House Price Prediction using Linear Regression and Prediction Techniques

Steven Duchene

Colorado State University Global

MIS470: Data Science Foundation

Osama Morad

August 2, 2020

Housing Price Prediction Using Linear Regression and Prediction Techniques

The ability to predict housing prices is a complicated process that includes many different types of variables; ranging from height to age and everything in between. In real estate, predicting the actions of the housing market is crucial. These types of predictions are typically completed using what is called “regression analysis.” Regression analysis offers a scientific approach for the valuation and prediction of real estate properties. While typical models use factors such as comparable sales (comps), income, and cost when valuing a property; there are additional factors that should be accounted for when predicting real estate values. These incude factors such as lot size, condition of the property, number of bathrooms, year constructed, and many others.

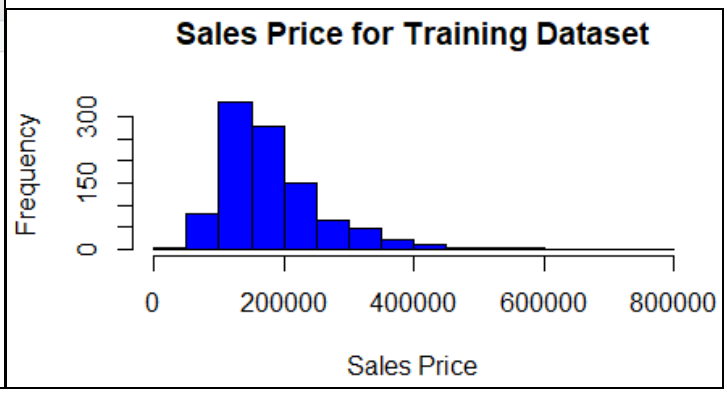
More often than not, the process of real estate valuation comprises of series of metaphorical actions by the real estate agent. They come to the property, look around, kick a few tires, and then produce a real estate valuation; a more qualitative analysis. Contrarily, linear regression and prediction takes a more quantitative approach to the real estate valuation process; one that can then be supplemented with the realtor’s qualitative analyses and local market knowledge. The goal of this report is to utilize 24 different quantitative, explanatory variables which describe residential homes in Ames, IA. The ultimate goal of this project is to predict home prices. To achieve this, a regression analysis will be completed.

Review of the Training Dataset.

In supervised learning techniques, also known as machine learning techniques, training datasets consist of data which will be used to guide the prediction process. Specifically, a training dataset is the data that is used to train an algorithm and produce predictive results (Gonfalonieri, 2019). In a prior report, the analyst discussed the basic properties of the training dataset with respect to the sale prices of the homes. The results are presented below. The shape

of the distribution appears to be normally distributed with a slight skew to the right. A majority of the datapoints lie near each other indicating that these are the most common house prices in the dataset. A right skew also indicates that the mean is likely greater than the median. This can be confirmed by the summary statistics displayed. Based on a quick review of the range, standard deviation, and mean, it is apparent that the house prices on the right of the histogram are outliers. These home prices may become an issue in future regression analysis activities.

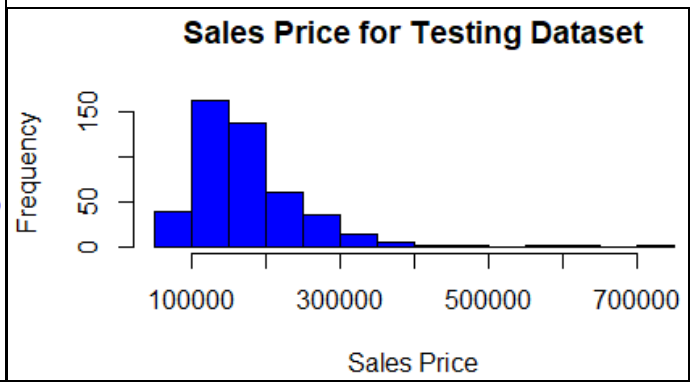
```
Console  Terminal  Jobs
~/
> range(MIS470_housing_training$SalePrice)
[1]  34900 755000
> min(MIS470_housing_training$SalePrice)
[1] 34900
> max(MIS470_housing_training$SalePrice)
[1] 755000
> mean(MIS470_housing_training$SalePrice)
[1] 182284.5
> median(MIS470_housing_training$SalePrice)
[1] 163995
> sd(MIS470_housing_training$SalePrice)
[1] 80291.22
> Sys.Date()
[1] "2020-07-05"
>
```



Evaluation of the Testing Dataset

As the training dataset is used to train an algorithm and produce predictive results, the testing dataset is utilized to determine how well the training algorithm is able to predict for any given response variable. As a rule of thumb, the testing dataset should be comprised of 20 percent of the data while the training dataset and the validation dataset (not discussed in this report) should be comprised of 60 percent and 20 percent of the data respectively. More easily stated, a 3:1:1 ratio should suffice (Shah, 2020). Presented below you will find the summary statistics and a histogram representing the sale price variable of the testing dataset. The summary statistics are comprised of the mean, median, minimum, maximum, and standard deviation values. These procedures were completed using the R Studio statistical platform where a custom function was created for calculation of the summary statistics.

```
Console   Terminal   Jobs
~/
+
+      Mean <- mean(X)
+      Median <- median(X)
+      Min <- min(X)
+      Max <- max(X)
+      StandDev <- sd(X)
+
+      print(Mean)
+      print(Median)
+      print(Min)
+      print(Max)
+      print(StandDev)
+ }
> SumStat(MIS470_housing_testing$SalePrice)
[1] 177957.6
[1] 161750
[1] 52500
[1] 745000
[1] 77569
> Sys.Date()
[1] "2020-07-27"
```



Much like the summary statistics and histogram presented for the training dataset, the shape of the distribution appears to be normally distributed with a skew to the right. A majority of the datapoints lie near each other indicating that these are the most common house prices in the dataset. A right skew also indicates that the mean is likely greater than the median. This can be confirmed by the summary statistics displayed. Based on a quick review of the range, standard deviation, and mean, it is apparent that the house prices on the right of the histogram are outliers. It should also be noted that the frequency values between the testing and training dataset appear to be significantly different; something that can be attributed to the differences in the size of the respective datasets.

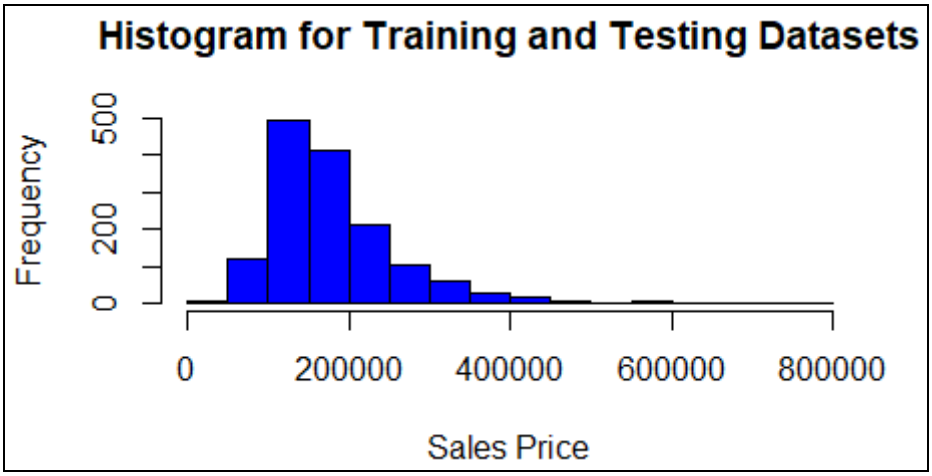
Comparison of the Training and Testing Datasets

Upon review of both the training and testing datasets, many similarities are observed. Both of the datasets exhibit a right-skewed distribution and share many of the same characteristics. Both of the datasets appear to have a majority of the values lying near each other; both similar in values. Interestingly enough, the datasets appear to have different ranges. The range of the training dataset was calculated to be $720,100 with minimum and maximum values

of $39,400 and $755,000 respectively. Alternatively, the range of the testing dataset was calculated to be $692,500 with minimum and maximum values of $52,500 and $745,000 respectively. The standard deviations for the datasets were separated by $2722 with the training dataset having a slightly larger value for standard deviation. A similar pattern was observed for the mean values of both datasets.

Combination of the Datasets

Using the merge() function in R Studio, the datasets were combined in order to develop the histogram presented below. When compared to the histograms for the testing and training datasets separately, the combined dataset exhibits many of the same properties. It is skewed to the right and exhibits the same outliers. Regarding the differences between the combined and separated datasets, when combined, the dataset exhibits a greater range than when separated.



Development of the Linear Regression Model Using the Training Dataset

Regression analysis mathematically sorts out what variables have an impact and how strong that impact is. Furthermore, regression analysis allows for a focus on associations between the variables, the ability of one variable to predict another variable, and the amount of agreement between the variables (Zou et al, 2003). Utilizing the training dataset, a linear regression model was fit using the lm() function in R Studio. The sale price was set as the

dependent variable and the 24 remaining variables of the dataset were set as the predictor variables. The results of the linear regression model are presented below.

```
> summary(fit1)

Call:
lm(formula = MIS470_housing_training$SalePrice ~ MIS470_housing_training$MSSubClass +
    MIS470_housing_training$LotFrontage + MIS470_housing_training$LotArea +
    MIS470_housing_training$OverallQual + MIS470_housing_training$OverallCond +
    MIS470_housing_training$YearBuilt + MIS470_housing_training$YearRemodAdd +
    MIS470_housing_training$MasVnrArea + MIS470_housing_training$TotalBsmtSF +
    MIS470_housing_training$GrLivArea + MIS470_housing_training$FullBath +
    MIS470_housing_training$HalfBath + MIS470_housing_training$BedroomAbvGr +
    MIS470_housing_training$KitchenAbvGr + MIS470_housing_training$TotRmsAbvGrd +
    MIS470_housing_training$Fireplaces + MIS470_housing_training$GarageYrBlt +
    MIS470_housing_training$GarageCars + MIS470_housing_training$GarageArea +
    MIS470_housing_training$WoodDeckSF + MIS470_housing_training$OpenPorchSF +
    MIS470_housing_training$MoSold + MIS470_housing_training$YrSold)

Residuals:
    Min      1Q  Median      3Q     Max
-377516  -17374   -1943   15760  236851

Coefficients:
                                        Estimate    Std. Error t value             Pr(>|t|)
(Intercept)                          -1176930.6722  1862093.1330  -0.632             0.527549
MIS470_housing_training$MSSubClass       -100.4922       37.1543  -2.705             0.006992 **
MIS470_housing_training$LotFrontage        -7.5249       67.6455  -0.111             0.911456
MIS470_housing_training$LotArea             0.7184        0.1480   4.853    0.000001481427163 ***
MIS470_housing_training$OverallQual     15279.9626     1634.5077   9.348 < 0.0000000000000002 ***
MIS470_housing_training$OverallCond      6573.6008     1534.8682   4.283    0.000020860598766 ***
MIS470_housing_training$YearBuilt         428.9194       92.1154   4.656    0.000003809409618 ***
MIS470_housing_training$YearRemodAdd      167.6588       94.0670   1.782             0.075101 .
MIS470_housing_training$MasVnrArea         32.4241        7.3511   4.411    0.000011812894272 ***
MIS470_housing_training$TotalBsmtSF        25.9663        4.4042   5.896    0.000000005646690 ***
MIS470_housing_training$GrLivArea          47.3152        6.2402   7.582    0.000000000000101 ***
MIS470_housing_training$FullBath        -2625.1576     3782.1077  -0.694             0.487836
MIS470_housing_training$HalfBath          704.7875     3376.2485   0.209             0.834701
MIS470_housing_training$BedroomAbvGr   -14906.0238     2338.9558  -6.373    0.000000000324323 ***
MIS470_housing_training$KitchenAbvGr   -27864.4605     7165.7819  -3.889             0.000110 ***
MIS470_housing_training$TotRmsAbvGrd     9115.8338     1658.4829   5.496    0.000000053195361 ***
MIS470_housing_training$Fireplaces       6198.0837     2359.1325   2.627             0.008784 **
MIS470_housing_training$GarageYrBlt       -92.7863       98.9927  -0.937             0.348905
MIS470_housing_training$GarageCars       5263.7413     3863.9563   1.362             0.173524
MIS470_housing_training$GarageArea         46.9568       13.3093   3.528             0.000444 ***
MIS470_housing_training$WoodDeckSF         20.4834       11.1860   1.831             0.067475 .
MIS470_housing_training$OpenPorchSF         0.4117       21.3592   0.019             0.984628
MIS470_housing_training$MoSold           -851.9375      470.9885  -1.809             0.070880 .
MIS470_housing_training$YrSold             56.3498      928.9655   0.061             0.951647
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33740 on 747 degrees of freedom
  (229 observations deleted due to missingness)
Multiple R-squared:  0.847,     Adjusted R-squared:  0.8423
F-statistic: 179.8 on 23 and 747 DF,  p-value: < 0.00000000000000022

> Sys.Date()
[1] "2020-07-28"
```

Determining the Significant Variables with Respect to Sale Price

Before attempting to determine the significant variables in the training dataset, it is first important to determine the goodness of fit of the linear regression model. One method of conducting this activity is through an examination of the R-squared value. The R-squared value measures the strength of the relationship between the predictor variables and the dependent variable; the dependent variable being, in this case, sales price (Frost, 2019). Typically, if the

independent variables are statistically significant, the R-squared value should be expected to be relatively high. However, it should also be noted that if the model exhibits a relatively low R-squared value, it does not necessarily indicate that the variables are not significant. Contrarily, it typically indicates that the independent variables are correlated with the dependent variable but may not explain the variation of the dependent variable (2019). The linear regression model displayed above was set to equal "fit1" in order to be used in subsequent calculations. The R-squared value for the linear model was calculated to be 0.8423 or 84.23%. As R-squared is a value that ranges between zero and one or zero percent and one percent, it is apparent that the independent variables explain a great portion of the variability seen in the home sale price.

Once the determination is made as to the goodness of fit of the linear regression model, the analyst can then conduct an analysis of the significant independent variables. In order to do so, one must look at the P-values associated with each independent variable. While P-values are often used in the processes of rejecting and failing to reject the null hypotheses, P-values can also be utilized to determine whether or not a particular variable is associated with changes in the response variable. A preliminary review of the linear regression model, conducted using the summary(fit1) function, was completed. Through this review, the following independent variables were determined to significantly influence the response variable. The variables in italics were shown to have a greater influence on variability.

- Lot frontage
- *Lot area*
- *Overall quality*
- *Overall condition*
- Year built
- *MasVnr Area*
- *Basement square footage*
- *GrLiv area*
- *Bedrooms above ground*
- *Kitchen above ground*

- *Total rooms above ground*
- *Garage area*
- Fireplaces

Now that we have identified the significant independent variables, it is important to discuss how these factors related to the response variable; sale price. This is completed via the regression coefficient. The sign of the regression coefficient indicates whether the effect of the predictor variable is positive or negative. The value of the regression coefficient indicates the amount that the dependent variable changes given a change in the independent variable. For example, for every one-unit change in Overall Quality variable, it would be expected to have a $15280 effect on the sale price. This same procedure can be completed using the rest of the independent variables. It is also important to ensure that all other variables are held constant. In doing so, the analyst can assess the effect of each variable on that of the dependent variable (Frost, 2019).

Prediction of Sale Price

In order to conduct the prediction functions available in the R Statistical Platform, specific activities must be completed. One of these activities included the removal of N/A values from the training and testing datasets. This was initially attempted using the complete.cases() function in R. Unfortunately, this was not successful. As a result, the na.omit() function was utilized on both the testing and training datasets. Once this task was completed, a new linear regression model was fit; fit1. The R function, predict() was then utilized to conduct the prediction task using the following code: predict(fit1, complete3). Complete3 was the name given to the testing dataset once the N/A values, and their associated rows, were removed.

```
> predict <- predict(fit1, complete3)
Warning message:
'newdata' had 20 rows but variables found have 771 rows
> print(head(predict, 20))
        1         2         3         4         5         6
224564.86 189722.43 214681.57 181510.39 292569.61 165791.51
        7         8         9        10        11        12
266668.55 185170.60  77158.03 109122.96 333119.58 255013.24
       13        14        15        16        17        18
153264.41  70681.10 156750.38 126980.70 304752.74 133160.85
       19        20
268205.89 141928.43
> Sys.Date()
[1] "2020-07-30"
```

Below, you will find the actual sale prices listed in the testing data set for the first twenty values; with the N/A values removed.

```
Console  Terminal  Jobs
~/
> complete3 <- na.omit(MIS470_housing_testing)
> print(head(complete3$SalePrice, 20))
 [1]  82000  86000 232000 181000 149900  88000 240000 135000 165000  85000 119200 227000
[13] 203000 213490 176000 194000  87000 191000 112500 167500
> Sys.Date()
[1] "2020-08-01"
>
```

Comparison of the Predicted Sales Price and the Actual Sales Price

There are several ways to compare the predicted sales prices to the actual sales prices listed in the testing dataset. The first method used will be the averaging of all 20 values in the dataset. The predicted average sales price was calculated to be $184,040 while the actual sales price was determined to be $156,730. On initial review, the averages above do not appear to be all that similar. Another method to compare the predicted versus actual sales prices would be to compare them individually. For the sake of time, this report will focus on a few of the most interesting results. Firstly, for the actual prices of the lowest priced homes, the predictions were highly inaccurate; the greatest of these differing by over $200,000. Secondly, it appears that the home values which lied closest to the mean, a majority of the values, the predictions were relatively accurate.

One reason that the predictions of the home values on the higher and lower ends of the spectrum were inaccurate is likely due to the presence of outliers. Often times in statistical analyses, the presence of outliers can be problematic as statistical tests are sensitive to their presence. The presence of outliers can affect both the results of the predictions and the assumptions made about the data. While simply removing the outlying values could solve the problem of prediction inaccuracy, one runs the risk of losing valuable datapoints in the dataset and subsequently affecting the ability of the model to accurately predict values.

One way to resolve this issue, while not completed in the activities conducted throughout this report, is to transform the data. Techniques such as square root and log transformations allow the analyst to "reel" in the outlying datapoints. In doing so, this can make the predictive model work better if the outlier is a dependent variable such as sale price (Martin, 2020). In conducting a log transformation, for example, the results would not only become clearer, but they would also become more accurate by reducing the residual standard error and strengthening the predictive abilities of the model.

Conclusion

As the above sections have discussed, the predictive abilities of regression models can provide invaluable insight to otherwise confusing and unclear data. By applying the regression model using the sales price as the response variable, the subsequent prediction activities were able to predict sales prices with moderate accuracy. However, issues arose when it came to predicting sales prices when the actual sales price was an outlier with regards to the rest of the dataset. The report discussed ways in which one can deal with outlying datapoints such as applying a series of transformations to the data. While liner regression models possess great predictive abilities, environments such as real estate, are often influenced by many other factors.

These factors can include supply and demand ratios, comparative home prices, economic crises, and, as we have all come to realizes, public health emergencies. While these factors are often hard to measure, let alone predict when they will occur, they are often not accounted for in predictive models. Further research into areas such as the factors listed above would serve to provide greater predictive abilities.

References

Frost, J. (2019, May 30). How To Interpret R-squared in Regression Analysis. Retrieved June 28, 2020, from https://statisticsbyjim.com/regression/interpret-r-squared-regression/.

Gonfalonieri, A. (2019, February 14). *How to Build A Data Set For Your Machine Learning Project*. https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac.

Martin, K. (2020, January 16). *Outliers: To Drop or Not to Drop*. https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/.

Shah, T. (2020, July 10). *About Train, Validation and Test Sets in Machine Learning*. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. Radiology, 227(3), 617-628.