Predicting Housing Prices using the Ames Dataset

David Almonte

Emory University

QTM 220 Final Project

Abstract

Our research seeks to build models to identify predictors of housing costs in Ames, Iowa. Specifically, we focused on factors including Proximity to main road/railroad, Basement Quality, Fireplace quality, and Heating quality/condition, while controlling for seasonal effects. Including these factors in regression adds to the existing literature by including atypical determinants of housing pricing. We find surprising results in variables like Fireplace quality and Basement, which substantially predict consumer preferences. Besides Seasonal effect, we conclude our discussion with the fact that every additional explanatory variable introduced is statistically significant, implying the accuracy of our model.

Keywords: Housing, Ames Dataset, Linear Regression

Introduction

Predicting Housing Prices using the Ames Dataset

Much empirical research exists assessing determinants of housing prices using both supply and demand-side factors. While most of the research addressed macroeconomic determinants of housing prices, including GDP income of the country, taxation, demographics, and loans, (Panagiotidis) research on amenities as determinants of housing prices remains minimal. Using the Ames dataset, we identify a series of amenities which may be relevant for housing prices. Of particular interest are; the effects of winter seasons, basement quality, externalities such as noisy railroads, fireplace insertion, and HVAC quality (heating, ventilation, and air conditioning).

First, we seek to identify whether main road activity has substantial externality effects. If proximity to a road negatively predicts prices we may either infer that the presence of nearby traffic devalues property or that nearby traffic reflects in housing price. Our slope estimate will be positive in this case. Secondly, we see the "basement quality" as a proxy for other desirable/undesirable home characteristics. Houses with nice basements likely also have nice kitchens. Since weather in Iowa, heating is surely crucial when purchasing an assets. Thus we expect a positive relationship between "Heating quality and condition" and sale price. In Mats Wilhelmsson's article 'Journal of housing Economics' published on March 2008, it mentioned that heating system's condition has significant effect on housing depreciation rate, influencing the housing price sold according (Wilhelmsson 3). We expand upon this analysis by testing the effects of the inclusion of a fireplace. We will compare if there is any relative difference between having a fireplace and HVAC quality. Thus in our third research question, we are interested in how heating condition of a house affecting the housing sold price.

Additionally, we try to figure out if the psychology of "omitting when not imminent" applied to house-purchasing. That is, when buyers are looking into the markets during winter and spring months(Sept - April), they may emphasize on the heating condition more because the experience with cold is fresh. In contrast, people seeking new house during warm summer may ignore this situation. Thus, we expect the effects of heating quality and condition to be more significant during winter/spring months. In this way, we can identify whether the salience of a housing factor affects the value customers place on it.

Our research can be later used to guide real estate developers in identifying value-maximizing investments in the housing stock.

Methods

We use OLS Multiple Regression to construct a model to predict housing prices. Variables will be added in the order of expected explanatory effect in the model, using our intuition. Variables which are added and fail to improve model prediction significantly, as determined by F-test statistics, are excluded. This approach allows us to align our model with real estate theory, makes us conduct more statistical tests, and ensures predictors with minimal explanatory effect remain outside of the model.

In addition to OLS regression, we use various statistical tests to infer conclusions and arguments about both our model(s) and data. Since numerous explanatory variables are coded qualitatively, we convert the variables factor levels into binary variables. The F test helps test the joint significance of variables: This proves especially valuable when finding the significance of some our of explanatory variables, such as the season sold, conditional housing proximity to railroad, and quality of the HVAC.

Data

We use the Ames Housing dataset, which describes the characteristics of properties sold between January 2006 and July 2010 in Ames, Iowa. While limited to one city, the data represents one of the largest publicly available collections of data on real estate sales and property characteristics, with 72 variables available for 1440 observations. Limiting geographic scope allows us to assume that typical fundamental determinants of housing prices, including population and real income, stay within a relatively minimal band. The number of yearly transactions is distributed: 619 in 2006; 692 in 2007; 622 in 2008; 647 in 2009; however, there are only 339 transactions for 2010 because only the first 7 months in 2010 were recorded. The

Sales data includes observations beyond normal market sales, including the sale of foreclosed homes and within family sales. Because the effects of foreclosure and family sales on prices are outside the modeling objectives of this paper, we exclude both kinds of sales from final analysis. De Cock (2011) further notes that there may be non-homogenous variances for sale prices, with increasing variation at higher prices; this is typical of prices which are logarithmically distributed. Data transformations, including log scale adjustment of our independent variables are incorporated into this paper.

Above ground area

Each home has a listed housing area in square footage. We anticipate that total square footage would be the largest determinant of housing prices. Because we are interested in percentage changes in square footage, and because we anticipate square footage will have a diminishing marginal effect on home prices, we convert are to a log scale for use in linear regression.

Lot Area

In addition to transforming total above ground area to log scale, we transform total lot area (i.e yard area + ground floor) to log scale as well.

Garage Area

Garage area in square feet is another variable present in the data. Because we don't expect the distribution of garage area to be as right-skewed as those of lot area and total above ground area, there is less of a justification to transform garage area to log scale. This means that our estimated coefficient will likely be nominally small.

Externalities

The variable "Condition" appears in the dataset, with levels describing relevant potentially detrimental neighborhood characteristics for a home. Characteristics including proximity to railroad and proximity to heavy traffic thoroughfares are included as factor levels. To test the hypothesis that the presence of such externalities negatively affects housing prices, we regress home prices on the reported *absence* of externalities (i.e Condition = normal).

Basement Quality

The data contain four levels of reported basement quality; Excellent, Good, Fair, and Poor. Poor is used as a reference level for our regression analysis. Since the data does not include metrics for above ground amenity quality, and since we expect basement quality to be highly correlated with the quality of overall home amenities, basement quality acts as a potential proxy for the overall quality of home fixtures.

Fireplaces

The dataset includes information on the number of fireplaces. Fireplaces are both an aesthetically desirable and utility maximizing housing fixture, since they provide heat. We

include the number of fireplaces in our regression to assess wether this effect is statistically significant.

HVAC Quality

The Ames housing dataset includes an indicator variable representing HVAC quality, with four reported levels. The quality of a homes housing system was either described as Excellent, Fair, Good, or Poor. From these classifications, we identify dummy variables for each level. Regression results are reported using Poor as a reference level.

Seasonal Adjustment

We split home sales into four quarters, with each quarter roughly representing one season. Rather than using fiscal quarters, we aligned the beginning and end of each quarter with the beginning and end of each season (i.e a quarter represents December through February for winter). This adjustment allows us to test the hypothesis that homes prices vary seasonally. We use winter as a reference season, and include the remaining quarters in regression analysis.

Results

OLS multiple linear regression analysis of the above independent variables on housing prices yielded statistically significant relationships between each. In each of the models, the strongest predictor of housing price was, unsurprisingly, total above ground living area, with a 1% increase in square footage associated with an 0.46% increase in sale price ($ß = 0.46$, $p < .001$). Basement quality was further strongly predictive of housing price; an basement rated "Excellent" by appraisers yielded an 36% increase in housing price relative to a basement appraised "Poor," all else equal ($ß = 0.36$, $p < .000$). Evidence for the externality hypothesis appears in our regression results, as homes with "Normal" conditions (i.e no externalities) sold for 8.8% more on average ($ß = 0.088$). Fireplaces seemed a good selling point as well; an additional fireplace was associated with an 8.3% increase in sale prices ($ß = 0.083$). The reported effect size of additional garage space has a very small coefficient in Figures 1-8 ($ß = 0.0004$), but this effect is statistically significant ($P < .001$) and may be nominally significant when interpreted as the effect of an additional square foot, since feet are small (i.e a 100 sqft increase in garage size would correspond to a 4% increase in home value). Lastly, homes with high quality heating appeared to sell for more; homes rated "Excellent" could expect to sell for 11% more than a similar home ($ß = .117$, $P < .001$) with poor heating quality (fireplaces and other amenities held constant). Our model was highly predictive; the included variables explained 80% of the total variation in housing prices ($R^2 = 0.799$). Notably, the inclusion of seasonal adjustments in our data revealed minimal effects on our model ($F = 0.9531$, $F \sim \mathbf{F}_{3,\,1434}$), as reported in Figure 4.

Note: Above results of linear regression are reported from Figures 8.

Discussion

We observe that the square footage has the greatest partial effect on housing price. This result is unsurprising, since the underlying value of a housing stock is strongly determined by the underlying value of the land; more land implies more value. The results of this paper indicate that housing price models should be based around the overall area encompassed by a tract of real estate.

Notably, the strongest non-area determinant of housing value was the quality of the basement. We expect that basement quality is strongly related to the overall quality of amenities; there are likely few houses with "Excellent" basements with relatively poor amenities otherwise. The result that Excellent conditions are associated with a 40% rise in sale price relative to a poor basement are similarly unsurprising. Surprisingly, homes with "Fair" quality basements sell for less, on average, than similar houses with no or poor basements. This effect may be explained by differing preferences among different groups of buyers. We implicitly assume that poorer buyers are more likely to buy (and sell) homes of poorer overall quality. If the lower subset of basement quality is marketed towards a poorer population, they are indifferent to those two basement qualities: The poor population would rather spend money on other housing qualities, thus lowering the value of a marginal increase in basement quality.

The relationship between HVAC quality, fireplaces, and housing prices were similarly unsurprising. Residents in Ames, Iowa (which experienced snow as recently as early April) likely place a high value on the quality of heating fixtures.

We anticipated that seasonal effects would influence housing prices, but this result dis not bear out in our model. Figure 4 reports the inclusion of quarterly adjustments, with winter as a reference quarter; quarterly adjustments are not jointly significant in our model, as evidenced by

an F statistic of less than 1. This means the deals made in winter and spring are not revealing a greater effects on the deal price.

Figure Index

Figure 1

	Dependent variable:
	lnSalePrice
lnGrLivArea	0.877***
	(0.021)
Constant	5.650***
	(0.156)
Observations	1,440
R^2	0.538
Adjusted R^2	0.538
Residual Std. Error	0.272 (df = 1438)
F Statistic	1,674.833*** (df = 1; 1438)
Note:	*p<0.1; **p<0.05;** p<0.01*

Observations	1,440
R^2	0.663
Adjusted R^2	0.662
Residual Std. Error	0.233 (df = 1436)
F Statistic	939.667*** (df = 3; 1436)
Note:	*p<0.1; **p<0.05;** p<0.01*

Figure 4

	Dependent variable:
	lnSalePrice
lnGrLivArea	0.624***
	(0.022)
lnLotArea	0.059***
	(0.013)
GarageArea	0.001***
	(0.00003)
QSold_2	0.004
	(0.021)
QSold_3	0.003
	(0.020)
QSold_4	-0.003
	(0.024)
Constant	6.610***
	(0.163)
Observations	1,440
R^2	0.663
Adjusted R^2	0.661
Residual Std. Error	0.233 (df = 1433)
F Statistic	468.928*** (df = 6; 1433)
Note:	*p<0.1; **p<0.05;** p<0.01*

Figure 5

	Dependent variable:
	lnSalePrice
lnGrLivArea	0.632***
	(0.021)
lnLotArea	0.069***
	(0.013)
GarageArea	0.001***
	(0.00003)
Condition1_Norm	0.156***
	(0.017)
Constant	6.348***
	(0.160)
Observations	1,440
R^2	0.680
Adjusted R^2	0.679
Residual Std. Error	0.227 (df = 1435)
F Statistic	763.222*** (df = 4; 1435)
Note:	*p<0.1;* ***p<0.05;*** p<0.01

Figure 6

	Dependent variable:
	lnSalePrice
lnGrLivArea	0.522***
	(0.019)
lnLotArea	0.099***
	(0.011)
GarageArea	0.0004***
	(0.00003)
Condition1_Norm	0.094***
	(0.015)
BsmtQual_Ex	0.427***
	(0.022)
BsmtQual_Gd	0.213***
	(0.012)
BsmtQual_Fa	-0.140***
	(0.034)
Constant	6.916***
	(0.139)
Observations	1,440
R^2	0.768
Adjusted R^2	0.767
Residual Std. Error	0.193 (df = 1432)
F Statistic	676.452*** (df = 7; 1432)
Note:	*p<0.1;* ***p<0.05;*** p<0.01

Figure 7

	Dependent variable:
	lnSalePrice
lnGrLivArea	0.462***
	(0.019)
lnLotArea	0.082***
	(0.011)
GarageArea	0.0004***
	(0.00003)
Condition1_Norm	0.092***
	(0.015)
BsmtQual_Ex	0.416***
	(0.021)
BsmtQual_Gd	0.210***
	(0.012)
BsmtQual_Fa	-0.126***
	(0.033)
Fireplaces	0.082***
	(0.009)
Constant	7.464***
	(0.148)
Observations	1,440
R^2	0.781
Adjusted R^2	0.780
Residual Std. Error	0.188 (df = 1431)
F Statistic	637.455*** (df = 8; 1431)
Note:	*p<0.1;* ***p<0.05;*** p<0.01

Figure 8

	Dependent variable:
	lnSalePrice
lnGrLivArea	0.442***
	(0.019)
lnLotArea	0.087***
	(0.010)
GarageArea	0.0004***
	(0.00003)
Condition1_Norm	0.088***
	(0.014)
BsmtQual_Ex	0.360***
	(0.021)
BsmtQual_Gd	0.171***
	(0.012)
BsmtQual_Fa	-0.139***
	(0.032)
Fireplaces	0.083***
	(0.009)
HeatingQC_Ex	0.117***
	(0.012)
HeatingQC_Gd	0.053***
	(0.015)
HeatingQC_Fa	-0.106***
	(0.028)
Constant	7.532***
	(0.142)
Observations	1,440
R^2	0.799
Adjusted R^2	0.797
Residual Std. Error	0.180 (df = 1428)
F Statistic	515.005*** (df = 11; 1428)
Note:	*p<0.1;* ***p<0.05;*** p<0.01

References

Wilhelmsson, Mats. March 2008. Journal of Housing Economics, Volume 17, Issue 1, Page 88-01

Panagiotidis, Theodore and Printzis, Panagiotis (2015) *On the macroeconomic determinants of the housing market in Greece: a VECM approach.* GreeSE papers (88). Hellenic Observatory, European Institute, London, UK.