Mariama SOUMAHORO

REAL ESATATE MARKET

ABSTRACT

Many different factors affect the real estate market. This study examines the market historical data set of real estate valuation. A multiple linear regression model was fit to the data in order to predict the Price per square unit of a house by establishing a statistically significant linear relationship with the corresponding factors. After fitting the regression model using the information given, transaction date, age of the house, distance to the nearest MRT station, number of convenience store, longitude and latitude are shown to be important determinants. The techniques used, and the importance of these variables give helpful insights into the current real estate market.

1 DATA CHARACTERISTICS

Data Collection

The data presented to us is part of the Market historical data set of real estate valuation. The data set was collected from Sindian Dist., New Taipei City which is located in Taiwan. It contains a representative sample of 414 observations (price of a house per unit area). The dataset describes how houses are priced in the real estate market based on some factors. The outcome of interest is the price of a house per unit area. The data frame consists on 6 characteristics that we will be using for our analysis. The 6 independent variables or predictors are:
X1 = transaction date: the date the house was bought, for example: 2013.250=March,2013
X2= house age: the house's age in years
X3=distance to the nearest MRT station in meter: the distance between the house and the nearest Medical Response Team like a Fire station or hospital. It is taken into account while setting the price of a house. The closer you are, the higher your property value will be.
X4= number of convenience stores in the living circle on foot: It represents how close you are from different stores. It increases the value of the property if it is close to attractions and stores and coffee shops. The closer you are the higher your property value can be.
X5= latitude helps locate the property on a map, geographic coordinates of a property(degree)
X6 =longitude: geographic coordinates of a property (degree)

2.DATA VISUALISATION AND EXPLORATION

Diving into the data and taking a closer look at the information given will help us get more familiar with the data in question. We first look at a summary statistic of all the variables like the mean, median, minimum, and maximum values organized by the n= 414 observations. The median of the price of house per unit square is 38.45; the minimum and mean are respectively 7.60 and 37.98 which is close to the median. The highest that a price of a house per unit square can go is 117.50. We proceed by dividing the transaction date into 3 groups to facilitate the analysis: "=<2012","2012-2013",">2013". We then plot the data to see if there exists a relation between the dependent variable price per square unit and independent variables.

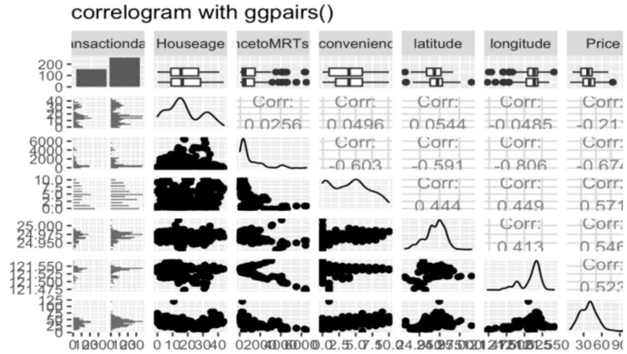


Figure 1 represents the scatterplots of each pair of numeric variables which are drawn on the left part of the figure. Pearson correlation is displayed on the right.
For both latitude(r=0.55) and longitude(r=0.52), although there is a somewhat strong positive relation between them and the price of the house, it might not be linear.
Variables: distance to the nearest MRT station with correlation r=-0.6736 and age of a house with r=-0.21 have a negative relation with the dependent variable. This relation might also not be linear since both lines are curving. For distance to the nearest MRT station, the furthest the house is from an MRT station, the lowest the price per unit square of that house will be. Looking at the relation between again the price and the number of convenience stores, we observe that there is a linear relation between the two variables. The correlation which is equal to 0.57 emphasizes that fact and demonstrates that we are in a presence of a strong linear relation.
After viewing the data, the next step is to fit a multiple linear regression.

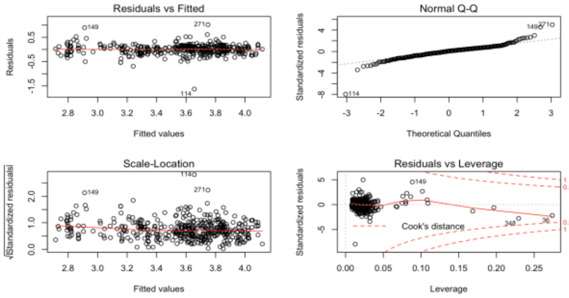
3 MODEL SELECTION AND INTERPRETATION

Section 2 established that there were some patterns between the price per unit square of a house and the different predictors. This section aims to summarize those patterns using a multiple linear regression modeling. We check the distribution of the dependent variable: price per unit square of a house. We see that 'log 'does a good job to transform the variable's distribution close to normal. Therefore, we fit a multiple regression model using the transformation and with all the predictors included. We get F=147.6 which is far greater than 1, and so it can be concluded that there is a relationship between predictor and response variable. Based on the 'p-value' we can conclude that all predictors are significant for the model besides longitude. R^2 (multiple-R-squared) value indicates how much variation is captured by the model. R2 closer to 1 indicates that the model explains the large value of the variance of the model and hence a good fit. In this case, the value is 0.6857 (closer to 1) meaning 68.57% of the data is explained by the model. Therefore, we have a good model. Despite the fact that "longitude" seems to not play a role in the model, it is important to note that we need both "longitude" and "latitude" to get coordinates, leading us to keep longitude in the model. Continuing with the modeling, we plot the residuals against each predictor to see if the predictors require any transformations in the model. The results show us that we might need to include polynomial terms in the model. We therefore refit the model. A better R^2(0.7298) and Adjusted R^2(0.7238) are found meaning that 72.98% of the data is being explained. Although, there is a decrease for the F-statistic which is 121.3 for this model, it is still far greater than 1. The AIC is -120.345 which is very low. All the p-values of the predictors variables are significant. All those observations are indication that the model fitted is the best model. Thus, the final model is as follow:

$\boldsymbol{log(\hat{Y})}$ ***= 9.272e+04***
+ 8.380e-02
(transaction date>2013) -2.517e-02
(House age) + 1.855e-02
(number of convenience stores) - 2.760e-04 (distance to MRTstation) -7.441e+03
(latitude)+ 6.840e-01 (longitude)+ 4.504e-04I(Houseage^2) +2.695e-08((distance to MRTstation^2) + 1.492e+02(latitude^2)

The coefficient of the transaction date is the average difference in the price per square unit of a house between the different groups. This means that the average of the price of a house is greater by 1.09 for the years after 2013 than for the other years while holding everything constant. There is a (e^(-2.575e-02)-1) = 0.05% decrease in the age of the house (years) for every 1% increase in the price per unit square while holding everything else constant. We also get a 1.01% increase in the number of convenience stores for every 1% increase in the price of a house. For every one - unit increase in the distance to a MRT station, the price of a house reduces by 0.99 while holding everything constant. Regarding longitude, there is a 1.98% increase in longitude for every 1% increase in the price of the house while holding the other variables constant. As for the polynomial factors, it is important to note that the price increases and decreases as the predictor increases.

We continue our analysis by checking for the residual plots which help us detect if there is an anomality with the model chosen.



Starting with the Residual vs fitted graph for the Nonlinearity of the data, we observe that the line is approximately horizontal at 0. It suggests that there is a linear relationship between the predictors and the outcome variable.

Non constant variance: using the scale location plot, we have that the residuals are scattered around the line at the beginning and then, there is a strong concentration at the end as the values of the fitted outcomes variable increase. This suggests that there is a non-constant variance in the residual's errors.

Q-Q plot shows whether the residuals are normally distributed. In the above plot, we see that most of the plots are on the line except at towards the end. Also using the Shapiro test, the p-value is significant, so we conclude that the residuals are not normally distributed.

The residuals vs Leverage help us find the influential points. Indeed, we have more than 15 influential points. Among them, we observe 3 outliers (observations 114,271,149).

LIMITATIONS

At the end of our study, after performing a multiple linear regression, we were able to determine a good model to fit the data and understand the relationship between the dependent variable "price per unit square of a house" and the different factors. All the predictors have a significant impact on the dependent variable. Around 72.98% of variation in the Real Estate data was explained. There is still a great portion of the data that has yet to be explained as well. The great number of influential points found, and the non-normality of the residuals are an indication that this model although good, needs more work. Perhaps more predictors should be added or more transformations of the data are required. It still has scope for improvement as we can apply advanced techniques Random Forest and to check whether the accuracy can be further improved for the model.