Cody Olivotto, Adeeb Salim

Professor Scrogin

October 24, 2021

Real Estate Model Evaluation

**Objective:**

For this analysis, we set out to determine the model with the lowest mean squared error, more commonly referred to as MSE. Our goal is to create a reusable model that can take in several exogenous variables and use them to predict the value of the endogenous variable.

**Process:**

First, we used SQL to create a database from the separate CSV files. We then used the join function to combine the different elements of each CSV and then exported them to a new CSV file named Sales.CSV. The Sales CSV originally included one endogenous variable, Price, and eight exogenous variables were: Year, Home size, Parcel size, Beds, Age, Pool, X\_coord, and Y\_coord.

    Next, we imported the Sales data set into Python, where we began the process of model estimation. As a starting point, the data was randomized, and intertools were used to determine an exhaustive list of all possible combinations of exogenous. After that, empty dictionaries were created to store the variable combination of x as keys and the MSE’s as values.

    Once the data is prepared, and the structures are created, the process of estimating the models begins. The Ordinary Least Squares, or OLS, method was the first to be conducted. A for loop was created to iterate through all the different possibilities. Inside of the loop, the variable combinations were called in conjunction with polynomial features with a parameter of 2. To circumnavigate the issues with Polynomial Features not retaining heading names, we wrote a loop named ‘polypairs’ that ensures the exogenous variables are stated in a clear and understandable manner. Linear Regression is then running inside the loop, with normalization turned on. Normalization reduces all values to a normal distribution. After the model is called, it is then cross-validated using a cv of 10 and is set to return MSE, r2, test, and training parameters. The combination that provides the lowest MSE is stored in the previously created dictionaries and then stores the associated MSE value.

   Following the completion of OLS, we then repeated the process to include Ridge and Lasso regressions. Much of the code is the same for all three, but ridge and lasso have a few key differences. Ridge and lasso, both take a tuning parameter, which is denoted as alpha. Prior to the loop, alphas are set to a normalized range provided by linspace that has 100 different values for the tuning parameter. Inside the loop, another loop is nested to ensure that the regression iterates through all the alpha values when estimating the model. Additionally, the Lasso function has a unique parameter in this case, which is maxed iterations. Max iterations were set to 100,000 to set a limit to the time it takes for the computer to crunch the data.

   Lastly, we hardcoded the appropriate values from the optimal model estimate that provided the lowest MSE and reran it with the full sample to ensure results. Upon completion, the full sample and the cross-validated sample were placed in an array and exported to a CSV file to create the tables in the latter part of the document.

**Table 1 Results:**

From the 3-model estimation and cross-validation, we determined that linear regression is the superior model.

| Model | Exogenous | Alpha | Train MSE | Train R^2 | Test MSE | Test R^2 |
| --- | --- | --- | --- | --- | --- | --- |
| Linear (OLS) | Beds, Year\*Beds, Parcel\_size\*Beds, Home\_size\*\*2 | NA | 11022.37 | 0.7395 | 12006.77 | 0.6881 |
| Ridge | Year, Home\_size, Parcel\_size, Beds, Age, Y\_coord | 49 | 17773.97 | 0.5815 | 18508.72 | 0.5671 |
| Lasso | Year, Home\_size, Parcel\_size, Beds, Age, Pool, X\_coord, Y\_coord | 49 | 22331.60 | 0.4735 | 22983.96 | 0.5189 |

**Full Set Table 2 Results:**

These Results confirm that linear regression is the most optimal model given the dataset and exogenous variables. Shown below

| | Full Sample Test | | | | --- | --- | --- | | Exogenous: | MSE: | R^2: | | Beds, Beds\*Year , Parcel\*Pool , Home\_size^2 | 10711.2783 | 0.7476 | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |
|  |  |  |

**Findings:**

In conclusion, the model we created that has the lowest MSE and thus the greatest potential for future usability in predictions is the Linear Regression  Model. The optimization process has enabled the removal of variables that weaken the prediction capabilities like the coordinates. Additionally, the interaction between variables is taken into account, for example, it makes sense that pool and parcel size would have some interaction since pools commonly lie inside of the parcel at most residential locations. Similarly, beds and year may have interactions that stem from a decrease in average family size across much of America over the last 20 years. During the process, it became clear that certain models have a bottleneck, and may only run on one core, which causes them to take longer to crunch.