**Factors affecting Conway’s house prices**

Saliou N’golo Ouattara

Regression Analysis

**Abstract**

Determining the factors that influence banks, real estate agents, and house owners in their decision to set the price of a house is of great importance in the light of family, buyers and sellers. This study focus on discovering which major statistical categories or variables (i.e. bedrooms, bathrooms, area, etc.) significantly contribute in setting the house prices in Conway. Thus, we collected the data of ninety-seven house for sale in Conway using the Zillow website, Google maps, and created an Excel spreadsheet. The data includes nine variables where two are qualitative and the seven others are quantitative. Using the software R, a statistical analysis of this data set was carried out. Results indicated that the number of bathrooms, area, and the age of the house contributed significantly to house prices in Conway. The standardized regression coefficients showed that area of the house had the greatest impact on its price. Implications of these findings are discussed in this investigation.

**Introduction**

This project is meant to apply regression analysis on a real-life situation. People look for houses that best fit their interest and in doing so they prioritize certain criteria over others. Also, specific features of the house may determine its price. According to Zillow, their data through September 2016 indicate that the median price of house is $ 164,650. A remarkable change occurs in the Conway house values since 2007, it has increased by 2.9 % over the past year. Moreover, the Little Rock Metro average list price per square foot is $91 while Conway has $94, which is slightly higher, and Zillow predicts an increase of 3.8% within the next year. In this study, we focus on 6 different variables that we believe are really important element in selling or purchasing a house and test their relationship with the House price. What are the factors affecting house prices in Conway? In order to tackle this question, we build a linear regression model through some statistical analysis.

**Data and Model**

To lead our analysis which deals with the factors that influence house price, we 8 variables that we believe in crucial in this study. Those variables are Beds, Baths, Area, Year built, Address, Church, Parking, and Yard. First of all, according to Collins English dictionary, house prices are defined as the sums in money for which houses may be bought or sold. The house prices at which banks, real estate, people are willing to pay and sell their property are display in the Zillow website for advertising and statistical data.

Second, bedrooms are the number of bedrooms inside a house. It is always along with the bathrooms, which are the numbers of bathrooms inside a house. There are usually one more bedrooms than the bathrooms or the same.

Next, Area is the “size” of a house, it is referred by Zillow as one the most confusing and misleading metrics in home building and house selling business. The reason for that is everyone measures differently the area, as there is no adopted standard. To avoid any confusion, we decide to go with the data provide by Zillow which measure this variable in square foot.

The Year.built our data refers to the age of the house. We used the year in which the house was built and subtracted 2016. Furthermore, Address is the location at which the house is built. It describes the street number, avenue, the city where the house is located, Conway, and the zip code. It is one of the qualitative variables in our data. Also, Church is the distance between the address and the nearest church in Conway. Church is measured in miles.

Parking is the variable that describes the space inside the house that can keep a specific number of cars. The Carport, the front yard, and the free space around the house are not considered in this variable.

Lastly, Yard is defined as a small usually walled and often-paved area open to the sky, or ground immediately surrounding a house that are usually with grass. It is the second qualitative variable in our data. It describes every house in our data that has a yard or not. We randomly collected our data from Zillow website and created an Excel spreadsheet. In addition, we collected the Church data separately, by using the address of every single house and putting them in Google maps, searching for the nearest Church regardless of the religious preferences. Since the house prices were big numbers we divided it by thousand to avoid high data number and accurately evaluate our model. We applied the same modification on Area data. In total, we have been able to collect 88 observations according to our 9 variables.

**Model**

From this data, we decided to build our model with House.Price\_thousand as dependent variable and we selected seven independent variables, which excluded Address because we just needed it to determine the Church variable. First, we did a descriptive summary (see annex) and we did pairwise correlation method to see if there is a significant correlation between the quantitative independent variables with House.Price\_thousand. We obtained a strong positive linear relation between the dependent variable and Area\_thousand. Also, the dependent variable has a weak linear relation with Year.built, Beds, Baths, and Parking. We did also a scatter plot matrix (see annex), which show a linear pattern between House.Price\_thousand and Area\_thousand. We, then build our initial full model, which ends up to be as follow:

The overall model is good with p-value of <2.2e-16, which is less than .05. We found out that all independent variables are have p-value greater than .05, which implies that we have a good fit. They did not totally respond to our expectation. Holding everything constant for every one-unit increase in Beds, or baths for a house its House.Price\_thousand will decrease by the estimate coefficient of whose variables, which is not realistic. However, every unit increase in the Area\_thousand will increase the House.Price\_thousand which is what we will expect holding everything else constant. Also, the for the Year.built the sign was what we were expecting. Holding everything else constant, every unit increase in the Year.built will lead to decrease of House.Price\_thousand by 1.56.

Next, we were not satisfied of this model so that we decided to check for the variance inflation factor and the coefficient of variance. The model did not show any multicollinearity, but the coefficient of variation was 28.48%, which is not good because 28.48% of the data can be explained by the residual. In addition, we did a residual analysis. It will decide if we have a good model. Furthermore, we looked for outliers and influential observations. We found that some observations affected the accuracy of the model, so there are some observations in the data that we need to exclude and obtain the best model. Using cook’s distance, standardized residual, and leverages, the model showed, after multiple trials, many outliers that were also influential outliers. As whose outliers can affect drastically the data, we removed them and rebuilt the model from the beginning. We repeat the process until we did not find anything that can affect the data. Therefore, we removed some observations and ended up with sixty observations for our best model.

Also, in building our model we have insert some interactive terms and some square term but all of them were kicked out as we were removing the outlier and using the stepwise method to generate the best model. After countless of trail we finally decided to keep the upcoming model as the final best model. Discussion on the model is done in the next section

**Results & Analysis**

*House.Price\_thousand =15.2 + 11Baths + 96Area\_thousand – 2.45 Year.built*

The overall model is still good with the same p-value lower than .05. The three remaining independent variables have a greater lower p-value except Baths, still it is a good fit. Again, to make sure that we have the best model, we started by checking the variance inflation factor. It showed that there is no multicollinearity for all the independent variables. Also, we checked the coefficient of variation which indicated that the value of the standard deviation of the least squares line is 11.35% of the value of the sample mean of House.Price\_thousand. It means that 11.35% of the data can be explain by the residual. It is half better of the previous one.

Next, for a better approach of our study, the model needs to pass the residual analysis. We set about the assumptions on the residuals. First, we look for outliers and influential observations. Using cook’s distance, standardized residual, and leverages, the model did not show any outliers as well as influential observations. We continued by making a subjective decision of the normality test using a Q-Q plot for residual (see annex) which showed that all the data points lie within the red lines, the residuals are normally distributed. In addition, the Shapiro test give us a p-value of 0.7679 which is greater than .05. It concludes that the sample is from a normal distribution. Next, we used Breusch-Pagan test for checking constant variance of residuals. It has a p-value of 0.8663 which is greater than .05 and suggest that the residuals are constant so there is homoscedasticity. Moreover, we used Durbin-Watson test for checking independence of residuals. It gives DW of 1.4133 with a p-value of 0.01322. Since the p-value of the Durbin-Watson test was greater than .05; it indicates that there is residual correlation. That may be due to important variables that was not included in our model or given the small size of our data we couldn’t test for some influential variables. We can consider that the model passed the residual analysis. Finally, we keep this model has our final model. It indicates that holding everything constant every increase in one-unit of Area\_thousand will increase House.Price\_thousand by $ 96,277 and every increase in one-unit of Year.built will decrease House.Price\_thousand by $ 2,449. After standardized the coefficient of the model, we found that Area\_thousand had the greatest coefficient; it is the most influential variable.

**Conclusion**

The variation in the house price depends on several aspects. In this study, we try to understand what are the factors that determine the house price in Conway. This project shows that the Area\_thousand, Year.built increases the House Price, which matches our expectation. Area increases the house price because the wider the square feet of the house the costly it should be. Year.built indicate that the older the house is the more it price will go down. However, testing the relationship between House price and the predicators baths, beds, parking, yard, and Distance from the house to the church seem to be not statistically significant according to our analysis. Overall, the results show that the determinants of house prices in Conway are the age of the house and the area of house, where the latter is the most influential one.

**Further Extension**

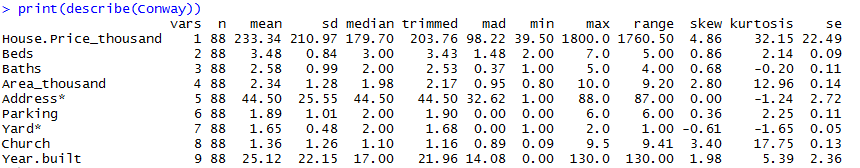
The work we have done here is very limited because of the data sample we are using and the limited control variables. Area is the most important variable from our study but this outcome may change if some new variables are introduced in the model. For future work, we could add more variables on the model based on other factors that can affect the house prices such as crime rate, distance from high school, college or downtown. As one may also find an important variable that we may have not noticed and included, and come out with a more robust analysis. Furthermore, the data can be collect in many other cities in Arkansas and a comparison between cities can be carry out. Moreover, we can have a model that a buyer will use to locate the address where he can have a house according to his financial situation and that the bank can use in their attribution of loan.

**References**

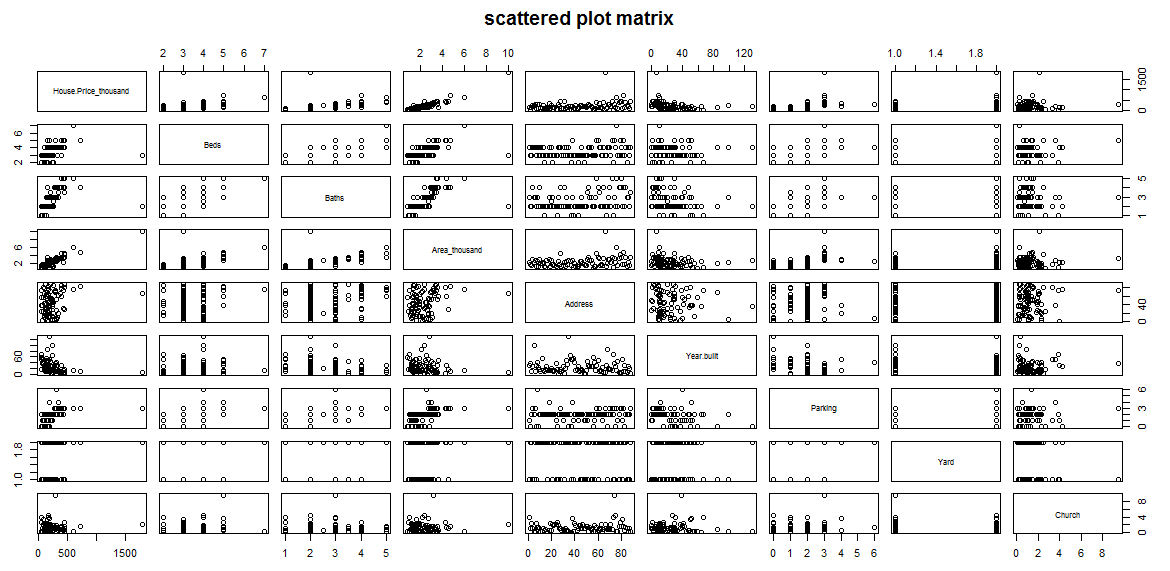
* <http://www.zillow.com/homes/for_sale/Conway-AR_rb/?fromHomePage=true&shouldFireSellPageImplicitClaimGA=false&fromHomePageTab=buy>
* <http://www.collinsdictionary.com/dictionary/english/house-prices>
* <http://www.merriam-webster.com/dictionary/yard>
* <http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis>

ANNEX

Descriptive Statistics



Scattered Plot Matrix



Q-Q plot

