**FINAL EMPIRICAL PAPER**

Authors: Khoa N. Tran and David R. Jacques

Instructor: Nayoung Lee PhD.

Course: ECON 8010 Econometrics

Date: 12/04/2020

A Hedonic Pricing Model for Midcentury Boston Housing

**ABSTRACT:**

This empirical paper attempts to construct a hedonic pricing model for the housing market in Boston city during the 1970s using econometrics methods. The model will consider factors regarding structure, neighborhood, accessibility, and air pollution, which may have impacts on housing prices. Additionally, the fit and shortcomings of the model will be discussed in length. In the end, the paper will show how the model can be used to interpret the influences of the factors on the housing prices and to predict housing prices.

**TABLE OF CONTENTS PAGE**

1. Introduction 3
2. About the Dataset 3
3. Model Specification 4
4. Addressing Endogeneity and Heteroskedasticity 7
5. Conclusion 8
6. Works Cited 9
7. **Introduction:**

Housing is a critical part of everyone’s life. It is where we go home every night. It is where we feel most safe. It affects where, and often how, our kids grow up. Homes are often the largest asset of families and are a key to building generational wealth. With the vast importance of housing in our lives, we wanted to take a deeper look at the factors that affect housing prices, which can exclude so many from building home equity. Initially, our plan was to evaluate the importance of housing stability in obtaining advanced degrees, but after gathering data from IPUMS, it accumulated to a size of several gigabytes that our computers were not able to effectively use, so we decided to take a broader look at what makes housing so valuable in some areas, and not-so-valuable in others.

Our project expands on the work done by David Harrison and Daniel Rubenfeld in their flagship 1979 publication *Hedonic Housing Prices and the Demand for Clean Air.* We took inspiration from their work and sought to see if we could make a better fitting model by taking into account different economic reasoning and incorporating the concept of diminishing returns.

1. **About the Dataset:**

Our data set was downloaded from Kaggle.com and sourced primarily from the 1970 United States tract. Additionally, we made slight adjustments to the variables. We removed the PART variable, which was redundant to the NOX variable listed below. Additionally, we used a revised B variable in lieu of the original BLK variable, to better account for the lasting impact of housing segregation. Our data consists of 506 complete observations of these 17 variables. You will find our variables below:

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Source (Original) |
| Dependent Variable | | |
| MEDV | Median value of owner-occupied homes in $1000's | 1970 U.S. Census |
| Structural Variable | | |
| RM | Average number of rooms per dwelling | 1970 U.S. Census |
| AGE | Proportion of owner-occupied units built prior to 1940 | 1970 U.S. Census |
| Neighborhood Variables | | |
| B | =1000\*(Bk - 0.63)^2 where Bk is the proportion of blacks by town | 1970 U.S. Census |
| LSTAT | % lower status of the population | 1970 U.S. Census |
| CRIM | Crime rate per capita by town | FBI (1970) |
| ZN | Proportion of residential land zoned for lots over 25,000 sqft | Metropolitan Area Planning Commission (1972) |
| INDUS | Proportion of non-retail business acres per town | Vogt, Ivers, and Associates [33] |
| TAX | Full-value property-tax rate per $10,000 | Massachusetts Tax- payers Foundation (1970) |
| PTRATIO | Pupil-teacher ratio by town | Massachusetts Dept. of Education (1971-1972) |
| CHAS | = 1 if tract bounds Charles river; 0 otherwise | 1970 U.S. Census Tract maps |
| Accessibility Variables | | |
| DIS | Weighted distances to five Boston employment centers | Srhnare [29] |
| RAD | Index of accessibility to radial highways | MIT Boston Project |
| Air Pollution Variables | | |
| NOX | Nitric oxides concentration (parts per 10 million) | TASSIM |

1. **Model Specification:**

We began our project with a simple linear regression of these variables:

OLS1 <- MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT

Almost all variables were considered significant with a R-squared value of .617. INDUS was not considered significant at any level. We then decided to run a logarithmic model, and received similar significance levels, and an R-squared value of .6846.

* lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT

Once again, INDUS was not statistically significant. Because of the limited impact on the significance of our variables and the higher R-squared value, we decided to proceed with the log model since it is a better fit for the data at hand.

Once we had our model determined, we decided to make some further specifications to the data originally used by Harrison and Rubenfeld. As mentioned earlier, we made changes to the variable used to measure the proportion of Black residence in a survey area. The relationship between Blk and housing values is parabolic. The intuition is that at low to moderate levels of Blk, an increase in Blk should have a negative influence on housing value if Blacks are regarded as undesirable neighbors by Whites. However, market discrimination means that housing values higher at the very high levels of Blk. Thus, in this dataset, one expects a linear relationship between B and MEDV. The logarithmic specification of LSTAT implies that socioeconomic status distinctions mean more in the upper brackets of society than in the lower classes. DIS and RAD variables are entered in logarithmic form in adherence to the traditional theory of urban land rent gradients. Lastly, we included a quadratic term (RM^2) to capture the parabolic relationship between the number of rooms in a home and the housing price. Our reasoning behind including this parabolic term in our model is the understanding that the price of a house is positively correlated to the price to a point, but an especially large number of rooms could be seen as burdensome, as most people’s families are not that large. With these new specifications, we form the following model:

* lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT)

In this model, we see all our specifications deemed statistically significant, and our R-squared value increased to .7372, signaling that this model fits the data better than our previous iterations.

Next, we wanted to focus on specifying our NOX variable in particular. We suspect that this variable would follow a linear relationship with our dependent variable, housing price. Based on economic reasoning and intuition, at low levels of NOX, the marginal change in NOX level should only have a mild effect on housing price, since it is unlikely to have significant health effects. For medium levels of NOX, we suspect higher marginal effects, as slight increases in NOX concentration could become hazardous and decrease demand for a dwelling. At high levels of NOX, we expect the marginal effect to decrease since many of these areas were not originally suited for inhabitance in the first place. So we suspect that NOX would affect housing price through the second and third power of NOX, and as a result, will compare 3 different models: (NOX, up to NOX2 included, and up to NOX3 included). These additional models are listed below.

* lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+I(NOX^2)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT)
* lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+I(NOX^2)+I(NOX^3)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT)

Comparing these three models reveals that the model including up to the third power of NOX is the best fitting model based on the high adjusted R-squared (.7419 vs the .7381 of the NOX2 model). Both models’ variables are statistically significant, with the exception of INDUS.

One thing of interest across all regressions that we ran is that our INDUS variable never once registered as statistically significant based on its p-value. This suggests that the number of nonretail business acreage in a community does not impact the price of the surrounding homes. Because of this, we decided to relax this control from our model and observe if there is any change in our estimates.

* Table

  Description automatically generatedlm(MEDV\_log~CRIM+ZN+CHAS+NOX+I(NOX^2)+I(NOX^3)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT)

As you can see in the comparison to the right, relaxing our INDUS control had little to no effect on the remaining independent variables and had no impact at all on the goodness of fit of our model. For these reasons, we decided to relax the INDUS variable from our final model.

1. **Addressing Endogeneity and Heteroskedasticity:**

After having specified our model, we had to consider the issues of endogeneity and heteroskedasticity in our model. Regarding endogeneity, our model assumes all independent variables are exogenous. To the best of our knowledge and data, we can conclude that there is no endogeneity in our model. Regarding heteroskedasticity, since our data is from the census tract and not based on individual observations, we suspect that there may be some degree of heteroskedasticity. Thus, we perform the Breusch-Pagan test and White’s test.

When performing the Breusch-Pagan test, we received a p-value less than 0.01, suggesting that the error is likely heteroskedastic. We confirm this using White’s test, which produced similar results. Since the heteroskedasticity was detectable by the Breusch-Pagan test, there is likely a linear relationship. To resolve these issues, we introduced the robust standard error and the FWLS/FGLS method. After implementing these methods, we see that the issue of heteroskedasticity had seriously affected our model. Taking these issues into consideration, we were no longer able to claim significance for AGE (due to the robust standard error), ZN (due to FWLS), and I(NOX^3) (due to FWLS). These issues would be best resolved by having specific observations, as opposed to census tract data.

1. **Conclusion:**

The model that we developed greatly expanded upon the work pioneered by David Harrison and Daniel Rubenfeld. By introducing new economic reasoning and convention regarding the diminishing returns of rooms in a dwelling, housing discrimination, and the varying marginal effects of nitric oxide levels, we were able to construct a more well-fitting model than what was previously developed. We find that the largest determinant of housing prices is the concentration of NOX in the area, backing up Harrison and Rubenfeld’s earlier work regarding the importance of clean air and the cost of pollution. While there are issues regarding heteroskedasticity that could be better addressed, we believe that our model is useful for interpretation of the different factors that influence the price of housing by observing the estimates for those factors. Moreover, this model can serve as a foundation for further predictive methods due to its good fit.

**Works Cited**

Aslan, Ali. *Real Estate Dataset: Dragon Real Estate – Price Predictor.* Web. 04 Dec. 2020.

< https://www.kaggle.com/arslanali4343/real-estate-dataset>

Harrison Jr, David, and Daniel L. Rubinfeld. "*Hedonic housing prices and the demand for clean*

*air*." (1978).