Predicting a House’s Sale Price

Multiple Regression Analysis

George Mavroeidis and Sebastian Campusano-Revuelta Professor Sandi Wing May Mak

201-HTH-05 section 00001

Probability and Statistics

***TABLE OF CONTENTS:***

1. **ABSTRACT . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2**
2. **INTRODUCTION . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3-13**

**2.1) Multiple Linear Regression Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3-7**

**2.2) Least Squares Estimation of the Parameters . . . . . . . . . . . . . . . . . . . . . . . . 7-9**

**2.3) Hypothesis test in multiple regression . . . . . . . . . . . . . . . . . . . . . . . . . . . 10-13**

1. **METHODS AND ANALYSIS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .13-16**

**3.1) Summary of Theoretical Concepts pertinent to the analysis . . . . . . . . . . . 13-14**

**3.2) Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15**

**3.3) Computer Algebra Methods Used. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16**

1. **RESULTS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14-21**
2. **DISCUSSIONS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21-22**
3. **CONCLUSION . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22**
4. **REFERENCES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23**
5. **Abstract**

The purpose of this comprehensive assessment is to use a multiple linear regression model in order to estimate a home's sale price according to a number of different variables; in other words the objective is to determine the regression equation. Once this has been achieved it can be compared to the actual sale price on the current market. The use of multiple linear regression models, least squares, variance, hypothesis test, p-value and VIF help achieve an adequate regression equation which can then be used to approximate a home’s sale price as previously mentioned. After analyzing the data pools through different factors such as; outliers, p-values and VIF, the final R-squared adjusted is 91.98%; a decent increase from the initial R-squared adjusted of 85.07%. The final regression equation is ; **SalePrice** = 94326 + 1.858 LotArea - 553.2 Age + 19.87 Basement Finished Sq. Ft + 28.74 BasementUnfinishedSF + 53.37 GrLivArea - 6641 BedroomAbvGr - 25742 Kitchens + 7581 Fireplaces + 9916 GarageCars. This equation allows someone to get an approximate value of a home according to the number that each variable contains (ex: # of bedrooms). The use of linear regression model helps identify the relation of independent variables to an dependent variable; in this case the relation between lot area, age etc and the home’s sale price.

1. **INTRODUCTION**

In the contemporary world, housing prices vary and their ranking depends on their properties associated with it. The price is mostly determined by the relationship of “supply and demand”, which is what the service provides with what the customer desires. However, a balance between the supplier and the demander has to remain constant or else the prices fluctuate and predictability of housing prices becomes inaccurate. Therefore, when demanding increases, which means when people buy houses more often, the house prices have to increase in order to maintain a stable market. The factors of each house determines its final price.

The most important factor in a property is its location. This is what customers are looking first when it comes to buying a house. This also includes the conditions of the neighbourhood in the area. The location must be ideal, where different community lots are near the house. For buyers with children, the house usually is located near a fairly decent school. For workers, their preferable location is near their workplace or near any social centers, such as malls and parks. Another important factor is the number of bedrooms and bathrooms the property contains. The more and bigger the bedrooms are, the greater the cost of the house. Therefore, a family that requires a great amount of bedrooms and bathrooms has to search for a house with the needed rooms, meaning the house needed will have a greater expense.

In general, housing prices become complicated due to the uncontrollable factors that may change the value of a property over time. The predictable prices are not always fairly accurate and sometimes, being excessively accurate can lead to outliers, making the property unrelated to the majority of the data. In order to result in an accurate selling price, the sample and the factor list must be big enough to produce as many variables as possible to be tested. This also takes into account the amount of data that will be erased due to the parameters setup and the removal of outliers. When the resulting selling price is predicted, a formula will be given, indicating the proportion of each factor affecting the selling price. This shows also which factor is most considerable when consumers buy a house.

In this comprehensive assessment, the selling price of a house will be predicted using multiple regression analysis. Factors such as ground living size and the amount of bedrooms and bathrooms will help determine the average selling price of a house.

***2.1) Multiple Linear Regression Model:***

Regression analysis “is a statistical process for estimating the relationship among variables”. It helps identify the relation of independent variables to a dependent variable. A regression model that includes multiple regressor variables (more than one) is named “Multiple Regression Model”. The general equation for “Multiple Linear Regression Model with k regressor variables” is;

Where,

Y=Dependent variable or response correlated to k independent/regressor variables

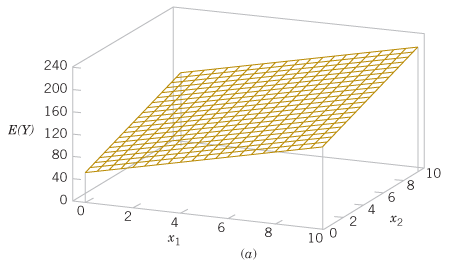
j= regressor coefficients where j=0,1,2,…,k

xj= regressor variables

In words; The parameter j shows the anticipated change in response “Y” per unit change in xj when all the remaining regressors are kept constant(xi ; i ).

The previous function is called a “Multiple ***Linear*** Regression Model” due to the fact that the above equation is a linear function with unknown parameters, β1, β2, until βk . These parameters are called regression coefficients and the “x” are the k variables of the linear regression model. These functions are usually used as estimating or approximating functions because the actual relationship between the dependent variable (Y) and the regressor variable (Xj) is unidentified. However, if there are a reasonable amount of regressor variables then the linear regression models gives an acceptable approximation. Some models contain **interaction effects** that are relatively common. Two variables (1 and 2) have an interaction if the effect of variable 1 changes depending on the level/magnitude of variable 2. It is important to note that interaction effect can occur with more than two variables. Interaction effects happen regularly when studying and analyzing real-word systems (ex: housing model).

The linear regression model can be graphed in a multiple dimensional system. For example, when the equation has two variables x1 andx2 , it can be represented in a three dimensional coordinate system. If the equation E(Y)=50+10x1+7x2 is to be graphed, it will result in a plane with a linear path.



* This is the regression plane of the equation E(Y)=50+10x1+7x2.

Since βο=50, the plane will have an E(Y)-intercept of 50. In reality, no model is 100% accurate. Thus, certain errors exist. The above ideal example was assumed that there were no errors in the model, meaning E(e)=0. The equation becomes:

* e= error term

More complicated models are not always linear, but they can be expressed using this system. This means the model may contain polynomials that form a curved structure with parabolic form. For example, if a model has a polynomial with power of 4, it will look like:

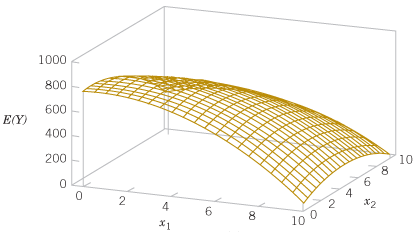
The variables can be replaced with the multiple regression terms, where x1=x, x2=x2, x3=x3 and x3=x4:

Sometimes, variables may interact and affect one another. For example, on a house`s cost, the number of rooms may affect the size of living space as well, resulting in more complex models. For a second-order model with interactions, the equation is represented like this:

* Note : Both terms x1 and x2 are represented in both first and second order and the x1x2 term represents the interaction in between.

Now replacing the terms with: β11= β3, x11=x3, β12= β4, x22=x4, β22= β5, x22=x4, the equations results in the form:

This means that the equation will not result in a linear form anymore and when graphed, the plane will be curved due to the different interactions between the variables:



* The shape isn’t always like that since interactions may vary differently between the variables.

***2.2) Least Squares Estimation of the Parameters:***

Least squares method can be used in order to approximate the regression coefficients (j) in a multiple regression model. If we assume that there are n>k observations and xij denotes the ith observation or level of variable xj, the estimation of each variable’s coefficient can be found. Least squares is used to find the best model for the given data with the least errors. The general equation is:

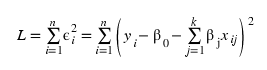


To simplify notations, it can be written as summation:



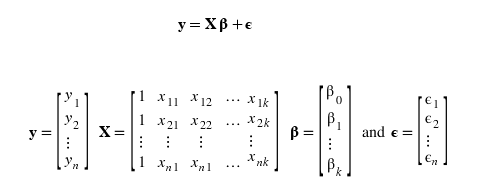
yi

The least square function L represents the errors produced by the data, giving the regression coefficients:



* The L has to be minimized in order to obtain the most accurate results possible.

When having to analyze a numerous amount of data, it gets crowded and calculations become tedious. Matrix notations simplify the calculations and the notations for least square analysis. The equation model can be written in a form of matrices:



The least square equation can be represented as a matrix form as well. The coefficients that are being estimated are expressed in a vector form, which is the solution for the equation:



* X= variable matrix
* ϐ= coefficient vector

The vector can be found by solving the normal equation. The normal equation minimizes the differences between each variable’s interactions with another.

The solution is:

* X’=transpose vector of variable vector X

Instead of working on countless variables and equations for each interaction, a matrix minimizes the work done and still provides the same results. These days, computer programs are able to solve a vast amount of data by only plugging in the equation once, which is able to solve for the rest of the coefficients.

When working with data, it is important to estimate the variance. Variance displays how far data are from the average, which shows the flow of the data. This is important, since variance is used to determine whether a variable is eligible enough in order to remain in the data pool, after setting a barrier (ex: deciding on what P-value the variables are considered clean). Variance is also labeled as a logical estimator. To obtain the variance in multiple regression models, the following equation is used:

* n-p= error or residual degrees of freedom
* SSe= sum of all errors in between variables

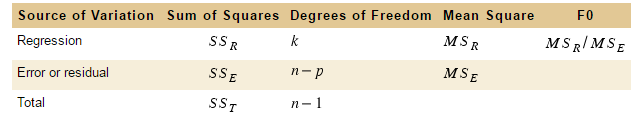
***2.3) Hypothesis test in multiple regression:***

There are hypotheses test conducted on different multiple regression models in order to determine its adequacy. This test is called “significance of regression”, which identifies if any linear relationships exist between the output y and the variables.

The null hypothesis determines the level of contribution of the variables to the model, where for at least one j, the coefficient. If hypothesis Ho is rejected, the variables are said to drastically affect the model. The sum of squares denotes a “measure of variation or deviation from the mean” (average). The summation of the squares of the differences from the mean is used to calculate the “sum of squares”. Two important conditions are taken in consideration when analyzing or calculating sum of square; the first is the “sum of squares so the deviation from the mean and secondly the randomness or error of the model. With this in mind, the total sum of squares can be expressed as the following;

* SST= total sum of square
* SSR= sum of squares due to the regression model
* SSE= sum of squares due to error

To test the hypothesis Ho, there is an equation that determines if the hypothesis is true. If it is true, the number of degrees of freedom for the chi-square random variable SSR/σ2 is the same as the number of the variables in the model. A similar process can be done for the SSE/σ2, which is a chi-squared random variable as well.



* Table that illustrates the variance used to test the level of significance for a multiple regression model. (Taken from: https://edugen.wileyplus.com/edugen/courses/crs7600/ebook/c12/bW9udGdvbWVyeTk3ODExMTg1Mzk3MTJjMTItc2VjLTAwMTAueGZvcm0.enc?course=crs7600&id=ref)

The following equation determines the significance of regression in a model:

* This provides the proportion of the regression’s and the error’s mean square, giving the significance of regression of the model.

When constructing a model, it is important to evaluate its accuracy. The coefficient of multiple determination R2 and adjusted R2 is used to determine a model’s statistical strength in percentage. The closer the percentage is to 100% (or 1 in the formula), the more accurately the data fall into the regression line of the model. It is important to note that the R2 value rarely decreases when adding values. If statisticians decide to set a minimum percentage on a data analysis, it is preferable to add variables in order to produce a better and accurate model. The formula for the R2 is the following:

* SST= total sum of square
* SSE= sum of squares due to error
* (n-p)=degrees of freedom for error
* (n-1)=degrees of freedom for total
* n=sample size

It is important to note that this is the formula for the adjusted R2 due to its better accuracy and simplicity. When using the R2 formula, the amount of regressors is determined, due to its affection by the terms added. In this comprehensive assessment, the R2 value determines whether it is right to add or remove certain house properties that determine the predicting house price.

In this comprehensive assessment, additional terms will be used to determine the housing model’s accuracy. The P value approach determines the probability a model is in favor of the null hypothesis Ho or the alternative hypothesis H1. When listing the data, there will be different probabilities for each term. When these probabilities are found, it should be noted whether the P-value is bigger, smaller or equal with the probability of rejecting the null hypothesis, “α”. There are two cases that determine the strength of a term:

* P-value is less or equal than “α”: reject null hypothesis Ho and move on with the alternative hypothesis H1. (sufficient data)
* P-value is greater than “α”: fail to reject null hypothesis Ho (insufficient data)

When testing a house model, a limit to the P-value will be stated in order to remain with the best terms needed for the model. For example, if the amount of pools in a house result in a Pvalue greater than “α=0.05” probability, the variable that determines the amount of pools will be removed from the model in order to construct a more accurate variable structure for the model.

Another factor that determines a variable’s strength in the model is the Variance Inflation factor or VIF. The VIF determines the stability of the data in the model. It can determine whether the variance of the regression coefficients is too high or too low. When the variance is high, multicollinearity is formed, which is the high correlation between the variables. When the P value does not bring accurate results, the VIF factor aids in minimizing the instability of a model. In this comprehensive assessment, the VIF factor is used to determine a variable’s correlation with the rest of the terms.

1. **Method of Analysis**

***3.1) Summary of Theoretical Concepts pertinent to the analysis:***

1. Multiple Linear Regression

A multiple linear regression model is “a statistical process for estimating the relationship among variables”. The general equation is given by;

1. Least Squares Estimation of the parameters

Least squares method can be used in order to approximate the regression coefficients (j) in a multiple regression model. It is commonly used in order to find the best model for the given data with the least’s errors. The general equation is;

The previous functions L represents the errors produced by the data. The goal is to find a minimal L in order to obtain the most accurate model possible. Matrices can be used when dealing with a large amount of data. The formula for least squares can then be written as;

1. Variance

It is important to estimate the variance when dealing with a large number of data. The variance indicates how far off data are from the average. In addition, the variance can be used to determine if a variable is eligible for the linear regression model( a certain limit/barrier). The following formula can be used to calculate the variance;

1. Hypothesis Test in Multiple Regression

Hypotheses test are conducted on different multiple regression models in order to determine its adequacy. This test is called “significance of regression”, which identifies if any linear relationships exist between the output y and the variables. The null hypothesis is;

The null hypothesis determines the level of contribution of the variables to the model, where for at least one j, the coefficient. The sum of squares denotes a “measure of variation or deviation from the mean” (average).  The summation of the squares of the differences from the mean is used to calculate the “sum of squares. The total sum of squares is given by;

1. P-Value and VIF

The P value approach determines the probability a model is in favor of the null hypothesis Ho or the alternative hypothesis H1 .When testing a house model, a limit to the P-value will be stated in order to remain with the best terms needed for the model. Another factor that determines a variable’s strength in the model is the Variance Inflation factor or VIF. The VIF determines the stability of the data in the model. It can determine whether the variance of the regression coefficients is too high or too low.

***3.2) Data Analysis:***

The housing data that will be analyzed contain factors that will be represented as a function for the multiple regression model. The factors will be treated as variables and the coefficient of each one will be determined. The resulting model will provide the linear regression function of the model, which is the predicting house price of the data. The factors are the following:

1. **Linear feet of street connected to property (Lot frontage):** the area of the lot in front of the house
2. **Lot size in square feet (Lot area):** the total area of the property in square feet
3. **Street conditions:** if the street is paved (=1) or not (=0)
4. **Age of construction:** the age of the building since it was constructed
5. **Area of finished basement in square feet:** the finished basement’s total area. If the area is 0, it means the basement is not renovated.
6. **Area of unfinished basement in square feet:** the unfinished basement’s total area. If the area is 0, it means the basement is renovated.
7. **General living area in square feet:** the total area of the building’s floor in all levels, including basement.
8. **Number of bathrooms:** the total number of bathrooms the building has
9. **Number of bedrooms which are above ground:** number of total rooms excluding rooms in basement
10. **Number of Kitchens:** total number of kitchens in the lot
11. **Number of rooms above ground:** the total number of rooms above ground, excluding basement
12. **Number of fireplaces:** the number of total fireplaces in the lot.
13. **Number of garages:** total number of garages
14. **Swimming pool:** The building’s status on having a pool (=1) or not (=0)
15. **Size of deck in square feet:** the size of the building’s deck in square feet

* For a greater visualization of the variables, see appendix that includes selling price of each house as well.

***3.3) Computer Algebra Methods Used:***

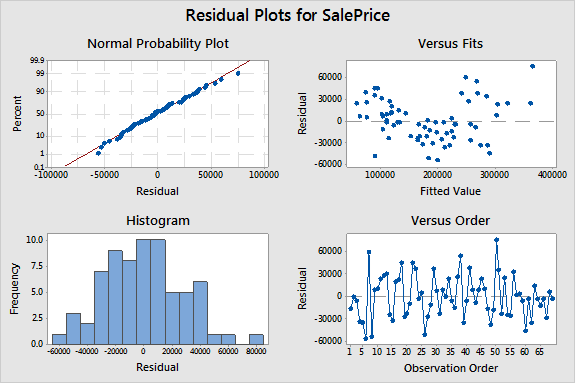
When doing multiple regression models, it is important to utilize a software in order to save time and get an accurate representation of the data analyzed, both tabulated and graphed. The software used in this data analysis was Minitab 2017. It is an easy to use program and similar functions with Microsoft Excel. Minitab was used for the majority of the data analysis, since it helped simplify calculation work and provide the most accurate results with the least errors. It also provides a visual form of the data. For example, the least square regression formula is plotted in a graph, in order for the user to check visually how close the data are to the regression line produced.

When the data was computed in the program and the multiple linear regression model was formed, the history of each procedure was shown by Minitab in the worksheet provided. The software was able to produce the mean, the variance standard deviation. In the analysis of variance table, the software depicts the results of each variable’s adjusted standard deviation, as well as the P-value. In the model summary section, the adjusted R-squared and R-squared were shown. Note: the adjusted R-squared was used in this comprehensive assessment. In the next section, the coefficients were plotted as well as the T-Value, P-Value and VIF. These are essential for the data analysis, since the statistician’s hypothesis testing depends on them. At the end, the regression equation is shown, where in this case, it is the predicted house’s sale price. Below that, the unusual observations (outliers) were shown, which notify the user to delete them for a more accurate model. In general, there were no calculations done since the software automatically processed the data and provided just the results. Therefore, no theoretical equations were used for the data at hand and the program used them instead.

1. **Results**

*Initial Data before any analysis:*

Plot set #1: Regression Analysis of Sale price initially



* This set was plotted after the removal of two unneeded variables and before the removal of outliers.

In this set of plots, all variables were included in the input of the analysis. It is the initial state of the regression model, therefore it is expected to include unusual observations (outliers), meaning that some variables do not fit the data. Before this state was computed, the variables “pool” and “pavement” were removed due to the creation of more than one regression equations, since these two variables create multiple cases, meaning more equations. When these two columns variables were deleted, the first state of linear regression model was computed.

**Model Summary**

 S    R-sq      **R-sq(adj)**    R-sq(pred)

31307.8   87.92% **85.07%**      73.94%

***Coefficients***

Term                        Coef  SE Coef  T-Value  **P-Value**   **VIF**

Constant                   38268   36170    1.06    0.295

LotFrontage                  -99   248 -0.40   0.692      1.76

LotArea                     2.14   1.00 2.13    0.037      2.39

Age                         -603   235 -2.56   0.013      3.53

Basement Finished Sq. Ft    29.1   14.4 2.02    0.048      3.52

BasementUnfinishedSF        43.0   13.8 3.12    0.003      2.93

GrLivArea                   38.3   19.9 1.92    0.060      6.33

Bathrooms                  14393  11460 1.26    0.214      5.00

BedroomAbvGr              -14818   7904 -1.87    0.066      2.66

Kitchens                  -24209  14598 -1.66    0.103      1.97

TotRmsAbvGrd                9325   6867 1.36    0.180      6.88

Fireplaces                 10760   8230 1.31    0.197      1.98

GarageCars                 16289   7712 2.11    0.039      2.42

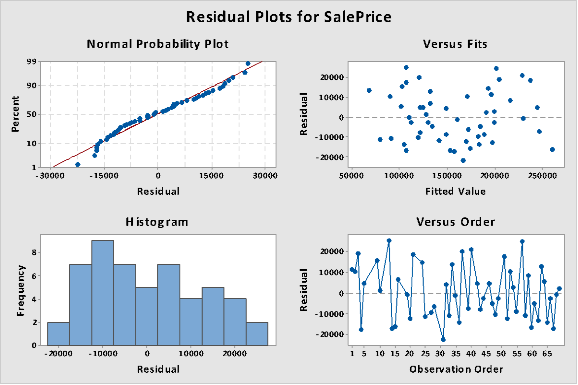
Deck                         3.7   29.5 0.13    0.901      1.24

**Regression Equation**

**SalePrice** = 38268 - 99 LotFrontage + 2.14 LotArea - 603 Age + 29.1 Basement Finished   Sq. Ft + 43.0 BasementUnfinishedSF + 38.3 GrLivArea + 14393 Bathrooms - 14818 BedroomAbvGr - 24209 Kitchens + 9325 TotRmsAbvGrd + 10760 Fireplaces + 16289 GarageCars + 3.7 Deck

In the first state of the linear regression model, meaning before any outliers were removed (See Appendix), the R-squared adjusted was obtained with a value of 85.07%. In this part, the columns that contain the variables of “pool” and “street condition” were removed before running the linear regression analysis.

Plot Set #2: Outliers were removed (1st step of Cleaning the pool of Data)



* This set of plots contains all the variables after all outliers of any analysis steps were removed.

In this set of plots, all outliers that were produced after each analysis step, were removed in order to produce a cleaner and more accurate model.

**Model Summary**

  S    R-sq   **R-sq(adj)**  R-sq(pred)

14603.0   92.89%     **91.36%**   87.60%

Coefficients

Term                     Coef      SE Coef  T-Value **P-Value VIF**

Constant                91364     17648     5.18 0.000

LotArea                 1.866     0.786     2.38 0.022  1.59

Age                    -561.6 88.3    -6.36   0.000  1.50

Basement Finished Sq. Ft   24.18      7.74    3.12 0.003  2.99

BasementUnfinishedSF    30.28      6.71     4.51 0.000  2.28

GrLivArea               53.30 7.26     7.34 0.000  2.13

BedroomAbvGr            -6738 3704    -1.82 0.076  1.43

Kitchens               -24450 8199    -2.98 0.005  1.42

Fireplaces               5170      4417     1.17 0.248  1.92

GarageCars              10676 3893     2.74 0.009  1.64

**Regression Equation**

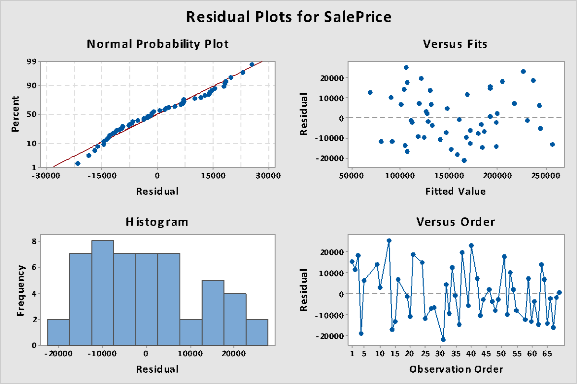
**Sale Price** = 91364 + 1.866 LotArea - 561.6 Age + 24.18 Basement Finished Sq. Ft

           + 30.28 BasementUnfinishedSF + 53.30 GrLivArea - 6738 BedroomAbvGr

        - 24450 Kitchens + 5170 Fireplaces + 10676 GarageCars

In this set of plots, all outliers that were produced after each analysis step, were removed in order to produce a cleaner and more accurate model. The R-squared obtained was 91.36%.

Plot Set #3: P-Value and VIF analyzed( final step of cleaning data)



**Model Summary**

  S    R-sq   **R-sq(adj)**  R-sq(pred)

13941.8   93.42%     91.98%   88.50%

Coefficients

Term                     Coef  SE Coef  T-Value  **P-Value VIF**

Constant                94326  16900     5.58   0.000

LotArea                 1.858  0.750     2.48   0.017  1.59

Age                      -553.2   84.4     -6.55    0.000  1.46

Basement Finished Sq. Ft   19.87  7.64 2.60    0.013  3.09

BasementUnfinishedSF    28.74  6.44      4.46  0.000  2.30

GrLivArea               53.37  6.93      7.70  0.000  2.13

BedroomAbvGr              -6641   3536     -1.88    0.068  1.43

Kitchens               -25742 7849     -3.28   0.002  1.43

Fireplaces               7581  4351      1.74    0.089  2.02

GarageCars               9916  3732      2.66    0.011  1.65

**Regression Equation**

**SalePrice** = 94326 + 1.858 LotArea - 553.2 Age + 19.87 Basement Finished Sq. Ft + 28.74 BasementUnfinishedSF + 53.37 GrLivArea - 6641 BedroomAbvGr - 25742 Kitchens + 7581 Fireplaces + 9916 GarageCars

* Final regression equation of predicting a house’s sale price.

In this final phase of the regression analysis, the variables that had a P-value bigger than 0.1 and a VIF of 5.0 were automatically removed from the data analysis test. The final regression equation listed above is the regression equation for predicting a house’s sale price.

1. **Discussion:**

In the beginning of the model, two columns with the variables of “pool” and “pavement” were removed due to the creation of more than one regression equation, and in order to simplify results, the two columns were not part of the list. When the model was being computed, the resulting R-squared was 85.07%. This is the percentage of the data that falls on the generated regression line of the model.

Since unusual observations (outliers) occurred, they were removed so the data produce greater results. The resulting R-squared was 91.36%. It increased since the beginning by 6.29%, indicating how closer the variables are now to the regression line. This produces new and more accurate coefficients for each variable. A new regression equation was displayed.

Finally, the last step was to remove any variables that did not meet the specified intervals for the P-value and the VIF. The P-value determines the probability the model is in favor of the null hypothesis or the alternative hypothesis. The VIF determines the stability of the variable in the model. The P-value limit was set at 0.1 and if one of the variables was close to another variable, the VIF factor was seen instead with a limit of 5.0. It is also noted that the constant of the regression equation (beta0) has no VIF factor, since it is not a variable. The resulting R-squared was 91.98%. The change is minor since the last model, but the coefficients are more accurate than before.

In general, the model resulted in great accuracy, with a R-squared value close to 100%. This means that the errors in the model are minor and the percentage is ideal for predicting a house’s sale price. However, some suggestions could be made for further improvements. The limit for the P-value and the VIF could be changed, possibly providing a greater R-squared. Also, more useful variables such as number of floors could be added so a more precise regression equation could be generated.

1. **Conclusion**

In conclusion, the purpose of this comprehensive assessment was to obtain the regression equation for predicting the sale price of a house. The model used for the data analysis was the multiple linear regression The initial and final R-squared obtained were 85.07% and 91.98% respectively. This was done by removing all the outliers from the variable list and setting limits for P-value and VIF. The limits were 0.1 and 5.0 respectively. This procedure helped obtain a more precise model and a greater R-squared value, meaning the gap between the variables and the regression line is minimized. In the end, the final regression equation was generated with its according coefficients. The equation is **SalePrice** = 94326 + 1.858 LotArea - 553.2 Age + 19.87 Basement Finished Sq. Ft + 28.74 BasementUnfinishedSF + 53.37 GrLivArea - 6641 BedroomAbvGr - 25742 Kitchens + 7581 Fireplaces + 9916 GarageCars. The model analysis turns out to be successful, since the adjusted R-squared was close to 100%, meaning the higher it is, the less the errors created.

1. **References**

Montgomery DC, Runger GC. Applied statistics and probability for engineers. Hoboken, NJ: Wiley; 2014 [accessed 2017 May 3]..

Factors Affecting University Entrants' Performance in High-Stakes Tests: A Multiple Regression Analysis [accessed 2017 May 3].

Brownie S. What factors determine house prices? National Housing Federation. 2015 Oct 26 [accessed 2017 May 3].

12.4 - Detecting Multicollinearity Using Variance Inflation Factors. PennState. [accessed 2017 May 16].

What is a variance inflation factor (VIF)? Minitab. [accessed 2017 May 16]. http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/model-assumptions/what-is-a-variance-inflation-factor-vif/