Christopher Lill, Garrett Yagade, Matthew Dumont, Jacob Dineen

SCM 651

Homework #1

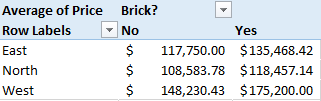
February 5, 2017

*Impact of Housing Prices on Known Predictor Variables*

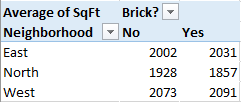
**1. Develop a categorization of your data using pivot tables. Develop two pivot tables of average price and average square feet by type of construction (brick) and neighborhood (20%)**

Using two basic pivot tables, we can see combinations of neighborhood, location and construction type and the resulting connection of price and square footage. Note that the column titles indicate whether the construction is brick, while the rows indicate the neighborhood.

Average Price by Construction Type and Neighborhood



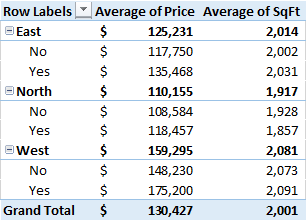
Average Square Footage by Construction Type and Neighborhood



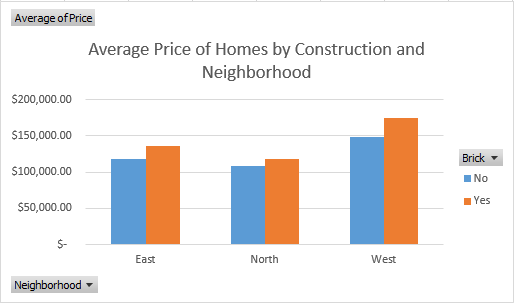
Based on a cursory look at the tables, it appears that homes built with brick are more expensive than homes made from other types of materials. Also, homes built in the West neighborhoods are both larger and more expensive.

The below table combines both price and square footage into one pivot table.

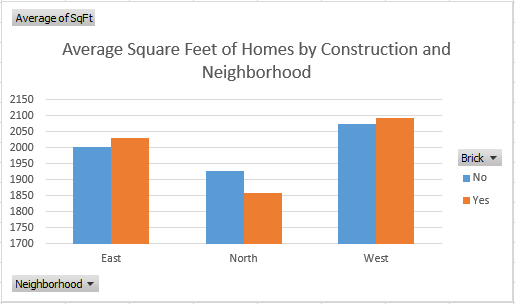
Price and Square Footage by Construction Type and Neighborhood



**2.Using the two pivot tables above, generate pivot charts for average price and average square feet by type of construction (brick) and neighborhood (10%)**

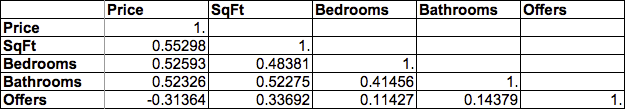


Using the pivot table above, a pivot chart can be made comparing the average price of homes in each neighborhood by construction type of the house. From the pivot chart, it can be seen that brick homes cost more across all neighborhoods, which logically makes sense because brick homes tend to cost more than wood or cement homes. From the graph, it can be seen that homes in the West neighborhood that are made of brick are the most expensive, while homes in the North neighborhood that are not made of brick are the lowest priced.



Using the pivot table above a pivot chart can be made comparing the average square feet of homes in each neighborhood by construction type. Both the East and West neighborhoods have higher average square feet per home, with the homes constructed of brick being slightly bigger. Interestingly, the North neighborhood has the smallest homes, with the brick homes being smaller in size than the non-brick homes. For the East and West neighborhoods, the brick homes are larger than the non-brick homes.

**3.Perform a correlation analysis of all quantitative variables except ID. Which two variables have the largest magnitude correlation? Which two variables have the smallest magnitude correlation? What does the largest magnitude imply if we perform a regression analysis next? Are there any negative correlations? Are these correlations intuitive? If not, why not? (20%)**

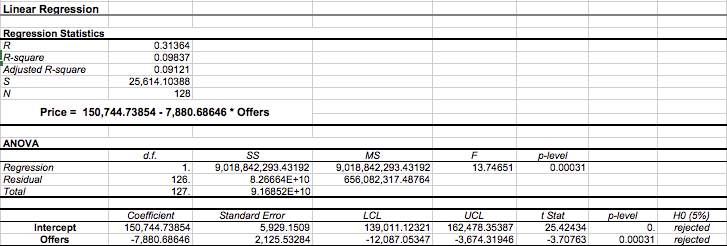


Price and square footage have the largest correlation in the data set. Conversely, the smallest correlation is between bedrooms and offers. The implication of the correlation between price and square footage is that as square footage increases the price will increase. Specifically, the correlation implies there will be a positive coefficient for the predictor variable, square footage.

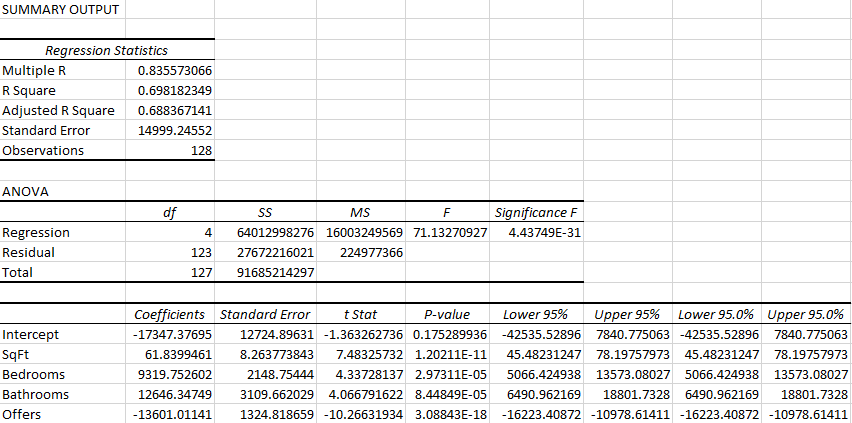
Regression analysis will allow us to understand specifically how the price is adjusted given changes in square footage.

The correlation between offers and house prices is not intuitive. Intuitively, as demand for a house increases the price should also increase. The increased demand should drive the price of the house up until no one is willing to submit a higher offer. If I am a homeowner and have 5 bids for my home, I would accept the highest bid. If there is a potential buyer that wants to outbid everyone then the price would increase. However, the correlation analysis demonstrates that a house with more offers will have a smaller price and a house with less offers will have a higher price. So, while the correlation data is logical it is not immediately intuitive.

Additionally, a univariate regression with offers as the predictor variable (input) and house prices as the predicted variable (output), highlights the inverse relationship between these variables.



**4.Perform an initial regression analysis of the quantitative variables excluding the ID. Which variables are statistically significant? What does each coefficient mean in a real-world sense? Are these coefficients intuitive? If not, why not? What does the R-squared mean? (25%)**



The significance of F falls far below ideal threshold of .05, which tells us that the statistical probability of having an error present in our equation is very small. The R2 tells us that 70% of the cost variability is measured by the present variables in this regression analysis, which are square feet, bedroom, bathrooms and offers. 30% of the cost variability cannot be determined by the present variables, and could best be explained by impacts outside of the realm of our predictor variables. These variables are all statistically significant because they all have p values that falls below the .05 confidence interval. An example of an undefined variable impacting our Y could be precise location. Our fixed cost is the coefficient of the intercept, which is defined as -17,347.37. The statistical significance of each of our inputs falls below the .05% confidence threshold, which tells us that these coefficients are statistically sound to the model.

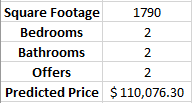
The regression equation can be defined as Y = Intercept Coefficient + X1 Coefficient \*X1 + X2 Coefficient \* X2 + X3 Coefficient \* X3 + X4 Coefficient \*X4

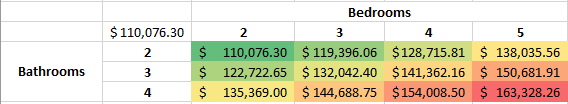
The previous analysis results in a regression equation as follows:

Price = - $17,347.37695 +$ 61.83995 \* SqFt + $9,319.7526 \* Bedrooms + $12,646.34749 \* Bathrooms - $13,601.01141 \* Offers

In a real-world sense, the coefficients do appear to be intuitive. What our regression equation tells us is the impending response on price as the variables increase/decrease. Looking above, we can see that Bedrooms and Bathrooms have the highest coefficients. However, when square footage increases we see the largest corresponding increase in our outcome (price), due to the magnitude of that value in comparison to bedrooms or bathrooms. Square footage has a positive relationship with price, but it doesn’t have the same weighted impact on the result as the aforementioned variables, as seen by the value of its coefficient. Lastly, offers has a negative relationship with price. With offers, it can be logically drawn that the negative coefficient tells us that as our y value increases, our x value (offers) decreases. As offers go up, demand goes down. As demand goes down, prices go down. If this variable was to be intuitive, price would move up as offers increased.

**5.Create a spreadsheet prediction of the model. Perform a two-way sensitivity analysis and use conditional formatting to highlight the results. (15%)**



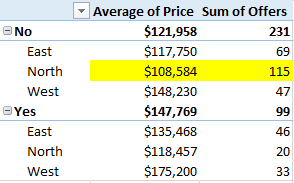


Using the regression equation, a two-way sensitivity analysis can be performed. However, this assumes that 2 variables are kept constant (this case square footage and offers) while the bedrooms and bathrooms are the two variables that are changing to see the effect on the predicted price of a home. Any two variables can be analyzed, but the chart must include only values that are within the minimum and maximum values of the data set.

Conditional formatting allows for an easier and quicker interpretation of the two-way sensitivity analysis. The green cells are the projected lowest priced homes with corresponding bedrooms and bathrooms, while the cells in red/orange are the homes with the highest projected price. The chart seems to be intuitive because the more bathrooms and bedrooms a home has, the higher the price will be, as demonstrated by the regression analysis.

**6.What would explain non-intuitive results in your regression using the data which you were provided? What additional data would assist you in explaining the non-intuitive results? (10%)**

While it seems logical that the number of bathrooms, number of bedrooms, and square footage are positively correlated with sales price, it is odd to see a negative correlation between the number of offers and price. Intuitively, more offers on a house would drive the price of the home up, at least to market equilibrium. The non-intuitive results might be explained by the limited number of variables included in the regression. For example, knowing the age of the home, the precise location (by an interstate, school, etc), and the condition might help to explain these results. See the below table showing that North neighborhood homes - not made of brick - had the most offers, yet the lowest selling price.



Also, the y-intercept is negative, meaning that the price of the home is negative if all other variables are zero. Again, this might indicate that other variables which are negatively correlated to price are missing from the equation.