1/21/2020

Real Estate Evaluation Using Regression

Instructor: Sir Mueen Azad

Submitted by:

Faisal Riaz (F2019313020)

Waseem Abbas (F2019313001)

Nauman Anwar (F2019313009)

Table of Contents

Real Estate valuation using Regression

Executive Summary:

The Multiple Regression Analysis can be used to predict the price value of real estate, provided that comprehensive dataset and all relative variables are available. If all the requirements are fulfilled, then any real estate company can successfully evaluate property value and can predict the future value as well. Accurate and well-built model using different methods can replace the Real Estate Evaluators for sure. For this purpose, the Multiple Regression Model can use any of the two methods, i.e. the Enter Method or the Stepwise Method. Through using any of these two methods, the price can be estimated. The price estimation can help the real estate builders in evaluating their profit margins in the buying and selling of their respective real estate property assets.

1. Introduction of Data Set and Purpose of the Project:

The price of a Real Estate property can be determined by Real Estate evaluators based on their experience and based on many other factors. There are many factors that contribute to price evaluation of property. Factors like geographical location, neighborhood, number of schools in the area, distance from market, no of bedrooms, bathrooms, parking space, basement and furnishing status of the property.

We have taken a basket case analysis Dataset and will predict the price of the property using Multiple Linear Regression using SPSS. Dataset consist of 545 properties including their price, area, bedrooms, bathrooms, stories, parking and furnishing status.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
price	Numeric	8	0	Priceof House(Rs)	None	None	8	Center	Scale	Input
area	Numeric	5	0	Covered SqFt	None	None	8	Center	Scale	Input
bedrooms	Numeric	1	0	Bedrooms	None	None	8	Center	Scale	Input
bathrooms	Numeric	1	0	Bathrooms	None	None	10	Center	Scale	Input
stories	Numeric	1	0	Stories	None	None	8	Center	Scale	Input
parking	Numeric	1	0	Parking	None	None	8	Center	Scale	Input
furnishingst...	Numeric	1	0	FurnishingStatus	{0, Unfurnished...	None	14	Center	Nominal	Input

All the variables have been explained in figure shown above. One *variable* named *"furnishing status"* has three values *"unfurnished ->0* ", "*semi furnished -> 1*" and "*furnished -> 2"*. Since *"furnishing status"* is a categorical variable so we will transform and recode *"furnishing statuses"* into three dummy variables as shown below:

unfurnished	Numeric	8	0	Unfurnished Dummy	None	None	13	Center	Nominal	Input
semifurnished	Numeric	8	0	Semifurnished Dummy	None	None	15	Center	Nominal	Input
furnished	Numeric	8	0	Furnished Dummy	None	None	11	Center	Nominal	Input

We will be using all above mentioned variables and conduct a series of multiple regression analyses that will eventually narrow down the best model predictor of *price* to *5 variables* and *n-1 dummy variables.*

1.1. Multiple Regression Equation for the Data

We will be predicting the price of house and our model is

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

$$Y = E(Y) + \varepsilon$$

$$Price = \beta_0 + \beta_1(area) + \beta_2(bedrooms) + \beta_3(bathrooms) + \beta_4(stories) + \beta_5(parking) + \beta_6(semifurnished) + \beta_7(furnished) + \varepsilon$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-292814.345	244972.751		-1.195	.232	-774036.718	188408.028		
	Covered SqFt	319.352	26.203	.371	12.188	.000	267.880	370.825	.845	1.183
	Bedrooms	147960.105	81415.898	.058	1.817	.070	-11972.587	307892.798	.757	1.321
	Bathrooms	1096881.397	116689.323	.295	9.400	.000	867657.892	1326104.903	.795	1.258
	Stories	535718.044	67617.702	.248	7.923	.000	402890.410	668545.677	.794	1.259
	Parking	339628.754	65917.744	.156	5.152	.000	210140.503	469117.006	.847	1.180
	Semifurnished Dummy	459881.910	123749.674	.121	3.716	.000	216789.110	702974.709	.733	1.364
	Furnished Dummy	664843.219	141972.644	.155	4.683	.000	385953.375	943733.063	.709	1.411

a. Dependent Variable: Priceof House(Rs)

Price = -292814.345 + 319.352(area)+147960.105(bedrooms)+1096881.397(bathrooms) + 535718.044(stories) + 339628.754(parking) + 459881.910(semi-furnished) + 664843.219(furnished)

2. Assumptions of Linear Regression:

2.1. Regression model is linear in parameters

Linearity in relationship means that dependent variable and all independent variables have linear relationship. Both the variables (x) and β are raised to power one only so the equation is linear in beta parameters and variables as well. So, this condition is satisfied.

2.2. Mean of the Residuals is zero

Residuals Statistics[a]

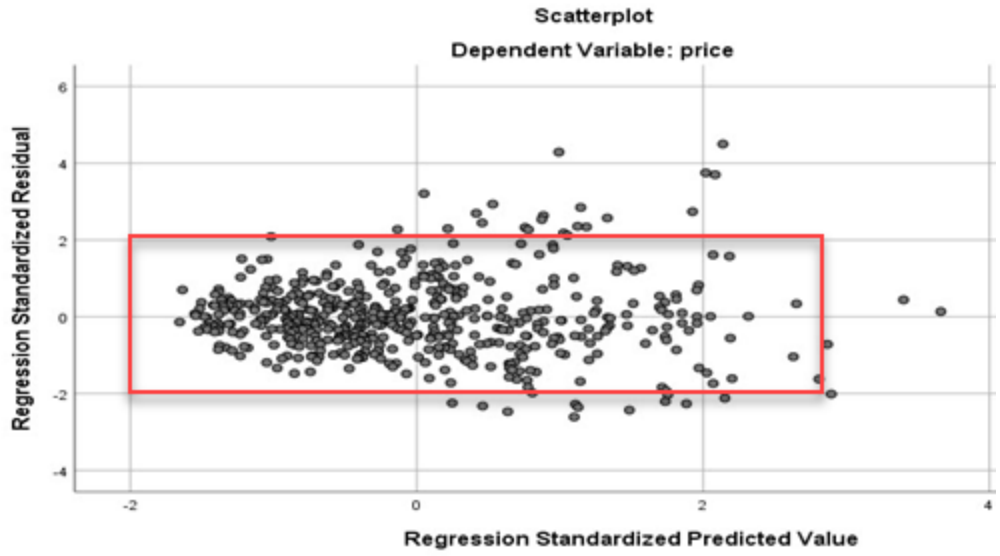
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2406402.50	11374550.00	4766729.25	1425145.161	545
Residual	-3182694.000	5486362.000	.000	1211406.466	545
Std. Predicted Value	-1.656	4.637	.000	1.000	545
Std. Residual	-2.610	4.500	.000	.994	545

a. Dependent Variable: Priceof House(Rs)

According to this assumption the mean of the residual should be zero for a particular model. As shown the mean of the residual for our model is zero highlighted above. We can conclude that this assumption is satisfied as well.

2.3. Homoscedasticity of residuals or equal variance

Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable. The impact of violating the assumption of homoscedasticity is a matter of degree, increasing as heteroscedasticity increases.



If we observed carefully the graph then it will be obvious that all the points lie in the shape of rectangle ranging from -2 to 2 along Y-axis and -2 to 3 along X -axis are not scattered significantly with an exception of outliers. We can conclude that this assumption is satisfied as well.

2.4. No autocorrelation of residuals

A rule of thumb is that test statistic values for *Durban Watson* in the range of *1.5 to 2.5* are relatively normal. In our case

Durban Watson =1.771. This assumption is satisfied as well.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.762[a]	.581	.575	1219276.474	.581	106.173	7	537	.000	1.771

a. Predictors: (Constant), Furnished Dummy, Bedrooms, Parking, Covered SqFt, Bathrooms, Stories, Semifurnished Dummy

b. Dependent Variable: Priceof House(Rs)

2.5. The number of observations must be greater than number of *X's*

This assumption states that Number of observations should be greater than number of independent variables. In Our dataset we have 1 *dependent variable* and 6 *independent variables* including a *categorical variable* with *2 dummy variables*. Total number of observations are 545. Therefore, this assumption is satisfied.

2.6. No perfect multicollinearity

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity

VIF is a metric computed for every X variable that goes into a linear model. If the VIF of a variable is high, it means the information in that variable is already explained by other X variables present in the given model, which means, more redundant is that variable. So, lower the

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-292814.345	244972.751		-1.195	.232	-774036.718	188408.028		
	Covered SqFt	319.352	26.203	.371	12.188	.000	267.880	370.825	.845	1.183
	Bedrooms	147960.105	81415.898	.058	1.817	.070	-11972.587	307892.798	.757	1.321
	Bathrooms	1096881.397	116689.323	.295	9.400	.000	867657.892	1326104.903	.795	1.258
	Stories	535718.044	67617.702	.248	7.923	.000	402890.410	668545.677	.794	1.259
	Parking	339628.754	65917.744	.156	5.152	.000	210140.503	469117.006	.847	1.180
	Semifurnished Dummy	459881.910	123749.674	.121	3.716	.000	216789.110	702974.709	.733	1.364
	Furnished Dummy	664843.219	141972.644	.155	4.683	.000	385953.375	943733.063	.709	1.411

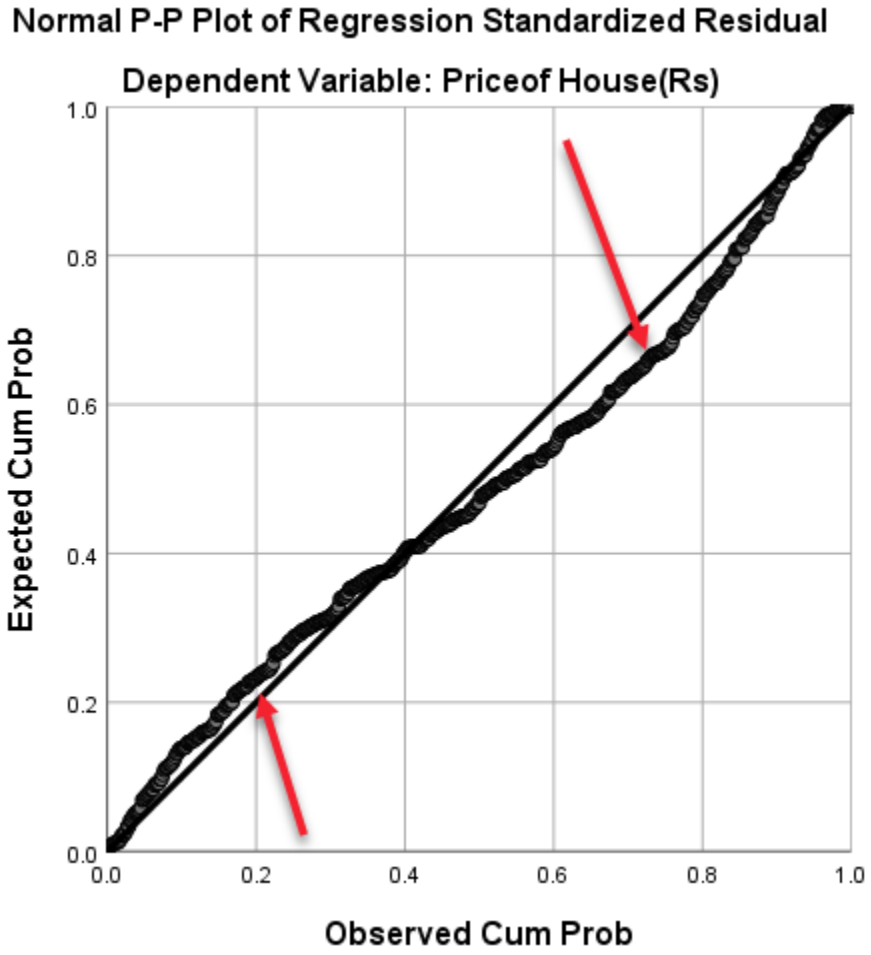
a. Dependent Variable: Priceof House(Rs)

VIF (<2) the better

As shown in the Fig above none of the *variables* have *VIF >2* so we can safely assume that there is no multicollinearity.

2.7. Normality of the Residuals

Residuals should be normally distributed. This also implies the *Y* and the *X's* are also normally distributed.



Points lie exactly on the line; it is perfectly *normal distribution*. However, some deviation is to be expected, particularly near the ends *(note the upper right and lower left)*, but the deviations should be small. We conclude that this assumption is satisfied as well.

8. The variability in X values is positive

This means the *X* (*independent variables*) values in a given sample must not all be the same (or even nearly the same). The variance in the *X variable* is much larger than 0 as shown in the table below. So, this assumption is satisfied.

Descriptive Statistics

	Mean	Std. Deviation	N
Priceof House(Rs)	4766729.25	1870439.616	545
Covered SqFt	5150.54	2170.141	545
Bedrooms	2.97	.738	545
Bathrooms	1.29	.502	545
Stories	1.81	.867	545
Parking	.69	.862	545
Semifurnished Dummy	.42	.493	545
Furnished Dummy	.26	.437	545

As we observe that *Std. Deviation is > 0* in other words *Variance* which is square root of *Std. Deviation* is also greater than 0. Therefore, this condition is satisfied as well.

3. Multiple Regression Model Using Enter Method:

We will use "***Enter***" Method to predict the price of property using all the *X'. level of significance = 5%* and base level for *dummy variables* is set to *unfurnished=0.*

a. **At the 5% significance level, determine if the model is significant or not**?

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.105E+15	7	1.578E+14	106.173	.000[b]
	Residual	7.983E+14	537	1.487E+12		
	Total	1.903E+15	544			

a. Dependent Variable: Priceof House(Rs)

b. Predictors: (Constant), Furnished Dummy, Bedrooms, Parking, Covered SqFt, Bathrooms, Stories, Semifurnished Dummy

F_{val} = *106.173*

P_{val} = *0.000*

df = (7,537)

Since *p-value is < 0.05(level of significance)* therefore we conclude that our Model is significant.

b. At the 5% level of significance, does it appear that any of the predictor variables can be removed from the full model as unnecessary?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-292814.345	244972.751		-1.195	.232	-774036.718	188408.028		
	Covered SqFt	319.352	26.203	.371	12.188	.000	267.880	370.825	.845	1.183
	Bedrooms	147960.105	81415.898	.058	1.817	.070	-11972.587	307892.798	.757	1.321
	Bathrooms	1096881.397	116689.323	.295	9.400	.000	867657.892	1326104.903	.795	1.258
	Stories	535718.044	67617.702	.248	7.923	.000	402890.410	668545.677	.794	1.259
	Parking	339628.754	65917.744	.156	5.152	.000	210140.503	469117.006	.847	1.180
	Semifurnished Dummy	459881.910	123749.674	.121	3.716	.000	216789.110	702974.709	.733	1.364
	Furnished Dummy	664843.219	141972.644	.155	4.683	.000	385953.375	943733.063	.709	1.411

a. Dependent Variable: Priceof House(Rs)

$\boldsymbol{\beta_0}$=-292814.345, t = -1.195, Pval = 0.232

$\boldsymbol{\beta_1}$=319.352, t = 12.188, Pval = 0.000

$\boldsymbol{\beta_2}$=147960.105, t = 1.817, Pval = 0.070

$\boldsymbol{\beta_3}$=1096881.397, t = 9.400, Pval = 0.000

$\boldsymbol{\beta_4}$=535718.044, t = 7.923, Pval = 0.000

$\boldsymbol{\beta_5}$=339628.754, t = 5.512, Pval = 0.000

$\boldsymbol{\beta_6}$=459881.910, t = 3.716, Pval = 0.000

$\boldsymbol{\beta_7}$=664843.219, t = 4,683, Pval = 0.000

Since *p-value for* $\boldsymbol{\beta_0}$ *= 0.232 is > 0.05(level of significance)* therefore we conclude that $\boldsymbol{\beta_0}$ is not significant. Similarly, $\boldsymbol{\beta_2}$*= 0.07 is > 0.05(level of significance)* therefore we conclude that $\boldsymbol{\beta_2}$ is not significant.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.762[a]	.581	.575	1219276.474	.581	106.173	7	537	.000	1.771

a. Predictors: (Constant), Furnished Dummy, Bedrooms, Parking, Covered SqFt, Bathrooms, Stories, Semifurnished Dummy

b. Dependent Variable: Priceof House(Rs)

c. Interpretation the coefficient of multiple R^2_{adj}

The coefficient of multiple determination is ***0.575***. Therefore, about 57.5 % of the variation in the price of the property is explained by all ***X's***. The value of R^2_{adj} is not quite close to 1 but coefficient of determination is always low once you introduce dummy variables. The Regression equation appears to be useful for making predictions.

d. Final Equation of Regression Model using Enter Method.

Price=319.352(area)+1096881.397(bathrooms)+535718.044(stories)+ 39628.754(parking) + 459881.910(semi-furnished) +664843.219 (furnished)

Though β_2 coefficient for Bedrooms is not significant but logically we cannot remove this independent variable. Whenever you plan to buy a house, the first question is the no. of bedrooms so we cannot ignore this variable in our final Model. Final model based on our understanding is

Price=319.352(area)+147960.105(bedrooms)+1096881.397(bathrooms)+535718.044(stories)+ 39628.754(parking) + 459881.910(semi-furnished) +664843.219 (furnished)

4. Multiple Regression Model Using Stepwise Method:

We will use "***Stepwise***" Method to predict the price of property using all the *X'. level of significance = 5%* and base level for *dummy variables* is set to *unfurnished=0.*

a. At the 5% significance level, determine if the model is significant or not?

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.468E+14	1	5.468E+14	218.884	.000[b]
	Residual	1.356E+15	543	2.498E+12		
	Total	1.903E+15	544			
2	Regression	8.852E+14	2	4.426E+14	235.623	.000[c]
	Residual	1.018E+15	542	1.878E+12		
	Total	1.903E+15	544			
3	Regression	1.010E+15	3	3.368E+14	204.112	.000[d]
	Residual	8.927E+14	541	1.650E+12		
	Total	1.903E+15	544			
4	Regression	1.062E+15	4	2.656E+14	170.604	.000[e]
	Residual	8.407E+14	540	1.557E+12		
	Total	1.903E+15	544			
5	Regression	1.078E+15	5	2.156E+14	140.872	.000[f]
	Residual	8.250E+14	539	1.531E+12		
	Total	1.903E+15	544			
6	Regression	1.100E+15	6	1.833E+14	122.793	.000[g]
	Residual	8.032E+14	538	1.493E+12		
	Total	1.903E+15	544			

Considering Model 6

F_{val} = *122.793*

P_{val} = *0.000*

df = (6,538)

Since *p-value is < 0.05(level of significance)* therefore we conclude that our *Model6* is significant.

b. At the 5% level of significance, does it appear that any of the predictor variables can be removed from the full model as unnecessary?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
6	(Constant)	-21175.063	194502.646		-.109	.913	-403252.788	360902.663		
	Covered SqFt	321.983	26.219	.374	12.281	.000	270.480	373.487	.848	1.180
	Bathrooms	1149931.475	113220.626	.309	10.157	.000	927522.782	1372340.167	.848	1.179
	Stories	575839.294	64048.773	.267	8.991	.000	450022.962	701655.625	.889	1.125
	Parking	346664.323	65944.638	.160	5.257	.000	217123.786	476204.860	.850	1.176
	Furnished Dummy	674315.904	142180.229	.158	4.743	.000	395019.454	953612.354	.710	1.409
	Semifurnished Dummy	473193.595	123796.789	.125	3.822	.000	230009.264	716377.925	.736	1.360

$\boldsymbol{\beta_0}$=-21175.063, t = -0.109, Pval = 0.913

$\boldsymbol{\beta_1}$=321.983, t = 12.281, Pval = 0.000

$\boldsymbol{\beta_3}$=1149931.475, t = 10.157, Pval = 0.000

$\boldsymbol{\beta_4}$=575839.294, t = 8.991, Pval = 0.000

$\boldsymbol{\beta_5}$=346664.323, t = 5.257, Pval = 0.000

$\boldsymbol{\beta_7}$=674315.904, t = 4.743, Pval = 0.000

$\boldsymbol{\beta_6}$=473193.595, t = 3.822, Pval = 0.000

Since *p-value for* $\boldsymbol{\beta_0}$ *= 0.913 is > 0.05(level of significance)* therefore we conclude that $\boldsymbol{\beta_0}$ is not significant.

Model Summary[g]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.536[a]	.287	.286	1580515.256	.287	218.884	1	543	.000	
2	.682[b]	.465	.463	1370520.929	.178	180.147	1	542	.000	
3	.729[c]	.531	.528	1284591.830	.066	75.936	1	541	.000	
4	.747[d]	.558	.555	1247766.822	.027	33.404	1	540	.000	
5	.753[e]	.566	.562	1237213.737	.008	10.251	1	539	.001	
6	.760[f]	.578	.573	1221883.023	.011	14.610	1	538	.000	1.780

a. Predictors: (Constant), Covered SqFt

b. Predictors: (Constant), Covered SqFt, Bathrooms

c. Predictors: (Constant), Covered SqFt, Bathrooms, Stories

d. Predictors: (Constant), Covered SqFt, Bathrooms, Stories, Parking

e. Predictors: (Constant), Covered SqFt, Bathrooms, Stories, Parking, Furnished Dummy

f. Predictors: (Constant), Covered SqFt, Bathrooms, Stories, Parking, Furnished Dummy, Semifurnished Dummy

g. Dependent Variable: Priceof House(Rs)

c. Interpretation the coefficient of multiple R^2_{adj}

The coefficient of multiple determination is ***0.573***. Therefore, about 57.3 % of the variation in the price of the property is explained by all ***X's***. The value of R^2_{adj} is not quite close to 1 but coefficient of determination is always low once you introduce dummy variables. The Regression equation appears to be useful for making predictions.

d. Final Equation of Regression Model using STEP Method.

Price *= 321.983(area) + 1149931.475(bathrooms) + 575839.294(stories) + 346664.323 (parking) + 473193.595 (semi-furnished) + 674315.904 (furnished)*

5. Conclusion:

We have used two different methods to predict the “Price” of Property using 5 continuous variables and one categorical with 3 levels using Regression. It is observed that “No. Of Bedrooms” is not a significant factor deciding the “Price” of a property. Furthermore, models built using “Enter Method” yields “57.5%” accuracy. In addition, it is pertinent to mention here that the “Stepwise Method” is accurate up to “57.3%” only. Either of these two methods, i.e. ‘Enter Method’ and ‘Stepwise Method’, can be used to predict the “Price”. Model can only explain 57% or total variation in Property. However, accuracy can be improved by adding more relative variables.