House Prices Multiple Regression Analysis

David Casado  
*Statistics for Data Data Analytics*

*National College of Ireland  
Dublin, Ireland  
x19204221@student.ncirl.ie*

*Abstract* — This study describes the steps to estimate a multiple regression model useful to predict the sale price of the houses in a US region. Descriptive statistics is used to give an insight of the data used and relationship between the variables. Multiple linear regression model is built on the second section and its e*fficacy is compared with the Gauss–Markov Assumptions for multiple linear regression.*

Keywords—multiple regression, population, linearity, correlation, homoscedasticity, residuals, variance

# Introduction (*Heading 1*)

This paper analyses the impact of different house characteristics and sales prices in a particular region in US. The dataset contains the sale price and characteristics of all houses which were sold in a given region in US. A descriptive analysis carried on the first portion of this project will give a first impression on how the houses characteristic affects the houses sale price. In addition, Data visualization will help to find trends and relationships between the different variables that might be difficult to find otherwise. Once the descriptive analysis of the data has been complete, a multiple lineal regression model will be built on the second portion of the project with the aim to have a better understanding, or even predict, how houses prices will vary depending on it characteristics

# Description of the Data

## Preparing the data for analysis

Firstly, integrity of the data in the dataset is analyzed to find any possible errors or missing values that could affect the analysis of the data. The Houses Price dataset does not present any NA value that we have to consider for our analysis so the conclusion is that the data is in good shape to start the analysis.

## Data description

The House price data set represent the population of houses on sale in a US region and a total of 1728 observations with 16 columns that corresponds to the following variables: **price** (price of the house in US Dollar), **lotSize** (in acres), **age** (age of the house in years), **landValue** (value of the land in US Dollars), **livingArea** (in square feet), **pctCollege** (percent of neighborhood that graduate from college), **bedrooms** (number of bedrooms), **fireplaces** (number of fireplaces), **bathrooms** (number of bathrooms), **rooms** (number of rooms), **heating** (type of heating system), **fuel** (fuel used for heating), **sewer** (type of sewer system), **waterfront** (whether the property has a waterfront), **newConsctruction** (whether the house is new construction) and **centralAir** (whether the house has central air).

## Descriptive statistics

Descriptive statistics refers to the methods of organizing, summarizing and presenting data in an informative way. Measures of central tendency are often referred as averages while measures of dispersion like range and standard deviation describes the spread of the data. Both measurements should be considered together as decisions made based on only one of the might arrive to inaccurate or misguided conclusions [1].

Descriptive statistics for the following variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | price | landValue | age | livingArea |
| Mean | 211967 | 34557 | 27.9 | 1755 |
| Median | 189900 | 25000 | 19.0 | 1634 |
| Range | 5000-775000 | 200-412600 | 0 – 225 | 616– 5228 |
| SD | 98441.39 | 35021.16 | 29.21 | 619.93 |

Table 1 Descriptive Statistics price, landValue, age and livingArea

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | bathrooms | lotSize | bedrooms | rooms |
| Mean | 1.9 | 0.50 | 3.15 | 7.04 |
| Median | 2 | 0.37 | 3.00 | 7 |
| Range | 0 – 4.5 | 0 – 12.2 | 1 – 7 | 2 – 12 |
| SD | 0.65 | 0.69 | 0.81 | 2.31 |

Table 2 Descriptive Statistics for variables bathrooms, lotSize, bedrooms and rooms

As the main objective of the project is to investigate how different characteristic of the houses influences the final price, the price distribution will be investigated on first instance.

Figure 1 represent the Histogram for the variable price. It can be observed that the distribution of the variable price is non-symmetrical or positively skewed to the right. In distributions which are highly skewed, the mean is not a good representative measure and we need to look at other parameters like its median or mode to obtain a more representative measure [1]. In addition, skewed distributions are a signal that a larger than desired number of outliers are included of the data.

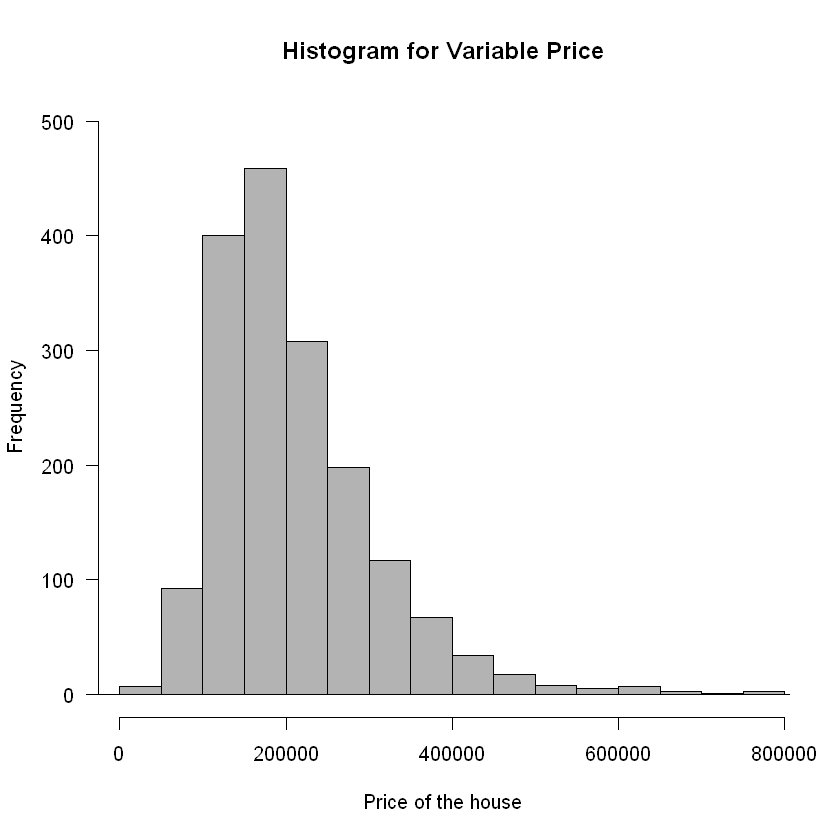


Figure 1: Histogram of the distribution of prices in the dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| 5000 | 145000 | 189900 | 211967 | 259000 | 775000 |

Table 3 Central tendency summary for variable price

In addition to this summary, the mode for variable price is 120000 USD. What is interesting about of the outliers presented in the variable price is that every one of them are given when the house falls under the category Not New Construction as it can be seen on Figure 2.

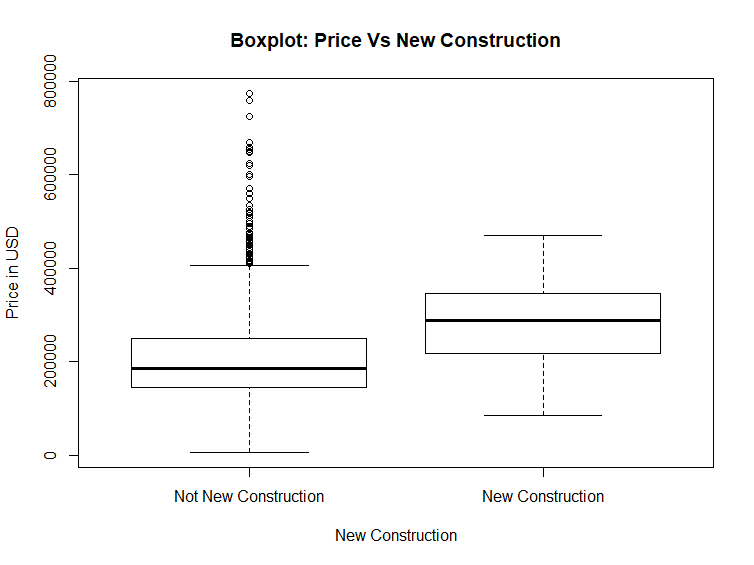


Figure 2: Boxplot for Price Vs. New Construction.

The fact that all outliers value for variable price occurs when the house is not new lead to conclusions like the price of the new houses are regulated by market standards while the price of houses on the market which are not new might be influenced by other factors like the cost of improvements made, sentimental value or a surge on the house price inflation. It is to expect that the price of a new house will be higher than the price of existing house with same characteristics unless compared with one of the outliers.

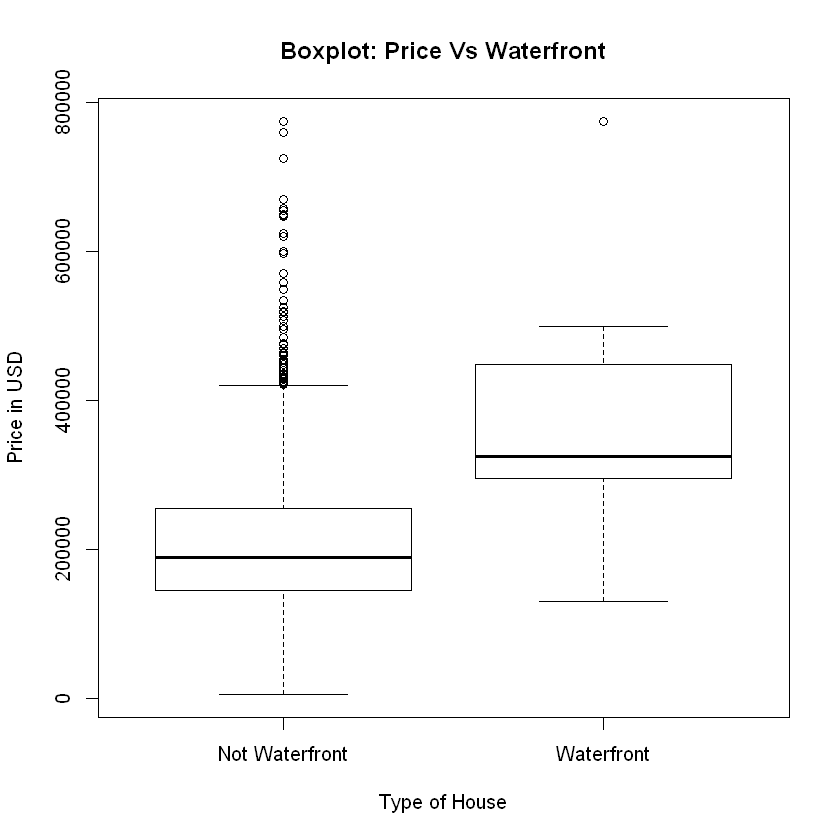


Figure 3: Boxplot for price Vs Waterfront

Computing the t-test with *t (df= 1727, p<0.001) = 90* the hypothesis that both variables have the same mean can be rejected. As observed in Figure 3, houses with waterfront tends to have a higher price that house without a waterfront.

# MULTIPLE LINEAR REGRESSION

Simple linear regression has the limitation of predicting the response variable value based on only one predictor variable. Multiple linear regression will allow to model a relationship between the response variable with more than just one predictor variables. For this analysis, the response variable will be the house price. The data used for the Multiple Linear Regression will be the same that was introduced on the previous chapter of this report and the first objective of constructing a multiple regression model would be to detect whether a relationship between response variable and the predictor variable exists. If the relationship exists, next step will be to decide which are the predictor variables that have more effect on the response variable. Each predictor variable will be given a slope coefficient. The multiple linear regression model will take the form:

Y = β0 + β1X1 + β2X2 + ··· + βpXp + ε

Where:

* β0 is the intercept, value of X when all Xp are zero.
* βi amount that Y changes when a Xi changes by one unit.
* ε errors. Coefficients βi should minimize errors.

## **Model building**

Given the number of variables in the data set, *Forward Selection* model will be chosen to find the predictor variables that will have a bigger effect on the response variable price. The starting point will be a null model with an intercept but nor predictor.

Next step is to create a new model with only one variable which will be the variable with the maximum R-squared value among all variables. This process keeps repeating until the Adjusted R Square does not increase after a new variable is added to the model. In addition, the model needs to be checked every time a new variable is added as the p-value for a one the variable already in the model could change after a new variable is added.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Adj R-Square | RSE | F (variables, df) | P-value |
| 0 | - | 234000 | - | - |
| 1 | 0.507 | 69100 | 1780 | P < .001 |
| 2 | 0.522 | 68100 | 944 | P < .001 |
| 3 | 0.618 | 60800 | 932 | P < .001 |
| 4 | 0.618 | 60800 | 700 | P < .001 |
| 5 | 0.622 | 60500 | 570 | P < .001 |
| 6 | 0.622 | 60500 | 475 | P < .001 |
| 7 | 0.626 | 60200 | 413 | P < .001 |
| 8 | 0.635 | 59500 | 377 | P < .001 |
| 9 | 0.648 | 58400 | 354 | P < .001 |

Table 4 Linear Models results

Firstly, variable ‘livingArea’ gets added to the model. At this point the model is a single linear model with only one variable as predictor for the price of the house. It reduces the residual standard error of the previous Null model and obtain a F(1, 1724 DF) = 1780 and a Adjusted R-squared = .507. Independent variables with the highest correlation with the response variable are added to the model. Next independent variable with the highest value is ‘centralAir.Yes’ which gets added to the model.

Variable ‘fireplace’ gets added to the model on model number 6. However, it does not have an impact on the total Adjusted R-squared which remains .622 and obtained a p = .40845 which is considered a high value. The addition of this variable to the model is questionable so the model building will continue without this variable.

Finally, the last two variables that are added to the model and that have an effect on the response variable price are waterfront.Yes. and age. The addition of waterfront.Yes lead to believe that houses with waterfront will have a higher price than other houses with similar characteristics but no waterfront. The linear model obtained after running forward selection and include the important variables and an increase on the R-squared adjusted value will also be the model with the lowest RSS. Table 6 below contains the summary of the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std Error | t-value | p-value |
| Intercept | 12616.81 | 6385.20 | 1.98 | 0.048 |
| livingArea | 70.15 | 4.51 | 15.56 | P < .001 |
| bathrooms | 22880.08 | 3308.67 | 6.92 | P < .001 |
| landValue | 0.9224 | 0.0462 | 19.95 | P < .001 |
| rooms | 3113.94 | 961.55 | 3.24 | P =.001 |
| bedrooms | -7695.50 | 2530.83 | -3.04 | P = .002 |
| lotSize | 7095.33 | 2045.83 | 3.47 | P < .001 |
| age | -148.86 | 54.89 | -2.81 | P = .007 |
| newConstructionYes | -42009.8417 | 7115.778 | -5.90 | P < .001 |
| waterfront.Yes | 121022.8816 | 15363.2722 | 7.88 | P < .001 |
| centralAir.Yes | 12954.4722 | 3237.5669 | 4.00 | P< .001 |

Residual standard error: 58300 on 1717 DF. Multiple R-Squared: 0651, Adj R-squared: 0.649, F-statistic: 320 on 9 and 1717 df, p-value < .001

Table 5 Linear Model – Forward Selection

# Model check

This section will analyze the linear model established on the previous section by checking whether the independent variables selected explain the dependent variable price. The main objective is that he predictors in the model minimize the error variance

## **Collinearity**

Firstly, the correlation between independent variables will be checked. Independent variables in a multiple linear regression model should be independent. It could cause problems when interpreting the results. Variables have been added to model consider how it affected the model Adjust R-Square value without considered the correlation between them.

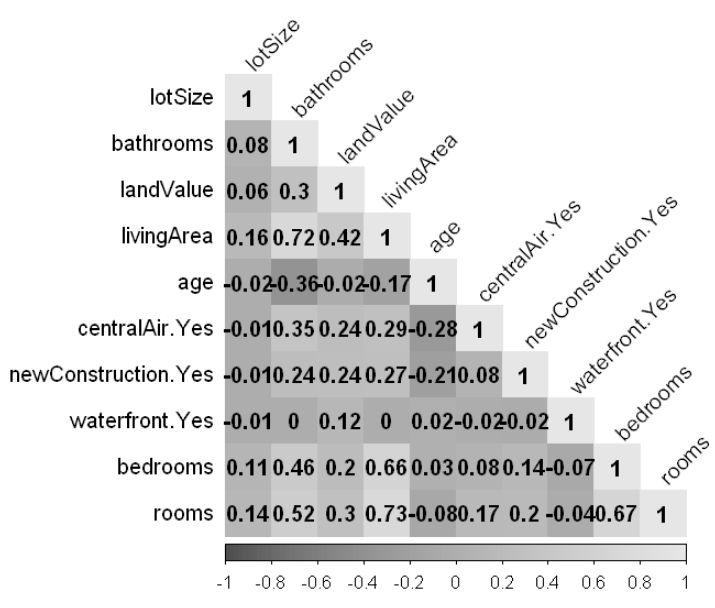


Figure 4 Correlation Matrix

For instance, Figure 4 reflects that variables livingArea and bedrooms are closed related to each other. This means that when one variable increases or decreases, the other will also increase or decrease at the same time. This fact will complicate to detect the individual effect on the response variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | EST | Std Err | t-Value | p-value |
| (Intercept) | 36667.90 | 6610.29 | 5.55 | p< 0.001 |
| livingArea | 125.40 | 3.53 | 35.56 | p< 0.001 |
| bedrooms | -14196.77 | 2675.16 | -5.31 | p< 0.001 |

Table 6 Correlation between livingArea and bedrooms

When both variables are included in the model, the coefficient of bedrooms can be considered as counterintuitive as it is negative. This means that the price of the house will decrease by -14196.77 for every bedroom that is added to the house. The correlation between livingArea and bedroom is r = .66 which could imply that both variables could have redundant information. One of the simplest solution when facing collinearity is to drop one of the problematic variables [2]. In this instance, variable bedroom will be removed from the model.

Similar circuanstances occur between variables livingArea and rooms. The coefficient of correlation between these two variables is *r = .73.* Therefore, variable rooms will also be removed from the model.

Without knowing more information about the characteristics of bedrooms and rooms, it will have to be assumed that the best predictor among this variables to predict the price of the house will be livingArea.

In addition, variable bathrooms will be removed from the model as it has a significant strong correlation level with livingArea of *r = .72*

Lastly, variable newConstruction.Yes will also be removed from the model as it present a counterintuitive behaviour as in the presence of the rest of the variables, it will decrease the house price by -37.582. Variable waterfront.Yes has also been removed as it is more significant for more expensive houses (outliers will be removed during assumptions check section).

At this point, the variable that form the linear model are livingArea, landValue, waterfront.Yes and centralAir.Yes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std Error | t-value | p-value |
| Intercept | 3019.778 | 4971.363 | 6.06 | p < .001 |
| livingArea | 85.246 | 2.703 | 31.54 | p < .001 |
| landValue | 0.937 | 0.047 | 19.93 | p < .001 |
| age | -220.681 | 52.998 | -4.16 | p < .001 |
| centralAir.Yes | 16334.848 | 3325.330 | 4.91 | p < .001 |

Table 7- Residual standard error: 61100 on 1723 DF. Multiple R-Squared: 0.615, Adj R-squared: 0.614, F-statistic: 689 on 4 and 1723 DF, p-value < 0.001

Written as equation, the model can be described as follows:

*price = β0 + β1  (living Area) + β2(landValue) -*

*β3(age) + β4(centralAir.Yes) + ε*

*with β0 = 3019.778 β1= 85.246 β2= 0.937 β3 = -220.681 β4 = 16334.84*

## **Transformations**

The result of previous section is a linear model with independent variables that are not significantly correlated between them. However, there are few steps to perform before the model is ready to meet all Gauss–Markov Assumptions for Multiple Regression.

* Outliers: as it has been described earlier, variable price contain a number of outliers which, given the fact that the dataset contain over 1500 rows, these will be removed.

* The histogram for response variable price presents a right positive skewness. Response variable will be converted to it square root form in order to reduce the skewness.
* Assumption of constant variance is not met with the current variables. In order to resolve this exception, the independent variables will be convert to its own squared root value. In the scenario that this conversion will not resolve the heteroscedasticity, there will be an important evidence that a significantly important variable might be missing from the model.

After mentioned transformations have been completed and the dataset has been modified accordingly, the linear model table is represented as shown on table 8:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std Error | t-value | p-value |
| Intercept | 131.57 | 9.4451 | 13.93 | p < .001 |
| livingArea | 6.4991 | 0.2206 | 29.47 | p < .001 |
| landValue | 0.3507 | 0.0197 | 17.76 | p < .001 |
| age | -4.0503 | 0.5797 | -6.99 | p < .001 |
| centralAir.Yes | 11.1934 | 2.9560 | 3.79 | p < .001 |

Table 8 - Residual standard error: 53.1 on 1639 DF. Multiple R-Squared: 0.586, Adj R-squared: 0.585, F-statistic: 579 on 4 and 1639 DF, p-value < 0.001

From this point onwards on this report, the linear model used to check the Gauss–Markov Assumptions for Multiple Regression will be the version with dependent and independent variables transformed represented by its equivalent equation defined:

*with β0 = 131.57 β1= 6.5 β2= 0.350 β3 = -4.05 β4 = 11.1934*

Each assumption will be presented and a conclusion explained when contrasted with the linear model.

## **Linearity**

The relationship between response variable price and the predictors is assumed to be linear. Residuals plots are a useful tool to identify non-linearity [2]. Figure 5 represent the plot for residuals value against the predicted values of the dependent variable. The red line is used to identify any relevant trend on the data. On this case, the red line is mainly horizontal and shows very little pattern in the residuals which is a good signal for a linear relationship.

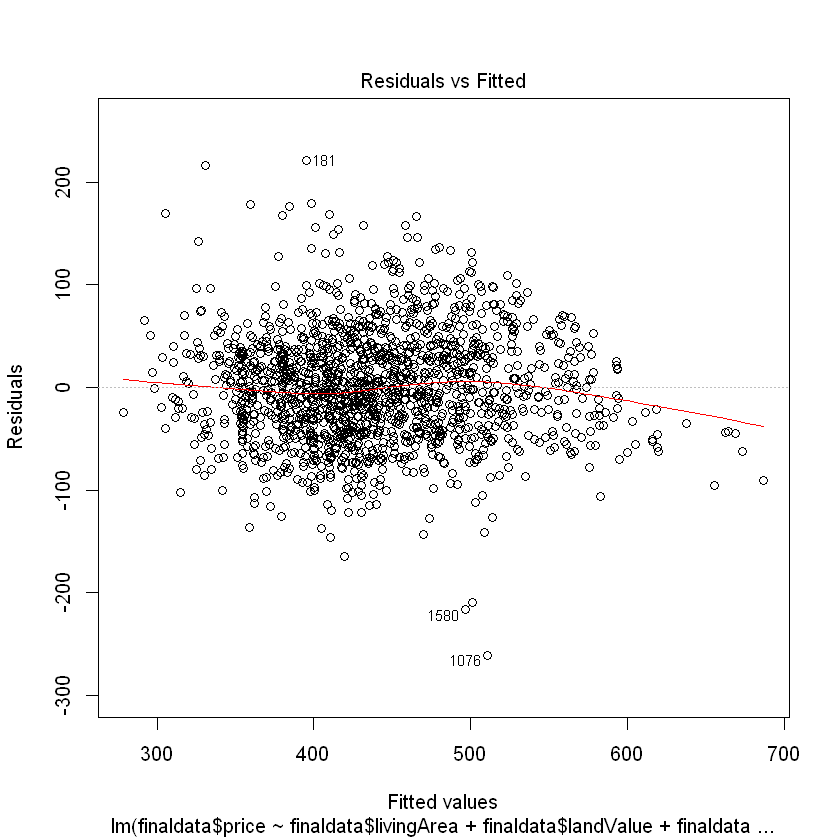


Figure 5- lm(price ~ livingArea + landValue + age + centralAir.Yes

## **Errors are normally distributed**

Figure 6 below represent a QQ Plot of the residuals in order to examine whether the errors are normally distributed. Almost every residual falls on a straight line with only few exceptions on both ends of the line.

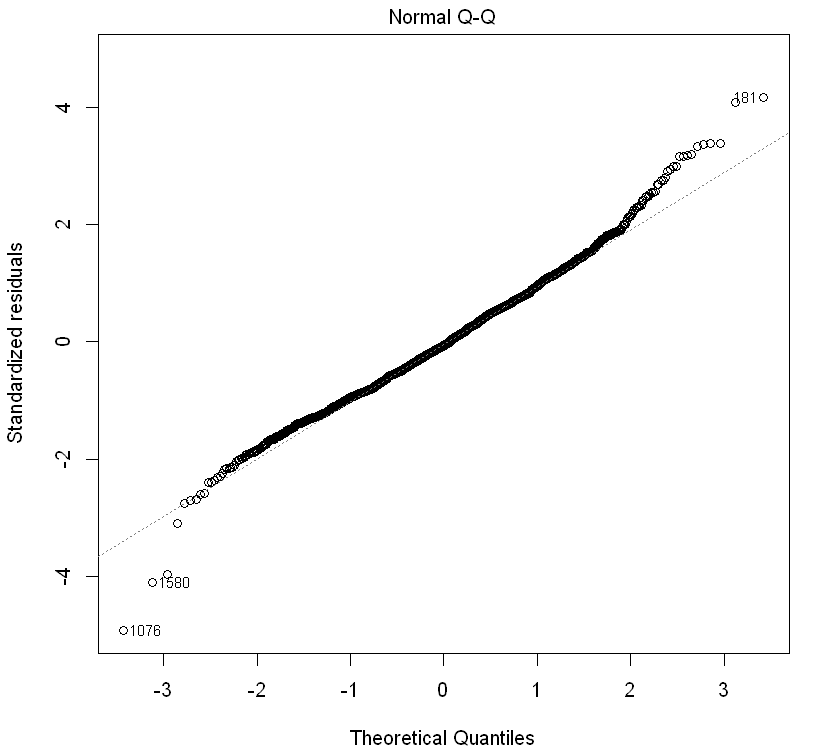


Figure 6 - lm(price ~ + livingArea + landValue + age + centralAir.Yes

## **Homoscedasticity**

Heteroscedasticity, or non-constant variances in the errors, is the violation of the assumption of homoscedasticity and it can be detected when the magnitude of the residuals tends to increase with the fitted value [2]. As it can be observed on Figure 7, all standardized residuals are below 2 with the exceptions of three points which are highlighted on the plot. All point are randomly distributed around the red line without any significant pattern, which is a good signal.

**ncvTest - Breusch-Pagan test**

The ncvTest run on the final linear model returned a relative high p-value of p = .53 (chi-square = 0.39, df = 1) which indicates that there is no reason to reject the null hypothesis that there is constant variance or homoscedasticity.

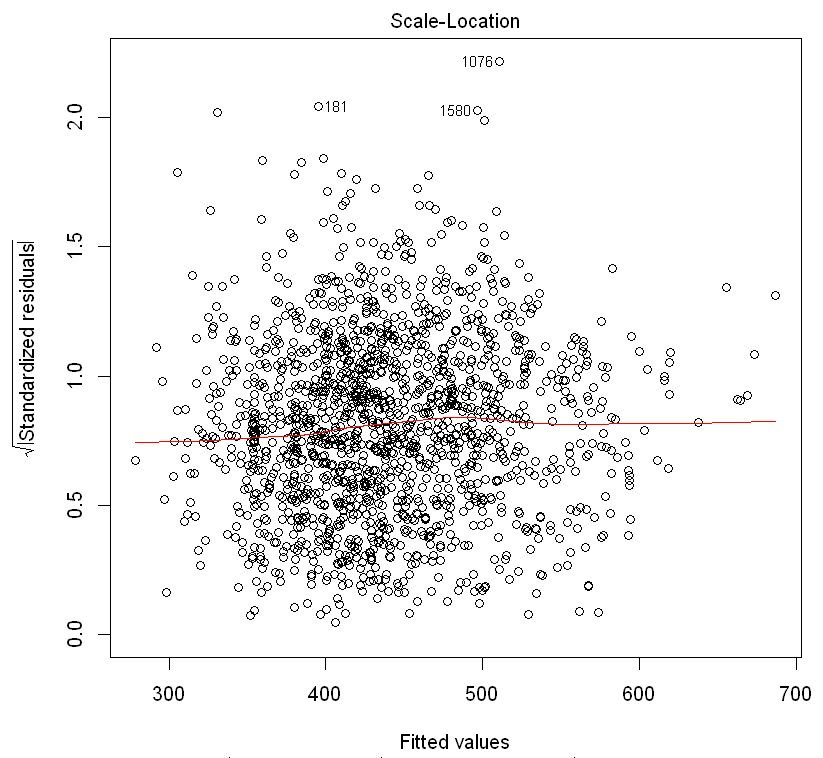


Figure 7- lm(price ~ + livingArea + landValue + age + centralAir.Yes

## **No autocorrelation between errors**

An important assumption of the linear regression model is that are uncorrelated. This will imply that the error made on one observation is completely independent of error made on a different observation.

**Durbin-Watson test**

The Durbin-Watson test notifies about whether the assumption of independent errors is justifiable. The closest the test result is to 2, the better indication of no autocorrelation detected. From 0 to close to 2 will mean positive autocorrelation and from close to 2 to 4 a negative correlation.

The Dublin-Watson returned a value of 1.97 with a p-value of p = .63 which indicates that there is practically no evidence of a relationship between the residuals in the model. In other words, the error made on one observation will be independent of the error made on a different observation.

## **Multicollinearity**

There occasions that correlation exist between three or more variable and it is not detected by the correlation matrix used previously. This “multiple” correlation is called multicollinearity and it is a more efficient way to detect hidden correlation between several variables [2]. It computes the variance inflation factor (VIF) which calculates how much a variable is contributing to the standard error.

|  |  |
| --- | --- |
| livingArea | 1.267 |
| landValue | 1.2 |
| age | 1.38 |
| centrailAir.Yes | 1.57 |

Table 9- VIF's values for independent variables

Absence of multicollinearity would be represented by VIF lowest possible value of 1. Values over 5 will indicate a problem of correlation among multiple variables. Based on the results obtained on the VIF calculation we can conclude that there is not multicorllinearity between the independent variables in the lineal model.

## **No influencial data points**

Considering that there is a large number of observation in the data set and that it is unknown how the data has been collected, it has been decided to remove outliers to reduce the impact of these on the fit of the model. These outliers can exists due to a number of different reasons like incorrect recording of the data or the use of wrong units.

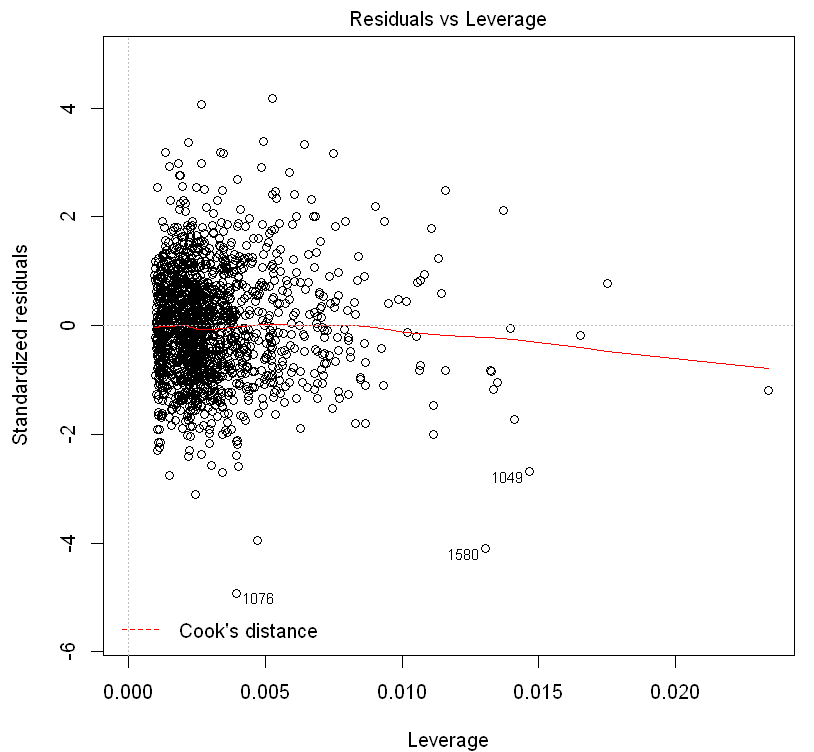
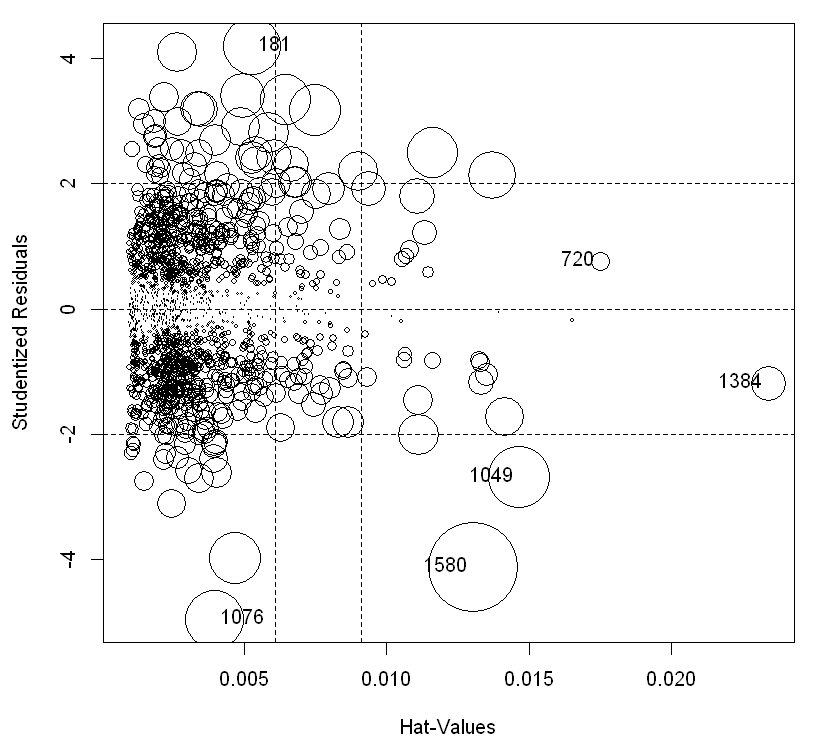


Figure 8- Leverage

Cook’s distance has been calculated for every observation in the data set. There is not any observation with a value of 1 or greater which it would have been an indicator to be considered influential. If there would have any observation with a Cook’s distance of 1 or larger, they could have been removed from the regression.

Figure 9 represent an influence plot of Studentized Residuals versus hat values. The area of the circle represent the proportion to the Cook’s distance value. In this case, the largest circle is number 1580 which correspond to a Cook’s distance value of .04459 which is lower than 1, therefore with not significant influence.

Figure 9 – Influence plot

# COnclusions and discussion

The overall evaluation of the fit of the linear model is based on its R-Squared adjusted value which is r: .586. This means that the 58.6 percent of the variation in the price of a house is explained by the regression model. The model F-statistic (4, 1639) = 579 with p-value < .001 this means at least one predictor has explanatory effect on the price variation.

As previously discussed, the dependent and independent variables were converted to it square root values.

This means that the square root of the house price is equal to the sum of the squared root values for the independent variables multiplied for their coefficient. This equation can be transformed back to its original form using the squared root values and then interpreted by squaring the result.

This study defines the linear regression model build and the assumptions that must be met in order for the model to be considered a valid model. Diagnostic of the assumptions have been completed by visualization of the residuals and analyzing the absence or presence of patterns. This is only a valid model of the many valid models that exists. The presence of some variables on the final model could be discussed base on its coefficients. In addition, observations highlighted on the influence plot represented on Figure 9 could be removed as the impact on the overall observation will be small. The existence of important variables that have been left out from the model could be studied and obtain different valid models as long as they meet the assumptions.

##### References

1. Douglas A.Lind, William G. Marchal, Samuel A. Wathen “Basic Statistics for Business & Economics” Mc Graw Hill, international edition 2005.
2. G. James, D. Witten, T. Hastie, R. Tibshirani “An Introduction to Statistical Learning with applications in R”. Second edition, Springer. 2013.