DePaul University

Final Report

Boom or Bust? Determinants that drive house prices

Online Graduate Students

Ryan Patrick
Akbar Aidarov
Amy Aumpansub

CSC423 Data Analysis and Regression

Professor Nandhini Gulasingam

November 21, 2017

Abstract

Understanding housing markets and how prices are determined is critical for consumers, industry professionals, and government entities. There are many macro and microeconomic forces that influence house prices and fluctuate as market conditions change. Defining a statistical approach to modeling housing prices is a complex task, and considerable research has been completed on this topic. Standard modeling approaches have been built, and the most well-established approach is known as the Hedonic Pricing Model. In this paper, we employ principles of the Hedonic model to develop and compare three separate models for predicting housing prices. Using identical datasets representing the Ames, Iowa housing market, variables were then examined and added or removed to build and improve each model. All three models had strong validation statistics, and this paper will compare them in order to identify the model that best fits the Ames, Iowa housing market dataset. The study was conducted on limited amount of data, and it is recommended that future analyses include a larger sample size to improve the final model.

Introduction

One of the primary causes for inflated prices that led to the 2007-2010 housing crisis were house appraisals made by appraisers who used unstandardized approaches for valuing houses. Many modern industry websites, including Zillow and realtor.com, base their entire business on their ability to accurately predict housing prices. It can therefore be argued that the need for consumer protection and a healthy housing market the industry suggests that regression analysis could be used as a scientific approach for assessing property values. This paper is a study in the application of regression analysis to model housing prices in the Ames, Iowa housing market. Three different models are compared in order to come up with one final model that best fits the dataset. The objectives of this analysis are to determine which model is most reliable for predicting housing prices, identify significant variables that influence housing prices, and to understand how location and structural characteristics influence prices. The methods and approaches taken are supported throughout the paper with references to the Open Journal of Statistics, as well as other academic white papers.

Methodology

The dataset used for this analysis was obtained from the predictive modeling competition website: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data. It is a sample of 1461 observations of housing data representing the Ames, Iowa housing market (file named 'uncleaned house data (ORIGINAL DATA).csv'). The dependent variable is a quantitative variable called "SalesPrice." There are 79 independent variables, of which 29 are quantitative and 50 are qualitative variables. These variables can be classified as structural, environmental, and location characteristics of the houses in addition to the variables representing periods of time, in which the constructions or significant transactions and modifications took place. 49 of independent variables are recoded as "dummy" variables in accordance with the dataset's original codebook (file named 'Original Codebook.txt') in order to represent the qualitative variables in the model. Variables representing months were grouped into seasons, while variables representing years were grouped into periods of one or two decades. Dummy variables allow for similar categories within a variable to be grouped together. It reduces the number of variables in the model and increases the number of observations per variable. Initial implementation of dummy variables in accordance with the original codebook increased the

number of total variables for consideration in the model to 299 (See Appendix E for description of all dummy variables). The file that was used by each of our team members is called 'housedata_new.csv'.

Due to a particular interest in Hedonic pricing characteristics and their examination with regards to house prices, this project was split into three parallel investigations of the data. Each investigation explored the data, cleaned it, transformed the variables and created new ones, if it was needed, and came up with the 'best' possible regression model. While approaches to each of these steps might have differed, the overall process of analysis was standardized and followed with the aim to compare the models afterwards. Each of the three analyses:

- *Checked for linearity of the associations with dependent variable and examined the distributions;*
- *Conducted analysis of the residuals and check for the extreme outliers;*
- *Investigated if there was spatial autocorrelation;*
- *Conducted model selection and validated it by performing training/test set split;*
- *Performed goodness of fit test and tested the model's predictive power on two random scenarios with unknown outcomes*.

Ryan Patrick's Analysis

The dataset for the Ames, Iowa housing market contains very detailed data points about environmental, location, and structural characteristics of 1460 houses within the town. For someone who is not familiar with the industry, forming a hypothesis to construct a regression model was challenging because the dataset contains 79 independent variables which may or may not appear to be related. Given our understanding of the Hedonic pricing model, it was important to identify variables that would express the interaction relationship between location/environmental characteristics and structural characteristics. Before beginning work to identify relationships in the data, the raw data is reviewed to determine variables that require coding for dummy variables due to qualitative data type, or the need to group levels of a variable into larger categories. As discussed in the introduction of the paper, 49 variables were coded as dummy variables. In keeping with the objective of identifying interaction relationships between location/environmental variables with structural characteristics of a house, a new dummy variable is also created called "NumNeighborhood_GRLivArea." This variable measures the interaction between the neighborhood a house is located in, and the square footage of the house. The hypothesis for creating this variable is that the price per square foot of a house is different depending on the neighborhood a house is located in. A second interaction variable, "MSZoning_LotArea," is a measurement of the relationship between building zone types and the size of the lot that a house is on. The hypothesis with this variable is that houses in urban areas will be priced higher if they have larger lot sizes. The assumption is that square footage is more valuable in certain neighborhoods compared to others. After reviewing the raw data and creating the dummy variables, exploratory analysis is conducted on the dataset to determine the shape of the distribution of the data as well as whether or not there is a linear relationship for the dependent variable, SalesPrice. The histogram for SalesPrice shows that the distribution is clearly skewed positively to the right. With a standard deviation of 79442, the maximum value lies 7 standard deviations away from the mean. The probability plot shows a positive convex curve, expressing a non-linear relationship and further supporting evidence for skewed data. As a result, a log transformation is performed on SalesPrice, which yields favorable results in the form of an even distribution of the log of SalesPrice(ln_SalesPrice).

With the Y variable now transformed, exploratory analysis is continued by visualizing all of the variables in a scatterplot matrix in order to identify potential multi-collinear relationships between the independent variables. This matrix, along with a Pearson correlation matrix show that there are no relationships between independent variables that meet or exceed a correlation value of .9. This indicates that there are no significantly correlated independent variables. While this is only an early

indicator of multicollinearity, it is sufficient evidence to begin the process of selecting a model with the variables in the dataset, followed by removal of influential points and outliers.

Using SAS to run a regression procedure, the initial model returned yields an F value of 64.66, which is significant at the .05 level with a p value of <.001. The R squared and adjusted R squared are .944 and .933, respectively, which are strong enough values to substantiate the model. The rstudentized residuals versus the predicted values show even distribution with no pattern or violation of model assumptions such as independence or constant variance. The predicted value versus the y variable ln_SalesPrice also shows a linear relationship. While these indicators are acceptable, this model can still be improved to ensure the best fitted model is identified. To improve on this and find an even better model, outliers and influential points must be removed. The studentized residuals reveal outliers while Cook's D identifies influential points. These indicators and charts are reviewed for the dataset and 62 observations which are found to be influential points or outliers are therefore removed from the dataset. Re-running the regression procedure shows that the model has improved after outliers and influential points were removed. The F value changes to 103.2 which is significant with a p value of <.001. The R2 and Adjusted R2 values have increased to .965 and .955, respectively, and the root mean square error has decreased to .079. While this is a significant improvement, the model can still be improved and SAS is used to run the Backward and Stepwise selection methods.

Before running the selection methods and testing the final model, the original dataset is split into a training set and a test set of data. 80% of the observations are then reserved for training the model during the model selection process, while the remaining 20% are used to test the validity of the final model. The observations split into each set must be randomly selected, so a "seed" of 899512 was selected to randomize the selection of observations for the split. A new variable called "train_y" is then designated as the y variable for the training set. The backward selection method begins with all variables in the model and eliminates them one by one was they are found to be insignificant for the model based on a significance level of .15. The backward selection process yielded 163 variables with an F value of 158, significant with a p value of <.001, and an R^2 of .964. These values are lower compared to the stepwise selection process which yields 102 variables with an F value of 215 and an R^2 of .95. It is possible that the stepwise model is stronger due to the lower significance threshold of .05 compared to the backward threshold of .05. The predicted values compared to the train_Y value show a linear relationship for residuals, and other residual charts show random distributions in compliance with all model assumptions. Therefore, the stepwise model is stronger than the backward selection model, and was chosen as the final model to run the test dataset on. The validation statistics of the stepwise model include the standardized estimate, which is used to determine which variables are the most significant for the model. The following are the top five variables identified by ranking by the standardized estimate: MSZoning3_lotArea, MSZoning4_lotArea, MSZoning1_LotArea, GrLivArea, and msZoning2_lot area. If the model can be validated during the testing process, these variables would seem to confirm the Hedonic model theory that housing prices are influenced by both location and structural characteristics of a house.

In the final step of the model selection process, the model selected from the stepwise method is run on the test dataset to determine if it performs well compared to the training set. The test model yielded a much different result than the training set. The R squared and adjusted R squared for the test model turned out to be .784 and .885, respectively, with a much higher root mean squared error of .196. The Phat value for ln_SalesPrice from the original dataset has a correlation value of .885 with train_y from the test set. The residuals are evenly distributed with no visible patterns, and predicted value has a linear relationship with train_y. The test set had a much lower performance compared to the training dataset and does not support the model the chosen model. While the training set yielded promising

validation statistics, the errors from the training set to the test set are significant, which indicates that this model should be reworked or eliminated. Further analysis could include improving the model by removing some of the variables that have lower standardized estimates, or increasing the sample size of the dataset. Another approach to improving this model could be to review the number of observations for various levels of qualitative variables. Levels within variables could be grouped if there are very few values available. This would decrease the number of variables in the model will increasing the number of observations per variable. Finally, future work should be done to further explore the effects of interaction between other independent variables in the dataset.

Akbar Aidarov's Analysis

Data Transformations and Cleaning

The first and foremost challenge associated with the data was the fact that it had too many variables after recoding it in accordance with the original codebook - 299 independent variables, of which only 29 are numeric. Limited number of observations (1460) also necessitated reconsideration of the codes and aggregation of some of the numeric variables along with grouping the dummy variables into larger categories. As a result, dummy variables for neighborhoods were grouped by geographic proximity (Appendix B-1). Types of dwelling (MSSubClass) were grouped into four more general categories instead of 16 very specific ones. Many other dummy variables were grouped and recoded in the same manner (see file 'Akbar's New Codebook.xlsx' for reference). Square footages of different parts of the house were summed up into one variable (TotalHouseSF) with the exception of unfinished areas. The same was done with the number of bathrooms around the house (TotalBaths). Aside from these changes, the dependent variable also was log-transformed (ln_saleprice) to solve the problem of its skewed distribution (see Appendix C-2). The number of independent variables dropped to 233.

Initial regression model including all new variables was run in SAS and examined for outliers - observations with significantly high residual and Cook's Distance values (red and blue arrowheads). After testing and re-running the model 9 times, all 24 outliers were deleted along with 27 parameters that had 0 contribution to the model. Such cleaning process resulted in data with 1436 observations and 206 independent variables. 21 of these variables are numeric and showed no multicollinearity (< 0.9) in the color coded Pearson Correlation table (Appendix B-2). The independent variable (ln_saleprice) had the strongest associations with TotalHouseSF, GrLivArea, GarageCars, GarageArea, TotalBaths (Appendix B-3). The scatterplot matrix of these variables shows linearity of the correlations (Appendix B-4). The statistics of the final regression model were as following: R^2=94.56%, Adj-R^2=93.65%, RMSE=0.0982, MSE=0.00964, F=103.09 (p<0.0001).

Model Selection and Validation

The cleaned dataset was further randomly split into training (80%) and testing (20%) datasets. The seed number used is '132408'. The training set (1141 observations) was used to fit a regression model using STEPWISE selection method with SAS's default configurations. It performed 112 steps and came up with a model with 88 independent variables (Appendix B-5) and following statistics: R^2=94.33%, Adj-R^2=93.85%, RMSE=0.09767, MSE=0.00954, F=198.79 (p<0.0001). However, not all of the parameters pass the significance test (alpha=0.05), which is why another STEPWISE selection was conducted with significance level at the entry (SLENTRY) and significance level to stay (SLSTAY) specified at 0.05. This time it performed only 60 steps and resulted in a model with 56 variables (Appendix B-6) and the following statistics: R^2=93.34%, Adj-R^2=92.99%, RMSE=0.10428, MSE=0.01087, F=271.16 (p<0.0001). For the model equation of the model refer to (Appendix B-7). The ten strongest parameters ranked by the standardized estimates are: 1) numOverallQual2 - high overall quality; 2) TotalHouseSF - total house square footage; 3) numPoolQC1 - good quality pool; 4) numOverallQual1 - average overall quality; 5)

GrLivArea - living area above ground; 6) numOverallCond2 high overall conditions; 7) numOverallCond1 - average overall conditions; 8) numExterQual3 - average quality of the exterior material; 9) numYearRemodAdd5 - last remodeled/constructed in 2001-2010; 10) numExterQual2 - good quality of the exterior material.

Given the significant p-value of F, we can reject the null hypothesis and claim that at least one of the variables has a significant effect on changes in the independent variable (tr_y). The analysis of the residual plots shows issues with constant variance and independence of the variables because of the funnel-shaped patterns (Appendix B-8). This is most likely due to a large amount of outliers and very small proportions of cases in the referenced categories. The first finding in this analysis is that more observations are needed to ensure compliance with the assumptions of constant variance and independence. As other assumptions were satisfied, the test of the trained model was conducted by fitting the model to the test dataset.

The computed predicted values (phat) were saved (287 observations) and used to calculate the model performance statistics: RMSE=0.11834 and MAE=0.09145 (Appendix B-9). The correlation coefficient of tr_y and phat is 0.94816, which means the R^2_{TEST}= 0.899. The Adj-R^2_{TEST} is equal to 0.8744, because 1 - ((1-0.899)*(287-1)) / (287-56-1) = 0.8744. As expected, we can conclude that the model performs better on the training data. Also, the R^2_{CV} = $|R^2_{TRAIN} - R^2_{TEST}|$= 0.9334 - 0.899 = 0.035. This means that, since $R^2_{CV} < 0.3$, the model has a good predictive performance.

Predictions on New Data

The data points for the two new observations and prediction results are shown in the Appendix B-10. For the first case, the model predicted tr_y = ln(SalePrice) = 11.6515 with 95% confidence interval (11.3373, 11.9657) and with 95% prediction interval (11.2766, 12.0264). This means the predicted SalePrice is $114,863.53 with 95% C.I. ($83.893.21, $157.266.96) and with 95% P.I. ($78,952.36, $167,108.74). For the second case, the model predicted tr_y = ln(SalePrice) = 12.3547 with 95% C.I. (12.1183, 12.5912) and with 95% P.I. (12.0421, 12.6674). This means the predicted SalePrice is $232,048.11 with 95% C.I.($183,193.82, $293,960.30) and with 95% P.I. ($169,753.05, $317,235.60).

Discussion and Conclusion

As we applied Hedonic approach to house pricing, we paid attention to three general aspects of house valuation such as location, environment, and structural features of the house. Aside from the detailed features of the houses, we had variables representing the zonings, neighborhoods, and descriptions of the surroundings such as proximity to public amenities and transportation. Given the final model of this analysis, the strongest predictors of the house prices in Ames, Iowa turned out to be exclusively tied to the quality and conditions of the house as well as its square footage. This very much correlates to the findings claimed by Zainodin and Khuneswari in their article (2009), where they also thoroughly examined configurations of the selection methods ($p < 0.05$) and chose the restricted linear regression model. Continued work on this project would suggest looking into dealing with the issues of residuals' constant variance and independence through either increasing the amount of observations or logarithmic transformation of the independent variables.

Reference

Zainodin, Haji and Gopal Khuneswari. 2009. A Case Study on Determination of House Selling Price Model Using Multiple Regression. *Malaysian Journal of Mathematical Sciences* 3(1): 27-44.

Amy Aumpansub's Analysis

The original dataset has 1460 observations along with house features such as structural, environmental, and location factors. It contains "SalePrice" as a dependent variable, 79 independent variables including 29 numeric variables, and total 299 independent variables after creating dummy variables. The analysis uses a hedonic house model as a framework.

Methodology

Data exploration

- The histogram in Appendix C-1 shows a skewed-right distribution of Sale Price, which the mean of $180,921 is greater than the median and the mode. The skewness of 1.99 and kurtosis of 6.5 confirms the distribution is positively skewed, which most values fall within the lower range.
- The "SalePrice" (y) is transformed using lognormal to ln_saleprice. The histogram displays a normal distribution of the ln_saleprice variable with standard deviation of 0.39 and the mean of 12.02, which is close to the median of 12.0 and the mode of 11.84 (see Appendix C-2).
- The scatterplots and Pearson correlation coefficient tables display a positive and strong linear association between the ln_saleprice (y) and GrLivArea (Above ground living area) and GarageCars (size of garage in car capacity), and GarageArea (size of garage in car capacity) (see Appendix C-3). Additionally, all correlation values are less than 0.9, which does not present a potential multicollinearity problem.

Data Cleaning and Residual Analysis

- The regression model is fitted using all 299 predictors to identify potential outliers, influential points, and multicollinearity and analyze residuals. Appendix C-4 shows that residuals satisfy three assumptions of linearity, constant covariance, and independence. All points are randomly scattered around zero line. The normality assumption is not violated if the outliers are eliminated as a histogram of residuals shows a normal distribution (see Appendix C-4).
- 17 major outliers and influential points with both red and blue arrowheads are eliminated. The VIFs indicate that few predictors such as Pool Quality have a multicollinearity with VIF greater than 10. After cleaning data, the model is improved, which the F-value increases from 69.45 to 95.09 and the adjusted R^2 increases from 0.9277 to 0.9464 (See Appendix C-5).

Model Validation

- The cleaned dataset, "housedata_new.csv" is split into train/test sets with a seed number of 731425. A training set contains 80% of total observations or 1155 obs. A test comprises 20% or 288 obs. Training set is used to fit the model. Test set is used to test predictive performance.
- The regression model is fitted using a training set with a stepwise selection method. The SAS output (Appendix C-7) displays the model from step 174 or the last step of stepwise method. This model has train_y as a dependent variable and 118 predictors. The output also shows the partial adjusted R^2 that each predictor contributes to the model's adjusted R^2.
- The new model is fitted with only strong predictors, which materially contribute to the model's adjusted R^2. The new model excludes the predictors with low partial adjusted R^2 or less than 0.01. As a result, the final model consists of only 79 predictors (Appendix C-8).
- The final model contains only statistically significant predictors with p value <0.001 (less than alpha of 0.05). It has a high F-value of 181.31 with the p-value of <0.001, so we reject the null hypothesis and conclude that at least one parameter has a significant effect on changes in log sale price. The final model has a high R^2 and adjusted R^2 of 0.9304 and 0.9253 and a small RMSE of 0.107. The residual plots show that the final model satisfies all assumptions (Appendix C-8).

- The predictors in final model are ranked with STB (Appendix C-9). The strongest predictors include: 1st PoolArea (in sq.ft.), 2nd numPoolQC4 (with/without pool), 3rdNumMSZoning3 (Industrial Zoning), 4th GrLivArea (above ground living area), and 5th numOverallQual2 (quality).
- The positive parameter estimates indicate that these variables positively affect the log of sale price (y-var). For example, Beta of numMSZoning3 is 0.43: [exp(0.43)-1]*100 = 53.72%. Thus, the sale price of a house located in industrial zoning is 53.72% higher than a price of house not in industrial zoning, assuming other predictors are constant.
- The predicted values, "phat" are computed for the test set. Then, the model performance on test set is computed based on phat. The performance statistics in Appendix C-10 display that a test set has a RMSE of 0.105 and MAE of 0.08. The R^2 of test set is computed by squaring the correlation value between phat and ln_saleprice. The R^2 of test is $0.96^2 = 0.9219$. The cross validated R^2 is calculated as an absolute value of the difference between R^2_{train} and R^2_{test}. The R^2cv is 0.008, which is less than 0.03, so the final model has a good predictive performance.
- Then, the performance statistics of train and test sets are compared. The adj R^2 of training set of 0.93 is slightly higher than adj R^2 of test set 0.92. Both sets have similar RMSE of 0.105. Thus, the model performs relatively better on a training set (Appendix C-11).

Amy's Final Model Please see Appendix C-9 for model equation and Appendix E for dummy variables.

New Predictions The final model is used to predict prices of a house in 2 scenarios (Appendix C-12).

- **Case 1:** The predicted log of sale price is 13.30. Thus, a house with 1,000 sq.ft. of above ground living area, 300 sq.ft pool, and located in industrial zoning has a predicted sale price of $597,195 with 95% C.I. of ($412,503, $873,269) and 95% P.I. of ($388,481, $918,043).
- **Case 2:** The predicted log of sale price of is 13.17. Thus, a house with 1,200 sq.ft. of above ground living area, 200 sq.ft pool, and located in industrial zoning has a predicted sale price of $524,394 with 95% C.I. of ($388,481, $700,815) and 95% P.I. of ($362,217,$751,630).

Conclusion and Future Studies

- The house with pool, high quality of materials, in very good condition, and located in the industrial zoning tends to have a higher percentage change in price. Comparing to environmental factors, the interior house features and location have a stronger effect on the housing prices.
- According to Feng et al. (2012), the hedonic model is used with 10 first-order or higher order predictors. Rather than testing the model, the Maximum Information Coefficient (MIC) is implemented to find the strength of the relationship between the model and the target value and can be used in amy function types. Box-Cox analysis is used in data exploration. Both Amy's and Feng's model transform Y to log(y).
- Given the limited number of observations, the future study should obtain at least 6,000 observations for 299 predictors. The predictors with similar characteristics should be grouped into same categories to include only strong predictors in analysis. The model can be improved by using the polynomial model and conduct MIC to explore the association between x and y.

Reference

Feng, W., Hu, G., and Wang J. 2012. Multivariate Regression Modeling for Home Value Estimates with Evaluation using Maximum Information Coefficient. *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing,* SCI (443):69-81.

Team's Final Model

Model Comparison

Three models are compared based on their accuracy, goodness of fit test, and performance on training and test sets. All 3 models perform better on a training set (Appendix D). Considering a training set, Ryan's model has the highest R^2 and Adjusted R^2 of 0.9545 and 0.9501, respectively. Amy's and Akbar's models have similar R^2 of 0.93, but Akbar's model has the lowest number of predictors. For the test set, Akbar's model has the highest adjusted R^2, the lowest RMSE, and the lowest number of predictors. The cross-validated R^2 of all three models are less than 0.03, which indicate a strong predictive performance of these statistical models. As a result, the team's final model comes from Akbar's model because it has the fewest predictors, highest F-value, lowest RMSE, and high Adj R^2. Although, the final model does not satisfy all of the assumptions of residual analysis, such as independence and constant variance, the analysis does suggest how such problem could solved via log-transformation of the independent variables. The model contains log of sale price as a dependent variable and 56 predictors including interior and exterior house features, location, environment. The major determinants that are most significant in predicting housing prices are overall quality and conditions of the house; total indoor square footage, especially above ground level; pool quality and the quality of the material on the exterior. The remaining 46 variables that drive the sale price of a house also include location and environmental factors such as the neighborhood, zoning and proximity to local amenities. The positive parameter estimates of independent variables indicate that they are positively associated with percentage change in prices.

Conclusion

Our study uses the hedonic price model as a framework and considers three major determinants including structural characteristics of a property, location, and environmental factors to identify and understand how these factors influence housing prices. According to our analysis and findings, a house with greater total indoor square footage, good or high overall quality and condition of the house, and higher pool quality, etc. tends to have a higher percentage change in sale prices. Compared to location and environmental factors, the interior and exterior house features or structural factors seem to have a stronger significant effect on the changes in sale prices of a house.

Limitations and Future Studies

There are several limitations of this study. The original dataset comprises 1460 observations along with 79 independent variables and contains 299 explanatory variables after implementing dummy variables. Additionally, this analysis does not take into account the effect of economic cycle on housing prices. When the economy is not stable, the sale prices are not accurately predicted by using only house features in dataset. Considering extrapolation, the dataset contains only housing prices between 2006 and 2011. Therefore, predicted value is less accurate if the new data point falls outside the range of our data (beyond that period). The study was conducted on limited amount of data, and it is recommended that future study include a larger sample size to at least 6000 observations and combine similar predictors in order to ensure a good representation of the observed data and improve the final model.