How to predict the selling price of a house

Executive Summary An analysis was done to find an equation that predicts the selling price of a house. The data used in this research analysis to predict the selling price of a house is shown in the Bryant/Smith Case 28 (See Appendix 1). The null hypothesis stated that there is not a relationship between the selling price of a house and its characteristics.

The alternate hypothesis stated that there is a relationship between the selling price of a house and its characteristics. A 95% confidence level was chosen and a prediction interval which is a confidence interval estimate of a predicted value of the selling price used. The MegaStat output of a Regression Analysis of the Bryant/Smith Case 28 data was used as the basis to calculate the multiple regression equation as the prediction point. The point prediction of the selling price of a house corresponding to the combination of values of the independent variables is; Y = -12. 5988 + 0. 0383(X1) + 4.

3573(X2) -14. 5371(X3) + 16. 0610(X4) + 11. 3576(X5) – 1. 2168(X6) given on the MegaStat output. The MegaStat output tells us that the p-value associated with the variables (Square Foot, Garage, Basement and Age) are less than 0.

1 level of confidence, therefore we have very strong evidence that these variables are significantly related to the selling price and thus, are very important in this model. Also, since the p-value associated with Bed and Heat was 0. 0248 and 0. 0199 respectively, we have a close to strong evidence that they are important. The results from the data calculation indicated that the null hypothesis should be rejected and the alternate hypothesis should

be accepted. The purpose of this research is to find an equation that predicts the selling price of a house.

Developments in housing prices are of great interest to householders, policy-makers and those involved in the housing industry. This has been the case in a number of countries where house price developments are having significant macroeconomic impacts. However, the construction of aggregate measures of housing prices is not a straightforward exercise, and involves addressing a number of conceptual and practical issues. This paper aims to provide a computationally simple method of addressing some of these issues. While the focus of this paper is on predicting the selling price of a house in Eastville, Oregon, the method outlined in this paper would also be feasible and readily adaptable for data from other areas or countries.

One major problem in measuring housing price growth results from the infrequency of transactions and the heterogeneous nature of the housing stock. To be meaningful, price data should be based on transactions prices rather than valuations. One of most important things you need to know when selling a house is the maximum you should pay for a property so that you can make your desired profit. The key to determining your maximum cash offer is knowing how to predict the value without relying on Realtors. There are many different house price indexes that can be obtained to get the latest information on property prices and the patterns and trends of growth. In essence, there are so many different guides with so much differing information that it becomes almost an impossible task to know which one you can trust to be accurate.

A perfect house price would only report on the actual purchase price of every completed property. To further enhance this information the type of property and any seasonal adjustments should be included. At this time this information is not obtainable and such an index does not exist and coupled with the effects of short term house price inflation or property price volatility, the house price index becomes a very complex equation. There are many problems with predicting house prices due to the nature of the market where no sale is the same and a house that is identical to another can sell for a different price for any number of reasons. This could be due to the location or condition of the property, the extra work that had been done on the interior or an extension that had been built.

Even houses on the same road and in identical condition can sell for different prices due to the negotiations that take place between buyer and seller on the price and the property valuations that vary from agent to agent. In a case study presented by Cain and Janssen, their approach was illustrated with an actual situation in the pricing of townhouses. According to Cain and Janssen " A serviced parcel of land in a suburban location is proposed for a 16-unit townhouse development. The developer has acquired the land, prepared plans and drawings, secured zoning approval, applied for bank financing and commenced excavation. Construction is set to begin. A major decision is the setting of prices for the units.

This is of critical importance to the success of the project. In particular, if the prices are perceived by the market as too high, sales will be slow, and profits eroded. If on the other hand prices are too low, profits may be foregone.

Overpricing, however, is more serious than under pricing in the present instance".

One key starting point for a House Price Index is a fixed housing stock. However, the quality of the housing stock is likely to rise as a result of newly built homes. This, in turn, causes the mean or median price to continue to rise even when individual properties are not appreciating (Bailey et al, 1963). For example, if a disproportionate number of high priced homes were sold in a given period, the mean or median price would rise, even though not a single house had increased in value (Case and Shiller, 1987).

If the underlying principle of random selection does not hold and different types of houses or different regions show different price changes, a bias will occur in the sample. These shortcomings led to the development of alternative methods such as the simple Multiple Regression analysis. The main question of interest is how to predict the selling price of a house The Multiple Regression analysis is used in this research to predict the selling price of a house using one variable (the dependent variable) on the basis of other variables (the independent variables). The main idea of a multiple regression analysis is to understand the relationship between the independent variables and the dependent variable.

A model of the relationship is hypothesized as; The Null Hypotheses – there is not a relationship between the selling price of a house and its characteristics. The Alternate Hypotheses – There is a relationship between the selling price of a house and its characteristics. The estimates of the parameter values are used to develop an estimated regression equation. The multiple regression equation used to describe the relationship is: Y = B0 + B1X1+ B2 X 2+ B3 X 3+ B4 X 4+ B5 X 5+ B6 X 6+…….

BkXk. It is used to estimate Y given selected B values and X independent variables. The data used in this research analysis to predict the selling price of a house is shown in the Bryant/Smith Case 28 (See Appendix 1). The objective of the analysis is to predict the selling price of a house using the characteristics recorded in Case 28 data sheet.

For each house the following variables will be measured: •Square Feet – Total square feet in the house •Beds – Number of Bedrooms •Baths – Number of bathrooms •Heat – 0= gas forced air heating; 1= electric heating Style – the architectural style of the home: 0= trilevel, 1= two-story, 2= ranched styled •Garage – Number of cars that can fit in garage •Age – Age of homes in years •Fire – 0= no fireplace present, 1= at least 1 fire place present •Basement – 0= no basement, 1= basement •Price – selling price in thousands: 0= Eastville District School; 1= Apple Valley school •School – district There are two components of a multiple regression analysis that are important; the coefficient of correlation (multiple R) and coefficient of determination (R?). The coefficient of correlation is the degree to which two or more independent variables are related to the dependent variable. The coefficient of determination is " the proportion of the total variation in the dependent variable that is explained by the independent variable. That is how well the multiple regression equation fits the sample data.

A perfect fit would result in R? = 1. A very good fit results in a value near 1. A poor fit results in a value of R? close to 0. The selling price of a house depends on the variables or characteristics. Furthermore, people will pay a fixed premium if the house is on a desirable location.

Cost of housing is rising rapidly and prospective home owners need to be able to predict possible housing costs based on given criteria in order to plan for financing. Analysis and Methods Section The method chosen for calculating the selling price of a house is to make a regression estimate of the selling price of a house on the basis of ten characteristics. The MegaStat regression analysis and data output in Tables 1 and 2 shows the level of significance in predicting the selling price of a house. Table 1 – MegaStat Regression Analysis Regression Analysis R? 0. 826Adjusted R? 0. 808 n 108 R 0.

909 k 10 Std. Error 11. 594 Dep. Var.

PRICE ANOVA table SourceSS df MSFp-value Regression 61, 703. 8105 10 6, 170. 3811 45. 912. 21E-32 Residual 13, 037. 8312 97 134.

4106 Total 74, 741. 6417 107 From the output we see that r2 = 0. 826, which means that 82. 6% of the variation of y can be explained using the ten independent variables (Square Foot, Beds, Baths, Heat, Style, Garage, Basement, Age, Fire and School). The MegaStat display shows the F-value of 45.

91 with a corresponding p-value of 2. 21E-32; p-value is below 0. 05, 0. 01 and 0.

01 levels of significance. In this case the small p-value of 2. 21E-32 suggests that the multiple regression equation has a good overall significance and is

usable for prediction. That is, it makes sense to predict the selling price of a house based on the independent variables; Square Foot, Beds, Heat, Garage, Basement and Age. Table 2 – MegaStat Regression Output Regression outputconfidence interval variables coefficientsstd.

error t (df= 97)p-value95% lower95% upper Intercept-15. 2124 9. 8179 -1. 549 . 1245-34. 6982 4.

2734 SQ_FT0. 0376 0. 0036 10. 365 2.

19E-170. 0304 0. 448 BEDS4. 9237 1.

9647 2. 506 . 01391. 0244 8. 8231 BATHS-2.

9115 3. 0240 -0. 963 . 3380-8. 9132 3. 0902 HEAT-12.

9097 6. 1009 -2. 116 . 0369-25. 0184 -0.

8010 STYLE2. 2877 1. 6437 1. 392 . 1672-0.

9746 5. 5501 GARAGE15. 7593 3. 8246 4. 121 . 00018.

1686 23. 3501 BASEMENT9. 0772 3. 4454 2. 635 .

00982. 2390 15. 9154 AGE-1. 0342 0. 2813 -3. 676 .

0004-1. 5925 -0. 4758 FIRE5. 3054 3. 9794 1. 333 .

1856-2. 5927 13. 2035 SCHOOL4. 6217 2.

5341 1. 824 . 0713-0. 4079 9. 6513 The p-value of 2. 21E-32 results from the test of the null hypothesis that Bath= Style= Fire= School= 0.

Rejection of Bath= Style= Fire= School = 0 implies that at least one of the variables (Bath, Style, Fire and School) is not = 0 suggesting that this regression equation is effective in determining the selling price of a house. Looking at the MegaStat output in Table 2, the p-values of 2. 19E-17, 0. 0001, 0. 0098, and 0. 0004 are shaded dark yellow.

This indicates that we can reject the null hypothesis at the 0. 01 level of significance. We can conclude that Square Footage, Garage, Basement, and Age contribute significantly to the prediction of the selling price of a house. The p- values of 0. 0139 and 0.

369 shaded light yellow also indicates that the null hypothesis can be rejected at 0. 05 level of significance; which can be concluded that we have a close to strong evidence that Beds and Heat contribute to the selling price of a house. The p-values 0. 1672, 0. 1856 and .

0713 are greater than 0. 05 level of significance. Since the p-values are greater than 0. 05 we do not reject the null hypothesis and therefore it has no effect on the selling price of a house; which means that Bath, Style, Fire and School can be ignored in the final prediction equation, leaving Square Footage, Garage, Basement, Age, Beds and Heat.

The prediction equation is Y = B0 + B1X1+ B2 X 2+ B3 X 3+ B4 X 4+ B5 X 5+ B6 X 6 where: B0 is the Y-intercept or the value of Y when all the predictor variables are 0. B1, B2 …. Bk are the coefficients of the independent variables x1, x2, ….. xk.

X1, X2, ….. Xk are the independent variables. The following independent variables are: X1 = Square Foot, X2 = Beds, X3 = Heat, X4 = Garage, X5 = Basement and X6 = Age Therefore prediction equation Y = -15.

2124 + 0. 0376(X1) + 4. 9237(X2) -12. 9097(X3) + 15.

75793(X4) + 9. 0772(X5) – 1. 0342(X6). A new MegaStat analysis was performed, dropping out the independent variables Bath, Style, Fire and School and the outputs are shown in Tables 3 and 4. Using the " new MegaStat output for the independent variables Square Footage, Garage, Basement, Age, Beds and Heat we see that the r2 = 0. 808, which means that 80.

8% variation in Y can be explained using the six independent variables and the new equation developed. This means that the regression model for Y can explain 80. 8% of the Y-values given. Table 3 – MegaStat Regression Analysis Regression Analysis R? 0.

808Adjusted R? 0. 797 n 108 R 0. 899 k 6 Std. Error 11. 922 Dep. Var.

PRICE ANOVA table SourceSS df MSFp-value Regression 60, 387. 1261 6 10, 064. 5210 70. 825. 79E-34 Residual 14, 354.

5156 101 142. 1239 Total 74, 741. 6417 107 Again we can see in MegaStat Regression Output in Table 4 that the p-values of the independent variables Square Footage, Garage, Basement, and Age are less than 0. 01 level of significance and they contribute significantly to regression model for the selling price of a house and should be kept. Likewise, Beds and Heat are less than 0. 5 levels of significance and contribute as well to the regression model.

Table 4 – MegaStat Regression Output Regression outputconfidence interval variables coefficientsstd. error t (df= 101)p-value95% lower95% upper Intercept-12. 5988 9. 4172 -1. 338 .

1839-31. 2799 6. 0823 SQ_FT0. 0383 0.

0032 11. 976 4. 25E-210. 0319 0.

0446 BEDS4. 3573 1. 9124 2. 278 .

02480. 5636 8. 1510 HEAT-14. 5371 6. 1467 -2. 365 .

0199-26. 7305 -2. 3437 GARAGE16. 0610 3. 9271 4.

090 . 00018. 2706 23. 8513 BASEMENT11. 3576 3.

2806 3. 462 . 00084. 8497 17.

8655 AGE-1. 2168 0. 2810 -4. 331 3. 1E-05-1.

7742 -0. 6595 The new prediction equation is Y = -12. 5988 + 0. 0383(X1) + 4. 3573(X2) -14.

5371(X3) + 16. 0610(X4) + 11. 3576(X5) – 1. 2168(X6). Therefore a typical example with the use of the equation to predict the price for a house using predictor values in data from Bryant/Smith Case 28 in Appendix I; the following independent variables are used: Square footage – 1238, Beds – 3, Heat – 0, Garage – 1, Basement – 1 and Age – 12. Y = -12.

5988 + 0. 0383(1238) + 4. 3573(3) -14. 5371(0) + 16.

0610(1) + 11. 3576(1) – 1. 2168(12) Y = -12. 5988 + 47.

4154 + 13. 0719 – 0 + 16. 610 + 11. 3576 – 14. 6016 = 60.

7 Predicted selling price = 60. 7 In order to predict the selling price of a house the data shown in the Bryant/Smith Case 28 (See Appendix 1) was used to develop two regression analysis using MegaStat. In the first analysis the p-value results from the MegaStat output (Table 1) was used to investigate whether the variables were significant in predicting the selling price of a house and the regression model used to predict the results. The null hypothesis (Ho) would be rejected in favor of the alternate hypothesis (Ha) at the level of significance ? f the following equivalent condition holds: P-value < ? (? = 0.

05 level of significance) Using Table 2 – MegaStat Regression Output, four independent variables; Bath, Style, Fire and School were rejected as they had no effect on the dependent variable (Y). A second regression analysis was performed weeding out the independent variables that did not have an effect on the dependent variables. See outputs in Tables 3 & 4. The independent variables – Square Foot, Bed, Heat, Garage, Basement and Age in the regression analysis data from the MegaStat output with the p-values less than 0.

5. Using the results of the data in attempting to find the best equation for predicting the price of a house was the basic prediction equation Y = B0 + B1x1+ B2 x 2+ B3 x 3+ B4 x 4+ B5 x 5+ B6 x 6. Conclusions and Summary

Section The analysis revealed that the selling price of a house is based on its predictor variables such as the Square footage, Beds, Heat, Garage, Basement and Age. The p-value of 5. 79E-34 results also revealed that the null hypothesis is rejected and suggest that multiple regression equation Y = B0 + B1x1+ B2 x 2+ B3 x 3+ B4 x 4+ B5 x 5+ B6 x 6 is very effective in etermining the selling price of a house. It also reveals that 80.

8% variation in Y can be explained using the six independent variables and the multiple regression equation for Y can explain 80. 8% of the Y-values given. Overall, the best equation for predicting the selling price a house using the six independent variables (Square footage , Beds, Heat, Garage, Basement and Age) is Y = -12. 5988 + 0.

0383(X1) + 4. 3573(X2) -14. 5371(X3) + 16. 0610(X4) + 11.

3576(X5) – 1. 2168(X6). The analysis was done to find an equation to predict the selling price of a house. This paper looked at a way of predicting the selling price of a house using the multiple regression analysis.

This was to understand the relationship between the independent variables (Square Foot, Beds, Baths, Heat, Style, Garage, Basement, Age, Fire and School) and the dependent variable (Price). That is, the Null Hypotheses (there is not a relationship between the selling price of a house and its characteristics) and the Alternate Hypotheses (There is a relationship between the selling price of a house and its characteristics). It was discovered from the data given in Bryant/Smith Case 28, predicting the selling price of a house in Eastville, Oregon; Square Foot, Beds, Heat, Garage, Basement, Age and Price are the key characteristics used in predicting the equation Y = -12. 5988 + 0.

0383(X1) + 4. 3573(X2) -14. 5371(X3) + 16. 0610(X4) + 11.

3576(X5) – 1. 2168(X6). I think that the weeding out of independent variables; Bath, Style, Fire and School was a surprise in that the number of baths and the style of a house are key characteristics in the purchasing of a home (This is from a Jamaican context). The main question of interest was how to predict the selling price of a house. The Multiple Regression analysis was used to predict the selling price of a house using the dependent variable on the basis of other independent variables using data from the MegaStat regression analysis. Two MegaStat regressions were done.

The p-values and F-values results from the MegaStat outputs were used to investigate whether the variables were significant in predicting the selling price of a house and the regression model used to predict the results. Six independent values (Square Foot, Bed, Heat, Garage, Basement and Age) with p-values less than 0. 1 and 0. 5 levels of significance reject the null hypothesis.

It was concluded that Square Foot, Garage, Basement and Age contribute significantly to the selling price of a house. Also, it was concluded that Heat and Bed has a close to strong evidence contributing to the selling price of a house. The first MegaStat regression analysis weeds out the independent variables which has no effect on the selling price of the house. The best equation for predicting the selling price a house using the six independent variables (Square footage , Beds, Heat, Garage, Basement and Age) is Y = - 12. 5988 + 0. 0383(X1) + 4.

573(X2) -14. 5371(X3) + 16. 0610(X4) + 11. 3576(X5) – 1. 2168(X6) which can explain 80.

8% of the Y-values given. Therefore I think my question was answered. The results were as expected. Since the results of the research were significant and the functional forms of the relationships were clearly identified, the results could be adapted for predicting the selling price of homes.

References Bailey, Martin J. , Richard F.

Muth and Hugh O. Nourse, “ A Regression Method for Real Estate Price Index Construction”, Journal of the American Statistical Association 58 (1963): 933#150; 42. Case, Karl E. , and Robert J. Shiller. Prices of single-family homes since 1970: New indexes for four cities”, New England Economic Review (September/October 1987), Pp.

45-56. Extract from Article on “ An application of decision analysis to the estimation of the market price of individual townhouse units where there is asymmetry in the loss function” from http://web. ebscohost. com/ehost/pdf?vid= 5&hid= 105&sid= af4496ea-7f04-49f0-a1dc-e1898233add5%40sessionmgr107 Appendix 1 Bryant/Smith manual Case 28 SQ_FTBEDSBATHSHEATSTYLEGARAGEBASEMENTAGEFIREPRICESCHOOL 123832001112159. 91 1707321020131640 129642002117066. 50 13203200211166.

50 12103200106166. 90 1296320021171680 76532002120068. 50 1725430021120690 179442002118070. 950 1294320020131710 13723200219072.

6921 11623200108172. 8010 199642002013175. 2071 1764421021131760 141632002080761 173042002015177. 50 13923200218179. 91 166433002011179.

90 1332320021140811 175233002018182. 80 216733102113184. 90 166432002091850 1973430020131860 13843200205189. 280 14313200217189. 91 1950530020131900 145232002141920 182932002010192.

4391 16524300217194. 6461 151632002110197. 2931 199843002117198. 11 19844300219198. 1490 1840330021911050 1823430021311101 2150530021121111.

90 2096330021911131 2212430031171124 2375430021111331 2809430031611950 12420121191590 81632012119061. 50 100852012117063. 50 91242012116166. 951 1008520121171680 83842011111068. 6941 1008430121191690 112032012111170. 4521 1242430121120750 91242012118176.

91 131632012181830 1490320121131880 111843012116188. 8791 127832012111189. 3470 149032012110189. 71 1418420121141901 125043012117190. 81 18814301219191.

51 1355320121719551 1518320121220981 160643012113199. 50 21243201211411070 2099430121131113. 751 2232530121911251

16085301311311321 1320320221131761 16803302207179. 91 143433022112179.

91 166432022171851 13793202218187. 1581 163433022110187. 21 14003302201191. 51 166444022116191. 51 15483302218192.

1 204043022115193. 50 179243022191941 186533022112196. 70 18333302217197. 41 16273302213198.

51 2198430221121990 1707330221121990 209633022171102. 41 20044302211311040 18184302211911071 1935330221111107. 91 23813302201711091 1977330221911101 231753022191110. 51 20724302211111121 210833022171112.

90 206044022191113. 61 206433022151114. 51 214854022181114. 71 21706402212111151 19844312211111211 21304302212011211 20494302211911251 2035330231611281 2264330231711301 22643302311011311 2362430221611320 2262530221121132. 51 235343022151132. 50 229833023121144.

51 228243022191150. 5771 254243022161153. 81 2380430221611551