Predicting Housing Prices

By Brendan Peek

Introduction

The goal of the study is to determine which housing characteristics are useful in predicting house prices. Data was gathered on 522 houses and the characteristics collected for each house included price, square footage, number of bedrooms, number of bathrooms, whether the house had air conditioning, size of garage, whether the house has a pool, its age, quality, lot size, and whether the house is close to a highway. Other variables that were thought to be important were created. The first is whether the house has both air conditioning and a pool. This is important because both of these features are valuable in hot weather. For this reason, they would be more valuable together. If a house is in a hot region and doesn't have one of these features then the value of the house should go down compared to having both. The other variable is the ratio of square footage to lot size. As this ratio increases the house takes up more of the lot which shrinks yard space and is expected to decrease the value of the house. More square feet are valuable and a larger lot is valuable but if the house size increases at the expense of the open space then the property is likely to go down compared to a property with a more balanced house size to lot size ratio.

Transformations and Testing Assumptions

The first model that was tested included all the variables and squares of all of the quantitative variables to make sure and capture any nonlinearities on the relationships between the explanatory variables and the prices. After running the model, three desirable properties were tested for. The Shapiro-Wilk test of normality rejected the null of normality with a p-value of < 0.0001 so it was concluded that the data is not normal. Next the Breusch-Pagan test was used to test for constant variance. The null of constant variance was rejected with a p-value of < 0.0001 so the data does not have a constant variance. Finally, the goodness of fit test was used to determine if the model fits the data well. With a p-value of 0.2802 we failed to reject the null of good fit and concluded that the model is a good fit for the data. Since two of the 3 tests failed we decided to try and improve the model.

The next model used the natural log of the price to try to get more desirable statistical properties. This is a common first transformation when dealing with the problems outlined in the paragraph above. The model was run and the same tests were used to test for normality, constant variance, and goodness of fit. All three tests improved. The Shapiro-Wilk p-value increased to 0.0083 which means the data is closer to normal with the log transform than without it. The p-value for the Breusch-Pagan test of constant variance is still very small but the test statistic is less than half of what it was for the non-transformed data which also means it is getting closer to constant variance. While the goodness of fit test concluded that the model was a good fit with non-transformed data the p-value for the log transformed data is even better at 0.5625.

While the log transformation improved the properties of the data, the tests for normality and constant variance still failed. Another common transformation was used to try and fix this. The

square root of price was calculated and the model was run. The properties were tested and they all were worse than the log transformation. The Shapiro-Wilk p-value was < 0.0001. The Breusch-Pagan p-value was < 0.0001. And the goodness of fit p-value decreased to 0.4196 which still concludes that the model is a good fit but it is not as strong as the result for the log transformed data.

When dealing with problematic data like this where the common transformations do not work the Box-Cox power transformation can help. The process was run and the suggested transformation for price was to raise price to the -0.3 power. The results from the model with this transformation were the best of the group. The Shapiro-Wilk test concluded the data was normal with a p-value of 0.1213. The goodness of fit test had the strongest evidence of good fit with a p-value of 0.6364 which is the highest of all the models run. The Breusch-Pagan test improved with the statistic shrinking further to 55.4 but it still concluded that the variance was not constant.

This was the best it was going to get with the methods available and normality is the most important property anyway so we decided to move forward with the transformation suggested by the Box-Cox method.

Model Selection

With the data transformation selected it was time to figure out which variables were the most useful for predicting the price of a house. Two selection procedures were used. The first procedure was stepwise selection. This method adds variables to the model one at a time and tests the explanatory power of each variable in the model in the presence of all of the other variables on each iteration. The resulting model is,

$$\widehat{bc.price} = \beta_0 + \beta_1 sqft + \beta_2 age + \beta_3 lot + \beta_4 high.qual + \beta_5 med.qual + \beta_6 bath + \beta_7 ac + \beta_8 cars + \beta_9 ac.pool + \beta_{10} pool + \beta_{11} sqft.lot$$

With this model data only needs to be gathered on 8 characteristics of each house. The next model selection method used was Adjusted R^2. This method tests whether adding an additional variable to the model is worth it given the extra explanatory power added by the new variable. This method suggested the model,

$$\widehat{bc.price} = \beta_0 + \beta_1 sqft + \beta_2 sqft^2 + \beta_3 bed^2 + \beta_4 bath + \beta_5 bath^2 + \beta_6 ac + \beta_7 cars + \beta_8 pool + \beta_9 age + \beta_{10} age^2 + \beta_{11} lot + \beta_{12} lot^2 + \beta_{13} highway + \beta_{14} med.qual + \beta_{15} high.qual + \beta_{16} ac.pool + \beta_{17} sqft.lot$$

The only variable not included in the above model is number of bedrooms but because bedrooms squared is included we need to include bedrooms which gives the full model. The goal of this model selection process was to find a model that balances efficiency and explanatory power. We ran both suggested models and looked at their respective adjusted R^2 values to see if the simpler model might be good enough. R^2 is a measure of how well the model explains the variation in price. The simpler model suggested by stepwise selection has an R^2 of 0.8314 while the full model suggested by the Adjusted R^2 selection process has an R^2 of 0.8597. The difference in explanatory power between the two models is negligible and the difficulty and cost involved in collecting the extra data points may not be worth it to get that little bit extra. For that reason, we settle on the simpler model for the predictions of interest.

The final model with the estimated parameters is,

$$\begin{aligned}\widehat{bc.price} = \; & 0.02847 - 0.00000195sqft + 0.00002708age - 0.000000032lot \\ & + 0.0023high.qual + 0.00158med.qual - 0.00035bath - 0.00043ac \\ & - 0.00025cars - 0.0044ac.pool + 0.0039\beta_{10}pool + 0.00287sqft.lot\end{aligned}$$

The signs look opposite of what we would expect but that is a product of the transformation we used to get the desired properties earlier in the analysis. When the value of the transformed price shrinks the untransformed dollar value gets larger. For this reason, increases in good characteristics like square footage, lot size, garage size, and ac and pool have a negative effect on the transformed price which results in an increase in the untransformed price as we would expect. With this in mind the signs of the parameters are what we would expect except for the values on the quality parameters. The increase in quality appear to have a negative effect on price but that may be because the other variables already explain the positive increases that would come with a higher quality house. The increase in quality is probably paired with being a larger house on a larger lot with more bathrooms and other positive characteristics already covered by the model.

With the simpler model we predicted the prices of the two houses of interest and provided 95% prediction intervals for both. For a new observation of price on a medium quality 1,950 square foot house with 4 bedrooms, 3 bathrooms, air conditioning, a 2 car garage, and no pool, that was built in 1980 on a 25,000 square foot lot and not near a highway we predict a value for the transformed price of 0.0242 with a 95% prediction interval of (0.0217, 0.0266). Reversing the transformation (by raising the value to the -10/3 power) to get the price in dollars gives a predicted value of $243,939.49. We are 95% confident that a new observation with the above characteristics will have a price that falls in the interval ($177,989.57, $350,860.29).

For a new observation of price on a low quality 3,500 square foot house built in 1975 on a 10,000 square foot lot with 3 bedrooms, 1 bathroom, no air conditioning, a 1 car garage and a pool that is not near a highway we predict a value for the transformed price of 0.0262 with a 95% confidence interval of (0.0225, 0.0299). Reversing the transformation to get the price in dollars gives a predicted value of $187,209.99. We are 95% confident that a new observation with the above characteristics will have a price that falls in the interval ($120,530.24, $310,974.62).

These prediction intervals are likely wider than a confidence interval because they deal with predictions of single new observations instead of the means of samples estimating the population mean price of houses with those particular characteristics.

Conclusion

For the purpose of predicting house prices reasonably well while minimizing the need for gathering large amounts of a variety of data the above model will work well. It is still important to consider the output of this model as a starting point only. The width of the confidence intervals for the predicted prices in the last section show that there is still a need for local expertise to narrow the range of reasonable prices. The complementary inputs of expert opinion and objective values from the model should result in stronger, more accurate valuations for single family homes.

Appendix: Code

```
data housing;
input price sqft  bed   bath  ac    cars  pool  year  quality     style lot
      highway;
cards;
…
. 1950 4 3 1 2 0 1980 2 0 25000 0
. 3500 3 1 0 1 1 1975 1  0 10000 0
run;
proc means data=housing;
run;


data housing2;
set housing;
age = 1998-year;
ac_pool = ac*pool;
sqft_lot = sqft/lot;
if quality=2 then med_qual = 1;
else med_qual = 0;
if quality=3 then high_qual = 1;
else high_qual = 0;
log_price = log(price);
root_price = sqrt(price);
sqft2 = sqft*sqft;
bed2 = bed*bed;
bath2 = bath*bath;
lot2 = lot*lot;
age2 = age*age;
bc_price = price**(-0.3);
run;
/*full model with no transformation - test assumptions */
```

```
proc reg data=housing2;

model price = sqft sqft2 bed bed2 bath bath2 ac cars pool age age2 lot lot2
highway med_qual high_qual ac_pool sqft_lot / lackfit;

output out=myout r=res p=yhat;

run;

proc univariate normal plot;

var res;

run;

proc model;

parms beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7 beta8 beta9 beta10
beta11 beta12 beta13 beta14 beta15 beta16 beta17 beta18;

price = beta0+beta1*sqft+ beta2*sqft2+ beta3*bed+ beta4*bed2+ beta5*bath+
beta6*bath2+ beta7*ac+ beta8*cars+ beta9*pool+ beta10*age+ beta11*age2+
beta12*lot+ beta13*lot2+ beta14*highway+ beta15*med_qual+ beta16*high_qual+
beta17*ac_pool+ beta18*sqft_lot;

fit price /breusch = (1 sqft sqft2 bed bed2 bath bath2 ac cars pool age age2
lot lot2 highway med_qual high_qual ac_pool sqft_lot);

run;

quit;

/* full model with log transformation - test assumptions */

proc reg data=housing2;

model log_price = sqft sqft2 bed bed2 bath bath2 ac cars pool age age2 lot
lot2 highway med_qual high_qual ac_pool sqft_lot / lackfit;

output out=myout2 r=res2 p=yhat2;

run;

proc univariate normal plot;

var res2;

run;


proc model;

parms beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7 beta8 beta9 beta10
beta11 beta12 beta13 beta14 beta15 beta16 beta17 beta18;

log_price = beta0+beta1*sqft+ beta2*sqft2+ beta3*bed+ beta4*bed2+ beta5*bath+
beta6*bath2+ beta7*ac+ beta8*cars+ beta9*pool+ beta10*age+ beta11*age2+
beta12*lot+ beta13*lot2+ beta14*highway+ beta15*med_qual+ beta16*high_qual+
beta17*ac_pool+ beta18*sqft_lot;
```

```
fit log_price /breusch = (1 sqft sqft2 bed bed2 bath bath2 ac cars pool age
age2 lot lot2 highway med_qual high_qual ac_pool sqft_lot);

run;

quit;


/* full model with square root transformation - test assumptions */

proc reg data=housing2;

model root_price = sqft sqft2 bed bed2 bath bath2 ac cars pool age age2 lot
lot2 highway med_qual high_qual ac_pool sqft_lot / lackfit;

output out=myout3 r=res3 p=yhat3;

run;

proc univariate normal plot;

var res3;

run;

proc model;

parms beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7 beta8 beta9 beta10
beta11 beta12 beta13 beta14 beta15 beta16 beta17 beta18;

root_price = beta0+beta1*sqft+ beta2*sqft2+ beta3*bed+ beta4*bed2+
beta5*bath+ beta6*bath2+ beta7*ac+ beta8*cars+ beta9*pool+ beta10*age+
beta11*age2+ beta12*lot+ beta13*lot2+ beta14*highway+ beta15*med_qual+
beta16*high_qual+ beta17*ac_pool+ beta18*sqft_lot;

fit root_price /breusch = (1 sqft sqft2 bed bed2 bath bath2 ac cars pool age
age2 lot lot2 highway med_qual high_qual ac_pool sqft_lot);

run;

quit;

proc transreg data=housing2 test;

      model boxcox(housing2 / lambda= -1.5 to 1.5 by 0.1) = identity(sqft
sqft2 bed bed2 bath bath2 ac cars pool age age2 lot lot2 highway med_qual
high_qual ac_pool sqft_lot);

run;

proc reg data=housing2;

model bc_price = sqft sqft2 bed bed2 bath bath2 ac cars pool age age2 lot
lot2 highway med_qual high_qual ac_pool sqft_lot / lackfit;

output out=myoutBC r=resBC p=yhatBC;

run;

proc univariate normal plot;
```

```
var resBC;

run;

proc model;

parms beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7 beta8 beta9 beta10
beta11 beta12 beta13 beta14 beta15 beta16 beta17 beta18;

bc_price = beta0+beta1*sqft+ beta2*sqft2+ beta3*bed+ beta4*bed2+ beta5*bath+
beta6*bath2+ beta7*ac+ beta8*cars+ beta9*pool+ beta10*age+ beta11*age2+
beta12*lot+ beta13*lot2+ beta14*highway+ beta15*med_qual+ beta16*high_qual+
beta17*ac_pool+ beta18*sqft_lot;

fit bc_price /breusch = (1 sqft sqft2 bed bed2 bath bath2 ac cars pool age
age2 lot lot2 highway med_qual high_qual ac_pool sqft_lot);

run;

quit;

/*************************************************************************
**********/

/* model selection procedures */

/*************************************************************************
******/


proc reg data=housing2;

model bc_price = sqft bed bath ac cars pool age lot highway med_qual
high_qual ac_pool sqft_lot / selection = stepwise;

run;

/*stepwise gives 11 variables: sqft age lot high_qual med_qual bath ac cars
ac_pool pool sqft_lot */

proc reg data=housing2;

model bc_price = sqft sqft2 bed bed2 bath bath2 ac cars pool age age2 lot
lot2 highway med_qual high_qual ac_pool sqft_lot / selection = adjrsq;

run;

/*adjR2 gives 17 variables: sqft sqft2 bed2 bath bath2 ac cars pool age age2
lot lot2 highway med_qual high_qual ac_pool sqft_lot */

proc reg data=housing2 alpha=0.05;

model bc_price = sqft age lot high_qual med_qual bath ac cars ac_pool pool
sqft_lot / lackfit cli;

output out=myoutFinal r=resFinal;

run;
```

```
proc univariate normal plot;

var resFinal;

run;


proc model;

parms beta0 beta1 beta2 beta3 beta4 beta5 beta6 beta7 beta8 beta9 beta10
beta11 beta12 beta13;

bc_price = beta0+beta1*sqft+ beta2*age+ beta3*med_qual+ beta4*high_qual+
beta5*lot+ beta6*cars+ beta7*pool+ beta8*ac+ beta9*bath+ beta10*highway+
beta11*ac_pool+ beta12*sqft_lot beta13*bed;

fit bc_price /breusch = (1 sqft age med_qual high_qual lot cars pool ac bath
highway ac_pool sqft_lot bed);

run;

quit;
```