|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  | |  | |
| **Econometrics Assignment** | | | |
|  |  | |  |
|  |  | |  |

Econometrics Assignment

From the dataset, the variables are described below.

Note: Excel is used for analyzing the below variables.

Description of Variables

|  |  |
| --- | --- |
| Sl. No. | Serial Number of the house |
| Price | Price of the house in Dollars |
| Living Area | Living Area in Sq. Ft. |
| Bathrooms | Number of Bathrooms |
| Bedrooms | Number of Bedrooms |
| Lot Size | Size of the Plot (in acres) |
| Age | Age of the house (in months) |
| Fireplace | Is there a fireplace in the house? – Yes or No |
| Locality | City Zone (there are three zones – A, B and C) |

The main objective is to build an appropriate regression model for predicting the price of the house. For that I took the data set given in excel sheet and ran regression using Data Solver in excel. The results of the regression equation without adding any new variables are as below,

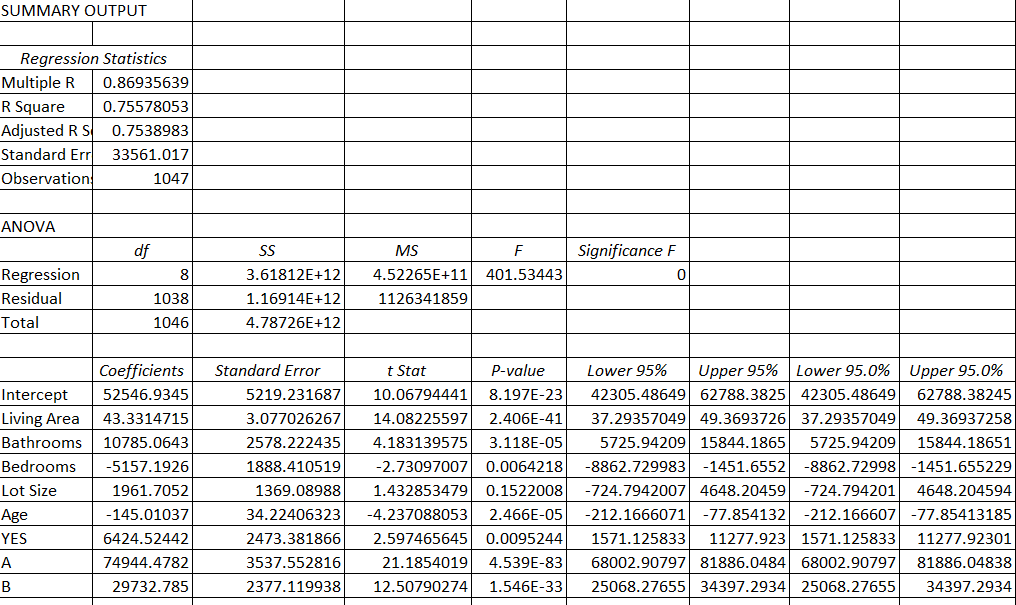
R2 = 0.7557

Dependent variables (Y) = Price

Independent variables = Living Area, Lot size, Bathrooms, Bedrooms, Age, Fireplace(YES), Location A, Location B

Regression equation is as below,

Price =52546.93+ 43.3314 Living Area + 10785.0643 Bathrooms – 5157.1926 Bedrooms + 1961.7052 Lot Size – 145.0103 Age + 6424.5244 Fireplace (YES) + 74944.4782 Location A + 29732.785 Location B



From the above summary table, below interpretations are made,

1. The coefficient 43.3314 Living Area indicates that for every increase in square feet in Living Area you can expect Price to increase by 43.3314 times.
2. The coefficient 10785.0643 Bathrooms indicates that for every increase in Bathroom you can expect Price to increase by 10785.0643 times.
3. For every additional increase in Bedroom, we expect 5157.1926 unit decrease in Price, holding all other variables constant.
4. The coefficient 1961.7052 Lot Size indicates that for every increase in square feet in Lot Size you can expect Price to increase by 1961.7052 times.
5. For every Additional increase in Age, we expect 145.0103 unit decrease in Price, holding all other variables constant.
6. The coefficient 6424.5244 Fireplace (YES) indicates that for every increase in Fireplace you can expect Price to increase by 6424.5244 times.
7. The coefficient 74944.4782 Location A indicates that for every additional house in Location A you can expect Price to increase by 74944.4782 times.
8. The coefficient 29732.785 Location B indicates that for every additional house in Location A you can expect Price to increase by 29732.785 times.
9. From the regression equation, we see that the intercept value is 52546.9345. If all the other coefficients are zero, the regression equation predicts that the Price of a House is Rs. 52546.9345. But all the remaining variables can’t be zero as there will be no house without any bathroom, bedroom, lot size 😊.
10. Adjusted R square is nothing but it is calculated R square from only those variables whose addition in the model which are significant. So always while we do multivariate linear regression we look at adjusted R square instead of R square.

In the above summary tables we got Adjusted R2 = 0.75389

Goodness of fit is normally measured by R2, it talks about how good the regression equation is, as our value is 0.7538 which means 75% of the variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line. 75% of the values fit the model.

1. Anova Table Interpretations: Analysis of Variance

Total sum of squares = 4787263161368.61 i.e. Total variation in Price

Error sum of squares = 1169142849869.51 i.e. Variation in errors

Regression Sum of squares = 3618120311499.1 i.e. Variation of estimated values of Price

1. Mean square error value is 1126341859.21918 which means on an average how much the predicted values are deviating from the actual values. Given, 3000 sq. ft. Living area, with a fireplace, 4 bedrooms, 2.5 bathrooms with a Lot Size of 1.5 acres, by substituting the above values we get,

Predicted Value = 145695.3872 which is the Price of a House with fireplace for Location C.

Price of House with fireplace for Location B= 175428.1721 (145695.3872 + 29732.785)

Price of House with fireplace for Location A= 220639.8653 (145695.3872+ 74944.4782)

Addition New Variables:

Tried adding the new variables to check whether the R2 value might increase, New variables are

1. Living Area / bedrooms ( R2 = 0.755712)
2. Average number of bathrooms per bedroom (R2 = 0.756454)
3. Price/Living Area ( R2 = 0.9471)

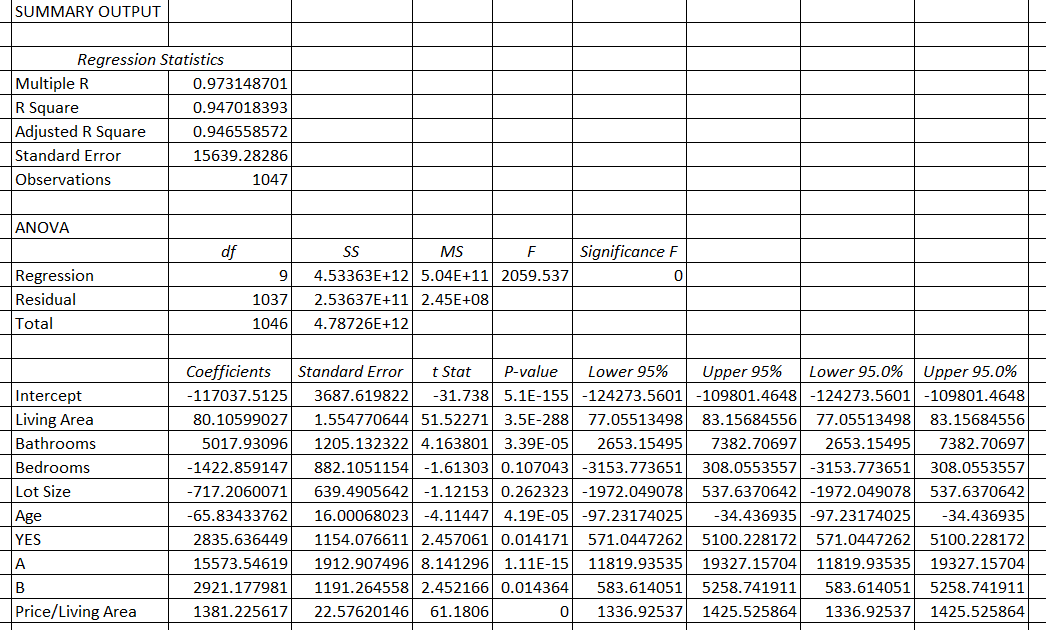
Hence the final model will be by adding the new variable Price/Living Area, we are not taking a) and b) variables because they are not significant.

From the below table, Predicted value of Price is ,

Predicted Value = 248931.2which is the Price of a House with fireplace for Location C.

Price of House with fireplace for Location B= 251852.4

Price of House with fireplace for Location A= 264504.7

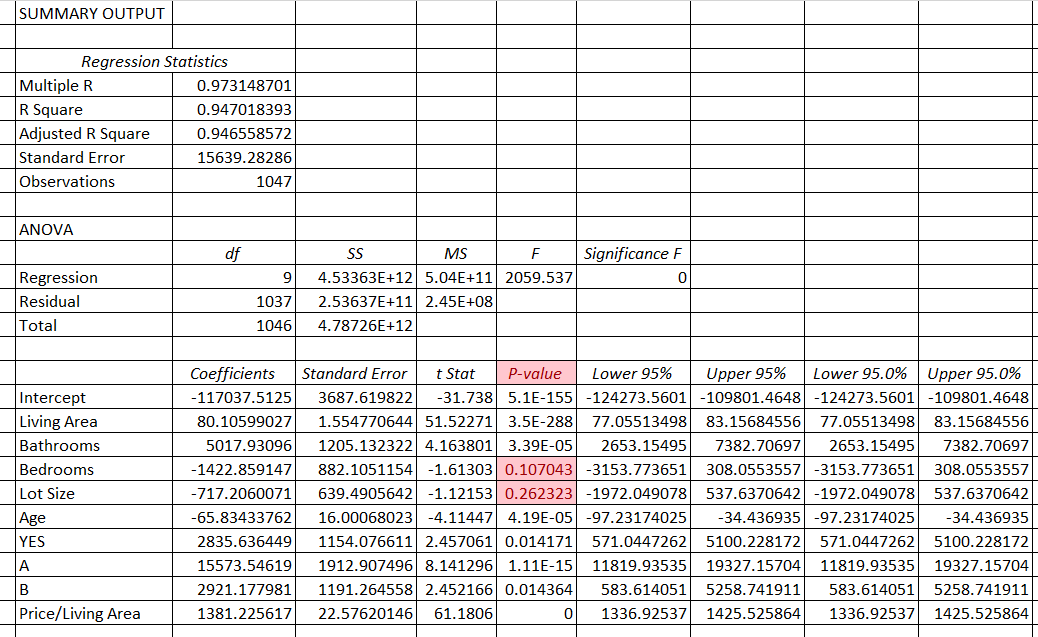


Interpretation of F Value:

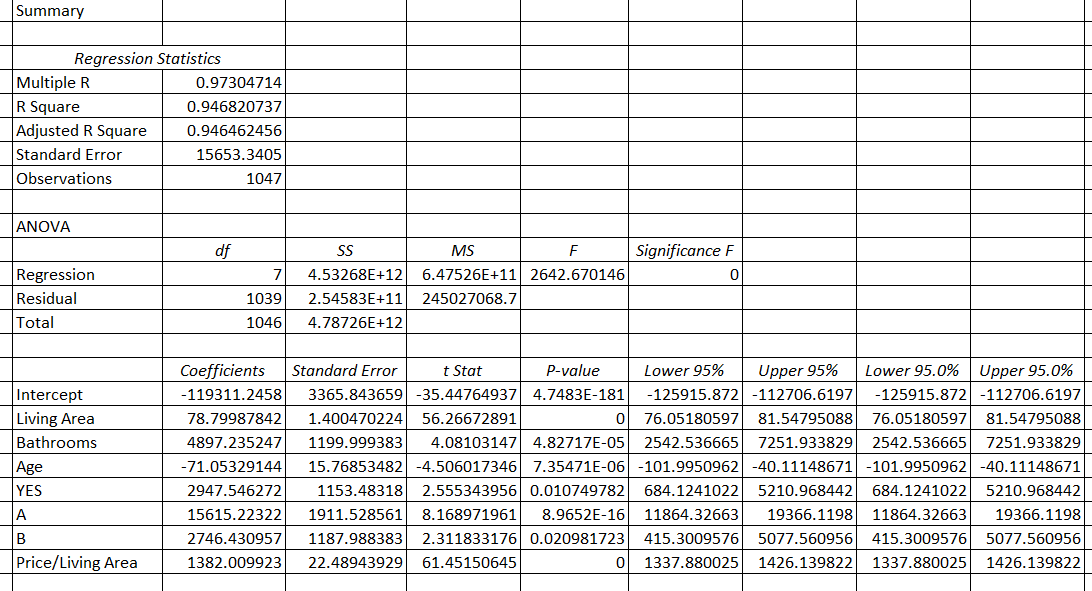
1. Normally F-tests can compare the fits of different models, [test the overall significance in regression models](https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/), test specific terms in linear models, and determine whether a set of means are all equal.

F= {\displaystyle \frac {{\text{between-groups variance}}}{{\text{within-group variance}}}}

1. F value is 2059.537. A high F value means that our data does not well support null hypothesis Or in other words, the alternative hypothesis is compatible with observed data.
2. The P value,  A low probability indicates that our sample data are unlikely when the null hypothesis is true. The larger the F value the greater the relative variance among the group means. The p value tells you the probability of obtaining an F value as extreme or more extreme as the one observed under the assumption that the null hypothesis is true.
3. If our p – values are greater than 0.05, it means that the variables corresponding to greater than 0.05 p value is not significant, hence we can remove it.



After we remove the variables Lot Size and Bedrooms, I ran the regression again to check the value of R2 . Results in the below summary table,



We get the R2 value as 0.946, now from the P- values, it is clear that all the variables are significant except Lot Size and bedrooms hence we removed it.

For the above summary table we get the price of houses as,

Price of House with fireplace for Location A = 251306.1

Price of House with fireplace for Location B = 254052.5

Price of House with fireplace for Location C = 266921.3

# To check the Heteroscedasticity, plotted scatter plot against Predicted Price and Residual2, the graph is as below, A scatterplot of these variables will often create a cone-like shape, as the scatter (or variability) of the dependent variable (DV) widens or narrows as the value of the independent variable (IV) increases. The inverse of heteroscedasticity is homoscedasticity, which indicates that a DV's variability is equal across values of an Independent variable. A test of homoscedasticity of error terms determines whether a regression model's ability to predict a DV is consistent across all values of that DV.

Let us now see the using Mallows Test- Where Mallows Cp helps to identify the right number of explanatory variables,

Calculating Cp value for partial model (excluding lot size), no. of variables is equal to 7 and Cp value is also 7 , hence the predicted variables which we took are correct.

|  |  |
| --- | --- |
| Diagnostics by using Mallow's model |  |
| SSE/MSE | 1039 |
| Cp =SSE/MSE-(n-2p) | 7 |

Multicollinearity : Sometimes high correlation between predictor variables could be problematic.

Variance Inflation Factors are typically used to confirm this suspicion,

VIF= 1/ (1-R2)

VIF for living area as dependent variable and bathrooms, bedrooms, lot size, age of the house, fireplace and locality as independent variables = 3.528004

VIF for Bathrooms as dependent variable and living area, bedrooms, lot size, age of the house, fireplace and locality as independent variables = 2.538583

VIF for Bedrooms as dependent variable and living area, bathrooms, lot size, age of the house, fireplace and locality as independent variables = 1.864152

VIF for Age as dependent variable and living area, bathrooms, lot size, Bedrooms, fireplace and locality as independent variables = 1.322392

VIF for Fireplace as dependent variable and living area, bathrooms, lot size, Age of the house, Bedrooms, fireplace and locality as independent variables = 1.322392

A rule of thumb for interpreting the variance inflation factor:

1 = not correlated.

Between 1 and 5 = moderately correlated.

Greater than 5 = highly correlated.

Hence all the above variables are moderately correlated as its range is in between 1 and 5.

In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above.

Let us now see the 95% Confidence Interval for the predicted price,

After calculation of TSTAT \* SE/SQRT(n) and adding, subtracting it from the predicted price we get the below results,

|  |  |  |
| --- | --- | --- |
|  | lower limit | upper limit |
| Price for location C | 250356.7903 | 252255.3226 |
| Price for location B | 253103.2213 | 255001.7536 |
| Price for location A | 265972.0136 | 267870.5458 |

Therefore, the final model has no autocorrelation and multicollinearity. The no. of variables has been proved by Mallows Model.

THANK YOU