**Find Your Dream House in a Flash**

***-Using Regression Model& Clustering to Find model for Unit House Price Associated with House Condition Variables.***

Name:

Yong Qiao

Zhenghao Yu

Bangchen Xu

Contact Address:

9784899667

E-mail addresses:

[yong.qiao@uconn.edu](mailto:yong.qiao@uconn.edu)

[zhenghao.yu@uconn.edu](mailto:zhenghao.yu@uconn.edu)

[bangchen.xu@uconn.edu](mailto:bangchen.xu@uconn.edu)

**Abstract/Summary**

In this project, we are focused on building regression model and using clustering to find the model for house price of unit area associated with house condition variables. Independent variables include house age, distance to nearest transportation stations, number of nearby stores and geographic coordinates information.

In the first part, in order to build a relatively simple but efficient model, we build a Linear Regression model and used stepwise algorithm to find the better model by AIC values. Although we get the better model by the above method, it did no go well when we test assumptions for this Multiple Linear Regression model: Normality assumption, homoscedasticity assumption and Multicollinearity assumption are not met for this model. Therefore, we started to consider using other methods to find suitable model.

In the second part, we focused on using clustering to find the model for house price of unit area associated with house condition variables. Working throughout three phases, observing, clustering and validating, we made unit house prices and geographic coordinates ordinal and found the ordinal regression model based on them and other existing independent variables. The accuracy is 0.53-0.67, which is relatively high for a small dataset with size of 414 and the multicollinearity assumption and ordinal variables assumptions are met for the model, which means the model meets our demands quite well.

**Section 1: Introduction**

Our project aims to give potential house customers credible predictions about the price of their dream house when they can provide house location information. In this way, they can find their dream house fast.

There are two models we have worked on: a multiple linear regression model of unit house price associate with house condition variables and an ordinal logistic regression model of rankings of unit house prices associated ordinal geographic information and other house condition variables. After testing assumptions and accuracy and thinking from customers’ perspective, we found the second model is more useful and realized the method in which we build the model is insightful.

Another significant point is that, the emphasis of this project is to provide a methodology for customers. That methodology can predict information about intervals of prices for their dream house as soon as they input their location requirements into the model for any part of the world with large dataset for unit house prices in a short duration. Therefore, we can say that this methodology is highly potential and waiting for being extending.

**Section 2: Description of Data**

Dataset Citation: The market historical data set of real estate valuation collected from Sindian Dist., New Taipei City, Taiwan, posted on website of UCI Center of Machine Learning and Intelligent Systems, http://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set .

Original Source: Owner and Donor Name: Prof. I-Cheng Yeh in Department of Civil Engineering, Tamkang University, Taiwan.

The independent variables(X1- X6) and response variable(Y) in data:

* Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)
* X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
* X2=the house age (unit: year)
* X3=the distance to the nearest MRT station (unit: meter)
* X4=the number of convenience stores in the living circle on foot (integer)
* X5=the geographic coordinate, latitude. (unit: degree)
* X6=the geographic coordinate, longitude. (unit: degree)

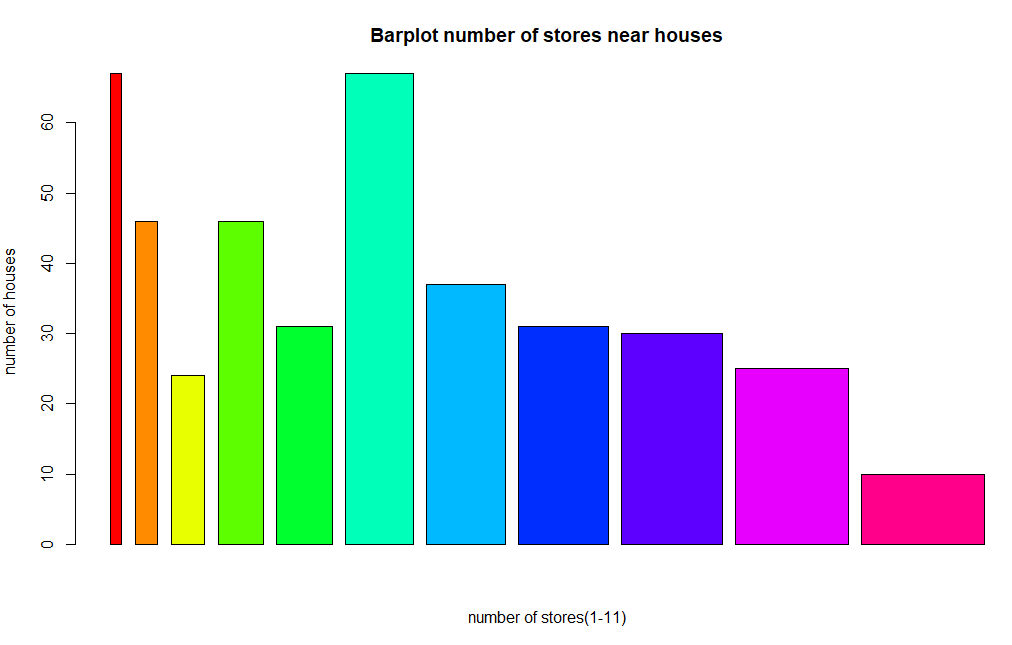
There are two reasons why we can exclude X1(the transaction date) before building the model.

* The first reason is that data we used were collected during 2013-2014, which is helpful since we do not have to worry about the price changed much in this short duration.
* We thought about the season of transaction dates could have impact on unit house price. However, from Graph 1, we can see that the season did not affect the unit house price much overall.

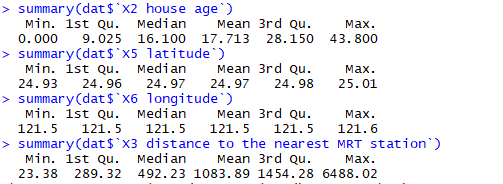
图片包含 文字, 地图

描述已自动生成

(Graph 1)



(Graph 2)



(Graph 3)

From Graph 2 and 3, we can see that in 414 datasets, house age, number of stores, distances to nearest transportation stations could be quite different. However, two geographic coordinates columns(X5-X6) could not give us much information. This is one of reasons why we spent more time on analyze those two variables in part 2.

**Section 3. Methods& Models:**

Part 1: Multiple Linear Regression Model:

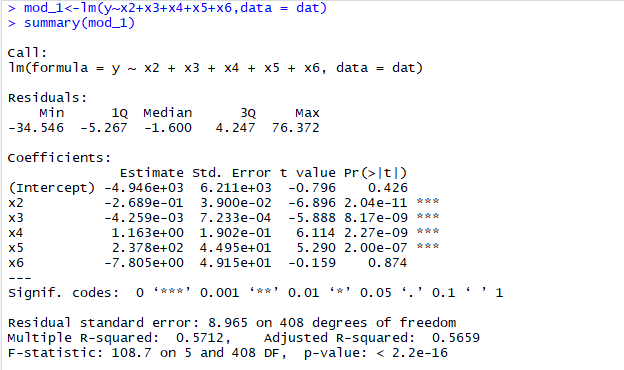
Model 2(After Stepwise Model Selection by AIC):

Part 2: Ordinal Logistic Regression model:

* Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)
* X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
* X2=the house age (unit: year)
* X3=the distance to the nearest MRT station (unit: meter)
* X4=the number of convenience stores in the living circle on foot (integer)
* X5=the geographic coordinate, latitude. (unit: degree)
* X6=the geographic coordinate, longitude. (unit: degree)
* Y\_rank= levels of house price of unit area (level 1: Y<=27.7, Level 2: 27.7 <Y<=38.45, Level 3:38.45 <Y<=46.6 and Level 4: Y >46.6)
* X7=Location ID number, which is the ordinal geographic ordinates, obtained from X5 and X6 and clustering procedures, including 4 levels.

**Section 4. Analysis of Data:**

Part 1: 1. First of all, after eliminating transaction data variable for two reasons we claimed in Section 2, we want to test the model 1: .



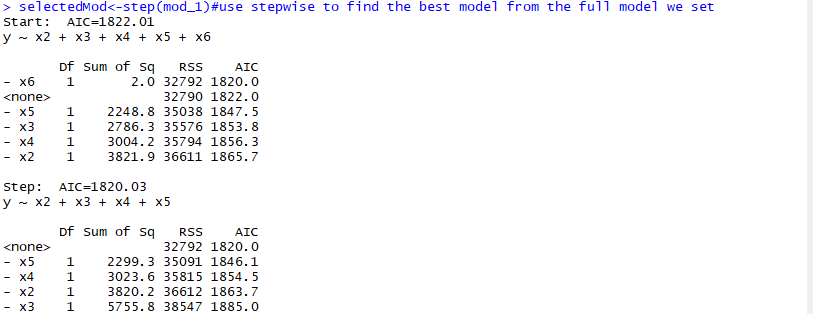
(Table 1)

We want to test if each independent variable has significant effect on Y(Unit house price).

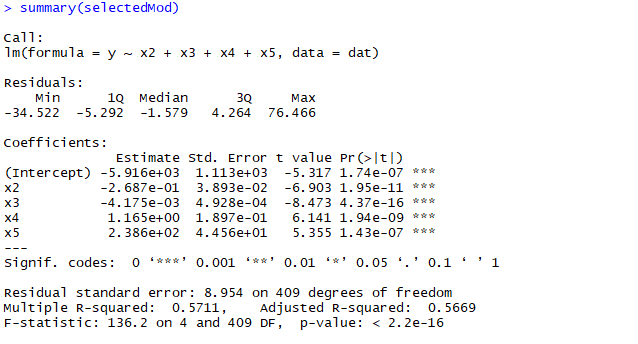
Hypotheses: H

From Table 1, under level of significance = 0.05, comparing the p-values with 0.05, we can see that the x2-x5 have significant effect on Y and X6 does not.

Then, for get a better model, we decide to use Stepwise Algorithm in R and find a better model:



(Table 2)



(Table 3)

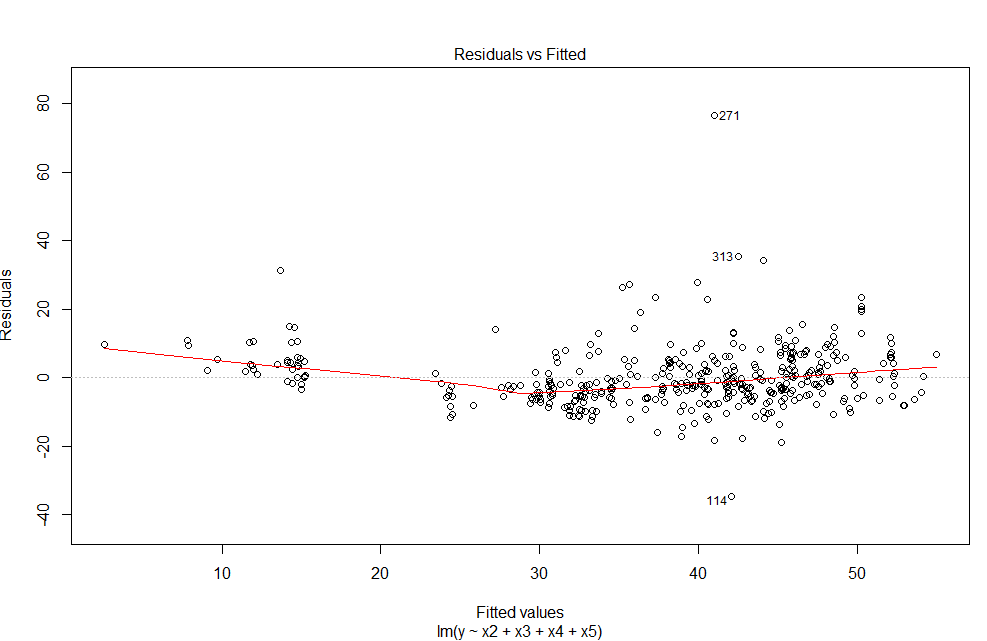
The table 2 and table 3 showed us the procedure of stepwise model selection (By AIC) and we finally got the better model(model 2):

Then, we decided to test three assumptions for this improved model: Homoscedascity, normality and multicollinearity.

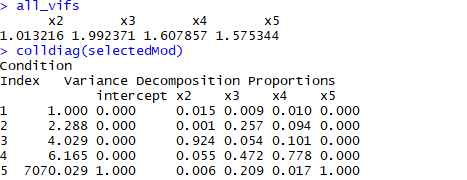
图片包含 屏幕截图

描述已自动生成

(Table 4)



(Graph 4)



(Table 5)

From KS-test, plot of residual vs. fitted values and condition index in table 4, graph 4 and table 5, we could see that p-value is too low to prove normality under level of significance = 0.05, the clear signal of heteroscedasticity and high condition index on X5. Therefore, three assumptions are not met.

For the above reasons, considering the utility of models we tried to build from customers’ perspective and the accuracy of models, we decide to explore more on the raw data and find another model for unit house price associated with other location variables.

Part 2:

In this part, we aimed to study deeply into datasets again and find certain grouping patterns among different levels of house prices, which indicates certain grouping has the relationship with unit house price if reorganized carefully.

The first step is observing the geographic coordinates on a google map. We first combined latitudes and longitudes as one new variable(combined geographic ordinates) and use the google maps to see the pattern of it and the environment of corresponding coordinates.

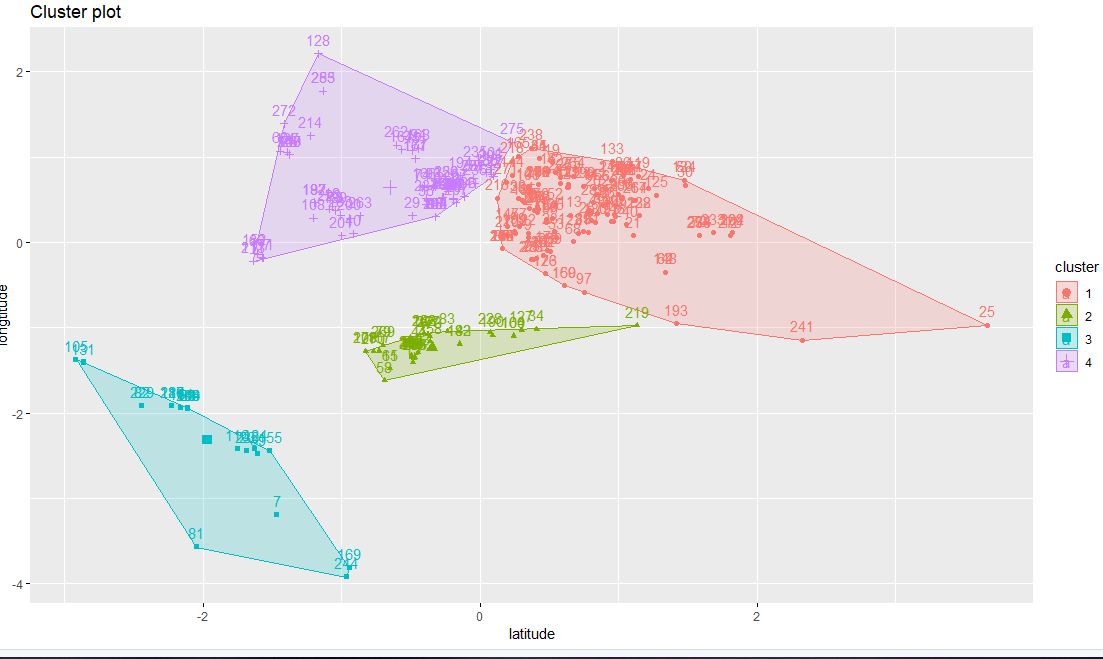
The google map is important in our observing procedure to have a rough idea about the number of clustering, which we will use in k-means procedure.

图片包含 文字, 地图

描述已自动生成

(Graph 5)

In graph 5, after observing the coordinate layout in google map, we have the rough idea that there could be about four groups of those coordinates’ information.

The second step is clustering. Standing from the perspective of customers, we decide to classify house price of unit area into four levels, so there will be a new response variable Y\_rank, intervals of house price of unit area. After that, we started clustering by K-means and obtained X7- Location ID number, which is the ordinal geographic ordinates, with four levels(as in graph 6).

(Graph 6)

Then we can get Ordinal Logistic Regression model of Y\_rank associated with X2,X3,X4 and X7.

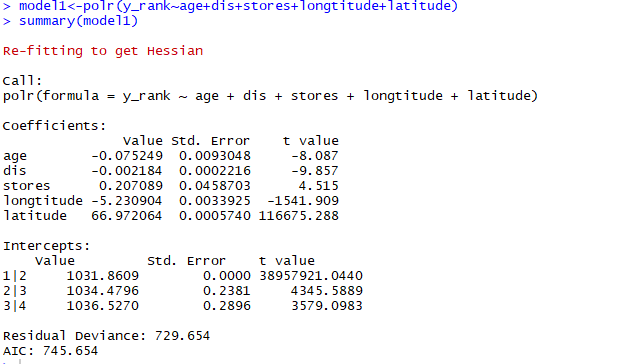
图片包含 屏幕截图

描述已自动生成

(Table 6)

We can see summary this ordinal logistic regression model in Table 6. In this way, we have already built a model to help people find the unit house price for their dream house as soon as they can provide their above location-related demands to this model.

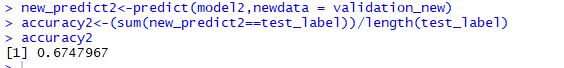
This model has been improved a lot by replacing latitudes and longitudes by location ID numbers. As we can see in Table 7,



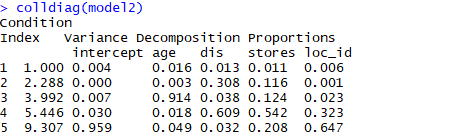
(Table 7)

Comparing with model with x5 and x6 in Table 7 through AIC, the model in Table 6 is clear the better one.

The third part is validation. Before stepping into the first stage of part 2, we had already split datasets into two parts randomly by ratio 7:3: training part and validation part. We used the training part to go through the first two steps and validated the third step by predict() function in R.



(Table 8)

Since the size of datasets is not large, the accuracy we got can be regarded as high. Therefore, we can say that this Ordinal Logistic Regression model of Y\_rank associated with X2,X3,X4 and X7 with summary in table 7 is the good model for local house customer.

(Table 9)

From (Table 9) and what we did to build the model, the multicollinearity assumption and ordinal variables assumptions are met for the model, which means the model meets our demands quite well.

**Section 5. Concluding Remark:**

First, assuming the model is a linear regression model, we built a Linear Regression model, improved it into the better model by AIC values, but three assumptions were not met for this model. Then, we focused on using clustering to find the model for house price of unit area associated with house condition variables. Working throughout three phases, observing, clustering and validating, we made unit house prices and geographic coordinates ordinal and found the ordinal regression model based on them and other existing independent variables. The accuracy is relatively high for these datasets and the multicollinearity assumption and ordinal variables assumptions are met for the model, which means the model meets our demands quite well.

What we can extend from our existing analysis and models:(1)We used 0 prior information to cluster the location data, which in real life we can cluster better based on commercial and geometrical information.(2) Inverse of this function, which means if we have utility for the unit house price, can directly give us the corresponding locations, which can significantly reduce the computational power of the problem and concluded to subproblems.(3)We can utilize this regression and its own sub-regressions to build an equilibrium system, which helps dealer better with sells and buyers find better houses and in a unseen fast speed.

**Reference:**

1. Data Source: The market historical data set of real estate valuation collected from Sindian Dist., New Taipei City, Taiwan, posted on website of UCI Center of Machine Learning and Intelligent Systems, <http://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>

2. assumptions of ordinal logistic regression model:

<https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php>