Spring 2017

Housing Price Models for Essex County Massachusetts

Steven Bourque
Merrimack College, bourques@merrimack.edu

Housing Price Models for Essex County Massachusetts

Steven Bourque

Merrimack College

Abstract

This paper analyzes indicators of local median housing prices in the Essex County, Massachusetts area. The housing market is a significant sector of the United States economy and therefore it is important to develop models, which can predict and identify movements in their prices. This paper contains an overview of housing price indexes and provides a short overview of previous housing price studies. Two models, based on 2010 to 2015 data, are presented and analyzed. The specification of each model was driven by possible variables that might affect housing price and availability of data. For example a subset of communities, for the second model, were left out due to lack of available data for specific variables. Analysis of the broader sample indicates that median housing prices are positively related to the labor force, educational attainment, median income, travel time to work, the percent of the population female, where, through an interaction variable, travel time dampens the positive effect of the labor force on median housing prices. The second model shows that Math MCAS scores (standardized math test scores for 10^{th} grade students), the over 65 population, the under 18 population, the percent of the population female are all positive indicators of median housing prices. The property crime rate is shown to be a negative indicator.

Introduction: Literature Review

According to Jordan Rappaport (2007), there are three methodologies to develop a housing price index. The first is a basic average of, either the mean or median, of the housing sales in the area for which you are interested. While this is the quickest technique it does not account for the heterogeneity or differences in houses, making it the least

accurate of the three. Some difference a house might have to others is where it is located or its features; not all houses are same.

The second model takes into account the fact that houses have different values based on other factors by measuring the repeat sales of the same house (Rappaport, 2007). Two of the most popular housing price indices are the Housing Price Index (HPI) and the S&P/Case-Shiller. The HPI takes data from Fannie Mae and Freddy Mac on repeat sales and transactions of single-family house mortgages from different regions or Metropolitan Statistical Areas (MSAs) to develop a more accurate measure of housing prices based on the area and a houses attributes. The Federal Housing and Finance Agency publishes this index. The S&P/Case-Shiller index is similar to the HPI except for a few differences. The Case-Shiller index does not include refinancing appraisals like the HPI and uses a weighted average, giving more expensive houses greater influence on the index. The data they use comes from county accessory and recorder offices and does not include 13 of the states (Federal Housing Finance Agency, 2016).

The third approach, which is the technique used in this paper, is the hedonic method. This method uses statistical analysis, taking a collection of attributes of the houses' or area's characteristics, and develops an equation to predict the median housing price for the specified region. Like the second approach, the hedonic method is able to take the variation of house features and location into consideration. Although, one problem with this process is you need a lot of data to create the equation and it is often difficult to acquire the necessary information. Only one hedonic index is published regularly, the Census Constant Quality Index, because it is difficult to acquire the data for the whole of the United States housing market. This index uses 12 housing attributes to calculate the mean

value of a fictional house, adjusting for inflation by using the base year of 1996 (Rappaport, 2007).

One study done in 2010 by Wei-Shong Lin, Jen-Chun Tou, Lin Shu-Yi, and Ming Yih Yeh used the hedonic approach, finding a linear model of socioeconomic factors in different regions of the United States. They identified different MSA's, the Northeast, West, Midwest, and South, just as the HPI because different areas may have different factors that affect housing prices. They found population, percent elderly, percent Asian, median housing income, and rent to income ratio have positive correlations with the median housing price. Adversely, mortgages, vacancy rates, violent crime, and foreclosure rates all had a negative correlation to housing prices. Along with these correlations, they found the Northeast MSA had stronger correlations for population, percent elderly, and rent to income ratio than the South. The Northeast had stronger correlations with population, percent elderly, and foreclosure rates than the West. Finally, they had stronger correlations with population and violent crime rates and a weaker correlation with mortgages than the Midwest (Lin, 2014). This study highlights the differences the area being studied can have on the correlation factors of housing prices.

Earlier research done by Kenneth Rosen and Lawrence Katz in 1979 looked at housing price data of a much smaller region of 63 San Francisco, California suburban communities. The main focus of the study was to find the effects community growth control had on housing prices. They, using the hedonic model, found it had a positive correlation to housing prices in this area. Some of the variables they used to develop their model were number of baths in the house, property area in square mile, and commute time to downtown San Francisco. In their paper, they found there was a negative correlation of

travel time to housing prices (Katz, 1987). Looking at the two hedonic studies, the former probably does a better job of generalizing what affects housing prices overall in larger regions while the latter is very specific to the San Francisco area. Because the study covers a small area it may show some general factors that affect housing prices everywhere but it could also have correlations specific to the region.

Data Summary and Analyses

In my study, I collected data on Essex County, Massachusetts's cities and towns. The data used is from 2010 to 2015, assuming the data did not change drastically in the five years. Using the program Stata, I came up with two statistically significant hedonic equations that can be used to predict housing prices in the Essex county area. The cities and towns data used in the first model were Amesbury, Andover, Beverly, Boxford, Groveland, Danvers, Georgetown, Gloucester, Hamilton, Haverhill, Ipswich, Lawrence, Lynn, Lynnfield, Manchester, Marblehead, Merrimac, Methuen, Middleton, Newbury, North Andover, Peabody, Rockport, Rowley, Salem, Salisbury, Saugus, Swampscott, Topsfield, and Wenham. The second model, since some towns did not have crime or MCAS score data, excludes Merrimac, Middleton, Newbury, Rowley, Salisbury, and Topsfield.

The process of choosing these variables was mostly trial and error from the data set collected, but to begin I looked at scatter plots of the variables to see which ones might have a correlation to the median housing price. The scatter plots for this model can be found under graph 1 in the appendix. Looking at the scatter plots also helped to see the regression model might be a mixed log-linear because some of the correlations between medhouseprice and the variables looked closer to a logarithmic relationship than a linear. The log-linear model did have a higher R-squared value and therefore was selected.

The first model is a log-linear mixed regression equation. The dependent variable, median housing price, is measured as a natural log. With the dependent variable thus specified, the estimated coefficients reflect percent changes. The indicator variables are specified in Table 1 below.

Table 1: Equation 1 Variables and Description

Variable Name	Description
Medhouseprice	The median housing price in the city or town
Labforce	Percent of the population in the labor force
Bsdegree	Percent of the population with a bachelors degree or higher
Medinc	The median income for the area
Travtime	The mean travel time to work of people over the age of sixteen
Perfemale	Percent of the population that is female
Labforcetravtime	An interactive variable between travtime and labforce, where the mean travel time to work has an effect on the percent of the population in the labor force

The summary of the variable statistics and the regression for equation one can be found in table 2 and table 3.

Table 2: Summary Statistics for Equation 1 (Source: census.gov/quickfacts)

Variable	Obs	Mean	Std. Dev.	Min	Max
medhousprice	30	407746.7	117249.7	211900	738300
labforce	30	.6806333	.0362876	.571	.759
bsdegree	30	.4489667	.1565359	.119	.705
medinc	30	86351.2	23647.04	34496	127813
travtime	30	29.86	3.148026	22.7	35.6
perfemale	30	.5183333	.0177925	.44	.547

Table 3: Regression Equation Statistics for Equation 1 (Source: census.gov/quickfacts)

Source	SS	df	MS		
				Number of obs	= 30
				F(6, 23)	= 29.39
Model	2.08256491	6	.347094152	Prob > F	= 0.0000
Residual	.27164288	23	.01181056	R-squared	= 0.8846

```
-------------+----------------------------------   Adj R-squared   =    0.8545
       Total |  2.35420779        29  .081179579   Root MSE        =    .10868

------------------------------------------------------------------------------
 lnmedhouseprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
        labforce |   12.01975   6.137897     1.96   0.062    -.6774587    24.71695
       bsdegree |   .4239429   .2912421     1.46   0.159    -.1785374    1.026423
          medinc |   7.40e-06   1.94e-06     3.82   0.001     3.40e-06    .0000114
       travtime |   .3462045   .1378503     2.51   0.019     .0610395    .6313696
      perfemale |   3.679702   1.568803     2.35   0.028     .4343859    6.925017
labforcetravtime |  -.4825255   .2058786    -2.34   0.028    -.9084179   -.0566331
           _cons |   1.445038   4.244045     0.34   0.737    -7.334438    10.22452
------------------------------------------------------------------------------
```

All of the variables were statistically significant except for the constant. The R-squared value for the model is .8846, meaning 88.46% of the time a movement in the median housing price can be attributed to these variables.

As shown in the equation labforce has a positive correlation with housing price. Originally labforce had a negative correlation. Since this was not expected it was possible there was missing variable bias. The missing variable bias was confirmed by a Ramsey-Reset test, and the equation was re-estimated with various nonlinear and interaction variables. The missing variable bias seemed to be resolved by the inclusion of the interaction variable labforcetravtime (where travel time dampened the effect of the labor force on median housing prices).

Table 4: Ramsey Test Results for Equation 1

```
Ramsey RESET test using powers of the fitted values of lnmedhouseprice
       Ho:  model has no omitted variables
                  F(3, 20) =      0.26
                  Prob > F =      0.8535
```

There was also concern that the medhousprice was a determinant of medinc rather than the other way around. To test for this issue, medinc was made into the dependent variable and medhouse price was placed on the right-hand side of the equation, resulting in

a lower R-squared and giving confidence medinc is a determinant of medhouse price (This test is available upon request). One final test, shown in table 5 was run to ensure the error terms have a constant variance.

Table 5: Breusch-Pagan / Cook-Weisberg Test Results for Equation 1

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnmedhouseprice

         F(1 , 28)     =     0.00
         Prob > F      =   0.9456
```

This test confirms model 1 is homoscedastic and has constant error terms.

The variables labforce, bsdegree, and medinc all have a positive correlation to housing price as might be expected. A greater percent of people with a bachelor's degree, a greater percent in the labor force, and greater median incomes in an area would all likely result in the people of that area having greater wealth, causing them to contribute to the betterment of their community and having more money to make home improvements resulting in higher median housing prices. Perfemale also had a positive correlation. A paper written by Katrin Elborgh-Woytek (2009/2013) gives two statistics about women: they tend to spend more money on their children's education than men and they spend about twice the amount of time as men on household work. This can explain why the percent of the population that is female has a positive effect on housing prices. Spending more time on the upkeep of a house will help maintain the value of the housing and caring about there child's education will cause them to search for neighborhoods near better schools. Travtime is one of the more interesting variables; it has a positive correlation to housing price. In Lawrence Katz and Kenneth Rosen's (1987) study they found travel time

in the San Francisco area to have a negative correlation to housing prices. This is one of the variables that is likely to vary widely with the region of study and could not be generically used for the whole country. As for Essex County, most of the high paying jobs are in Boston but a lot of the higher valued, less populated cities and towns are located further away from Boston, which would result in longer commuting times. As a result, the longer you have to commute to work in Essex County, the greater the median housing price in your area, in general. While it makes perfect scene for the percent in the labor force to have a positive effect on housing prices, it is not as clear why the coefficient for labforcetravtime is negative. The negative correlation of this variable essentially means as the commuting time increases it dampens the effect the labor force has on the median house price. One possible explanation for this outcome is an increase in travel time reduces the benefits of living in the suburbs while working in urban areas.

Since the variable labforcetravtime resulted in a negative coefficient and travtime dampens the effect the labor force has on median housing price it is possible to find the travel time and percent in the labor force that is related to maximum housing price in the Essex County area. To find the percent in the labor force that generates maximum median housing prices, hold all of the variables constant except for labforce and travtime. Then by taking the partial derivative of the equation with respect to travtime and setting it equal to zero you obtain the equation:

$$0=.346-.483(\text{labforce}) \rightarrow \text{labforce} = .71$$

The result of labforce equaling .71 means most likely the percent in the labor force that will have the maximum effect on housing prices is 71%, which is slightly over the mean of the

area. This has implications for the government where if they want to raise the median housing price they could target a 71% labor force participation rate. Likewise to find the optimal travel time to work you would hold all the variables constant except for labforce and travtime and take the partial derivative with respect to labforce and setting that equal to zero, resulting in the equation:

$0=12.020-.483(travtime) \rightarrow travtime = 25.26$

Therefore the optimal travel time for maximizing median housing price in the Essex County area is 25.26 minutes. While this has less implication for the government because you cannot alter the travel time to work drastically, it can help to show you where to invest in housing based on how far away the city is from Boston in the area.

The second model is also a log-linear mixed model using five variables as determinants of median housing price. The five variables and the descriptions can be found in table 6:

Table 6: Equation 2 Variables and Description

Variable Name	Description
Medhouseprice	The median housing price in the city or town
Mcasmath	Average math scores on the MCAS test
Ov65	Percent of the population over the age of 65 years old
Un18	Percent of the population under the age of 18 years old
Propcrime	Number of property crimes per 1000 people

The summary of these variables' statistics and the regression for equation 2 can be found in table 7 and table 8:

Table 7: Summary Statistics for Equation 2 (Sources: Census.gov/quickfacts, Metrowestdailynews.com/article, and profiles.doe.mass.edu/state_report/sat_perf.aspx)

Variable	Obs	Mean	Std. Dev.	Min	Max

ov65	30	.1496333	.0321135	.086	.231
un18	30	.2308667	.0323288	.172	.29
perfemale	30	.5183333	.0177925	.44	.547
propcrime	29	13.60621	9.086079	1.12	29.21
mcasmath	22	522.4545	41.19261	411	601

Table 8: Regression Equation Statistics for Equation 2 (Sources: Census.gov/quickfacts, Metrowestdailynews.com/article, and profiles.doe.mass.edu/state_report/sat_perf.aspx)

Source	SS	df	MS	Number of obs	=	22
Model	1.82232015	5	.36446403	F(5, 16)	=	41.48
Residual	.140599968	16	.008787498	Prob > F	=	0.0000
Total	1.96292012	21	.093472386	R-squared	=	0.9284
				Adj R-squared	=	0.9060
				Root MSE	=	.09374

lnmedhouse~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mcasmath	.004455	.0007848	5.68	0.000	.0027913	.0061186
ov65	2.47914	.8205502	3.02	0.008	.7396512	4.218629
un18	2.958898	.9507004	3.11	0.007	.9435028	4.974292
perfemale	4.567082	2.674197	1.71	0.107	-1.101961	10.23613
propcrime	-.0081942	.0033715	-2.43	0.027	-.0153415	-.0010469
_cons	7.220472	1.521976	4.74	0.000	3.994028	10.44692

The same method was used in determining these variables as was used for model 1, using logic, scatter plots, and trial and error to find a statistically significant equation with no missing variables. The graph of the scatter plots can be found in the appendix under graph 2.

All of the variables in this model are statistically significant, including the constant. It has an R-squared value of .9284, meaning 92.84% of the time a movement in median housing price can be attributed to a change in these variables. Once again, the Ramsey-Rest test for missing variables was used and concluded there are likely no missing variables. This test can be found in table 9:

Table 9: Ramsey Test Results for Equation 2

```
Ramsey RESET test using powers of the fitted values of lnmedhouseprice
       Ho:  model has no omitted variables
                  F(3, 13) =      0.77
                  Prob > F =      0.5287
```

Once again, this model needed to be tested for heteroskedasticity and the same test was used. The test results confirmed that model 2 is homoscedastic and can be seen in table 10:

Table 10: Breusch-Pagan / Cook-Weisberg Test Results for Equation 2

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnmedhouseprice

         F(1 , 20)     =     1.63
         Prob > F      =   0.2160
```

All of the variables in the second model are positively correlated, except for propcrime. Logically, it makes sense for property crime rates to negatively affect housing prices because people want to live in safer neighborhoods. Prior to using property crime, I tested the model with violent crimes per 1000 people and resulted in a lower R-squared value, showing property crime had more of an effect on housing prices than the violent crime. Mcasmath had a positive correlation to housing prices most likely because people want to live in areas with high performing schools. Out of the MCAS reading, writing, and math scores the math scores had the strongest correlation to housing prices. The higher the percent of the population that is over 65 results in higher median housing price probably because people over 65 have been able to accumulate wealth over their lifetime and have invested some in improving their houses and moving to higher quality areas. As for the percent of the population under 18 having a positive correlation could be due to the family atmosphere of the city or town. If there is a high number of under 18-year-olds that means there are probably more families and the demand for housing in a nice family

friendly town or city could drive up housing demand and therefore housing prices. Finally, perfemale's positive correlation was already discussed earlier in the paper.

While both model 1 and model 2 are statistically significant, they both have their pros and cons. Model 1 has more sample points, using 30 cities and towns in Essex County, which gives more confidence that the regression is accurate because of the increased number of points. Model 2 only has 22 observations but does have a higher R-squared value. Model 2 could also be considered parsimonious, meaning it forms a similar assumption using fewer variables. Model 6 uses six variables and a constant while model 2 uses 5 variables and a constant. Also one of the former's six variables is an interactive variable, which makes it a little more complex. The only similar variable between the two models is the percent of the population that is female. It is impossible to say which model is a better predictor of median housing prices in Essex county because of the difference in the number of observations but they are both considered statistically significant models.

I was satisfied with the data I was able to collect for this study but if perfect conditions were met there would have been a few changes. First, it would have been better to compile all of the data points from all 33 cities in Essex County but that was not possible. Three of the towns did not have any data in the census, presumably because they are too small. Also, eight of the towns did not have posted MCAS scores and one town did not have 2014 crime data. Second, the tests would have benefitted from having all of the data from a single year but it was difficult to find this data for the same year, considering most of the census data was presented in averages and other data was not freely accessible. Lastly, I would have like to include more variables if I was able to find data for them on the city and town level. Some of the variables would have been: environmental measures(e.g. air

quality, water quality), percent of the population that is married, and attributes of the houses(e.g. average number of bathrooms, pools per 1000 people). Overall the models did have statistically significant results but there is always more that can be done and limits to a hedonic model.

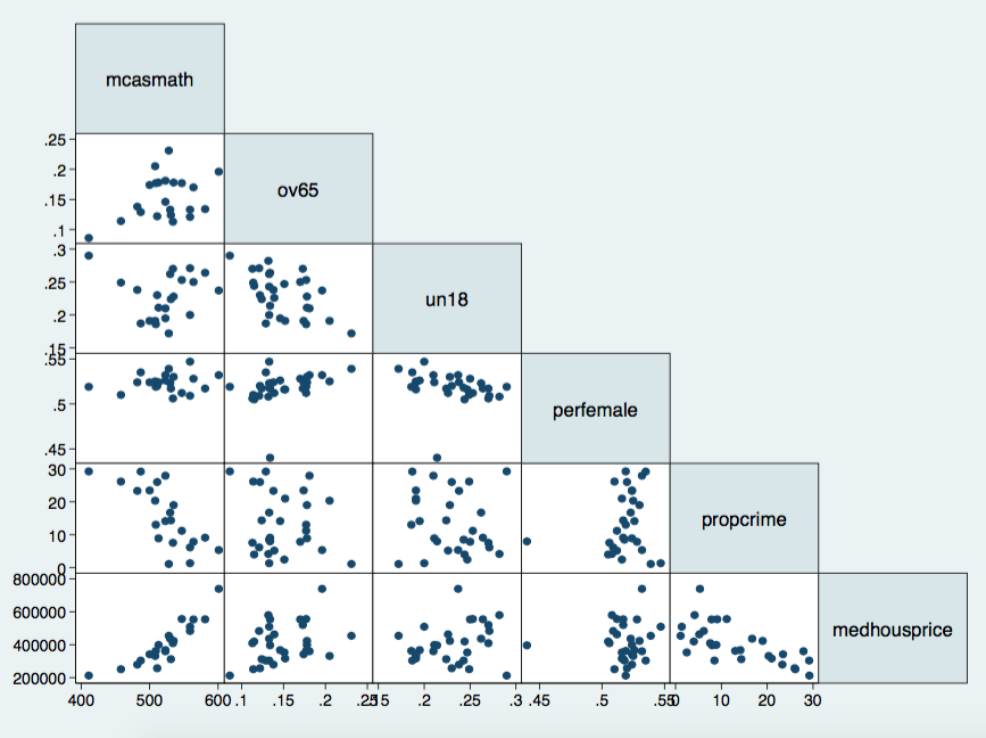
Conclusion

Housing price indices can be useful for predicting housing price trends in both specific and large regions using the different types of models. These two models give information about the housing prices in Essex county and would be most useful for this area but other parts of the state or country could have similar variables that affect their housing prices. Knowing these variables can have an effect on prices if you hear news that one of these variables is increasing or decreasing you may be able to determine the change in the value of houses in your area. Models such as these can also influence government policies. If you see MCAS scores and the percent of the population over 65 have large positive effects on housing prices and property crimes per 1000 people has less of a negative effect than the government might apply more money toward the schools system or luxury housing for the 65 and older community and less toward policing. It is important to track shifts in housing prices because homes are a major market in society and are considered a part of peoples' assets.

Appendix

Graph 1: Scatter Plot for Equation 1 (Sources: Census.gov/quickfacts)

Graph 2: Scatter Plots for Equation 2 (Sources: Census.gov/quickfacts, Metrowestdailynews.com/article, and profiles.doe.mass.edu/state_report/sat_perf.aspx)



Literature Review Sources

Elborgh-Woytek, Katrin, Monique Newiak, Kalpana Kochhar, Stefania Fabrizio, Kangni Kpodar, Philippe Wingender, Benedict Clements, Gerd Schwartz. (09/2013) *Women, Work, and the Economy: Macroeconomic Gains from Gender Equity.* International Monetary Fund

Federal Housing Finance Agency. (11/23/2016) *Housing Price Index Frequently Asked Questions.* www.fhfa.gov

Katz, Lawrence F., Kenneth T. Rosen. (1987) *The Interjurisdictional Effects of Growth Controls on Housing Prices.* Journal of Law and Economics 30(1): 149-160

Lin, Wei-Shong, Jen-Chun Tou, Lin Shu-Yi, Ming Yih Yeh. (2014) *Effects of Socioeconomic Factors on Regional Housing Prices in the USA.* Retrieved from International Journal of Housing Markets and Analysis

Rappaport, Jordan. *A Guide to Aggregate House Price Measures.* (2007) The Pennsylvania State University CiteSeerX Archives

Data Sources

United States Census Bureau. 2015. https://www.census.gov/quickfacts/

Essex County MA GenWeb. November 16, 2011. http://essexcountyma.net/towns.htm

Haddadin, Jim. “FBI Releases 2014 Crime Data for Massachusetts Cities, Towns”. October 8, 2015. http://www.metrowestdailynews.com

Massachusetts Department of Elementary and Secondary Education. “2015-16 SAT Performance Report (District) all Students”. http://profiles.doe.mass.edu/state_report/sat_perf.aspx