**Analysis of predicting selling price for a house in Hartwood**

Prepared for:

Daniel XXX

Submitted by:

Wanqi Su

April 18, 2017

**Executive Summary**

In the process of buying real estate, making a good purchase offer plays an important role. Daniel XXX, a real estate agent, is helping a client to buy a house in the Hartwood neighborhood of Langley, BC. His client is ready to make an offer on a house and need his guidance on its value. Therefore, Mr. XXX wishes to provide a report to his client that gives a reasonable range for the likely selling price of the house.

Mr. XXX’s question of interest is the likely selling price of his client’s chosen house in Hartwood and a reasonable range for the likely selling price of the house.

To address Mr. XXX’s question of interest, we set up a linear regression model using the available data collected by Mr. XXX and used the model to predict the selling price of the chosen house. Then, we calculated a 95% prediction interval to estimate the interval in which the future selling price will fall with 95% certainty. According to the results, the predicted selling price of the chosen house is $682,256 with the prediction interval ($544,873, $819,638).

In the dataset, some values are missing. We need to know why data are missing to make a more accurate prediction. Also, the size of the dataset is small. The results of the analysis may not have a high accuracy. This may be a limitation of this study. A large sample size is required to make a narrower prediction interval.

**Objectives**

In the process of buying a house, it is important to make a good purchase offer. Putting in a high offer on a house may risk wasting money and putting in a low offer may risk losing the house. Mr. XXX wants to make a good purchase offer to help his client get the dream house at the best possible price.

Therefore, the question of interest of Mr. XXX is the selling price of his client’s chosen house based on the recently house sell data in Hartwood and a reasonable range for the likely selling price of the house.

The purpose of this report is to answer Mr. XXX’s question of interest by using a linear regression model to predict the selling price based on the recently house sell data in Hartwood and get a 95% prediction interval of the predicted selling price. In this report, we analyze the data, and find out the predicted selling price of the chosen house and the 95% prediction interval of the predicted selling price.

**Data Collection and Methods of Analysis**

Mr. XXX collected some data on 29 homes that have recently sold in Hartwood. In the dataset, there are 8 predictors:

1. “DaysOnMarket”— number of days the house spent on the market before being sold
2. “Age”— age of house (years)
3. “HouseSize” — size of house (square feet)
4. “LotSize” — size of lot (square feet)
5. “Bedrooms” — number of bedrooms
6. “Bathrooms” — number of bathrooms
7. “Renovated” – indicator of whether the house has been recently renovated (Yes or No)
8. “Suite” – indicator of whether the house has a suite (Yes or No)

In addition, the response variable, “SellPrice”, is the selling price of house. The selling price is a numeric variable that may take on any value in an interval.

The chosen home has the following characteristics:

1. 14 days on the market so far
2. 20 years old
3. 3000 square feet
4. 7200 square foot lot
5. 6 bedrooms
6. 4 bathrooms
7. Not recently renovated
8. No suite

However, some values are missing in the dataset. The nature of the missing values (why data are missing) is not provided so we assume that these missing values are missing completely at random (MCAR). MCAR means that the probability that an observation is missing is unrelated to the value of the observation or to the value of any other variables. To deal with the missing data, we consider two methods: Listwise Deletion and Regression Imputation. Listwise Deletion deletes all the observations with any missing predictor variable values. By using Listwise Deletion to deal with the MCAR data, the predicted selling price will be unbiased. Regression Imputation replaces the missing values with predicted scores based on the non-missing data. A nice feature of Regression Imputation is that it produces a lower mean squared error (MSE) of the predicted selling price than the Listwise Deletion method. Generally, the Listwise Deletion method is used to deal with MCAR data because of the unbiased estimate. However, we decided to use the Regression Imputation method to deal with the missing data in our case. Because the Regression Imputation provides a narrower 95% prediction interval, the prediction interval is more useful to answer the question of interest of Mr. XXX.

When the dataset was completed, we used a linear regression model to predict the selling price of the chosen house and calculated a 95% prediction interval of the predicted selling price to address Mr. XXX’s question of interest. We modeled the selling price as a function of all the 8 predictors. After developing this model, the predictor variable values of the chosen house were substituted into the fitted model to get the prediction of the selling price. Then, we calculate the 95% prediction interval of the predicted selling price. The prediction interval is an interval in which the future selling price will fall with 95% certainty.

**Results**

According to the results of the regression analysis, the predicted selling price of the chosen house is $682,256. The 95% prediction interval for the predicted selling price is ($544,873, $819,638), which means that we are 95% confident that the selling price for the chosen house is between $544,873 and $819,638.

However, the analysis may not fully answer the question of interest as the prediction interval is wide. The wide prediction interval may be caused by the small size of the dataset. Decreasing sample size increases the residual standard error.

We make four assumptions to justify the use of linear regression models for purpose of prediction. These four assumptions are:

1. Linear relationship between the dependent variable and the independent variables
2. Statistical independence of the errors
3. Constant variance of the errors
4. Normal distribution for errors

We did not find any violation of the assumption of the regression model. Hence, these four assumptions might be appropriate.

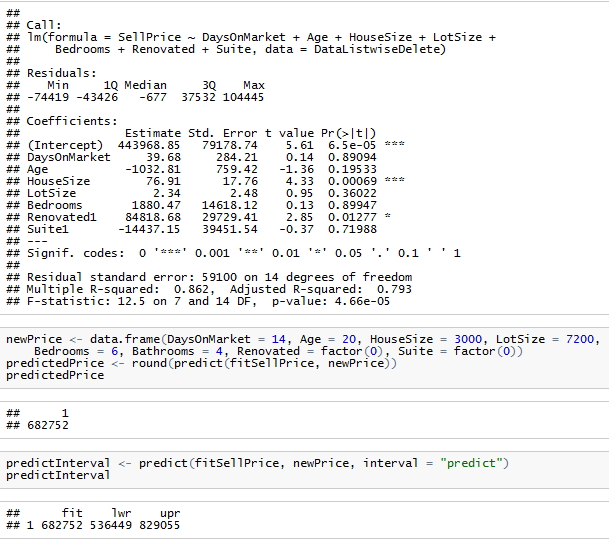
**Summary and recommendations**

Mr. XXX’ client wants to make a good purchase offer (close to the likely selling price) to get the dream house at the best possible price. We are helping Mr. XXX to guess the selling price of the chosen house and to give a reasonable range for the likely selling price of the house. According to the results, the predicted selling price of the chosen house is $682,256 with the prediction interval ($544,873, $819,638). The predicted selling price of the chosen house is reasonable.

However, the prediction interval of the predicted selling price may be too wide to fully answer the question of interest. The small size of the dataset may be the limitation of this study. To avoid this problem, a larger sample size is required. A further study is needed to investigate other methods to increase the sample size.

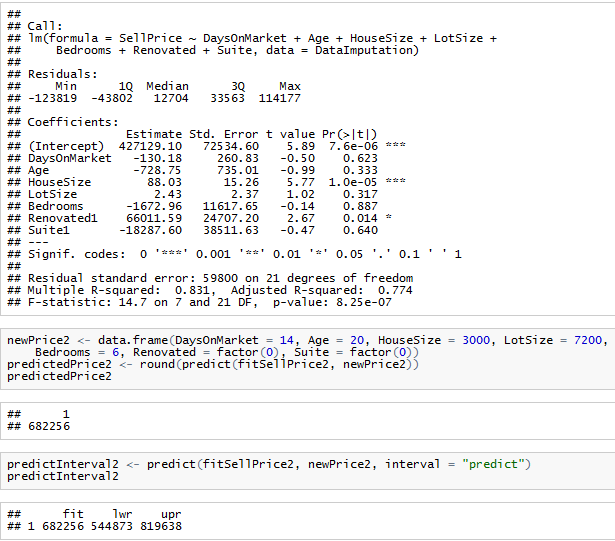
**Appendix**

Below is the result by using the Listwise Deletion to deal with the missing data.



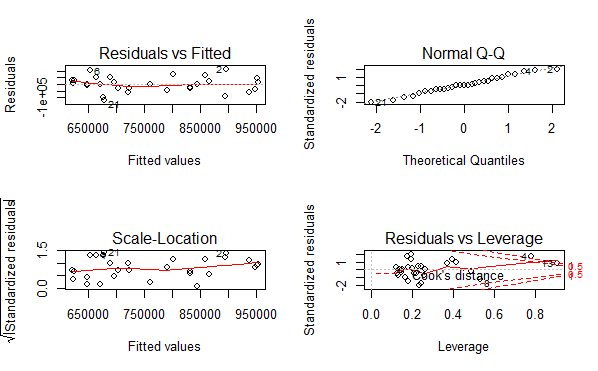
The predicted selling price of the chosen house is $682,752 and the 95% prediction interval for the predicted selling price is ($536,449, $829,055)

This is the result by using the Regression Imputation to deal with the missing data.



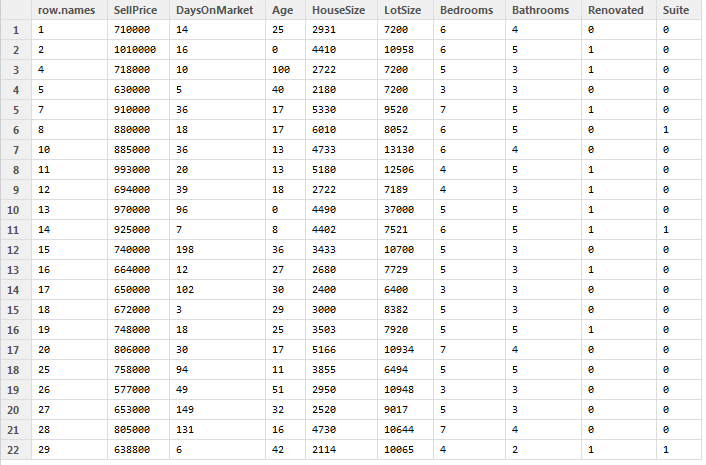
The predicted selling price of the chosen house is $682,256 and the 95% prediction interval for the predicted selling price is ($544,873, $819,638)

Check assumption:



Based on the results of the fitted linear regression model, the selling price may have a linear relationship with all the predictors as the coefficient of determination “R-squared” is 0.831. Also, there is no obvious pattern in the plots. Therefore, the assumptions may be appropriate.

Dataset after Listwise Deletion:



Dataset after Regression Imputation:

