March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

Statistical Approach for House Price Estimation by Integrating Crucial Factors using Step-wise Multiple Linear Regression(MLR) Model

Umair Mansoor[1], V. Sivakumar[2]
[1]*tp053860@mail.apu.edu.my*, [2]*dr.sivakumar@apu.edu.my*

Article Info
Volume 83
Page Number: 216 - 222
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 12 March 2020*

Abstract
Every business wants to be competitive in the market. Similar situation can observe in the housing business. This study aims to design a statistical model to estimate the price of a house by obtaining a dataset. The proposed model will be evaluated by incorporating various factors which will be proved beneficial for the housing industry in the coming future as well. If a buyer is informed with the characteristics of any home in terms of the total allocated area, making a decision for buying is still lot confusing. This process of buying certainly needs an efficient statistical approach to estimate the reasonable price of a house while integrating some of the crucial factors about it.

Keywords: *statistical model, correlation, logistic regression, Multi linearity.*

1. Introduction

In this age of digital universe shopping for any necessity has become very competitive. Every business wants to be competitive in the market. Similar situation can observe in the housing business. Purchasing of a house for an individual is a crucial task and no one wants to drain their money while making such decision. As majority of the people does not get a second chance in their lifetime to but more than one house. From this observation it is very easy to understand the significance of the decision which involves the purchasing of a house. Therefore, this study aims to design a statistical model to estimate the price of a house by obtaining a dataset. The selling price will be addressed by employing a couple of statistical models which will provide some assistance in the decision-making process for any purchaser. The proposed model will be evaluated by incorporating various factors which will be proved beneficial for the housing industry in the coming future as well.

2. Housing Price Estimation

Housing is one of the biggest industries across the world and for an individual buying a house in not as simple as it seems. It is very difficult to estimate or understand a house price even in a particular neighbourhood. If a buyer is informed with the characteristics of any home in terms of the total allocated area, making a decision for buying is still lot confusing. This process of buying certainly needs an efficient statistical approach to estimate the reasonable price of a house while integrating some of the crucial factors about it.

Following are the proposed objectives of this research study.

- To develop a regression model which can estimate the selling price of a house.
- To identify the significant attributes by developing the proposed model.

Following figure 1 is showing the proposed hypothesis for this study.

March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

Hypothesis	Ho	H1
1	Lot area does not affect the house sale price.	Lot area does affect the house sale price.
2	Basement area does not affect the house sale price.	Basement area does affect the house sale price.
3	Built Year does not affect the house sale price.	Built Year does affect the house sale price.
4	Garage area does not affect the house sale price.	Garage area does affect the house sale price.
5	Living area does not affect the house sale price	Living area does affect the house sale price.

Figure 1: Proposed Hypothesis

Experimentation and Analysis

This section of the study will elaborate the regression model development for the estimation of the sale price of houses while elaborating each of the step which was involved in this process.

Dataset

The first phase of this model development was the collection of the data and for that purpose, a dataset comprised upon various features was obtained from Kaggle data repository. The obtained dataset fulfils the required conditions of having at least seven matric independent and one matric dependent variable. The following figure 2 is representing the structure of the described dataset.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
5	Basement	Num	8	BEST.	Basement
4	Built_Year	Num	8	BEST.	Built_Year
1	Condition	Num	8	BEST.	Condition
6	First_Floor	Num	8	BEST.	First_Floor
2	Front_Road	Num	8	BEST.	Front_Road
9	Garage	Num	8	BEST.	Garage
8	Living_Area	Num	8	BEST.	Living_Area
3	Lot	Num	8	BEST.	Lot
10	Sale_Price	Num	8	BEST.	Sale_Price
7	Second_Floor	Num	8	BEST.	Second_Floor

Figure 2: Obtained Dataset

Moving forward, it can be observed in figure 3 that the obtained dataset was comprised upon 149 observations.

The CONTENTS Procedure

Data Set Name	WORK.IMPORT	Observations	149
Member Type	DATA	Variables	10
Engine	V9	Indexes	0
Created	08/04/2019 16:27:43	Observation Length	80
Last Modified	08/04/2019 16:27:43	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 3: Number of Observations in the Dataset

The next figure 4 is representing the first ten head rows of the obtained dataset. It can be clearly observed that all the attributes are of the matric data type.

Obs	Condition	Front_Road	Lot	Built_Year	Basement	First_Floor	Second_Floor	Living_Area	Garage	Sale_Price
1	5	65	8450	2003	856	856	854	1710	548	208500
2	5	68	11250	2002	920	920	866	1786	608	223500
3	5	84	14260	2000	1145	1145	1053	2198	836	250000
4	5	85	11924	2006	1175	1182	1142	2324	736	345000
5	5	101	14215	2006	1158	1158	1218	2376	853	325300
6	5	108	13418	2005	1117	1132	1320	2452	691	309000
7	5	24	2645	2000	970	983	756	1739	480	172500
8	5	68	13682	2006	1410	1426	1519	2945	841	438780
9	5	76	9591	2005	1143	1143	1330	2473	852	317000
10	5	74	10141	1998	832	885	833	1718	427	185000

Figure 4: Head Rows of the Dataset

Stepwise Multiple Linear Regression Model Development

This modelling process is comprised upon stepwise selection of independent variables for a multiple linear regression process using SAS. The first step of this stepwise MLR process can be observed in figure 5 where the first independent variable will be incorporated in the model because of its significance and correlation to the dependent variable. The first independent variable which entered in the model was "Living Area".

Stepwise Selection: Step 1

Variable Living_Area Entered: R-Square = 0.6464 and C(p) = 60.9876

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.230537E12	1.230537E12	268.75	<.0001
Error	147	6.73083E11	4578795600		
Corrected Total	148	1.90362E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-66442	19867	51213165242	11.18	0.0010
Living_Area	153.68126	9.37453	1.230537E12	268.75	<.0001

Figure 5: Stepwise Selection - Step 1

March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

The second step can be observed in figure 6 where the second independent variable will be incorporated in the model because of its significance and correlation to the dependent variable. The second independent variable which entered in the model was “Garage”.

Stepwise Selection: Step 2

Variable Garage Entered: R-Square = 0.7263 and C(p) = 16.4446

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.38262E12	6.913101E11	193.73	<.0001
Error	146	5.209994E11	3568488976		
Corrected Total	148	1.90362E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-91651	17959	92943557974	26.05	<.0001
Living_Area	95.43459	12.16949	2.194579E11	61.50	<.0001
Garage	250.04880	38.30238	1.520836E11	42.62	<.0001

Figure 6: Stepwise Selection - Step 2

The third step can be observed in figure 7 where the third independent variable will be incorporated in the model because of its significance and correlation to the dependent variable. The third independent variable which entered in the model was “Built Year”.

Stepwise Selection: Step 3

Variable Built_Year Entered: R-Square = 0.7401 and C(p) = 10.4156

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.408855E12	4.696185E11	137.63	<.0001
Error	145	4.947641E11	3412166063		
Corrected Total	148	1.90362E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2191964	757657	28559580626	8.37	0.0044
Built_Year	1067.87991	385.11850	26235311362	7.69	0.0063
Living_Area	97.38774	11.92078	2.277347E11	66.74	<.0001
Garage	195.01194	42.38826	72220643842	21.17	<.0001

Figure 7: Stepwise Selection - Step 3

The fourth step can be observed in figure 8 where the fourth independent variable will be incorporated in the model because of its significance and correlation to the dependent variable. The fourth independent variable which entered in the model was “Basement”.

Stepwise Selection: Step 4

Variable Basement Entered: R-Square = 0.7479 and C(p) = 7.8566

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.423753E12	3.559381E11	106.81	<.0001
Error	144	4.79867E11	3332409400		
Corrected Total	148	1.90362E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1960810	756689	22376610324	6.71	0.0105
Built_Year	955.61514	384.27698	20607991062	6.18	0.0140
Basement	52.81110	24.97774	14897125595	4.47	0.0362
Living_Area	67.10389	18.54553	43629015224	13.09	0.0004
Garage	192.20880	41.91091	70089139217	21.03	<.0001

Figure 8: Stepwise Selection - Step 4

The fifth step can be observed in figure 9 where the fifth independent variable will be incorporated in the model because of its significance and correlation to the dependent variable. The fifth independent variable which entered in the model was “Lot”.

Stepwise Selection: Step 5

Variable Lot Entered: R-Square = 0.7523 and C(p) = 7.3314

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.432004E12	2.864008E11	86.84	<.0001
Error	143	4.716157E11	3298012017		
Corrected Total	148	1.90362E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1987523	752963	22978899486	6.97	0.0092
Lot	1.47380	0.93176	8251235143	2.50	0.1159
Built_Year	973.60762	382.45776	21372395111	6.48	0.0120
Basement	49.17840	24.95441	12808744252	3.88	0.0507
Living_Area	56.11794	19.71361	26725334373	8.10	0.0051
Garage	194.65897	41.72281	71788347802	21.77	<.0001

Figure 9: Stepwise Selection - Step 5

Interpretation of MLR Model

The following figure 10 is showing the obtained parameters of stepwise MLR model in which total

March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

of five independent variables were selected and the total of 75.23% of variance were explained by this model for the dependent variable as the value of R-Square is 0.7523.

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

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Living_Area		Living_Area	1	0.6464	0.6464	60.9876	268.75	<.0001
2	Garage		Garage	2	0.0799	0.7263	18.4446	42.62	<.0001
3	Built_Year		Built_Year	3	0.0138	0.7401	10.4156	7.69	0.0063
4	Basement		Basement	4	0.0078	0.7479	7.8566	4.47	0.0362
5	Lot		Lot	5	0.0043	0.7523	7.3314	2.50	0.1159

Figure 10: MLR Model Results

Assessment of MLR Model Assumptions

As there are four assumptions linearity, normality, homoscedasticity and multicollinearity for any MLR model and each of those assumptions are elaborated in the subsequent study.

Linearity

The following figure 11 is showing the residual plots for each of the significant independent variable. As it can be observed that a couple of the plots are showing a linearity while the rest of the attributes have a nonlinear relation.

Figure 11: Linearity Assumption

Normality

For the developed model, normality assumption was assessed, and four different tests were performed. From figure 12 it can be seen that the model is fulfilling the assumption of the normality successfully.

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.717961	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.157305	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.4205	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	8.290414	Pr > A-Sq	<0.0050

Figure 12: Normality Assumption

Moreover, the normality assumption can also be evaluated by the Q-Q normality plot which has been showed in the mentioned below figure 13.

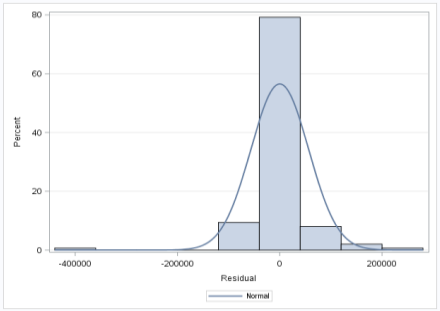

Figure 13: Normality Curve

Homoscedasticity

For the of the homoscedasticity assumption assessment following scatter plot between residual and predicted values were obtained and it can be observed in figure 14. From this plot it can be seen that most of the behavior of the model was homoscedasticity while there is a presence of some outliers in the obtained data.

March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

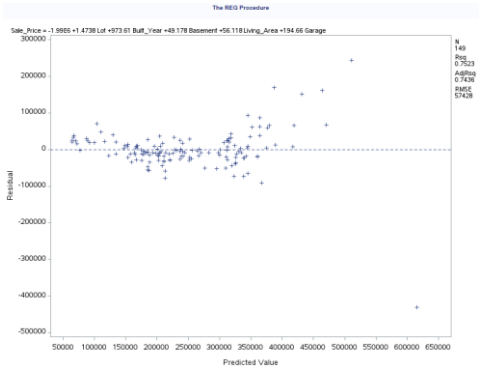

Figure 14: Homoscedasticity Assessment

Multicollinearity

The multicollinearity assumption can be assessed from the variance inflation parameter. All those attributes will be non-multicollinear which have the value of more than 5 for the variance inflation parameter and all this can be observed from figure 15.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-2015023	761412	-2.65	0.0091	0
Condition	Condition	1	12428	7917.83062	1.57	0.1188	1.33761
Front_Road	Front_Road	1	-116.19290	242.43070	-0.48	0.6325	2.04199
Lot	Lot	1	1.20567	0.94722	1.27	0.2052	2.07376
Built_Year	Built_Year	1	951.31397	386.31880	2.46	0.0150	1.59230
Basement	Basement	1	99.99878	37.95962	2.63	0.0094	10.90713
First_Floor	First_Floor	B	4.74123	38.30058	0.12	0.9017	10.70477
Second_Floor	Second_Floor	B	77.41417	24.43965	3.17	0.0019	2.02834
Living_Area	Living_Area	0	0	.	.	.	.
Garage	Garage	1	197.75627	44.97219	4.40	<.0001	3.25541

Figure 15: Multicollinearity Assumption

Model Significance

The following section of the study will elaborate whether the developed MLR model is significant or adequate while elaborating the obtained parameters mentioned in figure 16.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1.446157E12	1.807696E11	55.32	<.0001
Error	140	4.574625E11	3267589241		
Corrected Total	148	1.90362E12			

Root MSE	57163	R-Square	0.7597
Dependent Mean	246307	Adj R-Sq	0.7460
Coeff Var	23.20797		

Figure 16: Model Assessment

The first permanent which can be used to determine the significance of the developed MLR model is the p-value. If that p-value is equal to 0.05 then the model is considered as significance. Therefore, this model will be considered as significant as mentioned in figure 16 the p-value is <0.0001. While the other parameter which is R-Square represents the amount of variance explained by the incorporated independent variables in the model which is in this case is about 75.9 percent which also enforces the significance of this model.

Hypothesis Testing

In this section of the study, the proposed hypothesis will be tested on the basis of their significance to the model and dependent variable selling price of the house.

- **H1:** The first hypothesis will be rejected as the lot area does affect the sale price of the house.
- **H2:** The second hypothesis will be rejected as the basement area does affect the sale price of the house.
- **H3:** The third hypothesis will be rejected as the built year does affect the house sale price
- **H4:** The 4th hypothesis will be rejected as the garage area does affect the the house sale price
- **H5:** The 5th hypothesis will be rejected as the living area does affect the house sale price'

3. Factor analysis

Purpose of Factor Analysis

The most appropriate and substantial method to reduce the dimensionality of a dataset or a problem is achieved by the employment of factor analysis. By reviewing the correlation among various attributes, some comprehensive factors are formed by combining them. Therefore, various number of variables will be grouped into some factors which are explaining the same variance of the targeted variable.

March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

Exclusion of Non-matric Independent Variables from Factor Analysis

The factor analysis basic functionality is to calculate the correlation of independent variables with each other and then grouping them in a same factor which is only possible if the variable is matric as for non-matric variables correlation calculation is not possible.

Performing Factor Analysis

The following section of the study is illustrating the results of the factor analysis. Figure 17 is representing the eigenvalues for all the attributes. The cutoff eigenvalue is set to the default value of 1.

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.22824305	4.05046829	0.5809	0.5809
2	1.17777476	0.39865459	0.1309	0.7118
3	0.77912017	0.12515293	0.0866	0.7983
4	0.65396724	0.14882994	0.0727	0.8710
5	0.50513730	0.14163204	0.0561	0.9271
6	0.36350526	0.12734156	0.0404	0.9675
7	0.23616370	0.18007517	0.0262	0.9938
8	0.05608853	0.05608853	0.0062	1.0000
9	0.00000000		0.0000	1.0000

Figure. 17: Eigenvalues

Moreover, the next figure 2 is representing the factor pattern, variance explained by each factor and final communality estimates of the factor analysis. We can observe in figure 18 that there are two factors formed by the factor analysis on the provided dataset. The variance explained by the first factor is 5.22 while 1.77 by the second factor.

Factor Pattern

		Factor1	Factor2
Condition	Condition	-0.17086	0.83534
Front_Road	Front_Road	0.71969	0.26032
Lot	Lot	0.74708	0.30183
Built_Year	Built_Year	0.51698	-0.48132
Basement	Basement	0.90021	-0.05463
First_Floor	First_Floor	0.90430	0.10750
Second_Floor	Second_Floor	0.77133	0.05598
Living_Area	Living_Area	0.96772	0.09855
Garage	Garage	0.83434	-0.24912

Variance Explained by Each Factor

Factor1	Factor2
5.2282431	1.1777748

Final Communality Estimates: Total = 6.406018

Condition	Front_Road	Lot	Built_Year	Basement	First_Floor	Second_Floor	Living_Area	Garage
0.72697976	0.58571551	0.64923547	0.49894084	0.81337015	0.82930946	0.59807995	0.94619771	0.75818897

Figure. 18: Factor Pattern and Explained Variance

Grouping in Factor Analysis

The following figure 19 shows the orthogonal transformation matrix, the result of the employment of the grouping of the factor. As the eigenvalue is set 1 as default, the rotation method known as VERIMAX was applied to achieve the rotated matrix.

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2
1	0.90417	0.42718
2	-0.42718	0.90417

Figure. 19: Rotation Method Employment

4. Eigenvalue of the Obtained Factors

Calculation of Eigenvalue

The first two factors` eigenvalue has been calculated as below.

Factor 1: $(-0.17086^{2)} + (0.71696^2) + (-0.74708^2) + (0.51698^2) + (0.90021^2) + (0.90430^2) + (0.77133^2) + (0.86772^2) + (0.83434^2) = 5.2282431$

5.2282/100 = 0.52282 = 52.28%. This means each factor in group 1 have 52.28% of information can explain the factor 1.

Factor 2: $(0.83534^2) + (0.26032^2) + (0.030183^2) + (-0.48132^2) + (-0.05463^2) + (0.10750^2) + (0.05598^2) + (0.09855^2) + (0.24912^2) = 1.1777748$

1.1777748/100 = 11.77%. This means each factor in group 2 have 11.77% of information can explain the factor 2.

Total amount of factor 1 and factor 2 :(5.2282+1.1777) = 6.4059

March - April 2020
ISSN: 0193-4120 Page No. 216 - 222

Interpretation of Eigenvalue

Based on calculate the eigenvalue for the first 2 factors. It is indicating the relative important for first factor in secretarial for variance associated which is 5.2282 with the variables set. The important for second factor in accounting for variance associated which is 1.1777 with the set of variables. The factor one accounting for the most variance. Compare to the factor 2. Factor 2 is slightly less than factor 1. The sum of factor one and factor two is 6.4059, it indicates the total of two eigenvalues. by the factor solution it is represent the total amount of variance extracted.

5. Factorability, Factor Cross-Loading and its Resolution

a. Factorability of the Dataset

To testing factorability of dataset can improved prior to factor analysis have many methods. The one use for this dataset is that to use Bartlett's and KMO table. Usually if the KMO more than 0.6, means the factor analysis of the dataset can accept and adopt to do factor analysis. Besides that, from correlation matrix also can testing the factorability. If the overall factorability as well as measure of sampling adequacy is more than 0.5, which means is the factorable between the variables which is existence correlation among variables if less than 0.5, means there are not factorable between variable which means no correlation among variables.

b. Factor Cross-Loading

Cross loading occurs when an attribute appears in more than one factor with a similar amount of correlation. In this study, no factor cross loading appears in the factor analysis.

c. Solution for Factor Cross-Loading

The most fundamental method to avoid factor cross loading is by the implementation of the rotation method. Cross loading can also be avoided by removing such factors as well.

6. Conclusion and Recommendation

It is observed from this stepwise MLR model development that the obtained results are very much satisfactory. By the employment of the stepwise selection process of the attributes, it was found that five of the independent variables were significant to the and that made five of the hypotheses rejected in this study. Furthermore, the model assumption was also performed where all of the four assumptions which were linearity, normality, homoscedasticity and multi collinearity were assessed. In terms of the recommendation, the obtained data was just comprised upon 150 observations and only 9 attributes. For more comprehensive study on the housing prices assessment, larger sample size is required with various other attributes as well.

7. References

[1] Kaggle.com. (2019). *House Prices: Advanced Regression Techniques | Kaggle*. [Online] Available at: https://www.kaggle.com/c/house-prices-advanced-regression-techniques [Accessed 5 Aug. 2019].

[2] Medium. (2019). *5 Data Science Principles (for predicting house prices)—#100DaysOfMLCode*. [Online] Available at: https://medium.com/datadriveninvestor/5-data-science-principles-for-predicting-house-prices-100daysofmlcode-84bb47d7a99a [Accessed 5 Aug. 2019].

[3] Medium. (2018). *House hunting — the data scientist way*. [Online] Available at: https://medium.com/geoai/house-hunting-the-data-scientist-way-b32d93f5a42f [Accessed 5 Aug. 2019].