STAC67 Case Study: A model for predicting average housing values

Group 30: Ying Xie / Mayuri Kakkar / Li Monique / Chengli Yang / Shamsun Nila

2019-04-02

Abstract

Housing values are one of the standard metrics used by many shareholders in the economy. The median values help homebuyers and insurance companies to make more efficient choices. There are many aspects that cause fluctuation of housing values. To detect what factors contribute to the housing values, we attempt to create a model with elements chosen through extensive data analysis. We use R for statistical analysis to find a relationship between the median housing values and the considered factors in the model.

Background & Significance

The housing market has always been an influential factor for inhabitants in a particular region. Various aspects of the properties and the impression of the surrounding neighborhood shape the housing values. Interpreting what alters housing prices can reveal what people look for in long term residences and this can assist politicians, homebuyers, homeowners and insurance companies in making better decisions. The average number of rooms per home and property-tax rate are taken as significant predictors for housing prices due to high correlation with median housing values. In this study, we strive to understand the connection between the variables and the housing prices in the suburbs of Boston; we attempt to produce a model that can accurately describe the relationship based on the available data.

Exploratory Data Analysis

Crime rate (CR) Per capita crime rate represents the number of police-reported incidents for each person in the town. Calculation of the rate considers dividing the number of occurrences of crime by the total population of the city. From the available data set, we find average crime rate is 3.61 crimes per person, and the mode is 0.015 crimes per person.

Residential Land Zone (RLZ) Residential land zoned represents the numeric vector of the proportion of residential land zoned for lots over 25,000 sq. ft. (square feet), where residential property is specifically for homes and is divided into lots to build subdivisions. The average proportion is 11.36 and mode is 0 for lots over 25,000 sq. ft.

Non-retail business acres (NRBA) Non-retail business is the sale of goods happening outside the traditional retail channel or physical retail space. The type of non-retail business includes direct marketing, direct sales and automatic vending. The average proportion of non-retail business is 11.13 acres per town, and the mode is 18.10 acres per town.

Charles River (CHR) The Charles River dummy variable refers to the river that runs through eastern Massachusetts. The data is categorical and is given one if the river runs nearby and 0 otherwise. This variable represents the relationship between the river presence nearby with housing prices.

Nitric Oxide Concentration (NOC) In general, nitric oxide concentration has negative correlations with housing prices due to it being a pollutant, which automatically decreases the value of a home. From our dataset, we attain the average nitric oxide concentration in owner-occupied houses to be 0.555 ug/mˆ3, micrograms per cubic meter while the mode is 0.538 ug/mˆ3.

Average number of rooms (ANR) The average number of bedrooms per dwelling has a strong correlation since it is considered as a proper measure for housing values. The mean number of rooms is 6.3 while the mode is 5.7 rooms per housing property.

Built prior to 1940 The proportion of owner-occupied homes constructed before 1940 indicates all the old houses in the suburbs of Boston. We calculate, in average, 69 homes were built before 1940, while the most number of homes built before 1940 is 100.

Distances to employment centres The weighted distances to the five Boston employment centers are taken an essential predictor for housing values due to the reluctance of people to travel far for work and a higher number of job opportunities in urban areas, (i.e., negative correlation with housing prices). The mean distance, measured in kilometers is 3.79, while the mode for weighted distances is 3.49.

Accessibility to highways (ATH) Another significant predictor for housing values is the index for accessibility to radial highways since it can have a positive effect due to the utility of having more efficient travel along with an adverse impact due to noise, pollution, and traffic. The mean and mode for the index for accessibility to radial highways is to be 9.5 and 24 respectively.

Property-Tax Rate (PTxR) The full-value property-tax rate refers to the estimated property tax rate per $1000 in Boston; however, in this case, it is per $10,000. The mean and mode of the property tax rate are $408.23 and $666 per $10,000 in assessed value respectively.

Pupil-Teacher Ratio (PTR) The pupil-teacher ratio represents the average number of students per teacher in a town. Usually, if there are lesser teachers per student, the lower the housing value tends to be. We find that there are approximately 18 students for every teacher (18.45:1). In average, the maximum number of students per teacher is 20.

African Americans (AA) Given the formula 1000(B - 0.63)ˆ2, we can calculate the proportion of African Americans by the town. In this formula, B represents the proportion of African Americans by the city. From our dataset, we find that the values for mean and mode for the proportion of African Americans are 356.67 and 396 respectively.

Lower Status Population (LSP) The lower status population refers to the percentage of people under the poverty line. We find that in average 12.65 % of the population belongs to the lower status and the mode is 8.05 %.

Median Value of Housing Properties (MVH) The response variable, housing values, represents the cost of the residences located in the suburbs of Boston. The mean value of housing prices is 22.53 thousands of dollars while the mode of housing prices is 50 thousands of dollars

---

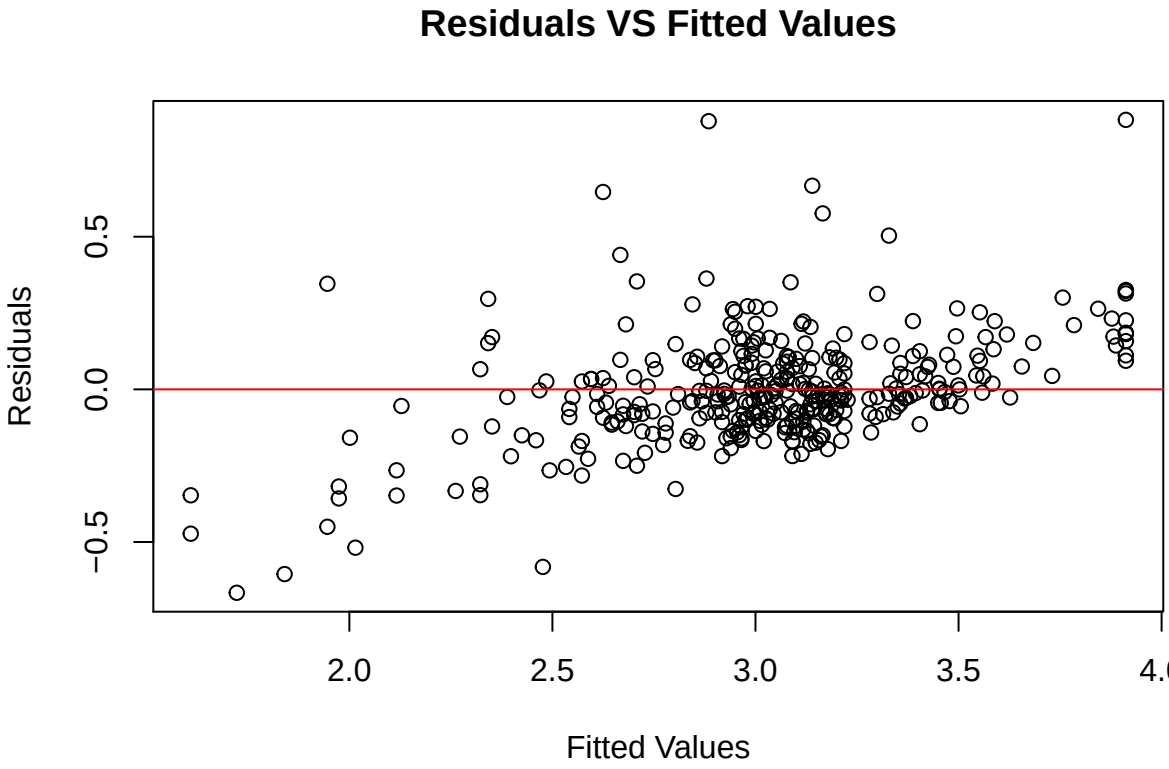
Model

Model Selection From the given data set, we use 70% of the observational data to build our model. We use the natural logarithm value of our median housing prices (response variable; MVH) to have a smoother normal distribution of the data. Through our data analysis, we find two collinearities of 0.76 and 0.91 between variables of non-retail business acres (NRBA) and nitric oxide concentration (NOC) and between variables of the index of accessibility to highways (ATH) and property-tax rate (PTxR), respectively. Additionally, we obtain high p values for the corresponding variables of NRBA and houses built prior to 1940 (BPR). We decide to drop the variables for RLZ, NRBA, and BPR after obtaining the lowest Akaike's Information criterion (AIC_10) of the model missing these variables. The final model is log(MHV) ~ CR + CHR + NOC + ANR + DEC + ATH + PTxR + PTR + AA + LSP. We use the remaining 30% of the observational data to validate the model. We calculate the Mean Square of Residuals to be 0.322 and the Mean Squared Prediction Error to be 0.4717; since the two mentioned values are quite near to each other, we conclude that the model is valid.

A summary of the model

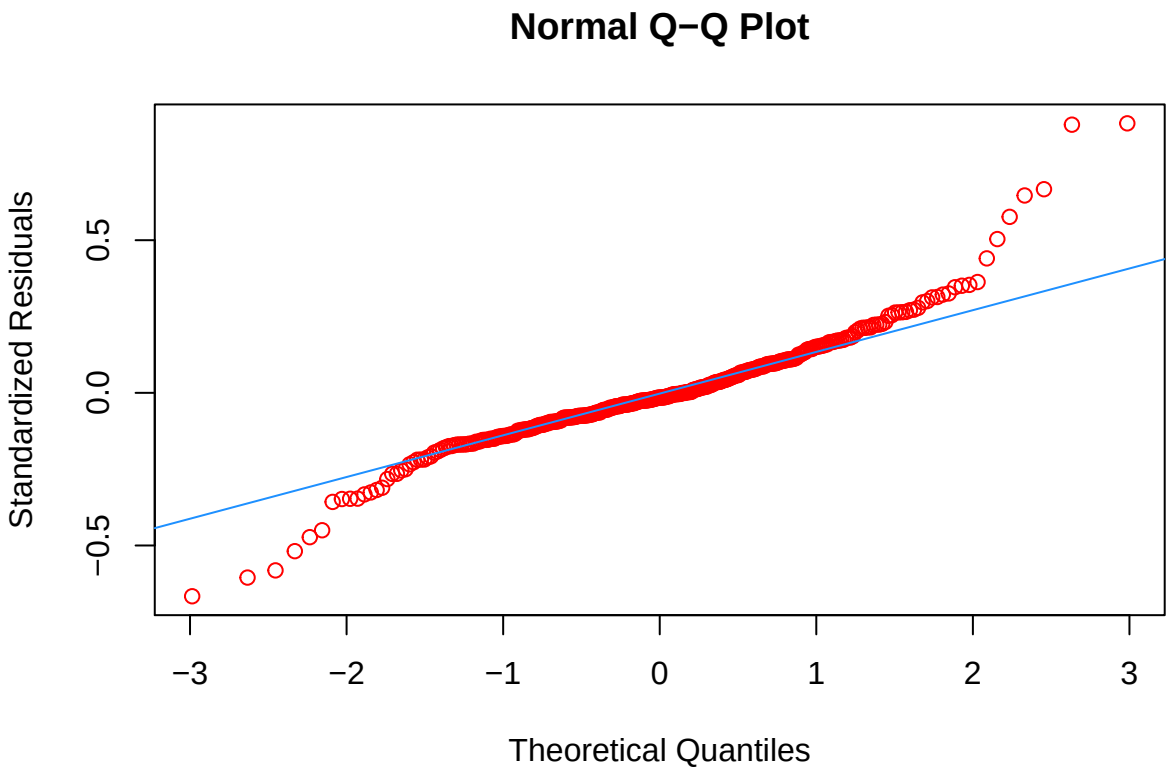
```
##
## Call:
## lm(formula = MVH ~ CR + CHR + NOC + ANR + DEC + ATH + PTxR +
##     PTR + AA + LSP, data = model_building)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66632 -0.09470 -0.01559  0.08961  0.88271
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5878433  0.2363374  15.181  < 2e-16 ***
## CR          -0.0111825  0.0014843  -7.534 4.41e-13 ***
## CHR          0.0617156  0.0401544   1.537 0.125226
## NOC         -0.5187985  0.1589590  -3.264 0.001210 **
## ANR          0.1356907  0.0195781   6.931 2.08e-11 ***
## DEC         -0.0296539  0.0071330  -4.157 4.07e-05 ***
## ATH          0.0126752  0.0030076   4.214 3.21e-05 ***
## PTxR        -0.0005920  0.0001554  -3.810 0.000165 ***
## PTR         -0.0380024  0.0056594  -6.715 7.82e-11 ***
## AA           0.0003876  0.0001260   3.075 0.002271 **
## LSP         -0.0231797  0.0022153 -10.464  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1837 on 343 degrees of freedom
## Multiple R-squared:  0.7908, Adjusted R-squared:  0.7847
## F-statistic: 129.7 on 10 and 343 DF,  p-value: < 2.2e-16
```

Model Diagnostics



Residuals vs Fitted

The residuals spread quite equally around the horizontal line without random patterns. However, we find a slight problem with the variance. There is huge variability in the range of the lower values of our response variable, while an extensively lower variability in the scope of the higher response values. Thus, the assumption of equal variance does not hold.

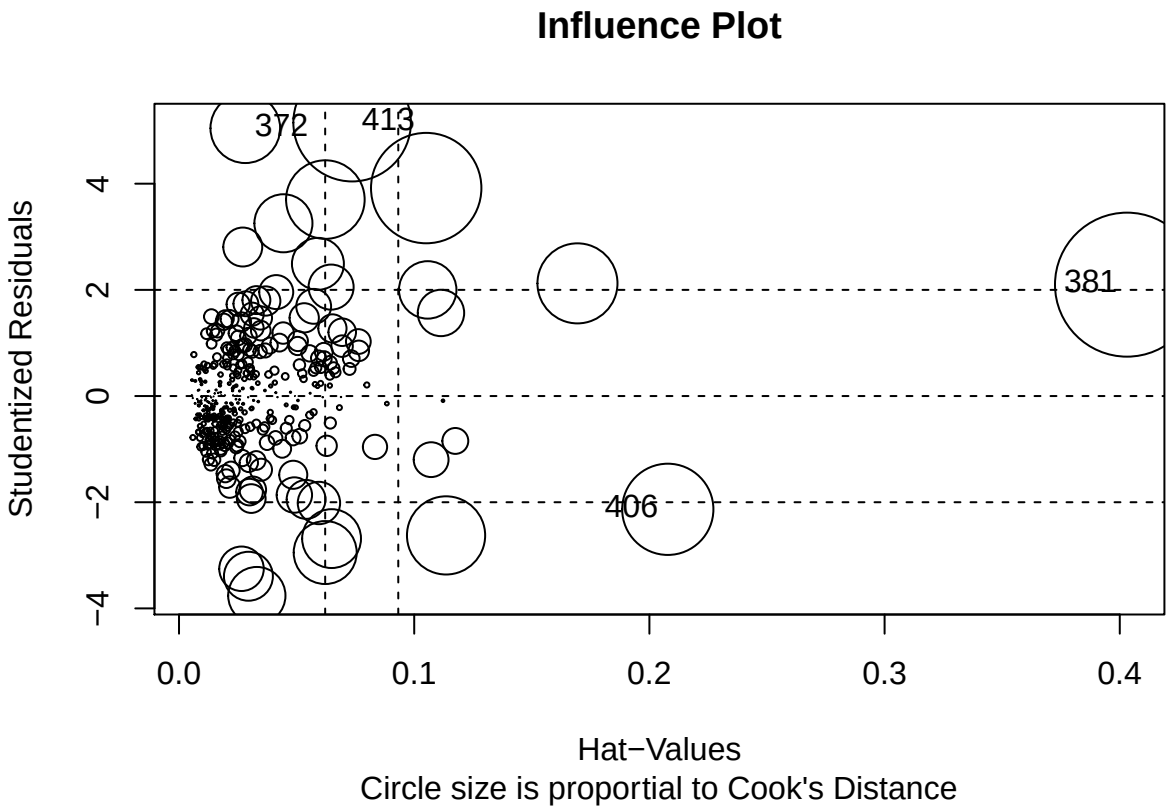


Normal Q-Q Plot

The Normal Q-Q plot represents the standardized residuals against the theoretical quantiles to check if the random errors are normally distributed. Since the graph shows a clutter near 0 with a straight line, we can use the normality assumption. However, there are some points which are outside the line - 413th, 372nd, 373rd, 369th, 405th and 428th observations.

Outlying Y Observations The ith observation is a Y outlier if the studentized residuals are larger than the studentized critical point, which is 3.925 in our case. We find the 413th, 372nd, 373rd and 369th observations to be outliers since their studentized residuals were higher than the threshold, studentized critical point.

Influence



The DFFITS test results show that the 413th, 372rd and 369th (outliers) have an absolute DFFITS value greater than 1. Addtionally, we calculate the Cook's distance for the 413th, 405th, 413th, 372nd, and 369th observations to be significantly smaller than the 10th percentile of Fisher distribution, F(0.2, 11, 495) = 0.484. Since our data set is relatively small, the DFFITS test and comparison of Cook's distance with the Fisher's 10th percentile help us conclude that these observations are not influential.

Conclusion

Our purpose is to build a model which can predict the median housing values from the available data. Through our study, we find a linear relationship of the median property values with crime rate, Charles river, nitric oxide concentration, average number of rooms, distances to employement centres, accessibility to highways, property-tax rate, pupil-teacher ratio, proportion of African Americans and lower status population. Considering buying a home is a long-term investment, our findings can extensively help homebuyers and insurance companies make appropriate decisions. Additionally, since there are 13 possible predictor variables, one of the limitations is to check every possible combination of the predictor variables to built a model due to the high number of combinations. A possible area of future study can be to extend the research to greater regions, and help homebuilders to choose a location.

References

Types of Zoning. Retrieved March 30, 2019, from
https://realestate.findlaw.com/land-use-laws/types-of-zoning.html

Non-Store-Retailing. (2018, February 6). Retrieved March 30, 2019, from
https://www.marketing91.com/non-store-retailing/

Statistics Canada (2015, Nov 27). Section 1: The Crime Severity Index. Retrieved from https://www150.statcan.gc.ca/n1/pub/85-004-x/2009001/part-partie1-eng.htm

Per Capita, Rates, and Comparisons. Retrieved March 30, 2019, from https://www.robertniles.com/stats/percap.shtml