INFX 573: Problem Set 6 - Regression

Matthew Peters

Due: Tuesday, November 15, 2016

Collaborators:

Derrick Priebe

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.

```r
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
library(ggplot2) # For plots
library(dplyr) # For piplined queries
```

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.

```r
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 ...
```

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

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 is the *medv* variable, which would be the result of some market function on the other variables. Where variable values reflect increased desireability of property, I would expect the *medv* value to increase. Similarly, I would expect the *medv* value to decrease when undesireable qualities increase. For example, I would expect an increase in *crim* to correlate negatively to *medv*, whereas an increase in *rm* would correlate to an increase in *medv*.

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.

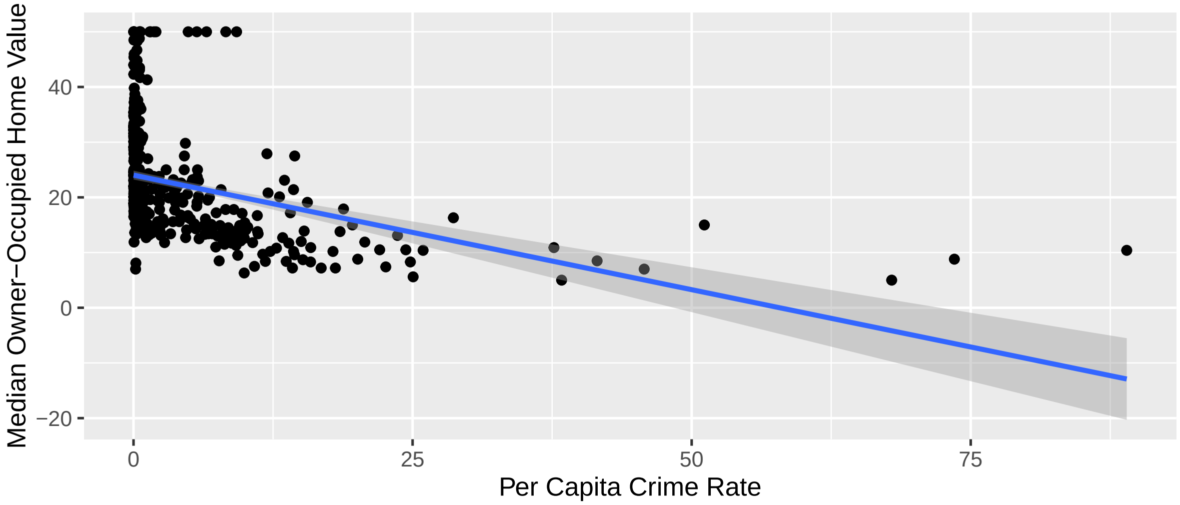
```r
# Construct linear regression models for each of the predictor variables over the
# the response variable
crim.medv <- lm(medv ~ crim, Boston)
zn.medv <- lm(medv ~ zn, Boston)
indus.medv <- lm(medv ~ indus, Boston)
chas.medv <- lm(medv ~ chas, Boston)
nox.medv <- lm(medv ~ nox, Boston)
rm.medv <- lm(medv ~ rm, Boston)
age.medv <- lm(medv ~ age, Boston)
dis.medv <- lm(medv ~ dis, Boston)
rad.medv <- lm(medv ~ rad, Boston)
tax.medv <-lm(medv ~ tax, Boston)
ptratio.medv <- lm(medv ~ ptratio, Boston)
black.medv <- lm(medv ~ black, Boston)
lstat.medv <- lm(medv ~ lstat, Boston)
```

Is Median Owner-Occupied Home Value Correlated to Crime Rate?

```r
# Summarize the crim.medv model
summary(crim.medv)
```

```
##
## 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
```

```r
# Construct a dataframe with the variables we are studying
# I am using this methodology for reliable repetition
study <- data.frame(y=Boston$medv, x=Boston$crim)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) + geom_point() +
    stat_smooth(method="lm", se=TRUE) +
    ylab("Median Owner-Occupied Home Value") +
    xlab("Per Capita Crime Rate")
```



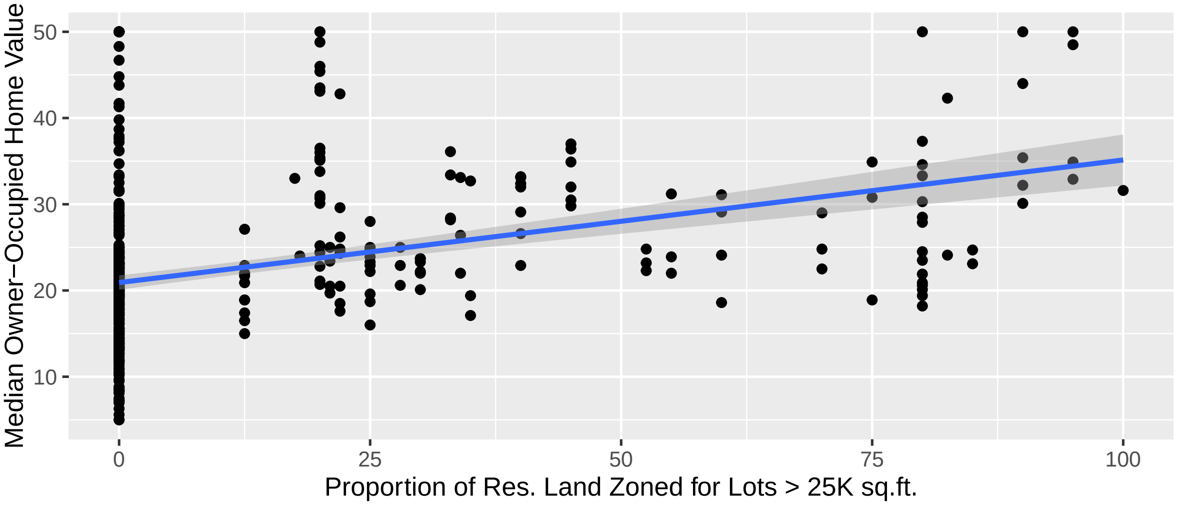
As suggested above, there is a negative correlation between the crime rate in a town and the median value of owner-occupied homes. The affect of *crim* on *medv* is statistically significant with a t-statistic of -9.46, a p-score lower than $2e - 16$. The correlation slope of -0.42.

Is Median Owner-Occupied Home Value Correlated to the Proportion of Residential Land Zoned for Lots Over 25K sq.ft.?

```
# Summarize the zn.medv model
summary(zn.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$zn)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) + geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Proportion of Res. Land Zoned for Lots > 25K sq.ft.")
```



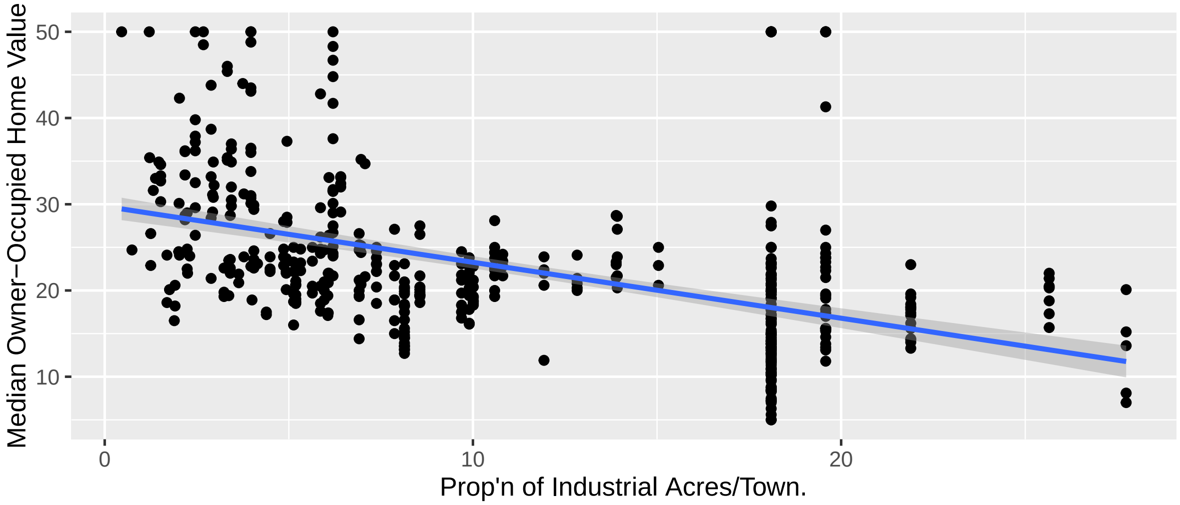
The t-statistic is 8.675; the p-score is lower than $2e-16$ and a correlation slope of 0.14.

Is Median Owner-Occupied Home Value Correlated to the Proportion of Non-Retail Business Acres per Town.?

```
# Summarize the indus.medv model
summary(indus.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$indus)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Prop'n of Industrial Acres/Town.")
```



The t-statistic is -12.41; the p-score is s lower than $2e - 16$. There is a strong negative correlation slope of -0.65.

Is Median Owner-Occupied Home Value Correlated to Proximity to the Charles River?

This is a Bernoulli categorical variable and will not map well against a straight-line regression model.

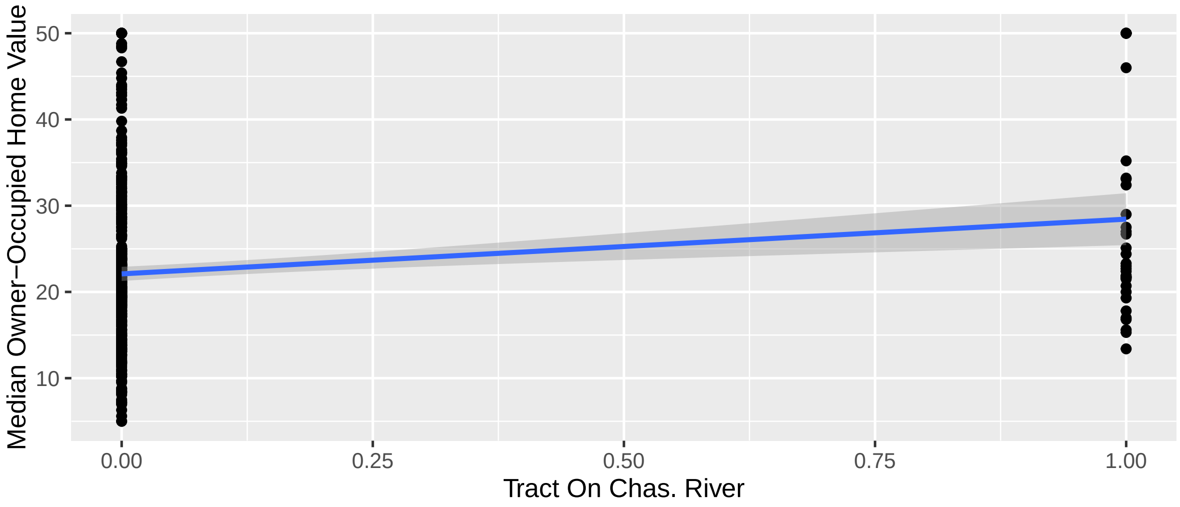
```
# Summarize the chas.medv model
summary(chas.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$chas)
```

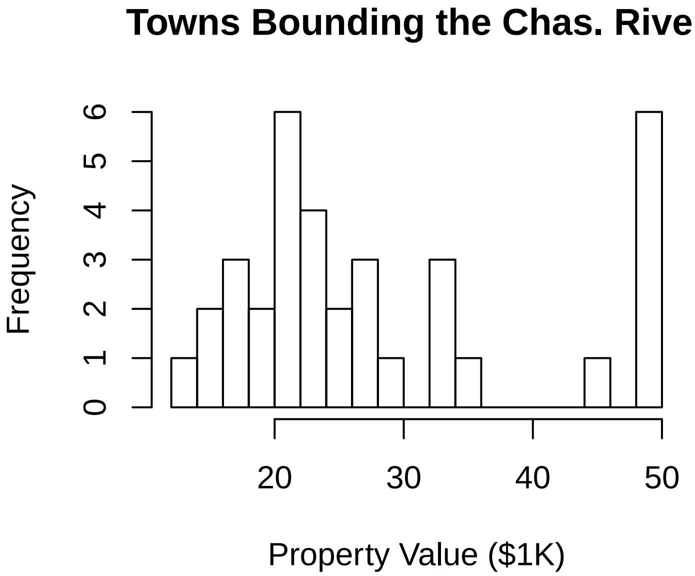
There is a very slight positive correlation with a slope of 6.3. The t-statistic is 3.996 and the p-score is $7.39e - 05$, which is statistically significant.

```
# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) + geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Tract On Chas. River")
```



```
# Lets look at only River-Front tracts ...
river.front <- study %>% dplyr::filter(x==1)
```

```r
# ... and draw a histogram of residential property values
hist(river.front$y, breaks = 25,
     xlab = "Property Value ($1K)",
     main = "Towns Bounding the Chas. River")
```



Tracts that do not bound the Charles River span the entire sample set. Tracts that bound the Charles River range in value between $15K and $50K, and exhibit a bi-modal distribution with modes at $20K and $50K. It is reasonable that being on the river front may have some impact on property value, but *where* on the river the town is located is probably important as well. For example, upstream property values might enjoy a cleaner river, whereas at the mouth of the river we might see more industry and less desireable residential property.

Is Median Owner-Occupied Home Value Correlated to Nitrous Oxide (NOx)?

- Nitrogen oxides and volatile organic compounds react in the atmosphere in the presence of sunlight to form ground-level ozone (smog).

- Health and environmental effects from high levels of ozone include:
  - Moderate to large (well over 20%) decreases in lung function resulting in difficulty in breathing, shortness of breath, and other symptoms;

  - Respiratory symptoms such as those associated with bronchitis (e.g., aggravated coughing and chest pain);

  - Increased respiratory problems (e.g. aggravation of asthma, susceptibility to respiratory infection) resulting in more hospital admiss ions and emergency room visits;

  - Repeated exposures could result in chronic inflammation and irreversible structural changes in the lungs that can lead to premature aging of the lungs and other respiratory illness.

  - Damage to forest ecosystems, trees and ornamental plants, and crops

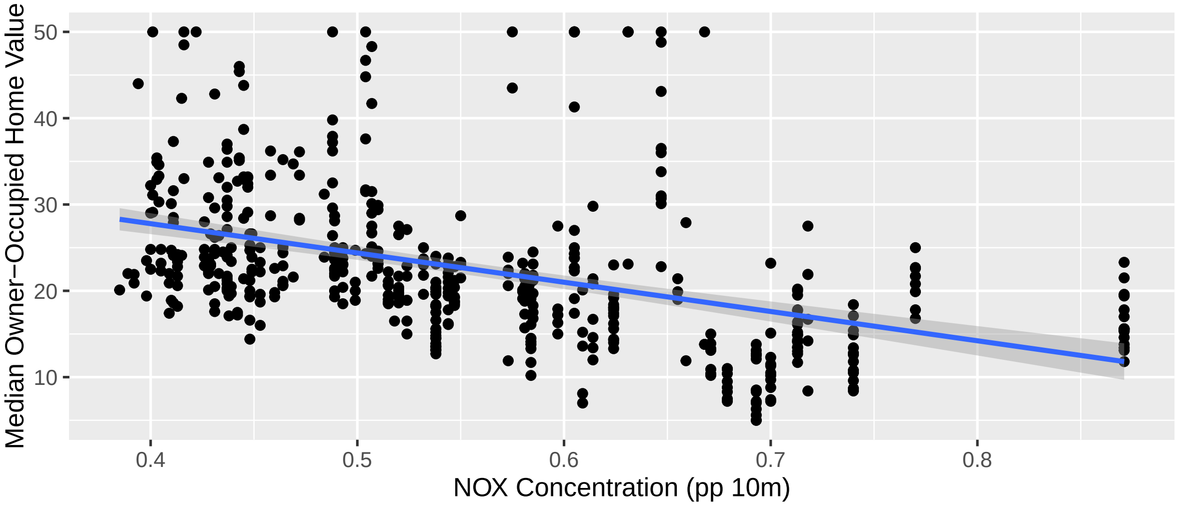
(US Environmental Protection Agency, "Overview of the Human Health and Environmental Effects of Power Generation: Focus on Sulfur Dioxide (SO2), Nitrogen Oxides (NOX) and Mercury (Hg)", June, 2002)

When asked as "Is the median owner-occupied home value correlated to smog levels?" I think it is fair to expect a strong negative correlation.

```
# Summarize the nox.medv model
summary(nox.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$nox)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("NOX Concentration (pp 10m)")
```



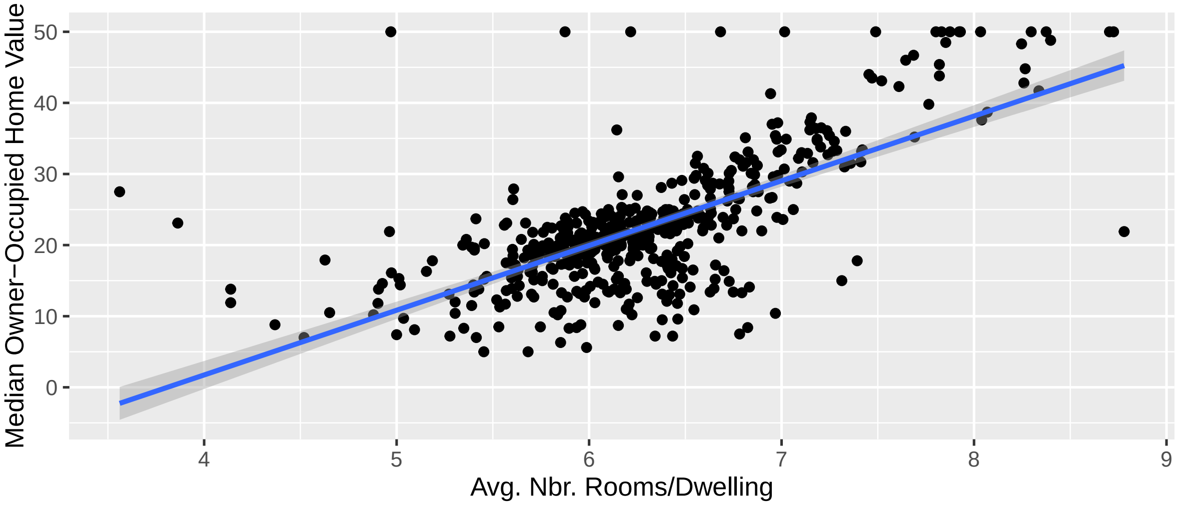
The t-statistic is -10.61 and the p-score is statistically significat at less than $2e - 16$. The negative correlation slope is -33.91. The plot does a good job of illustrating how higher levels of Nitrous Oxide correlate to lower property value

Is Median Owner-Occupied Home Value Correlated to the Average Number of Rooms per Dwelling

```
# Summarize the rm.medv model
summary(rm.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$rm)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Avg. Nbr. Rooms/Dwelling")
```



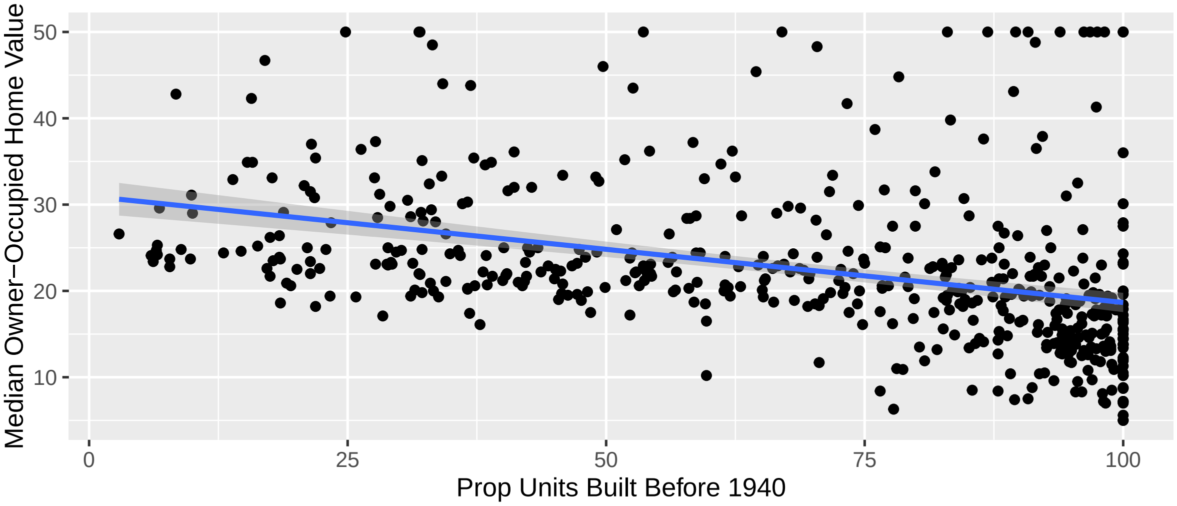
This is a well correlated pair of variables. The positive correlation slope is 9.1, the t-statistic is 21.72, and the p-score is less than $2e - 16$. The plot illustrates a nice fit between the linear regression line and the data.

Is Median Owner-Occupied Home Value Correlated to the Poportion of Units Built Prior to 1940

```r
# Summarize the age.medv model
summary(age.medv)
```

```
##
## 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
```

```r
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$age)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Prop Units Built Before 1940")
```



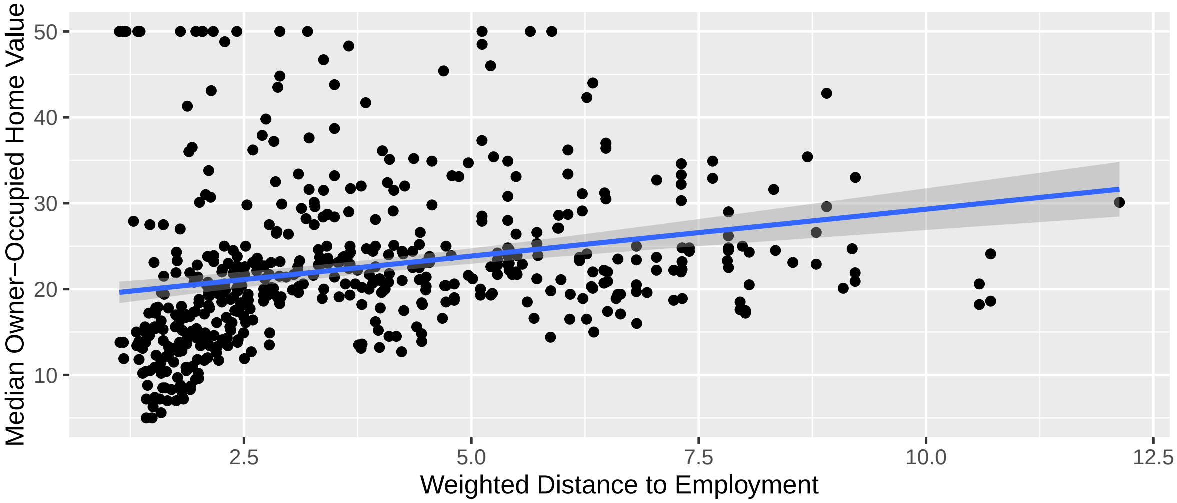
There is a negative correlation between the proportion of WWII-era buildings and newer buildings and the median home value with a slope of -0.123. The correlation is statistically significant with a t-statistic of -9.137 and a p-score of less than $2e - 16$.

Is Median Owner-Occupied Home Value Correlated to Distances to Five Boston Employment Centres

```r
# Summarize the dis.medv model
summary(dis.medv)
```

```
##
## 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
```

```r
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$dis)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Weighted Distance to Employment")
```



Here we have a positive correlation with a slope of 1.09. The t-statistic is 5.795 and the p-score is statistically significant at $1.21e-08$. But this again is an example where the simple line model does not really explain the correlation - the highest valued properties are within the closest distance index from eployment centers. Not distinguished in these numbers are which of the five the employment centers are closest, nor is there a distance per center, which might provide insight into employment flexibility. The plot suggests that some towns achieve their proprty value by being distant from employment centers. I would like to try to correlate this value with the *chas*, *rm*, and other variables. It would also be interesting to see if it negatively correlates to the *nox* variable.

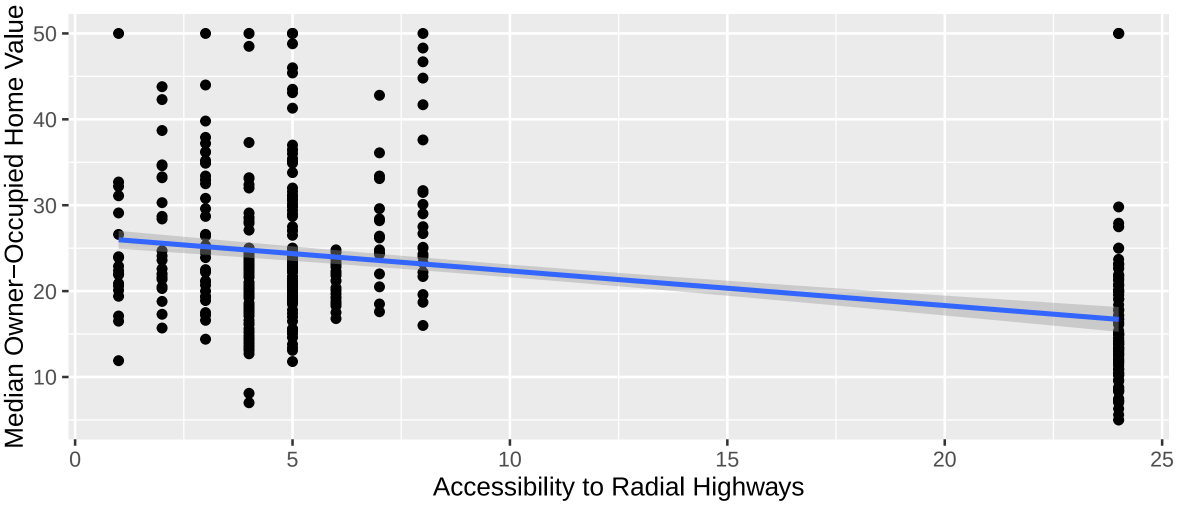
Is Median Owner-Occupied Home Value Correlated to Accessibility to Radial Highways

Boston has three highways that ring it. From the city center out they are: the 95, the 495, and the 190/146. We are not provided with the names of the town, or the highway(s) that the town is in proximity to.

```
# Summarize the rad.medv model
summary(rad.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$rad)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Accessibility to Radial Highways")
```



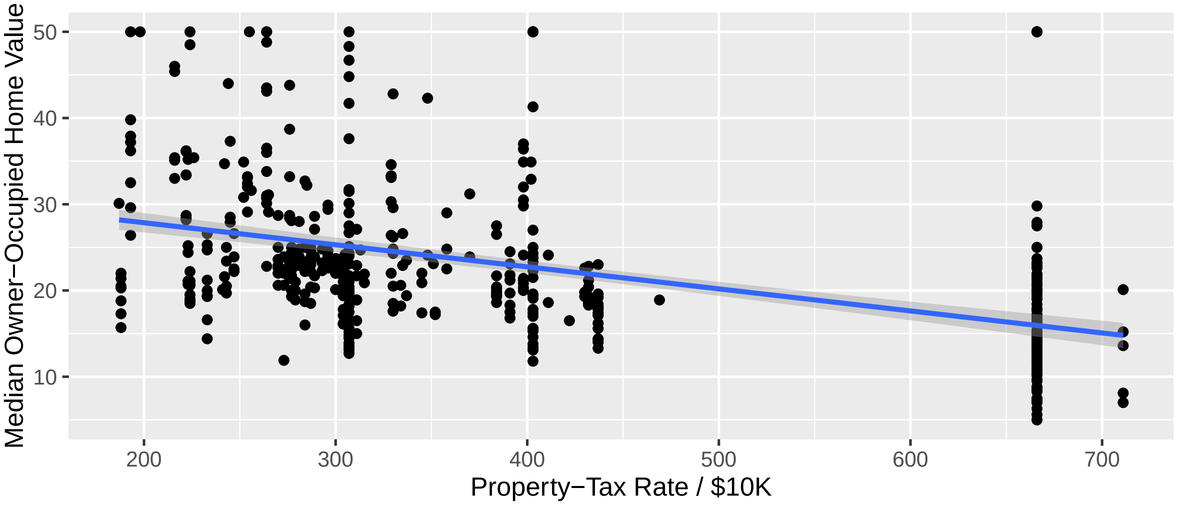
This data is bimodal, with one mode at about $15K with a distance of about 24 from the radial highways. The other mode is at about $22K with a distance between 1 and 7 from the radial highways. While it is true that of the six towns with the highest medain property value, five are no more than 10 from the radial highways, the sixth is about 24 from them. From $30K to $50K, proximity to the radial highways appears to support property value. The negative correlation between these variables has a slope of -0.4 The t-statistic is -9.269 and the p-score is $2e-16$, making the correlation statistically significant.

Is Median Owner-Occupied Home Value Correlated to Full-Value Property-Tax Rate per $10,000

```
# Summarize the tax.medv model
summary(tax.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$tax)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Property-Tax Rate / $10K")
```



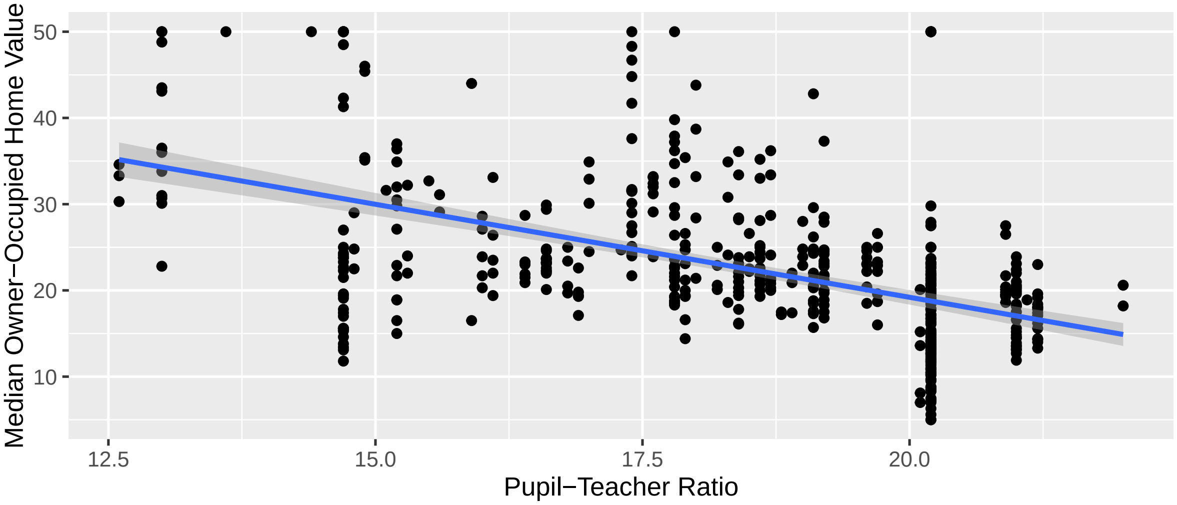
This data behaves much like the *rad.medv* model: for towns where the median property value is between $30K and $50K, we see tax-rate values between $200 and $400. The data are bimodal below $30K – there is one modal cluster at about $15K where the property tax rate is between $620 and $700, and another mode around $24 where the tax rate is between $200 and $400. The variables are negativelu correlated with a slope of -0.03. The t-statistic is -11.91 and the p-score is less thatn $2e - 16$, indicating statistical significance.

Is Median Owner-Occupied Home Value Correlated to Pupil-Teacher Ratio by Town

```
# Summarize the ptratio.medv model
summary(ptratio.medv)
```

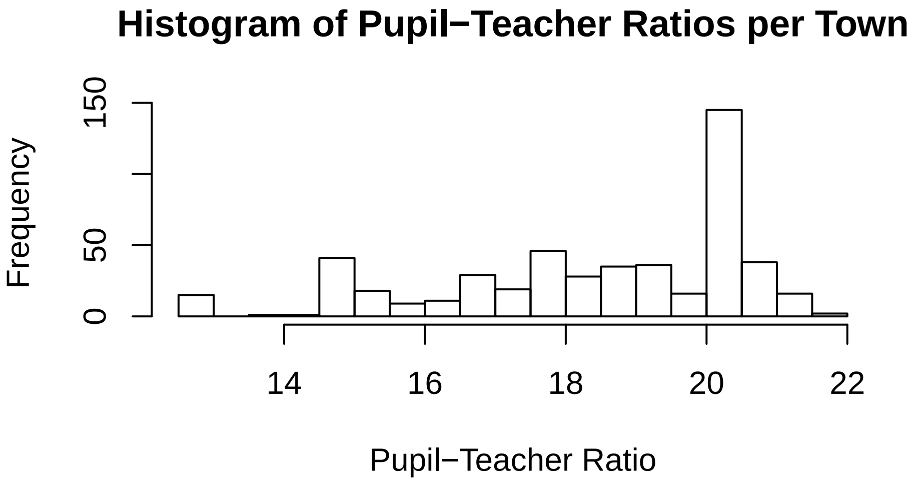
```
##
## Call:
## lm(formula = medv ~ ptratio, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.345      3.029   20.58   <2e-16 ***
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$ptratio)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Pupil-Teacher Ratio")
```



The ratio of pupil to teacher is commonly thought to be an indicator of educational quality: the fewer students a teacher has, the more time the teach can dedicate to each student. We would be reasonable to expect a negative correlation between the ratio of student to teacher and median property value - homes with better education would be worth more. This is what the data shows, with a negative correlation slope of -2.16. The t-statistic is -13.23 and the p-score is less than $2e-16$, meaning the relationship is statisitcally significant.

```
# Lets look at a distribution of P:T ratios
hist(Boston$ptratio, breaks = 30,
     xlab="Pupil-Teacher Ratio",
     main = "Histogram of Pupil-Teacher Ratios per Town")
```



The majority of towns maintain a P:T ratio of 20, and none fall below 12.5 (there is a cost to schools for having a low ratio, as federal funding is frequently on a per-student basis)

Is Median Owner-Occupied Home Value Correlated to The Proportion of Blacks by Town

The *black* variable is a weighted proportion such that:
$black = 1000 \times (Bk - 0.63)^2$ where Bk is the proportion of blacks by town

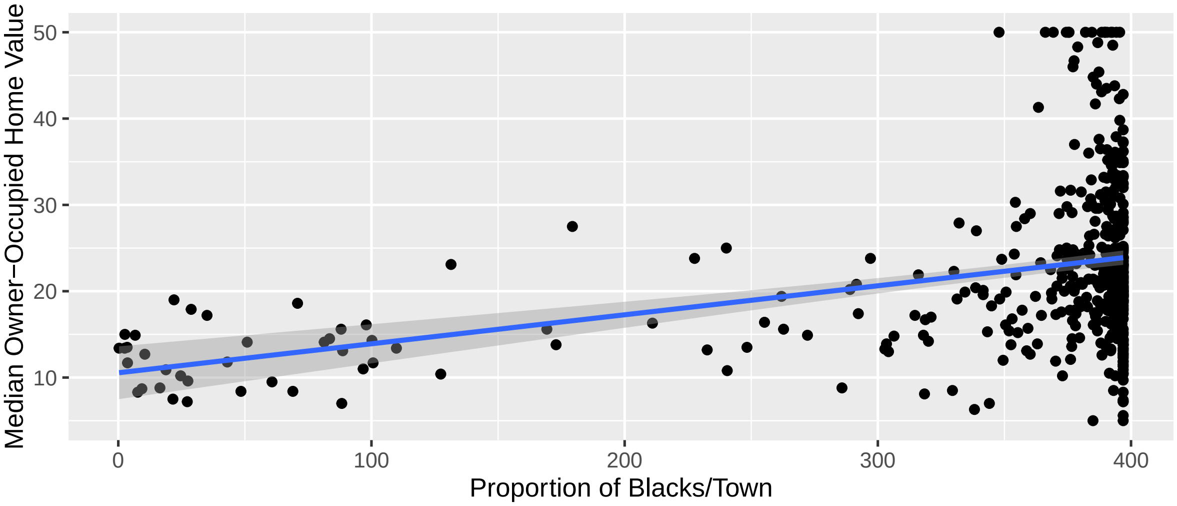
This means that smaller values indicate proportion of blacks by town that are closer to 0.63. As the proportion increases and decreases away from 0.63, the variable value increases. The data appears to have a hard maximum at 400 (the square root of .4 is 0.63). This suggests that when the values are at 400, the proportion of blacks to others in the town are either 0 or 1.26.

```
# Summarize the black.medv model
summary(black.medv)
```

```
##
## 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
```

```
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$black)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Proportion of Blacks/Town")
```



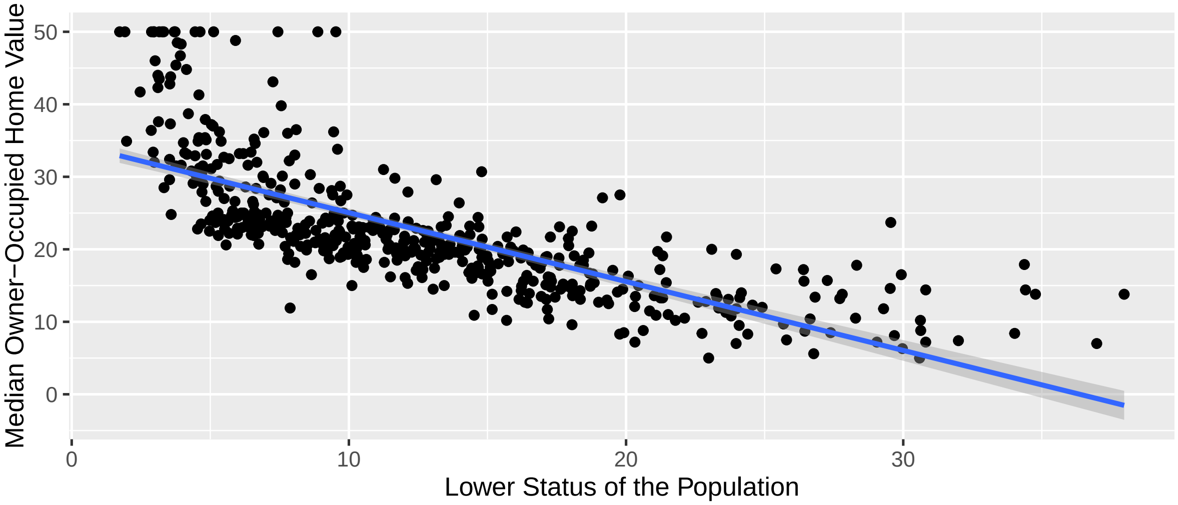
This data exhibits statistical significance and positive correlation with a slope of 0.03, a t-statistic of 7.94, and a p-score of $1.32e - 14$. The problem that I have with this data is the way the data is manipulated. Should we interperet this model as suggesting that an increase in the proportion of black people living in a town correlates positively to an increase in median property value? Should we interpret it as the further away the proportion is from some magic ratio of 0.63, the more the median property value increases. I suspect that the answer is that this manipulation is a means of masking a negative correlation: that property values increase as the proportion of black residents drop below 0.63. I do not know if this manipulation is an attempt to retain some ethical norm, if so, it does so in a questionable way.

Is Median Owner-Occupied Home Value Correlated to Lower Status of the Population (percent)

```r
# Summarize the lstat.medv model
summary(lstat.medv)
```

```
##
## 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
# Construct a dataframe with the variables we are studying
study <- data.frame(y=Boston$medv, x=Boston$lstat)

# Plot a scatter-plot and a linear-regression model of our study
study %>% ggplot(aes(x=x, y=y)) +  geom_point() +
    stat_smooth(method="lm", se=TRUE)+
    ylab("Median Owner-Occupied Home Value") +
    xlab("Lower Status of the Population")
```



These data are negatively correlated with a slope of -0.95, a t-statistic of -24.53 and a p-score of less that $2e - 16$. This means that as the median property value increases, the percentage of lower status population decreases.

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 $H_0 : \beta_j = 0$?

```
multi.regress.model <- lm(Boston$medv ~ Boston$crim + Boston$zn +
                              Boston$indus + Boston$chas + Boston$nox +
                              Boston$rm + Boston$age + Boston$dis +
                              Boston$rad + Boston$tax + Boston$ptratio +
                              Boston$black + Boston$lstat)

summary(multi.regress.model)
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$crim + Boston$zn + Boston$indus +
##     Boston$chas + Boston$nox + Boston$rm + Boston$age + Boston$dis +
##     Boston$rad + Boston$tax + Boston$ptratio + Boston$black +
##     Boston$lstat)
##
## 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 ***
## Boston$crim    -1.080e-01  3.286e-02  -3.287 0.001087 **
## Boston$zn       4.642e-02  1.373e-02   3.382 0.000778 ***
## Boston$indus    2.056e-02  6.150e-02   0.334 0.738288
## Boston$chas     2.687e+00  8.616e-01   3.118 0.001925 **
## Boston$nox     -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## Boston$rm       3.810e+00  4.179e-01   9.116  < 2e-16 ***
## Boston$age      6.922e-04  1.321e-02   0.052 0.958229
## Boston$dis     -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## Boston$rad      3.060e-01  6.635e-02   4.613 5.07e-06 ***
## Boston$tax     -1.233e-02  3.760e-03  -3.280 0.001112 **
## Boston$ptratio -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## Boston$black    9.312e-03  2.686e-03   3.467 0.000573 ***
## Boston$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
```

```
# SPOILERS!!!
# We are going to need the adj.r.squared of this model for question # 7!
question.4.adj.r.squared <- summary(multi.regress.model)$adj.r.squared
```

Based on the p-scores, we can reject the null hypothesis H_0 on each of the variables except *indus* and *age*. That means we can reject the null hypothesis on *Boston$crim*, *Boston$zn*, *Boston$chas*, *Boston$nox*, *Boston$rm*, *Boston$dis*, *Boston$rad*, *Boston$tax*, *Boston$ptratio*, *Boston$black*, and *Boston$lstat*.

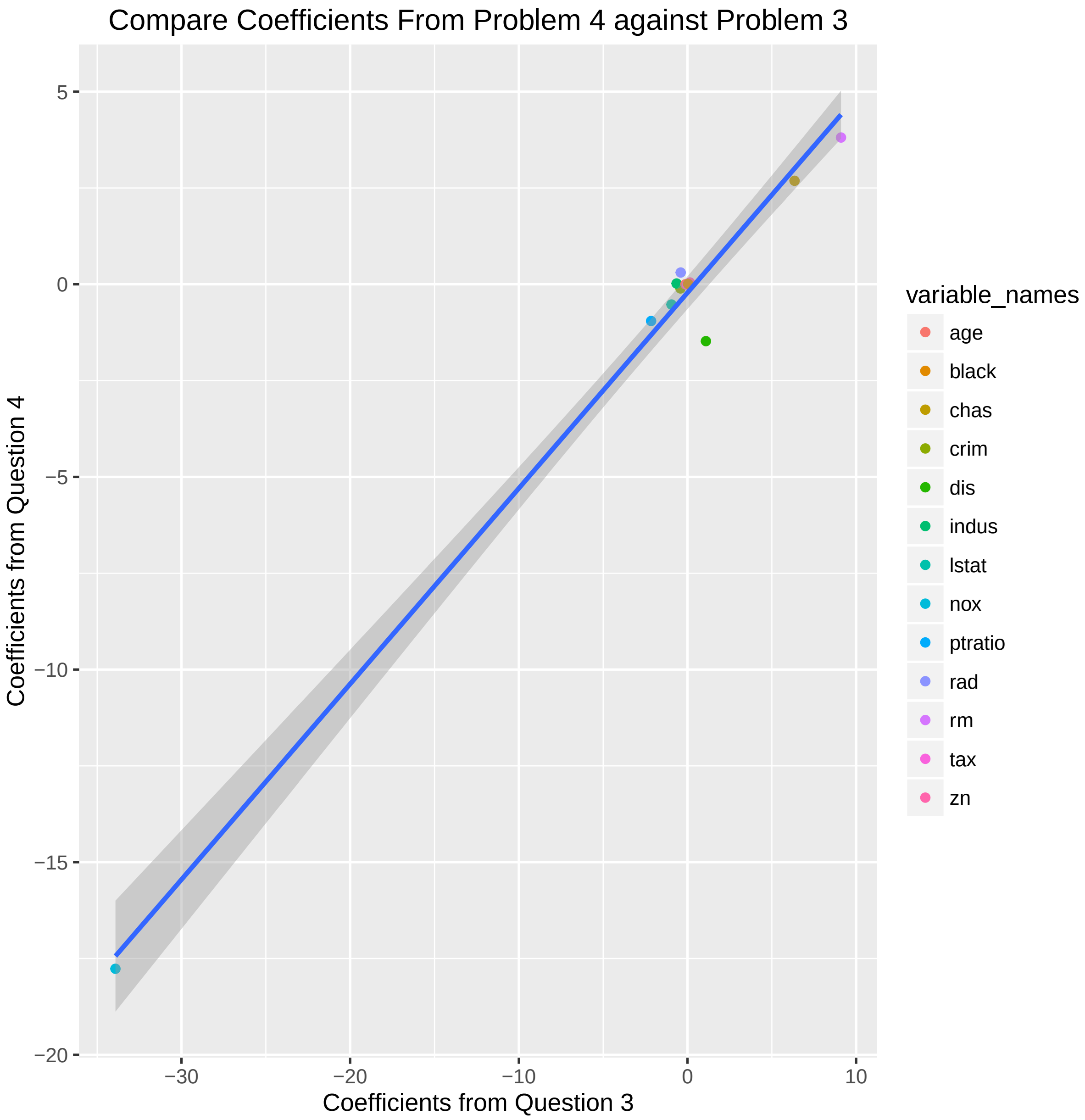
5. How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response.

```r
q3.coef <- data.frame(coefficients(crim.medv)[2], coefficients(zn.medv)[2],
                      coefficients(indus.medv)[2], coefficients(chas.medv)[2],
                      coefficients(nox.medv)[2], coefficients(rm.medv)[2],
                      coefficients(age.medv)[2], coefficients(dis.medv)[2],
                      coefficients(rad.medv)[2], coefficients(tax.medv)[2],
                      coefficients(ptratio.medv)[2], coefficients(black.medv)[2],
                      coefficients(lstat.medv)[2])

q4.coef <- coefficients(multi.regress.model)[2:14]

compare.coef <- data.frame(x=unlist(q3.coef),
                           y=unlist(q4.coef),
                           variable_names=factor(names(Boston)[1:13]))

compare.coef
```

```
##                                          x             y variable_names
## coefficients.crim.medv..2.     -0.41519028 -1.080114e-01           crim
## coefficients.zn.medv..2.        0.14213999  4.642046e-02             zn
## coefficients.indus.medv..2.    -0.64849005  2.055863e-02          indus
## coefficients.chas.medv..2.      6.34615711  2.686734e+00           chas
## coefficients.nox.medv..2.     -33.91605501 -1.776661e+01            nox
## coefficients.rm.medv..2.        9.10210898  3.809865e+00             rm
## coefficients.age.medv..2.      -0.12316272  6.922246e-04            age
## coefficients.dis.medv..2.       1.09161302 -1.475567e+00            dis
## coefficients.rad.medv..2.      -0.40309540  3.060495e-01            rad
## coefficients.tax.medv..2.      -0.02556810 -1.233459e-02            tax
## coefficients.ptratio.medv..2.  -2.15717530 -9.527472e-01        ptratio
## coefficients.black.medv..2.     0.03359306  9.311683e-03          black
## coefficients.lstat.medv..2.    -0.95004935 -5.247584e-01          lstat
```

The *compare.coef* data frame shows each of the coeffients from the univariate models (field x) and the multivariate model (field y). The *variable_names* field was added for the benefit of the plot legend below. Here, and in the plot, we can see that the multivariate model has changed coefficients. If the coefficients had not changed, all of the values would be the same, and they would all lie on the line $y = x$. In fact, the slope of the linear regression model of these coordinages is 0.50773, and not 1. As for the extent of change, the RSE of coefficients of Question 4 over Question 3 is 0.6855.

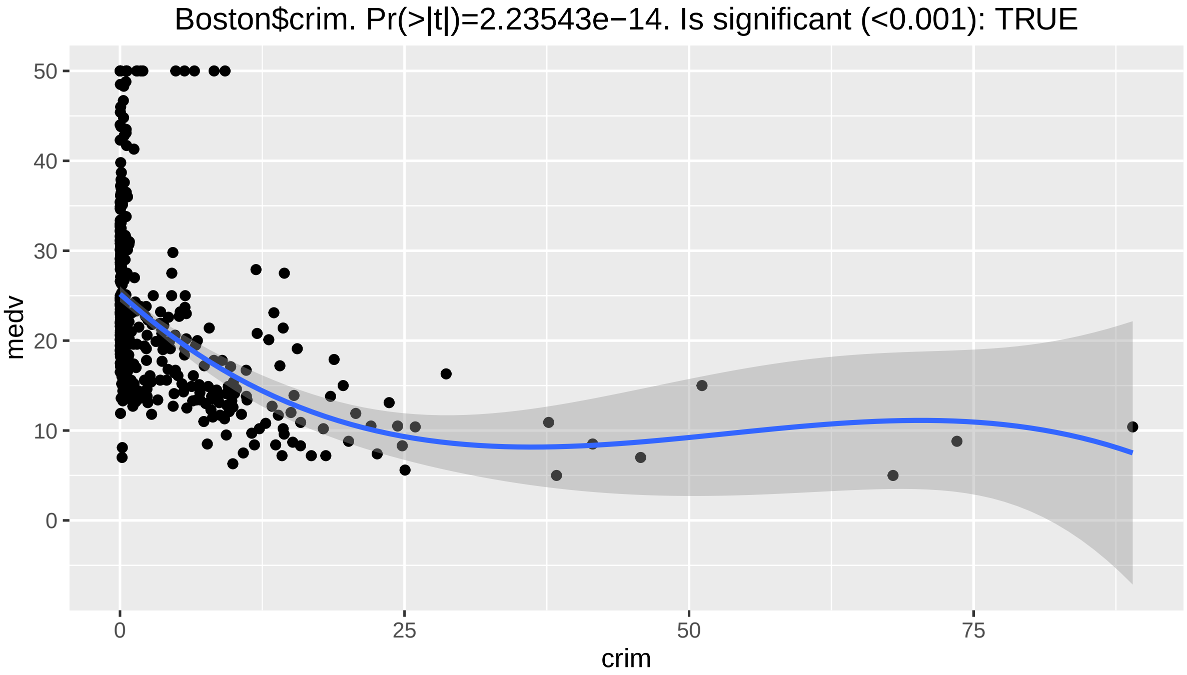
```r
compare.coef %>%
    ggplot(aes(x=x, y=y, color=variable_names)) +
    geom_point() +
    stat_smooth(method="lm", se=TRUE, inherit.aes = FALSE, aes(x=x, y=y))+
    xlab("Coefficients from Question 3") +
    ylab("Coefficients from Question 4") +
    ggtitle("Compare Coefficients From Problem 4 against Problem 3")
```



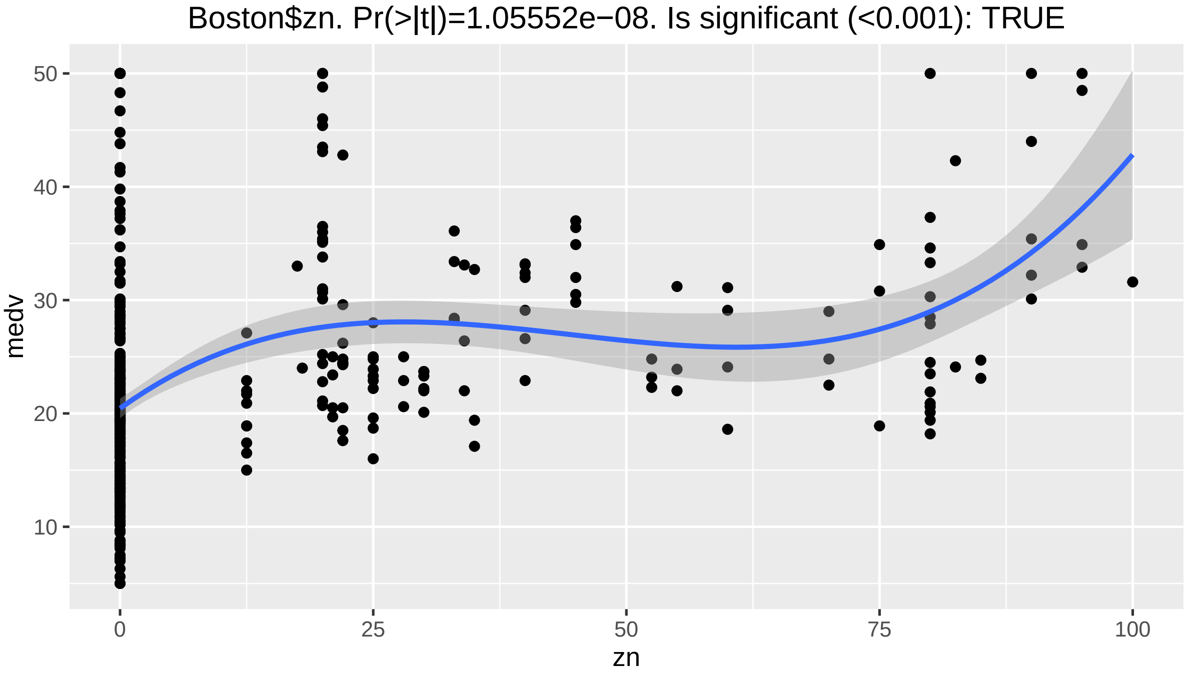
6. Is there evidence of a non-linear association between any of the predictors and the response? To answer this question, for each predictor X fit a model of the form:

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

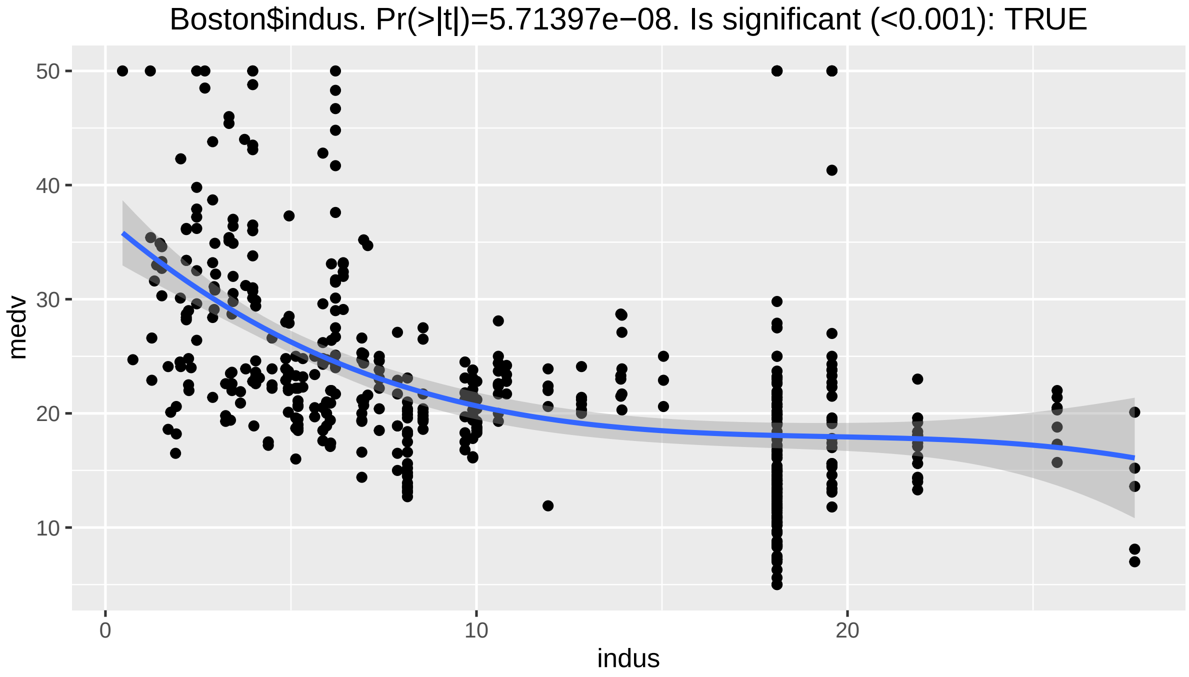
```
# This method is effective at producing plots, but hard to format.  For clarity,
# we will print a line between plots.
a="============================================================================"

# For each variable in Boston, except medv
for(i in seq(1,13)){
    # construct a dataframe consisting of the variable and the medv variable
    test.df = data.frame(Boston$medv,
                         Boston[i])
    # Change the variable names for predictability
    names(test.df) <- c("response", "predictor")

    # Generate a polynomial model between the predictor and response variable
    # NOTE: we forgo the use of poly() so we can attempt to plot the Boston$chas
    # variable, which only has two values
    s <- summary(lm(formula = response ~ predictor + I(predictor^2) + I(predictor^3),
                    data=test.df))

    # Use a report of the p-score as the plot title
    p.title <- paste("Boston$",
                names(Boston)[i],". Pr(>|t|)=",
                gettextf("%g",coefficients(s)[2,4]),
                ". Is significant (<0.001): ",
                coefficients(s)[2,4] < 0.0001,
                sep="")

    # Generate a plot of the polynomial regression model to see if
    # it looks like a good fit
    plt <- test.df %>%
        ggplot(aes(x=predictor, y=response)) +
        geom_point() +
        stat_smooth(method="lm",
                    formula = y ~ x + I(x^2) + I(x^3),
                    se=TRUE) +
        ylab("medv") + xlab(names(Boston)[i]) +
        ggtitle(p.title)

    # Due to known bug in knitr, we have to print plots within loops
    print(plt)
    print(a)
}
```


Boston$crim. Pr(>|t|)=2.23543e−14. Is significant (<0.001): TRUE
medv
50
40
30
20
10
0
0
25
50
75
crim

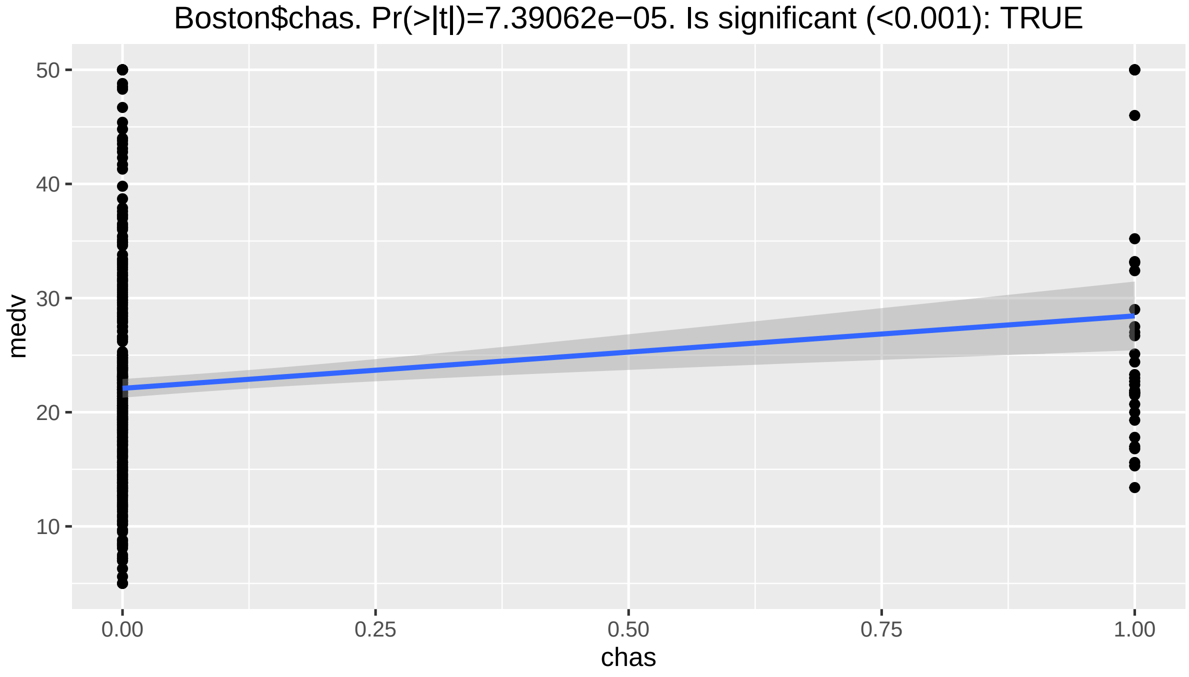
```
## [1] "=================================================================================="
```


Boston$zn. Pr(>|t|)=1.05552e−08. Is significant (<0.001): TRUE
medv
50
40
30
20
10
0
25
50
75
100
zn

```
## [1] "=================================================================================="
```

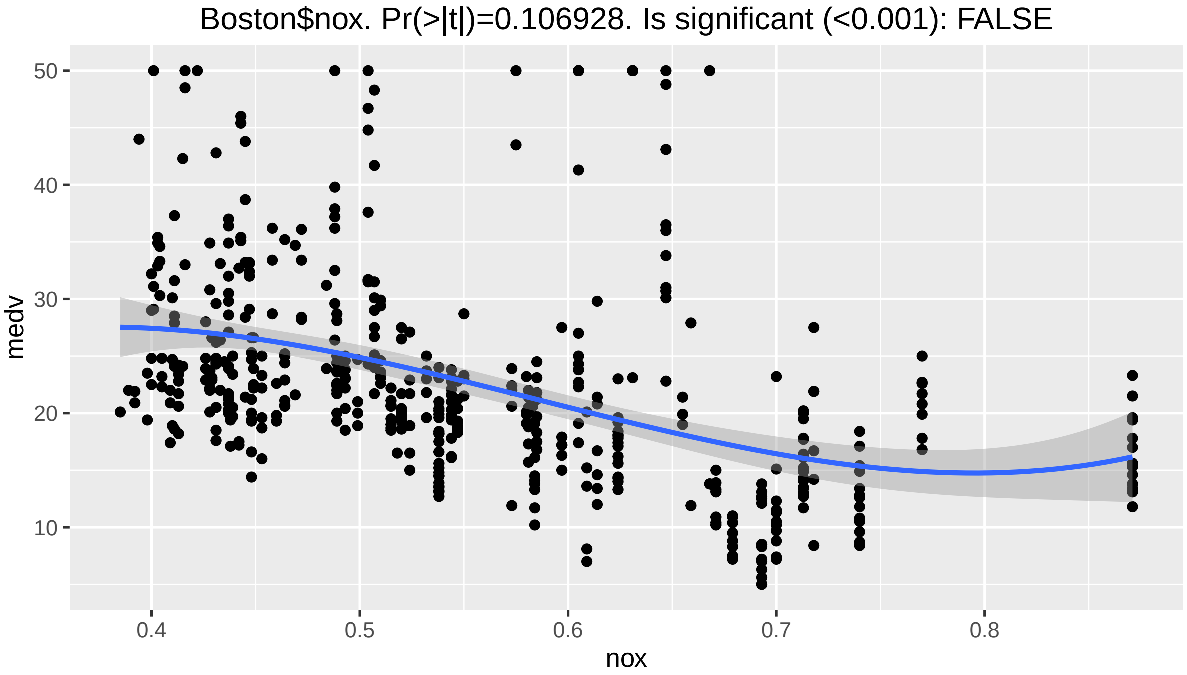

Boston$indus. Pr(>|t|)=5.71397e−08. Is significant (<0.001): TRUE
medv
50
40
30
20
10
0
10
20
indus

```
## [1] "==============================================================================="
```


Boston$chas. Pr(>|t|)=7.39062e−05. Is significant (<0.001): TRUE
medv
50
40
30
20
10
0.00
0.25
0.50
0.75
1.00
chas

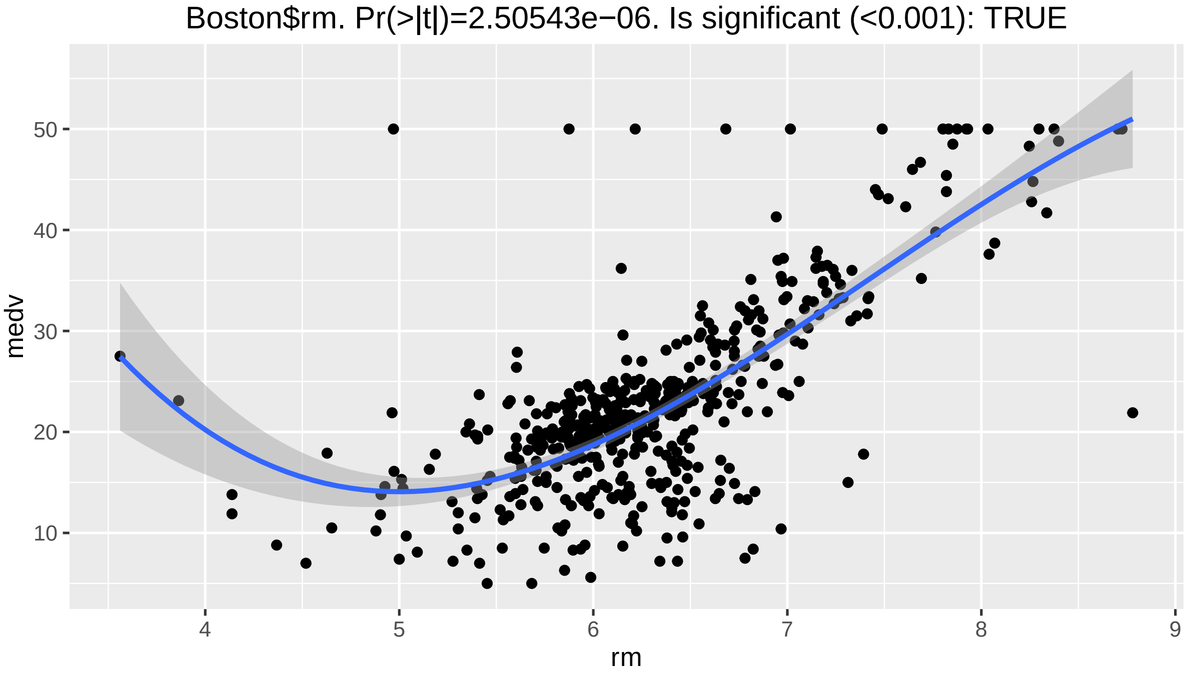
```
## [1] "==============================================================================="
```

Boston$nox. Pr(>|t|)=0.106928. Is significant (<0.001): FALSE

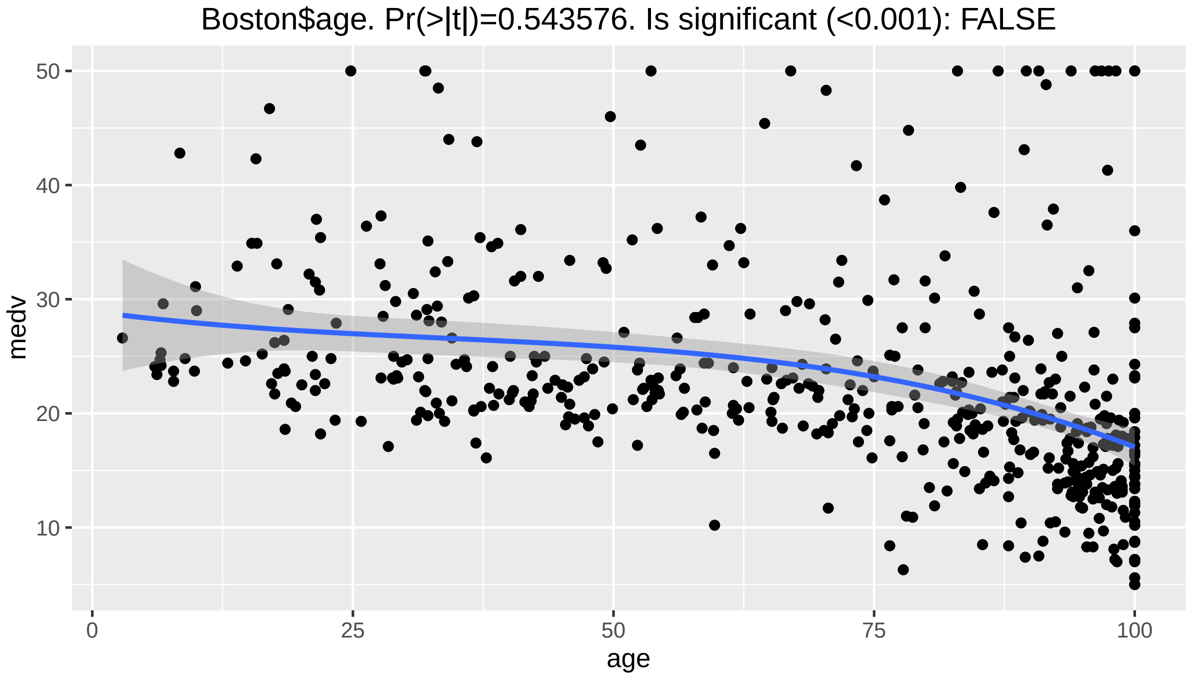


```
## [1] "========================================================================================"
```

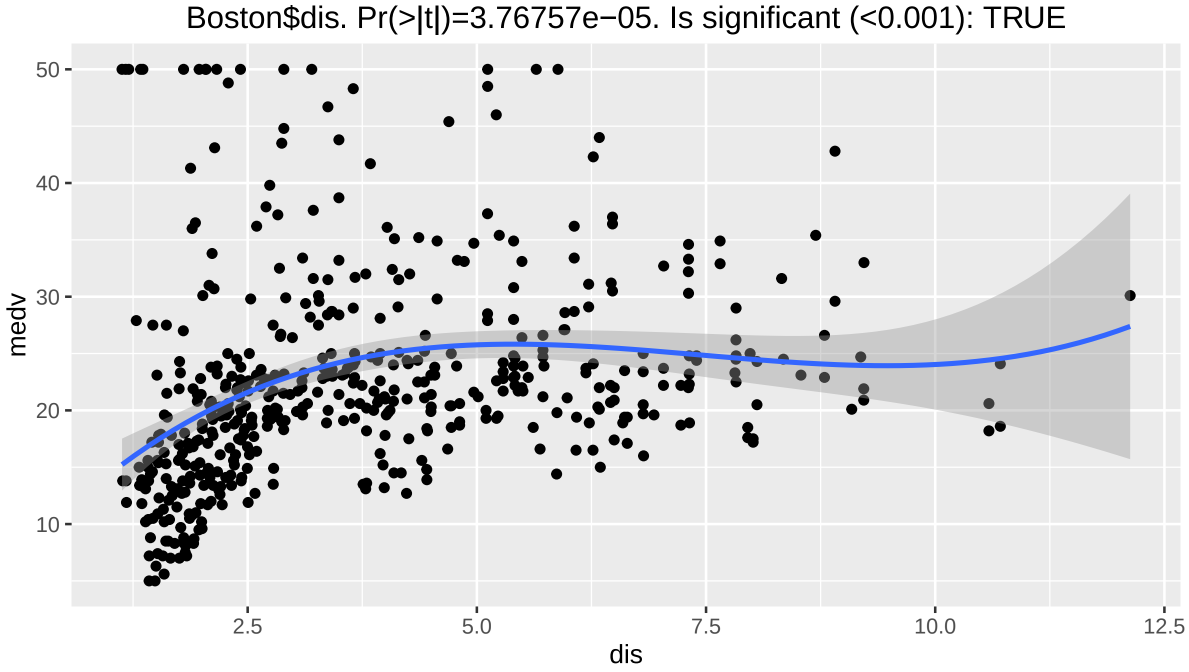
Boston$rm. Pr(>|t|)=2.50543e−06. Is significant (<0.001): TRUE



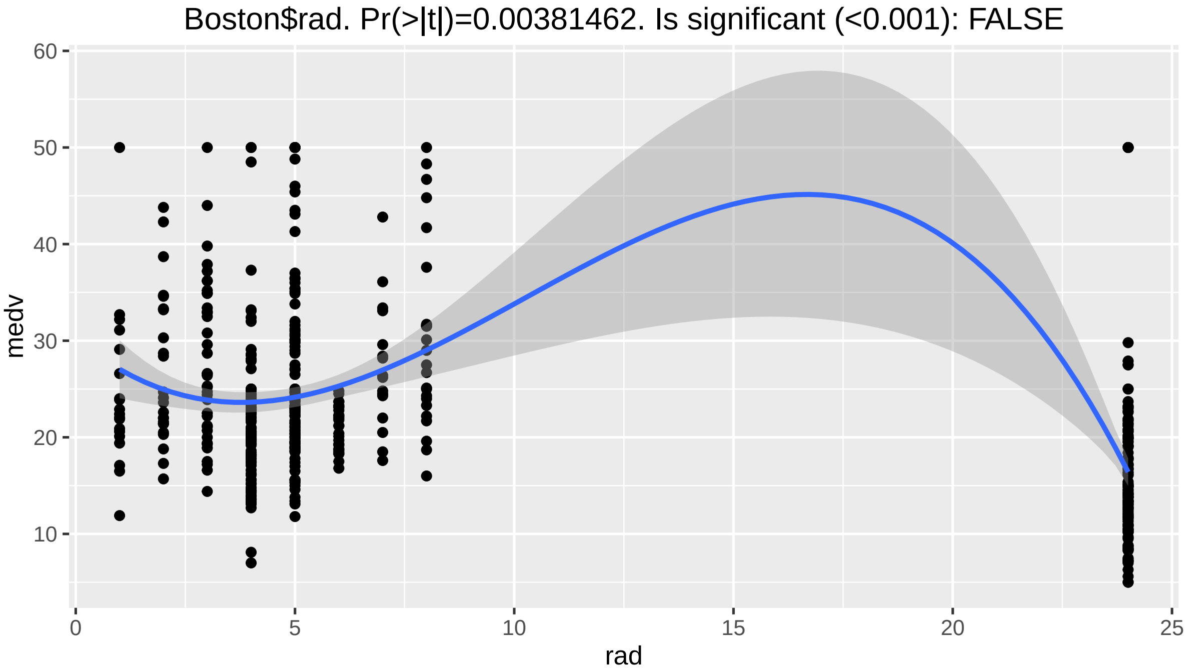
```
## [1] "========================================================================================"
```


Boston$age. Pr(>|t|)=0.543576. Is significant (<0.001): FALSE
medv
50
40
30
20
10
0
25
50
75
100
age

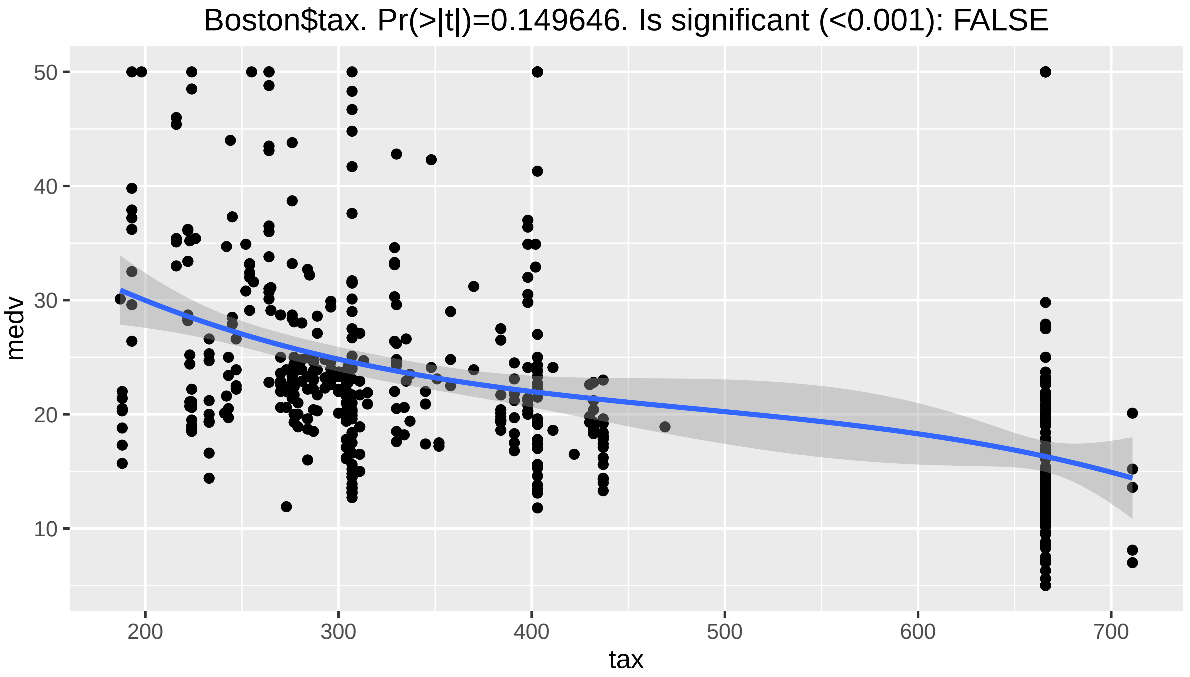
```
## [1] "==========================================================================="
```


Boston$dis. Pr(>|t|)=3.76757e−05. Is significant (<0.001): TRUE
medv
50
40
30
20
10
2.5
5.0
7.5
10.0
12.5
dis

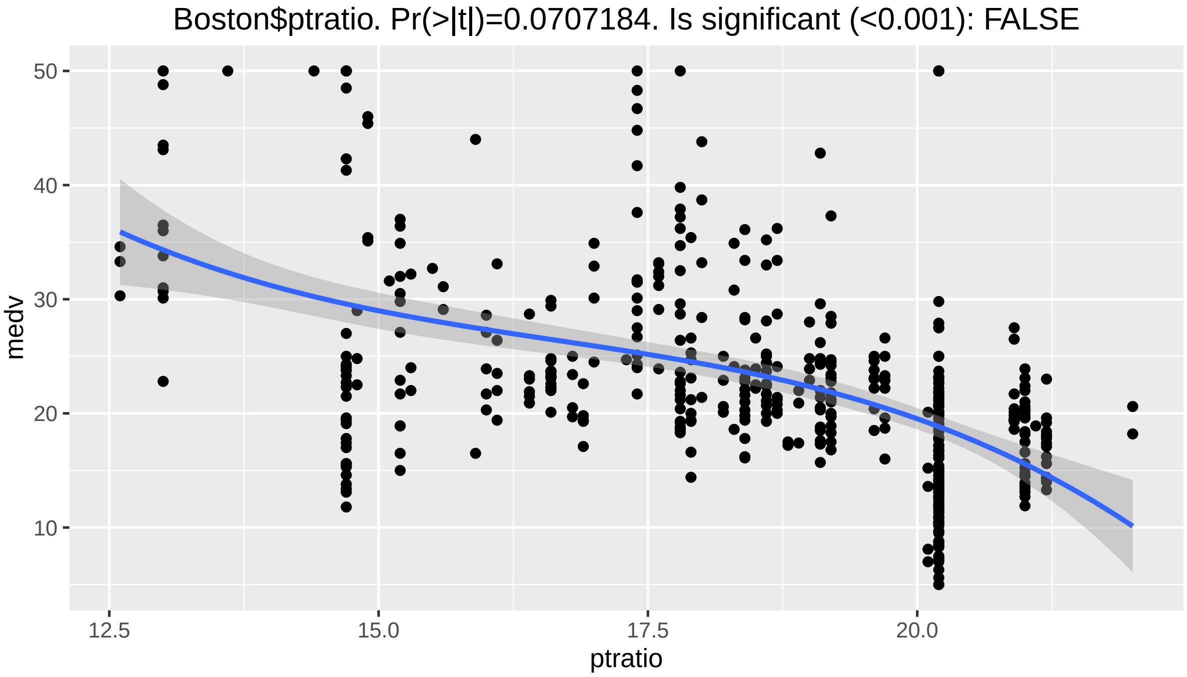
```
## [1] "==========================================================================="
```


Boston$rad. Pr(>|t|)=0.00381462. Is significant (<0.001): FALSE

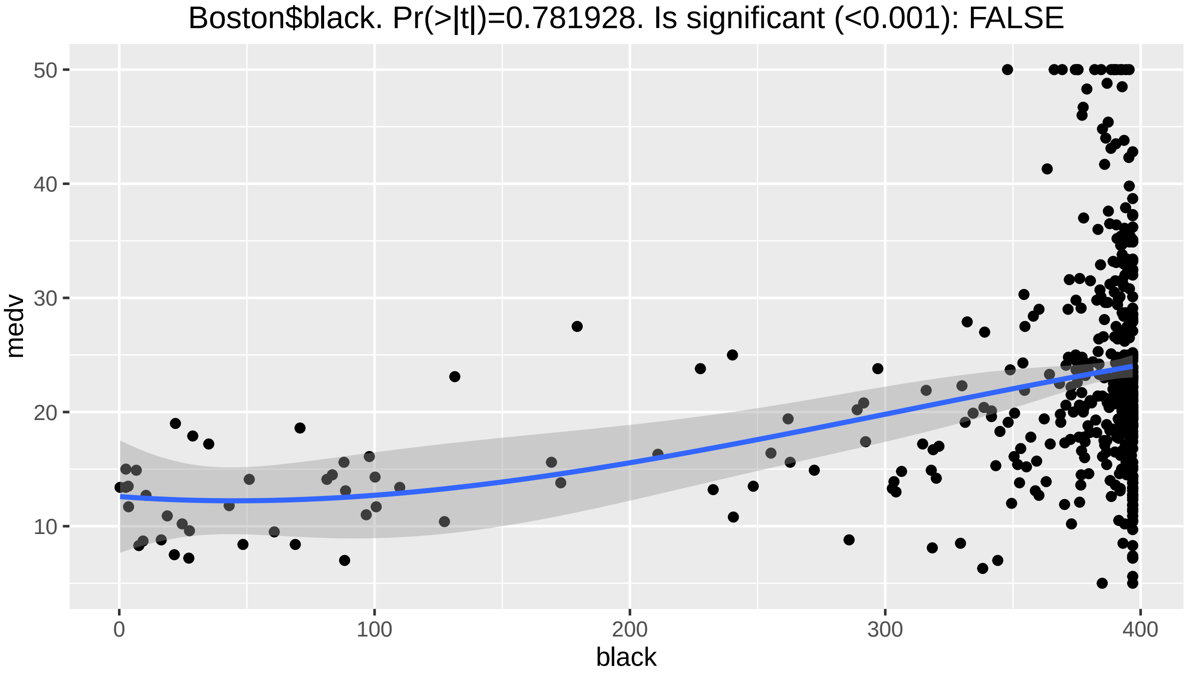
```
## [1] "==========================================================================="
```


Boston$tax. Pr(>|t|)=0.149646. Is significant (<0.001): FALSE

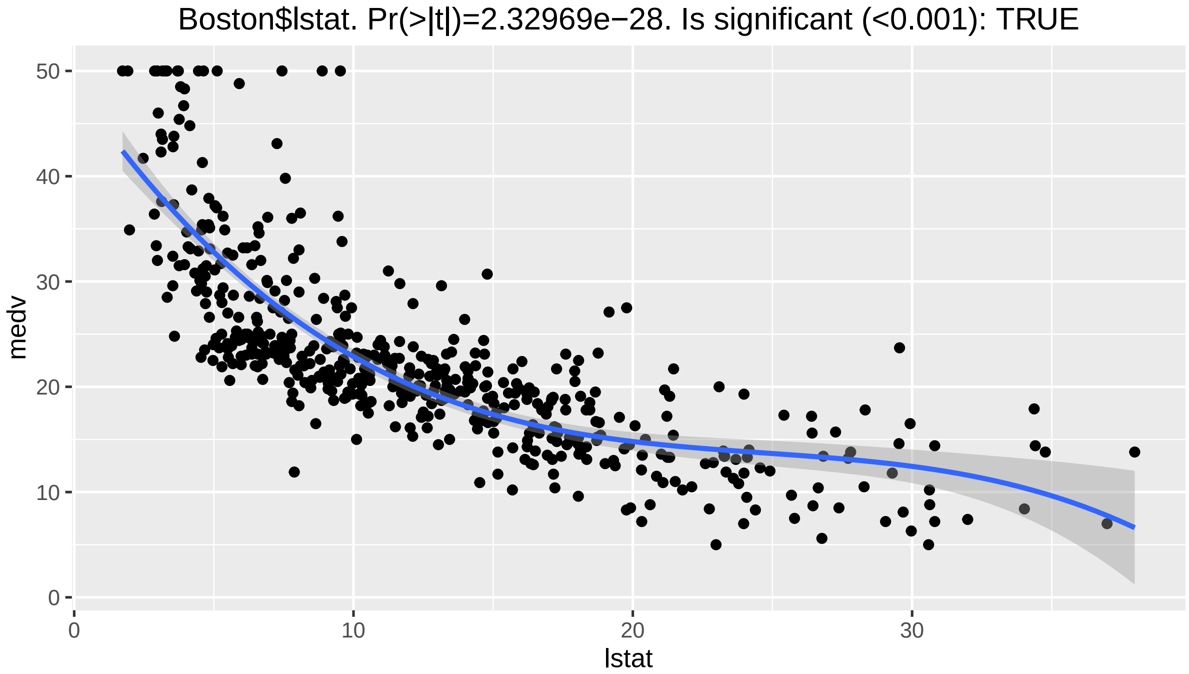
```
## [1] "==========================================================================="
```


Boston$ptratio. Pr(>|t|)=0.0707184. Is significant (<0.001): FALSE

```
## [1] "==========================================================================="
```


Boston$black. Pr(>|t|)=0.781928. Is significant (<0.001): FALSE

```
## [1] "==========================================================================="
```



```
## [1] "=============================================================================="
```

There is evidence of non-linear associations between many of the predictors and *medv*. The following variables produce a p-score below 0.001: *crim*, *zn*, *indus*, *chas*, *rm*, *dis*, *lstat*. The following produce a p-score between 0.01 and 0.001: *rad*. Even when the p-score is not as low as we might want, many of the fitted linear regression lines in plots are compelling.

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)?

I personally like the *Backward elimination* approach to stepwise model selection. The strategy is to remove one variable at a time based on whichever removal results in the hightest R^2_{adj}. When removing a variable cannot produce a highter R^2_{adj} than the last model, we are done.

```
# To begin with, our baseline model has an adj.r.squared value that we want to
# improve on by eliminating variables.

score.to.beat <- question.4.adj.r.squared

# initialize a couple of helpful structures
variables.removed <- c()
test.formula.templ <- "medv ~ . "

# Loop once per variable, minus one (we might remove all but one variable)
for (test.each in seq(1,12)){
    # initialize a vector to hold scores - default values to 0
    scores <- vector(mode="integer", length=13)

    # test the removal of each (remaining) variable
    # and put the adj.r.squared into the corresponding scores vector
    for(var.to.remove in seq(1,13)){

        # we will just skip over variables already removed
        if (var.to.remove %in% variables.removed) {next}
        # Produce a test formula that removes the var.to.remove and those
        # variables.removed
        test.formula <- paste(test.formula.templ,
                              "-",
                              names(Boston)[var.to.remove])

        for (already.removed in variables.removed){
            test.formula <- paste(test.formula,
                                  "-",
                                  names(Boston)[already.removed])
        }

        # capture the score of the test
        scores[var.to.remove] = summary(
            lm(formula=test.formula, data=Boston)
            )$adj.r.squared
    }

    # Decide which score is the new score.to.beat, and add the corresponding
    # index to variables.removed.  If the best score is not an improvement,
    # we are done

    best.idx = which.max(scores)

    if (scores[best.idx] > score.to.beat) {
        score.to.beat <- scores[best.idx]
        variables.removed <- append(variables.removed, best.idx)
    }else{
        print(paste("Ending because our latest best score of",
                    scores[best.idx]))
        print(paste("is not better than our score to beat of",
                    score.to.beat))
        break
    }
}
```

```
## [1] "Ending because our latest best score of 0.729914928077186"
## [1] "is not better than our score to beat of 0.734805772327457"
```

```
# How did we do?
question.4.adj.r.squared
```

```
## [1] 0.7337897
```

```
score.to.beat
```

```
## [1] 0.7348058
```

```
# So, what is the formula for the best model?
final.formula <- test.formula.templ
for (already.removed in variables.removed){
    final.formula <- paste(final.formula,
                           "-",
                           names(Boston)[already.removed])
}
final.formula
```

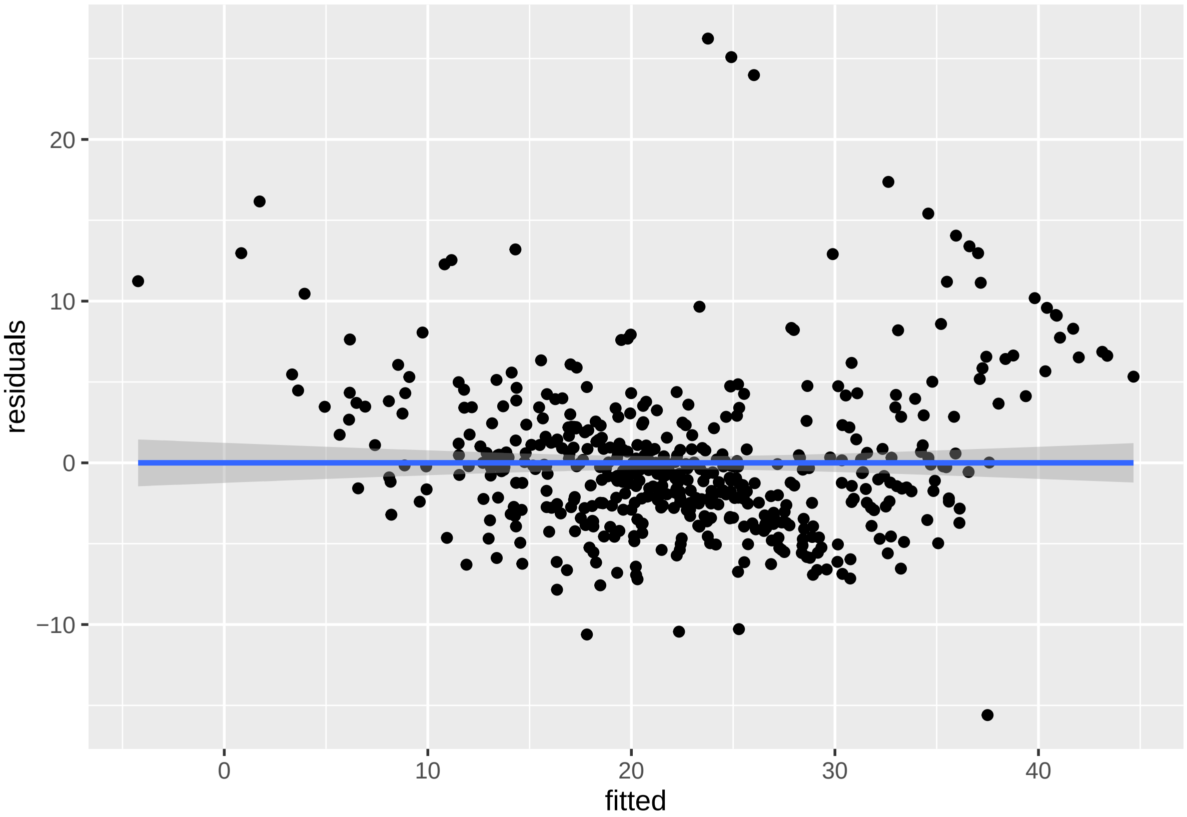
```
## [1] "medv ~ .  - age - indus"
```

```
# Summarize the stepwise model
stepwise.model <- lm(final.formula, data=Boston)

summary(stepwise.model)
```

```
## 
## Call:
## lm(formula = final.formula, 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
```

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.

```
# Let's plot the residuals against fitted values
to.plot = data.frame(fitted = stepwise.model$fitted.values,
                     residuals = stepwise.model$residuals)
to.plot %>%
    ggplot(aes(x=fitted, y=residuals)) +
    geom_point()+
    geom_smooth(method="lm", se=TRUE)
```

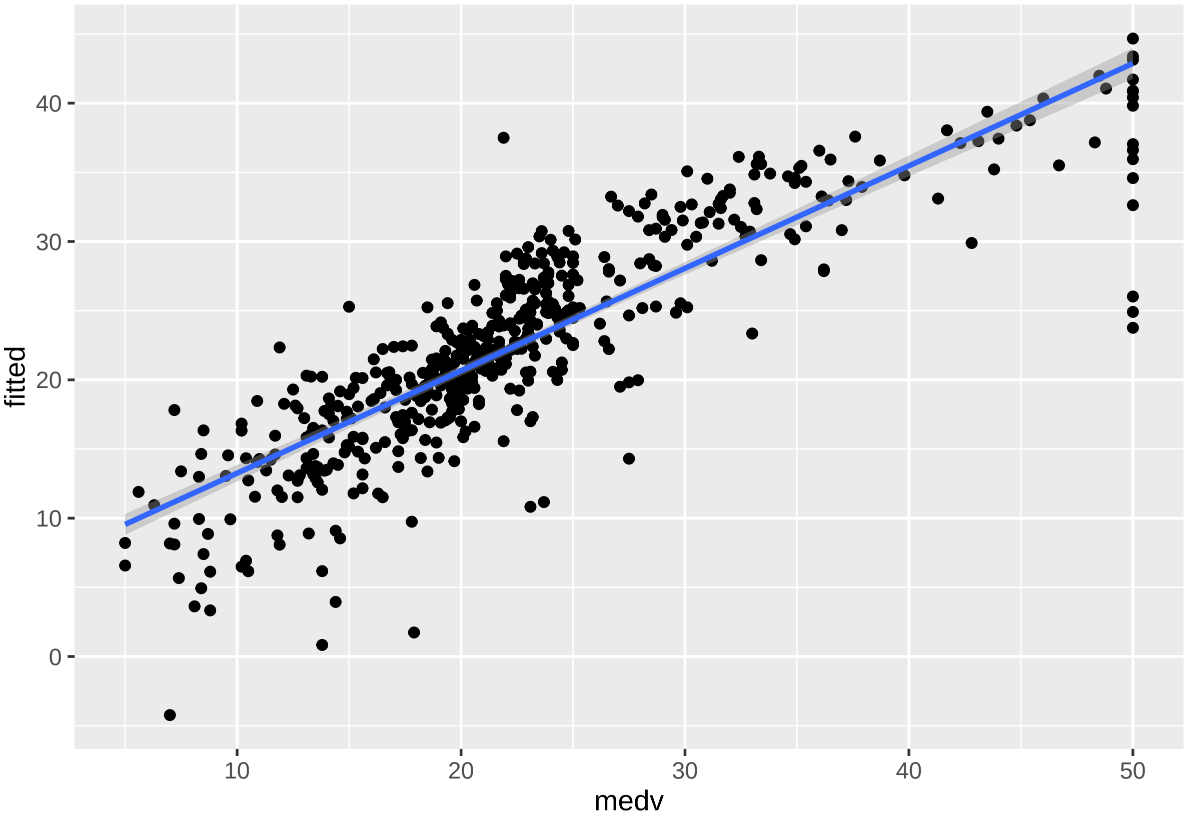


Our residual values appear evenly distributted over our fitted values, such that a linear regression line over the plot shows no slope ($\beta_1 = 0$).

It is interesting that the residuals greater than are much more diffuse in both x and y dimenstions than those that are less than zero - these range between x=5 to 37, y=-10 to 0. There is also an interesting pattern at the upper-left of the plot where a series of residuals appear to share a negative correlation to fitted values in a well-defined line. They may be related to the points on the next plot that are at *medv* = 50, as they share a similar range along the *fitted* dimension.

The stepwise method successfully removed the two variables that were shown to be statistically insignificant in question 4: *indus* and *age*.

```
# Let's plot the fitted values against Boston$medv
to.plot = data.frame(medv=stepwise.model$model$medv,
                     fitted=stepwise.model$fitted.values)

to.plot %>%
    ggplot(aes(y=fitted, x=medv)) +
    geom_point()+
    geom_smooth(method="lm", se=TRUE)
```



This appears to be a fairly good fit (note the points at $medv = 50$, as mentioned above).