## Introduction

The goal of this project was to create a regression model for housing prices in the Los Angeles area. It was my introduction into time-series datasets, and how to introduce time-fixed effects into a regression model.

## Dataset

The dataset was a merged set of real estate transactions in the LA area during the first months of 2016. The datasets are: residentialsales1 and residentialsales2. They come from my Applied Econometrics course at the University of San Diego. I merged the datasets on the sr_unique_id variable. The resulting merged dataset had 5080 entries with 17 variables.

## Analyzing the Dataset

The first step I took was looking over the data to form a high level understanding. First, I counted the sales that took place in each month, finding that 1565, 1497, and 2018 sales took place in January, February, and March, respectively. The importance of this was the even distribution, ensuring that there weren't any outlier months. Then I counted the number of distinct zip codes within the data, and the number of distinct zip codes within each sale month. There were 111 distinct zip codes, with 108, 109, 109 distinct zip codes in the months of January, February, and March, respectively.

Next, was removing any outliers. This included any houses that were less than $200,000 (of which there were 46).

Finally, I created a descriptive statistics table.

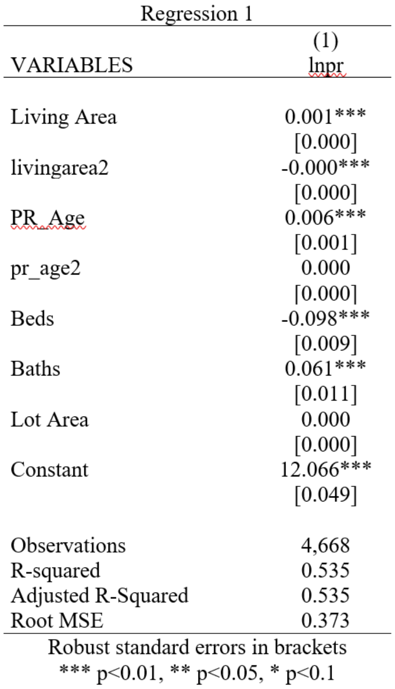
|  | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| VARIABLES | N | mean | sd | min | max |
| Sold Price | 4,656 | 742,860 | 582,896 | 200,000 | 9.605e+06 |
| Beds | 5,034 | 3.307 | 0.950 | 0 | 9 |
| PR_Age | 5,002 | 58.12 | 20.85 | 1 | 132 |
| SaleMonth | 5,034 | 2.089 | 0.836 | 1 | 3 |
| SaleYear | 5,034 | 2,016 | 0 | 2,016 | 2,016 |
| Living Area | 5,034 | 1,886 | 981.0 | 0 | 11,759 |
| Baths | 5,034 | 2.310 | 1.080 | 0 | 10 |
| Lot Area | 5,008 | 11,286 | 22,629 | 1,528 | 971,827 |
| Year Built | 5,034 | 1,946 | 157.1 | 0 | 2,016 |

## Building the Regression Model

Since the goal was to predict the housing price, which is a continuous variable, I used a regression model. The first model I created used the following variables:

lnpr livingarea livingarea2 pr_age pr_age2 beds baths lotarea where lnpr is the natural logarithm of the sales price, livingarea2 is livingarea squared, and pr_age2 is the property age squared.

The results are displayed here:



Regression 1

| VARIABLES | (1)<br>lnpr |
|---|---|
| Living Area | 0.001*** |
| | [0.000] |
| livingarea2 | -0.000*** |
| | [0.000] |
| PR_Age | 0.006*** |
| | [0.001] |
| pr_age2 | 0.000 |
| | [0.000] |
| Beds | -0.098*** |
| | [0.009] |
| Baths | 0.061*** |
| | [0.011] |
| Lot Area | 0.000 |
| | [0.000] |
| Constant | 12.066*** |
| | [0.049] |
| Observations | 4,668 |
| R-squared | 0.535 |
| Adjusted R-Squared | 0.535 |
| Root MSE | 0.373 |

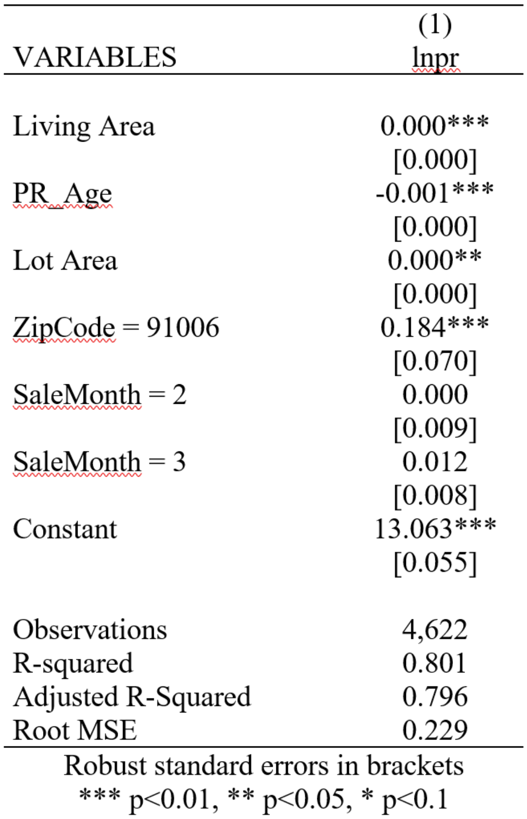
Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Looking at the table, I saw that the lotarea and pr_age2 were not significant at the 5% level. I decided to drop pr_age2, but to keep lotarea, since it is a relevant variable in the context of house price.

Since it was a time series dataset, I added dummy variables for the sales months of February and March for the next iteration of the regression. I also added the ZIP code dummy variable, since ZIP codes could describe the desirability of the location, such as proximity to the beach, etc. Here are the results of that regression:

The final step was to ensure that there was no multicollinearity among the variables. I used a correlation matrix and a VIF chart to analyze any correlation across the variables. I found the variables of livingarea, baths, and beds to be highly correlated. Removing baths and beds lead to the final regression model.

The final regression is here:

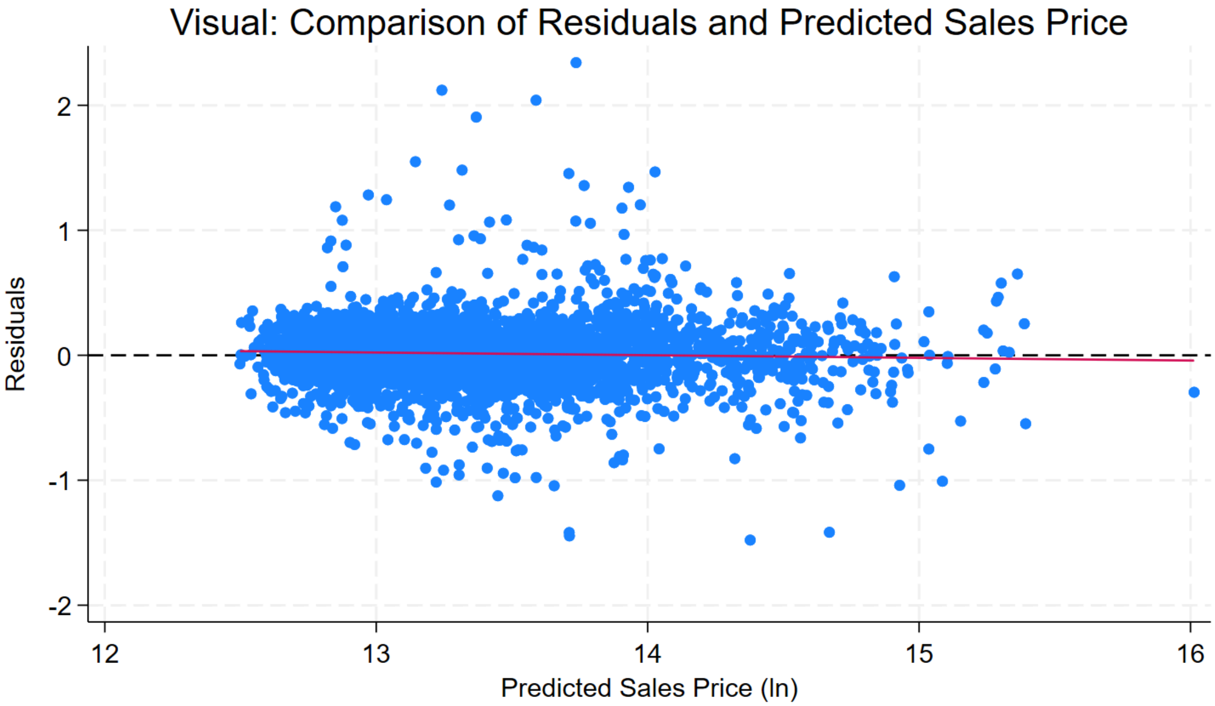


| VARIABLES | (1)<br>lnpr |
|---|---|
| Living Area | 0.000*** |
| | [0.000] |
| PR_Age | -0.001*** |
| | [0.000] |
| Lot Area | 0.000** |
| | [0.000] |
| ZipCode = 91006 | 0.184*** |
| | [0.070] |
| SaleMonth = 2 | 0.000 |
| | [0.009] |
| SaleMonth = 3 | 0.012 |
| | [0.008] |
| Constant | 13.063*** |
| | [0.055] |
| | |
| Observations | 4,622 |
| R-squared | 0.801 |
| Adjusted R-Squared | 0.796 |
| Root MSE | 0.229 |

Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note that I didn’t include every ZIP code here for readability. The full results can be found in the google doc.

Using a VIF and a correlation matrix, I saw that there was no multicollinearity in the model. Plotting the residuals, I concluded that there was no heteroskedasticity, as seen here:



And with that, I had created a regression model with an r-squared value 0.801 to predict the price of houses in the Los Angeles area.