Simple Linear Regression: Boston Housing Market

Sean O'Malley

3/25/2017

ISLR: Process Explained

Simple Linear Regression is a very straightforward approach for predicting a quantitative response Y on the basis of single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. It is apporoximately modeled as regressing X on Y.

$$Y = B0 + B1x$$

B0 and *B1* are two unknown constants that represent the intercept and slope terms in the linear model. Together, *B0* and *B1* are known as the model coefficients or parameters. Once we have used our training data to produce estimates for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing.

EG: X may represent TV advertising and Y may represent Sales. Then we can regress sales onto TV by fitting the model

How do you define a null hypothesis and an alternative hypothesis?

In statistics and experimental design the null hypothesis is the initial statistical claim that the population mean is equivalent to the claimed. Conversely, the alternative hypothesis is only accepted when the null hypothesis is rejected due to statistical significance. We most often define statistical significance with a test that surpasses a pre-defined p-value (calculated probability) threshold; often with the standard of 0.05.

What are good hypotheses? Can you give at least one example of a good hypothesis?

A good hypothesis includes the following things: * Before you make a hypothesis, you have to clearly identify the question * Ensure your hypothesis is an educated, testable prediction about what will happen * Write in a clear, simple language * Define your hypothesis with easy-to-measure terms, like who the participants are, what changes during testing, and the effect of the changes * Ensure hypothesis is testable * Research similar projects that have existed * Ensure hypothesis is a specific statement relating to a single experiment

An example of a hypothesis we've used recently in marketing data science in regard to A/B testing CTA size of creative banners for digital advertising: * By enlarging the CTA it will be more clear where the CTA is and users will be more likely to convert

Hypothesis / Overview: **Boston Housing Data**

At the time of the publication of this Boston housing market data, the data scientist involved wanted to see if the poor air quality had a significant effect on the housing prices in the area. The data is a series of attributes that those involved could help provide for insight on the Boston Housing market effects, specifically that of pollution.

Specifically considering the proximity to the Charles River, distance to the main employment centers, pupil-teacher ratio in schools, and levels of crime. In regard to pollution, we will look towards the nitric oxcide levels of the area.

I'd prefer to attack this some what systematically, first looking at the data and exploring anything that I personally feel stands out in the data. After gaining this more scattered knowledge I will explore the hypothesis posed by the scientist.

- **H0** : Air pollution does not have a significant effect on housing prices in the Boston Area
- **H1** : Air pollution significantly lowers housing prices in the Boston area

```
glimpse(housing)
```

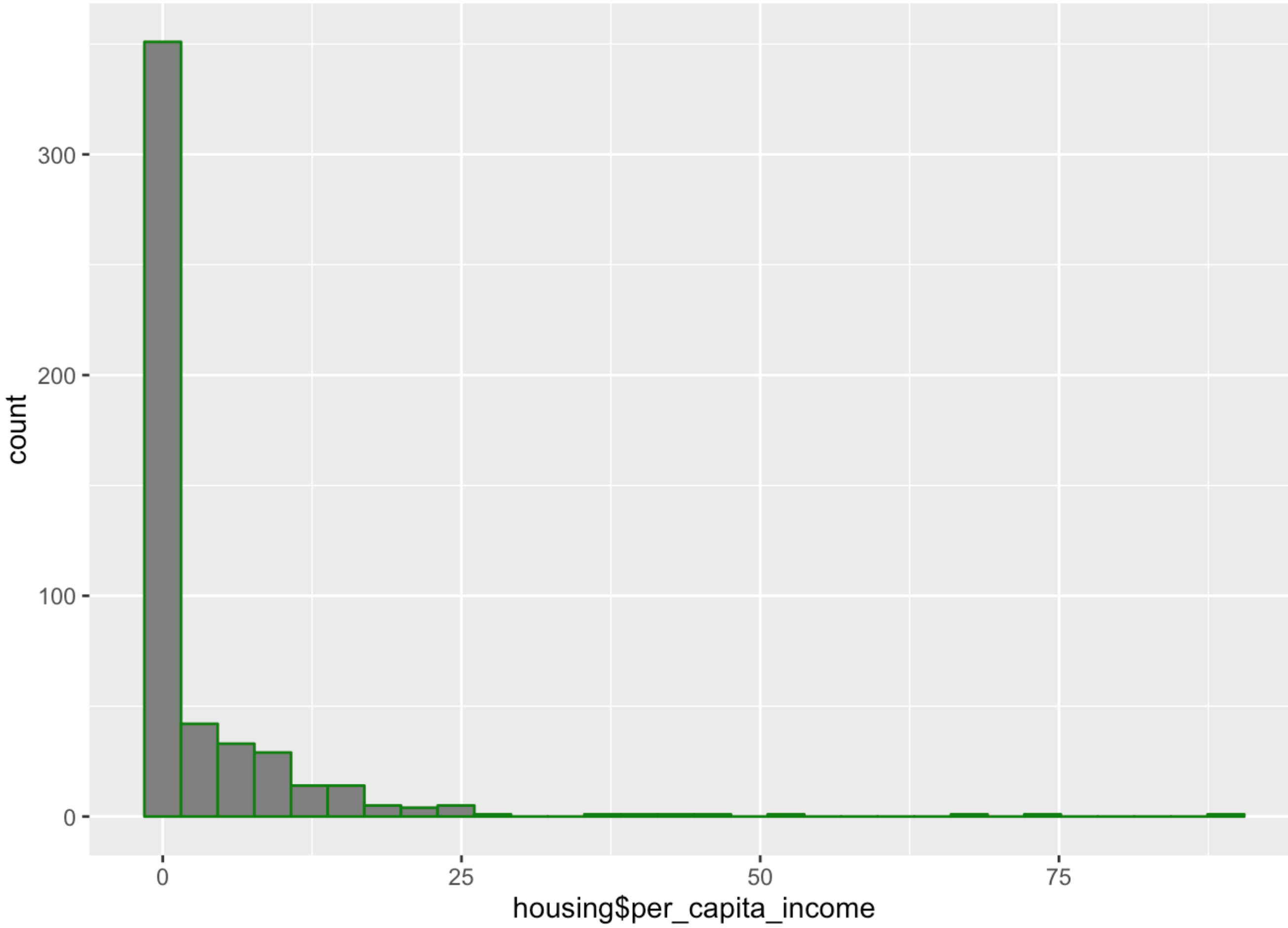
```
## Observations: 506
## Variables: 14
## $ per_capita_income     <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06...
## $ land_zoned_proportion <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12....
## $ non_retail_proportion <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87...
## $ charles_riv_dummy     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ nitric_oxcide_conc    <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458...
## $ rm_per_house          <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430...
## $ old_house             <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6...
## $ dis_to_work           <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, ...
## $ hwy_access            <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4...
## $ tax                   <dbl> 296, 242, 242, 222, 222, 222, 311, 311, ...
## $ pt_ratio              <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2...
## $ blacks_per_town       <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, ...
## $ l_status              <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.4...
## $ med_val_1k            <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9...
```

```
summary(housing)
```

```
##  per_capita_income  land_zoned_proportion non_retail_proportion
##  Min.   : 0.00632   Min.   :  0.00        Min.   : 0.46
##  1st Qu.: 0.08204   1st Qu.:  0.00        1st Qu.: 5.19
##  Median : 0.25651   Median :  0.00        Median : 9.69
##  Mean   : 3.61352   Mean   : 11.36        Mean   :11.14
##  3rd Qu.: 3.67708   3rd Qu.: 12.50        3rd Qu.:18.10
##  Max.   :88.97620   Max.   :100.00        Max.   :27.74
##  charles_riv_dummy nitric_oxcide_conc  rm_per_house     old_house
##  Min.   :0.00000   Min.   :0.3850     Min.   :3.561   Min.   :  2.90
##  1st Qu.:0.00000   1st Qu.:0.4490     1st Qu.:5.886   1st Qu.: 45.02
##  Median :0.00000   Median :0.5380     Median :6.208   Median : 77.50
##  Mean   :0.06917   Mean   :0.5547     Mean   :6.285   Mean   : 68.57
##  3rd Qu.:0.00000   3rd Qu.:0.6240     3rd Qu.:6.623   3rd Qu.: 94.08
##  Max.   :1.00000   Max.   :0.8710     Max.   :8.780   Max.   :100.00
##   dis_to_work       hwy_access          tax           pt_ratio
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
##  blacks_per_town     l_status       med_val_1k
##  Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
##  1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
##  Median :391.44   Median :11.36   Median :21.20
##  Mean   :356.67   Mean   :12.65   Mean   :22.53
##  3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :396.90   Max.   :37.97   Max.   :50.00
```

```
ggplot(data=housing, aes(housing$per_capita_income)) + geom_histogram(fill = "#808080",
col = "#008000")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

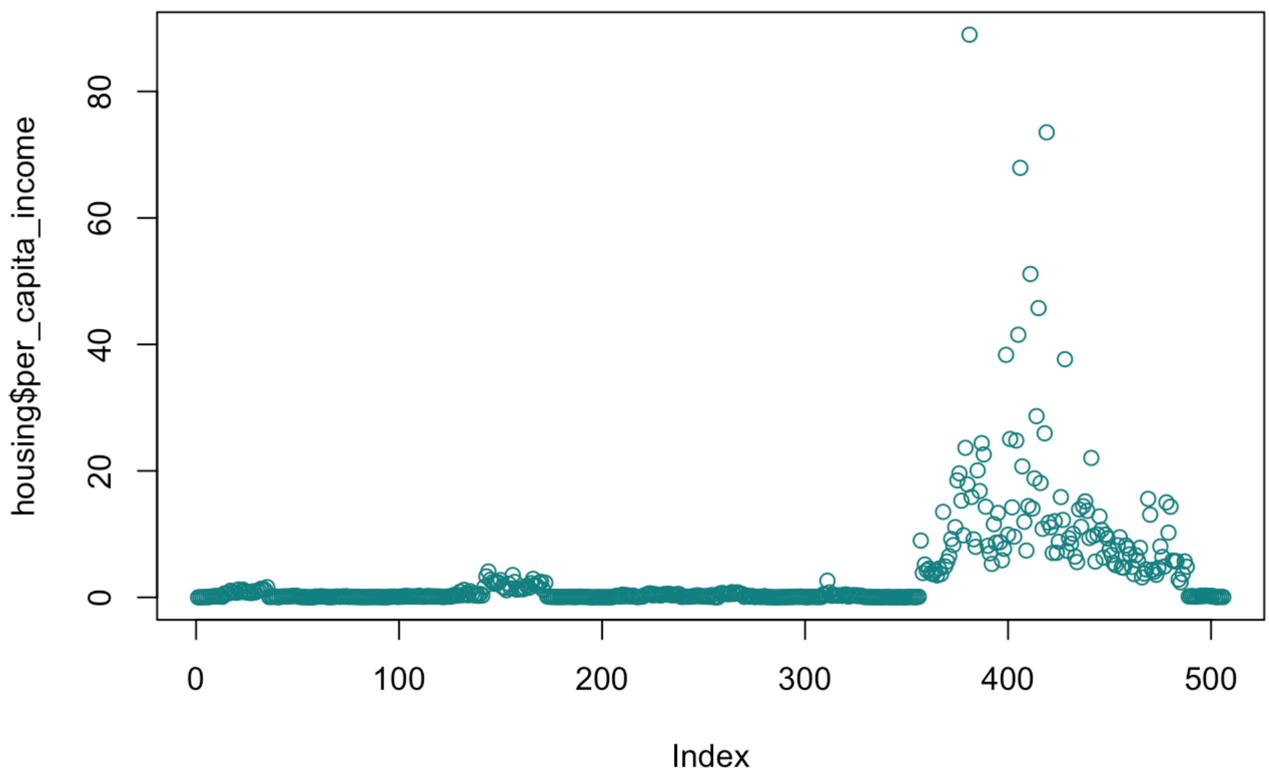


That histogram of per capita income was certainly not normal and interested me, let me explore it further…

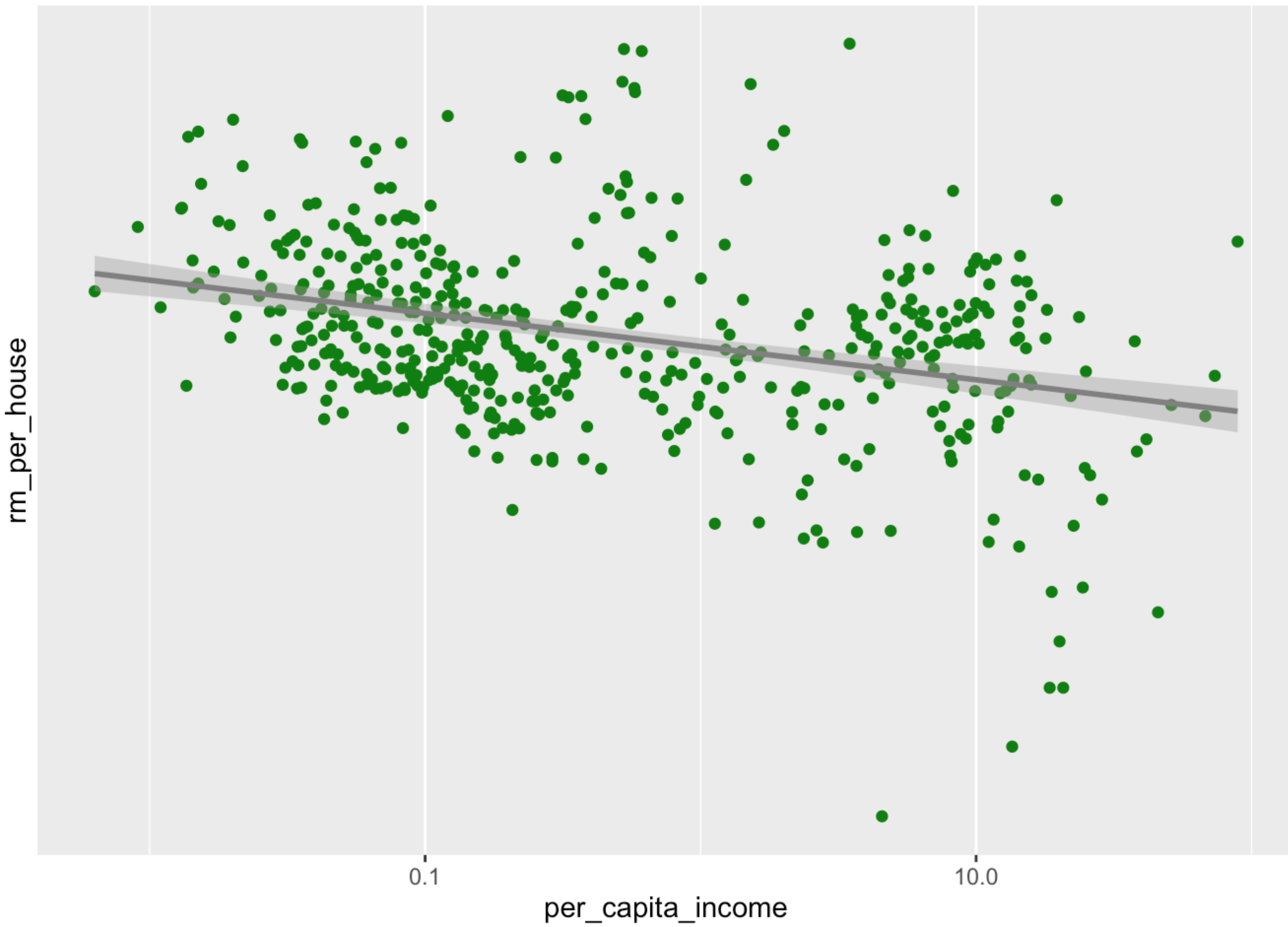
```
summary(housing$per_capita_income)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25650  3.61400  3.67700 88.98000
```

```
plot(housing$per_capita_income, col = "#008080")
```



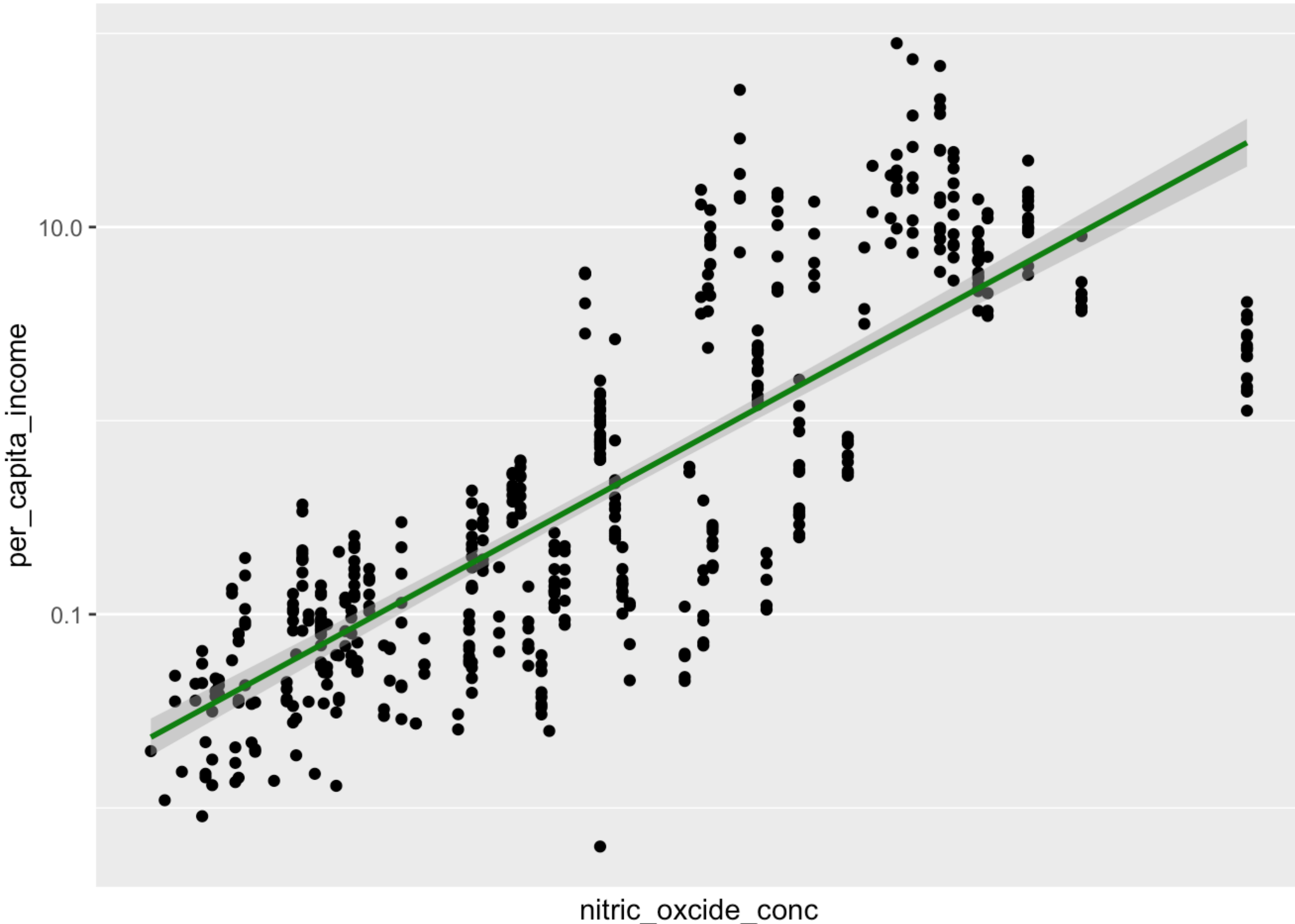

```
ggplot(data=housing, aes( x= per_capita_income, y = rm_per_house)) +
      geom_point(color = "#008000") +
      geom_smooth(method = "lm", col = "#808080") +
      scale_x_log10() +
      scale_y_log10()
```



After looking at many of the factors that (all of which I could include in a Multiple Linear Regression), I've decided that nitric oxide concentration is the most prevalent and interesting factor for a Simple Linear Regression. Lets visualize.

Visualizing the effect of nitric oxide as the independent variable on the dependent variable, per capita income; we see a strong positive, linear, fairly concentrated relationship.

```
ggplot(data=housing, aes(y = per_capita_income, x = nitric_oxcide_conc)) +
      geom_point() +
      geom_smooth(method = "lm", color = "#008000") +
      scale_x_log10() +
      scale_y_log10()
```



Now, lets see mathematically the relationship of nitric oxcide concentration regressed on per capita income

Correlation:

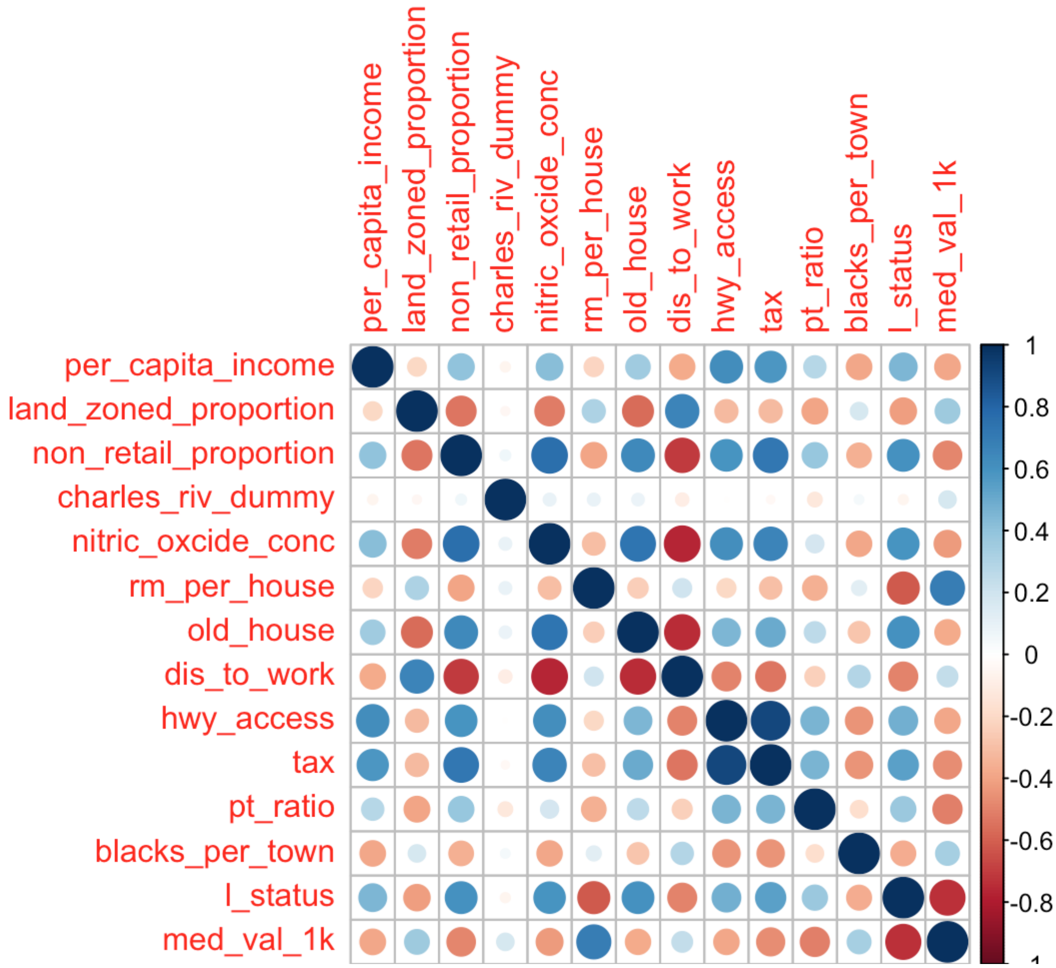
```
cor(housing$per_capita_income, housing$nitric_oxcide_conc)
```

```
## [1] 0.4209717
```

Now, the primary problem I am interested is the nitric oxcide concentration, but that correlation is pretty low, let me build a correlation matrix of all the variables at play and see if there is a stronger choice in simple linear regression to predict per capita income.

```
housing_cor <- cor(housing)

corrplot(housing_cor, method = "circle")
```



Hmm, interesting. We see the strongest relationships are with nitric oxcide and highway access. The other strongly correlated variables to per capita income is tax bracked and whether they are lower status or not. We cannot use the latter two because those figures are more than likely based on income to begin with, thus biased. It is also interesting to note the tax bracket correlation with highway access.

After looking at this correlation matrix I'd still like to proceed with using nitric oxcide to predict per capita income in a simple linear regression, simply out of the fact that it is a much more interesting / non-obvious problem to solve.

Build Model:

```
model <- lm(formula = per_capita_income ~ nitric_oxcide_conc, data = housing)
```

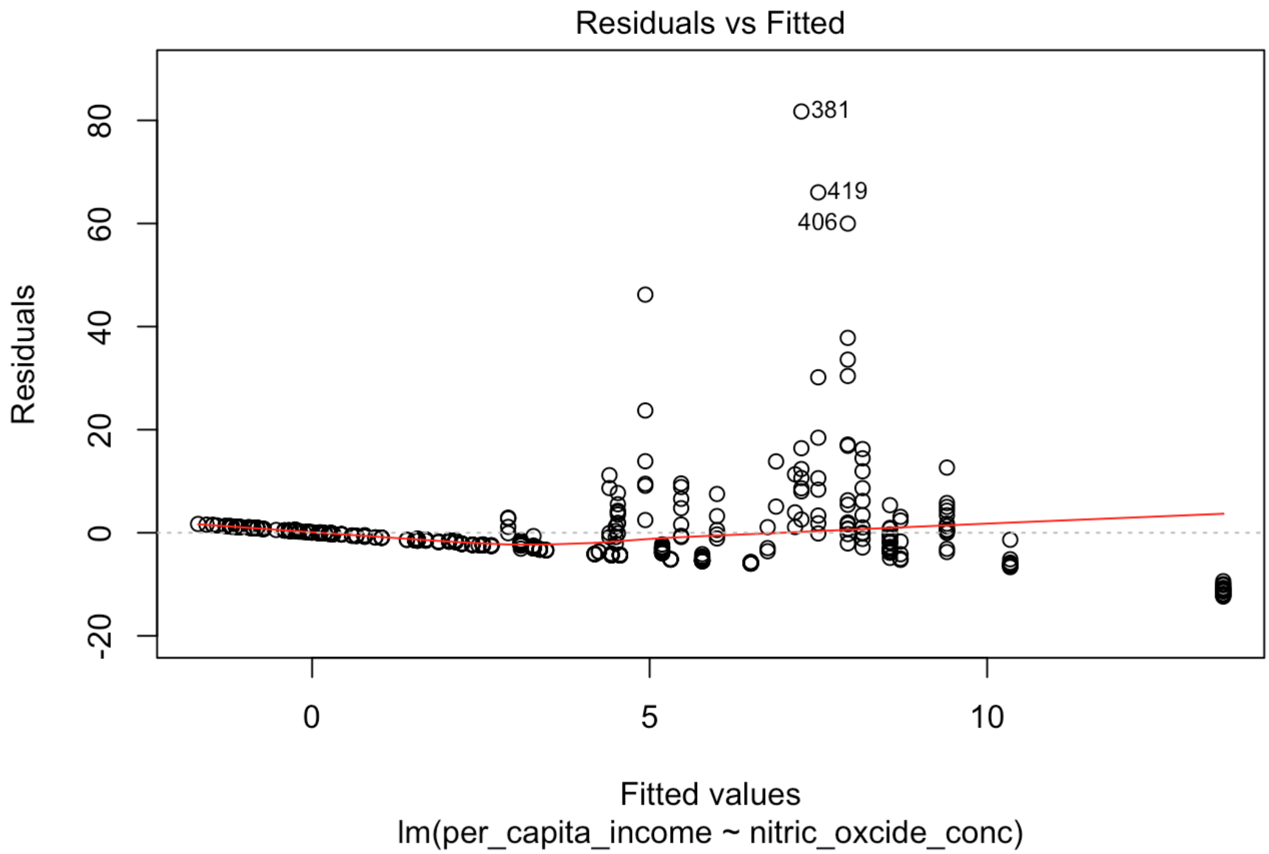
Evaluate:

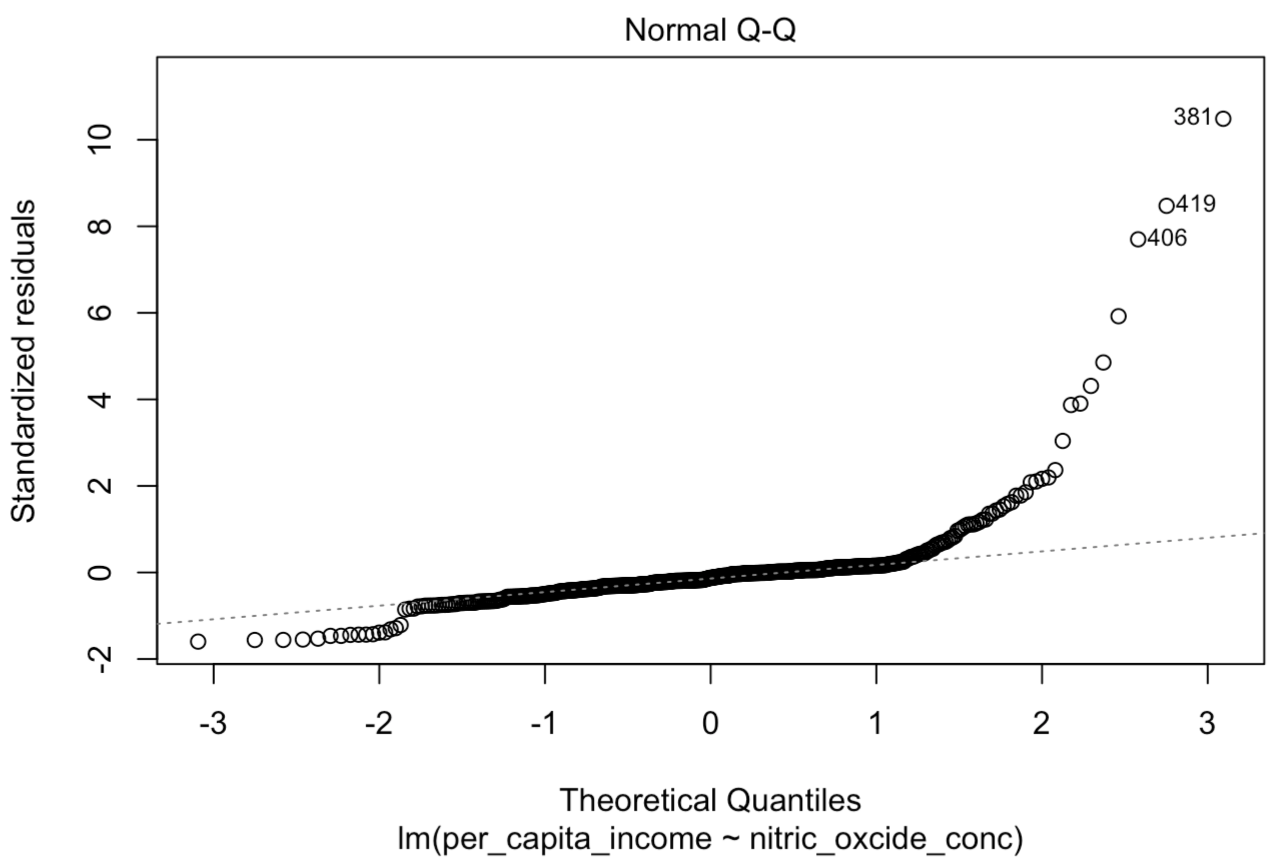
```
summary(model)
```

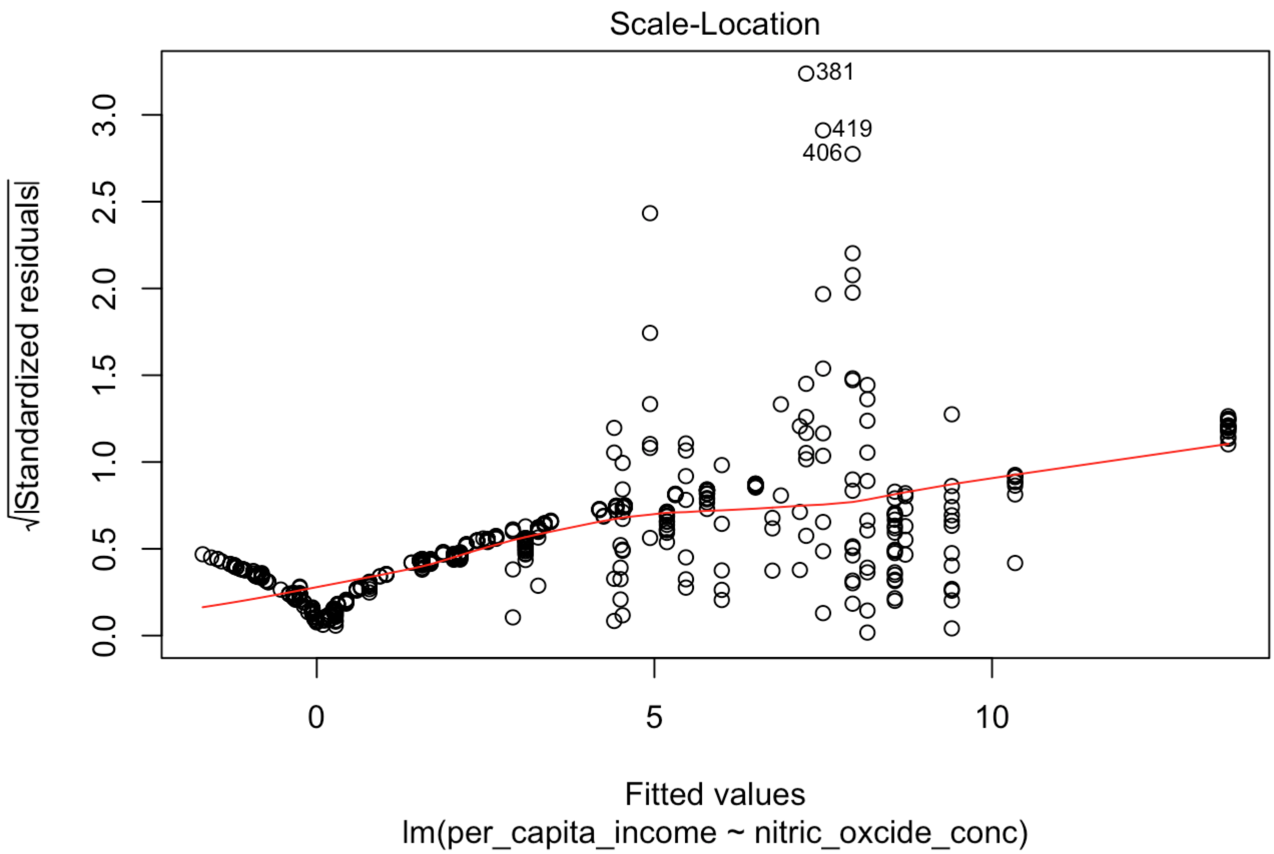
```
##
## Call:
## lm(formula = per_capita_income ~ nitric_oxcide_conc, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -13.720      1.699  -8.073 5.08e-15 ***
## nitric_oxcide_conc   31.249      2.999  10.419  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

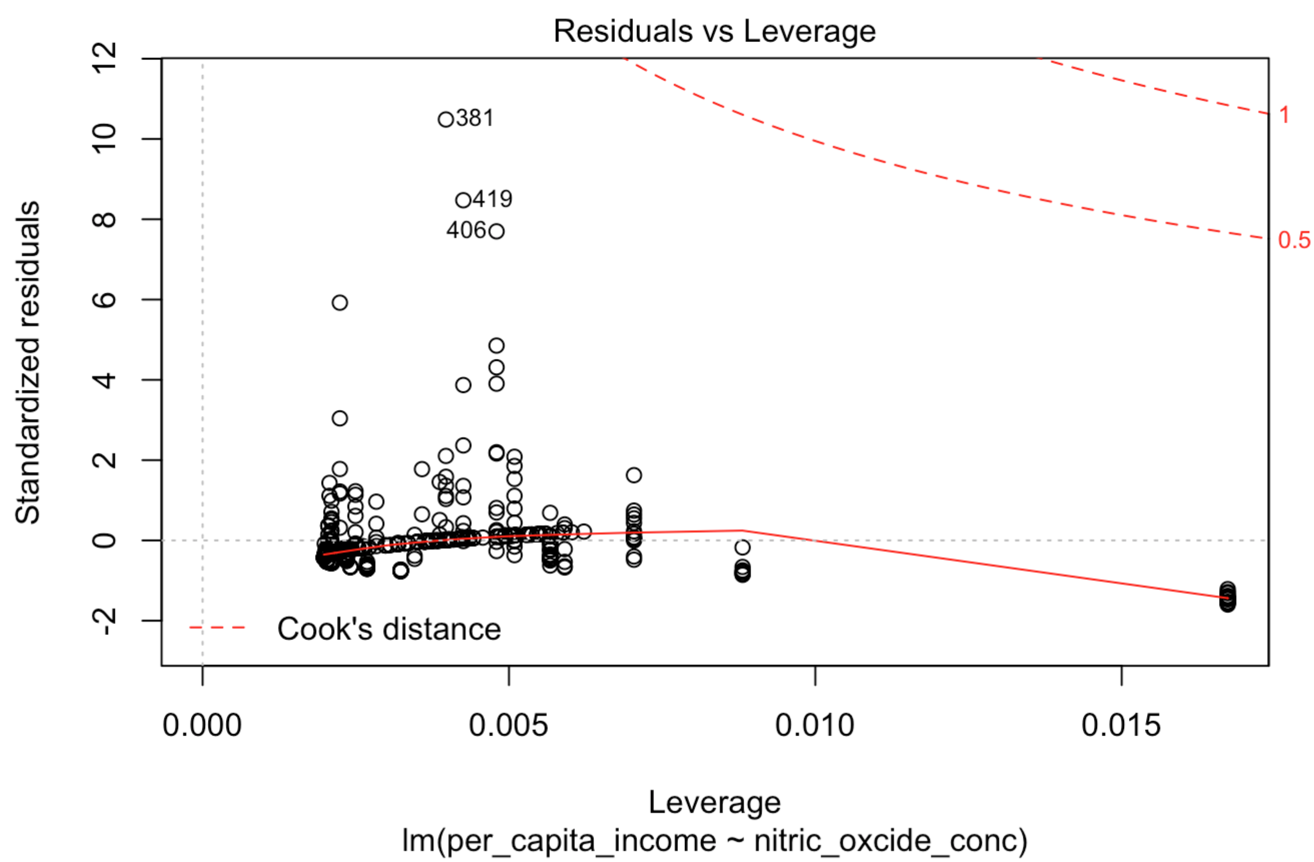
Plot:

```
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(per_capita_income ~ nitric_oxcide_conc)
381
419
406
80
60
40
20
0
-20
0
5
10


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(per_capita_income ~ nitric_oxcide_conc)
381
419
406
10
8
6
4
2
0
-2
-3
-2
-1
0
1
2
3





View Predictions and Residuals:

```
augment(model)
```

```
##     per_capita_income nitric_oxcide_conc      .fitted   .se.fit
## 1             0.00632             0.5380  3.091827476 0.3507876
## 2             0.02731             0.4690  0.935678823 0.4319744
## 3             0.02729             0.4690  0.935678823 0.4319744
## 4             0.03237             0.4580  0.591944980 0.4523813
## 5             0.06905             0.4580  0.591944980 0.4523813
## 6             0.02985             0.4580  0.591944980 0.4523813
## 7             0.08829             0.5240  2.654348039 0.3591934
## 8             0.14455             0.5240  2.654348039 0.3591934
## 9             0.21124             0.5240  2.654348039 0.3591934
## 10            0.17004             0.5240  2.654348039 0.3591934
## 11            0.22489             0.5240  2.654348039 0.3591934
##           .resid        .hat   .sigma      .cooksd    .std.resid
## 1    -3.085507476 0.002017389 7.816519 1.580767e-04 -0.3954718878
## 2    -0.908368823 0.003059265 7.817627 2.081967e-05 -0.1164871582
## 3    -0.908388823 0.003059265 7.817627 2.082059e-05 -0.1164897230
## 4    -0.559574980 0.003355137 7.817692 8.669956e-06 -0.0717692772
## 5    -0.522894980 0.003355137 7.817697 7.570582e-06 -0.0670648191
## 6    -0.562094980 0.003355137 7.817692 8.748221e-06 -0.0720924843
## 7    -2.566058039 0.002115231 7.816893 1.146571e-04 -0.3289097739
## 8    -2.509798039 0.002115231 7.816929 1.096846e-04 -0.3216985325
## 9    -2.443108039 0.002115231 7.816971 1.039330e-04 -0.3131504044
## 10   -2.484308039 0.002115231 7.816945 1.074679e-04 -0.3184312993
## 11   -2.429458039 0.002115231 7.816980 1.027748e-04 -0.3114007875
##  [ reached getOption("max.print") -- omitted 495 rows ]
```

Conclusion:

As we look down through the model I am mildly disappointed. Looking to the R^2 we see that the model can only explain around 17 percent of variance in per capita income. This is poor, even for a simple linear regression model. As we look at the p-value we see that incredibly low, confirming that with this simple linear regression model we fail to reject the null hypothesis that nitric oxcide does not have a significant effect on per capita income in the Boston area. To round out our poor prediction ability we see a correlation coefficient in the 0.001 range. Graphically we see that the data is far sparse to gain significant insight via a simple linear model.

I have personally concluded that this problem would best be solved with Multiple linear regression, being that housing, income and opportunity are not narrowed down to one factor, but are often the effect of many socioeconomic, physical and political reasons.