Valparaiso University
ValpoScholar

Business Faculty Publications | College of Business

10-2016

Multiple Linear Regression Applications in Real Estate Pricing

Ceyhun Ozgur
Valparaiso University

Zachariah Hughes
Valparaiso University

Grace Rogers
Valparaiso University

Sufia Parveen
Valparaiso University

Follow this and additional works at: https://scholar.valpo.edu/cba_fac_pub

Recommended Citation
Ozgur, Ceyhun; Hughes, Zachariah; Rogers, Grace; and Parveen, Sufia, "Multiple Linear Regression Applications in Real Estate Pricing" (2016). *Business Faculty Publications*. 61.
https://scholar.valpo.edu/cba_fac_pub/61

This Article is brought to you for free and open access by the College of Business at ValpoScholar. It has been accepted for inclusion in Business Faculty Publications by an authorized administrator of ValpoScholar. For more information, please contact a ValpoScholar staff member at scholar@valpo.edu.

International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 4 Issue 8 || October. 2016 || PP-39-50

Multiple Linear Regression Applications in Real Estate Pricing

Ceyhun Ozgur, Ph.D., CPIM[1], Zachariah Hughes[2], Grace Rogers[3], Sufia Parveen[4]

[1]*Professor, Valparaiso University College of Business Information & Decision Sciences Urschel Hall 223 – Valparaiso University Valparaiso, IN 46383*
[2]*Undergraduate Research Assistant Valparaiso University Finance – College of Business Economics – College of Arts and Sciences*
[3]*Undergraduate Research Assistant Valparaiso University Actuarial Science - College of Arts and Sciences Business Analytics - College of Business*
[4]*Graduate Research Assistant Valparaiso University M.S. in Analytics and Modeling*

ABSTRACT: *In this paper, we attempt to predict the price of a real estate individual homes sold in North West Indiana based on the individual homes sold in 2014. The data/information is collected from realtor.com. The purpose of this paper is to predict the price of individual homes sold based on multiple regression model and also utilize SAS forecasting model and software. We also determine the factors influencing housing prices and to what extent they affect the price. Independent variables such square footage, number of bathrooms, and whether there is a finished basement,. and whether there is brick front or not and the type of home: Colonial, Cotemporary or Tudor. How much does each type of home (Colonial, Contemporary, Tudor) add to the price of the real estate.*

I. INTRODUCTION

This paper is about 480 houses selected as a representative sample of all real estate sold in the State Indiana in 2014. The information used in this paper was taken from realtor.com. The purpose of this paper is to develop a relatively good regression equation for predicting the price of these houses. It is known that there are many factors to influence housing price, but we do not know for certain what factors will influence the price of the houses and to what extent these factors will impact the price. As a buyer, one can judge whether the price of a house of interest is rational or not. As a seller, one would be able to select a rational price according to the equation helping to optimize potential sales and correctly forecast demand.

According to one of the similar papers by Anpalaki J. Ragavan(2008), independent variables were categorized as two types: a) continuous independent variables (e.g: number of bedrooms, number of bathrooms, size of the property), and b) indicator independent variables that provide supporting information about the unit in the form of an item or facility that is either present (1) or not present (0) in the unit (e.g: built in dishwasher (DW), refrigerator (fridge), laundry facilities (WD)). There were 6 continuous independent variables namely, i) number of bedrooms (BED), ii) number of bathrooms (BATH), iii) square footage or size (SIZE), iv) sale price (PRICE), v) age of the property (AGE), and v) size of the yard (LOT).

According to Leslie A. Christensen, a linear model has the form Y = b0 + b1X + ε. The constant b0 is called the intercept and the coefficient b1 is the parameter estimate for the variable X. The ε is the error term. ε is the residual that cannot be explained by the variables in the model. Most of the assumptions and diagnostics of linear regression focus on the assumptions of ε. The following assumptions must hold when building a linear regression model.

1. The dependent variable must be continuous. If we are trying to predict a categorical variable, linear regression is not the correct method. We can investigate discrim, logistic, or some other categorical procedure.
2. The data we are modeling meets the "iid" criterion. That means the error terms, ε, are: a) independent from one another and b) identically distributed. If assumption 2a does not hold, we need to investigate time series or some other type of method. If assumption 2b does not hold, we need to investigate methods that do not assume normality such as non-parametric procedures.
3. The error term is normally distributed with a mean of zero and a standard deviation of σ 2, N(0,σ 2). Although not an actual assumption of linear regression, it is good practice to ensure the data we are modeling came from a random sample or some other sampling frame that will be valid for the conclusions we wish to make based on our model.

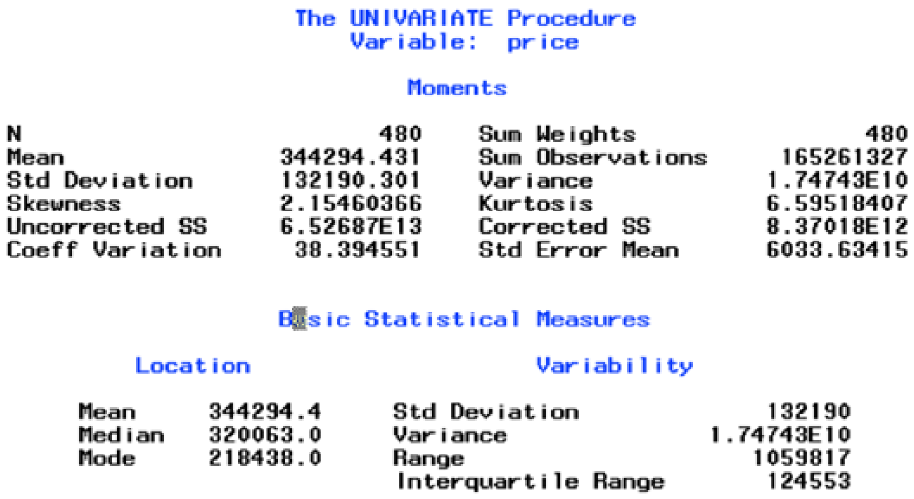
Following the above procedures in this paper, the response variable selected is the “price” of these houses and the price is a numeric variable. At the same time, there are 12 variables used as the potential predictors and they are: Size (in square feet), Region (Urban, Suburban or Rural) Type (Condominium, Townhouse or SFH), Yard (in square feet), Bedrooms (number of), Bathrooms (number of), Garage (Attached, Detached or No), Floors, (number of), HOA (Homeowner Association in dollars), Tax (in percentage), Basement (Yes or No), Age (in years). The variables region, type, garage and basement are specifically numeric variables and are measured by digits, while the remainder are character variables. These character variables can be changed to numeric variables for the purpose of performing regression analysis. In this study, there are 480 houses and they are observational subjects.

To assess the validity of the regression assumptions, residual plots are used. Appropriate transformations of the dependent or independent variable can be done by making changes to the model until they are improved upon from previous problems. Estimated regression equation is then reported. Any outliers or influential values are being noted. If we find any influential values, we remove them from the dataset and a new report of estimated regression equation is made without the influential values. Finally comparison of the new estimated regression equation to the original one is done. We have used the Sherwin Rosen (1974) paper on hedonic pricing and implicit markets as the foundation of this paper regarding pricing in real estate. We also used the paper by Lipsey and Rosenbluth (1971). Both of these papers, showed that the Real Estate market is competitively priced.

II. PRELIMINARY DATA EXPLORATION

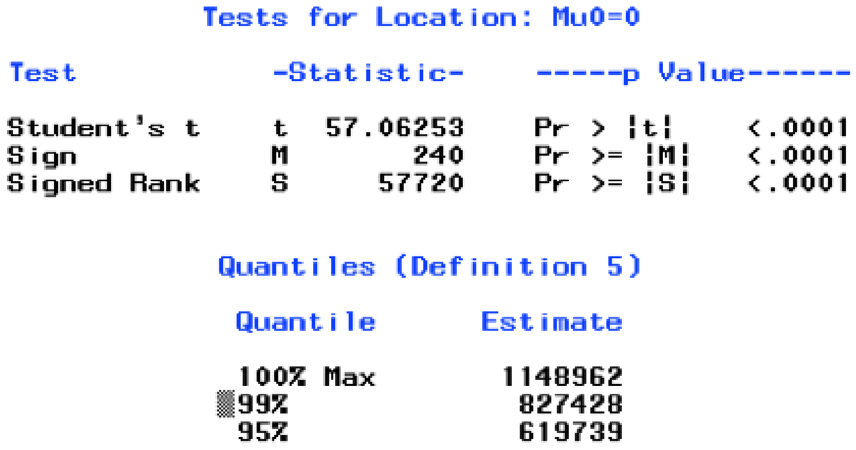
For quantitative predictors, scatterplot and correlation can be used to assess; for categorical predictors, boxplots can be used to assess. At the same time, the method of correlation can be used to find if there are some high correlated potential predictors.

2.1The Basic Data

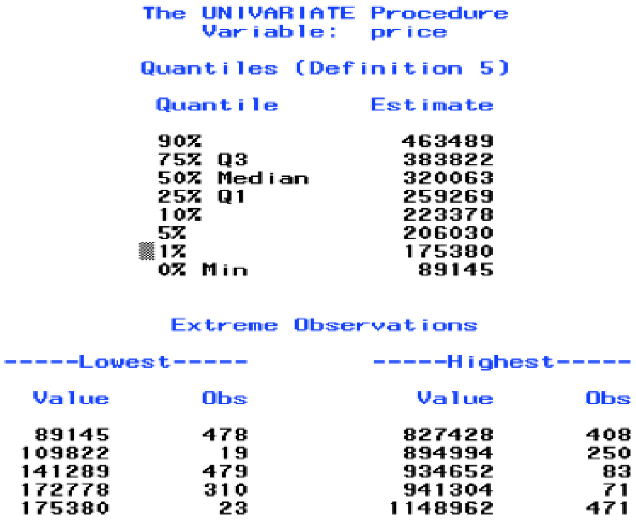


```
                     The UNIVARIATE Procedure
                        Variable:  price

                            Moments

N                          480    Sum Weights                480
Mean                344294.431    Sum Observations     165261327
Std Deviation       132190.301    Variance            1.74743E10
Skewness            2.15460366    Kurtosis            6.59518407
Uncorrected SS      6.52687E13    Corrected SS        8.37018E12
Coeff Variation      38.394551    Std Error Mean      6033.63415


                  Basic Statistical Measures

        Location                    Variability

    Mean     344294.4     Std Deviation                132190
    Median   320063.0     Variance                 1.74743E10
    Mode     218438.0     Range                       1059817
                           Interquartile Range          124553
```

Table 2: Tests for Location



```
              Tests for Location: Mu0=0

Test             -Statistic-    -----p Value------

Student's t      t  57.06253    Pr > |t|    <.0001
Sign             M       240    Pr >= |M|   <.0001
Signed Rank      S     57720    Pr >= |S|   <.0001


              Quantiles (Definition 5)

               Quantile        Estimate

               100% Max         1148962
               99%               827428
               95%               619739
```

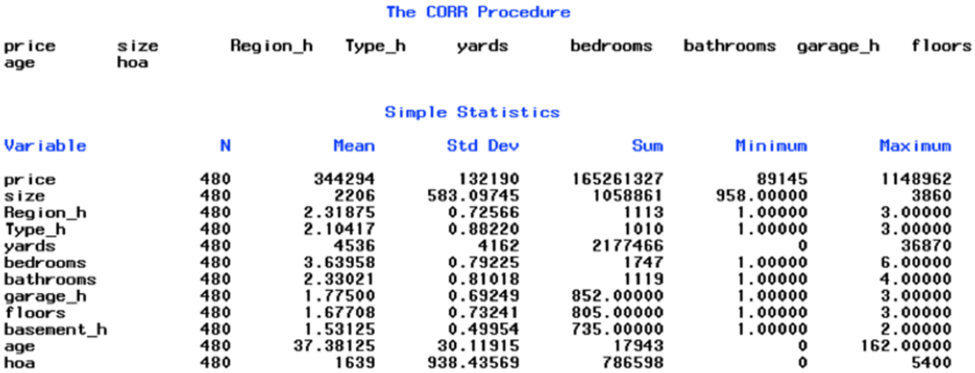


```
                 The UNIVARIATE Procedure
                    Variable:  price

                 Quantiles (Definition 5)

                  Quantile          Estimate

                  90%                463489
                  75% Q3             383822
                  50% Median         320063
                  25% Q1             259269
                  10%                223378
                  5%                 206030
                 ▒1%                 175380
                  0% Min              89145

                   Extreme Observations

        -----Lowest-----          -----Highest-----

          Value       Obs           Value       Obs

          89145       478          827428       408
         109822        19          894994       250
         141289       479          934652        83
         172778       310          941304        71
         175380        23         1148962       471
```

From this result, we can get some basic data about the response variable (the price of these houses). Specially, Mean=344294.4 Median=320063.0 Standard deviation=132190

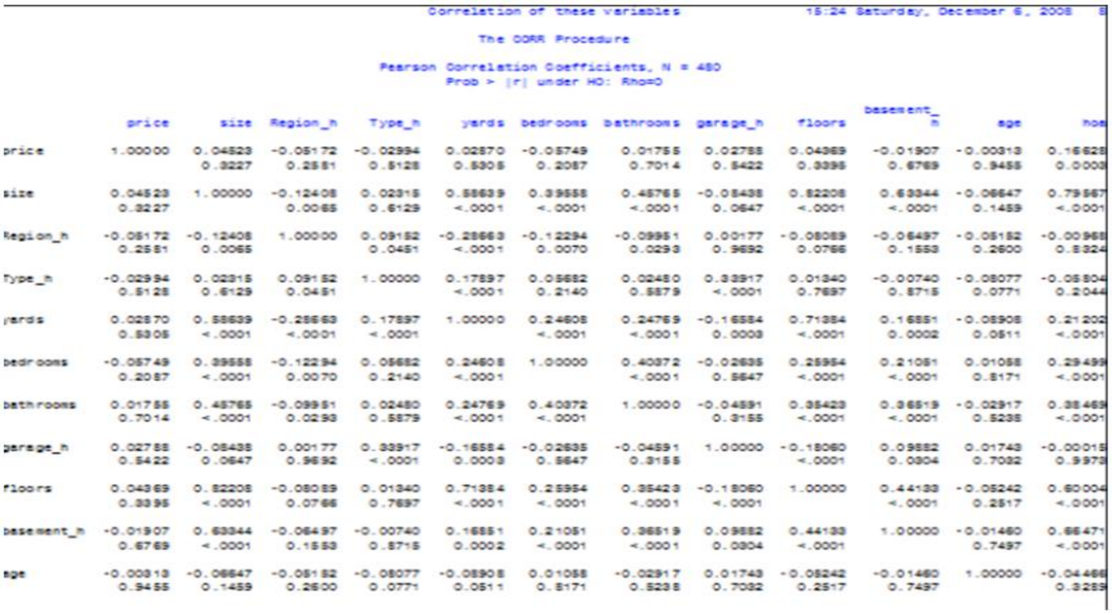
2.2: Correlation between variables

Table 3: Correlation between variables



```
                                 The CORR Procedure

price      size        Region_h   Type_h     yards      bedrooms   bathrooms  garage_h   floors
age        hoa

                                 Simple Statistics

Variable             N         Mean      Std Dev          Sum      Minimum        Maximum

price              480       344294       132190    165261327        89145        1148962
size               480         2206    583.09745      1058861    958.00000           3860
Region_h           480      2.31875      0.72566         1113      1.00000        3.00000
Type_h             480      2.10417      0.88220         1010      1.00000        3.00000
yards              480         4536         4162      2177466            0          36870
bedrooms           480      3.63958      0.79225         1747      1.00000        6.00000
bathrooms          480      2.33021      0.81018         1119      1.00000        4.00000
garage_h           480      1.77500      0.69249    852.00000      1.00000        3.00000
floors             480      1.67708      0.73241    805.00000      1.00000        3.00000
basement_h         480      1.53125      0.49954    735.00000      1.00000        2.00000
age                480     37.38125     30.11915        17943            0      162.00000
hoa                480         1639    938.43569       786598            0           5400
```

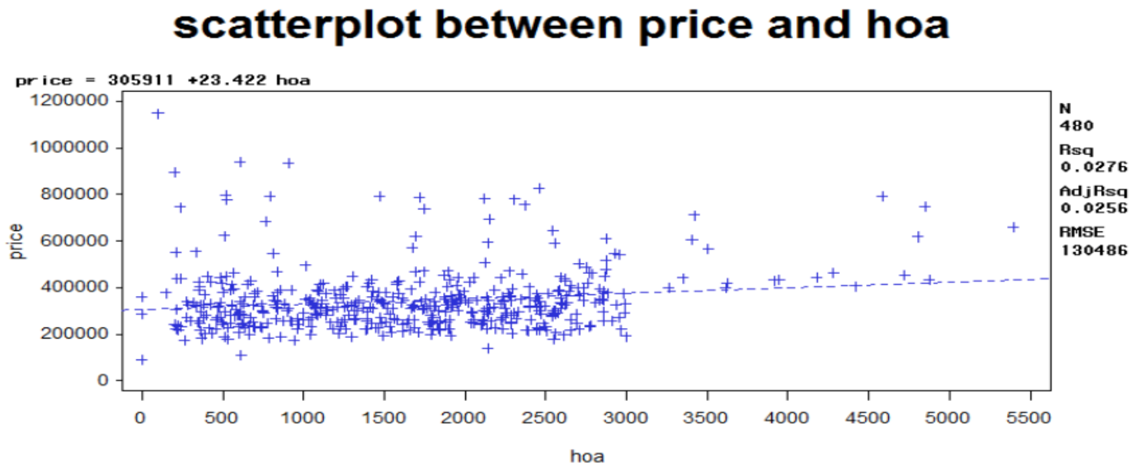
This procedure gives information about the 12 variables with their simple statistics and Pearson's correlation coefficient.



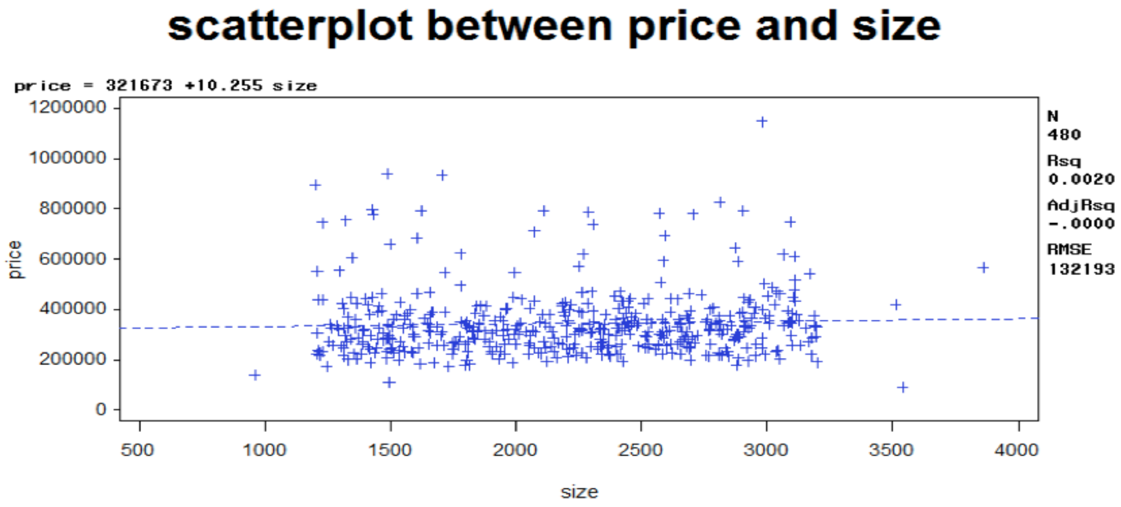
```
                                  Correlation of these variables             15:24 Saturday, December 6, 2008   8

                                      The CORR Procedure

                              Pearson Correlation Coefficients, N = 480
                                      Prob > |r| under H0: Rho=0

                                                                                                    basement_
             price      size  Region_h    Type_h     yards  bedrooms  bathrooms  garage_h    floors          h        age       hoa

price      1.00000   0.04523  -0.05172  -0.02994   0.02870  -0.05749   0.01755   0.02788   0.04369   -0.01907   -0.00313   0.16628
                      0.3227    0.2551    0.5128    0.5305    0.2087    0.7014    0.5422    0.3395     0.6769     0.9455    0.0003

size       0.04523   1.00000  -0.12408   0.02315   0.58639   0.39558   0.45765  -0.08438   0.82208    0.63344   -0.06647   0.79567
            0.3227              0.0065    0.6129    <.0001    <.0001    <.0001    0.0647    <.0001     <.0001     0.1459    <.0001

Region_h  -0.05172  -0.12408   1.00000   0.09152  -0.28663  -0.12294  -0.09951   0.00177  -0.08089   -0.06497   -0.05152  -0.00965
            0.2551    0.0065              0.0451    <.0001    0.0070    0.0293    0.9692    0.0766     0.1553     0.2600    0.8324

Type_h    -0.02994   0.02315   0.09152   1.00000   0.17897   0.05682   0.02480   0.33917   0.01340   -0.00740   -0.08077  -0.05804
            0.5128    0.6129    0.0451              <.0001    0.2140    0.5879    <.0001    0.7697     0.8715     0.0771    0.2044

yards      0.02870   0.58639  -0.28663   0.17897   1.00000   0.24608   0.24769  -0.16584   0.71384    0.16851   -0.08908   0.21202
            0.5305    <.0001    <.0001    <.0001              <.0001    <.0001    0.0003    <.0001     0.0002     0.0511    <.0001

bedrooms  -0.05749   0.39558  -0.12294   0.05682   0.24608   1.00000   0.40372  -0.02635   0.25954    0.21051    0.01058   0.29499
            0.2087    <.0001    0.0070    0.2140    <.0001              <.0001    0.5647    <.0001     <.0001     0.8171    <.0001

bathrooms  0.01755   0.45765  -0.09951   0.02480   0.24769   0.40372   1.00000  -0.04591   0.35423    0.36519   -0.02917   0.38469
            0.7014    <.0001    0.0293    0.5879    <.0001    <.0001              0.3155    <.0001     <.0001     0.5238    <.0001

garage_h   0.02788  -0.08438   0.00177   0.33917  -0.16584  -0.02635  -0.04591   1.00000  -0.18060    0.09882    0.01743  -0.00015
            0.5422    0.0647    0.9692    <.0001    0.0003    0.5647    0.3155              <.0001     0.0304     0.7032    0.9973

floors     0.04369   0.82208  -0.08089   0.01340   0.71384   0.25954   0.35423  -0.18060   1.00000    0.44133   -0.05242   0.60004
            0.3395    <.0001    0.0766    0.7697    <.0001    <.0001    <.0001    <.0001               <.0001     0.2517    <.0001

basement_h -0.01907  0.63344  -0.06497  -0.00740   0.16851   0.21051   0.36519   0.09882   0.44133    1.00000   -0.01460   0.66471
            0.6769    <.0001    0.1553    0.8715    0.0002    <.0001    <.0001    0.0304    <.0001                0.7497    <.0001

age       -0.00313  -0.06647  -0.05152  -0.08077  -0.08908   0.01058  -0.02917   0.01743  -0.05242   -0.01460    1.00000  -0.04466
            0.9455    0.1459    0.2600    0.0771    0.0511    0.8171    0.5238    0.7032    0.2517     0.7497               0.3283
```

2.3: Scattered plots

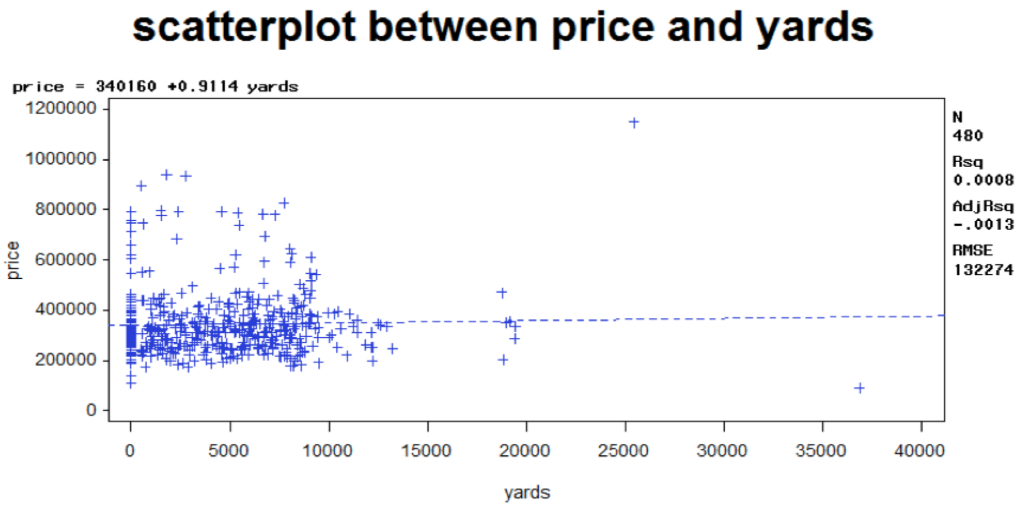
Graph 1: Scatter plot between Price and hoa



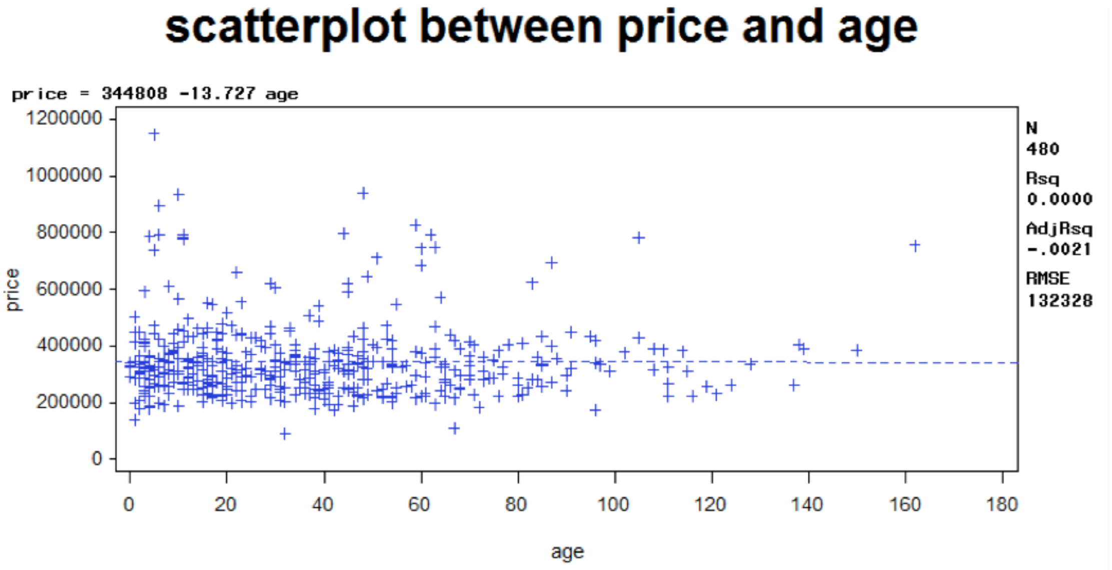
Graph 2: Scatter plot between Price and size:



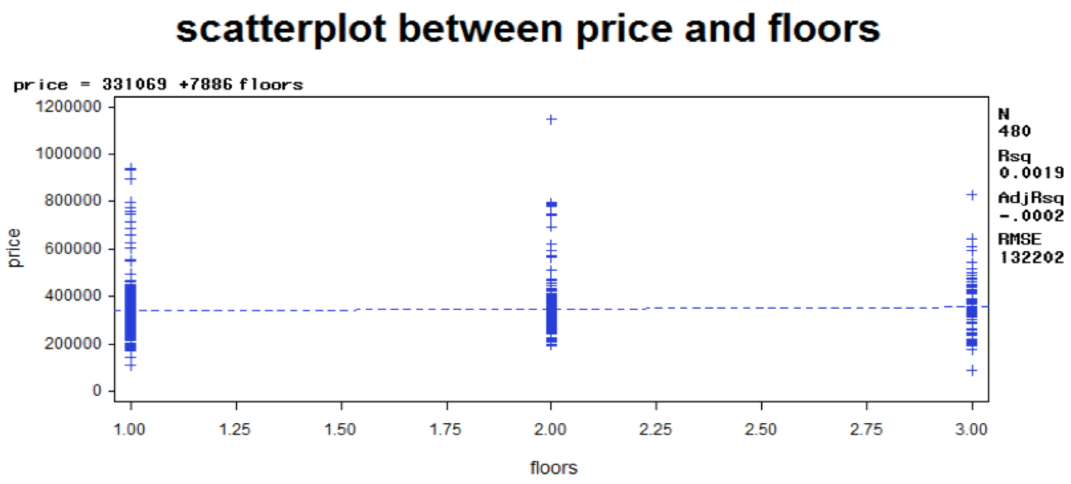
Graph 3: Scatter plot between Price and yards:



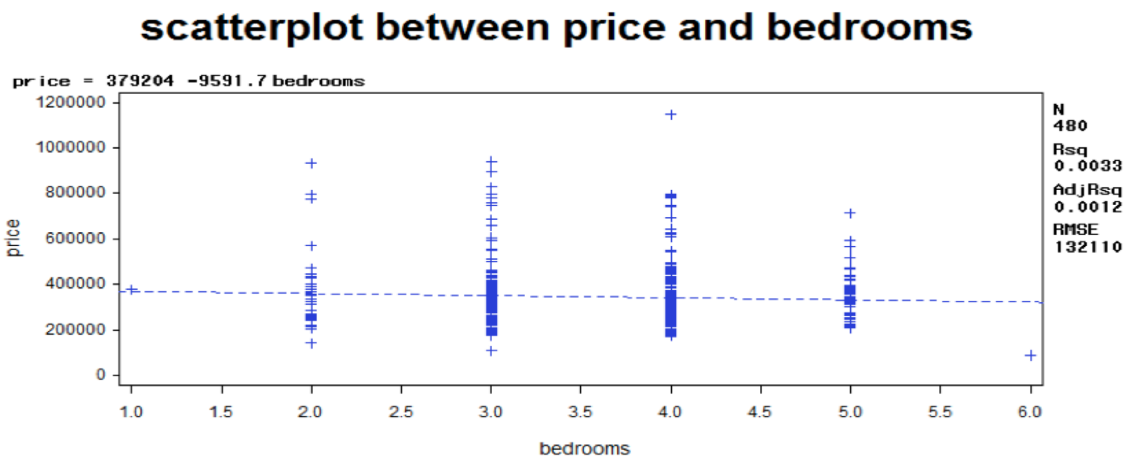
Graph 4: Scatter plot between Price and age:



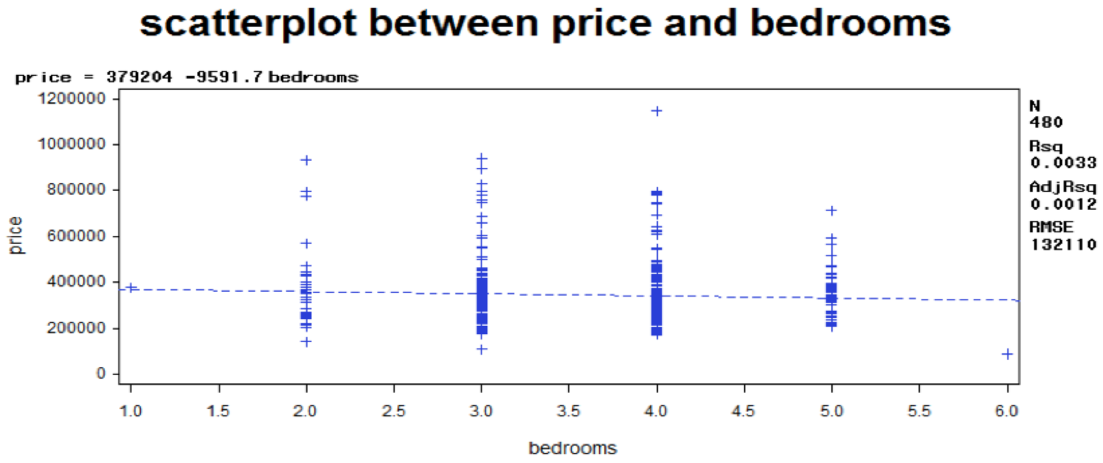
Graph 5: Scatter plot between Price and floors:



Graph 6: Scatter plot between Price and bedrooms:



Graph 7: Scatter plot between Price and bathrooms:



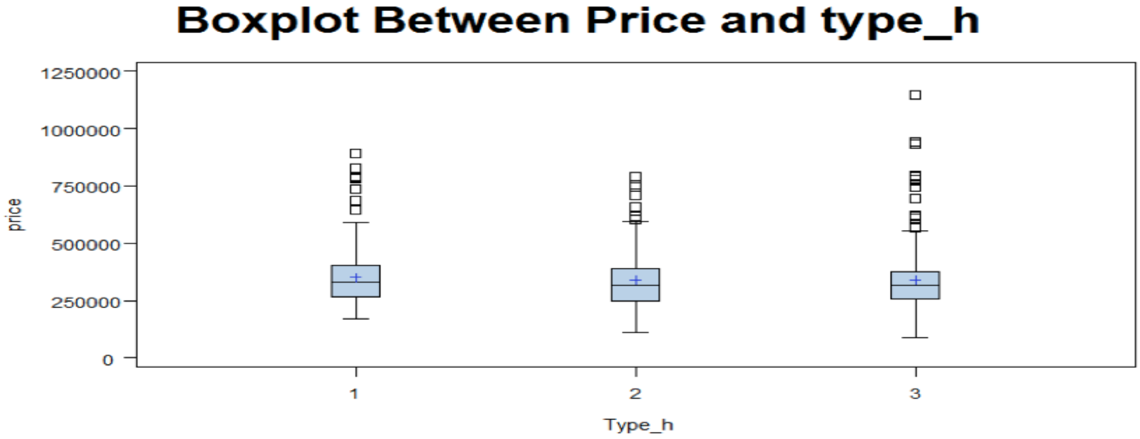
By observing the correlation between the variables and the scatterplots the following conclusion was drawn. The correlation coefficient between price and HOA is 0.16628 and the scatterplot shows they have positive relationship. From the second scatterplot, we can find that the price and size have positive relationship; the price increases as the size increases. In detail, the correlation coefficient between price and size is 0.04523. From the third scatterplot it has been observed that price and yards have positive relationship just like it has with size. The correlation coefficient between price and yards is 0.028, indicating that the correlation is significant. The correlation coefficient between price and age is -0.00313, indicating that the correlation between price and age is not significant and the scatterplot between them shows a negative relationship, that is, as the age of the house increases the price obviously decreases.

The correlation coefficient between price and floors is 0.04369 and the scatterplot shows there is a significant positive correlation between them. The correlation coefficient between price and bedrooms is -0.05749 and the scatterplot shows as the number of bedrooms increases the price gradually decreases which is a negative correlation. Finally, the correlation coefficient between price and bathrooms is 0.01755 and the scatterplot indicates there is a positive correlation meaning, as the number of bathrooms increases the price increases initially but later decreases with an increase in number of bathrooms.

There also exist relationships among the other potential predictors through the table. For example the correlation coefficient between size and yards is found to be 0.58639 and the correlation coefficient between garage_h and type_h is 0.33917, these two absolute values of correlation coefficient are both greater than 0.3 and less than 0.8, so they have moderate correlation. Other absolute values of correlation coefficients, which were found to be less than 0.3, were found to have low correlation relationships.

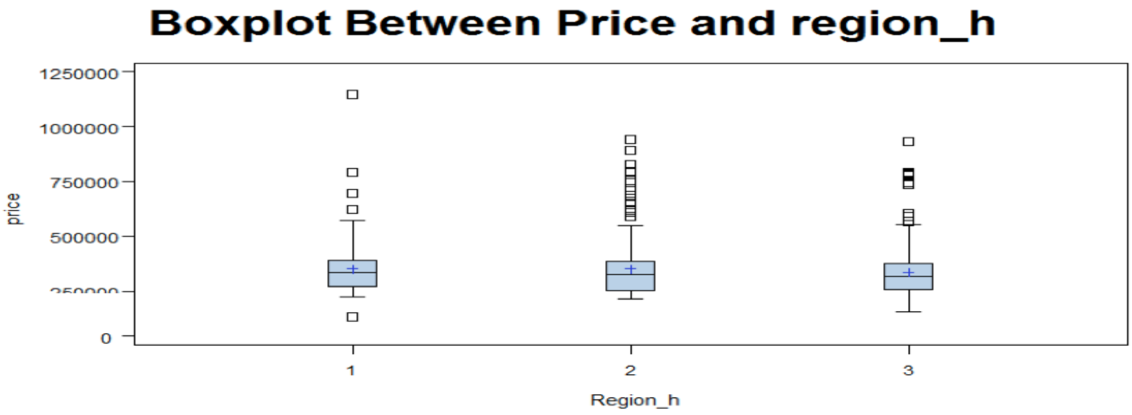
2.4: Box plots

Graph 8: Box plot between Price and type:



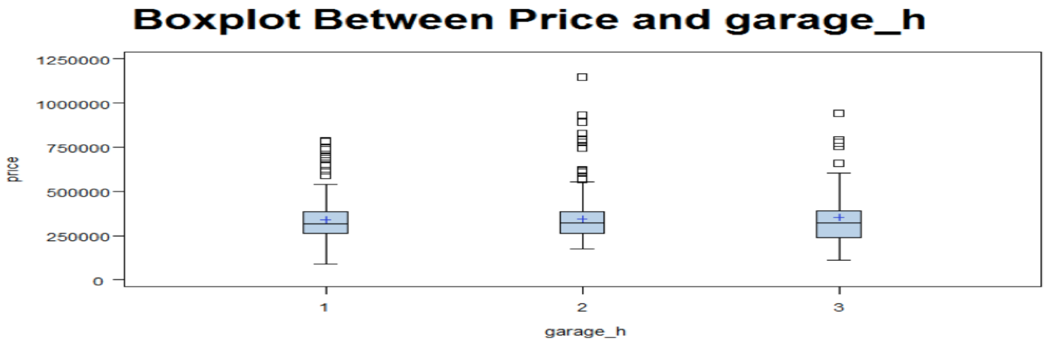
From the box-plot above, it can be observed that when price range is between 20,000 to 60,000, the type of the house available is townhouse, when it is between 15,000 to 65,000 the house type is SFH and when the price range is between 10,000 to 55,000 the house type available is condominium. All the three types of houses have some outliers.

Graph 9: Box-plot between Price and region h:



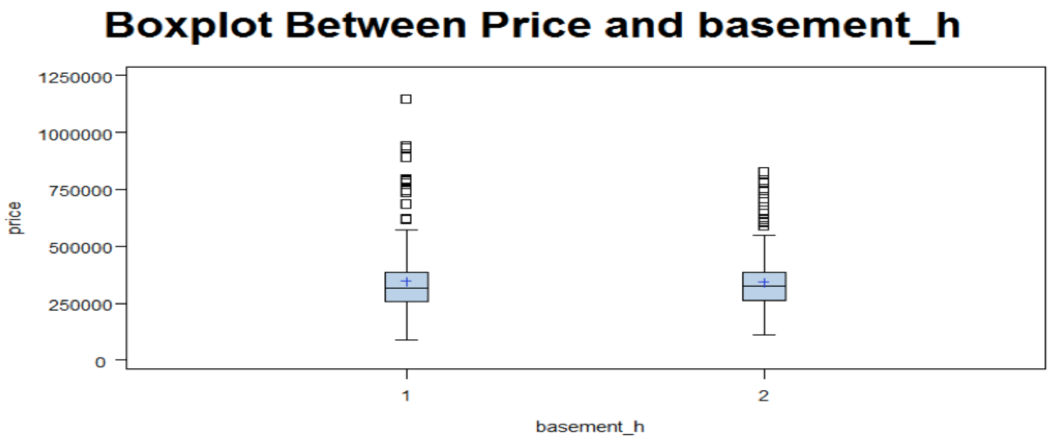
From the box-plot between price and region_h, it has been observed that when the price range is between 20,000 and 55,000 the region is rural, when the price range is between 20,000 and 53,000 the region is urban, and when the price range is between 15,000 and 55,000 the region is suburban. All three regions have some outliers.

Graph 10: Box-plot between Price and garage_h:



From the box-plot between price and garage_h it has been observed that when the price range is between 10,000 and 55,000 the garage is detached, when the price range is between 20,000 and 57,000 the garage is attached, and when the price range is between 11,000 and 60,000 there is no garage. All three garage types have outliers.

Graph 11: Box-plot between Price and basement_h:



From the box-plot between price and basement_h it has been observed that when the price range is between 10,000 and 60,0000 there is a basement and when the price range is between 12,000 and 58,000 there is not a basement. Both of the basement types have some outliers.

As we can see from the figures, the various categorical predictors all have the outliers. Now as we did some primary analysis for different predictors, we cannot give up any predictors from the information above.

III. REGRESSION

3.1. Based on all the information and outputs above, our initial model would be:

Price= $\square_0$size+ $\square_1$region_h+ $\square_2$type_h+ $\square_3$yards+ $\square_4$age+ $\square_5$floors+ $\square_6$hoa+ $\square_7$bedrooms+$\square_8$bathrooms+ $\square_9$garage_h+ $\square_{10}$basement_h+ $\square$.

From the scattered plots, it is clearly seen that the linear relation is negative between price and age. The normal thinking is that the price will go down with the age goes up. That is to say, the coefficient between price and age is negative. Although the dots are dispersed and the linear relation is not obvious, we still put that in into the initial model. Secondly, the relationship between price and size is positive, as observed in the scatterplot, so we include this in the initial model. We also include all other numeric predictors into the initial model. As for the four categorical predictors, we use the side-by-side box plots to see the distribution. By observing the boxplots, some outliers were existent. But we still include them in the initial model. Initially, we include all potential predictors in the model, assuming that they may have an influence on the price.(Articles &Statistical Papers, December 1992)

3.2. Then let us begin to do the multiple regression analysis.

Graph 12: Scatter plot between predicted value and student zed residual

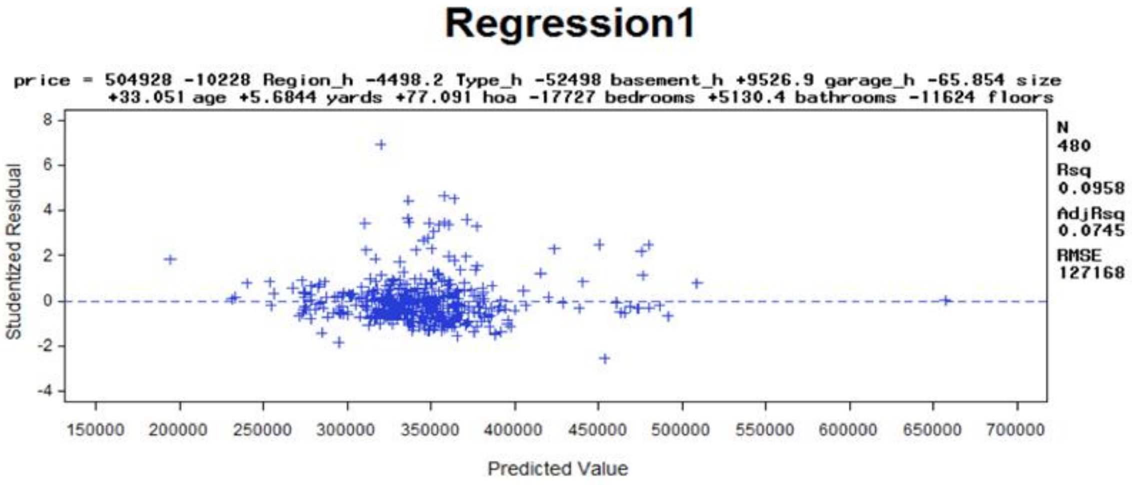
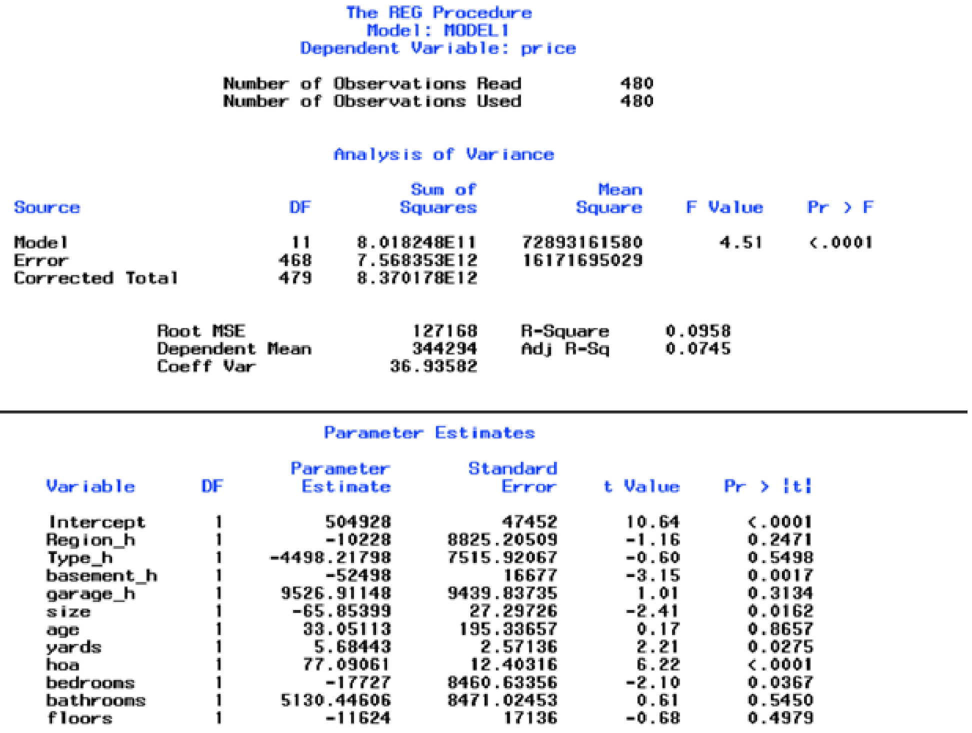


Table 4: Regression Analysis of Real Estate predictor variables



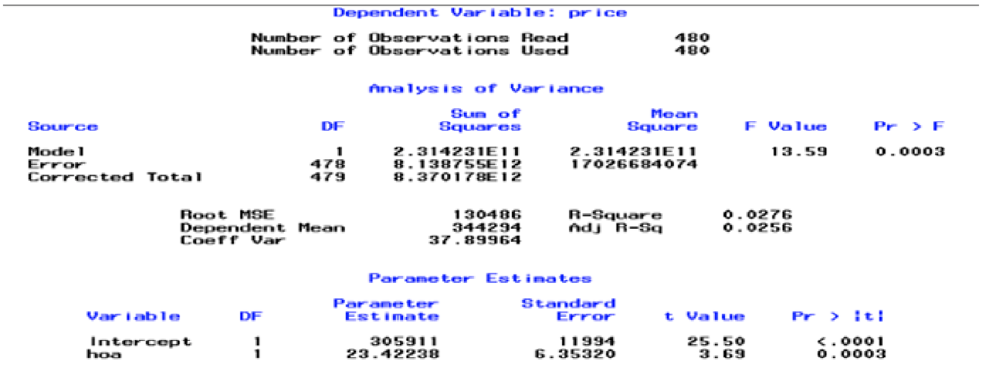
```
                           The REG Procedure
                             Model: MODEL1
                       Dependent Variable: price

              Number of Observations Read          480
              Number of Observations Used          480


                         Analysis of Variance

                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F

Model                    11    8.018248E11    72893161580      4.51    <.0001
Error                   468    7.568353E12    16171695029
Corrected Total         479    8.370178E12


           Root MSE              127168    R-Square     0.0958
           Dependent Mean        344294    Adj R-Sq     0.0745
           Coeff Var           36.93582

                          Parameter Estimates

                        Parameter       Standard
  Variable      DF       Estimate          Error    t Value    Pr > |t|

  Intercept      1         504928          47452      10.64      <.0001
  Region_h       1         -10228     8825.20509      -1.16      0.2471
  Type_h         1    -4498.21798     7515.92067      -0.60      0.5498
  basement_h     1         -52498          16677      -3.15      0.0017
  garage_h       1     9526.91148     9439.83735       1.01      0.3134
  size           1      -65.85399       27.29726      -2.41      0.0162
  age            1       33.05113      195.33657       0.17      0.8657
  yards          1        5.68443        2.57136       2.21      0.0275
  hoa            1       77.09061       12.40316       6.22      <.0001
  bedrooms       1         -17727     8460.63356      -2.10      0.0367
  bathrooms      1     5130.44606     8471.02453       0.61      0.5450
  floors         1         -11624          17136      -0.68      0.4979
```

3.3.*The fitted quadratic model:*

Price=504928-10228Region_h-4498.217Type_h-52498Basement_h+9526.911 Garage_h-65.854Size+33.051Age+5.6844Yards+77.091Hoa-17727Bedrooms +5130.446Bathrooms-11624Floors

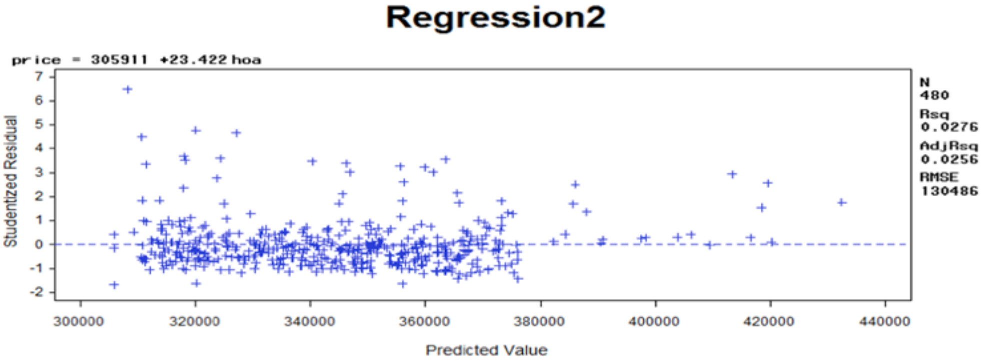
The adjusted R^2 is 0.0745, which indicates this model fits the data well. The F-test (ANOVA) tests whether $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$. The p-Value is < 0.0001, indicating that overall the explanatory variables are significant to the response variable.

Parameter estimates and standard errors of the estimates, along with t-tests and p-values are also given. The t-test tests whether each explanatory variable is zero. Hence, only HOA is significant because the p-value < 0.0001. We can then put the HOA into this model.



```
                         Dependent Variable: price

              Number of Observations Read          480
              Number of Observations Used          480

                            Analysis of Variance

                                       Sum of           Mean
Source                    DF          Squares         Square    F Value    Pr > F

Model                      1      2.314231E11    2.314231E11      13.59    0.0003
Error                    478      8.138755E12    17026684074
Corrected Total          479      8.370178E12

           Root MSE                   130486    R-Square     0.0276
           Dependent Mean             344294    Adj R-Sq     0.0256
           Coeff Var                37.89964

                            Parameter Estimates

                           Parameter       Standard
   Variable       DF        Estimate          Error    t Value    Pr > |t|

   Intercept       1          305911          11994      25.50      <.0001
   hoa             1        23.42238        6.35320       3.69      0.0003
```

Table 5: Analysis of Variance for dependent variable price



Graph 13: Scatter plot between predicted value and student zed residual

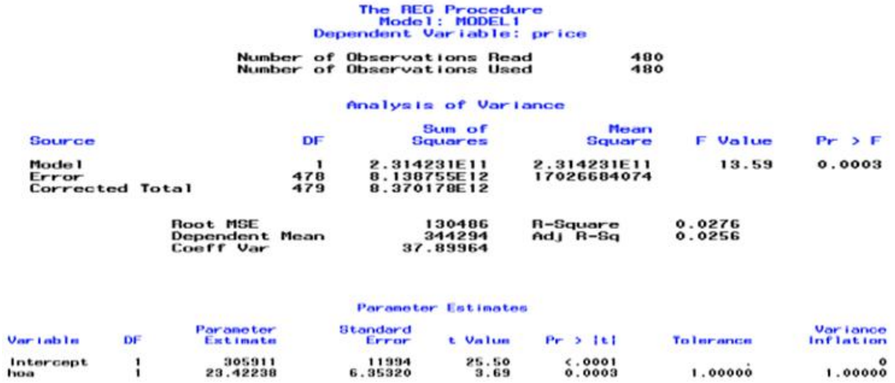
From the result, the p-value=0.0003, which means that HOA variables are significant in relation to the response variable. At the same time, the adjusted R-Sq=0.0256, which means that it explains 2.56% of the change in price. Finally, we get the following equation:

Price=305911+23.422hoa.

However, the above residual plot shows that the assumption of constant variance is not met. So we need to check for multicollinearity as well as influential observations.

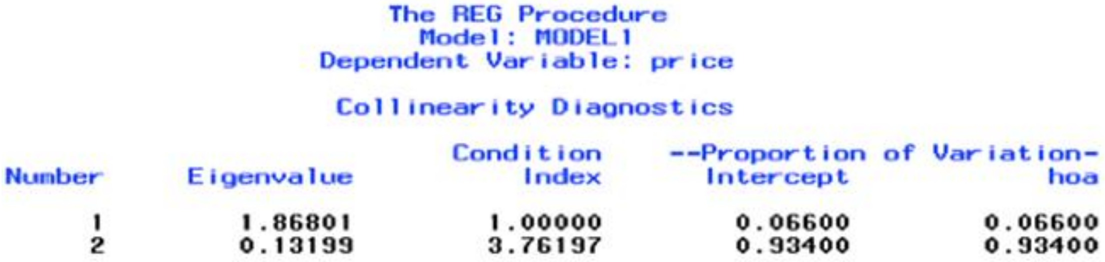
3.4: Multicollinearity

Table 6: Multicollinearity



```
                            The REG Procedure
                             Model: MODEL1
                        Dependent Variable: price

              Number of Observations Read          480
              Number of Observations Used          480

                            Analysis of Variance

                                       Sum of           Mean
Source                    DF          Squares         Square    F Value    Pr > F

Model                      1      2.314231E11    2.314231E11      13.59    0.0003
Error                    478      8.138755E12    17026684074
Corrected Total          479      8.370178E12

           Root MSE                   130486    R-Square     0.0276
           Dependent Mean             344294    Adj R-Sq     0.0256
           Coeff Var                37.89964

                                   Parameter Estimates

                      Parameter     Standard                                          Variance
Variable      DF       Estimate        Error   t Value   Pr > |t|      Tolerance     Inflation

Intercept      1         305911        11994     25.50     <.0001              .             0
hoa            1       23.42238      6.35320      3.69     0.0003        1.00000       1.00000
```

Table 7: Regression Procedure collinearity diagnostics



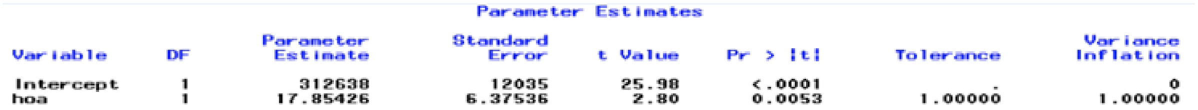
```
                         The REG Procedure
                          Model: MODEL1
                    Dependent Variable: price

                      Collinearity Diagnostics

                                  Condition    --Proportion of Variation-
Number        Eigenvalue              Index      Intercept              hoa

     1           1.86801            1.00000        0.06600          0.06600
     2           0.13199            3.76197        0.93400          0.93400
```

Tolerance is the proportion of each variable's variance not shared with the other explanatory variables. Small tolerance values indicate collinearity. In general, we should ensure the tolerance is greater than 0.2. If we notice the tolerance value of HOA its 1.000 which is greater than 0.2, this variable has no tolerance problem.

Then we need to find the outliers. Studentized residuals greater than 1.5 or less than -1.5 should be deleted because they are outliers. From these 480 data, #13 #24, #62, #63, #351, #371 should be deleted. In the second round, we need to delete #87, #96, #106, #108, #109, #118, #398. In the third round, we delete #135 #153, #157, #170, #211, #218, #224, #436. In the fourth round we delete #236, #242, #244, #268, #270, #284, #293, #313, #437, #217 and finally in total, we delete 26 data of 480.

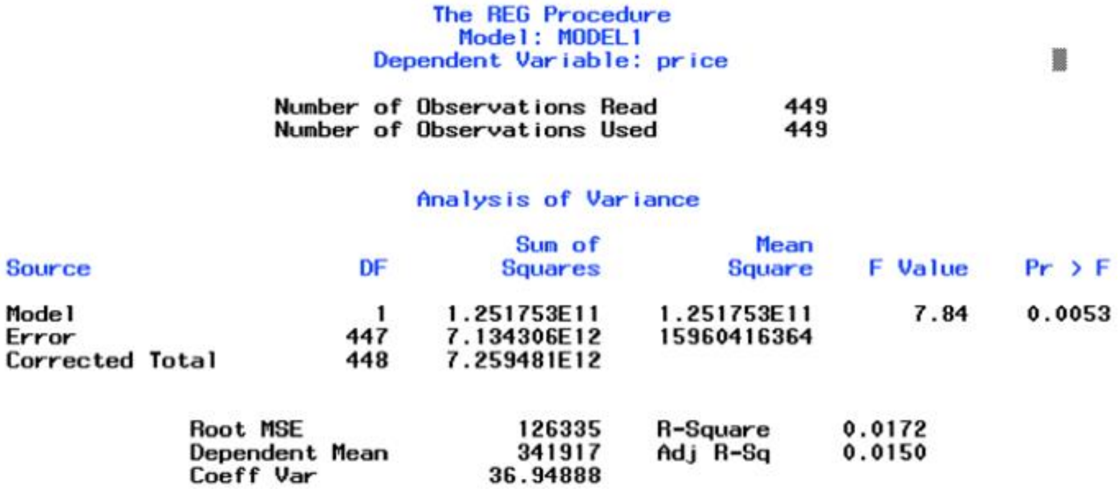
After that, we need to check for multicollinearity again.

\



```
                                   Parameter Estimates

                        Parameter     Standard                                          Variance
Variable      DF         Estimate        Error    t Value    Pr > |t|     Tolerance    Inflation

Intercept      1           312638        12035      25.98      <.0001             .            0
hoa            1         17.85426      6.37536       2.80      0.0053       1.00000      1.00000
```

```
                The REG Procedure
                 Model: MODEL1
           Dependent Variable: price

             Collinearity Diagnostics
```

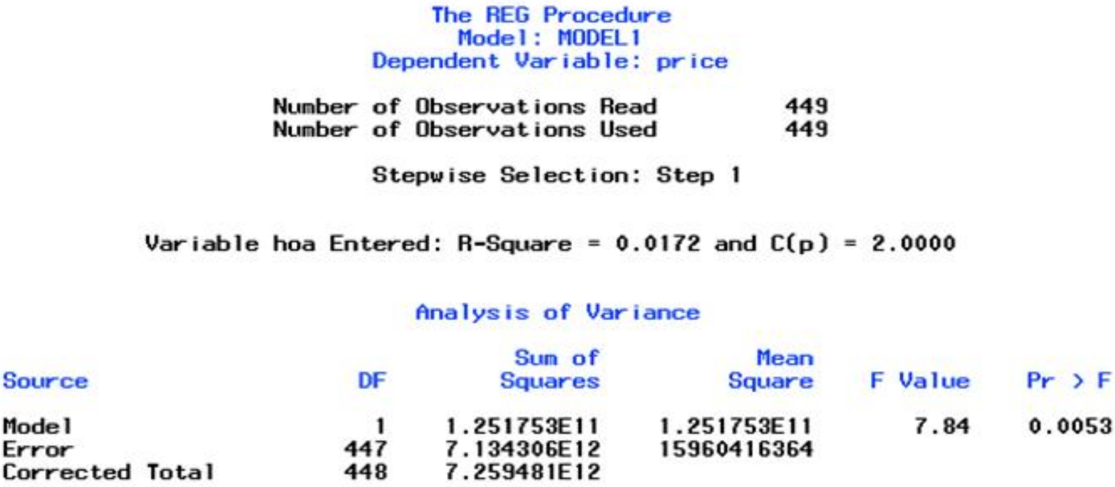


```
                         The REG Procedure
                          Model: MODEL1
                    Dependent Variable: price

              Number of Observations Read          449
              Number of Observations Used          449

                        Analysis of Variance

                                      Sum of           Mean
Source                  DF           Squares         Square    F Value    Pr > F

Model                    1        1.251753E11    1.251753E11       7.84    0.0053
Error                  447        7.134306E12    15960416364
Corrected Total        448        7.259481E12

          Root MSE                  126335    R-Square     0.0172
          Dependent Mean            341917    Adj R-Sq     0.0150
          Coeff Var               36.94888
```

Table 8: Regression procedure dependent variable price

From the data, the tolerance of the HOA variable is greater than 0.2, which indicates that it has no tolerance problem.

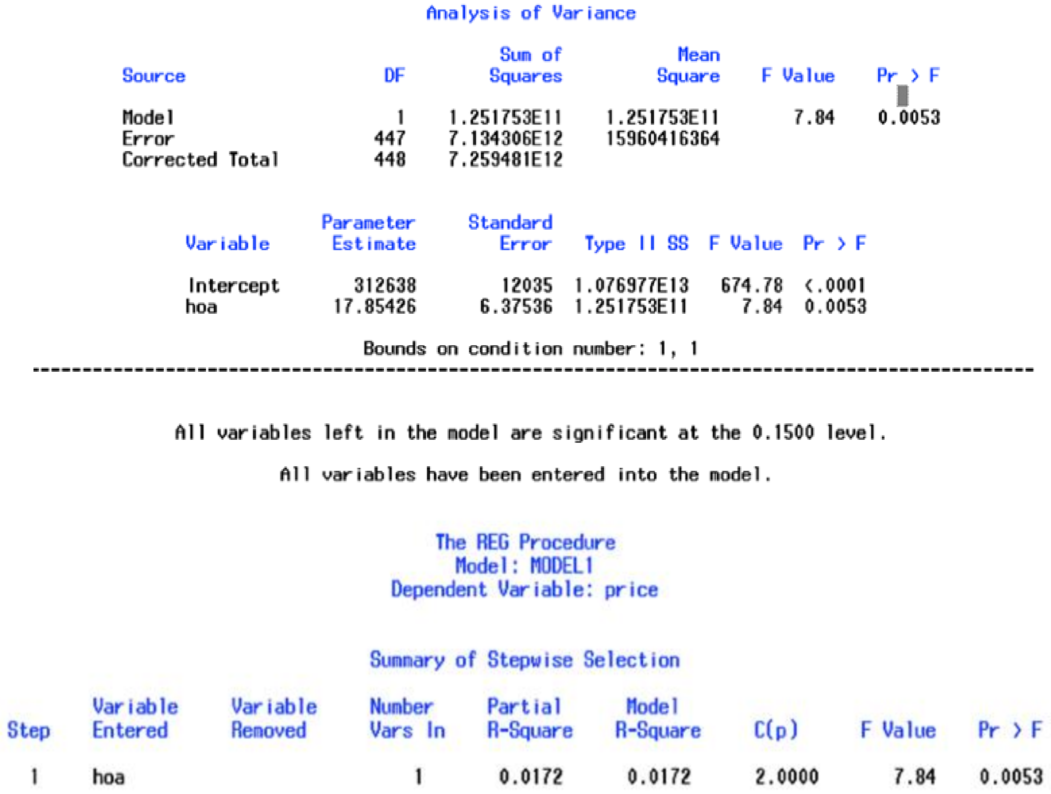
Table 9: Regression Procedure Dependent Variable price



```
                          The REG Procedure
                           Model: MODEL1
                      Dependent Variable: price

              Number of Observations Read          449
              Number of Observations Used          449

                    Stepwise Selection: Step 1

        Variable hoa Entered: R-Square = 0.0172 and C(p) = 2.0000

                         Analysis of Variance

                                    Sum of         Mean
Source                   DF        Squares       Square    F Value    Pr > F

Model                     1     1.251753E11  1.251753E11      7.84    0.0053
Error                   447     7.134306E12  15960416364
Corrected Total         448     7.259481E12
```

3.5. Then we use the Stepwise Regression method to get the equation:

Table 10: Analysis of Variance



```
                              Analysis of Variance

                                        Sum of          Mean
Source                      DF         Squares        Square    F Value    Pr > F

Model                        1     1.251753E11   1.251753E11       7.84    0.0053
Error                      447     7.134306E12   15960416364
Corrected Total            448     7.259481E12

                 Parameter     Standard
      Variable    Estimate        Error   Type II SS  F Value  Pr > F

      Intercept     312638        12035  1.076977E13   674.78  <.0001
      hoa         17.85426      6.37536  1.251753E11     7.84  0.0053

                        Bounds on condition number: 1, 1
-----------------------------------------------------------------------------------------

             All variables left in the model are significant at the 0.1500 level.

                       All variables have been entered into the model.

                                The REG Procedure
                                 Model: MODEL1
                            Dependent Variable: price

                          Summary of Stepwise Selection

       Variable   Variable   Number    Partial    Model
Step   Entered    Removed    Vars In   R-Square   R-Square     C(p)    F Value    Pr > F

   1   hoa                         1     0.0172     0.0172   2.0000       7.84    0.0053
```

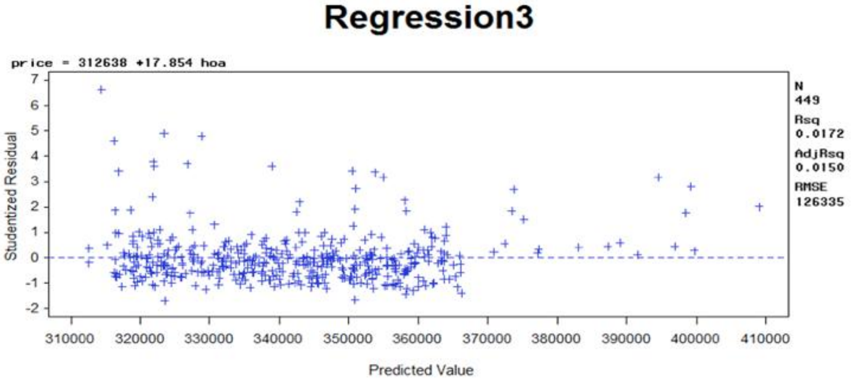
Table 11: Regression results, stepwise model

The intercept for HOA is not significant because the p-value is 0.053 which is > 0.001.
From the multiple regression model we can get the equation as:

Price=312638+17.854Hoa.

Finally, we need to check the variance.

Graph 14: Scatter plot between predicted value and student zed residual



Through this diagram, it can meet the requirement of constant variance in general. So the best model is: **Price=312638+17.854Hoa**.

IV. CONCLUSION

This paper is about what factors would have influence on the price of the houses. Firstly, we did the preliminary analysis and chose 12 variables to predict the price. Then we used the method of multiple linear regression to analyze how these factors affect the price of the houses. After the analysis, we chose seven variables (HOA, size, age, yards, floors, bedrooms, bathrooms) to include in our model. There were many outliers, and finally, we included HOA in our model. Because the easiest model is the best, we chose to use one variable to for prediction. Finally, we got the equation: **Price=312638+17.854Hoa**

Once we solved this paper question, we got a clear idea that HOA is one of the most useful indexes to estimate the price of house. At the same time, it has most important influence on the price. When we look forward to buying a house, we can include the information of HOA in this regression model, and construct an estimate quote for the price and compare that with the real price. Then we can decide whether to buy or not.

However, when you decide to buy a house you should consider other factors that influence the price, including the region, age, size, area, yards, number of bedrooms, floors and bathrooms etc. All of these elements will have a large or small influence on the price. If we want to get a more precise model, many other factors should be considered. However, it may consume too much time to collect the data and find the best regression model.

REFERENCES

[1]. A procedure for stepwise regression analysis (Articles &Statistical Papers, December 1992, Volume 33, Issue 1, pp 21-29

[2]. How to use SAS® to fit Multiple Logistic Regression Models Anpalaki J. Ragavan, Department of Mathematics and Statistics, University of Nevada, Reno, NV 89557(369-2008)

[3]. Introduction to Linear Regression Analysis (By Douglas C. Montgomery, Elizabeth A. Peck, G.)

[4]. Lipsey R.G. & Rosenbluth G. "A Contribution to the New Demand Theory: A Rehabilitation of the Giffen Good." *Canadian Journal Econ*. 4 (May 1971), pp. 131-163.

[5]. The Little SAS Book for Enterprise Guide 4.2 (By Susan J. Slaughter, Lora D. Delwiche)

[6]. Muth R. "Household Production and Consumer Demand Functions Econometrica 34 July 1966 pp. 699-708

[7]. S. Rosen "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition" Author(s): Journal of Political Economy, Vol. 82, No. 1 (Jan. - Feb., 1974), pp. 34-55Realtor.com. http://www.realtor.com

[8]. Regression with SAS (http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter1/sasreg1.htm)

[9]. ReliaSoft's Experiment Design and Analysis Reference

[10]. SAS System for Regression (Third Edition, By Rudolf J. Freud, Ph.D., Ramon C. Littell, Ph.D.)

[11]. Statistics Using SAS Enterprise Guide (By James B. Davis, Ph.D.)

[12]. Using Multivariate Statistics (FIFT H EDITION, Barbara G. Tabachnick California State University, Northridge Linda S. Fidell California State University, Northridge