Hanna Grossman

Modeling House Sale Price in Ames, Iowa

Abstract:

A number of factors, including size, year built, and neighborhood go into predicting what price a house will sell for. For this research study, we analyze housing prices in Ames, Iowa, using a dataset with 2,500 observations, each representing a house, and 81 variables containing information about each house. The goal of this research study is to create a valid multiple linear regression model to predict the prices of houses in Ames, Iowa.

This model was then used to predict sale prices for a testing data set, and these predictions were submitted to Kaggle to be analyzed for accuracy. My kaggle name for this project was "Hanna Grossman Lec 1". The R2 of my model using the training data was 0.9105, while the R2 of my testing model was 0.9217. My Kaggle final rank was 71, including lectures 1 and 2, with a valid model containing 6 predictors and an interaction term (11 betas total).

Introduction:

We began this project with a training data set containing 81 variables, and 2,500 observations. This data contains information about houses in Ames, Iowa, including predictors such as the neighborhood, year built, year remodeled, and areas of each house. This data set was originally created by Dean De Cock, but edited to match column names afterwards.

After reading this data into R Studio, I began with an exploratory analysis of the data. Here I observed the data's dimensions and began to understand what types of predictor variables were present. Then, I looked at where missing values were present in the data and filled in the ones that were supposed to be 0s. This left me with 933 missing values in the train data set, which I then used Mice, an R package, to impute. After this, I decided to add a few new variables to the data set that I believed may be good predictors. This included a variable age, or year sold minus the year built, and a variable total area, or a sum of the important areas in each house. I also created a variable called neighborhood class, which separated the variable

neighborhood into four new categories based on the prices of houses in each area. In addition, for variables like age and basement square feet, I changed 0s into 1s in order to allow for potential transformation. This did not change the model or predictions, as having a basement of 0 or 1 square foot, or having a house 0 or 1 years old, does not change how much one would pay for the house. This ultimately left we with a new dataset, free of missing values, and with 2,500 observations and 89 variables after transformations.

Methodology:

With this new data set, I began exploring the potential predictors in order to create the ultimate model. Through this, I used step and regsubset functions to see which predictors minimized AIC and BIC, while maximizing the adjusted R2. After finding my chosen predictors through this process, I then used inverse response plot and power transform to see which of my numerical predictors, and my response variable, should be transformed. Through this, I decided to transform Sale Price, Total Area, Age, Lot Area, and Basement Square Footage 1 using the suggested lambdas. After transforming my variables, I looked at the diagnostic plots of my model.

As seen below, the first plot shows that the errors are randomly distributed as the red line has a relative slope of zero. In addition, the second plot shows that the model is approximately normal, besides perhaps a few points in the left hand corner. The third diagnostic plot shows that the variances are fairly constant. In addition, there are 176 bad leverage points observed. Although this may seem like a large amount, as the data contains 2,500 observations, this is only about 7% of the data. I did minimize this number my adding variables like age that helped to explain low priced houses, as this variable is negatively correlated with sale price. In addition, the VIF for each predictor in the model is well below 5, and therefore there is not a multicollinearity problem present.

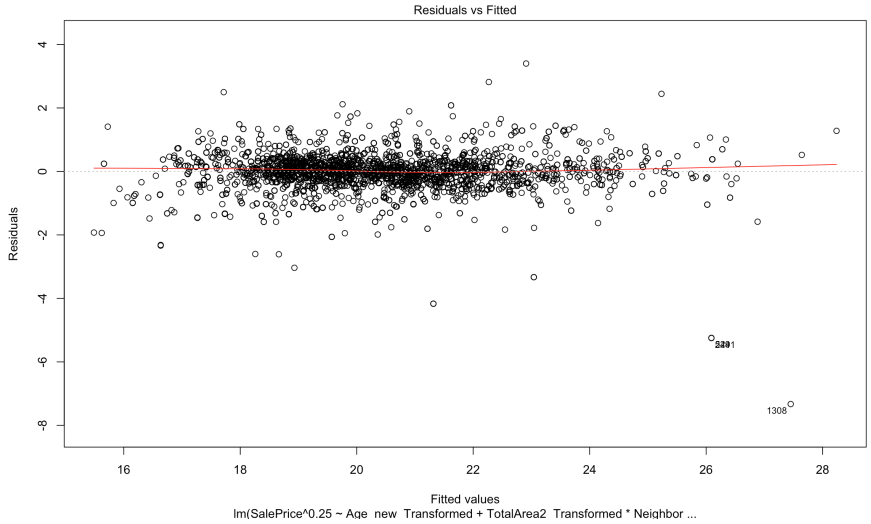

Fig 1: Diagnostic Plot 1 - Fitted Values vs Residuals

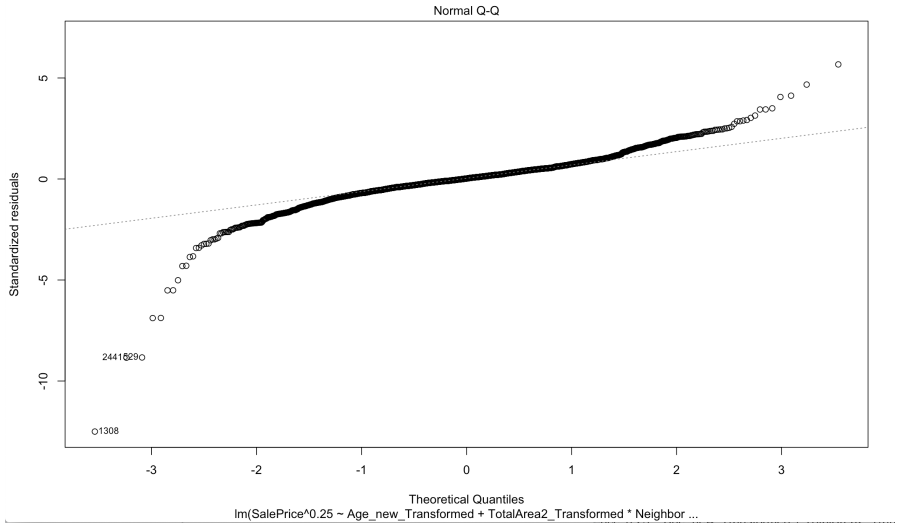

Fig 2: Diagnostic Plot 2 - QQ Plot

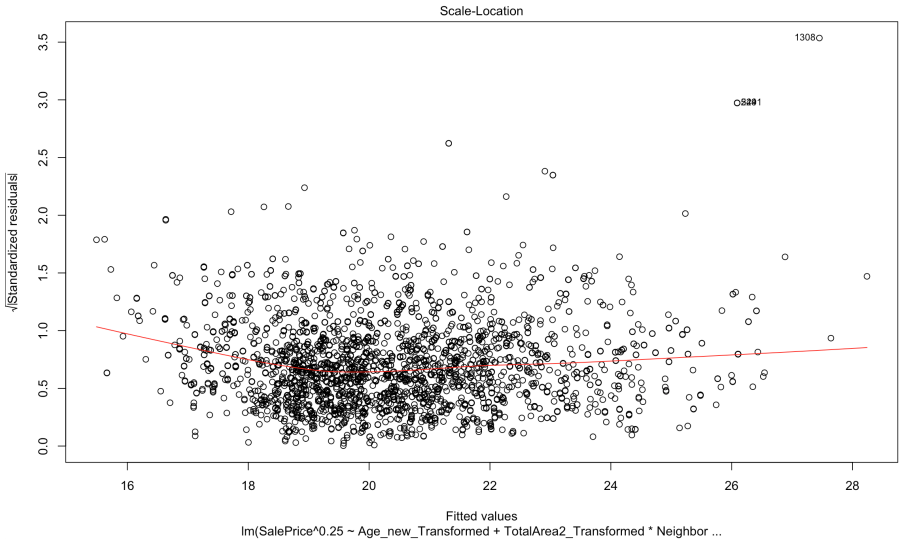

Fig 3: Diagnostic Plot 3 - Fitted Values vs Square Root of Standardized Residuals

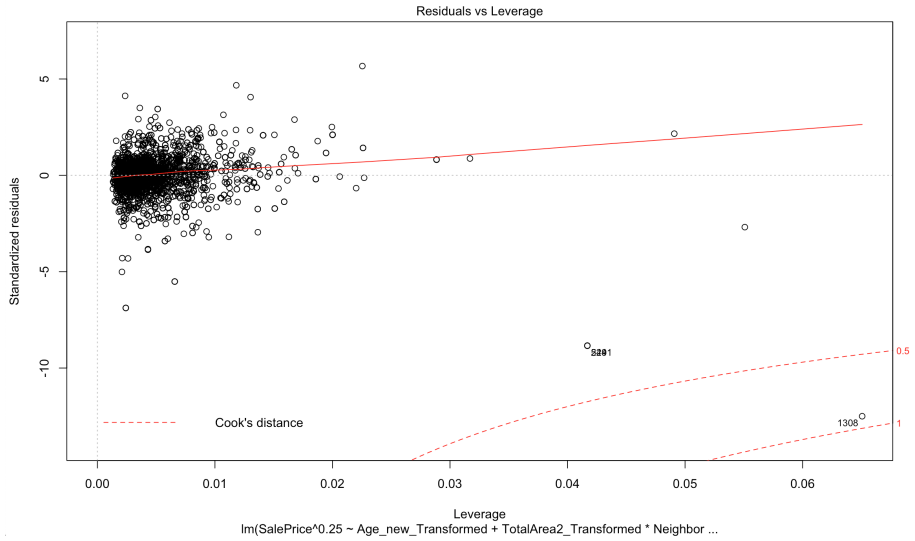

Fig 4: Diagnostic Plot 4 - Leverage vs Standardized Residuals

After analyzing the validity of the model, I then analyzed the correlation of each predictor to the other predictors. This allowed me to see if any interaction terms should be added. In these additions, I considered which interaction terms were both statistically and practically significant, as the adjusted R2 needed to improve the model in order for the addition to be worth it. In the end, I added an interaction term between Total Area and Neighborhood Class, which improved the model from an R2 of 0.9036 to an R2 of 0.9105. With the addition of this interaction term, this completed my model. From here, I used the step functions and regsubsets

one last time to make sure it was best to keep all of the variables and the interaction term in my model. Finally, I used the predict function in R to predict sale price for the test data, and then uploaded these predictions onto Kaggle.

My Model:

```
model_final3 <- lm(formula = SalePrice^.25 ~  Age_new_Transformed  +
TotalArea2_Transformed *  Neighborhood_Class  + OverallQual
+LotArea_Transformed+ BsmtFinSF1_New_Transformed, data = train_imp)
```

Results:

This model created with age, total area, neighborhood class, overall quality, lot area, and basement area square footage as predictors, in the end received an R2 of 0.9105 for the training data and 0.9217 for the test data. This means that the model explains the variation in the sale price extremely well. In addition to receiving a high R2, the model also produced an adjusted R2 of 0.9101 for the training data. As this value is extremely close to the R2, this shows that the data is likely not overfit by having too many predictors. This model also is valid, with randomly distributed errors, approximately normal data, and approximately constant variance in errors. In conclusion, this model is valid and performed well in predicting housing prices for houses in Ames, Iowa, with an R2 of 0.9105 for the training data, and 0.9217 for the test data.

Discussion:

In summary, we successfully achieved our goal of the study to create a model that predicted house prices in Ames, Iowa using predictors about the houses. Imputing missing values and adding new potentially useful variables allowed us to analyze and get the most out of our data. The step and regsubset functions allowed us to see which predictors would be useful in our model. Then, the inverse response plot and power transform functions allowed us to transform our numerical variables in order to create a valid model. Finally, the addition of an interaction term allowed us to explain interactions in our predictor variables, thus raising our R2.

In the future, additional variables could be added to this model to further explain low housing prices. This would reduce our number of bad leverage points, and therefore improve our predictions. However, when trying to do this, there was no clear variable, other than age, that lowered the number of bad leverage points by a significant amount. Perhaps using the data to create a new predictor would be a better approach in future studies.

In addition, future studies could look at other cities across the U.S. and abroad to see how important predictors change across the globe, in addition to which variables hold constant regardless of location.

Limitations and conclusions:

It is important to note that as this model was created using data collected from only Ames, Iowa, certain variables may be less or more important depending on the location. For example, a fireplace might be much more important somewhere cold, than somewhere that stays somewhat warm year round lik Los Angeles. Therefore, if looking at housing prices, each variable's correlation to the house price should be explored again before creating a model.

In addition, this model did contain 176 bad leverage points. Although this is only 7% of the observations, this could be reduced by adding a predictor that helps to explain low priced houses.

In conclusion, through this model, age, total area, neighborhood class, overall quality, lot area, and basement area square footage predicted sale price of houses in Ames, Iowa with an R2 of 0.9105 for training data, and 0.9217 for test data. The diagnostic plots reveal that the model is valid, although the number of bad leverage points could be minimized. Therefore, we have completed our goal of creating a valid model to predict housing price in Ames, Iowa using the given set of predictors.

References:

Almohalwas, Akram. "Introduction to Data Analysis and Regression." *Course: Introduction to Data Analysis and Regression*, ccle.ucla.edu/course/view/19W-STATS101A-1?section=0.

Almohalwas, Akram. "Stats 101a Lectures." 2019, Los Angeles , California.

"Mice." *Function | R Documentation*,
www.rdocumentation.org/packages/mice/versions/3.4.0/topics/mice.

"Sale Price of Houses in Ames, Iowa." *Kaggle*, www.kaggle.com/c/stat101ahouseprice/data.

Sheather, Simon J. *A Modern Approach to Regression with R*. Springer, 2009.