Module 3 CT Option 1

Ty Cook

4/4/2020

## R Markdown

# Read into Data Frame

housing.df <- read.csv("BostonHousing.csv", header = TRUE)

# Summary

summary(housing.df)

## CRIM ZN INDUS CHAS   
## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000   
## 1st Qu.: 0.08204 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000   
## Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000   
## Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917   
## 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000   
## Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000   
## NOX RM AGE DIS   
## Min. :0.3850 Min. :3.561 Min. : 2.90 Min. : 1.130   
## 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100   
## Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207   
## Mean :0.5547 Mean :6.285 Mean : 68.57 Mean : 3.795   
## 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188   
## Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127   
## RAD TAX PTRATIO LSTAT   
## Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 1.73   
## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.: 6.95   
## Median : 5.000 Median :330.0 Median :19.05 Median :11.36   
## Mean : 9.549 Mean :408.2 Mean :18.46 Mean :12.65   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:16.95   
## Max. :24.000 Max. :711.0 Max. :22.00 Max. :37.97   
## MEDV CAT..MEDV   
## Min. : 5.00 Min. :0.000   
## 1st Qu.:17.02 1st Qu.:0.000   
## Median :21.20 Median :0.000   
## Mean :22.53 Mean :0.166   
## 3rd Qu.:25.00 3rd Qu.:0.000   
## Max. :50.00 Max. :1.000

# Data Exploration Functions

data.frame(mean=sapply(housing.df[-c(14)],mean,na.rm=TRUE),  
sd=sapply(housing.df[-c(14)],sd,na.rm=TRUE),   
min=sapply(housing.df[-c(14)],min,na.rm=TRUE),  
max=sapply(housing.df[-c(14)],max,na.rm=TRUE),  
median=sapply(housing.df[-c(14)],median,na.rm=TRUE),  
length=sapply(housing.df[-c(14)],length)  
)

## mean sd min max median length  
## CRIM 3.61352356 8.6015451 0.00632 88.9762 0.25651 506  
## ZN 11.36363636 23.3224530 0.00000 100.0000 0.00000 506  
## INDUS 11.13677866 6.8603529 0.46000 27.7400 9.69000 506  
## CHAS 0.06916996 0.2539940 0.00000 1.0000 0.00000 506  
## NOX 0.55469506 0.1158777 0.38500 0.8710 0.53800 506  
## RM 6.28463439 0.7026171 3.56100 8.7800 6.20850 506  
## AGE 68.57490119 28.1488614 2.90000 100.0000 77.50000 506  
## DIS 3.79504269 2.1057101 1.12960 12.1265 3.20745 506  
## RAD 9.54940711 8.7072594 1.00000 24.0000 5.00000 506  
## TAX 408.23715415 168.5371161 187.00000 711.0000 330.00000 506  
## PTRATIO 18.45553360 2.1649455 12.60000 22.0000 19.05000 506  
## LSTAT 12.65306324 7.1410615 1.73000 37.9700 11.36000 506  
## MEDV 22.53280632 9.1971041 5.00000 50.0000 21.20000 506

# Select Variables

selected.var <- c(1:13)

# Partition Data

set.seed(1)   
train.index <- sample(c(1:1000), 600)  
train.df <- housing.df[train.index, selected.var]  
valid.df <- housing.df[train.index, selected.var]

# Lm Function

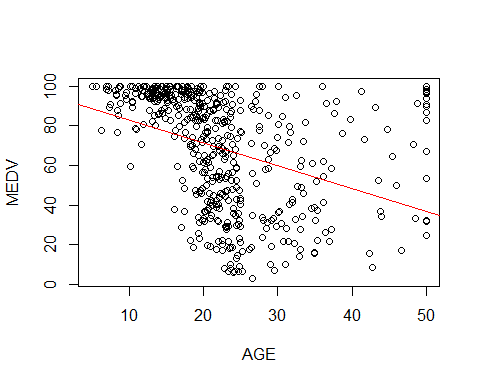
housing.lm <-lm(MEDV ~ ., data = train.df)

# Options

options(scipen = 999)  
summary(housing.lm)

##   
## Call:  
## lm(formula = MEDV ~ ., data = train.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.2678 -2.5763 -0.4978 1.6110 27.7560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.180971 6.090107 6.105 0.0000000033008735 \*\*\*  
## CRIM -0.133516 0.036739 -3.634 0.00033 \*\*\*  
## ZN 0.037370 0.016321 2.290 0.02276 \*   
## INDUS 0.039124 0.071204 0.549 0.58311   
## CHAS 2.426175 1.128640 2.150 0.03241 \*   
## NOX -19.065745 5.003594 -3.810 0.00017 \*\*\*  
## RM 4.349536 0.530736 8.195 0.0000000000000082 \*\*\*  
## AGE -0.013462 0.015790 -0.853 0.39463   
## DIS -1.528485 0.253201 -6.037 0.0000000048181172 \*\*\*  
## RAD 0.324728 0.078585 4.132 0.0000471086479119 \*\*\*  
## TAX -0.014765 0.004461 -3.310 0.00105 \*\*   
## PTRATIO -0.867761 0.158513 -5.474 0.0000000953653006 \*\*\*  
## LSTAT -0.498855 0.062854 -7.937 0.0000000000000461 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.465 on 289 degrees of freedom  
## (298 observations deleted due to missingness)  
## Multiple R-squared: 0.7564, Adjusted R-squared: 0.7463   
## F-statistic: 74.77 on 12 and 289 DF, p-value: < 0.00000000000000022

# Scatter Plot



**Summary**

For this week’s critical thinking assignment, I choose to work on the Linear Regression Model option which had us working with something that I had no prior experience or knowledge of called R Markdown. From we learned from the assigned readings and other research on, R Markdown is format that allows you make dynamic documents using R which can contain bits of code, output, and plotted graphs into an easy to read document.

My First step was to understand the syntax for R Markdown so that I could see what goes into making the Word Document such as what allows the different headers to be created with the # Command, and the comments needed in order to show both the code and the output for each segment ```{r} and ```{r echo=FALSE}.

Next was to start working on the code for the assignment first as per the directions of the assignment I created a new R Markdown file in R Studio following the instructions of the assignment. After the first function I ran was the summary statistics of the BostonHousing.CSV in order to explore the data. From there we are able to see that CAT..MEDV is a categorical variable and would need to be removed in order to see significant data calculations such as: Standard Deviation, Minimum, Maximum, Median, and Length.

After that I needed to process linear model function or lm() in order to process the t-test, f-test, r-squared, residual, and significance values. From there I was able to see that AGE was statically significant so I decided to proceed to plot a scatter plot diagram against MEDV and AGE using MEDV as Y axis variable and X axis variable as AGE using the Plot command. After that it was just simply using theabline(lm(housing.df$AGE ~ housing.df$MEDV), col="red") in order to plot the regression line.