Prabhakar kumar jha

Housing Price Prediction Model

Summary

The purpose of this study is to predict residential home sales prices in a midwestern city from various characteristics of the home and surrounding property. From 522 transactions, a random sample of 300 is taken and used for the analysis. Tests are run on 13 characteristics to determine possible correlation between sale prices and these variables. The study shows that sale prices are not independent of these variables and that there is correlation between sale prices and many of the variables.

**Introduction:**

The city tax assessor in a midwestern city was interested in predicting residential home sales prices as a function of various characteristics of the home and surrounding property. To conduct this study, data on 522 arms-length transactions were obtained for home sales in 2002.

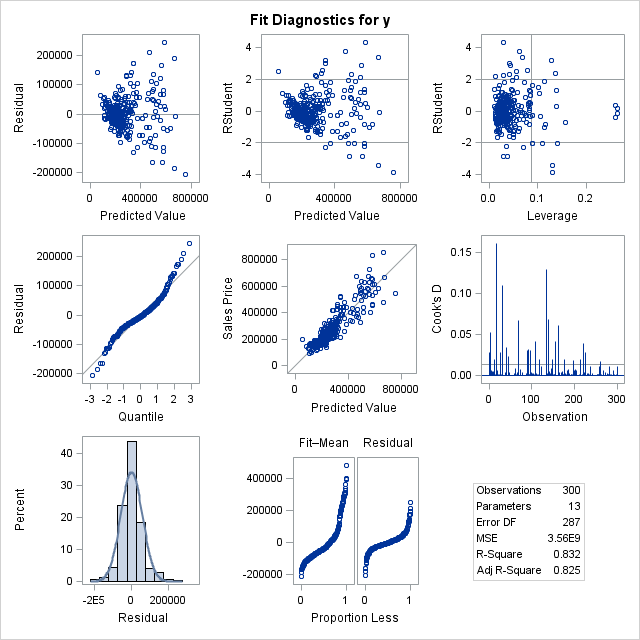
**Study Design and Procedure:**

The purpose of this study is to make predictions on the sale prices of homes based on the demographics of several variables. A random sample of 300 from 522 observations has been used for this analysis, and the study begins with testing on 13 possible predictor variables, including total square footage, number of bedrooms and bathrooms, the presence or absence of air conditioning or a pool, garage and lot size, year built, quality (high (1) or medium (2)), and whether or not the property is adjacent to a highway.

**Descriptive Statistics:**

For the Original Regression Model:

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | **1** | -1868718 | 550222 | -3.40 | 0.0008 |
| **x1** | Finished Square Feet | **1** | 104.30347 | 10.30501 | 10.12 | <.0001 |
| **x2** | Number of Bedrooms | **1** | -6781.42809 | 4353.63919 | -1.56 | 0.1204 |
| **x3** | Number of Bathrooms | **1** | 7274.05050 | 5339.79084 | 1.36 | 0.1742 |
| **x4** | Air Conditioning | **1** | -5861.27081 | 11064 | -0.53 | 0.5967 |
| **x5** | Garage Size | **1** | 15112 | 7326.48943 | 2.06 | 0.0401 |
| **x6** | Pool | **1** | 22732 | 15514 | 1.47 | 0.1439 |
| **x7** | Year Built | **1** | 1015.90095 | 279.64183 | 3.63 | 0.0003 |
| **x8** | Quality Indicator 1 | **1** | -147007 | 19385 | -7.58 | <.0001 |
| **x9** | Quality Indicator 2 | **1** | -130951 | 14355 | -9.12 | <.0001 |
| **x10** | Style | **1** | -7818.63993 | 1824.75263 | -4.28 | <.0001 |
| **x11** | Lot Size | **1** | 1.33156 | 0.31614 | 4.21 | <.0001 |
| **x12** | Adjacent to Highway | **1** | -25746 | 30367 | -0.85 | 0.3972 |

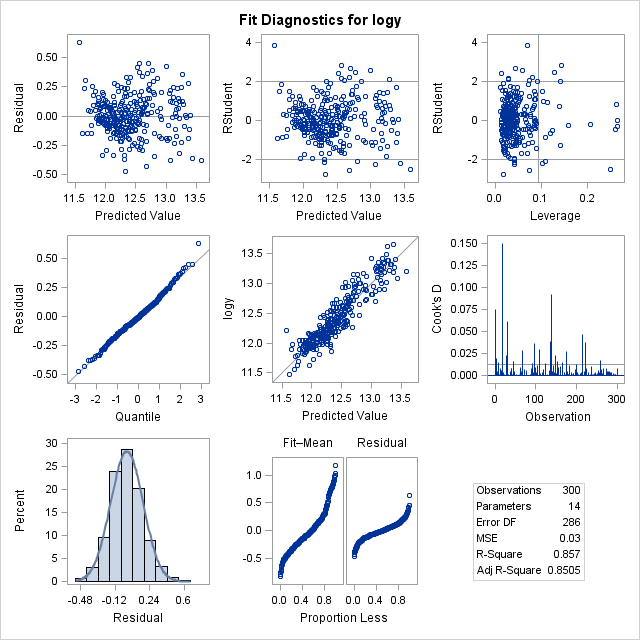


Original Regression Equation: +

For the Transformed Regression Model:

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Type I SS** | **Type II SS** |
| **Intercept** | Intercept | **1** | 1.18347 | 1.23558 | 0.96 | 0.3390 | 46209 | 0.02749 |
| **x2** | Number of Bedrooms | **1** | -0.00195 | 0.01256 | -0.16 | 0.8764 | 13.82388 | 0.00072577 |
| **x3** | Number of Bathrooms | **1** | 0.72471 | 0.27269 | 2.66 | 0.0083 | 17.89303 | 0.21162 |
| **x4** | Air Conditioning | **1** | -0.00932 | 0.03236 | -0.29 | 0.7735 | 0.68839 | 0.00249 |
| **x5** | Garage Size | **1** | 0.05791 | 0.02131 | 2.72 | 0.0070 | 5.23390 | 0.22120 |
| **x6** | Pool | **1** | 0.09922 | 0.04532 | 2.19 | 0.0294 | 0.24558 | 0.14359 |
| **yb2** | ’Year Bought Squared’ | **1** | 8.464027E-7 | 2.076665E-7 | 4.08 | <.0001 | 2.62957 | 0.49771 |
| **x8** | Quality Indicator 1 | **1** | -0.37000 | 0.05629 | -6.57 | <.0001 | 0.37651 | 1.29448 |
| **x9** | Quality Indicator 2 | **1** | -0.31833 | 0.04159 | -7.65 | <.0001 | 5.31052 | 1.75508 |
| **x10** | Style | **1** | -0.01975 | 0.00532 | -3.71 | 0.0003 | 0.02464 | 0.41211 |
| **x11** | Lot Size | **1** | 0.00000464 | 9.209599E-7 | 5.04 | <.0001 | 1.37488 | 0.76192 |
| **x12** | Adjacent to Highway | **1** | -0.11852 | 0.08813 | -1.34 | 0.1797 | 0.09465 | 0.05419 |
| **sqf2** | ’Square Footage(ln)’ | **1** | 1.04378 | 0.12102 | 8.62 | <.0001 | 3.44546 | 2.22867 |
| **bsf** | ’Bathrooms by Square Footage’ | **1** | -0.08903 | 0.03500 | -2.54 | 0.0115 | 0.19391 | 0.19391 |

New Regression Equation: ln(+

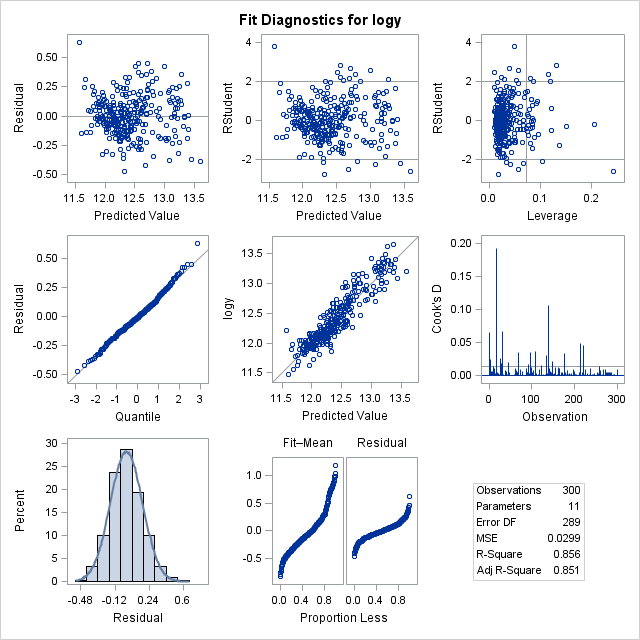


Summary of Stepwise Selection:

| **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** | sqf2 |  | ’Square Footage(ln)’ | 1 | 0.7312 | 0.7312 | 241.440 | 810.62 | <.0001 |
| **2** | yb2 |  | ’Year Bought Squared’ | 2 | 0.0429 | 0.7741 | 157.708 | 56.37 | <.0001 |
| **3** | x10 |  | Style | 3 | 0.0207 | 0.7947 | 118.378 | 29.81 | <.0001 |
| **4** | x11 |  | Lot Size | 4 | 0.0152 | 0.8100 | 89.9619 | 23.61 | <.0001 |
| **5** | x5 |  | Garage Size | 5 | 0.0081 | 0.8181 | 75.6851 | 13.16 | 0.0003 |
| **6** | x3 |  | Number of Bathrooms | 6 | 0.0043 | 0.8224 | 69.0566 | 7.12 | 0.0080 |
| **7** | x9 |  | Quality Indicator 2 | 7 | 0.0055 | 0.8280 | 59.9677 | 9.41 | 0.0024 |
| **8** | x8 |  | Quality Indicator 1 | 8 | 0.0230 | 0.8510 | 15.8986 | 45.00 | <.0001 |
| **9** | bsf |  | ’Bathrooms by Square Footage’ | 9 | 0.0025 | 0.8535 | 12.9334 | 4.92 | 0.0274 |
| **10** | x6 |  | Pool | 10 | 0.0025 | 0.8560 | 9.9836 | 4.97 | 0.0266 |

Model Given in Final Step of Stepwise Selection:

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Type I SS** | **Type II SS** |
| **Intercept** | Intercept | **1** | 1.29405 | 1.20301 | 1.08 | 0.2830 | 46209 | 0.03455 |
| **x3** | Number of Bathrooms | **1** | 0.70136 | 0.26995 | 2.60 | 0.0099 | 31.37225 | 0.20153 |
| **x5** | Garage Size | **1** | 0.05698 | 0.02110 | 2.70 | 0.0073 | 5.98237 | 0.21769 |
| **x6** | Pool | **1** | 0.10055 | 0.04511 | 2.23 | 0.0266 | 0.24323 | 0.14830 |
| **yb2** | ’Year Bought Squared’ | **1** | 8.329783E-7 | 2.025721E-7 | 4.11 | <.0001 | 2.61957 | 0.50482 |
| **x8** | Quality Indicator 1 | **1** | -0.36978 | 0.05586 | -6.62 | <.0001 | 0.37731 | 1.30831 |
| **x9** | Quality Indicator 2 | **1** | -0.32115 | 0.04090 | -7.85 | <.0001 | 5.29100 | 1.84087 |
| **x10** | Style | **1** | -0.01919 | 0.00528 | -3.63 | 0.0003 | 0.05131 | 0.39410 |
| **x11** | Lot Size | **1** | 0.00000477 | 9.051156E-7 | 5.27 | <.0001 | 1.40250 | 0.82901 |
| **sqf2** | ’Square Footage(ln)’ | **1** | 1.03431 | 0.11792 | 8.77 | <.0001 | 3.75120 | 2.29702 |
| **bsf** | ’Bathrooms by Square Footage’ | **1** | -0.08621 | 0.03466 | -2.49 | 0.0134 | 0.18474 | 0.18474 |

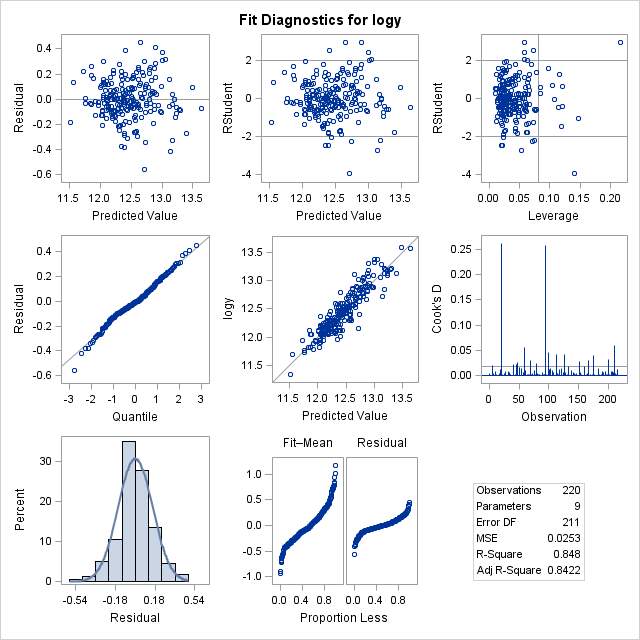


Summary of Stepwise Selection (for training set):

| **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** | sqf2 |  | ’Square Footage(ln)’ | 1 | 0.7163 | 0.7163 | 172.843 | 550.29 | <.0001 |
| **2** | yb2 |  | ’Year Bought Squared’ | 2 | 0.0567 | 0.7730 | 97.1345 | 54.20 | <.0001 |
| **3** | x11 |  | Lot Size | 3 | 0.0203 | 0.7932 | 71.3719 | 21.16 | <.0001 |
| **4** | x10 |  | Style | 4 | 0.0094 | 0.8027 | 60.4227 | 10.30 | 0.0015 |
| **5** | x9 |  | Quality Indicator 2 | 5 | 0.0065 | 0.8092 | 53.4702 | 7.33 | 0.0073 |
| **6** | x8 |  | Quality Indicator 1 | 6 | 0.0304 | 0.8396 | 13.7882 | 40.39 | <.0001 |
| **7** |  | x10 | Style | 5 | 0.0013 | 0.8383 | 13.6352 | 1.79 | 0.1824 |
| **8** | x4 |  | Air Conditioning | 6 | 0.0057 | 0.8440 | 7.8365 | 7.77 | 0.0058 |
| **9** | x3 |  | Number of Bathrooms | 7 | 0.0024 | 0.8463 | 6.5619 | 3.30 | 0.0708 |
| **10** | x10 |  | Style | 8 | 0.0017 | 0.8480 | 6.2879 | 2.30 | 0.1306 |

Model Given in Final Step of Stepwise Selection (for training set):

| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 1.88501 | 1.09028 | 0.07559 | 2.99 | 0.0853 |
| **x3** | 0.03997 | 0.01959 | 0.10527 | 4.16 | 0.0426 |
| **x4** | 0.08901 | 0.03277 | 0.18662 | 7.38 | 0.0071 |
| **yb2** | 0.00000113 | 2.15215E-7 | 0.69371 | 27.43 | <.0001 |
| **x8** | -0.32356 | 0.05758 | 0.79847 | 31.58 | <.0001 |
| **x9** | -0.29169 | 0.04157 | 1.24476 | 49.22 | <.0001 |
| **x10** | -0.00894 | 0.00589 | 0.05825 | 2.30 | 0.1306 |
| **x11** | 0.00000417 | 0.00000102 | 0.42158 | 16.67 | <.0001 |
| **sqf2** | 0.80876 | 0.07561 | 2.89335 | 114.42 | <.0001 |



Final Model:

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Type I SS** | **Type II SS** |
| **Intercept** | Intercept | **1** | 0.68760 | 1.41108 | 0.49 | 0.6266 | 34167 | 0.00622 |
| **sqf2** | ’Square Footage(ln)’ | **1** | 0.94311 | 0.13636 | 6.92 | <.0001 | 25.14404 | 1.25210 |
| **yb2** | ’Year Bought Squared’ | **1** | 0.00000119 | 2.230842E-7 | 5.33 | <.0001 | 1.99066 | 0.74453 |
| **x10** | Style | **1** | -0.00891 | 0.00609 | -1.46 | 0.1453 | 0.53735 | 0.05593 |
| **x11** | Lot Size | **1** | 0.00000369 | 0.00000105 | 3.52 | 0.0005 | 0.50557 | 0.32475 |
| **x5** | Garage Size | **1** | 0.00970 | 0.02002 | 0.48 | 0.6285 | 0.09007 | 0.00615 |
| **x3** | Number of Bathrooms | **1** | 0.41322 | 0.31564 | 1.31 | 0.1919 | 0.13708 | 0.04486 |
| **x9** | Quality Indicator 2 | **1** | -0.30443 | 0.04498 | -6.77 | <.0001 | 0.29595 | 1.19909 |
| **x8** | Quality Indicator 1 | **1** | -0.34300 | 0.05854 | -5.86 | <.0001 | 0.89153 | 0.89855 |
| **bsf** | ’Bathrooms by Square Footage’ | **1** | -0.04807 | 0.04045 | -1.19 | 0.2361 | 0.03610 | 0.03696 |
| **x6** | Pool | **1** | 0.01902 | 0.03964 | 0.48 | 0.6319 | 0.00602 | 0.00602 |

Final Regression Equation:

**Methodology:**

A correlation matrix of the original model was created to view possible multicollinearity problems, and predictor variables with a correlation coefficient greater than 0.7500 were examined more closely. Multiple residual plots were also used to come to the conclusion that a transformation would be necessary to obtain the best possible model, and the plots showed that a log transformation should be tested first. Hence, the box-cox procedure was used to find the best lambda for the log transformation as -0.25.

After the transformation, a correlation matrix was created for the new model to take any newly created multicollinearity problems into consideration. Multiple methods were used to find the best possible subsets of the regression model. From the R-Square Selection Method (using C(p) and R-Squared), the best subset of the model is one that includes 11 predictor variables: {Intercept, sqf2, yb2, x10, x11, x5, x3, x9, x8, bsf, x6}, which are the intercept, square footage, year bought squared, style, lot size, garage size, number of bathrooms, quality (if high (1) or medium (2)), bathrooms by square footage, and the presence or absence of a pool. The same results are obtained by the Stepwise Selection Method. The PRESS value of 9.54 is reasonably close to the SSE of 8.63 supporting the validity of this model and of MSE=0.0299 as an indicator of the predictive capability of this model.

To obtain greater support that these predictor variables make up the best possible regression model, the same analysis is run on a training sample. From both the R-Square Selection Method and the Stepwise Selection Method, the best possible regression model includes {Intercept, x3, x4, yb2, x8, x9, x10, x11, sqf2}. The PRESS value of 6.33 is reasonably close to the SSE of 5.47 supporting the validity of this model and of MSE=0.0262 as an indicator of the predictive capability of this model.

The similar selected predicted variables and the relative closeness of coefficient values of variables in the final models for both datasets add validity to the obtained best regression model in the original data set. Also, the MSPE of 0.0282 for the training set is close to the MSE of 0.0262 for the training set. Since MSPE is approximately MSE in the training set, the selected regression model on the training data set is not seriously biased and gives an appropriate indication of the predictive ability of the model. Therefore, the best regression model from the original data set [ is valid.

Finally, two observations are selected randomly and the variables of each are plugged into the final regression model and the predicted values are compared to the observed values. Based on the regression model, the predicted values for observations 73 and 502 are 13.44 and 12.10, respectively; the actual values were 13.73 and 12.20 for observations 73 and 502, respectively. The closeness of the observed and predicted values further supports that this is a good regression model.

**Results and Conclusions:**

The analysis done on this data set to predict sale prices in a midwestern city based on several characteristics of the home shows that square footage, year bought, style, lot size, garage size, number of bathrooms, quality (if high (1) or medium (2)), bathrooms by square footage, and the presence or absence of a pool are all strong influencers of sale prices. Therefore, the final model includes these variables. The final regression model is , and the analysis run in this study supports the validity of this model.

**Discussion:**

From the final regression model, the characteristics of houses in this midwestern city that are most influential to the sale prices of these homes are square footage, year bought, style, lot size, garage size, number of bathrooms, quality (if high (1) or medium (2)), bathrooms by square footage, and the presence or absence of a pool are all strong influencers of sale prices. Many of these variables involve the size of the home and property—square footage, lot size, garage size, and bathrooms by square footage—showing that buyers are willing to pay more for a larger home and/or more property. The price of a house can also increase with the number of bathrooms; this may also be related to larger homes often having more bathrooms. Also, newer homes usually cost more than older ones. Many buyers would pay more for a home with a pool. Most of these variables are taken into account by any potential buyer, so it is not unrealistic that a model made up of these predictor variables could make accurate predictions of the sale price for a home in the city from the study or a completely different area.