**Date-11/23/2021**

**Name – Sylvia and Sam**

**Saratoga Data Project**

**Introduction**

Data about the pricing of Saratoga, New York's 1728 single-family residences was gathered, although it was somewhat altered for this assignment. In contrast to the subjective technique used by real estate agents, regression analysis gives a more objective method for valuing a property based on its attributes (size, amenities, location, etc.) and the price of comparable properties nearby. A model for predicting a Saratoga, New York home's worth must be developed using the data supplied in the "Saratoga" data table. Data has been shown in the following tables for the reference in creating the below tables, each information above is clear of the values and diagrams also represented in the pictographically for the clear sense. This data is collected for the reference of creating the values for the tables which have been created in the different modules to represent the data in the clear view. We are going to analyze which of the variables predict the Sale prices in the data.

**Methods**

We are analyzing data on 1728 homes sold in Saratoga, NY. We are interested in creating a model to estimate sale price by these independent variables: Size in thousands of sq feet, lot size in acres, age of home in years, value of land in dollars, square footage of living area, # of bedrooms, # of fireplaces, # of bathrooms, total # rooms, type of heating(electric, hot air, hot water), source of fuel(electric, gas, or oil), sewer type(none, public, septic), if waterfront property or not, if new construction or not, if home has central air or not. This is summarized in Table 1.

|  |  |
| --- | --- |
| **Table 1: Summary of Variables** | |
| **Variable** | **Description** |
| y | Sale Price ($) |
|  | Lot Size(acre) |
|  | Building age (year) |
|  | Land Value ($) |
|  | Living area size (sq ft) |
|  | # of bedrooms |
|  | # of fireplaces |
|  | # of bathrooms |
|  | # of rooms |
|  | = 1 if electric heating  = 0 if other |
|  | = 1 if hot air heating  = 0 if other |
|  | = 1 if oil source  =0 if other |
|  | = 1 if gas fuel source  = 0 if other |
|  | = 1 if no sewer  = 0 if other |
|  | = 1 if public sewer  = 0 if other |
|  | = 1 if waterfront property  = 0 if not |
|  | = 1 if new construction  = 0 if not |
|  | = 1 if home has central air  = 0 if not |
| Default values for dummy variables | Hot water heating, electric fuel source, septic sewer |

We start with exploratory analysis of the data. Examining histograms of each variable, we observe high skewness in price, lot size, age, land value, and living area. To normalize the data, we transform each variable, preferring to use a natural log transformation, but we note lot size and age have some zero values. Performing a ln transformation will result in undefined data, so we opt for a sqrt transformation for these two variables. These transformed variables are summarized in Table 2.

We also notice how the variables recording counts of a feature can seem to be grouped as a dichotomous variable easily. We examine graphs of price against # of bedrooms, # of fireplaces, # of bathrooms, and # of rooms to ensure this won’t distort the data. We consider simplifying these variables as such: # of bedrooms to if bedrooms are greater than 4, # of fireplaces to if fireplaces greater or equal to 1, # of bathrooms to if 2 or more bathrooms, # of rooms to if rooms greater than or equal to 7. These transformed variables are summarized in Table 2.

|  |  |
| --- | --- |
| **Table 2: Summary of Transformed Variables** | |
| **Variable** | **Description** |
|  | Ln of Price |
|  | Sqrt of Lot size |
|  | Sqrt of Age |
|  | Ln of Land value |
|  | Ln of Living area |
|  | =1 if Bedrooms >= 4  =0 otherwise |
|  | =1 if Fireplaces >= 1  =0 otherwise |
|  | =1 if Bathrooms >= 2  =0 otherwise |
|  | =1 if Rooms >= 7  =0 otherwise |

We create a table of summary statistics to better visualize the data, as well as graphs plotting price vs each independent variable to better grasp what order model we should use. We note the lack of discontinuity in the graphs and rule out needing to utilize piecewise regression. We also utilize stepwise regression to see which variables contribute most to prediction accuracy. We consider our own insight of the situation to apply knowledge of interactions. Yet with all this research we notice in our data just how spread it is.

Our first model will be a first order model utilizing each variable as a baseline. Model 1 is as follows:

With this basic model we can examine the plot of residuals vs predicted. We note that variance in the residuals, how it is generally even. Ultimately, we can’t use this model for comparison as the skewness of the response variable breaks the equal variation of residuals assumption for least squares regression.

For our second model we see if normalizing the independent variables will improve predictive power, utilizing the variables in Table 2. Model 2 is as follows:

We note that the ln of land value and living area add the most to the predictive ability to the models, and as such expand on this including a higher order model for these terms. With this comes as many as 80 variables, so we create a more concise model using stepwise regression using a 5 Max K-fold RSquare stopping rule to validate the model. To keep the assumption of equal variance of the residuals for Least Squares Regression true, we must continue using the ln transformation on price. Model 3 is as follows:

We note the RSquare K-Fold is similar to the model’s RSquare and conclude the predictive power of this model should perform well for prediction. We also note the graph of residuals vs predicted that spread of the data does seem uniform and random.

We then compare the models, primarily models 2 and 3 to see if the complexity of model 3 is necessary for better predictions by using a partial F-tests at a 0.05 significance level. We record and compare the fit of each model by each model’s RSquare adjusted, S, and global F statistic, and some more statistics necessary to calculate F statistics. We then compare if these hypothesized models predict better than model 1 which uses each variable linearly. Once we have our best fitting model we will check for influential data and mitigate if needed, as well as validate the model.

**Results**

In describing the problem, we create histograms of the independent variables to visualize the data, as shown in Table 3. We include both original data and transformed data in this table. We note that many of the variables are very skewed. Our response to this is to attempt to normalize the data, creating the variables mentioned in Table 2. These histograms show many outliers in their respective data space, which is going to affect the predictive power of each model hypothesized. It should be noted that resulting transformations do not do much to mitigate these outliers.

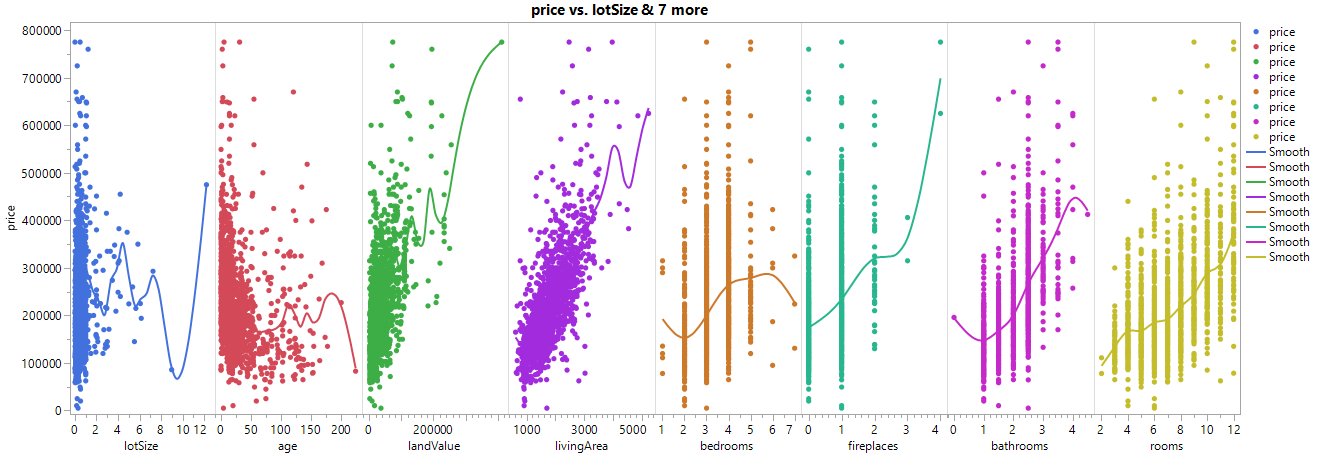
|  |  |
| --- | --- |
| **Table 3: Histograms of the data** | |
| **Original data** | **Transformed data** |
| **price** | **Ln price** |
| **Lot size** | **Square root lot size** |
| **Age** | **Square root age** |
| **Land value** | **Ln land value** |
| **living area** | **In living area** |
| **Bedrooms** | **Bedrooms > 3** |
| **Fireplaces** | **Fireplaces > 1** |
| **Bathrooms** | **Bathrooms >=2** |
| **Rooms** | **Rooms >= 7** |

We also include histograms of the dummy variables and note the rarity of waterfront and new construction homes can align with our expectations of reality. That being that there isn’t nearly as much water to build around, and that most homes are already built. These histograms are included in Table 4.

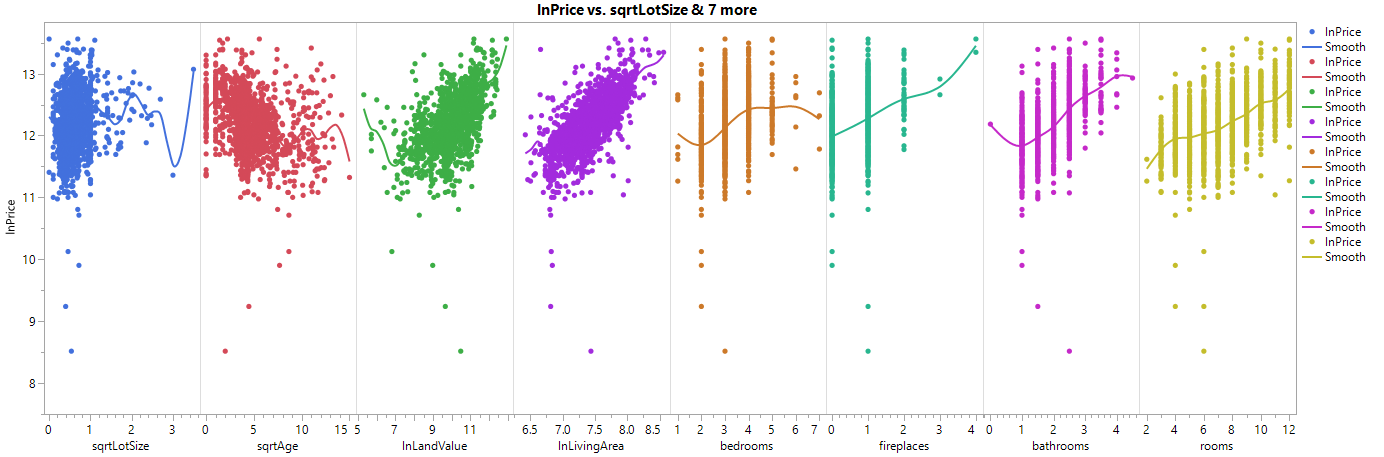
|  |  |
| --- | --- |
| Table 4: Histograms of Dummy variables | |
| Heating | Fuel |
| Sewer | Waterfront |
| New Construction | Central Air |

We include Graph 1, our base variables plotted against price to visualize any trends in the data to help us create our models. Once again, we notice the skewness and spread of the data, helping us in our assumption that we will have a wide margin of error with our final model. We also include Graph 2 to visualize the transformed data.

**Graph 1: Price vs Discrete and Continuous Variables**

****

**Graph 2: ln Price vs Transformed Continuous Variables**

****

We note in both graphs that, considering the spread of the data, could possibly be modeled using a linear model. We have only the slight possibility ln of land value and living area could be a higher order model

We proceed to compare the models we’ve hypothesized. In Table 5 summarize the characteristics of our 4 models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 5: Summary of Models | | | | | | |
| Model | RSquare adj | Global F | S | MSE | SSE | # of Betas |
| 1 | 0.6501 | 186.71 | 58300 | 3.3989e+9 | 5.7169e+12 | 17 |
| 2 | 0.5810 | 139.56 | 0.29 | 0.086 | 144.68631 | 17 |
| 3 | 0.6205 | 127.28 | 0.279 | 0.0779 | 130.63856 | 22 |

Note models 2 and 3 are measuring the ln of price, thus the change in units for S, MSE, and SSE. We also note that each of these models account for roughly the same amount of variation in price, as denoted by the RSquare adj with Models 1 and 3 accounting for the most variation.

Note that we cannot complete a partial F-test comparing two models with different response variables, so we must compromise and extrapolate our inferences between the similarities in the various models.

Comparing Model 2 to Model 3 , we can perform a partial F-test to test if the higher order model provides more information.

We calculate the F statistic utilizing the columns of Table 5, which is found to be 36.07. Comparing that to the threshold 2.21, we can reject the null hypothesis and conclude that the added terms in Model 3 contribute more information than Model 2.

We consider the differences between Model 1 and 3, as we do not have a statistical test to compare these models. Model 1 utilizes a skewed response variable, which will violate an assumption of Least Squares Regression. We also see there is not much loss between the RSquare adj values. With the inclusion of 5 more beta values, we can feel confident confirming that Model 3 is the best predictor.

We proceed to determine if Model 3 is a good predictor as is. We sort the data values by Cook’s D Influence, and see no value exceeds 0.05. We can say there are no influential data values and can proceed.

We next proceed to verify the assumptions of Least Squares Regression. These assumptions being:

1) The mean of the residuals = 0.

* These models are built around this assumption.

2) The variance is constant for all settings of the independent variables.

* We can examine the residuals vs the predicted values to determine if the variance appears constant for all predicted values.

3) The residuals follow a normal distribution.

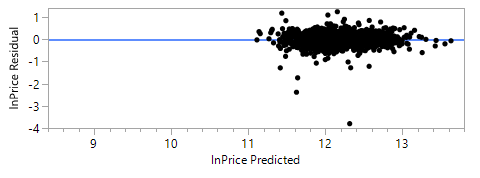
* We will examine a histogram of the residuals to observe normality.

4) The residuals are independent.

* Each sale is unrelated to any other, so this can be assumed to be true.

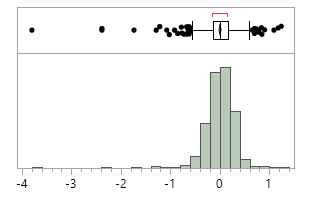
For point 2, we provide Graph 3 of the residuals to examine if it is reasonable to claim the variance is equal. We can clearly see that the residuals seem spread at random with no observable pattern of tapering, jumps, higher order models, trigonometric models, etc. There is the exception of a few outliers but note we have already claimed these outliers are not influential. The majority of the 1700 data values lie within 2 Std. Deviations of the mean, between 11.4 and 13 ln(price). We can say this assumption is true for this model.

**Graph 3: Model 3 Residuals vs Predicted Values**



For point 3, we check the normality of the residuals by providing Graph 4, and what is shown is a strong bell shaped curve, with a mean of 0. We can say this assumption is true as well.

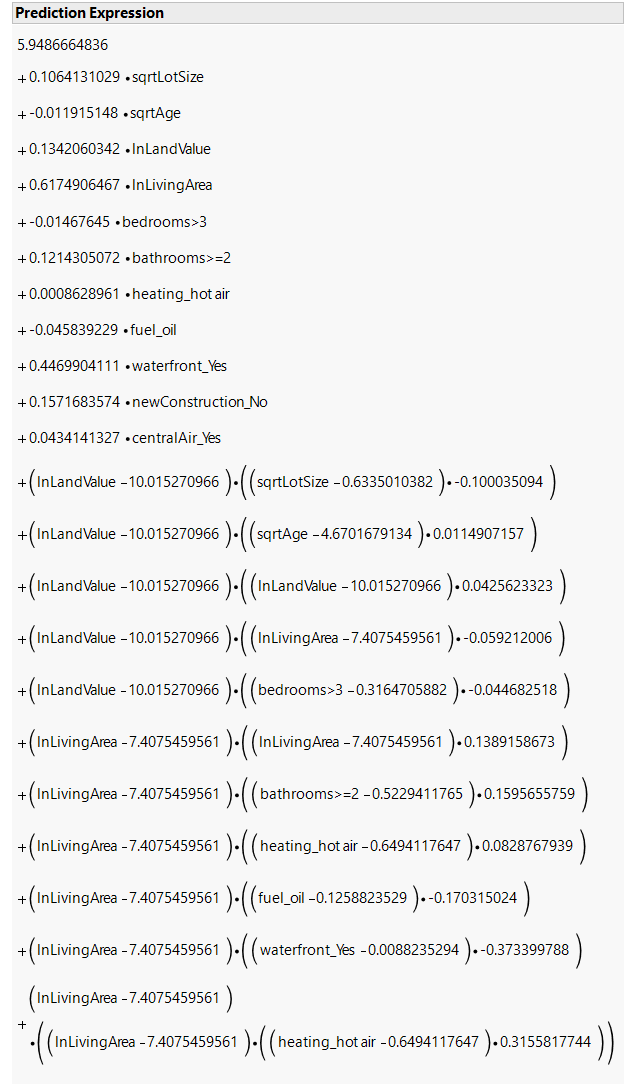
**Graph 4: Residuals Histogram**



We continue to prove the validity of the model. As mentioned prior we used stepwise regression with a stopping rule of Maximum RSquare K-fold. Ideally, we have a model RSquare Adj similar in value to the RSquare Kfold to signify that regardless of which old, we are predicting other values within the data set reasonably well. Model 3 RSquare Adj = 0.6205, with RSquare K-Fold = 0.6138. With less than .01 difference between the two, we conclude this model will predict as well as the RSquare Adj says it will.

With this, we can say Model 3 is our best predictor of price. We expand on this and provide the parameter estimates in Figure 1.

**Figure 1: Parameter Estimates**

****

We remind that this model is measuring the ln of price and requires reverse transformation for prediction. Each of the parameters can directly revert to their original output, allowing easy translation to what this means for the output.

Here we examine what these estimates mean in the context of this situation. The intercept coefficient is meaningless and is simply a correction factor. Using square root of lot size as an example, if all else is held constant, each additional square root acre will increase the price prediction by e raised to the parameter estimate of 0.106. Similarly for ln of land value, a $1 increase in the ln of land value will increase price by adding the parameter estimate of 0.134 to the exponent of e. For if the property is fueled by oil, the price will decrease by subtracting the parameter estimate of 0.0458 from the exponent of e, the default level is if the property is fueled by gas or electric. Similarly for waterfront properties, if the property is waterfront, the price will increase by adding the parameter estimate of 0.447 to the exponent of e.

Interpreting the interaction and higher order terms is the crux of this model, but we continue. For the square of ln living area, each square foot increase in ln of living area will increase the curvature of price slowly by the parameter estimate of 0.139. We interpret the parameter estimate for the square of ln living area multiplied by hot air heating as such: if the property has hot air heating, the curvature of price will further slowly increase for each additional square foot of ln living area at the rate of the parameter estimate of 0.3156. Further examining parameters in context will prove redundant, as reasoning is shared among similar parameters.

We admit the added headache from interpreting these parameter estimates might be cumbersome, but their inclusion increases the RSquare adj by almost 7%. We once again remind that the transformation of price was necessary to fit this model using least squares regression. Comparing Model 3 to Model 1 is a non-issue as Model 1 breaks an assumption of least squares regression, though it does have slightly higher RSquare adj.

**Conclusion**

After analyzing, we found that “the housing price” is the most predicted variable on this data. From the above Table 1: In the depicted table we see that histogram of each Price, lot size, age, land value, and living space all exhibit significant skewness, as do other variables. To standardize the data, we convert each variable, preferring to apply a natural log transformation, if possible, but keeping track of the lot size is important as well. Some of the values for age and lot size are zero. The use of a ln transformation will result in undefined data, thus avoid doing so, For these two variables, we choose a sqrt transformation to check this transformation output from the

Table 2 has a variable for summary transformed from table 1 We construct a table of summary statistics and graphs displaying price versus each independent variable to better understand what order model we should apply. The absence of discontinuity in the graphs indicates that piecewise regression is unnecessary. We also use stepwise regression to determine which factors have the greatest impact on prediction accuracy. To use interactional knowledge, we look at our own perceptions of the current circumstance as a starting point. However, our data shows how widely dispersed this research is. It is possible to see the relationship between residuals and predictions using this simple model. The variation in residuals is often even, and we notice this. In the end, we cannot use this model for comparison since the skewness of the response variable breaches the equal variation condition for least squares regression.

Is it possible to increase the prediction power of our second model by normalizing all of its independent variables which was insight and which we got a result for the a higher-order model for these parameters is necessary since we see that land value and living area ln have the greatest impact on the models' ability to forecast.We use stepwise regression with a 5 Max K-fold R Square stopping rule to generate a more succinct model because of the large number of variables. The ln transformation on price must be used throughout this method to maintain all of the assumptions for Least Squares Regression as true.

Table 3 shows the independent variables as histograms to help us visualize the data. Tables like this one show the original data, as well as the data that was changed. Several of the variables seem to be highly skewed. As a result, we've created the variables shown in Table 2 to try to standardize the data. \* These histograms exhibit numerous outliers in their respective data space, which will impact the prediction potential of each model hypothesized. These outliers aren't helped much by the adjustments that are made.

Additionally, we have included histograms of the dummy variables and noted that beachfront and new construction residences are not as common as we think. For these reasons, most residences have already been constructed, and there isn't nearly as much water to build around. In Table 4, you'll see histograms like this.

***We may do a partial F-test to see whether the higher-order model delivers more information by comparing Model 2 to Model 3.***

***According to hypothesis H0, Model 3 provides no more information than Model 2 does.***

***The Model 3 provides more data than the Model 2.***

***Table 5's columns are used to generate the F statistic, which is 36.07.***

***We may reject the null hypothesis and conclude that the extra words in Model 3 give more information than Model 2.***

***Because we don't have a statistical test to evaluate these models, we focus on the differences between Model 1 and Model 3. Skewed response variables are not allowed in least squares regression, hence model 1 will not work. Additionally, we can observe that there is little difference between the adj values of R Square. We can confidently conclude that Model 3 is the best predictor with the addition of 5 extra beta values.***

***Our next step will be to see whether Model 3 is accurate as is. Using Cook's D Influence, we find that no data value surpasses 0.05. no significant data values can be found; thus, we come to know all the models have a data representation that is different, but output has Same value to ensure either the data represented is accurate in case if we follow multiple methodologies also.***