STAC67 Report: Boston Housing Prices

Kenneth Leung

April 5, 2019

Abstract

In this case study, we are presented with a multitude of predictor variables that may account for changes in the housing prices of suburbs located in Boston. The dataset provided contains 506 observations on 13 predictors. The objective of this report is to investigate potential determining factors through research and statistical analysis using R to build a model to accurately predict Boston's suburb housing prices.

Background and Significance

In recent years, the development of real estate has become an important and essential part of economic growth. Buyers and sellers are constantly seeking for accurate valuations of housing property. We built a regression model by exploring the influence of several predictor variables, using historical data. Such data was obtained from a study conducted in 1978, comparing the relationship between housing prices in Boston and surrounding air quality (Harrison & Rubinfield, 1978). As a result, this model is an asset to home buyers who wish to purchase a home meeting their structural, demographic, accessibility, and environmental needs, alongside with validating its cost. Likewise, sellers are able to use this model to identify key attributes to legitimize the price level of different suburbs. The model will also aid in predicting housing market trends. The goal of this paper is to investigate which variables are the most significant and impactful to the housing prices in Boston and compiling them in a linear regression model.

Exploratory Data Analysis

The given data set contains 506 observation on 13 predictors that affect the house value which in this data is the 'Median value of owner-occupied homes' in thousands (MEDV).The 13 predictors are:

1. Per capita crime rate by town (x1) Crime rate is the proportion of criminal occurrences relative to the number of citizens in each town. The higher the crime rate, the less likely a potential buyer would purchase property in that area. The mean and the median of this variable is 3.613 and 0.2565 respectively.

2. Proportion of residential land zoned for lots over 25,000 sq. Ft. (x2) Residential land, measured in square feet, is the proportion of lots over 25,000 square feet for residential use over all zoned land over 25,000 square feet. In general the less dense areas show a trend of higher prices than the more dense areas. The mean and the median of this variable is 11.36 and 0.00 respectively.

3. Proportion of non-retail business acres per town (x3) Non-retail business land, measured in acres, is the proportion of non-retail business land relative to land size of the town. This includes real estate occupied for leisure, offices, healthcare, and residential use. The mean and the median of this variable is 11.14 and 9.69 respectively.

4. Charles River dummy variable (x4) This is a binary categorical variable indicating whether a suburb is in the general vicinity of the Charles River in Boston. The river has a history of being polluted and contaminated with bacteria and other chemicals (EPA.gov), potentially thwarting new buyers from purchasing property in this area. The mean and the median of this variable is 0.069 and 0.00 respectively.

5. Nitric oxide concentration (parts per 10 million) (x5) Nitric Oxide, also known as NO and is measured in parts per 10 mills, is a gas pollutant shown to cause many respiratory health conditions. This variable describes the quantity of airborne nitric oxide in the given suburb (Weinberger 2001). The higher the pollution, the poorer the living conditions. The mean and the median of this variable is 0.5547 and 0.5380 respectively.

6. Average number of rooms per dwelling (x6) Average number of rooms per dwelling, represents spaciousness and in a certain sense, quality of housing. Dwellings, that have more rooms will tend to have a higher cost. The mean and the median of this variable is 6.285 and 6.208 respectively.

7. Proportion of owner-occupied units built prior to 1940 (x7) This number refers to the number of houses that are still occupied, of which are built before the year 1940. Older property tent to have lower house values compared to newer ones. The mean and the median of this variable is 68.57 and 77.50 respectively.

8. Weighted distances to five Boston employment centers (x8) This variable measures the distances from the suburbs to five employment centers and then taking a weighted summary of those five distances. It just measures how close people in these suburbs are to where people may work. A buyer would prefer to buy property near there workplace. The mean and the median of this variable is 3.795 and 3.207 respectively.

9. Index of accessibility to radial highways (x9) This measures how close a property is too a highway. The more accessible the highway the more likely a buyer is to purchase the property. The mean and the median of this variable is 9.549 and 5.000 respectively.

10. Full-value property-tax rate per 10,000 (x10) This variable measures the town's property tax rate. Higher property taxes may mean, for instance that the school in an area are better because public schools are often funded by property taxes. The mean and the median of this variable is 408.2 and 330.0 respectively.

11. Pupil-teacher ratio by town (x11) This ratio is the number of students who attend school divided by the number of teachers in the town. If the ratio is high, then the connection between students and teachers is relatively weak. Generally speaking, low ratio represents good education condition. The mean and the median of this variable is 18.46 and 19.05 respectively.

12. 1000(B - 0:63) where B is the proportion of African Americans by town (x12) It is the proportion of the African Americans in the town. In the United States of America, African American populated neighbourhoods tend to be valued less than other neighbourhoods (NextCity). The mean and the median of this variable is 356.67 and 391.44 respectively.

13. A numeric vector of percentage values of lower status population (x13) This variable describes the percentage of the population in a given suburb that are of lower economic status. The mean and the median of this variable is 12.65 and 11.36 respectively.

Model

Model Selection

For our model, we decided to split the data into a training set of size 300 and a validation set of size 206. To get a sense of collinearity among our independent variables, we produced a correlation matrix using `cor(train)`. We discovered a couple of large values (≥ 0.7), which may indicate multicollinearity. Therefore, we used `stepAIC` to reduce our variable space.

```r
fit1 <- lm(y ~ ., data = train)
step = stepAIC(fit1, direction = "both")
```

```
## Start:  AIC=707.59
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + 
##     x12 + x13
## 
##        Df Sum of Sq    RSS     AIC
## - x3    1       2.7 2892.8  705.86
## - x4    1       8.2 2898.4  706.44
## - x9    1      15.7 2905.9  707.21
## - x2    1      18.0 2908.2  707.45
## <none>              2890.2  707.59
## - x13   1      44.1 2934.3  710.13
## - x5    1      46.8 2937.0  710.41
## - x1    1      60.6 2950.8  711.81
## - x12   1     113.2 3003.3  717.11
## - x10   1     173.7 3063.8  723.09
## - x7    1     197.7 3087.8  725.43
## - x8    1     331.3 3221.4  738.14
## - x11   1     374.7 3264.8  742.15
## - x6    1    4919.6 7809.7 1003.80
##
## Step:  AIC=705.86
## y ~ x1 + x2 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 +
##     x13
##
##        Df Sum of Sq    RSS     AIC
## - x4    1       9.5 2902.3  704.84
## - x9    1      13.8 2906.6  705.29
## - x2    1      16.5 2909.4  705.57
## <none>              2892.8  705.86
## + x3    1       2.7 2890.2  707.59
## - x13   1      42.8 2935.6  708.27
## - x5    1      44.2 2937.0  708.41
## - x1    1      62.9 2955.7  710.31
## - x12   1     113.5 3006.3  715.41
## - x10   1     177.1 3069.9  721.69
## - x7    1     201.2 3094.0  724.03
## - x8    1     344.9 3237.8  737.66
## - x11   1     373.5 3266.3  740.29
## - x6    1    4934.5 7827.3 1002.48
##
## Step:  AIC=704.84
## y ~ x1 + x2 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13
##
##        Df Sum of Sq    RSS     AIC
## - x9    1      16.8 2919.1  704.58
## - x2    1      17.5 2919.7  704.64
## <none>              2902.3  704.84
## + x4    1       9.5 2892.8  705.86
## + x3    1       3.9 2898.4  706.44
## - x13   1      41.2 2943.5  707.07
## - x5    1      45.9 2948.2  707.55
## - x1    1      66.0 2968.3  709.59
## - x12   1     114.0 3016.3  714.40
## - x10   1     190.0 3092.3  721.87
## - x7    1     199.9 3102.2  722.83
## - x8    1     360.8 3263.0  737.99
## - x11   1     386.3 3288.6  740.33
## - x6    1    4936.8 7839.1 1000.93
##
## Step:  AIC=704.58
## y ~ x1 + x2 + x5 + x6 + x7 + x8 + x10 + x11 + x12 + x13
##
##         Df Sum of Sq    RSS     AIC
## - x2    1       8.7 2927.9  703.48
## <none>              2919.1  704.58
## + x9    1      16.8 2902.3  704.84
## + x4    1      12.5 2906.6  705.29
## + x3    1       1.3 2917.8  706.44
## - x5    1      43.2 2962.3  706.99
## - x13   1      43.8 2962.9  707.04
## - x1    1      65.6 2984.7  709.25
## - x12   1     115.4 3034.5  714.21
## - x10   1     173.8 3092.9  719.93
## - x7    1     193.9 3113.0  721.87
## - x8    1     344.5 3263.7  736.05
## - x11   1     404.5 3323.6  741.51
## - x6    1    5167.0 8086.2 1008.24
##
## Step:  AIC=703.48
## y ~ x1 + x5 + x6 + x7 + x8 + x10 + x11 + x12 + x13
##
##         Df Sum of Sq    RSS     AIC
## <none>              2927.9  703.48
## + x4    1      12.4 2915.5  704.20
## + x2    1       8.7 2919.1  704.58
## + x9    1       8.1 2919.7  704.64
## + x3    1       0.9 2927.0  705.38
## - x13   1      40.9 2968.8  705.64
## - x5    1      43.6 2971.5  705.91
## - x1    1      62.7 2990.6  707.83
## - x12   1     114.6 3042.4  712.99
## - x10   1     165.6 3093.5  717.99
## - x7    1     216.1 3144.0  722.84
## - x8    1     356.2 3284.1  735.92
## - x11   1     499.4 3427.3  748.73
## - x6    1    5437.3 8365.2 1016.41
```

The model with the lowest AIC value was chosen, which took the following predictor variables:

```
Step:  AIC=703.48
y ~ x1 + x5 + x6 + x7 + x8 + x10 + x11 + x12 + x13
```

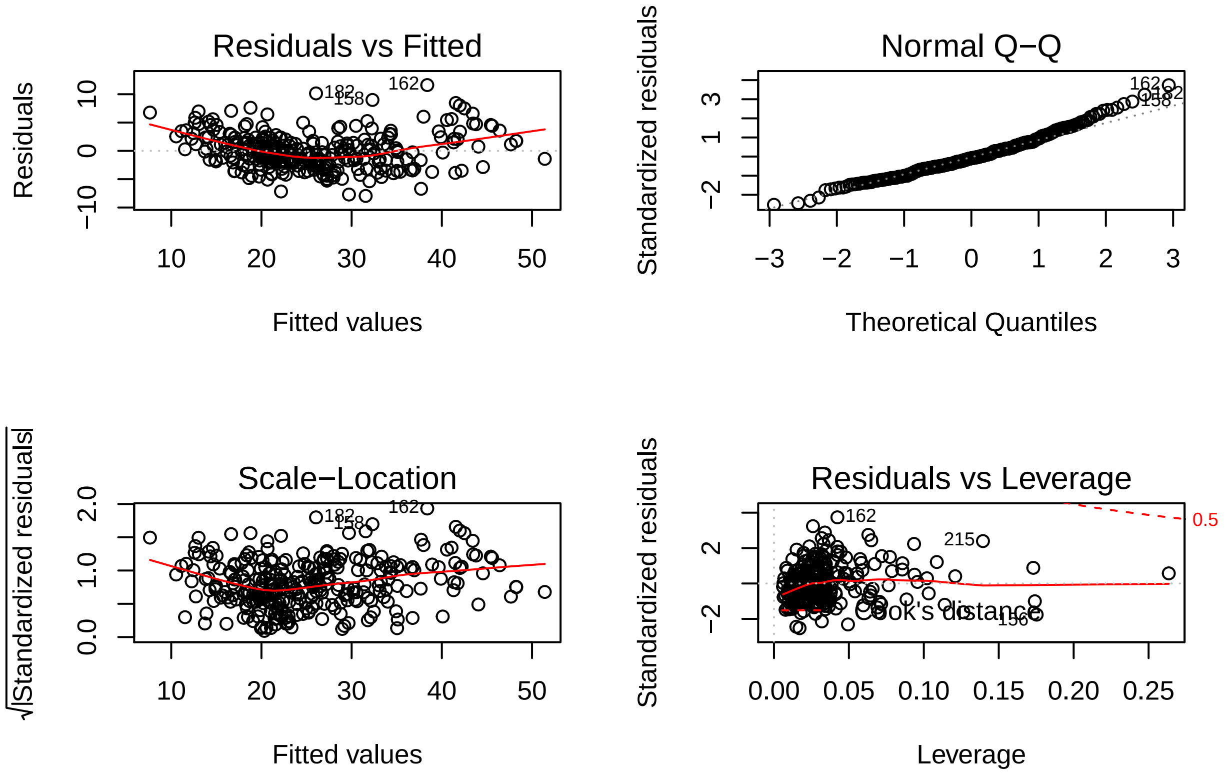
The variables that were removed were:

- x2 = Proportion of residential land
- x3 = Proportion of non-retail business acres
- x4 = Charles River dummy variable
- x9 = Accessibility to radial highways

A logical reasoning for the removal of these variables would be their mutual relation to Boston's industrialization at the time. Residential land and business land could be related to the abundance of factories in the area. Charles River had a history of pollution from artificial chemicals produced. Factories would have to be close to highways to be effective.

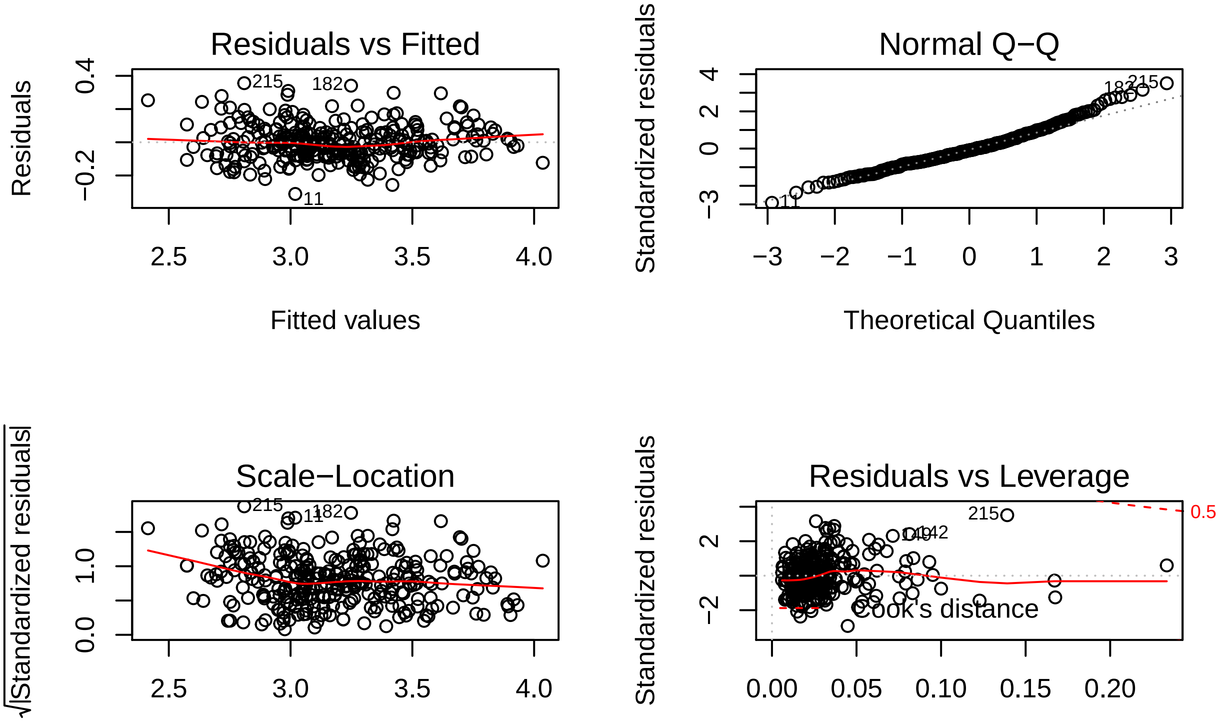
To test our model, we then plotted the residual vs fitted, normal Q-Q, scale-location, and residual vs leverage plots. Unfortunately, both the residual vs fitted plot and the scale-location plot had a curvature pattern, indicating potential non-linearity and heteroscedasticity. The other plots however, were fine though we did have a few concerns regarding outliers that we discuss later. The correlation matrix was also examined using `cor(f2)` and the model still had large correlation among variables.

```r
aicfit <- lm(y ~ x1 + x5 + x6 + x7 + x8 + x10 + x11 + x12 + x13, data = train)
aicfitvar <- c("x1","x5","x6","x7","x8","x10","x11","x12","x13")
f2 = train[aicfitvar]
```



In an attempt to solve our problems, we applied the natural log to our response variable ($\log(Y_i)$) and fitted another model with all of the original variables present. We then ran `stepAIC()` again, and obtained a new model with the lowest AIC with the variables $X1, X2, X3, X4, X9$ absent.

The plots of the log model are shown below.



The residual vs fitted values improved drastically as there was no longer a curvature pattern. However, the problem with the scale-location plot remained, and high correlation was still present in the `cor(f3)` matrix.

We then used model tests to further narrow down our variable space.

```
##        SSres  Rsq Rsq_adj      C      AIC
##   [1,]  3.44 0.89    0.89   9.12 -1320.93
##   [2,]  3.47 0.89    0.89  10.34 -1319.61
##   [3,]  3.55 0.89    0.88  15.07 -1314.78
##   [4,]  3.84 0.88    0.87  37.38 -1293.33
##   [5,]  8.20 0.74    0.73 406.53 -1063.74
##   [6,]  3.76 0.88    0.88  32.86 -1297.44
##   [7,]  4.03 0.87    0.87  55.31 -1276.90
##   [8,]  3.68 0.88    0.88  25.62 -1304.37
##   [9,]  4.41 0.86    0.85  87.22 -1249.94
##  [10,]  3.66 0.88    0.88  24.15 -1305.80
##  [11,]  3.94 0.87    0.87  45.73 -1285.68
##  [12,]  4.05 0.87    0.87  55.10 -1277.32
##  [13,]  3.86 0.88    0.87  38.86 -1291.96
##  [14,]  4.44 0.86    0.85  87.89 -1249.79
##  [15,]  3.84 0.88    0.87  37.38 -1293.33
##  [16,]  4.02 0.87    0.87  52.11 -1279.96
```

We tested the top 3 models given in the output. The model we chose to use was Model [3] because it was one of the contenders for the best AIC and R^2_{adj}. It also had the lowest VIF values and mean VIF (1.911) compared to the other ones. There was a minor discrepancy in R^2 by switching to Model [3], which was acceptable. All in all, it was the best combination between the VIF, AIC, and R^2_{adj} values. The resulting model was the same model from before with the exemption of variable $X5$.

A logical reasoning for taking out variable X5 (nitrous oxide concentration) would be its relation with industrialization. Factories would produce large quantities of NO, which means that the other predictor variables would have already provided an explanation for the data, rendering X5 redundant. This is also why R^2_{adj} did not go down as much.

However, this model still had the problem of potential heteroscedasticity shown by the plots (plots omitted). We also attempted to center our predictor variables to their respective means, but it had negligible effects. This is the most we could do.

Our final model is the model listed below:

```
fit5 <- lm(log(train$y) ~ x6 + x7 + x8 + x10 + x11 + x12 + x13, data = train )
```

Model Validation

To validate our model, we took our 206 entries from the original split and tested it against our model. The $MSPE$ and the $MSRES$ are given below, respectively.

```
## [1] 180.3428
```

```
## [1] 10.26968
```

This is good news, as $MSPE$ and $MSRES$ are both very close to each other, and both very low, indicating that the model constructed from our training set was accurate in predicting the values of the validation set.

Model Diagnostics

Functional Form

To check whether the functional form of the model was adequate, we plotted the residuals against the selected explanatory variables separately (plots omitted). Our analyses showed that all explanatory variables depicted randomness in the residuals except for x12. We attempted to find a suitable transofmration such as the inverse and log transformations on x12 and y to gain a better distribution of randomness of the residuals for x12, however there were no significant improvements to the residuals. Due to the laack and imporvement and significance of x12, we concluded that our model was of a proper functional form.

Outlying Y Values

We investigated the possibility for outlying y observations by investigating the significance of the Studentized deleted residuals. Thereafter testing the residuals, there were no outlying y observations.

```
## 162
## 162
```

Outlying X Test

A leverage value P_{ii} is considered to be large according to the guideline:

- $P_{ii} > 0.5$

We also investigated the leverage points in our model by evaluating $P_{i,i}$, measuring the distance between the X values of observations and the center of the x-space. The conclusion of testing $P_{i,i}$ suggested there were no observations far away from the mean of x values, under the guidline of $P_{i,i}$>0.5.

```
##   9  33  35  49 103 127 142 145 146 147 148 156 157 215 226 254 258
##   9  33  35  49 103 127 142 145 146 147 148 156 157 215 226 254 258
```

```
## named integer(0)
```

Two main guidelines of determining influence points which are the following:

- DFFITS
- Cook's Distance

We found observation 215 to be an influential point with a DFFITS 1.47, indicating it is influential in our model. Furthermore, we investigated the level of infleunce by measuring its Cook's Distance. The Cook Distance's value is 0.259 indicating the 215th observation has weak influence on the fitted values because its value is less than the 20th percentile of F(8,292). Therefore we decided to keep this obseration in our model.

Below are the results of the three tests, respectively.

```
## named integer(0)
```

```
## named integer(0)
```

Once again, some tests indicate influence points and some fail to reveal influence points. The existence of "true" outliers cannot be affirmed.

Multicollinearity

The formal method of testing for multicollinearity is the VIF test. We tested multiple models and `fit5` obtained the best score. Below is the output of the VIF test and the mean VIF, respectively.

```
##       x6       x7       x8      x10      x11      x12      x13
## 2.314536 2.615516 1.943063 1.197616 1.152337 1.160797 2.991531
```

```
## [1] 1.910771
```

Each individual VIF value as well as the mean VIF are all sufficiently low (≤ 3). Compared to the mean VIF of the model with all variables (2.59), the current model `fit5` has better performance. This means that this model handles multicollinearity decently well.

Discussion/Conclusion

The goal of the study was to construct an accurate model for predicting housing prices of suburbs given demographic, environmental, and accessibility information of the suburb. Our final model has the log transformed price as the response variable and the following variables as the predictors:

- average number of rooms per dwelling
- proportion of owner-occupied units built prior to 1940
- weighted distances to five Boston employment centers
- full-value property-tax rate per 10,000
- pupil-teacher ratio by town
- proportion of African Americans
- percentage values of lower status population

Although our model performed well on our validation set with low errors, the problem of heteroscedasticity was still present. This means that our model may not be accurate in predicting prices from certain ranges of data. Another problem might be the potential "outliers"; one guideline showed no outliers while the other showed a few. However, without knowing more information about the specific outlying suburbs, we cannot prematurely remove any data. Lastly, one other problem might be the context of the data. The source of the data was collected from a specific location (Boston) at a specific time (1970s). Historical data is not always indicative of future outcomes, especially when external factors that we discovered, such as industrialization, are present.

However, our model is an appropriate medium of finding relationships between demographics, environment, and housing prices, and will nevertheless serve to be an asset to potential buyers and sellers and providing housing price insight to both parties.

References

[1] About the Charles River. (2018, April 26). Retrieved from https://www.epa.gov/charlesriver/about-charles-river

[2] Harrison, D., Jr., & Rubinfield, D. L. (2004, July 28). Hedonic housing prices and the demand for clean air. Retrieved from https://www.sciencedirect.com/science/article/pii/0095069678900062

[3] Weinberger, Barry, Laskin, E., D., Laskin, & D., J. (2001, January 01). Toxicology of Inhaled Nitric Oxide. Retrieved from https://academic.oup.com/toxsci/article/59/1/5/1658774

[4] Why Black Neighborhoods Are Valued Less Than Other Neighborhoods. (n.d.). Retrieved from https://nextcity.org/daily/entry/why-black-neighborhoods-are-valued-less-than-other-neighborhoods