Housing Price Analysis

By Jacquie Furbish, Djamila Oliveira, Nika Faraji

Wentworth Institute of Technology

550 Huntington Ave., Boston, MA 02115

Professor: Dr.Semere Gebresilasie

April 20, 2021

Introduction

House hunting can be a stressful experience. There are many factors that come into play when people are searching for a house. How do people know how much to sell their house for? How do people know is a house they want to buy is overpriced? In order to provide some insight into the real estate market, a study was conducted. This study analyzes amenities, locations, and other details about houses in order to determine what influences the price of a home. Two models will be used to predict the cost of a house within the data set provided.

Multiple linear regression (MLR) is a very effective model used to predict a response variable. This model is commonly used when one is trying to predict a numerical output using several exploratory variables. In order to predict the price of the house, the data set is randomly divided into 80% of the houses. The model will be created as a formula that includes the most significant coefficients. This formula can then be used by entering details of each coefficient which will result in the predicted price of the given house. In order to test the accuracy of the model, the model will be tested against the remaining 20% of the houses and compared to the true prices of each given house.

Logistic regression is another effective model that can be used to predict a response variable. A key difference between a multiple linear regression model and a logistic regression model is that the logistic regression model can only have an output of 1 or 0. Therefore, for this model, the prices of the house had to be split and redefined into two categories. Once the data is redefined, it is randomly divided into 80% of the houses. The model is then created and tested on the remaining 20% of the houses in order to determine the accuracy of the model. This model can then predict weather the price of the house will fall into one of the two categories created.

Data Description & Exploratory Analysis

The goal of this project is to determine and use the most significant variables in order to predict the sale price of a house. We want to know what factors are most important when looking at the price of a house, and with that, create an improved model to best show the accuracy of certain factors of the house which are critical. These factors include variables such as the year it

was built, remodeled, area of the house, presence of parking garage and its size, kitchen quality, overall quality, presence of a basement, and other amenities.

We start by creating a multiple linear regression model to determine which variables are most significant in impacting the price of a house. Once we observe the significant codes which determine our important predictors, we can then create a multiple regression model using only the significant variables. This will improve our model and help us predict sale price more accurately.

We will then use logistic regression to create a model based on our categorical data for classification. Here, we will use our significant predictors to create our model so that we have a stronger prediction based on our results from our initial linear regression. Our goal is to use our statistical outcomes to predict whether the price of the house will result in a 1 or a 0. 1 is defined as the price of the house being above the mean price of the house which is $184,812. 0 is defined as the price of the house being below the mean price of the house. The model created from the training data set will then be tested on the testing data set in order to determine how accurate the model is in predicting whether the price of a house will be above $184,812 or below it.

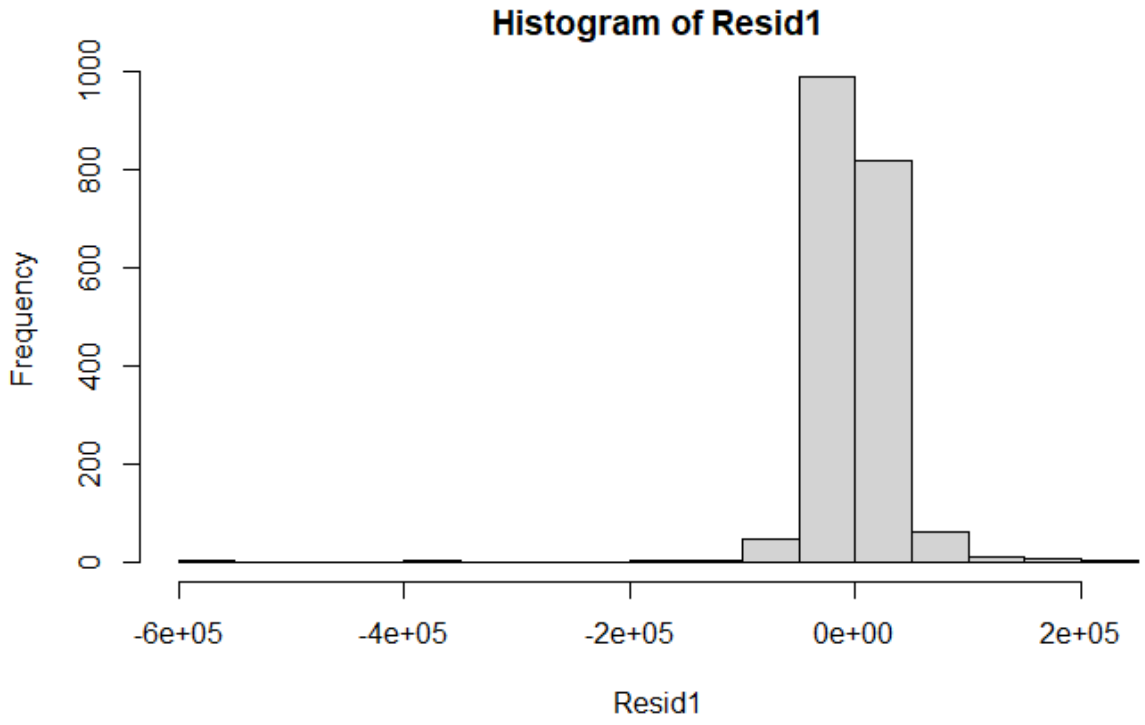
However, one thing to note is that many of the variables have various levels of conditions and traits that affect the accuracy of our model. Such as, the quality of a kitchen ranging from being excellent to poor, the garage being finished, unfinished, or having no garage at all, and the porch area, whether it be enclosed, screened, open, or for three-season. Being quantitative, we do not expect to see a big effect towards the model, but rather a low to no significance to our response variable. Once we create our different models, we conclude by executing our model on our testing data set to predict price outcomes and our model accuracies. The predictions and accuracies will determine whether we have created a strong model.

Methods and Analysis

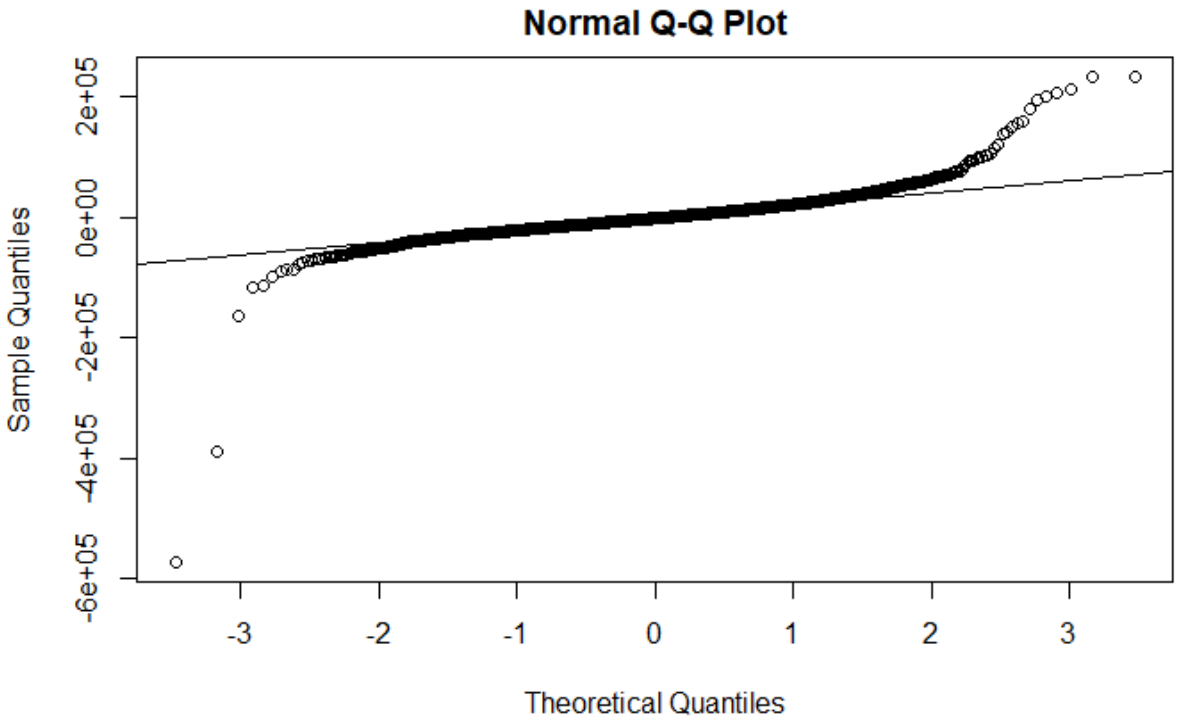
Multiple Linear Regression

The first method of implementation being multiple linear regression. This is a form of analysis that focuses on two main elements: seeing how well each predictor reacts with the

response and their estimates in the estimates of the beta coefficients. With the estimates, the relationship of both the dependent variable and multiple independent variables. We will use our training data for this method. Thus, we use sale price (SalePrice) being the response, and all other variables being the predictors, into our linear regression function to give us an output of the summary of the model, giving us the intercept, our values of the coefficients, etc. What is most important is reading the p-value of each variable, and the level of significance it has towards our response variable. If the p-value was less than our supposed given alpha of 0.05, then we presume it to be significant. From the model, the variables we focus on is lot area, overall quality and overall condition of the house, the year built, unfinished and finished basement, garage area, kitchen quality, and garage car capacity. With that, we can create a second model (MLR1) with these variables. We then check our assumptions and make sure we can use this model for our data set. We find the residuals, used to compare how much the regression line misses a point from the data, and detects the errors that are not explained from the regression line. Our results can be seen below, in the form of a histogram and a line plot of the residuals.



The histogram although does not look like a normal histogram, still shows to be relatively normal plot with a bell-shaped curve. The resulting plot can be seen still as linear and the error terms normal as well.



On our QQ plot, we can see that there is also normality, having there be outliers on both ends and trailing off in opposite directions from the beginning and end of the line. Our model was calculated to be:

$$\begin{aligned} SalePrice = {} & -844700 + 0.5087(LotArea) + 15250(OverallQual) + 4832(OverallCond) \\ & + 423.4(YearBuilt) + 33.98(BsmtFin) + 17.94(BsmtUnf) \\ & + 51.61(GrLivArea) - 54110(KitchenQualFa) - 59300(KitchenQualGd) \\ & - 44170(KitchenQualPo) - 66140(KitchenQualTA) \\ & + 13120(GarageCars) \end{aligned}$$

In order to confirm and check our accuracy of the model, we use the model to predict the prices of the houses within the testing data set, printing out each value from the 485 houses. With that, we can calculate the error by taking the absolute value of the difference between each sale

Price value of the testing data from the predicted value, all over each sale price value from the testing data again, multiplied by 100. Once again, we receive all 485 value outcomes, and take the overall mean of the error values in order to find the percentage error, being around 12%. Thus, the overall accuracy of using a multiple linear regression model is 88.27%, deeming a dependable and useful method to finding and predicting the sale price of selling a house.

Logistic Regression

For our second implementation, we have a logistic model using only our most significant variables for better accuracy. Logistic regression is another form of regression modeling which is often used for classification. However, for us to be able to use classification methods, our response variable must be binary where the output is either 0 or 1 based on what each value represents. Since the housing data set's response variable is "SalePrice", which is quantitative, we must manipulate the data accordingly. To do this, we can use an "if-else" statement to set all the set the column of price to 0s and 1s. We gave the value 0 to those prices less than the mean of all the prices and 1 to those greater than the mean which is $184,812. At this point we create a random data set of 80% of the data to be used to create the model, and 20% to test the accuracy of the model. Using our new data set, we can now run our logistic model with what was determined to be the five most significant predictors which are overall condition, the year it was built, presence of basement, and the area. Our final logistic model with the outputted coefficients is:

$$\frac{e^{-144.3+0.759(OverallCond)+0.06525(YearBuilt)+0.00237(BsmtFin)+0.002(BsmtUnf)+0.0053(GrLivArea)}}{1+e^{-144.3+0.759(OverallCond)+0.06525(YearBuilt)+0.00237(BsmtFin)+0.002(BsmtUnf)+0.0053(GrLivArea)}}$$

Using our model, we want to know the accuracy of it predicting the sale price of houses based on these explanatory variables. We do this by using the model made with the training data set to calculate the probability of our output equaling 1 within our testing data set. With these results, anything with a probability of 0.5 or higher was considered 1 while anything below a probability of 0.5 was considered 0. These results were then compared with the true value of each house. The accuracy was then calculated to be 92.78% accurate.

Conclusion

After creating and testing both models, the best results came from the logistic regression model with its accuracy of 92.78%. Although this model is more accurate, it is not capable of predicting an exact price but rather the price range the house will fall under. Due to this, both methods are valuable in predicting the price of a house. If it is desired to predict a precise price of a house, the multiple linear regression model will be the better model to use even those the accuracy is not as high, resulting with the 88.27% accuracy as previously stated. If someone only wants to know the price range that a house falls under then the logistic regression model would be the better model to use. This is because is it more accurate and if a precise value is not needed, it will be the better model to use. If we were to analyze this data in a different approach, we could find the correlation between each variable with the sale price, to see if there is any change in which predictor has a stronger more positive relationship. Variance inflation could also be another method to determining level of correlation. Another variable we could use as a response is the condition and quality of the house. This could have a different outcome based on how well each room or designated area of the house was worked on, as stated from our description of the data.

Appendix

Below is the R-script we made in order to create the models:

```
housing <- read.csv("housing.csv")
attach(housing)
#data <- sort(sample(nrow(housing),nrow(housing)*.8))
#training <- housing[data,]
#testing <- housing[-data,]
library("writexl")
#write.csv(training, file="trainingdata.csv")
#write.csv(testing, file="testingdata.csv")
training <- read.csv("trainingdata.csv")
```

```
attach(training)
testing <- read.csv("testingdata.csv")
attach(testing)
MLR <- lm(SalePrice ~ .,data=training)
summary(MLR)
## Call:
## lm(formula = SalePrice ~ ., data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -560568  -15160   -1005   12973  243777
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.366e+07  6.905e+06   3.427 0.000624 ***
## X                 9.099e+01  1.183e+02   0.769 0.442030
## ï..Order         -9.330e+01  1.016e+02  -0.918 0.358686
## LotArea           4.403e-01  1.021e-01   4.312 1.71e-05 ***
## Alleyyes         -6.081e+03  3.789e+03  -1.605 0.108710
## LotConfigCulDSac  1.151e+04  3.442e+03   3.345 0.000839 ***
## LotConfigFR2     -5.600e+03  5.075e+03  -1.103 0.269977
## LotConfigFR3     -1.272e+04  1.353e+04  -0.940 0.347203
## LotConfigInside   3.376e+03  1.978e+03   1.706 0.088099 .
## OverallQual       1.363e+04  1.005e+03  13.573  < 2e-16 ***
## OverallCond       6.290e+03  9.136e+02   6.885 7.95e-12 ***
## YearBuilt         3.522e+02  6.185e+01   5.695 1.44e-08 ***
## YearRemod        -1.578e+01  5.818e+01  -0.271 0.786266
## FoundationCBlock -1.219e+03  3.228e+03  -0.378 0.705750
## FoundationPConc   7.395e+03  3.764e+03   1.965 0.049604 *
## FoundationSlab    4.621e+03  8.470e+03   0.546 0.585468
## FoundationStone   9.663e+03  1.188e+04   0.813 0.416156
## FoundationWood   -1.435e+04  1.689e+04  -0.850 0.395421
## BsmtFin           2.864e+01  2.842e+00  10.078  < 2e-16 ***
```

```
## BsmtUnf          1.742e+01  2.986e+00   5.833 6.43e-09 ***
## ACY              3.878e+02  4.024e+03   0.096 0.923251
## GrLivArea        4.615e+01  3.490e+00  13.224  < 2e-16 ***
## HalfBath        -1.525e+03  1.877e+03  -0.812 0.416617
## FullBath         4.882e+03  1.726e+03   2.830 0.004713 **
## BedroomAbvGr    -6.462e+03  1.506e+03  -4.290 1.88e-05 ***
## KitchenQualFa   -5.754e+04  6.902e+03  -8.337  < 2e-16 ***
## KitchenQualGd   -5.584e+04  3.491e+03 -15.995  < 2e-16 ***
## KitchenQualPo   -2.748e+04  3.323e+04  -0.827 0.408480
## KitchenQualTA   -6.060e+04  4.091e+03 -14.814  < 2e-16 ***
## TotRmsAbvGrd     2.167e+03  1.030e+03   2.105 0.035470 *
## Fireplaces       4.421e+03  1.430e+03   3.092 0.002017 **
## GarageFinishRFn -5.253e+03  2.165e+03  -2.426 0.015343 *
## GarageFinishUnf -4.154e+02  2.524e+03  -0.165 0.869292
## GarageCars       9.077e+03  2.453e+03   3.701 0.000221 ***
## GarageArea       2.163e+01  7.877e+00   2.746 0.006086 **
## WoodDeckSF       2.031e+01  6.507e+00   3.121 0.001830 **
## PorchSF          1.829e+01  7.575e+00   2.415 0.015833 *
## YrSold          -1.211e+04  3.429e+03  -3.533 0.000421 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32550 on 1821 degrees of freedom
##   (81 observations deleted due to missingness)
## Multiple R-squared:  0.8513, Adjusted R-squared:  0.8483
## F-statistic: 281.8 on 37 and 1821 DF,  p-value: < 2.2e-16
```

```
MLR1 <-lm(SalePrice ~
LotArea+OverallQual+OverallCond+YearBuilt+BsmtFin+BsmtUnf+GrLivArea+KitchenQual+GarageCars,
data=training)

summary(MLR1)
```

```
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + OverallCond +
##     YearBuilt + BsmtFin + BsmtUnf + GrLivArea + KitchenQual +
##     GarageCars, data = training)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -572078  -15219   -1788   12682  233823
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.447e+05  7.513e+04 -11.244  < 2e-16 ***
## LotArea        5.087e-01  1.003e-01   5.070 4.35e-07 ***
## OverallQual    1.525e+04  9.562e+02  15.945  < 2e-16 ***
## OverallCond    4.832e+03  7.558e+02   6.393 2.03e-10 ***
## YearBuilt      4.234e+02  3.800e+01  11.143  < 2e-16 ***
## BsmtFin        3.398e+01  2.368e+00  14.350  < 2e-16 ***
## BsmtUnf        1.794e+01  2.577e+00   6.961 4.61e-12 ***
## GrLivArea      5.161e+01  2.046e+00  25.220  < 2e-16 ***
## KitchenQualFa -5.411e+04  6.277e+03  -8.621  < 2e-16 ***
## KitchenQualGd -5.930e+04  3.358e+03 -17.660  < 2e-16 ***
## KitchenQualPo -4.417e+04  3.366e+04  -1.312     0.19
## KitchenQualTA -6.614e+04  3.827e+03 -17.285  < 2e-16 ***
## GarageCars     1.312e+04  1.431e+03   9.167  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33270 on 1925 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8429
## F-statistic: 866.8 on 12 and 1925 DF,  p-value: < 2.2e-16
Resid1 <- MLR1$residuals
hist(Resid1)
qqnorm(Resid1)
qqline(Resid1)
plot(MLR1)
pred <- predict.lm(MLR1, testing)
```

```
Pred
Error <- (abs(testing$SalePrice-pred)/testing$SalePrice)*100
Error
MeanError <- sum(Error)/485
MeanError
Accuracy <- 100-MeanError
Accuracy
housing1 <- read.csv("housing1.csv")
attach(housing1)
#PriceYN <- ifelse(housing1$SalePrice>mean(housing1$SalePrice), 1, 0)
#housing1$PriceYN <- as.factor(PriceYN)
#PriceYN
#data1 <- sort(sample(nrow(housing1),nrow(housing1)*.8))
#training1 <- housing1[data1,]
#testing1 <- housing1[-data1,]
#library("writexl")
#write.csv(training1, file="trainingdata1.csv")
#write.csv(testing1, file="testingdata1.csv")
testing1 <- read.csv("testingdata1.csv")
training1 <- read.csv("trainingdata1.csv")
attach(training1)
attach(testing1)
log <-glm(training1$PriceYN ~ OverallCond+YearBuilt+BsmtFin+BsmtUnf+GrLivArea, data=training1,
family="binomial")
summary(log)
## Call:
## glm(formula = training1$PriceYN ~ OverallCond + YearBuilt + BsmtFin +
##     BsmtUnf + GrLivArea, family = "binomial", data = training1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.3072  -0.0715   0.2911   3.2711
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.443e+02  9.340e+00 -15.445  < 2e-16 ***
## OverallCond  7.591e-01  1.102e-01   6.888 5.66e-12 ***
## YearBuilt    6.525e-02  4.398e-03  14.836  < 2e-16 ***
## BsmtFin      2.368e-03  2.776e-04   8.530  < 2e-16 ***
## BsmtUnf      2.009e-03  2.772e-04   7.246 4.29e-13 ***
## GrLivArea    5.281e-03  3.007e-04  17.566  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2604.49  on 1938  degrees of freedom
## Residual deviance:  991.01  on 1933  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1003
##
## Number of Fisher Scoring iterations: 25

testing1$predict <- (exp(-1.443e+02+(7.591e-01*testing1$OverallCond)+(6.525e-02*testing1$YearBuilt)+(2.368e-03*testing1$BsmtFin)+(2.009e-03*testing1$BsmtUnf)+(5.281e-03*testing1$GrLivArea)))/(1+(exp(-1.443e+02+(7.591e-01*testing1$OverallCond)+(6.525e-02*testing1$YearBuilt)+(2.368e-03*testing1$BsmtFin)+(2.009e-03*testing1$BsmtUnf)+(5.281e-03*testing1$GrLivArea))))

testing1$predict

write.csv(testing1$predict, file="logisticResults.csv")
```