STAC67 Case Study: A Model for Predicting Median Value of Homes in Boston

Group 04 / Yuhan Wang 1002917169 / Yiming Fu 1002928246 / Arda Erturk 1002937799
Syed Zain Zafar 1002534705

April 5, 2019

Abstract

The median value of homes in a certain area is one of the popular indicators for monitoring the property market. It gives buyers and sellers a better indication of market trends, consumer sentiment and market conditions. This study explores the median value of homes in Boston area. We consider 13 potentially related factors, such as crime rate and the number of rooms. We used R and built the regression model to find a relationship between the median value of homes in Boston and those influential factors.

Background and Significance

The median sale price is the value in the middle of the data set when we arrange all the sale prices from low to high. It is a widely used index to estimate the listing price of a property. It is often said that the median house price is a better indicator than the mean house price within an area because it is not affected by outliers.

If the median sale price is trending down, it might take longer to sell a home and buyers might have more power to negotiate. If it is trending up, the market might be in demand and houses will likely be selling more quickly. Sellers will have the advantage when prices are going up and buyers will have less power to negotiate.

In North America, the real estate industry has grown rapidly in the past decade and the 2018 Housing and Mortgage Market Review estimates home prices will continue to rise for the next couple of years (Ramsey, 2019).

Therefore, to maximize the benefits, price judgment of the homes becomes a significant issue for both buyers and sellers. For example, if a seller's home price is underpriced, he/she may lose thousands of dollars. This can be easily improved by understanding the median price of the market as a reference.

Exploratory Data Analysis

The model has been built based on the dataset with 506 observations. Depending on the information we obtain from the data set, there are 13 predictors we consider that may have effect on the median value of homes in Boston.

Median Value of Owner-Occupied Homes in Thousands of Dollars (y)

These data are collected in the American Community Survey (ACS). The data are period estimates, that is, they represent the characteristics of the housing over a specific 60-month data collection period.

Per Capita Crime Rate by Town (x1)

The mean and mode of per capita crime rate by town is 3.613 and 0.015 respectively.

Proportion of Residential Land Zoned for Lots Over 25,000 sq. Ft. (x2)

This predictor is concerned about people's demand for the size of the floor space. Different demand could determine market competitiveness of the house and affect house price. The mean and mode of proportion of residential land zoned for lots over 25,000 sq. Ft. is 11.363 and 0 respectively.

Proportion of Non-Retail Business Acres Per Town (x3)

Non-retail business is the selling of goods and services outside the confines of a retail facility, such as vending, mobile vending and telemarketing, internet marketing. The mean and mode of proportion of non-retail business acres per town is 11.136 and 18.1 respectively.

Charles River Dummy Variable (x4)

It refers to the houses that are close to the Charles River. This is a binary variable(= 1 if tract bounds river; 0 otherwise). The mean and mode of Charles River dummy variable is 0.069 and 0 respectively.

Nitric Oxide Concentration (Parts Per 10 Million) (x5)

Nitric oxide is a toxic gas that is colorless and odorless. Negative side effects may include nausea or vomiting, headache, and/ or shivering. The mean and mode of nitric oxide concentration is 0.554 and 0.538 respectively.

Average Number of Rooms Per Dwelling (x6)

The mean and mode of average number of rooms per dwelling is 6.284 and 5.713 respectively.

Proportion of Owner-Occupied Units Built Prior to 1940 (x7)

The mean and mode of proportion of owner-occupied units built prior to 1940 concentration is 68.574 and 100 respectively.

Weighted Distances to Five Boston Employment Centres (x8)

The mean and mode of weighted distances to five Boston employment centres is 3.795 and 3.495 respectively.

Index of Accessibility to Radial Highways (x9)

The mean and mode of index of accessibility to radial highways is 9.549 and 24 respectively.

Full-Value Property-Tax Rate Per 10,000 (x10)

Property tax is assessed through the environment around the house and the value of the house. The mean and mode of full-value property-tax rate per 10,000 is 408.237 and 666 respectively.

Pupil-Teacher Ratio by Town (x11)

The proportion of teachers and students in this area, which also can reflect education level in the area in some extent. The mean and mode of pupil-teacher ratio by town is 18.455 and 20.2 respectively.

$1000(B - 0.63)^2$ Where B is The Proportion of African Americans by Town (x12)

The mean and mode of $1000(B - 0.63)^2$ where B is the proportion of African Americans by town is 356.674 and 396.9 respectively.

A Numeric Vector of Percentage Values of Lower Status Population (x13)

The mean and mode of a numeric vector of percentage values of lower status population is 12.653 and 8.05 respectively.

Model

Model Selection

We split the dataset into 2 sets by taking a 70/30 split. The 2 sets are: training and testing datasets, consisting of 354 and 152 observations, respectively. We started by using the training dataset to construct a model consisting of all variables and plotting the residuals against all 13 explanatory variables independently to observe any trends. We transformed x8 and x13 by taking their natural logarithm as their plots were better modeled by the natural logarithm function. We decided to transform the response variable by power of 0.5 to improve normality of residuals.

We utilized Akaike's Information Criterion to derive a better model. The final model (houseModel) is: lm(y^0.5 ~ x1 + x4 + x5 + x6 + log(x8) + x9 + x10 + x11 + x12 + log(x13)).

This model was verified by utilizing the testing data. The Mean Square Residual (MSRes) of the fitted model is 0.159 and the Mean Square Prediction Error (MSPE) is 0.167. Since the MPSE and MSRes are extremely close, this represents an appropriate indication of the predictive ability of the model and we conclude that the model is valid.

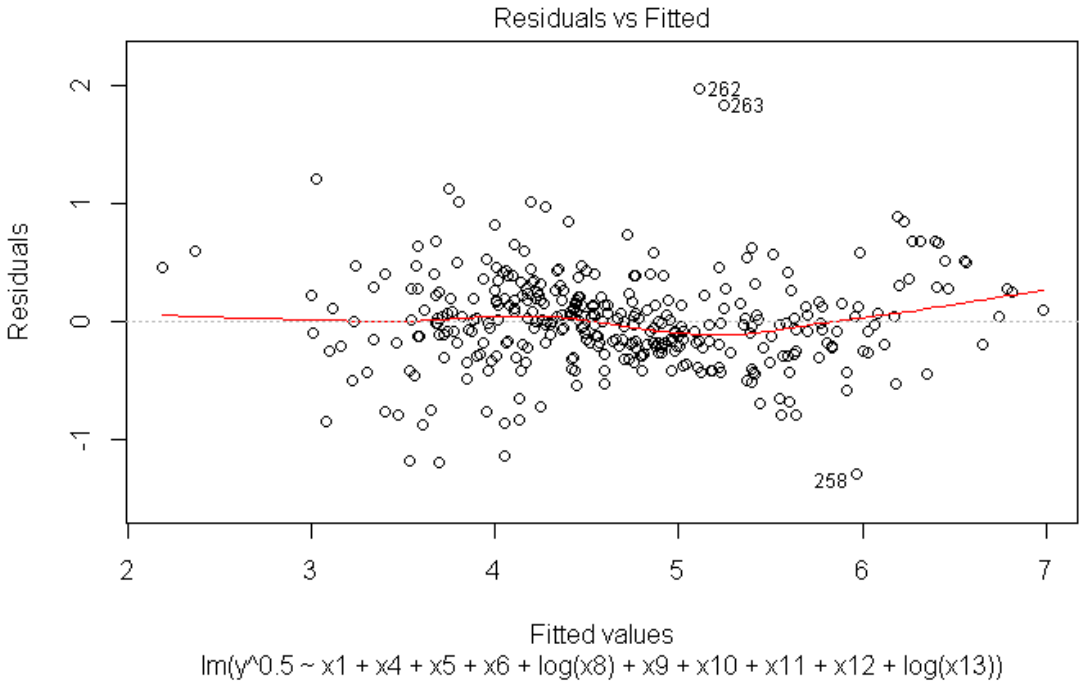
Model Summary

```
summary(houseModel)

##
## Call:
## lm(formula = y^0.5 ~ x1 + x4 + x5 + x6 + log(x8) + x9 + x10 +
##     x11 + x12 + log(x13), data = houseTrainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28558 -0.21478 -0.01176  0.20736  1.95913
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.0483812  0.5825512  15.532  < 2e-16 ***
## x1          -0.0198658  0.0033762  -5.884 9.52e-09 ***
## x4           0.2109534  0.0887174   2.378 0.017963 *
## x5          -2.4109638  0.3835818  -6.285 9.93e-10 ***
## x6           0.2285157  0.0429131   5.325 1.83e-07 ***
## log(x8)     -0.6906033  0.0758630  -9.103  < 2e-16 ***
## x9           0.0310942  0.0060300   5.157 4.26e-07 ***
## x10         -0.0013524  0.0003099  -4.364 1.69e-05 ***
## x11         -0.0905387  0.0122114  -7.414 9.65e-13 ***
## x12          0.0009690  0.0002867   3.380 0.000809 ***
## log(x13)    -0.8497866  0.0632198 -13.442  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4057 on 343 degrees of freedom
## Multiple R-squared:  0.8152, Adjusted R-squared:  0.8098
## F-statistic: 151.3 on 10 and 343 DF,  p-value: < 2.2e-16
```

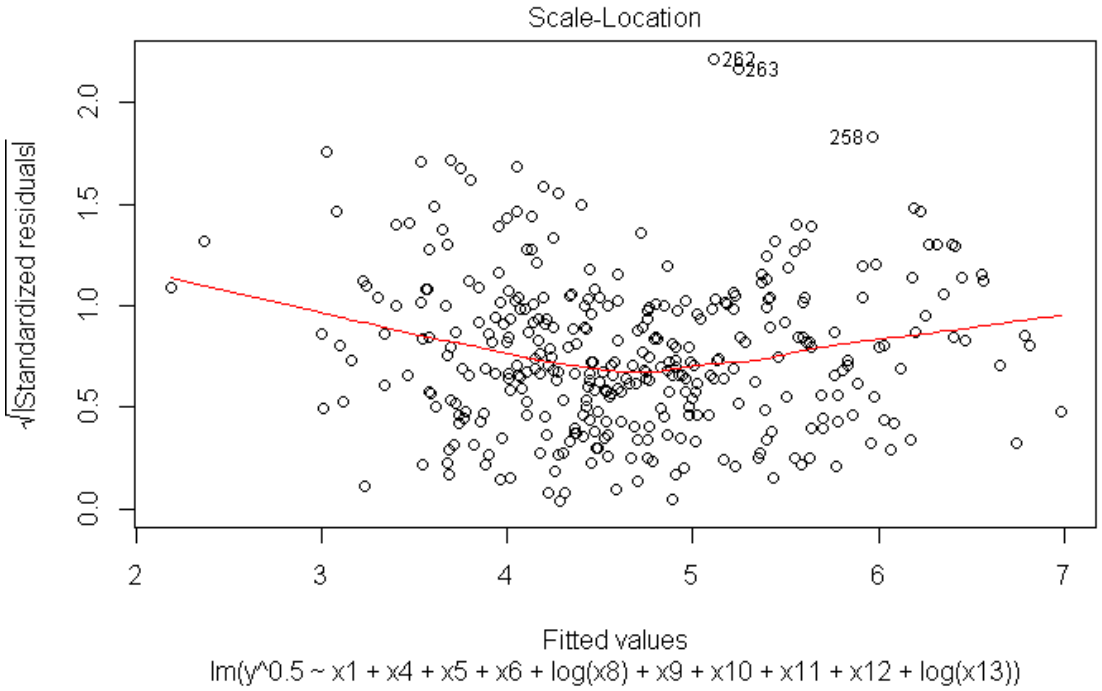
Model Diagnosis

Residuals Against Fitted Values



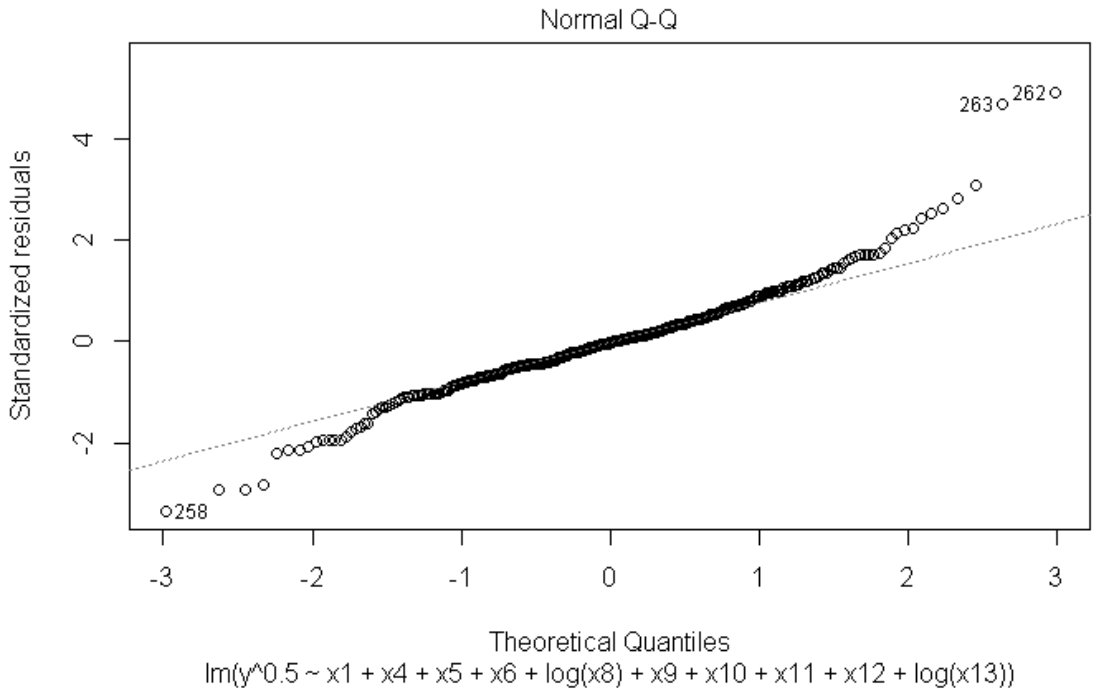
The residuals are randomly and evenly distributed along the red nearly-horizontal line. There is no indication of the linear assumption being violated.

Scale-Location Plot



The red line is not completely horizontal, but this should suffice. The points are randomly spread out and so the homoscedasticity assumption holds.

Normal Q-Q Plot



The Normal Q-Q plot of the residuals shows that majority of the observations are on the line, the only exceptions are 258, 263, and the 262 observations. Since majority of the points are on the line, the normality assumption holds.

Outlying Y Observations

The 262 and 263 observations stand out as outlying Y's as their studentized residuals are larger than the threshold (t_crit = 3.84950).

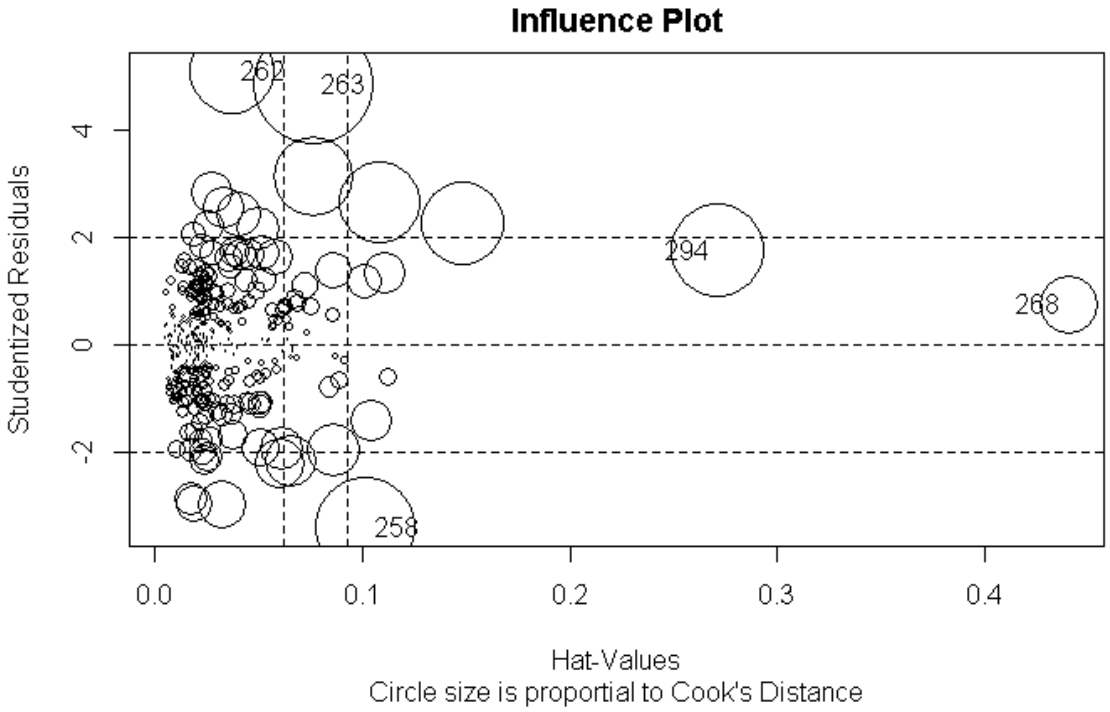
Outlying X Observations

All observations with a leverage greater than twice the mean leverage value (0.06214689) are: 67, 97, 98, 99, 103, 104, 105, 106, 107, 109, 110, 140, 150, 252, 253, 258, 259, 260, 261, 263, 268, 281, 285, 288, 289, 291, 294, 297, 312, 315, 341, 342, 343, 344, and 345.

There are no observations with leverage greater than 0.5.

The observations above are leverage points.

Influential Observations




```
##         StudRes        Hat      CookD
## 258 -3.3945524 0.10175407 0.11513446
## 262  5.0972814 0.03701878 0.08463618
## 263  4.8454126 0.07641452 0.16572975
## 268  0.7464541 0.43995032 0.03984296
## 294  1.7398371 0.27154516 0.10197764
```

The 10th, 20th, and 50th percentile of F(11, 343) are: 0.5045727, 0.633494, and 0.9419, respectively. None of the observations above have a Cook's Distance close to 0.9419 and all of the observations above have a smaller Cook's Distance than 0.5045727 and 0.6334948, in other words, no observations are influential.

By Cook's Distance and DFBETAS, no observations were found to be influential. However, by DFFITS, observations 258, 263, and 294 were found to be influential.

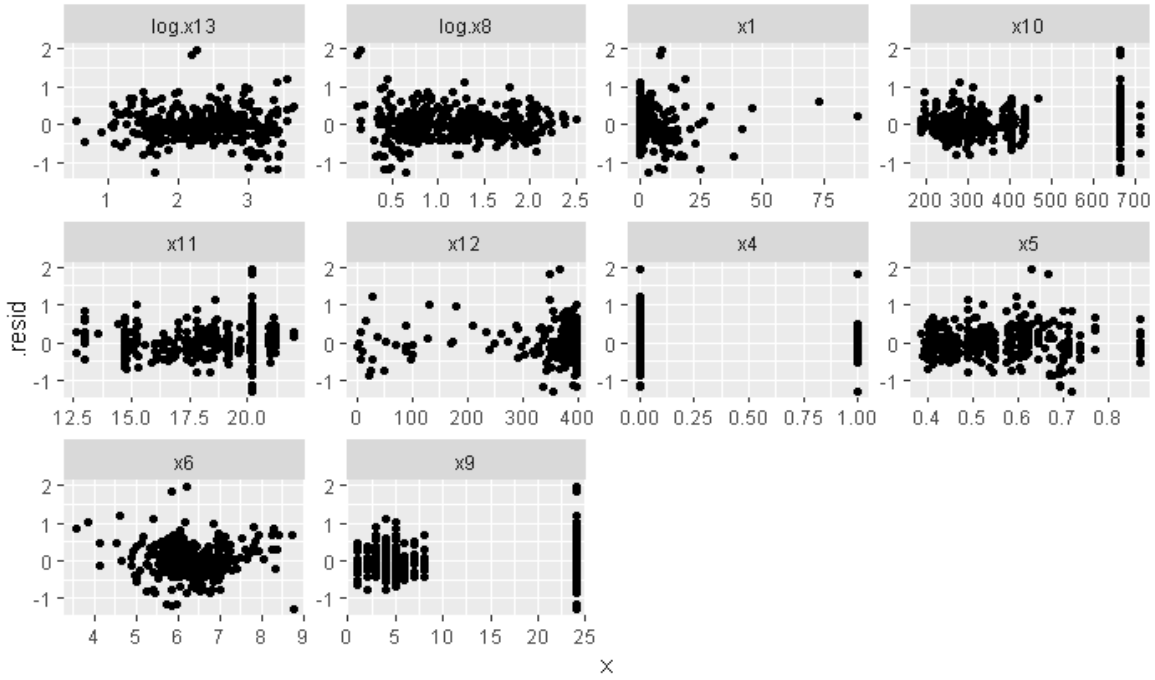
Overall, the 258 observation is a leverage point and is also influential. The 263 observation is an outlying y and is influential. The 294 observation is a leverage point and is also influential. The 262 observation is an outlier as it is an outlying y that is not influential. We dealt with this by removing this observation from the training data afterwards.

All other leverage points that are not influential can be removed, but we have decided not to throw away data as deletion has little effect on model.

Multicollinearity

The largest VIF value is 5.836977 which is less than 10 and the mean VIF value is 3.038436 which is not considerably larger than 1 and thus not indicative of serious multicollinearity.

Residuals Against Explanatory Variables



Most of the plots follow no pattern. Residuals against x4 and residuals against x9 plots are non-linear, but after trying various transformations, such as squaring, taking natural logarithm, and taking inverse of x4 and x9 to better the fit, nothing has seemed to improve the model.

Discussion/Conclusions

The goal of this study is to build a predictive model that best explains the median value of homes in Boston. During the study, we found that the square root of median value of homes (y^0.5) can be predicted by x1, x4, x5, x6, log of x8, x9, x10, x11, x12, and log of x13. The slopes indicate that y^0.5 increases as x4, x6, x9, and x12 increase and y^0.5 decreases as x1, x5, log of x8, x10, x11, and log of x13 increase. Our findings will impact the field by providing home buyers and sellers with a model that can predict median house prices. Home buyers and sellers can compute the predicted house price in the neighborhood to better judge the price they should buy/sell a house for. A limitation in our study is that the residuals against x4 plot has improper functional form as x4 is a binary variable. A possible area of future study is to increase the number of observations to make the predictive model more reliable.

References

Hastreiter, N. What's the Future of Real Estate? (2018, April 17). Retrieved from https://www.futureofeverything.io/ask-the-thought-leaders-whats-the-future-of-real-estate/

Ramsey, D. 2019 Real Estate Trends: What You Need to Know. (2019, January 05) Retrieved from https://www.daveramsey.com/blog/real-estate-trends

Why is it so Important that your Home is Priced Correctly? Retrieved from http://www.taramoorerealestate.com/pricing-properly/

Maxmino, M. The impact of crime on property values: Research roundup. (2017, February 16). Retrieved from https://journalistsresource.org/studies/economics/real-estate/the-impact-of-crime-on-property-values-research-roundup/

California Dental Association. Nitrous Oxide. Retrieved from https://www.cda.org/portals/0/pdfs/fact_sheets/nitrous_oxide_english.pdf.

American Community Survey (ACS) and Puerto Rico Community Survey (PRCS). Retrieved from https://www.census.gov/quickfacts/fact/note/US/HSG495217