STAC67 Case Study: Predictive Model For The Median Value Of Homes In Boston

Group 18 / Abbas Rai 1003957104 / Calvin Chan 999940364 / Huan Wang 1001707049/ Ruotong Zheng 1002586659/ Andrey Zhuravlev

# Abstract

Buying a home is one of the most important purchase decisions in oneâs life. There are many important factors for one to consider, especially the value of the home one is purchasing. In reality, the value of a home can be evaluated based on an in-depth analysis of the homeâs geographical location and several other environmental and socioeconomic factors. Our team has used a plethora of statistical tests and considered several influential variables associated with this important topic in order to develop a predictive model which best explains the value of homes in the city of Boston.

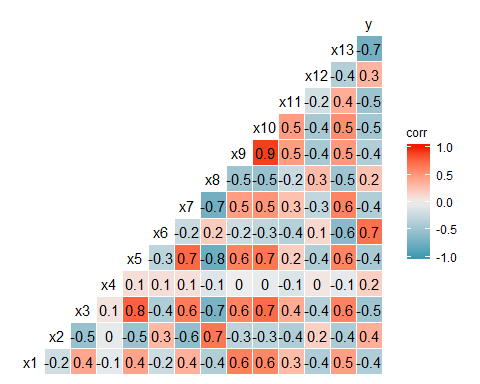
# Background and Significance

The debate surrounding purchasing the “ideal” home is one that is extremely prevalent within households across the globe. People work extremely hard to save up to buy a home, making it one of the most important investments in one’s lifetime. In fact, research shows that it takes the average American more than seven years to save up enough money just to deposit the down payment towards a home (Olsen, 2018). Moreover, future home buyers find themselves in an even more vulnerable position. Due to inflation and increasing price levels within the economy, there has been an increasing trend of the time needed to save up for a home in America, which has already risen by almost two years since 1988 (Olsen, 2018). The presence of these trends makes it even more important for individuals to make smart home purchases, in order to avoid facing severe economic repercussions. One of the fundamental factors which affects the value of a home is safety. A report published by the Center for American Progress concluded that âa 10% reduction in homicides would lead to a 0.83% increase in housing values the following year.â (Byloos, 2016). In addition, one must consider the presence of retail stores in the nearby area, which can actually lower the value of a house if located within close proximity (Matthews, 2006). We ultimately propose the following model to predict the median house price levels in Boston in order to provide home buyers and real estate agents with more holistic understanding of the factors which influence the value of homes and provide them with the necessary information to understand the housing market better and make better decisions overall.

## x1 x2 x3 x4 x5 x6 x7   
## 0.01501 0.00000 18.10000 0.00000 0.53800 5.71300 100.00000   
## x8 x9 x10 x11 x12 x13 y   
## 3.49520 24.00000 666.00000 20.20000 396.90000 8.05000 50.00000

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa



# Model

# Model Creation

Half the data set is choosen to create the model and other half is for validation. Seed is set to 67 for consistent results. We used stepwise regression with function stepAIC with 12 variables which checks which variables have an Akaike’s Information Criterion(AIC) lower than the threshold AIC. X4 is ignored since only 6% of the data set has X4 at 1.

The resulting model is Y ~ X2 + X5 + X6 + X8 + X9 + X10 + X11 + X12 + X13.

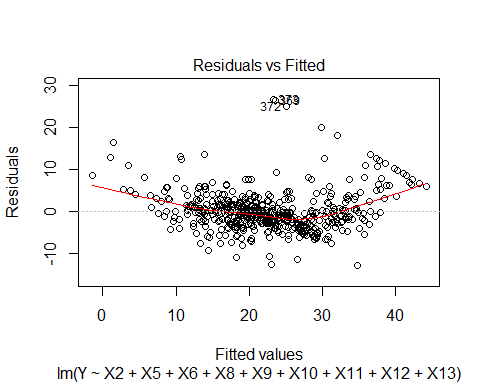
# Model validation

The remaining 50% of the data set is used for validation. We checked the difference between the Mean Square Prediction Error(MSPE) and the Mean Square Residual(MSRes).

We got 23.217 as the MSPE and 22.251 for the MSRes. The difference(0.966) is fairly close so we can validate this model

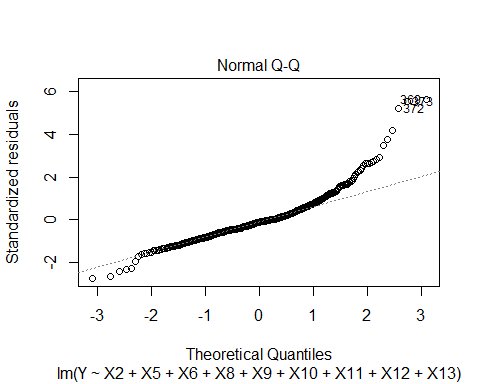
# Model diagnostics

## Residuals Vs. Fitted Values



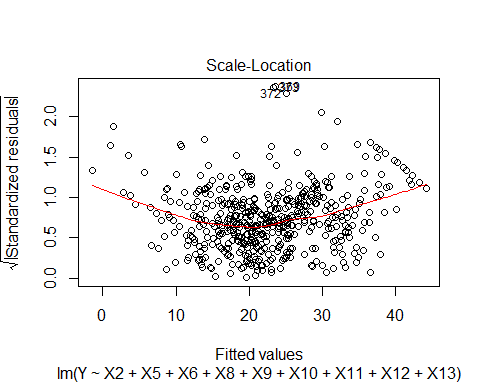
The Residuals vs Fitted plot shows a random spread points therefore it indicates a proper functional form.

## Normal Q-Q Plot



The normal Q-Q plot shows that the errors are normally distributed. Most of the residuals follow the straight line but there are many residuals off the line towards the ends. There are three noteworthy points of 369, 373, and 372. Since most of the residuals are along the line, there is no indication that normality is violated.

## Scale-Location



## Outlying Y observations

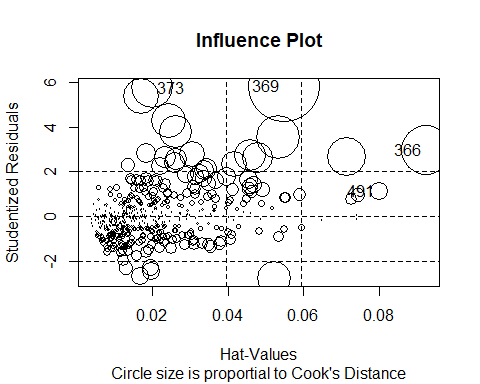
We found the threshold(t\_crit) to be 3.925493. Using studentized deleted residuals we found that observations 369, 370, 372,and 373 above the threshold and therefore are the outlying Y observations.

## Outlying X observations (Leverage)

No leverage points is with guideline 2 which is Pii > 0.5. With the guideline Pii > 2\*p\_prime/n there are 46 leverage points.

[9, 49, 55, 103, 142, 143, 144, 145, 146, 147, 148, 149 151, 152, 153, 154, 155, 156, 157, 160, 204, 205, 254, 258, 291, 292, 293, 352, 353, 354, 355, 356, 365, 366, 368, 369, 375, 411, 413, 415, 425, 489, 490, 491, 492, 493]

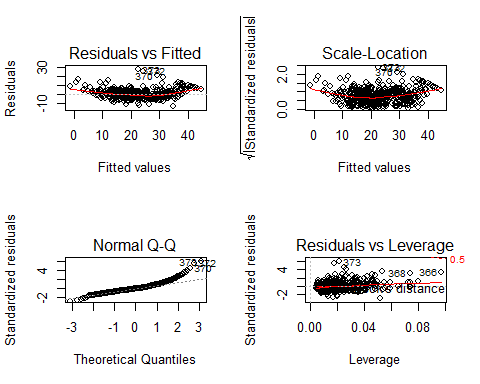
## Influence



## StudRes Hat CookD  
## 366 2.965550 0.09222629 0.08796623  
## 369 5.815164 0.05489718 0.18423487  
## 373 5.747572 0.02022483 0.06405406  
## 491 1.131307 0.07996003 0.01111687

There is only one data point that has a DFFITS greater than 1 which is 369.

There are no observations with a Cooks Distance greater than the 20th percentile of F(10,496) = 0.6168.

There are no observations with a DFBETAS greater than 1. 

# References

Olsen, S. (2018, October 22). Home Buyers Need 7.2 Years to Save Down Payments â 1.5 Years More Than in 1988. Retrieved from <https://www.zillow.com/research/how-many-years-down-payment-21734/>

Byloos, M. (2016). Research Crime Rates and the Impact on Home Values | Homes.com. Retrieved from <https://www.homes.com/blog/2016/05/secure-new-home-research-crime-rates-impact-home-value/>

Matthews, J. (2006). Retail Proximity and Residential Values or Do Nearby Stores Really Run Down Property Values?. SSRN Electronic Journal. doi: 10.2139/ssrn.989049