Residential Home Prices as Affected by Home Characteristics

By David Joe, Kaitlyn Boyle and Fiona Fu

PSTAT 126 Winter 2018

Introduction

In order to predict the residential home sales prices in a midwestern city, we want to find out which home characteristics are most influential. We initially have twelve variables and using statistical methods we want to find out which of the twelve variables are most influential in predicting home prices. Our twelve initial variables are:

1. Sales price: sales price of residence measured in dollars
2. Finished square feet: finished area of residence measured in square feet
3. Number of bedrooms: total number of bedrooms in residence
4. Number of bathrooms: total number of bathrooms in residence
5. Air conditioning: presence or absence of air conditioning; 1 if yes, 0 otherwise
6. Garage size: number of cars that garage will hold
7. Pool: presence or absence of swimming pool; 1 if yes, 0 otherwise
8. Year built: year property was originally constructed
9. Quality: index for quality of construction: 1 indicates high quality, 2 indicates medium quality, 3 indicates low quality
10. Style: qualitative indicator of architectural style
11. Lot size: lot size measured in square feet
12. Adjacent to highway: presence or absence of adjacency to highway: 1 if yes, 0 otherwise

Of these twelve variables, we have five categorical variables: air conditioning, pool, quality, style, and adjacent to highway. The other variables are all numeric. The data were collected based off the sales of 521 homes in a midwestern city in 2002.

Question of Interest

We want to find out what variables are most influential in predicting the inevitable transaction prices of purchasing a home in a midwestern city. Based on the data we have, we want to find out which of the twelve variables have a significant effect on the sales price of a residential home.

Regression Method

To answer the question of which variables will have a significant effect on home sales prices, we will use a multiple linear regression model with eleven predictors. The multiple linear regression model is as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i$$

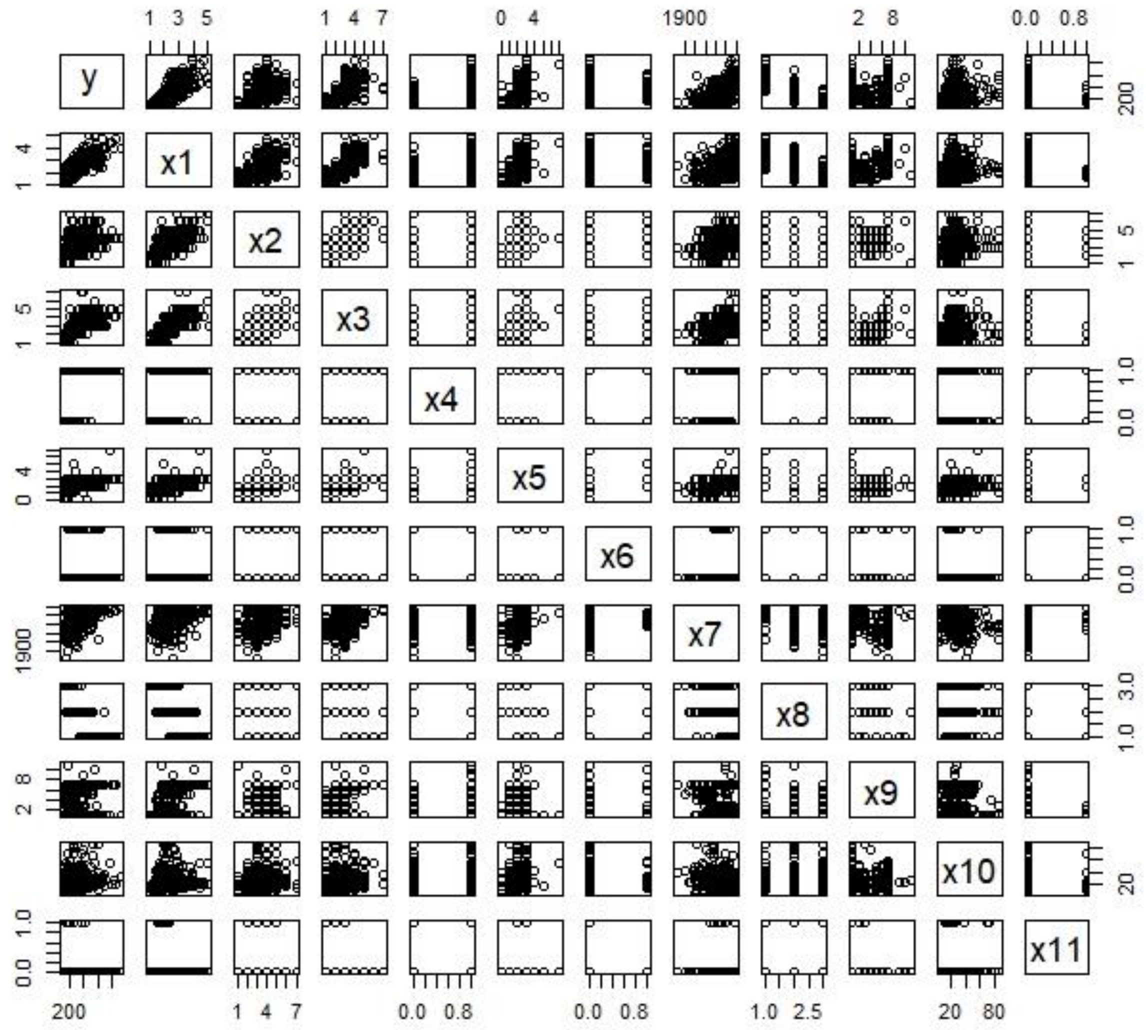
This is our population model that we want to predict. The assumptions associated with this model are that the ε_i are independent and normally distributed with mean 0 and variance σ^2, β_i are coefficients and β_0 is the y-intercept. For our prediction values, the model becomes:

$$Y^{\wedge}_i = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 + b_{10}x_{10} + b_{11}x_{11}$$

Where y = home sales prices, x_1 through x_{11} represent the predictor variables and Our residual error term will be $e_i = Y_i - Y^{\wedge}_I$, which is the difference between the actual value and its predicted value. Through this regression model, we wish to find the best model for predicting Y, meaning which predictor variables b_ix_i will best predict the home sales prices. We are wanting to predict home sales prices based on the eleven other predictor variables.

Regression Analysis, Results, and Interpretation

For our model and testing, we use the R software. In finding the best model for our prediction of home sales prices, we first used the pairs() function on all of our variables to see if we needed to use transformations based on the LINE conditions.



Based on the pairs() function, we decided to use BoxCox transformations to transform our response variable, sales prices. We found that the optimal transformation for this variable was a log transformation. We then used the powerTransform() function to find the best transformations of our other variables. We found that the only predictor variable that needed transformation was square feet, and the optimal transformation was the negative square root. Using this information and our newly transformed variables, we then used best subsets regression to find our best multiple linear model. Once we did this, we used Bayesian Information Criterion to find which of our predictor variables were most significant. We found that BIC limits the number of parameters for square feet, quality, style, year, and lot size, which are the most significant predictor variables in predicting the response, home sales price. This means that will be exclude the variables pool, highway proximity, number of bedrooms, number of bathrooms, and air conditioning in our new model.

Based on the BIC and looking at the data set, we decided not to use certain variables in our new model. The reason we chose not to include pool is because in a midwestern city, pools are not very common. In fact, only 7% of the homes in our data set had pools. For this reason, the predictor variable pool had little effect on sales prices and therefore was best left out of the new model. In the Midwest, the majority of homes are not near major highways: in our data set, only 11 homes out of 521 homes were near a highway, and therefore the predictor variable highway proximity could be left out of the new model. In deciding whether to include number of bedrooms and number of bathrooms, according to the best subsets regression, it was shown that these predictor variables did not have a significant effect on predicting home sales prices. This could be because square footage is the best predictor of home price and having number of bedrooms and number of bathrooms is extraneous and unnecessary to our model. Lastly, it was shown that air conditioning did not have a significant impact on home sales prices. This could be because only 17% of the homes in our data set had air conditioners, and the average annual high temperature in the general Midwest is around 60° Fahrenheit. For these reasons, our new model included square feet, quality, style, year, and lot size as predictor variables for the response variable of home sales prices.

Once we found our best linear model, we tested for interaction terms. We used F tests on our predictor variables to find which ones had a significant interaction. It was found that the interaction between square feet and year built was significant, as well as the interaction between quality and style. Therefore, our final model included a log transformed response variable of sales prices, a negative square root transformed predictor variable of square feet, and the predictor variables quality, style, year built and lot size, with significant interaction between square feet and year built as well as between quality and style. This is our best model so far with an optimized number of parameters. We then tested the normality of our fitted model by plotting the residuals and a normal Q-Q line. We found that the residuals tended to bounce randomly along 0, and that most of the points were on the normal Q-Q line with the exception of a few outliers.

Knowing that we had a few outliers, we then needed to analyze them. Using Cook's Distance, we found that there were no influential points where cd > 0.5. From here, we used DFFITS to analyze the outliers. Using the criterion, we found that points 513, 11, 202, and 24 have both internally and externally studentized values greater than 3. We then tested to find high leverage points. We found that where internally studentized residuals were greater than three, points 202 and 11 were significant. When we tested where DFFITS were greater than the criterion, we found that the points 202, 96, and 74 were high leverage values. The point 202 was a high leverage value based on both tests. Based on this, our points of interest were 202, 96, 74, and 11.

Conclusion

Our research question aimed to find out which predictor variables were significant for predicting the final sales prices of homes in a midwestern city. Using the methods of transformations, best subset regression, testing for interaction terms, and analyzing and removing outliers, we found that the best multiple linear regression model for prediction of home sales prices in a midwestern city included square feet, year built, quality, style, and lot size. By beginning with checking the LINE assumptions, we found that transformations of certain variables were necessary, which helped us to find a better model. After this, using best subset regression and testing interaction terms, we were able to come up with an optimized model for best predicting the home sales prices. Finally, finding, analyzing, and removing outliers allowed us to have the best possible model for prediction of home sales prices. We found that our best model did not include certain parameters, and based on the BIC, we were able to see why it was best to leave these predictor variables out, and why certain variables were more significant in predicting final home prices. The methods described in this report aided us in finding our best possible model for predicting final home prices based on our data.