Final Assignment Part 2

MSCA 31007 - Statistical Analysis

By: **Paresh Raut**

9th December 2022

In this report, I continue to examine and analyze real estate purchases within the borough of Brooklyn from 2016-2020, provided by the City of New York. In this Part of the final assignment, I will specifically try to estimate how prices changed from the third quarter of 2020 (Q3) to the last quarter of 2020 (Q4). The model I built in part 1 will be used as the base model here.

***Did the prices change?***

After creating a linear model with the square root of prices as the response variable and the following explanatory variables: neighborhood(factor), land square feet, square root of gross square foot, age (how old the property is), block, lot, zip code (factor), building class category (factor), quarter (factor - which year and quarter was the property sold in). The regression estimates tell us that there was a **significant increase of $2402.96(+ $175**) **in house prices** going from **Q3 2020 to Q4 2020**. This finding aligns with the U.S. House price index report – 2020 Q4 by the federal housing finance agency (FHFA HPI). The reason for the same cited by the FHFA HPI was *“Low mortgage rates, pent up demand from homebuyers, and a limited housing supply propelled every region of the country to experience faster growth in 2020 compared to a year ago despite the pandemic. In particular, house prices in western states and cities saw the highest rates of growth, where annual gains often rose above 10 percent."*

***How I selected the model***

Firstly, before building a model I followed data cleaning and filtered prices only for houses and not for theatres, commercial garages, factories etc. I made sure that I select the relevant data for my response variable, and so I only considered relevant prices such as non-zero prices and house prices that were not too low. Further, I wanted to include neighborhood and zip code for locality as a factor in my regression model and so I created 10 bins for them based on the mean prices in each neighborhood and zip code. I also wanted to consider how old the property was (in years) hence, created a column ‘age’. Age is the difference between the year property was built and the year of sale for the particular property.

In trying to find the change in prices from Q3 to Q4 in 2020, I had to make sure I consider all factors that might affect the prices(to control variation due to factors affecting price) such as land square feet area, gross square feet area, which tax block and lot does the property lie in , any previous seasonality in prices from the previous 4 years (this was taken care of by taking data from all previous 4 years and 2020), the building class category(as a factor) and of course what quarter and year is the observation from (as a factor). Also, to keep my predictions fairly accurate I had to make sure not induce bias by considering a small amount of data and therefore, kept number of observations for the model 13,494.

I chose to keep a good number of factors in my model to promote variability since OLS regression models like the variability in its factors.

***Model Performance***

Prior to running the model, I specified the reference category level to ‘Q3 2020’ to make it easier to measure the change in house prices between Q3 and Q4 of 2020. Here, this would simply be the square of the coefficient estimated for Q4 2020 in this model. Next, I ran the model and found the following:

1. Adjusted square error was 0.6746
2. RMSE of my untransformed price 123410.5
3. The coefficient for my factor ‘quarterq4 2020’ was 49.20, the standard error was 13.23 and was marked as significant. (We must consider square of these values since model was computed on square root of prices)
4. I plotted the residual v/s fitted values and the QQ plot below:

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

***Reliability and limitations to my model***

From the above plots we can certainly say that there was bias in the residuals for very high prices and very low prices. This might be because we do not have a lot of observations where prices were very high or were very low. This limits our model from asserting anything clearly for the very high-priced houses and houses with very low price. Further to confirm the same I ran a k.s. test for iid - normality of residuals and a brush-pagan test for checking if the model is violating the rules of OLS regression. The results of these tests show that the residuals are not iid-normal and there is heteroskedasticity in the model as well, asserting that this model is not very reliable.

This model was built primarily focused on finding if there was any change in the house prices from Q3 2020 to Q4 2020 about which we may have some idea about but unfortunately this only establishes associativity between change of house prices and change of quarters and does not identify causality behind the prices changing from Q3 to Q4 in 2020.

The model’s reliability is influenced by the selection of the exploratory variables, the transformations done on variables and the modeling choices I made in this analysis. If the model was not to select neighborhood or zip code as a factor in predicting house prices then we might be miss out a major reason which might be affecting the price of the house, which is the location. Also, we need to include how old the property is since that could be a reason If older houses are costlier or cheaper.

***Conclusion/ Key Takeaways***

There has been an increase in house prices in Brooklyn from Q3 to Q4 in 2020 but we cannot identify any reason as to why did the prices change. The model has a low RMSE and a fairly high adjusted R^2 value which suggest that this model can fit the dataset the best out of the potential models.

I also able to identify the following things from the model summary:

1) Brooklyn heights(neighborhood) had one of the most significant changes in house prices.

2) 11201 (zip code) had one of the most significant changes in house prices.

3) 12 condos-walkup apartments (building class category) had one of the most significant changes in house prices.

4) I discovered that house prices rose from the fourth quarter of 2019 to the fourth quarter of 2020. The probable reason for this might be because 2019 being adversely hit by the Covid-19 pandemic led to no homebuyers in that year and hence saw a faster growth in 2020.