# Regression Analysis of Boston Suburbs

*Mashhood Syed*

*November 1, 2014*

This dataset represents housing values in suburbs of Boston in the late 1970's. It was created in 1978 by D. Harrison and D.L. Rubinfeld. The dataset was found at UCI's machine learning repository.

Here is a description of each of the attributes (by column number):

1. per capita crime rate by town
2. proportion of residential land zoned for lots over 25,000 sq.ft.
3. proportion of non-retail business acres per town
4. Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. nitric oxides concentration (parts per 10 million)
6. average number of rooms per dwelling
7. proportion of owner-occupied units built prior to 1940
8. weighted distances to five Boston employment centres
9. index of accessibility to radial highways
10. full-value property-tax rate per $10,000
11. pupil-teacher ratio by town
12. 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. % lower status of the population
14. Median value of owner-occupied homes in $1000's

Here is what I want to predict: median_home_value (column 14)

Here are the 4 continuous variables I am using to predict median_home_value:

pupil_teacher_ratio (column 11) percent_lower_status_of_population (column 13) access_to_radial_highways (column 9) age_of_dwelling_since_1940 (column 7)

```
setwd("~/Desktop/PS3")
dataFile = "~/Desktop/PS3/housing.csv"
data = read.csv(dataFile, header=TRUE)
library(foreign)
library(ggplot2)
attach(data)
summary(data)
```

```
##    crime_rate        zoning        industrial    charles_river
##  Min.   : 0.01   Min.   :  0.0   Min.   : 0.46   Min.   :0.0000
##  1st Qu.: 0.08   1st Qu.:  0.0   1st Qu.: 5.19   1st Qu.:0.0000
##  Median : 0.26   Median :  0.0   Median : 9.69   Median :0.0000
##  Mean   : 3.61   Mean   : 11.4   Mean   :11.14   Mean   :0.0692
##  3rd Qu.: 3.68   3rd Qu.: 12.5   3rd Qu.:18.10   3rd Qu.:0.0000
##  Max.   :88.98   Max.   :100.0   Max.   :27.74   Max.   :1.0000
##  nitric_oxide_conc  num_of_rooms  age_of_dwelling_since_1940
##  Min.   :0.385     Min.   :3.56   Min.   :  2.9
##  1st Qu.:0.449     1st Qu.:5.89   1st Qu.: 45.0
##  Median :0.538     Median :6.21   Median : 77.5
```

```
##  Mean   :0.555     Mean   :6.29   Mean   : 68.6
##  3rd Qu.:0.624     3rd Qu.:6.62   3rd Qu.: 94.1
##  Max.   :0.871     Max.   :8.78   Max.   :100.0
##  distance_to_employment_centers access_to_highway tax_amt_per_10k
##  Min.   : 1.13                  Min.   : 1.00     Min.   :187
##  1st Qu.: 2.10                  1st Qu.: 4.00     1st Qu.:279
##  Median : 3.21                  Median : 5.00     Median :330
##  Mean   : 3.79                  Mean   : 9.55     Mean   :408
##  3rd Qu.: 5.19                  3rd Qu.:24.00     3rd Qu.:666
##  Max.   :12.13                  Max.   :24.00     Max.   :711
##  pupil_teacher_ratio proportion_of_blacks
##  Min.   :12.6        Min.   :  0.3
##  1st Qu.:17.4        1st Qu.:375.4
##  Median :19.1        Median :391.4
##  Mean   :18.5        Mean   :356.7
##  3rd Qu.:20.2        3rd Qu.:396.2
##  Max.   :22.0        Max.   :396.9
##  percent_lower_status_of_population median_home_value_in_1k_increments
##  Min.   : 1.73                      Min.   : 5.0
##  1st Qu.: 6.95                      1st Qu.:17.0
##  Median :11.36                      Median :21.2
##  Mean   :12.65                      Mean   :22.5
##  3rd Qu.:16.95                      3rd Qu.:25.0
##  Max.   :37.97                      Max.   :50.0
```

Step 1: Coefficient Expectation

I expect the coefficients to each have a negative sign. For a majority of the variables, their is a inverse relationship between the price of a home and things like crime, % of people in a lower social status living in the area, pupil to teacher ratio, number of older homes. As each predictor gets larger the price of the home goes down.

First we start with the Pupil to Teacher Ratio

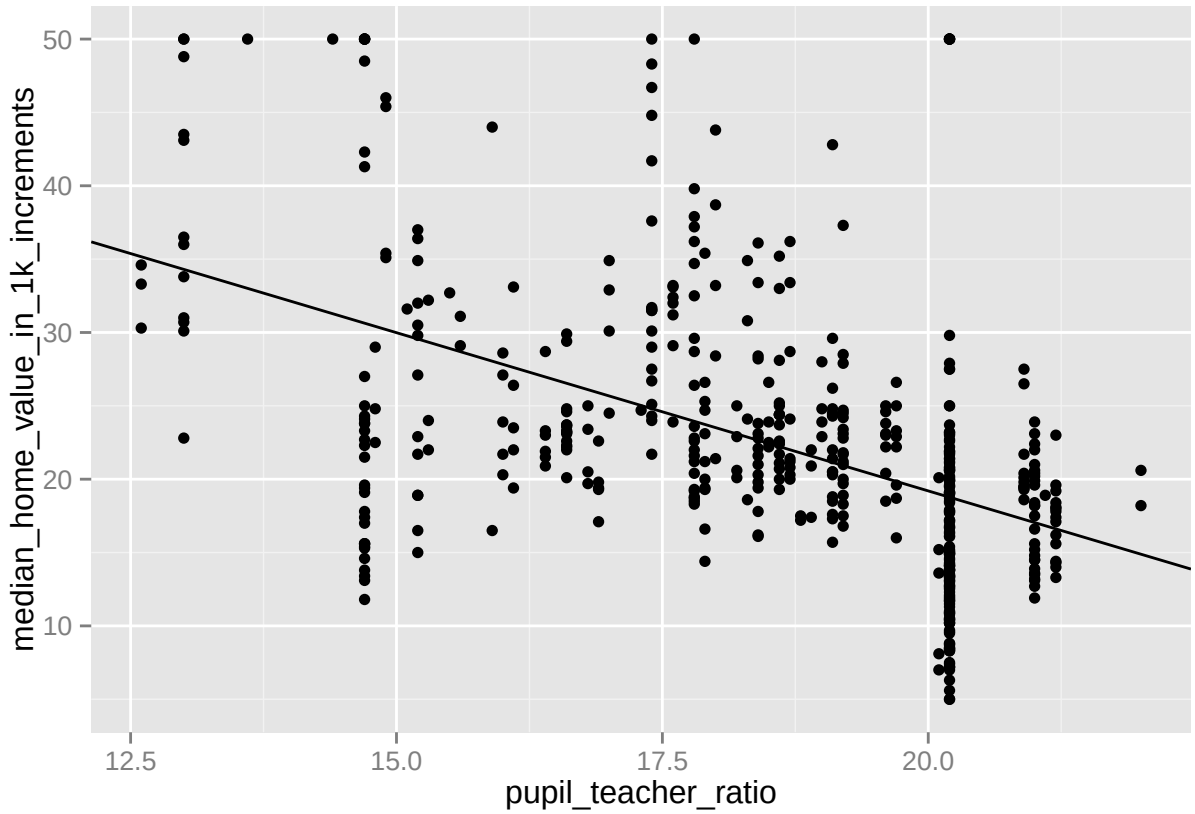
```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ pupil_teacher_ratio)
##
## Coefficients:
##         (Intercept)  pupil_teacher_ratio
##               62.34                -2.16
```

```
##             2.5 % 97.5 %
## (Intercept) 56.39   68.3
```

```
##                      2.5 % 97.5 %
## pupil_teacher_ratio -2.477 -1.837
```

```
## [1] "Sigma: 7.93100371044508"
```

```
## [1] "Adj R-squared: 0.256374792846544"
```




Now lets look at Percent of Population that is considered lower status

```
fit.2 = lm(median_home_value_in_1k_increments ~ percent_lower_status_of_population)
print(fit.2)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ percent_lower_status_of_population)
##
## Coefficients:
##                          (Intercept)  percent_lower_status_of_population
##                                34.55                               -0.95
```

```
confint(fit.2, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 33.45  35.66
```

```
confint(fit.2, "percent_lower_status_of_population", level = 0.95)
```

```
##                                    2.5 % 97.5 %
## percent_lower_status_of_population -1.026 -0.874
```

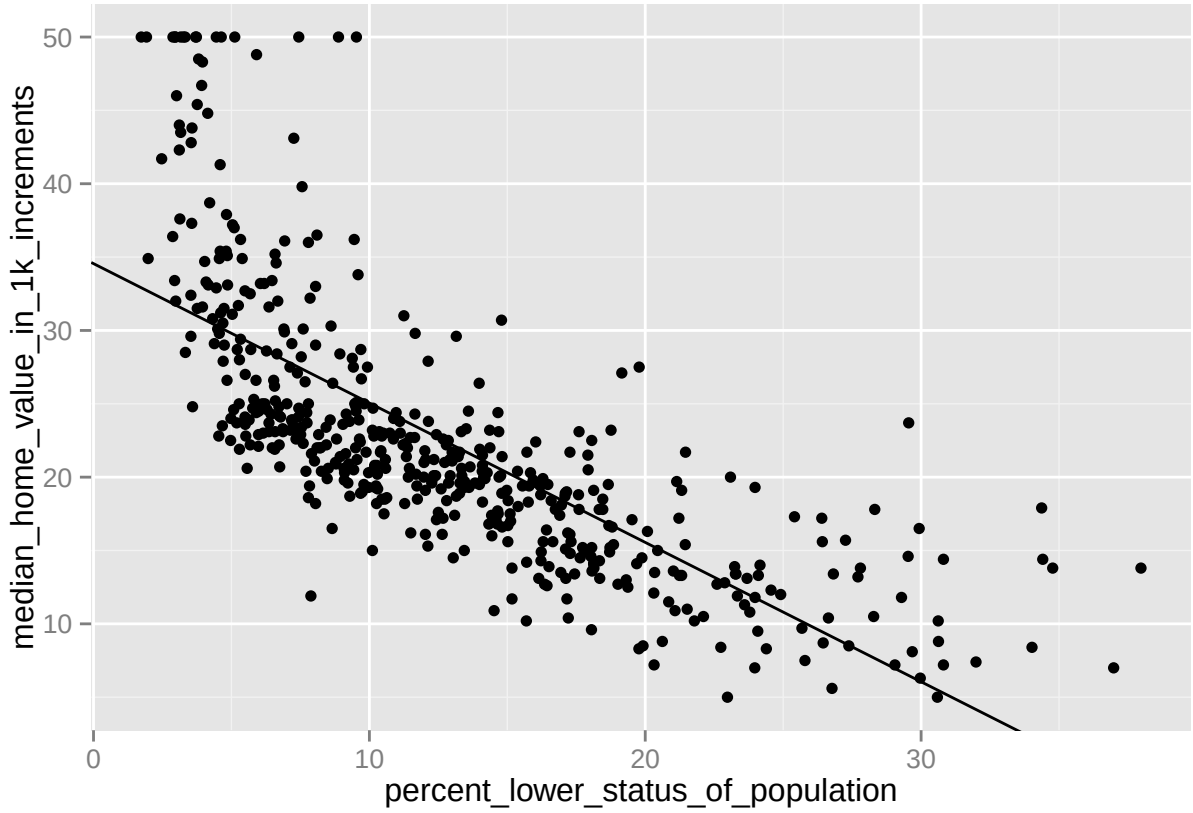
```
print(paste0("Sigma: ", summary(fit.2)$sigma))
```

```
## [1] "Sigma: 6.21576040539807"
```

```
print(paste0("Adj R-squared: ", summary(fit.2)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.543241825954707"
```

```
ggplot(data = data, aes(x = percent_lower_status_of_population, y = median_home_value_in_1k_increments)
```




Next we look at Access to highways

```
fit.3 = lm(median_home_value_in_1k_increments ~ access_to_highway)
print(fit.3)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ access_to_highway)
##
## Coefficients:
##       (Intercept)  access_to_highway
##            26.382             -0.403
```

```
confint(fit.3, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 25.28  27.49
```

```
confint(fit.3, "access_to_highway", level = 0.95)
```

```
##                       2.5 %  97.5 %
## access_to_highway -0.4885 -0.3177
```

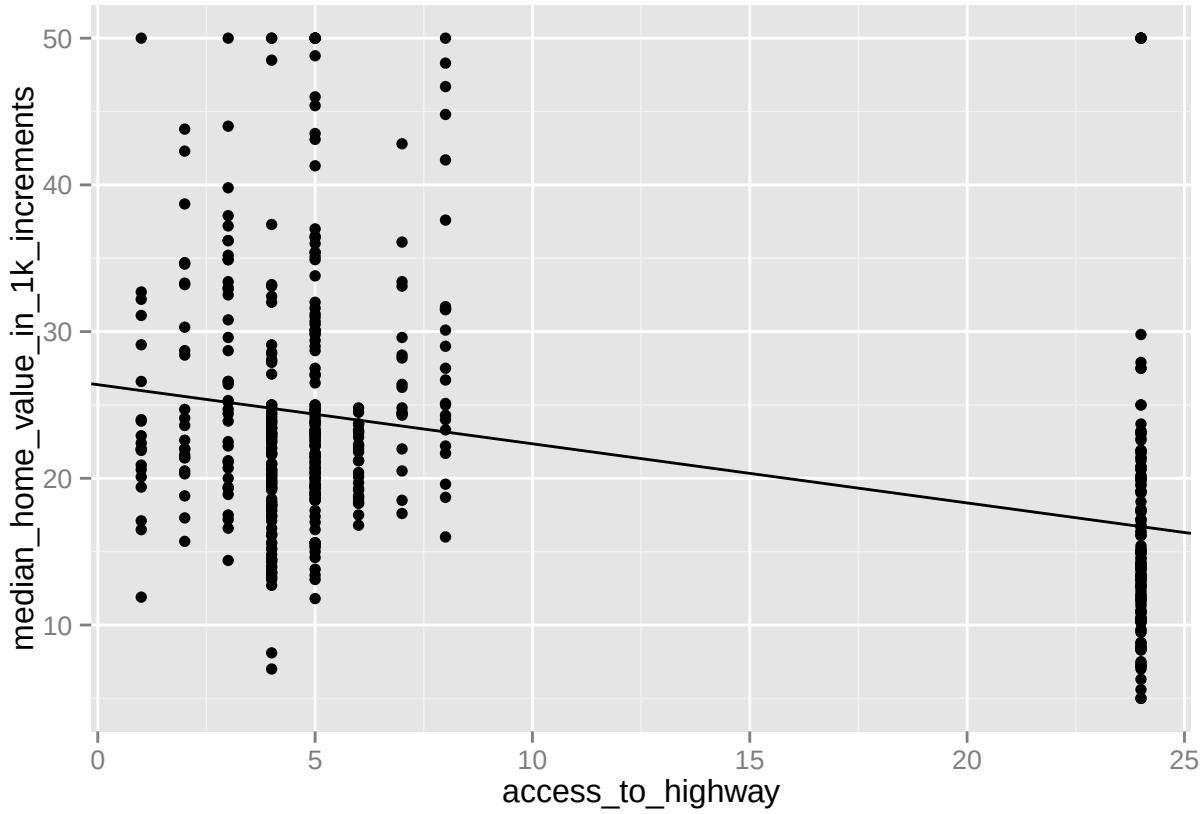
```
print(paste0("Sigma: ", summary(fit.3)$sigma))
```

```
## [1] "Sigma: 8.50946657856462"
```

```
print(paste0("Adj R-squared: ", summary(fit.3)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.143943418364532"
```

```
ggplot(data = data, aes(x = access_to_highway, y = median_home_value_in_1k_increments)) + geom_point() +
```




Finally lets look at Proportion of homes that were built prior to 1940

```
fit.4 = lm(median_home_value_in_1k_increments ~ age_of_dwelling_since_1940)
print(fit.4)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ age_of_dwelling_since_1940)
##
## Coefficients:
##                 (Intercept)  age_of_dwelling_since_1940
##                      30.979                      -0.123
```

```
confint(fit.4, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 29.02  32.94
```

```
confint(fit.4, "age_of_dwelling_since_1940", level = 0.95)
```

```
##                               2.5 %   97.5 %
## age_of_dwelling_since_1940 -0.1496 -0.09668
```

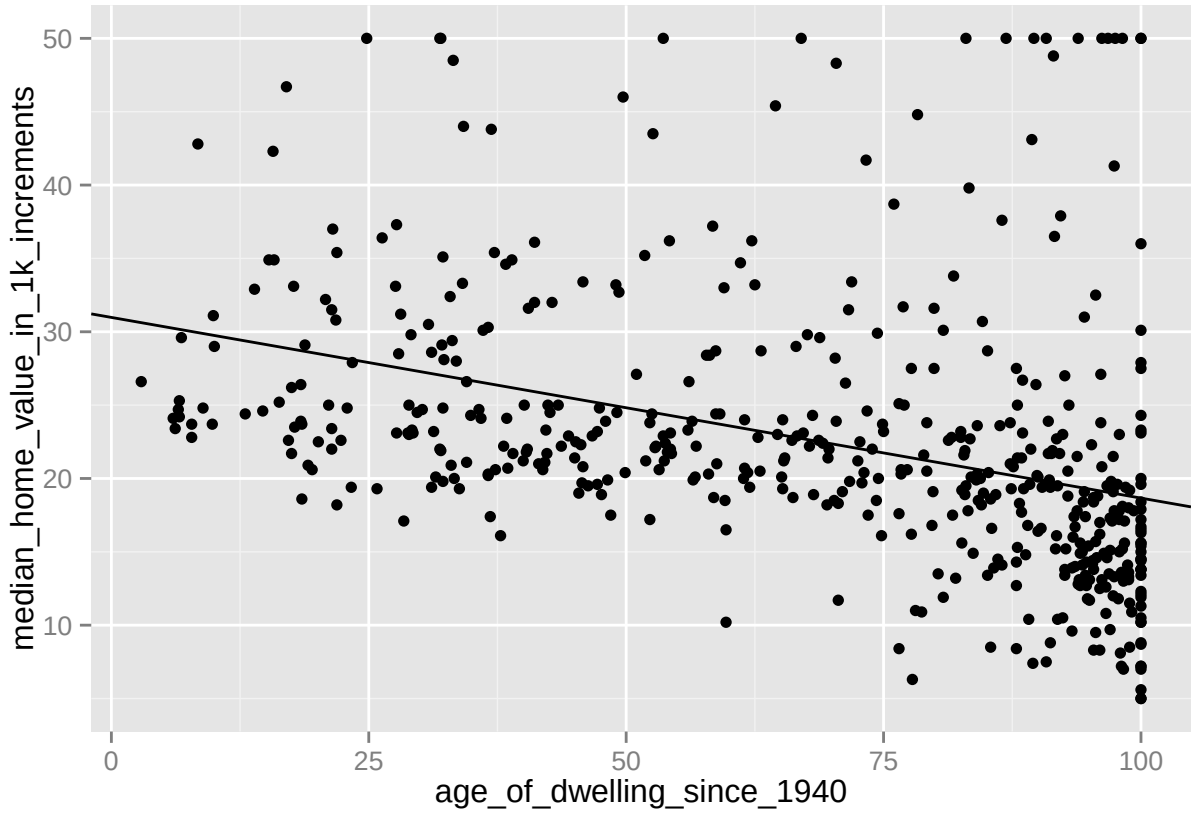
```
print(paste0("Sigma: ", summary(fit.4)$sigma))
```

```
## [1] "Sigma: 8.5270966734705"
```

```
print(paste0("Adj R-squared: ", summary(fit.4)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.140392551109705"
```

```
ggplot(data = data, aes(x = age_of_dwelling_since_1940, y = median_home_value_in_1k_increments)) + geom_
```




Step 2: Meaning of Regression Results Using Single Variable

Based on these 4 variables, it looks like the "percent_lower_status_of_population" is the closest single variable that explains the variation in the median priceof homes in Boston suburbs.

This tells us that for each unit change increase in the "percent_lower_status_of_population" variable, there is a average 6.21 point change decrease in the median price of a home in a Boston suburb.

The adjusted r-squared was 54% for this single variable.

Step 3: Adding variables and Data Transformation

I continue to expect to see coefficients with negative signs. Lets see what happens when we add the pupil teacher ratio and the crime rate

```
fit.1x = lm(median_home_value_in_1k_increments ~ pupil_teacher_ratio + crime_rate)
print(fit.1x)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ pupil_teacher_ratio +
##     crime_rate)
##
## Coefficients:
##         (Intercept)  pupil_teacher_ratio           crime_rate
##              57.378               -1.833               -0.281
```

```
confint(fit.1x, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 51.51  63.25
```

```
confint(fit.1x, "pupil_teacher_ratio", level = 0.95)
```

```
##                      2.5 % 97.5 %
## pupil_teacher_ratio -2.153 -1.513
```

```
confint(fit.1x, "crime_rate", level = 0.95)
```

```
##              2.5 %  97.5 %
## crime_rate -0.3621 -0.2008
```

```
print(paste0("Sigma: ", summary(fit.1x)$sigma))
```

```
## [1] "Sigma: 7.59193145446213"
```

```
print(paste0("Adj R-squared: ", summary(fit.1x)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.318599648263564"
```

Lets see what happens when we add the pupil teacher ratio and the Percent of Lower Status Population. We see that our regression improves up to 60.5%!

```
fit.2x = lm(median_home_value_in_1k_increments ~ pupil_teacher_ratio + percent_lower_status_of_populati
print(fit.2x)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ pupil_teacher_ratio +
##     percent_lower_status_of_population)
##
## Coefficients:
##                          (Intercept)                 pupil_teacher_ratio
##                                54.05                               -1.15
## percent_lower_status_of_population
##                                -0.82
```

```
confint(fit.2x, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 49.64  58.45
```

```
confint(fit.2x, "pupil_teacher_ratio", level = 0.95)
```

```
##                       2.5 %  97.5 %
## pupil_teacher_ratio -1.397 -0.8936
```

```
confint(fit.2x, "percent_lower_status_of_population", level = 0.95)
```

```
##                                     2.5 %  97.5 %
## percent_lower_status_of_population -0.8965 -0.7439
```

```
print(paste0("Sigma: ", summary(fit.2x)$sigma))
```

```
## [1] "Sigma: 5.77962721611035"
```

```
print(paste0("Adj R-squared: ", summary(fit.2x)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.6050906177960 29"
```

Lets see what happens when we add the pupil teacher ratio and Access To Highway

```
fit.3x = lm(median_home_value_in_1k_increments ~ pupil_teacher_ratio + access_to_highway)
print(fit.3x)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ pupil_teacher_ratio +
##     access_to_highway)
##
## Coefficients:
##         (Intercept)  pupil_teacher_ratio    access_to_highway
##              57.450               -1.790               -0.196
```

```r
confint(fit.3x, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept)  51.2   63.7
```

```r
confint(fit.3x, "pupil_teacher_ratio", level = 0.95)
```

```
##                      2.5 % 97.5 %
## pupil_teacher_ratio -2.146 -1.435
```

```r
confint(fit.3x, "access_to_highway", level = 0.95)
```

```
##                     2.5 %  97.5 %
## access_to_highway -0.2846 -0.1078
```

```r
print(paste0("Sigma: ", summary(fit.3x)$sigma))
```

```
## [1] "Sigma: 7.79284513078396"
```

```r
print(paste0("Adj R-squared: ", summary(fit.3x)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.28205712844728"
```

Lets see what happens when we add the pupil teacher ratio and Age of Dwelling

```r
fit.4x = lm(median_home_value_in_1k_increments ~ pupil_teacher_ratio + age_of_dwelling_since_1940)
print(fit.4x)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ pupil_teacher_ratio +
##     age_of_dwelling_since_1940)
##
## Coefficients:
##                (Intercept)         pupil_teacher_ratio
##                    62.8431                     -1.8660
## age_of_dwelling_since_1940
##                    -0.0856
```

```r
confint(fit.4x, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 57.15  68.54
```

```r
confint(fit.4x, "pupil_teacher_ratio", level = 0.95)
```

```
##                      2.5 % 97.5 %
## pupil_teacher_ratio -2.184 -1.548
```

```r
confint(fit.4x, "age_of_dwelling_since_1940", level = 0.95)
```

```
##                               2.5 %   97.5 %
## age_of_dwelling_since_1940 -0.1101 -0.06121
```

```r
print(paste0("Sigma: ", summary(fit.4x)$sigma))
```

```
## [1] "Sigma: 7.58891147938939"
```

```r
print(paste0("Adj R-squared: ", summary(fit.4x)$adj.r.squared))
```

```
## [1] "Adj R-squared: 0.31914164546236"
```

Step 4:

Now we will add a transformation and a interaction to see if we can beat 60.5%.
By using num_of_rooms along with percent_lower_status_of_population, pupil_teacher_ratio and distance_to_employment_centers:num_of_rooms_scaled we reach a Adjusted R squared of 68.7%!

```r
num_of_rooms_scaled = (num_of_rooms - mean(num_of_rooms))/(2 * sd(num_of_rooms))
fit.11x = lm(median_home_value_in_1k_increments ~ percent_lower_status_of_population + pupil_teacher_ra
print(fit.11x)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ percent_lower_status_of_population +
##     pupil_teacher_ratio + distance_to_employment_centers:num_of_rooms_scaled,
##     data = data)
##
## Coefficients:
##                                         (Intercept)
##                                              45.793
##                 percent_lower_status_of_population
##                                              -0.616
##                                pupil_teacher_ratio
##                                              -0.858
## distance_to_employment_centers:num_of_rooms_scaled
##                                               1.719
```

```r
confint(fit.11x, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 41.63  49.95
```

```r
confint(fit.11x, "percent_lower_status_of_population", level = 0.95)
```

```
##                                      2.5 %  97.5 %
## percent_lower_status_of_population -0.6925 -0.5401
```

```r
confint(fit.11x, "pupil_teacher_ratio", level = 0.95)
```

```
##                       2.5 %  97.5 %
## pupil_teacher_ratio -1.087 -0.6288
```

```r
confint(fit.11x, "distance_to_employment_centers:num_of_rooms_scaled", level = 0.95)
```

```
##                                                  2.5 % 97.5 %
## distance_to_employment_centers:num_of_rooms_scaled 1.427   2.01
```

```r
print(paste0("Sigma: ", summary(fit.11x)$sigma))
```

```
## [1] "Sigma: 5.14002255017721"
```

```r
print(paste0("(Adj) R-squared: ", summary(fit.11x)$adj.r.squared))
```

```
## [1] "(Adj) R-squared: 0.687659827763166"
```

Step 5: The variables with the largest coefficients are:
"pupil_teacher_ratio" "nitric_oxide_conc" "num_of_rooms" "industrial"

First we will scale these variables and see what our graph looks like:

```r
pupil_teacher_ratio_scaled = (pupil_teacher_ratio - mean(pupil_teacher_ratio))/(2 * sd(pupil_teacher_ra
nitric_oxide_conc_scaled = (nitric_oxide_conc - mean(nitric_oxide_conc))/(2 * sd(nitric_oxide_conc))
num_of_rooms_scaled = (num_of_rooms - mean(num_of_rooms))/(2 * sd(num_of_rooms))

fit.b = lm(median_home_value_in_1k_increments ~ pupil_teacher_ratio_scaled + nitric_oxide_conc_scaled +

print(fit.b)
```

```
##
## Call:
## lm(formula = median_home_value_in_1k_increments ~ pupil_teacher_ratio_scaled +
##     nitric_oxide_conc_scaled + num_of_rooms_scaled)
##
## Coefficients:
##                (Intercept)  pupil_teacher_ratio_scaled
##                      22.53                       -5.12
##   nitric_oxide_conc_scaled         num_of_rooms_scaled
##                      -3.94                        9.78
```

```r
confint(fit.b, "(Intercept)", level = 0.95)
```

```
##             2.5 % 97.5 %
## (Intercept) 22.02  23.04
```

```
confint(fit.b, "pupil_teacher_ratio_scaled", level = 0.95)
```

```
##                              2.5 % 97.5 %
## pupil_teacher_ratio_scaled -6.212 -4.027
```

```
confint(fit.b, "nitric_oxide_conc_scaled", level = 0.95)
```

```
##                            2.5 % 97.5 %
## nitric_oxide_conc_scaled -5.009 -2.866
```

```
confint(fit.b, "num_of_rooms_scaled", level = 0.95)
```

```
##                      2.5 % 97.5 %
## num_of_rooms_scaled 8.655  10.91
```

```
print(paste0("Sigma: ", summary(fit.b)$sigma))
```

```
## [1] "Sigma: 5.81555939459636"
```

```
print(paste0("(Adj) R-squared: ", summary(fit.b)$adj.r.squared))
```

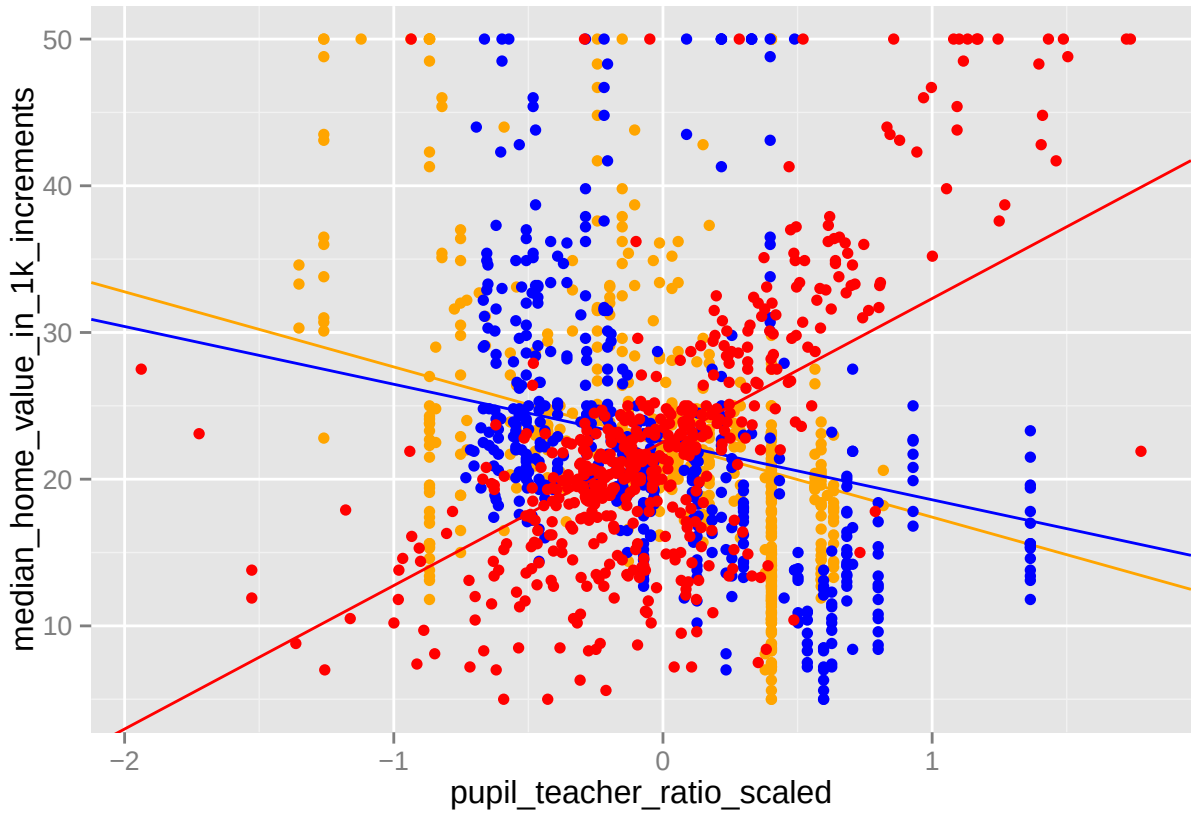
```
## [1] "(Adj) R-squared: 0.600165018397314"
```

We see three different colors with regression lines to match representing the relationships between the median home value and: pupil teacher ratio, nitric oxide concentration, and the number of rooms. The only predictor with a positive correlation is the number of rooms.

```
ggplot() +
  geom_point(data = data, aes(x = pupil_teacher_ratio_scaled, y = median_home_value_in_1k_increments), 

 #ggplot() +
  geom_point(data = data, aes(x = nitric_oxide_conc_scaled, y = median_home_value_in_1k_increments), co

#ggplot() +
geom_point(data = data, aes(x = num_of_rooms_scaled, y = median_home_value_in_1k_increments), colour =
```


median_home_value_in_1k_increments
50
40
30
20
10
−2
−1
0
1
pupil_teacher_ratio_scaled