Boston Housing - Homework

Aziz Isamedinov

February 13, 2019

Importing Packages

```
library(readr)
library(ggplot2)
library(corrplot)

## corrplot 0.84 loaded

library(mlbench)
library(reshape2)
library(caret)

## Loading required package: lattice

library(caTools)
library(sjPlot)
library(sjmisc)
library(car)

## Loading required package: carData
```

Read data from SCVDataset

```
#Reading CSV file
boston_h=read.csv(file="C:/Users/azizi/Documents/CSVDatasets/Boston_Housing.csv", sep = "
,")

#Attribute Information
str(boston_h)

## '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    : int  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 ...
##  $ B      : 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 ...
```

Descriptive statistics

```r
#summarize dataframe
summary(boston_h)

##       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             B
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      LSTAT            MEDV
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Data Exploration

1.a. Check for missing values in the dataset.

Answer: No missing values are found

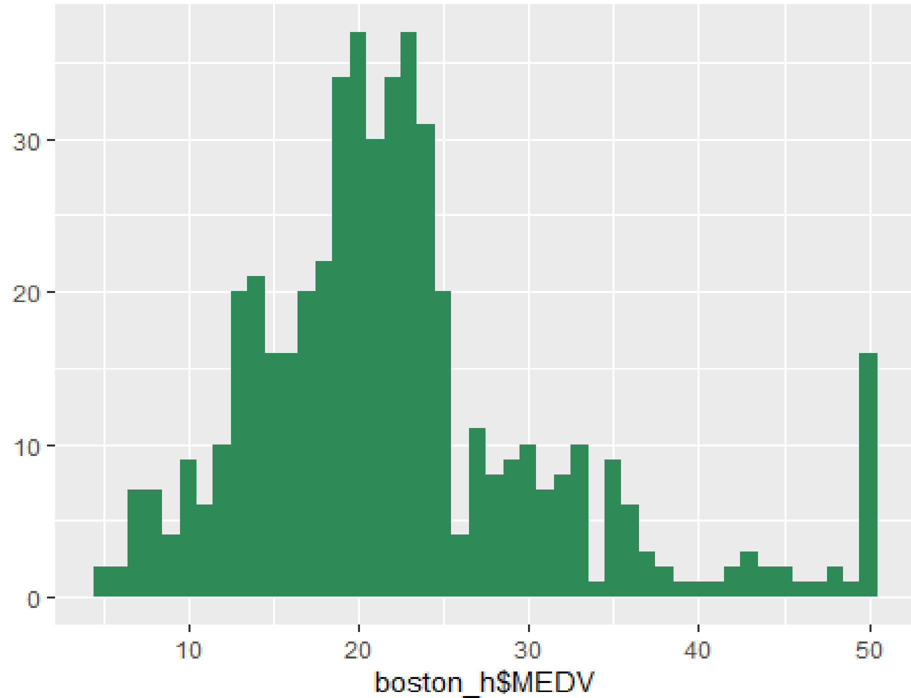
```r
#Searching NA in the dataframe
which(is.na(boston_h))

## integer(0)
```

1.b. Plot the distribution of MEDV. What do you observe?

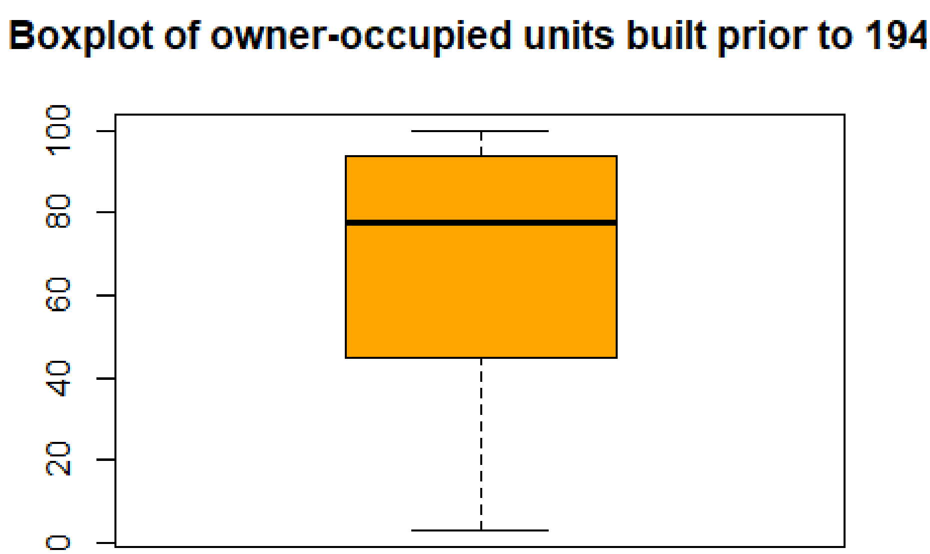
Answer: It can be seen that data is not normally distrubited around the mean of MEDV

```r
#Plot Histogram
qplot(boston_h$MEDV, geom="histogram", fill=I("seagreen"), binwidth=1)
```

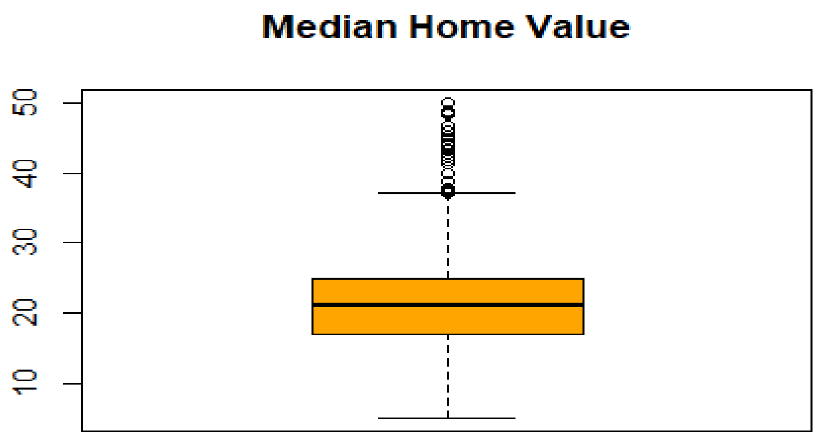


1.c. Generate box-plots of the AGE (proportion of owner-occupied units built prior to 1940) and MEDV (median home value) attributes and identify the cutoff values for outliers.

```
#Generate Boxplot from  AGE data
boxplot(boston_h$AGE,col=c("orange"),main="Boxplot of owner-occupied units built prior to
1940")
```



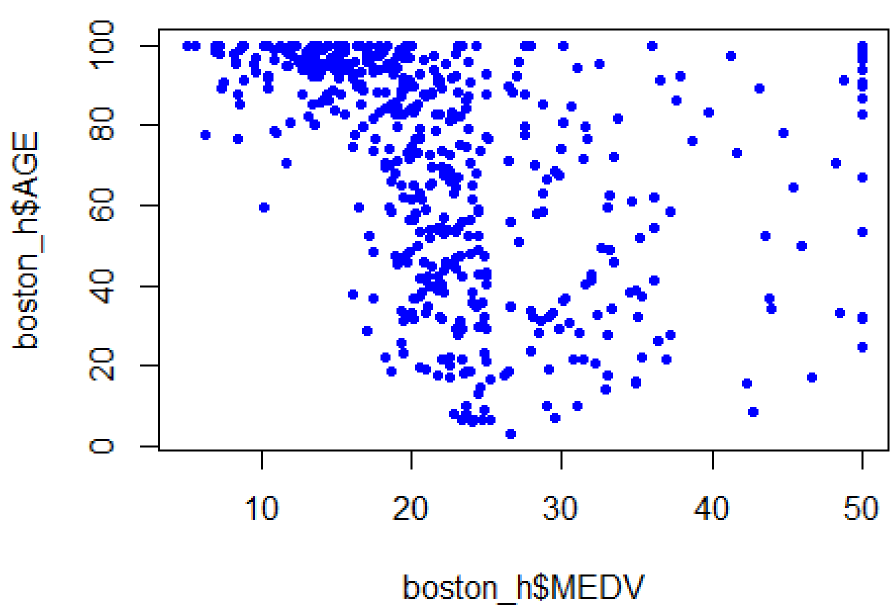
```
#Generate boxplot from  MEDV data
boxplot(boston_h$MEDV,col=c("orange"),main="Median Home Value")
```



1.d.Generate a scatterplot of MEDV against AGE; comment on how inclusion of the outliers would affect a predictive model of median home value as a function of AGE.

Answer: IT is affecting definetely not postively as there can be observed more noise in the data.

```
#Generate Scatter for MEDV vs AGE
plot(boston_h$MEDV, boston_h$AGE, type="p", pch=20, col="blue", lty=3)
```



2. Try to fit an MLR to this dataset, with MEDV as the dependent variable. MEDV has a somewhat long tail and is not so Gaussian-like, so we will take a log transform, (use LMEDV = log(MEDV)), and then predict LMDEV instead. (You should convince yourself that this is a better idea by looking at the histograms and quantile plots to assess normality; however, no need to submit such plots). Keep the first 356 records as a training set (call it Bostrain) which you will use to fit the model; the remaining 150 will be used as a test set (Bostest). Use only LSTAT, RM, TAX, AGE and ZN as independent (predictor) variables and LMEDV as dependent (target) variable as follows when constructing a linear regression model:

```
#Convert MEDV data to log and split data into two parts
boston_h$LMEDV = log(boston_h$MEDV)
bos_train<-boston_h[1:356,]
bos_test<-boston_h[357:506,]

#Run regression with the given model
fit1<-lm(LMEDV~LSTAT+RM+TAX+AGE+ZN, data=bos_train)
summary(fit1)

##
## Call:
## lm(formula = LMEDV ~ LSTAT + RM + TAX + AGE + ZN, data = bos_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37400 -0.08262 -0.01016  0.07685  0.45421
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.396e+00  1.119e-01  12.480  < 2e-16 ***
## LSTAT       -1.003e-02  1.990e-03  -5.041 7.46e-07 ***
## RM           3.251e-01  1.514e-02  21.478  < 2e-16 ***
## TAX         -5.077e-04  1.064e-04  -4.772 2.69e-06 ***
## AGE         -7.358e-04  3.522e-04  -2.089   0.0374 *
## ZN          -3.036e-05  3.187e-04  -0.095   0.9242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1291 on 350 degrees of freedom
## Multiple R-squared:  0.8255, Adjusted R-squared:  0.823
## F-statistic: 331.1 on 5 and 350 DF,  p-value: < 2.2e-16

anova(fit1)

## Analysis of Variance Table
##
## Response: LMEDV
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## LSTAT       1 18.1100 18.1100 1086.6838 < 2.2e-16 ***
## RM          1  8.9216  8.9216  535.3371 < 2.2e-16 ***
## TAX         1  0.4717  0.4717   28.3030 1.855e-07 ***
## AGE         1  0.0887  0.0887    5.3198   0.02167 *
## ZN          1  0.0002  0.0002    0.0091   0.92416
## Residuals 350  5.8329  0.0167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Do any variables have to be dropped because of multicollinearity?(Use VIF criteria to check for multicollinearity)

Answer: Favorably, all the values of VIF are way below critical point which implies that variables are independent and thus are not multicollinear.

```
# viffunction to check wether multicollinearity exists amoung  variables
vif(fit1)

##    LSTAT       RM      TAX      AGE       ZN
## 2.907403 2.218182 1.110602 2.146126 1.506499
```

4.Report the coefficients obtained by your model. Would you drop any of the variables used in your model (based on the t-scores or p-values)?

Answer: From estimates it is clear that, for every unit increase of LSTAT, TAX, AGE, ZN, the predicted value of MEDV would be around -0.0100329, -0.0005077, -0.0007358, -0.0000304 units lower respectively. However, for every unit increase in RM, the MEDV is predicted to be 3.25 unit higher. Let's test null hypothesis that the parameter is equal to zero. In our case coefficients having p values less than 0.05 or less LSTAT, TAX, AGE, ZN, RM are significant. This means that the slope and the t value are in the region of rejection for the null hypothesis.At the 0.05 level of signi???cance, there is evidence that the relationship between prior mentioned independent variables and MEDV indeed exists. While ZN is statistically insignificant as its p-values is higher than 0.05 therefore it is guilty one. Therefore I would drop ZN.

5.Rerun your regression model after removing variables (if any) based on your analysis in the previous question. What is the value of R2 ? What does it signify? What is the overall F-statistic and the corresponding p-value of this final model? What does it signify?

Answer: if we look at R square for goodness of fit, it shows around 83%. So, overall measure of strength association is relatively strong. This shows that the fit of the regression line to the points is fairly good. An R square of 0.83 means that 0.83or 83% of the variation in the values of Y can be explained on the basis of the regression line. The explanation is a statistical one, meaning that 83% of the differences in MEDV's rate in the different provinces are explained statistically by differences in selected predictors

6. What is the overall F-statistic and the corresponding p-value of this final model? What does it signify?

Answer: The important step in a multiple regression analysis is to compute F-statistics and investigate associated p-value. The p-value for F-test is very small, that is smaller than 0.05 and therefore significant. It implies that our group of independent variables reliably predict the dependent variable. In other words, at least one of the predictors is related to the response, then it is natural to wonder which the guilty ones are (for that we looked at T-test above).R

```r
#Regression analysis after removing variables
fit2= lm(LMEDV~LSTAT+RM+TAX+AGE, data=bos_train)
summary(fit2)
```

```
##
## Call:
## lm(formula = LMEDV ~ LSTAT + RM + TAX + AGE, data = bos_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37337 -0.08208 -0.01022  0.07642  0.45535
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.3971629  0.1110078  12.586  < 2e-16 ***
## LSTAT       -0.0100453  0.0019833  -5.065 6.61e-07 ***
## RM           0.3247872  0.0146845  22.118  < 2e-16 ***
## TAX         -0.0005086  0.0001058  -4.807 2.28e-06 ***
## AGE         -0.0007203  0.0003119  -2.310   0.0215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1289 on 351 degrees of freedom
## Multiple R-squared:  0.8255, Adjusted R-squared:  0.8235
## F-statistic: 415.1 on 4 and 351 DF,  p-value: < 2.2e-16
```

7. Report the MSE obtained on Bostrain. How much does this increase when you score your model (i.e.,predict) on Bostest?

Answer: As it is indicated below, There is a slight increase in MSE for test data(bos_test) compared to train data(bos_train)

```r
#Find MSE for the bos train
MSE_bostrain=mean(fit1$residuals^2)
MSE_bostrain
```

```
## [1] 0.01638446
```

```r
# model to predict MSE for bostest
test_predict = predict(fit1, bos_test)
MSE_bostest= mean((bos_test$LMEDV-test_predict)^2)
MSE_bostest
```

```
## [1] 0.1781114
```

(Bonus 1 point). Use the stepwise regression considering to reach your final model (LMEDV as dependent variable and all but MEDV as independent variables). Try different model section criteria (i.e., AIC, Cp, BIC, adj R^2) and see if you can come up with the same model even with the different criteria. Determine the best model if you get different models with different criteria? We will consider a model that gives the highest accuracy (in terms of MSE) in the test set as the best model.

Answer: The best model is as follows: fit4=lm(formula = LMEDV ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, data = bos_train), however based on the MSE statistics the first model(fit3) is prefered.

```
#Determine the best model
step(lm(LMEDV~.-MEDV,data=bos_train))

## Start:  AIC=-1578.21
## LMEDV ~ (CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT + MEDV) - MEDV
##
##           Df Sum of Sq    RSS     AIC
## - CRIM     1    0.0129 3.9212 -1579.0
## - INDUS    1    0.0191 3.9274 -1578.5
## <none>                 3.9083 -1578.2
## - CHAS     1    0.0284 3.9367 -1577.6
## - ZN       1    0.0302 3.9385 -1577.5
## - RAD      1    0.0618 3.9701 -1574.6
## - NOX      1    0.1036 4.0119 -1570.9
## - B        1    0.1719 4.0802 -1564.9
## - AGE      1    0.2807 4.1890 -1555.5
## - TAX      1    0.3537 4.2620 -1549.4
## - LSTAT    1    0.3994 4.3077 -1545.6
## - DIS      1    0.5407 4.4490 -1534.1
## - PTRATIO  1    0.7464 4.6547 -1518.0
## - RM       1    4.9822 8.8906 -1287.6
##
## Step:  AIC=-1579.04
## LMEDV ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##           Df Sum of Sq    RSS     AIC
## <none>                 3.9212 -1579.0
## - INDUS    1    0.0224 3.9436 -1579.0
## - ZN       1    0.0298 3.9510 -1578.3
## - CHAS     1    0.0300 3.9512 -1578.3
## - RAD      1    0.0647 3.9859 -1575.2
## - NOX      1    0.0929 4.0141 -1572.7
## - B        1    0.1592 4.0804 -1566.9
## - AGE      1    0.2814 4.2026 -1556.4
## - TAX      1    0.3435 4.2647 -1551.1
## - LSTAT    1    0.3929 4.3141 -1547.0
## - DIS      1    0.5296 4.4508 -1535.9
## - PTRATIO  1    0.7902 4.7114 -1515.7
## - RM       1    4.9839 8.9051 -1289.0
```

```
##
## Call:
## lm(formula = LMEDV ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS +
##     RAD + TAX + PTRATIO + B + LSTAT, data = bos_train)
##
## Coefficients:
## (Intercept)           ZN        INDUS         CHAS          NOX
##   2.2524769    0.0005252    0.0020575    0.0359232   -0.3095658
##          RM          AGE          DIS          RAD          TAX
##   0.2775077   -0.0016655   -0.0333158    0.0091700   -0.0005494
##     PTRATIO            B        LSTAT
##  -0.0266376    0.0005924   -0.0097952
```

```r
#Using the 1st model calculate MSE
fit3=lm(LMEDV ~ (CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +TAX + PTRATIO + B
+ LSTAT + MEDV) - MEDV, data=bos_train)
fit3_predict = predict(fit3, bos_train)
mse_fit3 = mean(bos_train$LMEDV-fit3_predict)^2

#Using the 2st model calculate MSE
fit4=lm(LMEDV ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTA
T, data= bos_train)
fit4_predict = predict(fit4, bos_train)
mse_fit4 = mean(bos_train$LMEDV-fit4_predict)^2

#Print out all models MSE
options(scipen = 100, digits=4)
mse_fit3
```

```
## [1] 0.00000000000000000000000000003758
```

```r
mse_fit4
```

```
## [1] 0.000000000000000000000000000007779
```