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 1003148961

April 4, 2019

# Abstract

When purchasing a home, there are many important factors for one to consider, especially the value of the home one is purchasing. The value of a home can be through analyzing geographical location and several other environmental and socioeconomic factors. We have used a plethora of statistical tests and considered several variables associated with this topic in to develop a predictive model which best explains the value of homes in the city of Boston.

# Background and Significance

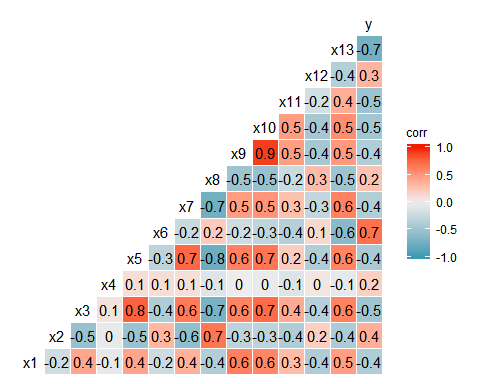
Purchasing the “ideal” home is pertinent to individuals across the globe and is one of the most important investments in one’s lifetime. Studies show that it takes the average American more than seven years to save up enough money just to deposit the down payment towards a home (Olsen, 2018). Due to 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). These trends make it necessary for individuals to make smart home purchases. A key factor affecting house prices is safety. In 2019, the median home value in Boston was $600,100 which increased by 5.7% from 2015. In fact, since 2013, housing prices in Boston area have risen significantly. The prices of detached houses rose by 18.8%, while prices of apartments increased by 22.8%. From the Boston real estate data, supply has been decreasing. The supply of detached houses has fallen by 38.2%, and the supply of apartments has fallen by 44% (Zillow 2019). 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 a more holistic understanding of the factors which influence the value of homes and to 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

## Warning: package 'GGally' was built under R version 3.5.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.5.3



# Exploratory data analysis

This data set contains 506 observations (median value of owner occupied homes in thousands, MEDV) and 13 other. predictors, which are per capita crime rate by town (x1), proportion of residential land zoned for lots over 25,000 sq.ft. (x2), proportion of non-retail business acres per town (x3), Charles River dummy variable (x4), nitric oxide concentration (parts per 10 million) (x5), average number of rooms per dwelling (x6), proportion of owner occupied units built prior to 1940 (x7), weighted distances to five Boston employment centres (x8), index of accessibility to radial highways (x9), full-value property-tax rate per 10,000 (x10), pupil-teacher ratio by town (x11), 1000(B-0.63)2 where B is the proportion of African Americans by town (x12), a numeric vector of percentage values of lower status population (x13).

# House value

Property(house) value is the worth(price) of a piece of real estate which both of the buyer and seller agree upon(Bankrate 2019). In this data, we use the median value of the house value of owner occupied homes in thousands (MEDV) dollars. What buyers are willing to pay for property depends on a number of factors which in our data there are 13 predictors that may affect the house value. In our data, the Min, Mean,Max,Mode of MEDV are 5, 22.53, 50, and 50 respectively.

# Per capita crime rate by town (x1)

Crime rate, measured in per capita by town. A report published by the Center for American Progress concluded that “a 10% reduction in homicides would lead to a 0.83% increase in housing values the following year.” (Byloos, 2016).The Mean and Mode of crime rate are 3.61352, and 0.01501 respectively. The minimum and maximum crime rates committed in town are 0.0632 and 88.9762 respectively.

# Lots over 25,000 sq.ft. (x2)

This concerns about the proportion of residential land zoned for lots over 25,000 sq.ft..The Min, Mean,Max of it are 0, 11.36, and 100 respectively.

# Proportion of non-retail business acres per town (x3)

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).The aces of town that is of non-retail business in proportion takes up on average 11.13678 aces. The min and max acres are 0.64 and 27.76.

# Charles River dummy variable (x4)

The Charles River dummy variable measures of either 1 or 0 to represent whether the property is located alongside the Charles River or not. Of the 506 observed locations, only 35 locations (6.917%) fulfill this variable.

# Nitric oxide concentration (parts per 10 million) (x5)

Properties locate within 0.5546951( parts per 10 million) nitric oxide concentration. The concentrate ranges from 0.385 to 0.871(parts per 10 million concentrate) for located housing.

# Average number of rooms per dwelling (x6)

The average rooms per dwelling is 6.284634 with the min and max dwellings being 3.561 and 8.78 respectively.

# Proportion of owner occupied units built prior to 1940 (x7)

Owner occupied units has raised to about 68.5749 percent prior to 1940. Since 1940, at lowest only 2.9 percent of owners occupied units while at max 100 percent of the units were occupied.

# Weighted distances to five Boston employment centres (x8)

3.79503 is the mean weighted distance to five Boston employment centres. At max is 12.1695 and at min the distance is 1.1296

# Index of accessibility to radial highways (x9)

We are on average 9.549407 accessible to radial highways.It also seems to have a strong correlation to X10.

# Full-value property-tax rate per 10,000 (x10)

The mean, min, max full-value property tax rate per 10,000 is 408.2372, 187, 711 respectively.

# Pupil-teacher ratio by town (x11)

The mean pupil to teacher range is about 18 students to 1 teacher by town. The median pupil to teacher ratio is 19 students to 1 teacher.

# African Americans by town (x12)

This represents 1000(B-0.63)^2 where B is the proportion of African Americans by town. The mean and mode are 356.67 and 396.9 respectively.

# A numeric vector of percentage values of lower status population (x13)

12.65306 of the housing are of lower status populated.

# 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.

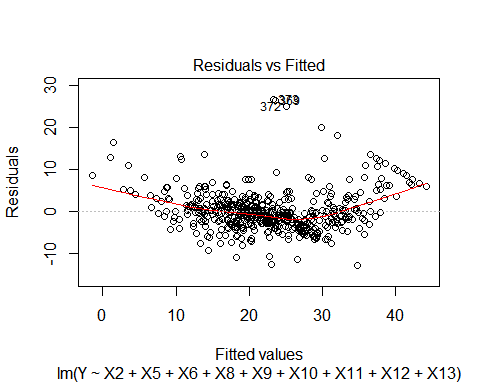
# Model validation

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

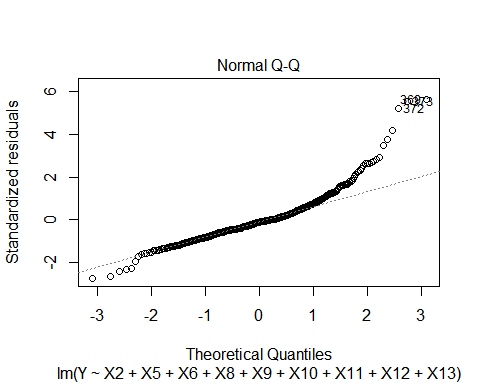
# Model diagnostics

## Residuals Vs. Fitted Values



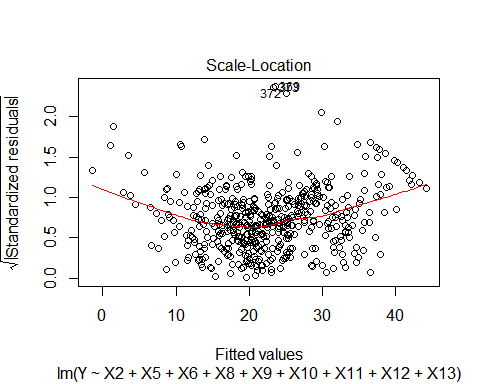
The Residuals vs Fitted plot shows a random spread points therefore it indicates a proper functional form.

## 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

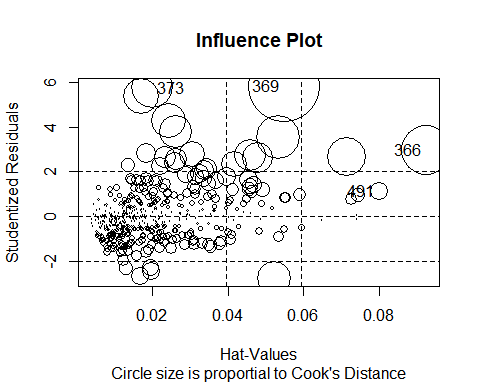
We found the threshold(t\_crit) to be 3.925493. Using studentized deleted residuals we found that observations 369, 370, 372,and 373 above the threshold and therefore are the outlying Y observations.

## 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]

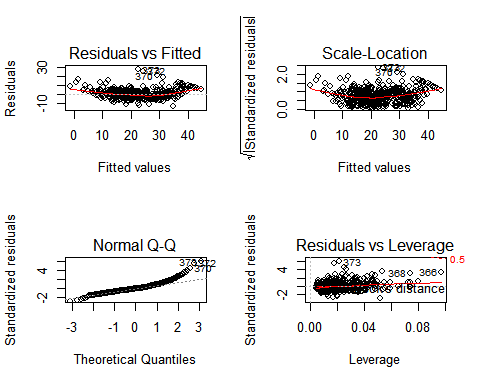
## Influence



## 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. 

# Conclusion

The goal of this case study was to build a predictive model that best explains the median value of homes in Boston. During the study, we found that the difference between the MSPE and the MSRes (0.966) was fairly close, allowing us to validate our model. In addition, most of the residuals follow a straight line, which removes the indication that normality is violated. Our findings will impact the field by giving home buyers and real estate agents a model that can understand the housing market better. If these individuals can understand the factors which can increase or decrease the value of a home, it will allow home buyers to purchase homes which will yield a better return in the long-term, while real estate agents will be able to predict the trends in the housing market by understanding the factors which affect house prices. Some limitations in our findings include the sample size being only 506 observations and the concentrated sample with data pertaining to house prices in Boston only. This makes it difficult to get a holistic understanding and being able to confidently confirm that these trends could apply to Americans in general or even to housing markets in other countries. A possible area of future study that can overcome our current limitations is expanding the sample size both in population and making the study a more globally inclusive one in order to gain more powerful insights and affect individuals across the world.

# References

Bankrate. (2019). Property value Definition | Bankrate.com. [online] Available at: <https://www.bankrate.com/glossary/p/property-value/> [Accessed 30 Mar. 2019].

Zillow.com. (2019). Boston Home Prices & Values. [online] Available at: <https://www.zillow.com/boston-ma/home-values/> [Accessed 30 Mar. 2019].

Olsen, S. (2018, October 22). Home Buyers Need 7.2 Years to Save Down Payments - 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