2.06633   4.158 3.77e-05 ***
## I(Distance^2) -1.24953    0.41235  -3.030  0.00257 **
## I(Distance^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
```

```r
# Median_Value = 7.038 + 8.59 * Distance - 1.25 * Distance^2 + 0.056 * Distance^3
```

```r
lm.out = lm(Median_Value ~ Rad + I(Rad^2) + I(Rad^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Rad + I(Rad^2) + I(Rad^3), data = BostonData)
##
## 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
```

```r
# Median_Value = 30.25 - 3.8 * Rad + 0.61 * Rad^2 - 0.02 * Rad^3
```

```r
lm.out = lm(Median_Value ~ Tax + I(Tax^2) + I(Tax^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Tax + I(Tax^2) + I(Tax^3), data = BostonData)
##
## 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
```

The summary suggests there isn't a strong non-linear relationship between Tax and Median_Value.

```r
lm.out = lm(Median_Value ~ PTRatio + I(PTRatio^2) + I(PTRatio^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ PTRatio + I(PTRatio^2) + I(PTRatio^3),
##     data = BostonData)
##
## 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
```

The summary suggests there isn't a strong non-linear relationship between PTRatio and Median_Value.

```r
lm.out = lm(Median_Value ~ Blacks + I(Blacks^2) + I(Blacks^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Blacks + I(Blacks^2) + I(Blacks^3),
##     data = BostonData)
##
## 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 ***
## Blacks      -1.703e-02  6.150e-02  -0.277    0.782
## I(Blacks^2)  2.036e-04  3.258e-04   0.625    0.532
## I(Blacks^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
```

```
# Median_Value = 0.126 - 0.01703 * Blacks + 0.0002036 * Blacks^2 - 2.22e-07 * Blacks^3
```

```
lm.out = lm(Median_Value ~ Lower_Status + I(Lower_Status^2) + I(Lower_Status^3), data=BostonData)

summary(lm.out)
```

```
##
## Call:
## lm(formula = Median_Value ~ Lower_Status + I(Lower_Status^2) +
##     I(Lower_Status^3), data = BostonData)
##
## 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 ***
## Lower_Status      -3.8655928  0.3287861 -11.757  < 2e-16 ***
## I(Lower_Status^2)  0.1487385  0.0212987   6.983 9.18e-12 ***
## I(Lower_Status^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
```

```
# Median_Value = 48.65 - 3.86 * Lower_Status + 0.15 * Lower_Status^2 - 0.002 * Lower_Status^3
```

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

```
#7
#Backward Selection
stepwise_model <- step(lm(Median_Value ~ Crime_Rate + Zoned_Land + Tract_Bound + NOX + Avg_Rooms +
Distance + Tax + Rad + PTRatio + Blacks + Lower_Status,
data = BostonData), direction = "backward")
```

```
## Start:  AIC=1585.76
## Median_Value ~ Crime_Rate + Zoned_Land + Tract_Bound + NOX +
##     Avg_Rooms + Distance + Tax + Rad + PTRatio + Blacks + Lower_Status
##
##                Df Sum of Sq   RSS    AIC
## <none>                      11081 1585.8
```

```
## - Tract_Bound   1    227.21 11309 1594.0
## - Crime_Rate    1    245.37 11327 1594.8
## - Zoned_Land    1    257.82 11339 1595.4
## - Blacks        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
## - Distance      1   1448.94 12530 1645.9
## - Avg_Rooms     1   1963.66 13045 1666.3
## - Lower_Status  1   2723.48 13805 1695.0
```

```r
summary(stepwise_model)
```

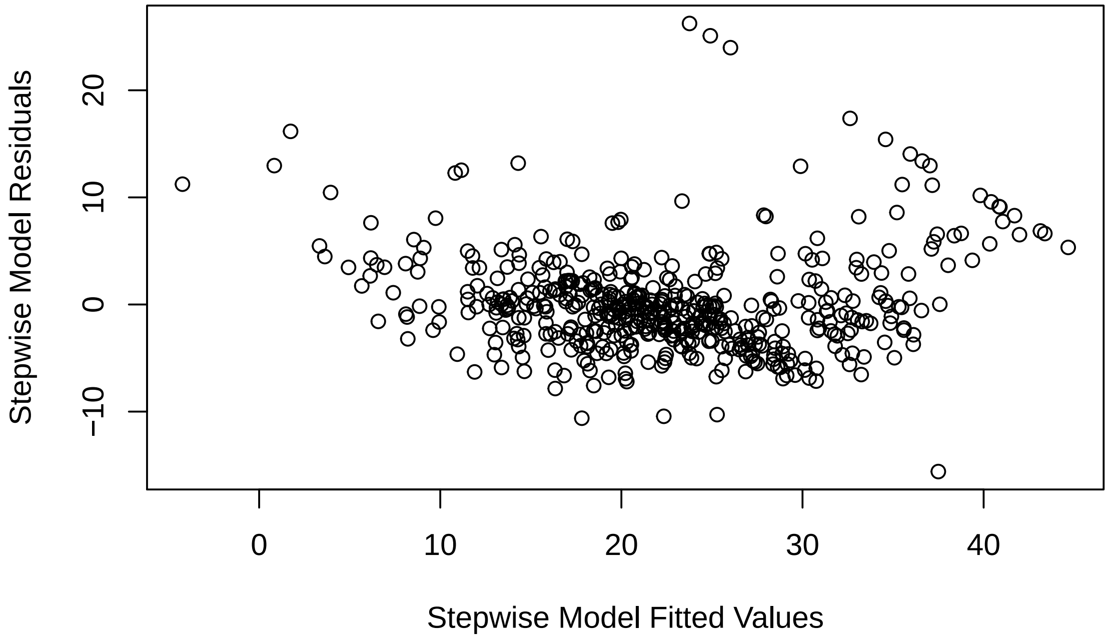
```
##
## Call:
## lm(formula = Median_Value ~ Crime_Rate + Zoned_Land + Tract_Bound +
##     NOX + Avg_Rooms + Distance + Tax + Rad + PTRatio + Blacks +
##     Lower_Status, data = BostonData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.341145   5.067492   7.171 2.73e-12 ***
## Crime_Rate    -0.108413   0.032779  -3.307 0.001010 **
## Zoned_Land     0.045845   0.013523   3.390 0.000754 ***
## Tract_Bound    2.718716   0.854240   3.183 0.001551 **
## NOX          -17.376023   3.535243  -4.915 1.21e-06 ***
## Avg_Rooms      3.801579   0.406316   9.356  < 2e-16 ***
## Distance      -1.492711   0.185731  -8.037 6.84e-15 ***
## Tax           -0.011778   0.003372  -3.493 0.000521 ***
## Rad            0.299608   0.063402   4.726 3.00e-06 ***
## PTRatio       -0.946525   0.129066  -7.334 9.24e-13 ***
## Blacks         0.009291   0.002674   3.475 0.000557 ***
## Lower_Status  -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
```

After performing a backward selection of predictors, we observe that none of the predictors is removed from the model suggesting all the predictors are significant. The most significant predictors are Lower_Status, Avg_Rooms, Distance and PTRatio.

This model is similar to (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.

```r
plot(stepwise_model$fitted.values, stepwise_model$residuals,
xlab = "Stepwise Model Fitted Values", ylab = "Stepwise Model Residuals")
```



The model residuals do not show a pattern and are mostly clustered around 0 which supports the assumption of linear regression. Also there is no overfitting of data which supports the assumptions of regression that the error terms musy be independent of each other.

One of the concerns is there are a lot of outliers which indicate this might not be the best model to represent this dataset.

As seen in (5), there are some predictor variables for which the coeefecients have oppsite signs for linear and multiple regression which might suggest there is some anomaly and needs more introspection.