Final Presentation

Steffen Rigby, Kendra DeLange, Joshua Baugh, Alvaro Garcia

2022-12-30

# Introduction

Our project’s goal is to examine the effects that determine how many days a house is on the Logan market. We hypothesize that the size and location of the house will influence the decision of potential buyers; namely, houses with certain qualities will sell more quickly. That being said, we will use the data to estimate these effects. For example, what are the primary causes that determine how long a house is on the market in Logan? How do houses located in the Logan School District fare on the market compared to houses in the Cache School District? These questions are relevant and important because almost everyone at one point will participate in the housing market. These questions are especially important to those selling homes in Logan.

In summary, our formal research question is: What are the primary factors that determine how long a house is on the market in Logan, Utah?

# Data

For this project, we used a data set that was complied on UtahRealEstate.com. Data includes variables from the Logan, Utah housing market in 2019. However, we have no further information regarding the collection of the data.

Our data set contained 4110 observations on 9 variables. We used all 9 variables, namely: DOM (days on the market: how long a house was up for sale); Sold\_Price (the final price that the house sold for, measured in dollars); Garage\_Capacity (how many cars could fit within the garage); Quadrant (the sector of Logan in which the house is located); School\_District (whether the house belonged in the Cache or Logan School District); Total\_Bathrooms (total number of bathrooms); Total\_Bedrooms (total number of bedrooms); Total\_SQ (the total square feet of the house, measured in feet); and month\_sold (the month in which the house sold, measured numerically between 1 and 12). Additionally, as we tested various regression models we used the log of DOM and the squares of Total\_SQ and Garage\_Capacity. We also tested the interaction terms between Sold\_Price and School\_District, as well as, Total\_Bedrooms and School\_District.

For our regression models we used DOM as the dependent variable, with the other 8 variables as independent.

The following table shows summary statistics for the variables. Some of the interesting observations include:

skim(Log\_hous)

Data summary

|  |  |
| --- | --- |
| Name | Log\_hous |
| Number of rows | 4110 |
| Number of columns | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Quadrant | 0 | 1 | FALSE | 4 | NE: 1164, SW: 1119, SE: 1027, NW: 800 |
| School\_District | 0 | 1 | FALSE | 2 | Cac: 2735, Log: 1375 |
| month\_sold | 0 | 1 | FALSE | 12 | 8: 475, 7: 471, 6: 468, 5: 413 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sold\_Price | 0 | 1 | 282125.25 | 117864.72 | 13000 | 214000 | 260000 | 324900.00 | 1900000 | ▇▁▁▁▁ |
| DOM | 0 | 1 | 40.55 | 51.57 | 1 | 6 | 23 | 56.00 | 712 | ▇▁▁▁▁ |
| Garage\_Capacity | 0 | 1 | 1.83 | 1.09 | 0 | 1 | 2 | 2.00 | 20 | ▇▁▁▁▁ |
| Total\_Bedrooms | 0 | 1 | 3.82 | 1.18 | 1 | 3 | 4 | 5.00 | 9 | ▁▇▂▁▁ |
| Total\_Bathrooms | 0 | 1 | 2.58 | 0.85 | 1 | 2 | 3 | 3.00 | 7 | ▇▇▂▁▁ |
| Total\_SQ | 0 | 1 | 2370.43 | 1049.69 | 520 | 1583 | 2141 | 2941.75 | 10890 | ▇▃▁▁▁ |

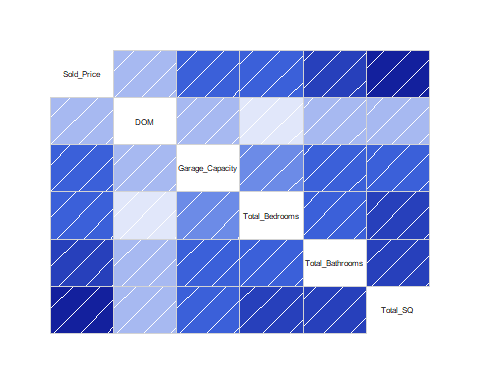
The mean Sold\_Price was $282,125, which seems very low in our opinion. That being said, this data set was complied in 2019. This was before the COVID 19 pandemic, which significantly increased housing prices due to material shortages.

The mean for month\_sold is 6.678. That suggests that the mean month in which these houses sold in 2019 was during the middle of June and even the surrounding months. This average makes sense as weather in Logan during the winter is more difficult to move in. Winter months are also holiday season and times when children would be in school. Therefor, high demand for buying homes in the summer months makes sense.

The max for total bedrooms was 9. That number suggests that this data set contains some very large houses. This is surprising to us, considering the location of Logan. It is not set up against the mountains like Providence or River Heights, in which there are many large houses overlooking Cache Valley. Furthermore, the max for total square feet was 10890. Upon doing the conversions, that means that the largest house in this data set was 0.25 acres large. That is also surprising, since according to the 2020 census the median household income in Logan, Utah is $43056.

Below we have a correlation diagram. This diagram includes all of the variables that we used in our analysis. From this diagram we were hoping to find that DOM was more highly correlated with the dependent variables. However, there is fairly low correlation observed. We can aslo see that there are some highly correlated variables in which we needed to check VIF for.

corrgram(Log\_hous)



# Base Model

Model Composed of all variables in our sample.

MRM<-lm(DOM~I(Sold\_Price/10000)+Garage\_Capacity+Quadrant+School\_District+Total\_Bedrooms+Total\_Bathrooms+I(Total\_SQ/100)+month\_sold, Log\_hous)

# Empirical Framework

In order for us to keep using the Ordinary Least Squares (OLS) technique, we need to prove that our Data satisfies the Classical Linear Model Assumptions.

MLR.1 Linear Parameters: Assumption met by our base model.

MLR.2 Random Sampling: We have no information regarding how our data was gathered or sampled, but assume that the dataset is randomly sampled as it comes from a trusted source, and has a relatively big size. However there are likely some limitations to the quality of the data, provided it’s mostly transactional data gathered by a seller.

MLR.3 No Perfect Colinearity: According to our correlation chart, there are some variables that are highly correlated but no variables exhibit perfect collinearity.

vif(MRM)

## GVIF Df GVIF^(1/(2\*Df))  
## I(Sold\_Price/10000) 3.847142 1 1.961413  
## Garage\_Capacity 1.575057 1 1.255013  
## Quadrant 1.243372 3 1.036972  
## School\_District 1.210137 1 1.100062  
## Total\_Bedrooms 1.911076 1 1.382417  
## Total\_Bathrooms 1.817124 1 1.348008  
## I(Total\_SQ/100) 4.528086 1 2.127930  
## month\_sold 1.048443 11 1.002153

As noted above, there is no need for concern of multicollinearity as the vif never surpasses 10.

MLR.4 Zero Conditional mean: There should be nothing in the error term that is correlated with both the explanatory and our independent variables. Satisfying this assumption can be challenging and its violation is normally caused by omiting significant variables in the model. We include most of the variables in our sample, however, there could still be other variables we don’t know of -external to the sample- causing some bias in our results.

MLR.5 Homoscedasticity: This assumption dictates that the variance across all the Xs in the model should be relatively constant.

We ran a Breusch-Pagan hypotheis test, with to see if we are able to fail to reject the null.

bptest(MRM)

##   
## studentized Breusch-Pagan test  
##   
## data: MRM  
## BP = 208.82, df = 20, p-value < 2.2e-16

As seen above, the p-value is too small, meaning we reject the null. In the empirical framework below we will explore various methods of solving this issue.

MLR.6 Normality: Checking for this assumption is unecessary due to the fact that we have such a large sample size -central limit theorem-.

Not all OLS assumptions were met, meaning our base model is not BLUE -Best Linear Unbiased Estimator-. In the next section of this report we’ll show the different measures used to correct them and achieve a significant final model.

# Tested Models

## Log-Level

We tried to use a log-level to correct the heteroskedasticity in our initial model, however that didn’t help. The results of a bp-test showed that our variance was even more homoskedastic than before.

## Garage Capacity Squared

$DOM=𝛽\_0 +𝛽\_1SoldPrice+ 𝛽\_2GarageCapacity+ 𝛽\_3 Quadrant+ 𝛽\_4SchoolDistrict+ 𝛽\_5TotalBedrooms+𝛽\_6TotalBathrooms+ 𝛽\_7TotalSQ+ 𝛽\_8Month\_Sold +𝛽\_9GarageCapacity^2 +u $

This model increased the significance of Garage Capacity and showed us that the variable has a diminishing effect over days on the market.

## SoldPrice:School District

The interaction added in the model above decreased the significance of other variables and the of our model. We decided to keep it out of the model.

## Sold Price Squared + Total SQ Feet Squared

This interaction increased the significance of the Total Squared feet in our model, corroborating the idea that this variable has a diminishing marginal effect on DOM. However, the significance of other variables decreased so much that we decided to keep this variable out of the model. This is likely due to multicolinearity between the interaction and other predictors.

We also added Sold Price squared to the model, and decided to keep it as it increased the significance of Sold Price without influencing other variables in our model.

# Results

Testing the models described above eventually leads to the conclusion that the model with the best fit for the equation, as measured by adjusted R2, is:

The results for this model are reported below.

MRM\_quad\_int<-lm(DOM~I(Sold\_Price/10000)+I((Sold\_Price/10000)^2)+Garage\_Capacity+I(Garage\_Capacity^2)+Quadrant+School\_District+Total\_Bedrooms:School\_District+Total\_Bedrooms+Total\_Bathrooms+I(Total\_SQ/100)+month\_sold, Log\_hous)  
  
robust\_se\_quad<-coeftest(MRM\_quad\_int, vcov= hccm(MRM\_quad\_int, type="hc0"))[,2]  
  
stargazer(MRM\_quad\_int,type = "text", se=list(robust\_se\_quad))

##   
## ===============================================================  
## Dependent variable:   
## ---------------------------  
## DOM   
## ---------------------------------------------------------------  
## I(Sold\_Price/10000) -0.994\*\*\*   
## (0.363)   
##   
## I((Sold\_Price/10000)2) 0.008\*\*   
## (0.004)   
##   
## Garage\_Capacity 4.670\*\*\*   
## (1.322)   
##   
## I(Garage\_Capacity2) -0.426\*\*\*   
## (0.158)   
##   
## QuadrantNW -1.053   
## (2.384)   
##   
## QuadrantSE -3.963\*   
## (2.193)   
##   
## QuadrantSW -4.410\*\*   
## (2.164)   
##   
## School\_DistrictLogan -19.901\*\*\*   
## (5.253)   
##   
## Total\_Bedrooms -6.390\*\*\*   
## (1.188)   
##   
## Total\_Bathrooms 3.340\*\*   
## (1.306)   
##   
## I(Total\_SQ/100) 1.694\*\*\*   
## (0.232)   
##   
## month\_sold2 1.998   
## (4.863)   
##   
## month\_sold3 -2.243   
## (4.678)   
##   
## month\_sold4 -16.734\*\*\*   
## (4.187)   
##   
## month\_sold5 -15.820\*\*\*   
## (4.176)   
##   
## month\_sold6 -20.946\*\*\*   
## (3.844)   
##   
## month\_sold7 -18.573\*\*\*   
## (3.919)   
##   
## month\_sold8 -18.414\*\*\*   
## (3.819)   
##   
## month\_sold9 -18.866\*\*\*   
## (3.909)   
##   
## month\_sold10 -16.459\*\*\*   
## (3.956)   
##   
## month\_sold11 -14.789\*\*\*   
## (4.431)   
##   
## month\_sold12 -2.466   
## (4.560)   
##   
## School\_DistrictLogan:Total\_Bedrooms 3.310\*\*   
## (1.404)   
##   
## Constant 49.241\*\*\*   
## (6.315)   
##   
## ---------------------------------------------------------------  
## Observations 4,110   
## R2 0.125   
## Adjusted R2 0.120   
## Residual Std. Error 48.383 (df = 4086)   
## F Statistic 25.330\*\*\* (df = 23; 4086)   
## ===============================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

This model has the highest adjusted R2 of any of our models: 0.12. So, the model can only account for 12 percent of the differences in wages among different people. This must mean that there are further variables that we do not have access to or data about that might contribute to a greater explanation of our dependent variable.

Except for QuadrantSW, QuadrantSE, and month\_sold2,3,12, all of our variables are significant at the 5% level.

As this is a level-level model, all changes in wage are measure in direct units. Please note that we changed the units for Sold\_Price and Total\_SQ to tens of thousands of dollars and hundreds of square feet respectively; this made it easier to explain the partial effects of these factors on our dependent variable.

Answering the original question, houses located in the Logan School District sell 20 days quicker than those located in Cache, as indicated by the estimated coefficient .

# Conclusion

Following are the interpretations of our most statistically significant variables:

 Price: An increase of $10,000 decreases the DOM by 0.994; Diminishing marginal effect

 Garage Capacity: The capacity to hold an extra car increases the DOM by 4.67; diminishing marginal effect

 nLogan School District: Houses in the Logan school are on the market almost 20 days less than houses in Cache

 Total Bedrooms: An extra bedroom will decrease the average days in the market by about 6 days

 Total Bathrooms: An extra Bathroom will increase the days in the market of the house by about 3 days on average.

 Total Square Feet: An extra 100 Square feet on a house will increase the house’s DOM by about 1.7 days on average.

The results of our regression show that it if attempting to sell a house as quickly as possible, it is preferable to build a house in the Logan School District and to sell in the summer months. As mentioned in the class the results of our model are relatively standard, and support pre-existing notions about possible variable interactions within the housing market.

For future reference, It’s important to recognize that OLS is a relatively limited technology, as it only works whenever a linear representation of the data can be achieved, and does not recognize other patterns. It would be beneficial for those interested in further exploring this data to use different techniques that make up for this constraint.