MESSY MLS DATA AND HEDONIC HOUSING PRICES

Class Project – OMB 3302 – Spring 2018

Brooklyn Crum

April 17, 2018 | Florida Gulf Coast University

Introduction

The major question being surveyed by this set of data is "What is the average marginal price effect of additional bedrooms in existing single-family detached houses in a specific housing market over the 2011-2013 time period?" The importance of the answer to this question is key in determining how important extra bedrooms are to the price level at which houses were being sold at. It is important for relators, housing contractors and others involved in the housing industry to look at previous data sets and findings in order to determine the effect of housing price based on the variables. For this specific set of data, it is important to find a model of best fit to find the effect of number of bedrooms on price so that analysts can look at the findings and determine if it is important to incorporate another bedroom into the floor plan or not in order to maximize its selling price. It is important for data analysists to take each possible variable into consideration and determine whether or not it has significance on the dependent variable. In the data provided from the "Messy MLS Data and Hedonic Housing Prices," it is important to analyze what information has an effect and what information and data can be removed and is unnecessary in determining the best possible model of the data. This project produces the best possible model for the set of data and answers the research question, "What is the average marginal price effect of additional bedrooms in existing single-family detached houses in a specific housing market over the 2011-2013 time period?"

Description of the Data

The raw data provided in the beginning by the real estate agents of the MLS began with 29,125 entries. Many of the entries had missing data which were not beneficial in building a model for the price effect. There were also many variables that were irrelevant in determining the price effect. For this set of data, outlier criteria were also set so that the data was not skewed. The outliers that were removed were:

- Sold Price greater than $950,000 and less than $50,000
- Bedrooms greater than 6 and less than 1
- Baths Full greater than 6 and less than 1
- Approx Living Area in square feet greater than 7,000 and less than 500
- Acres of land area greater than 5 and less than .01 or with missing information,
- A ratio of Sold Price and List Price greater than 2 and less than .5, and
- Age (calculated by subtracting Year Built from the year of sale) greater than 50.

After analyzing the data, removing the houses that were not single-family detached houses, the data with missing information and also the entries that were outliers, the sample was able to be cleaned up and reduced to 11,718 entries in order to build a best fit model. The cleaning up of data allows data analyzers to produce a model that is the best fit for the provided information. In this model, the dependent variable that is used is the natural log of the sold price. The independent variables used for the full model specification are Bedrooms, Baths Full, Baths Half, Garage Spaces, Acres, Age, Ocean Access, Pool, Waterfront, Location 1, Location 2, Location 4, Location 5, Location 6, and Trend. On the next page (*Figure 2*) there is a chart of the

	P-value
Intercept	0
Bedrooms	2.12165E-42
Baths Full	6.03E-129
Baths Half	3.0075E-85
Garage Spaces	3.09686E-60
Acres	8.22113E-36
age	1.9652E-212
Ocean Access	0
Pool	0
Waterfront	3.4124E-255
Location 1	0.740998205
Location 2	1.3777E-178
Location 4	4.41615E-38
Location 5	1.55937E-29
Location 6	7.0943E-148
Trend	0

Figure 1

statistical summaries for each independent variable, except Location 1, in the model. Location 1 is not included in this set of statistics as it is not included in the best reduced model more the set of entries provided. In the figure of data to the right (*Figure 1*), the data provided shows that Location 1 produces a P-value greater than an alpha value of .05, therefore it should be removed from the set of data for the best, reduced model. The independent variables included in the final reduced models are the ones whose statistical summaries are provided in *Figure 2*.

Bedrooms		*Baths Full*		*Baths Half*		*Garage Spaces*		*Acres*		*Age*		*Ocean Access*	
Mean	3.17069216	Mean	2.11146198	Mean	0.08338312	Mean	1.98935308	Mean	0.25149765	Mean	17.2053427	Mean	0.23564052
Standard Error	0.0045856	Standard Error	0.00333894	Standard Error	0.00263001	Standard Error	0.00391004	Standard Error	0.00077671	Standard Error	0.1121883	Standard Error	0.00392089
Median	3	Median	2	Median	0	Median	2	Median	0.23	Median	12	Median	0
Mode	3	Mode	2	Mode	0	Mode	2	Mode	0.23	Mode	6	Mode	0
Standard Deviation	0.49636917	Standard Deviation	0.36142413	Standard Deviation	0.28468576	Standard Deviation	0.42324228	Standard Deviation	0.08407555	Standard Deviation	12.1438331	Standard Deviation	0.42441659
Sample Variance	0.24638236	Sample Variance	0.1306274	Sample Variance	0.08104598	Sample Variance	0.17913403	Sample Variance	0.0070687	Sample Variance	147.472682	Sample Variance	0.18012944
Kurtosis	2.01788742	Kurtosis	10.9793383	Kurtosis	10.6557922	Kurtosis	8.9590929	Kurtosis	714.641463	Kurtosis	-0.30040356	Kurtosis	-0.44764175
Skewness	0.90574918	Skewness	2.8822863	Skewness	3.33249027	Skewness	-0.65467964	Skewness	20.7893855	Skewness	0.86837889	Skewness	1.24596737
Range	4	Range	5	Range	3	Range	5	Range	4.2375	Range	50	Range	1
Minimum	2	Minimum	1	Minimum	0	Minimum	0	Minimum	0.1025	Minimum	0	Minimum	0
Maximum	6	Maximum	6	Maximum	3	Maximum	5	Maximum	4.34	Maximum	50	Maximum	1
Sum	37151	Sum	24740	Sum	977	Sum	23309.25	Sum	2946.798	Sum	201595	Sum	2761
Count	11717	Count	11717	Count	11717	Count	11717	Count	11717	Count	11717	Count	11717

Pool		*Waterfront*		*Location 2*		*Location 4*		*Location 5*		*Location 6*		*Trend*	
Mean	0.50123752	Mean	0.33779978	Mean	0.13723649	Mean	0.14210122	Mean	0.11718017	Mean	0.17384996	Mean	18.039942
Standard Error	0.00461933	Standard Error	0.00436953	Standard Error	0.003179	Standard Error	0.00322572	Standard Error	0.00297148	Standard Error	0.00350128	Standard Error	0.09590807
Median	1	Median	0	Median	0	Median	0	Median	0	Median	0	Median	18
Mode	1	Mode	0	Mode	0	Mode	0	Mode	0	Mode	0	Mode	27
Standard Deviation	0.50001981	Standard Deviation	0.47298011	Standard Deviation	0.34411153	Standard Deviation	0.34916883	Standard Deviation	0.32164857	Standard Deviation	0.37899658	Standard Deviation	10.3815783
Sample Variance	0.25001981	Sample Variance	0.22371018	Sample Variance	0.11841274	Sample Variance	0.12191887	Sample Variance	0.1034578	Sample Variance	0.14363841	Sample Variance	107.777169
Kurtosis	-2.00031696	Kurtosis	-1.52968987	Kurtosis	2.44731397	Kurtosis	2.20432849	Kurtosis	3.66867878	Kurtosis	0.9634434	Kurtosis	-1.2665833
Skewness	-0.00495072	Skewness	0.68598197	Skewness	2.10876653	Skewness	2.05035418	Skewness	2.38076723	Skewness	1.72141772	Skewness	0.01524824
Range	1	Range	1	Range	1	Range	1	Range	1	Range	1	Range	35
Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	1
Maximum	1	Maximum	1	Maximum	1	Maximum	1	Maximum	1	Maximum	1	Maximum	36
Sum	5873	Sum	3958	Sum	1608	Sum	1665	Sum	1373	Sum	2037	Sum	211374
Count	11717	Count	11717	Count	11717	Count	11717	Count	11717	Count	11717	Count	11717

Figure 2

For the independent variable bedrooms, the mean number of bedroom for the sample is 3.17. The minimum number of bedrooms is two and the maximum is six. The standard deviation is .496. This standard deviation data value provides us with the information that the data for bedrooms is not to largely spread out or close and compact to the average. Another mean to specifically point out is the age. The average age for houses in this sample is 17.21 based on a minimum of zero to a maximum of fifty years. A higher average of houses had a pool instead of a waterfront view or ocean access. The average acreage for the houses in the sample was ¼ of an acre. Two garages, 2.11 full bathrooms, and .08 half baths were also averages for the houses in the data. The locations were all relatively equivalent in average. Location 2's average was .137, Location 4's was .142, Location 5's was .117, and Location 6's was .174. This is important as the housing data was taken from each area relatively equally and there is no biasness in our data as a result. Figure 3 shows the statistical summary for the dependent variable, the natural log of selling price and also the standard selling price of the houses.

ln(Sold Price)	
Mean	11.880952
Standard Error	0.0051484
Median	11.81303
Mode	11.561716
Standard Deviation	0.5572926
Sample Variance	0.3105751
Kurtosis	-0.22632
Skewness	0.4987191
Range	2.9318589
Minimum	10.821776
Maximum	13.753635
Sum	139209.12
Count	11717

Sold Price	
Mean	170945.78
Standard Error	1052.1563
Median	135000
Mode	105000
Standard Deviation	113890.75
Sample Variance	1.297E+10
Kurtosis	6.7203975
Skewness	2.1748139
Range	889900
Minimum	50100
Maximum	940000
Sum	2.003E+09
Count	11717

Figure 3

After creating a set of data that was effective in creating an analysis, a final reduced model is:

E[ln(soldPrice)] = 10.632 + 0.072(Bedrooms) + 0.175(Baths Full) + 0.162(Baths Half) + 0.1(Garage) +0.340(Acres) - 0.009(Age) + 0.480(Ocean) + 0.332(Pool) + 0.273(Waterfront) - 0.234(Location 2) - 0.098(Location 4) - 0.093(Location 5) - 0.198(Location 6) + 0.012(Trend)

This model provides all of the variables that provide a significance when determining the natural log of the sold price. The coefficients in this model come from the summary output of regression coefficients (Figure 6 on page 5). It is important to analyze the effect of the coefficients of the independent variables on the dependent variable and analyze how changes in each independent variable effect the dependent variable, the natural log of selling price. We are able to determine that this model is the best reduced models as all of the p-values are less than the alpha value of 0.05. This is shown in figure 4 to the right. By having a best, reduced model, we are able to analyze the goodness of fit, the standard error and make accurate predictions.

	P-value
Intercept	0
Bedrooms	2.172E-42
Baths Full	6.09E-129
Baths Half	3.031E-85
Garage Spaces	3.225E-60
Acres	7.649E-36
age	9.05E-255
Ocean Access	0
Pool	0
Waterfront	3.06E-255
Location 1	1.3E-187
Location 2	6.461E-44
Location 4	9.034E-31
Location 5	3.09E-153
Location 6	0
Trend	

Figure 4

Findings

The final reduced model discussed previously can be used to determine the results of the model and what the different coefficients and variables tell data analyzers. The final model also produces statistics that help explain the accuracy of the model and the goodness of fit of the model. *Figure 4* (shown below) provides the regression statistics for the model population. The adjusted R square is the value that will be used to analyze in order to determine the goodness of the model created. By definition the R square value is, "a measure of "fit" of the line to the data; the value of the R square will be between 0 and 1. The larger the value of R square, the better the fit" (Evans 614). For the analysis of this sample, the adjusted R square value will be used as it provides a more accurate goodness of fit value since there were variables that were removed, thus affecting the R square number. For this model, the Adjusted R Square value is 0.81307. This value tells us that the model that has been created with the independent variables, Bedrooms, Baths Full, Baths Half, Garage Spaces, Acres, Age, Ocean Access, Pool, Waterfront, Location 1, Location 2, Location 4, Location 5, Location 6, and Trend and the dependent variable, the natural log of selling price is a strong fit, since the value is far greater than 0.7. The standard error for our model is 0.2409. The standard error tells analysts the "standard deviation of the sampling distribution of the mean" (Evans 615). It also

Regression Statistics	
Multiple R	0.901830408
R Square	0.813298085
Adjusted R Square	0.813074719
Standard Error	0.240944672
Observations	11717

Figure 5

allows analysts to determine a confidence interval. A confidence interval is "a range of values between which the value of the population parameter is believed to be along with the probability that the interval correctly estimates the true (unknown) population parameter" (Evans 610). With the given mean of the natural log of the selling price of 11.88, analysts can say, with 95% confidence, the natural log of the selling price of the house will fall between (11.640,12.122). This confidence interval is constructed by using the average of the natural log of the selling price found in *Figure 3*, the standard error, in *Figure 4*, can be added and subtracted from the average in order to determine the confidence interval. Provided the adjusted R square and the standard error, our model can forecast and provide close to accurate predictions of the natural log of selling price. Therefore, the model created is of best fit for the given data.

Our final reduced model also helps answer the research question, "What is the average marginal price effect of additional bedrooms in existing single-family detached houses in a specific housing market over the 2011-2013 time period?" From the regression analysis provided in *Figure 6*, The regression coefficients are given for each independent variable as well as the intercept in the model. The intercept is the percentage increase in the natural log of selling price when all of the other independent variables are zero. The p-values of each of the independent variables are all below the alpha value of .05. Thus, they all have an effect, both positive and negative, on the selling price of the house. The p-values for the variables are shown in figure 4 above. As for this specific research question, the focus will be on the coefficient of Bedrooms. The coefficient is 0.07239. This value can be interpreted as, for every increase in number of bedrooms, the selling price increases by 7.239%. This is a positive effect on the price. Therefore, the number of bedrooms positively effects the selling price of the house. After acquiring the data from the population, cleaning up the data in the sample, and creating a model that is of best fit for the data, it can be determined that the addition of a bedroom to an existing single-family detached house in a specific housing market over the 2011-2013 time period, the average marginal price effect is 7.239%.

	Coefficients	*Standard Error*
Intercept	10.63211536	0.022429858
Bedrooms	0.072389819	0.005284152
Baths Full	0.174997188	0.007153454
Baths Half	0.161540951	0.008188717
Garage Spaces	0.099976776	0.006072924
Acres	0.340358197	0.027141466
age	-0.008740657	0.000250024
Ocean Access	0.479887755	0.009276814
Pool	0.332354211	0.005059426
Waterfront	0.27339909	0.007812868
Location 2	-0.234480905	0.007880317
Location 4	-0.098047832	0.007025276
Location 5	-0.093177077	0.008056314
Location 6	-0.19820267	0.007404886
Trend	0.012289596	0.000215964

Figure 6

Summary

After conducting analysis of data, the answer to the question, "What is the average marginal price effect of additional bedrooms in existing single-family detached houses in a specific housing market over the 2011-2013 time period?" was found to be that for every

increase in bedroom, the selling price of the house increased by 7.239%. The most difficult part of this data was how messy it was in the beginning. There were many entries with missing data values and also a lot of entries that were not applicable in answering the question regarding single-family detached houses. One way to provide a better analysis would be to have a set of even more data points that had all of the necessary information. This would lead to even more accurate results in the data and confirm what has been concluded with the current set of data. Another way to improve the analysis to better answer the research question is to test the outliers and their effect on the data. This may result in a more accurate model. Overall, the data provided was able to produce a model that was a strong fit and answer the research question at hand regarding the effect of additional bedrooms on the average marginal price.

References

Evans, J. R. (2016). *Business Analytics: Methods, Models, and Decisions* (2nd ed). Pearson.