**INTERMEDIATE ANALYTICS**(Beijing Housing Prices Final Report)

Feb 12, 2020

ALY 6015 - 20975

**SUBMITTED TO: PROFESSOR ROY WADA**

**SUBMITTED BY: REETIKA CHATURVEDI**

**INTRODUCTION**

For this project, we chose to analyse the “Beijing Home Prices” data set that contained information about the pricing and different characteristics of over 300,000 home sales in the Beijing region. Throughout our analysis of the data, we seek to better understand the relationship between the prices of the homes and the population density of the area; and the relationship between the prices of the homes and the number of living, bath, and drawing rooms. We believe that we can better understand these relationships by using the following three methods: linear regression, multiple regression, and lasso regression. All three of these regression analysis models are unique in their own way and will help us better understand the correlations and the impact each data point has on the pricing of houses within Beijing.

**METHODS**

**Linear Regression**

The linear regression technique models the relationship between one independent variable and one dependent variable in a dataset. In our case, we regressed the square feet of our homes (independent) and the prices of those homes (dependent). By analysing the Beijing Housing prices dataset, we realized that the square feet variable can be a reasonable independent variable to consider in relation to the prices of the homes. We chose this method because we believe it is an effective way to understand and analyse the correlation of two variables to each other, giving us an effective way of determining what factors are correlated to/have an impact on the prices of the houses in Beijing.

**Multiple Regression**

Multiple regression is a model that explains the relationship between multiple independent variables with one dependent variable. We chose this method of analysis because our dataset contains a number of different variables that we would like to understand the relationship between the designated variable and the price. For example, we studied the number of living, bath, and drawing rooms and their relation to the price of the homes in Beijing to better understand their correlation. In the analysis of this report, we will determine how great of an impact each variable has on the home prices.

**Lasso Regression**

Lasso regression is the Least Absolute Shrinkage Selection Operator. Lasso takes a number of factors into consideration and provides insight into which factor is the most impactful on a dependent variable. In our dataset, we used all available data categories when finding the most impactful factor with regards to the price. The reason we chose this method is because we believe it is extremely effective in sorting and analysing all factors of the data set quickly and accurately.

**ANALYSIS**

For our analysis, we will discuss the linear regression, multiple regression, and lasso regression methods and how we have applied them to our dataset. We will discuss our inputs (R codes), outputs (graphs and statistics), why we chose this method of analysis, and the conclusions that we can infer from our outputs.

**Linear Regression Analysis**

The coefficient of the correlation was 0.56. This indicates that there is a relationship between the two factors but it is definitely not a strong relationship. Further, we need to consider other variables to decide the variables that have an impact on pricing. We applied linear regression on the data set the code to which is as follows:

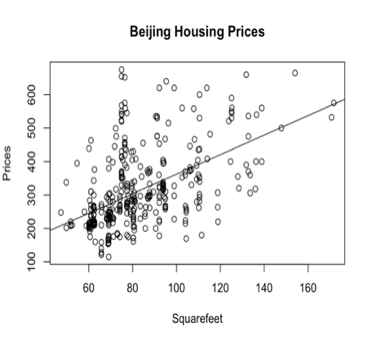
plot(Final$totalPrice,Final$square,main = "Scatterplot")

cor(Final$totalPrice,Final$square)

#not a strong correlation as per the coefficient

l\_model<- lm(formula=Final$totalPrice~Final$square)

l\_model

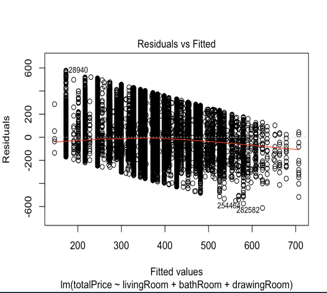
The linear regression used to explain the relationship between two variables. In this case, we set prices as the dependent variable(Y-axis) and the square-feet as the independent variable(X-axis). The linear regression graph below (Figure 1) shows that there is a positive relationship between prices and square feet, which indicates that the higher square feet always lead to higher prices..

**Multiple Regression Analysis**

In our dataset, we used the number of living, bath, and drawing rooms to understand the relationship they have on the total price. The diagnostic plots of the model using the plot () function provided a lot of details for the multiplication regression model.

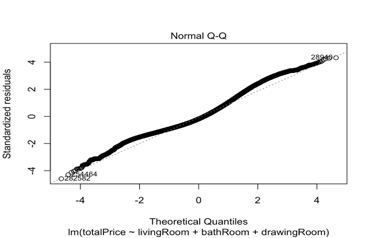
mr<-lm(formula=Clean\_GROUP\_HOUSING$totalPrice~Clean\_GROUP\_HOUSING$livingRoom+Clean\_GROUP\_HOUSING$bathRoom+ Clean\_GROUP\_HOUSING$drawingRoom)

plot(mr)



**Figure 3: Residuals diagnostic Plot for Multiple Regression**

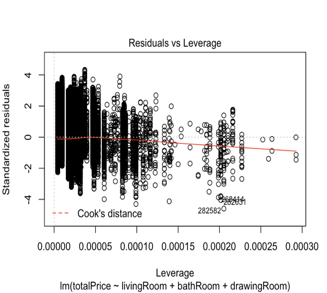
The residual plot provided the comparison of residuals of our model against the fitted values produced by our model. It is one of the most important plots as it helps provide the trends in our residuals, and the possible outliers we may have in the data. By observation, we could say that the model systematically underpredicted the higher values of the prices and systematically overpredicted the lower values of the prices. Although the residuals are not evenly spread around 0 for all the fitted values, we found no evidence for heteroskedasticity, which is, the presence of sub-population in the collection of random variables that have different variabilities from others.

Also, there are likely no outliers as there are no points on the plot well-separated from the rest.

**Figure 3: QQ diagnostic Plot for Multiple Regression**

The QQ plot showed that most of the points fell on the line which indicated that the residuals came from a normal distribution. But we saw some points strayed away from the line in the upper and lower quartiles of the plot. It could be possible that these points did not come from a normal distribution, but most of the points seemed to come from a normal distribution, so there

was not a lot to worry.



**Figure 4: Leverage diagnostic Plot for Multiple Regression**

The Leverage plot graphed the standardized residuals against their leverage. It contained the Cook’s distance boundaries which basically states that any point falling outside these boundaries would be an outlier in the x-direction. As we could hardly see the boundaries on our plot, we could conclude that there were no outliers.

|  |
| --- |
| ## Call: |
| lm(formula = price ~ livingRoom + drawingRoom + kitchen, data = housing) |
|  |
| Residuals: |
| Min 1Q         Median 3Q        Max |
| -52875    -15336 -4804         10249 117859 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 43419.29 364.54  119.107   <2e-16 \*\*\* |
| livingRoom   -485.82 56.31     -8.627 <2e-16 \*\*\* |
| drawingRoom -5090.47  84.07     -60.551 <2e-16 \*\*\* |
| kitchen  7096.41 362.39   19.582 <2e-16 \*\*\* |
| --- |
| Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error:  21520 on 318815 degrees of freedom |
|  |
| Multiple R-squared:  0.01735, Adjusted R-squared:  0.01734 |
| F-statistic:  1876 on 3 and 318815 DF, p-value: < 2.2e-16 |

Also, the summary statistics for the multiple regression stated a lot about the model.

The t-statistic values in the summary statistics for all the predictors had values significantly different from zero, which proved that the predictors were a valuable addition to the model. The R-squared value or the correlation coefficient for the model comes to approximately 0.6 which is close to 1. We also observed a significantly lower Relative Standard Error (RSE) value for the model, indicating the high accuracy of the model.

Altogether, we could conclude that there is a positive correlation between the number of living, bath, and drawing rooms and the total price of the houses in Beijing.

**Lasso Regression Analysis**

We felt that lasso is easier than ridge to interpret as it performs L1 regularization. In L1 regularization a penalty is added that equals the absolute value of the size of coefficients. It is used for models with fewer coefficients. Some coefficients can be eliminated from the model if they become zero. When penalties are large the value of coefficients is closer to zero, which is the best approach for producing simple models.

Lasso was used by us to perform not only regularization but also variable selection so that we are able to have a better understanding of the statistical model produced by us and also establish prediction accuracy. Lasso is the best for feature selection and since we had a lot of variables that affect the pricing of the houses, we decided to perform lasso regression to know the most important features.

We created a matrix of the data using the function matrix. Since the variables were odd in number (17 in total), we selected one row and seventeen columns. After this, we divided our data test into two parts- training and testing data. We selected 67% as training data and 33% as testing data. We trained the model using the training data. To apply lasso, we used the glmnet function. Alpha was chosen as 1. Specifically, our following inputs were:

X.bm<- matrix(data = Clean\_GROUP\_HOUSING, nrow = 1,ncol =17 , byrow = F)

str(X.bm)

dim(X.bm)

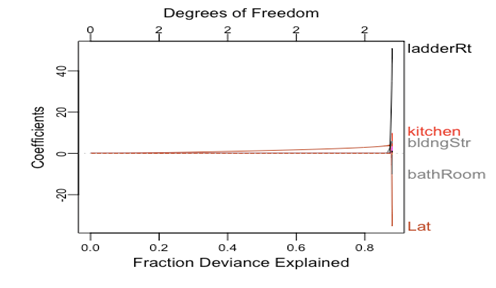
x <- model.matrix(totalPrice~.,Clean\_GROUP\_HOUSING)

y <-Clean\_GROUP\_HOUSING$totalPrice

train <- sample(1:nrow(x), nrow(x)\*.67)

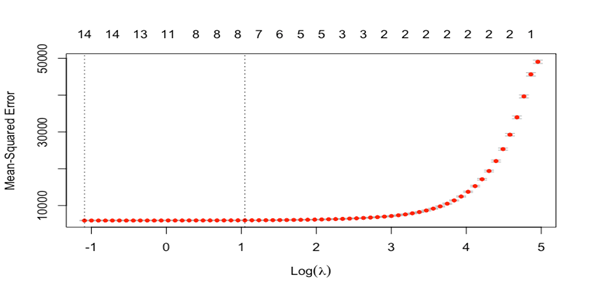
test <- (-train)

y.test <- y[test]

These inputs gave us the following output:

It can be seen that 90% of the variation is defined by our model. The five major elements that are essential to predict house pricing are on the right-hand side of the model. Lasso was not able to shrink these variables to zero. The higher the coefficient the higher the importance of the variable. It can be analysed that the most important factors are ladder ratio, building structure, number of kitchen and bathrooms and the latitudinal location of the house.

The next step was to choose the best lambda. We selected our lambda value as 1. The figure below represents the logged lambda values in terms of mean squared errors.

According to the figure, the best lambda values lie between -1 and 1. We can see two dotted lines, one is at the minimum and the other is at the standard error of the minimum. The least value is the optimal one but as we increase the lambda, we tend to get rigorous feature selection. Our selection of lambda value was based on the intention of deriving the best features amongst the many features available.

Lasso was one of the best ways to deal with this dataset as there were several independent predictors and all predictors do not hold the same relevance. The penalty added by it to the RSS minimized the RSS.

**CONCLUSION**

All in all, through the linear regression, multiple regression, and lasso regression methods, the dataset “Beijing Housing Prices” is now more understandable and better analysed. The conclusion that was derived from linear regression is that there is a positive correlation between the number of square feet in the home and the price for that home. Therefore, we fail to reject the null hypothesis and prove that square feet and price has a positive correlation, at least within our dataset. Through our multiple regression analysis, we have identified the relationship between living, bath, and drawing rooms with the price of the home in Beijing. We concluded that the information is normally distributed.Thirdly, through our lasso regression, we have concluded that the factors that have the most significant impact on price are: ladder ratio, latitude, building structure, number of kitchens, and number of bathrooms. This is important to understand when identifying important factors, predicting future prices, and understanding past prices. The three aforementioned methods of analysis have been very helpful as we analysed and understood the unique statistics, distribution, and regression of the data at hand. We have learned a lot from this project and will carry over our knowledge to future learning and employment.

**REFERENCES**

*Yaohui Zeng*. (2020). *Biglasso: Extending Lasso Model to Big Data in R*. Retrieved 27 January 2020, from https://github.com/YaohuiZeng/biglasso/blob/master/vignettes/biglasso.pdf

Yaohui Zeng. (2020). *The Biglasso Package: A Memory and Computation Efficient Solver for Lasso Model Fitting with Big Data in R*. Retrieved 27 January 2020, from <https://arxiv.org/pdf/1701.05936.pdf>

**R Statistics Blog. (n.d.). Retrieved from** [**https://www.rstatisticsblog.com/data-science-in-action/lasso-regression/**](https://www.rstatisticsblog.com/data-science-in-action/lasso-regression/)

**(n.d.). Retrieved from** [**https://www.youtube.com/watch?v=\_3xMSbIde2I&t=1056s**](https://www.youtube.com/watch?v=_3xMSbIde2I&t=1056s)