Forecast Market Cost of Houses using Multiple Regression Analysis

Priyanka
MSc Data Analytics
National College of Ireland
X20192037@student.ncirl.ie

Abstract - This paper forecast the prices of houses in the US Region using Multiple regression in Statistical Package for Social Sciences (SPSS) using data from a text file with details of the houses. The model reflects the factors involved in real estate development. The purpose of this analysis is to represent the relationship between the property characteristics and sale prices. The analysis using the dependent variable and independent variables aims to provide insights into real estate market trends to investors that establish trust between seller and buyer of the houses in the area.

Keywords - SPSS, multiple regression model, Correlation, descriptive statistics.

I. INTRODUCTION

House plays a major role in an individual daily life it is one of the major investments in a lifetime. The decision of paying for it includes various factors that vary from every other individual. The paper aims to find the sales price of the house that makes the price a dependent variable on two or more independent variables using multiple regression analysis to analyze the future trend of the US housing sales prices.

II. DATA DESCRIPTION

The Dataset of HouseDeatil.txt [1] has been loaded to SPSS it has total of **1728** observations and **16** variables of scale and nominal type. Data cleaning help in data processing for accuracy in the model prediction. The dataset has no missing value. Manually removed the auto created variable V17 in SPSS.

RECODE newConstruction ('yes'=1) (ELSE=0) INTO newlyConstructed.
VARIABLE LABELS newlyConstructed 'C'
RECODE heating ('electric'='1') (ELSE='0').
EXECUTE.

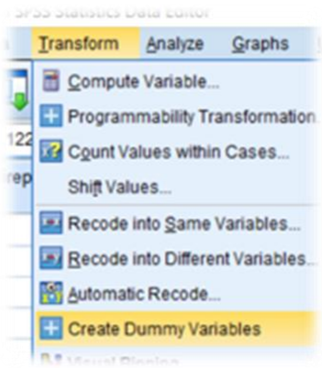
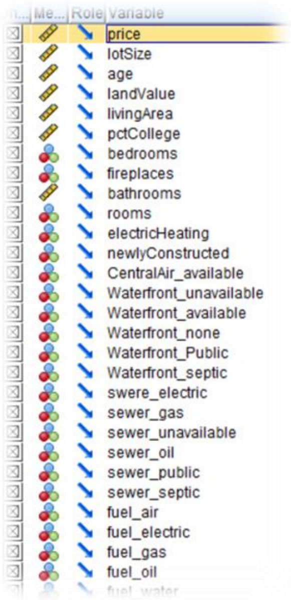

Transform
Analyze
Graphs
Compute Variable...
Programmability Transformation
Count Values within Cases...
Shift Values...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Create Dummy Variables

Figure 1 create dummy variable in SPSS.

Dummy variables are created for categorical variables for n values we have (n-1) dummy value in the factor.



Me... Role Variable
price
lotSize
age
landValue
livingArea
pctCollege
bedrooms
fireplaces
bathrooms
rooms
electricHeating
newlyConstructed
CentralAir_available
Waterfront_unavailable
Waterfront_available
Waterfront_none
Waterfront_Public
Waterfront_septic
swere_electric
sewer_gas
sewer_unavailable
sewer_oil
sewer_public
sewer_septic
fuel_air
fuel_electric
fuel_gas
fuel_oil
fuel_water

Figure 2 Variable list after creating dummy variables.

Multiple regression on the dataset contains two or more independent variables and one dependent variable which are continuous.The dataset considered in this analysis has one dependent continuous variable

i. Dependent variable: Price (US dollar)
ii. Independent variables:
   - LivingArea (square feet)
   - LivingValue (US dollar)
   - bathroom (half bathrooms have no shower or tub)

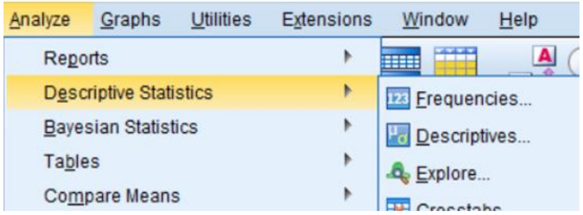
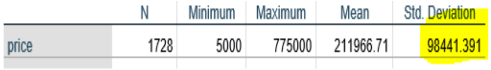


Figure 3 Descriptive Statistics in SPSS.

Price – Target variable data is not normally distributed as the descriptive statistics show the mean of the variable is 211966.71 and median is 189900.00 whereas the data range from 5000 to 775000 which is around 3.6 times of the mean and positively skewed as shown in the below histogram.



	N	Minimum	Maximum	Mean	Std. Deviation
price	1728	5000	775000	211966.71	98441.391

Figure 4 Descriptive Statistic Analysis of dependent variable price.

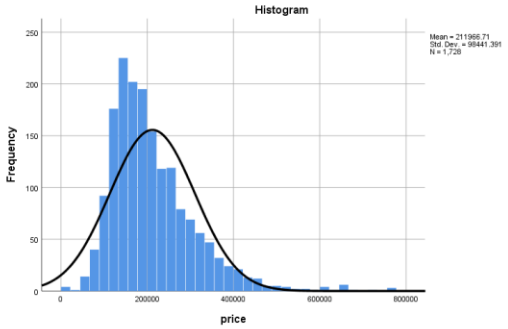
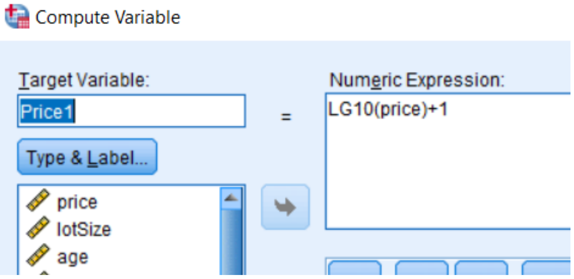


Figure 5 histogram of Salesprice.

Normalize the predictor for accurate prediction using the log transformation and values are stored in the newly created variable column named Price, the dependent variable for predicting the Sale price of houses.



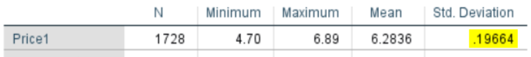
Compute Variable

Target Variable: Price1

=

Numeric Expression: LG10(price)+1

Figure 6 log transformation of Dependent variable Price.



	N	Minimum	Maximum	Mean	Std. Deviation
Price1	1728	4.70	6.89	6.2836	.19664

Figure 7 descriptive analysis of Price

- Sales Price data has been normally distributed

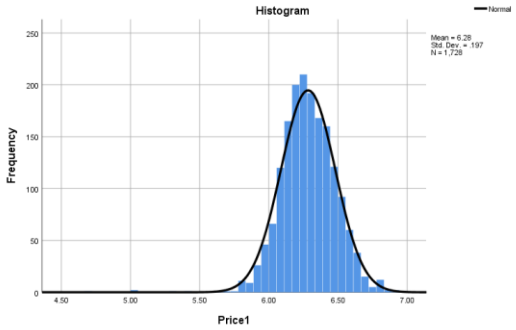


Figure 8 Normally distributed Sales Price

Independent variables are selected by checking the linearity using the scattered matrix graph and verified correlation matrix and then manually I removed the highly correlated variables and analyzed the impact on the model creation.

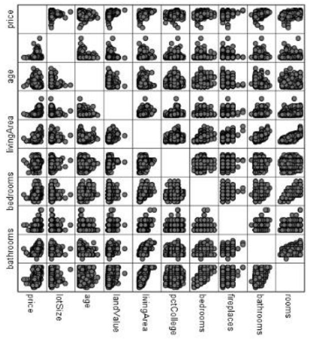


Figure 9 Matrix Scatter of variables.

III. MODEL BUILDING

Multiple regression model predicts the value of a variable when it is dependent on more than two variables.

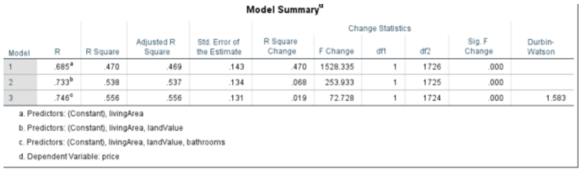
Regression analysis will be done on the two and more variables using the Pearson correlation coefficient of dependent variable Price with other factors involved in deciding the house price for checking the correlation of each independent variable with each other and as well as with dependent variable. It has a r-value range from -1 to +1.

We perform the variable selection by trying out the number of different models. Here till the time p-value is relatively less, we cannot consider any model. Also, an adjusted R square can be used to determine the quality of model. One of the approaches to perform this task is stepwise selection. In which we must start with all the variables and then eliminate the variable with the largest p-value with this we will get all the significant values.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

Figure 10 Mathematical model of multiple regression

Here, Y is the average sales price of US Houses, B is the beta coefficient, C is the constant, and X are the independent variables like bathroom, bedroom, etc.



Model Summary[d]

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	.685[a]	.470	.469	.143	.470	1528.335	1	1726	.000	
2	.733[b]	.538	.537	.134	.068	253.933	1	1725	.000	
3	.746[c]	.556	.556	.131	.019	72.728	1	1724	.000	1.583

a. Predictors: (Constant), livingArea

b. Predictors: (Constant), livingArea, landValue

c. Predictors: (Constant), livingArea, landValue, bathrooms

d. Dependent Variable: price

Figure 11 Model summary with a different set of predictors using stepwise.

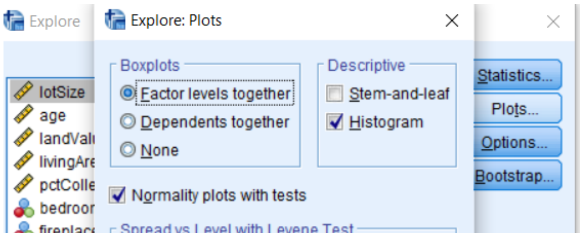
IV. Diagnostics and assumption checking

This model shows how each factor can contribute to the prediction. Multiple regression [1] Model must go through a set of eight assumptions and failing to pass a few of them is normal as we are dealing with real-world data. The researcher needs to decide as per the objective of the research whether violations need to be taken care by their remedial steps or pass.

i. Dependent variables should be measured on a continuous scale. (Price is a continuous variable)
ii. Two or more independent variables need to be there either continuous or categorical (livingArea, bedroom, livingValue, bathrooms, etc.)
iii. The data must not show multicollinearity.
iv. Independence of Observation needs to be there (Durbin-Watson statistics value should be closer to 2.)
v. Data needs to show homoscedasticity (the variances along the line of best fit remain similar as you move along the line.
vi. Multicollinearity between the independent variable should not be there.
vii. No Significant outliers
viii. Normal distribution of the data

- Check normality on the dependent variable price.

Analyze >Explore > Plots.



Boxplots

● Factor levels together
○ Dependents together
○ None

Descriptive

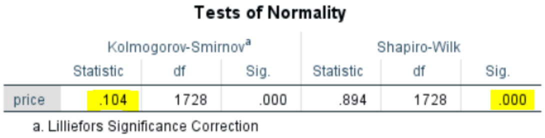
□ Stem-and-leaf
☑ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test

Figure 12 Normality plots for Sales Price

The T-test for the dependent variable shows that the Significant value is lesser than the statistics [Fig11.] Which is not normally distributed (Sig<Statistics).



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
price	.104	1728	.000	.894	1728	.000

a. Lilliefors Significance Correction

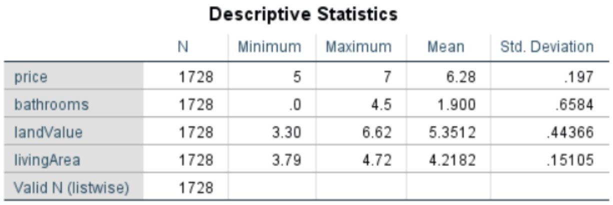
Figure 13 T-test for Sales Price

We need to check the distribution of the data across the variables for any missing value, abnormality, skewness etc. using the descriptive statistics with the help of plotting histogram.

Verify whether the sample size assumption is meet:
The Sample size needs to be calculated in multiple regression using N>50+8m

N = 1728, m =16
1728 > 50+8(16)
1728 > 338 (TRUE)

Refer to the below Descriptive statistics table for sample size Fig [14].



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
price	1728	5	7	6.28	.197
bathrooms	1728	.0	4.5	1.900	.6584
landValue	1728	3.30	6.62	5.3512	.44366
livingArea	1728	3.79	4.72	4.2182	.15105
Valid N (listwise)	1728				

Figure 14 Descriptive statistic

Verify the linear relationship between the dependent and independent variables.

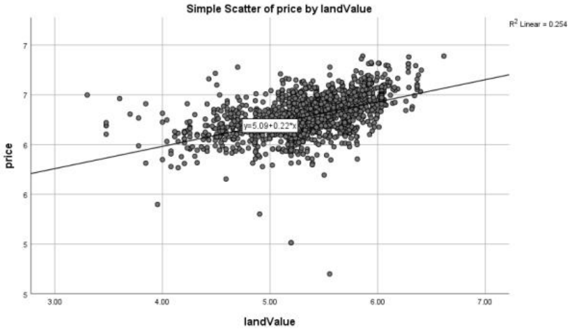


Figure 15 Sales Price has a Linear Relationship with LandValue

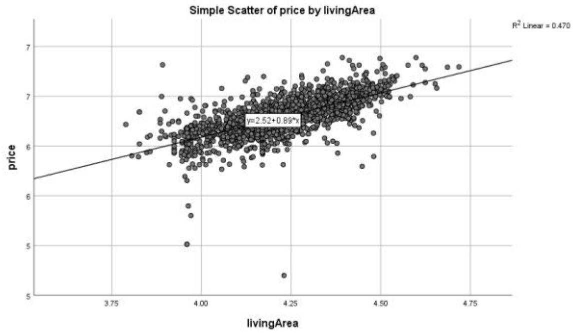


Figure 16 Sales Price has linear relationship with LivingArea

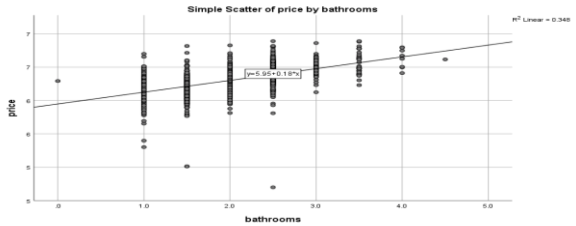
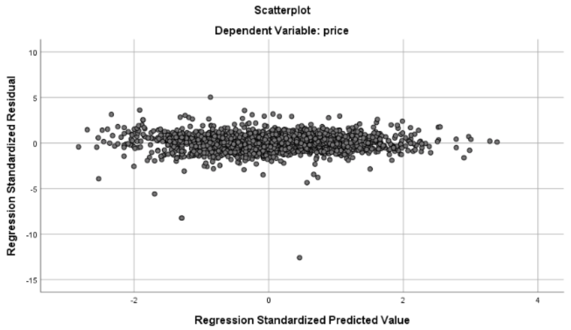


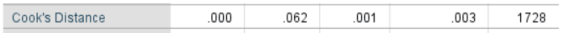
Figure 17 Linearity graph of bathrooms in Houses

Verify the Residual of observation using scatter plot.

Since the residuals are distributed around 0 within -3.3 to +3.3 there are no outliers, and no violation of linearity and homoscedasticity is present.



The cook's distance values are not more than 1,which represents the model contains no unusual cases.



Cook's Distance	.000	.062	.001	.003	1728

Figure 18 Cook's distance for model

Verify Normality

Using "The P-P plot compares the observed cumulative probability of the standardized residual to the expected cumulative probability of the normal distribution.

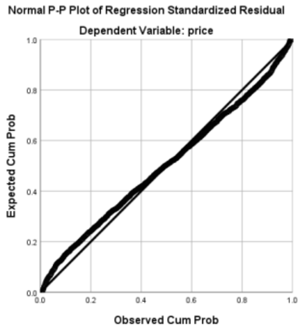


Figure 19 P-P plot for the normality of model

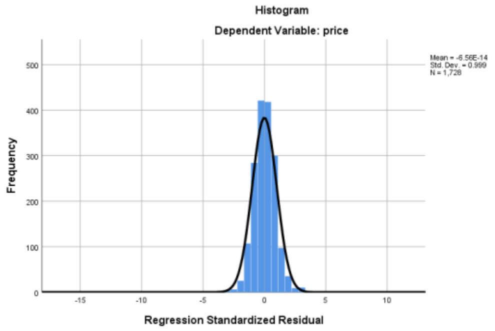
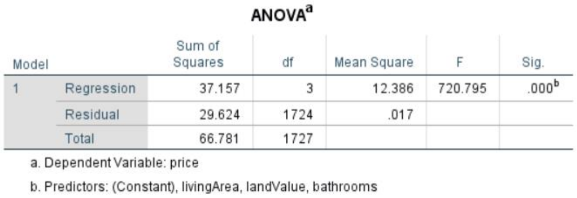


Figure 20 Histogram representation between Sales Price and independent variables

The independent variables should not be correlated.
Multicollinearity may be illustrated using:

- Correlation Matrix- The magnitude of Pearson's correlation coefficients should be less than .80.
- Variance Inflation Factor (VIF)- Value higher than 10 points indicates multicollinearity.
- Our model shows that the Sales Price has a positive correlation of .590, .504, .685 with bathroom, landValue, LivingArea, respectively Fig [20].



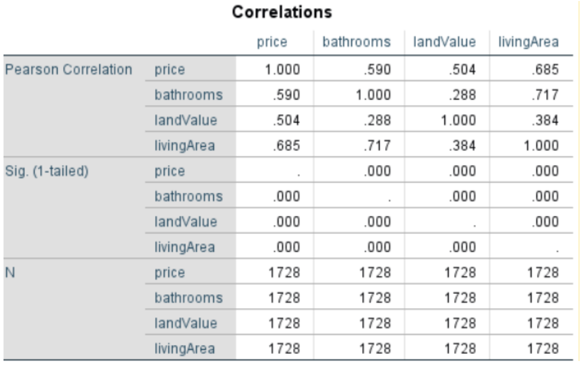
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37.157	3	12.386	720.795	.000[b]
	Residual	29.624	1724	.017		
	Total	66.781	1727			

a. Dependent Variable: price

b. Predictors: (Constant), livingArea, landValue, bathrooms

Figure 21 ANOVA

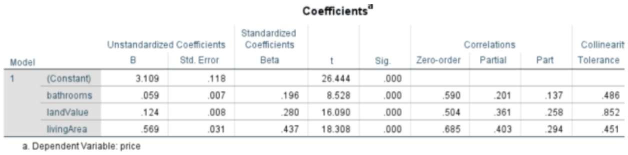


Correlations

		price	bathrooms	landValue	livingArea
Pearson Correlation	price	1.000	.590	.504	.685
	bathrooms	.590	1.000	.288	.717
	landValue	.504	.288	1.000	.384
	livingArea	.685	.717	.384	1.000
Sig. (1-tailed)	price	.	.000	.000	.000
	bathrooms	.000	.	.000	.000
	landValue	.000	.000	.	.000
	livingArea	.000	.000	.000	.
N	price	1728	1728	1728	1728
	bathrooms	1728	1728	1728	1728
	landValue	1728	1728	1728	1728
	livingArea	1728	1728	1728	1728

Figure 22 Correlation matrix between variables

It is recommended that an independent variable has a correlation value between .3 and .7 with the dependent variable. Our model satisfies that assumption as well.



Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearit
Model		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance
1	(Constant)	3.109	.118		26.444	.000				
	bathrooms	.059	.007	.196	8.528	.000	.590	.201	.137	.486
	landValue	.124	.008	.280	16.090	.000	.504	.361	.258	.852
	livingArea	.569	.031	.437	18.308	.000	.685	.403	.294	.451

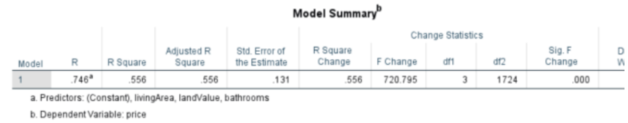
a. Dependent Variable: price

Figure 23 Coefficient matrix with Collinearity

Variance Inflation factor (VIF)is less than 10 for all the independent variables Fig [21].

The Residuals are independent.

Using regression analysis Durbin Watson test has been determined as 1.583 which is recommended to be less than 2 and indicates that there is positive autocorrelation with no violation of the assumption.



Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	D W
1	.746[a]	.556	.556	.131	.556	720.795	3	1724	.000	

a. Predictors: (Constant), livingArea, landValue, bathrooms

b. Dependent Variable: price

Figure 24 Variance Inflation factor

V. MODEL SUMMARY:

Using the Regression method stepwise model prediction option, the best fit model is created that has the highest accuracy of predicting the Sales Price of US Houses.A Model summary shows that the model is 55.6% of the variation in the Price can be considered a good predictor.

Multiple regression conducted to show the relationship between the Sales prices of Houses in US Region depends 55.6% on the Living Area, Living Value and number of bathrooms in a property in the final model Fig [23], the cook's distance value is less than 1, the VIF (Variance inflation factor) is less than 10 which shows that it is not collinear with other characteristics and the D-W statistics value is closer to 2 and prediction, p <.05.

Y = intercept + coefficient * X + error[1]

The regression equation is :

Price = 3.109 + .059(bathrooms) + .124 (landValue) + .569 (livingArea)

Price denotes Sales [rice for US houses in dollars.

VI. REFERENCES

[1] D. A. Lind, W. G. Marchal, S. A. Wathan, Basic Statistics for Business & Economics, 8th ed., United States: McGraw-Hill Education, 2012