**A statistical model for predicting housing price in Boston**

Duy Truong

MCS 242 Statistical Method- Traditional

1. **Introduction:**

Housing pricing has always been one of the most stressful things for families. Because of its high cost, one would need to deliberately measure not just the quality of the house itself, but its surrounding factors as well. Things such as the crime rate within the city, the distance to the center of the city… intuitively could contribute to the changes in the house pricing. However, mere guesses will not sufficiently convince a household to jump right in and purchase a house, as there are numerous variables that one could argue to be important factors when deciding house price. In order to accurately capture the essential variables contributing greatly to house price, a statistical model is mandatory. In 1978, Harrison,D and Rubinfeld came up with a housing price model to measure people’s willingness to pay for air quality. How accurate people’ willingness depends on how well the housing model is. This research paper hopes to support their guidelines for households choosing to buy a house in Boston. To further improve the quality of the research paper, the author also used another article called “A dymimic model of housing price determination” by Robert F. ENGLE, David M. LILIEN and Mark WATSON as reference for the model.

In this paper, we want to predict the housing price depending on couples of variables in Boston. Our first research question is:

‘How can we predict Boston’s housing price and with what variables?’

In the original dataset, there are a total of 14 variables, and so our hypothesis is to explore whether at least one of those variables has a strong correlation with the housing price.

1. **Materials and methods**

The data is from U.S Census Service relating to housing in the area of Boston Mass. StatLib archive contains this dataset and the link can be found in the reference section of this paper. The data was first published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978. This dataset has been used at great length as a benchmark algorithm.

The original dataset contains 506 cases with a total of 55 missing variables. 1 missing value is from the tax variable and 54 values are from the ratio variable between student and teacher. There is a total of 14 variables in the dataset. Our response variable is the median value of owner-occupied homes in 1000$.

By examining carefully with hypothesis testing (testing correlation between each variable and our response variable), 10 explanatory variables for our models are chosen, with a focus on 2 variables, room number and the proportion of lower status population. The descriptions of the chosen variables can be summarized in the Table 1.

**Table 1. Variable descriptions for the model**

|  |  |
| --- | --- |
| Variable | Description |
| MEDV | The median value of owner-occupied homes in 1000$.( Our response variable) |
| RM | Average number of room in 1 unit |
| AGE | Proportion of owner unit built prior to 1940 |
| LSTAT | Proportion of population that is lower status(adults without highschool education and male laborers) |
| CRIM | Crime rate by town |
| ZN | Proportion of a town’s residential land zoned for lots greater than 25000 square feet (zone restricts small house construction) |
| TAX | Property tax in 10000$, measuring the cost of public services |
| PTRATIO | Pupil-teacher ratio by school district |
| DIS | Distance to 5 employment centers in Boston region |
| RAD | Index of accessibility to radial highway |
| CHAS | Charles River dummy variables; 1 if tract is bounded by the river, 0 therwise. |

55 NA data, including 1 from the tax variable and 54 from the ratio between student and teacher variable, after careful consideration, will be omitted in the analysis. During the research process, the author considered dropping the ratio variable, but after extensive analyzing, it was concluded that the ratio variable was too important to drop out of the analysis. Therefore, omitting missing variable must be implemented so the ratio variable can be used inside the model.

Cook’s distance method is used on the final model to identify outliers. A total of 11 outliers are spotted and will be removed from the dataset, with a remaining of 440 observations.

The variable RAD, or the index of accessibility to highway in our model, will be used under a log transformation. The reason is that a higher index number has more influence over the price of housing comparing to a lower index number. Same reasoning and transformation technique are used for the ratio of pupil teacher. Also, the original dataset counts TAX as string variable, so a modified variable of TAX, called taxModified, was introduced with integer element.

Now we can get the summary statistics of our new data.

**Table 2. Variable means and standard deviations**

|  |  |  |
| --- | --- | --- |
| Variable | Mean | Standard Deviation |
| MEDV | 23.53 | 8.43 |
| taxModified | 373.37 | 148.63 |
| RM | 6.35 | 0.64 |
| CRIM | 1.35 | 2.44 |
| ZN | 13.04 | 24.57 |
| CHAS | 0.07 | 0.25 |
| AGE | 64.99 | 28.02 |
| DIS | 4.09 | 2.08 |
| LSAT | 11.37 | 5.94 |
| PTRATIO | 18.22 | 2.2 |

Multiple Regression model is used to measure the relationship between the response variable, median income, with our 10 explanatory variables. It is concluded that MRM is the best model because the response variable is a continuous quantitative variable. As mentioned previously, cook’s distance method is used to cut off outliers, thereby improving the accuracy of the prediction outcome. Because the relationships of our explanatory variables can be either positive or negative to our response variable, the P-values for the explanatory variables are two-sided.

To assess the condition of the model, which needs to satisfy the following: linear assumption, multivariate normality, no multicollinearity, and homoscedasticity. For the linear assumption condition, 2 most important variables are measured with the response variable to check for linearity in the model under scatterplot table. VIF method is used to check for multicollinearity of the model. As for the multivariate normality condition, QQ-plot is used to check for linearity of the residual. Lastly, a residual vs fitted plot is used to check for homoscedasticity.

**III. Result**

*The model*

Y= B0 + B1\*CRIM + B2\*CHAS +B3\*ZN +B4\*RM +B5\*AGE +B6\*DIS +B7\*log(RAD) +B8\*taxModified + B9\*PTRATIO +B10\*log(LSTAT) +e

The model shows that all the independent variable coefficients are highly significant at 5% significance level, with an exception of the small house restriction proportion variable sitting at 0.1 level

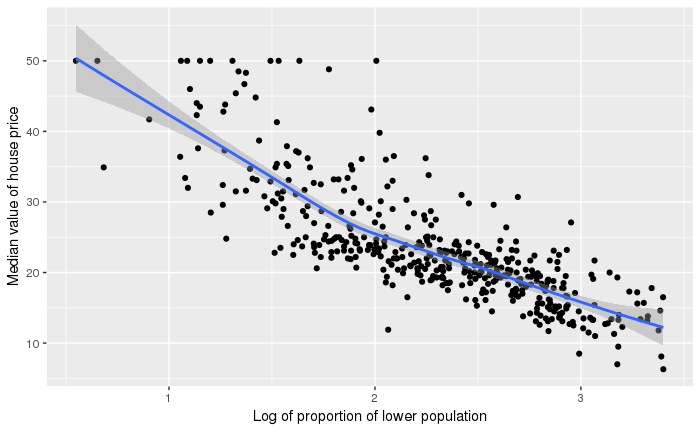
Crime rate, age, distance to the center Boston, tax, the log ratio of pupil-teacher, and the proportion of population of lower status have negative coefficients. This makes sense since all these variables intuitively have negative influence over the house pricing. Whereas the proportion of town’s residential land zoned over 25000 feets, the average rooms per unit, and the log of index access to radial highway have positive coefficients. Again, this is understandable, as more rooms mean bigger houses, bigger houses tend to be at land zoned over 25000 feet, and higher access to radial highway means more expensive housing. The details of the coefficients and their standard errors can be summarized in the table below:

**Table 3. The coefficients and standard errors of the variables**

|  |  |  |
| --- | --- | --- |
|  | Estimate | Std. Error |
| (Intercept) | 7.881201 | 3.64681 |
| CRIM | -0.62503 | 0.121842 |
| CHAS | 1.447673 | 0.60801 |
| ZN | 0.016078 | 0.009383 |
| RM | 6.768479 | 0.380314 |
| AGE | -0.02332 | 0.009296 |
| DIS | -0.77717 | 0.125182 |
| log(RAD) | 1.751951 | 0.351346 |
| taxModified | -0.00807 | 0.002038 |
| PTRATIO | -0.6422 | 0.08232 |
| log(LSTAT) | -4.54171 | 0.541248 |

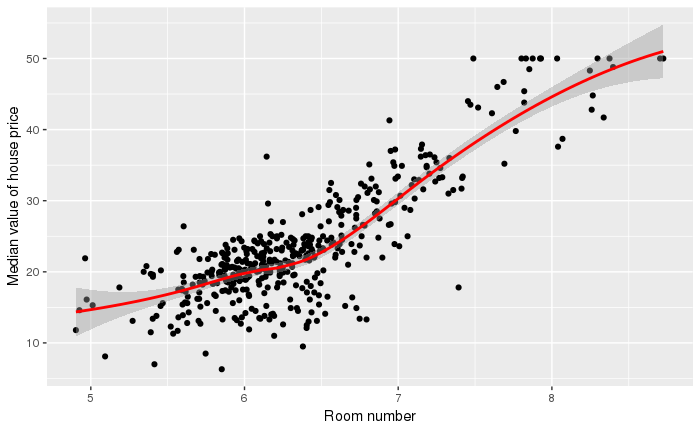
The model has proven that the LSTAT variable, which stands for the proportion of lower status is one of the most correlated with the median house price. Figure 1 showed the relationship between median value and log of the percentage of lower status. For every 1 percent increase in the proportion of the lower population, the price of housing will drop by 4541,707 USD

**Figure 1. Scatterplot of house price and proportion of lower education**



The second highest correlated variable with the response variable is the RM, or the room number of each house. Figure 2 shows a linear relationship between the 2 variables. For every 1 room increase inside a house, the price of that house is expected to increase by 6768,479 USD

**Figure 2. Scatterplot of house price and the room number**



Regarding the conditions of the model, the model fulfills all necessary conditions (See the graphs in the appendix).

With the result given, we can answer our hypothesis question. Our hypothesis:

Ho: B4= B10=0

Ha: At least one is different from 0

With a p-value of 2.2e-16, we reject the null hypothesis saying that all coefficients are equal to 0 at 5% significance level.

**4. Discussion**

The research captures some of the aspects that could affect housing decision in Boston, primarily how higher percentage of the lower population and a lower room number per house could lessen the housing price. Overall, we reject the null hypothesis (there is no relationship between housing price with the number of room and the proportion of lower population) and conclude that housing price is based primarily on its neighborhood and its number of rooms. When purchasing a new house, a household would want to place security on top of their checklist, and so the house sellers respond accordingly by measuring how educated the people in the area are and then adjusting the price. Another criterion, especially for household with children, is that they would want enough rooms for their children. Also, houses that have more rooms are likely to be more expensive, considering the costs of building the house and the cost of the land.

Comparing to the model by Harrison and Rubinfeld, although the model is less complex, the final result is very reasonable. An R-squared of 85%, F-statistics of 273, with all statistically significant variables that satisfy all the condition of an ordinary least squared model, prove that the model is usable.

The model will prove to household what important qualities to look at when buying a house, which are the room number and the proportion of lower population.

However, the model is not perfect. By looking at the rows containing missing variables, unusual trends are spotted, as for each missing variable of the PTRATIO variable correspond to constant values found in these following variables: ZN, INDUS, CRIM, DIS, and RAD. Possible confounding variable, such as rating of housing could change people’s perception on whether the quality of a house is satisfactory. Also, only limited variables are given in the original dataset, limiting the possible improvement for the model. On top of that, the data was collected in 1970, making data searching more difficult. Lastly, the sample size is relatively small, with only 500 observations. Although the model can be used to predict Boston housing price, it is not advised to use it to predict all house in Massachusetts. Such prediction requires larger sampling method. Lastly, because the dataset was from 1970, there could be changes with the explanatory variables. For example, property tax law could undergo significant changes that potentially change the coefficient from being negative to positive.

Overall, although there are flaws with the model, the result could still be used as a reference for similar research projects, specifically one with predicting house price within a city.

Reference

Engle, Robert F., et al. “A Dymimic Model of Housing Price Determination.” *Journal of Econometrics*, vol. 28, no. 3, 1985, pp. 307–326., doi:10.1016/0304-4076(85)90003-x.

Harrison, D. and Rubinfeld, D.L. `*Hedonic prices and the demand for clean air*', J. Environ. Economics & Management, vol.5, 81-102, 1978.