IBM Machine Learning Week 2 Peer-graded Assignment: Course Project - Peer Review

Finding Significant Predictors in the Ames Housing Data

By Thomas Arnold

Overview

I wanted to learn more about the factors that affect housing prices. We have a house on the Mississippi river in Little Falls MN. I have been doing some remodeling work and wondered if I will ever see a return on my investment of time and resources. We added a gas fireplace. Was that a good investment? We also purchased an adjoining lot and I am interested in determining whether lot size affects the sale price. As we upgrade the quality of the house by adding better quality windows, etc., will that help?

One of the course datasets contained real estate data from Ames Iowa. This project seemed like a good test of my newly acquired statistical powers.

Brief description of the data set and a summary of its attributes

The Ames Iowa Real Estate Data

I am working with the Ames Iowa real estate dataset (De Cock, 2011). This dataset was provided by the Ames Iowa assessor and partially cleaned by De Cock (2011). The dataset has a variety of data related to real estate along with sale prices.

Based on the df.info(), the Ames dataset contains 2,930 rows and 81 columns. If I run df.dtypes.value_counts(), I find there are 43 object columns, 28 integer columns, and 11 float columns. Running df.isna().any().sum() tells me that there are 27 columns with null values.

Initial plan for data exploration

I used a variety of techniques for examining the data. I started with the describe function. I then tried plotting the sale price and lot area values.

The Describe Function

My first attempt to look at the data involved the describe() function. I found that the describe() function does not return all columns. There were two problems, 1) the default setting returns numeric columns only, and 2) the output is truncated due to the large number of columns.

I resolved the first issue (only numeric columns) by using describe(include='all'), which returns both numeric and non-numeric columns. I resolved the second issue (no all columns being displayed) by adding iloc sections (i.e. df.iloc[:, 2:12], df.iloc[:, 12:22]) which could be used to limit the numbers of columns to display. I then looked at 10 columns at a time.

For example, this provides the first 10 data columns.

```
df.iloc[:, 2:12].describe(include='all')
```

Plotting the Sale Price

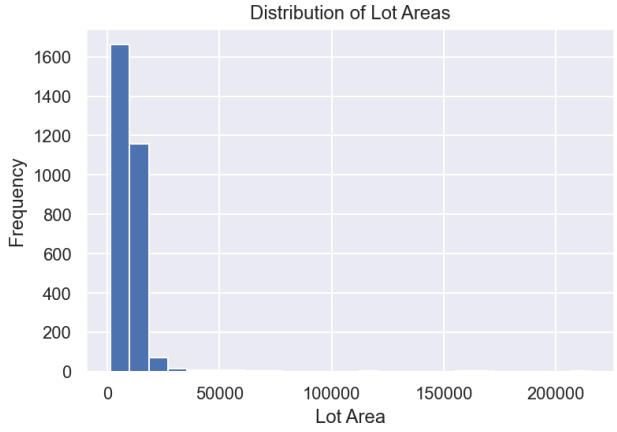
The next step was to plot the sale prices to see if the distribution was symmetric. It appeared that the prices had a positive skew.

Distribution of Sale Prices

Plotting the Lot Areas

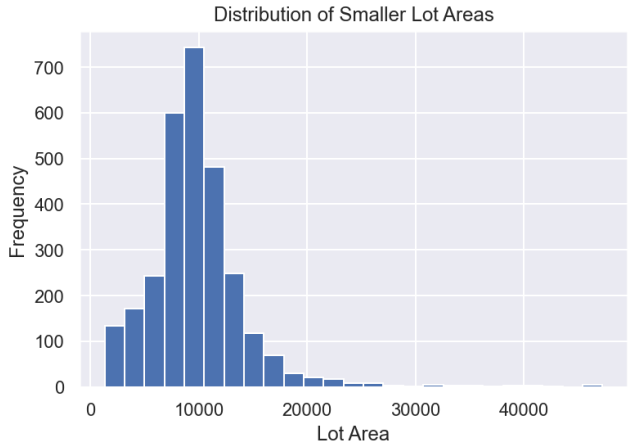
I then looked at the plot of the lot areas. These were even more positively skewed than the sale prices.

Distribution of Lot Areas



I removed the values over 50,000 and the data looked a little more symmetrical. I am not sure how to fix this.

Distribution of Smaller Lot Areas



Lot Areas vs Sale Price

In looking at the lot areas vs sale price, the data seemed to be highly skewed. There did not seem to be a strong relationship between lot size and sale price.

Lot Area vs Sale Price

The get_dummies Function

I tried the get dummies function on the entire dataset as suggested in the feature engineering module. While this function works to create dummy variables, it created 265 variables that I could not really understand. This might be a useful function when used with one variable at a time.

Actions taken for data cleaning and feature engineering

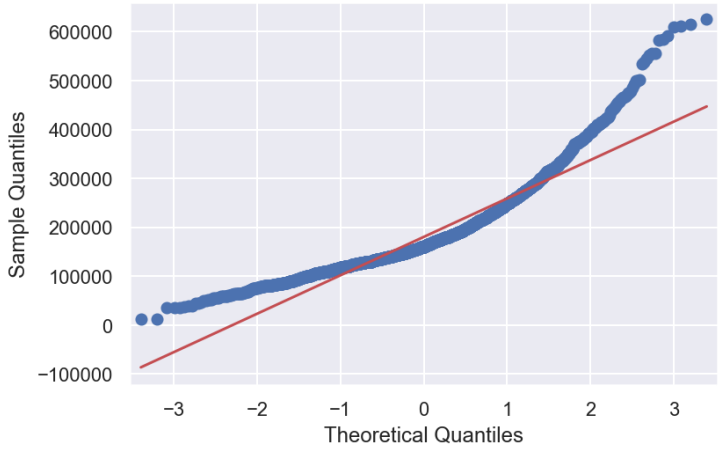
The next step was to clean the data. Since I was not interested in the columns with missing data, I did not need to worry about those. I was interested determining how these three variables affected sale price.

1. Fireplaces
2. Home Quality
3. Lot Size

Adjusting the Dependent Variable (Sale Price)

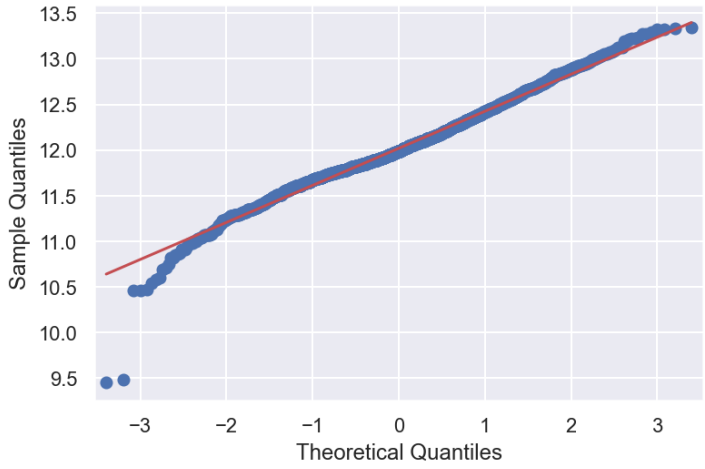
My output was sale price. However, it turned out that the sale price was not normally distributed. This would cause problems for my hypothesis tests, which were designed for normally distributed variables. When I created a QQ plot of the sale price, I found that the sale price was decidedly not normal.

QQ Plot of Sale Price



To fix this problem, I created a new variable LogPrice by taking the log of the SalePrice variable. The QQ plot for the LogPrice variable indicated that the LogPrice variable was more normally distributed.

QQ Plot of Log Sale Price



Adjusting the Independent Variables

The independent variables were created by adding dichotomous variables to the dataset. These were set up as follows. A house with one or more fireplaces had the variable HasFireplace set to 1. The quality score ranged from 0-10. Houses with a quality score from 6-10 had the HighQuality variable set to 1. Lots with areas above the median lot size had the BigLot variable set to 1.

Independent Variables

Variable	0 Rule	1 Rule	Mean
HasFireplace	No fireplace	One or more fireplaces	51.4%
HighQuality	0-5 quality score	6-10 quality score	62.1%
BigLot	Lot area <= median	Lot area > median	50.0%

Key Findings and Insights

There were a few things I learned.

I found that the sale price was skewed positive and the log of the sale price might look a little more normal. I decided to use the LogPrice variable for my dependent variable.

I found that the get_dummies function works properly. It gets rid of the original variable and it creates one less variable than there are unique values. However, I don't like the number of variables that it creates when it runs in automatic mode. The get_dummies function probably should probably be used sparingly.

I found that there are a few lot areas that are very large. When looking at the scatter plot between lot area and price, it seemed that some dwellings with large lots still had lower prices.

I was able to create three dichotomous variables for hypothesis testing.

1. FireplaceInHome
2. HighQuality
3. BigLot

Testing My Three Hypotheses

I predict that all three of the variables I created will be significantly associated with higher sale prices.

I set up the null hypothesis (H0) and the experimental hypothesis (H1).

- H0: Homes with fireplaces are not more expensive
- H1: Homes with fireplaces are more expensive

I set up the hypothesis test as shown in the 01e_DEMO_Hypothesis_Testing.ipynb notebook from the IBM ML course. They suggest looking at the upper and lower confidence limits for the 95% confidence level. The CI limits are computed using the following formula.

the mean (mu) +/- 1.96 * StdDev (sigma)

If the confidence intervals overlap, we fail to reject the null hypothesis.

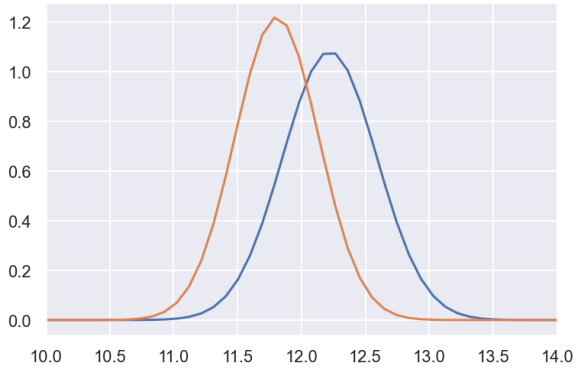
I came up with the following results. None of the three variables was significant predictor of log sale price. The confidence intervals overlapped in all three cases.

Results from Significance Tests

Sample	Mean	Std Dev	Upper 95% CI	Lower 95% CI	Result
With Fireplace	12.221	.369	12.945	11.498	Overlaps
Without Fireplace	11.807	.327	12.449	11.166	Fail to reject
With HighQuality	12.212	.340	12.879	11.545	Overlaps
Without HighQuality	11.705	.293	12.279	11.131	Fail to reject
With BigLot	12.166	.409	12.967	11.365	Overlaps
Without BigLot	11.874	.346	12.552	11.195	Fail to reject

The plots of the price distributions all indicated a high degree of overlap. The plot for the HasFireplace test is shown below.

Price Distributions for HasFireplace 0 and 1



I don't seem to be a very good predictor of significant factors affecting LogPrice. My three variables failed to be associated with a significant difference in LogPrice.

Suggestions for next steps in analyzing this data

Based on the results found, it would seem that finding a significant predictor of log price might be difficult when using one variable at a time. I might want to look at combining variables in future models. Perhaps a linear regression model?

I tried running a linear regression using all three variables and it appears that using all three at once produces a more accurate model with a 50.4% R Squared. All three variables are significant in the linear regression model.

Regression Output

Coefficient	B	S.E.	t	p
Intercept	11.570	.009	1316.285	.000
FireplaceInHome	.213	.011	18.929	.000
HighQuality	.397	.011	36.080	.000
BigLot	.187	.011	17.764	.000

Summary

The initial exploration of the Ames Housing dataset indicates that single variables tend to not be significant predictors of log sale price. However, adding multiple variables tends to improve predictive accuracy. There are many variables that were not tried in this example. These other variables could be cleaned up and added to the linear regression model. It also might be interesting to explore the use of dummy variables. The conversion to dummy variables should be done sparingly, and should proceed with some theoretical end in mind.

References

De Cock, Dean (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3):1-15. Downloaded from https://www.tandfonline.com/doi/pdf/10.1080/10691898.2011.11889627