Final Project - CEE 218X

Rex Shen

November 2021

1 Introduction

In the urban setting, housing prices have become a hot topic of interest, since it continues to impact the quality of life around the Bay Area. In the 1990s, the median price of a home in San Francisco was approximately around $300,000 [1]. In the 2000s, median house prices rose to $500,000. Yet, it wasn't until the likes of Twitter and Facebook, i.e. the Silicon Valley Boom, that skyrocketed the housing prices, where the median housing price is currently around $1,400,000 [1]. Importantly, the impact of higher housing prices remains vast. First, high house expenses mean that workers may need to live elsewhere in order to commute to their workplaces. Consequently, a longer commute time would imply using more gas, which would increase expenses in other categories besides housing. Second, high housing has a direct effect on the economy [2]. Particularly, places like Silicon Valley, which would like to hire talented individuals to join their workforces, would lose out on employees who find California to be an undesirable place to live. Ultimately, this warrants further research.

In the literature, many studies have found California to have regulations governing land use and residential construction [3]. As cities in California are free to set their rules independently, this encourages retail development over housing construction. Given the implications of housing prices, not only are low-income families affected by high housing prices but business owners are also affected by not retaining talent. In our report, instead of analyzing the impact of regulation, we analyze similar variables, such as unemployment rate, percent of students graduating college or higher, and number of total housing units. In particular, we utilize these particular set of variables to predict median housing prices for owner-occupied units across the Bay Area since a fundamental mechanism or underlying patterns exhibited may be classic supply-demand economics.

In this report, we illustrate key factors that contribute to underlying differences in housing prices across regions over time, particularly from 2013 to 2017. Note that there are lots of additional market forces that cause year-to-year instability that we will essentially ignore in this report. First, we run linear regression models using a combination of the three covariates: unemployment rate, percent of students graduating college or higher, and number of total housing units. Then, we run LASSO since many of the observations are correlated across different regions of the Bay Area. Finally, we interpret our results and suggest potential further extensions on this project.

2 Where We Got the Data

We obtained our results from the 2017 ACS 1-Year Comparison Profiles from the Census API. Particularly, we obtain the variables unemployment rate, percent of students graduating college or higher, number of total housing units, and median housing prices for owner-occupied units across the Bay Area between 2013 and 2017. We used PUMS to create the maps. Moreover, considering that we are using number of total housing units as a variable, we will not account for household weights in this report. Throughout this report, we assume that the results collected from the Census API are accurate enough for careful analysis.

3 Analysis & Interpretation of the R Shiny Plots

3.1 Map Plots

Note, we will only perform an analysis for 2013 since many of the years in 2014, 2015, 2016, and 2017 can be interpreted by similar means.

For percent that graduated college or higher, it appears that the highest percentages are those around downtown San Francisco. On the other hand, places on the outskirts of the main cities have lower percentages. For instance, the region relatively south of San Jose has one of the relatively lower percentages of $\approx 80\%$ for 2013. For the unemployment rate, it appears that the places with the highest unemployment rates are within the big cities, albeit in scattered places, such as a small subsection of downtown San Francisco. For the map of total housing units, we see that the places with the greater number of housing units are on the outskirts of the big cities, like San Francisco. For the map of median housing dollars for owner-occupied housing units, there is a significant greater median housing dollars in the regions between downtown San Francisco and San Jose. As illustrated in the introduction, this makes intuitive sense because cities encourage retail development over housing construction.

3.2 Univariate Regression

Again, we will perform an analysis on the univariate linear regression model for only 2013, as much of the analysis and interpretations translate to 2014, 2015, 2016, and 2017.

From the scatter plot, it appears there is a negative correlation between unemployment rate and median housing dollars for owner-occupied housing units, whereas there is a positive correlation between percent graduating college or higher and median housing dollars for owner-occupied housing units as well as between total housing units and median housing dollars for owner-occupied housing units. In terms of the residual plots, a common assumption is that the residuals should be approximately symmetric around the mean, with mean zero. Overall, in all plots, it appears that it is centered around zero and is symmetric. Hence, there is no need to perform some sort of log transformation or some other statistical technique to fix this non-existent issue in our scenario. For the linear regression summary results, we can interpret the coefficients as follows. That is, a one unit increase of the unemployment rate is associated with a decrease of \$59,568 median housing dollars for owner-occupied housing units, assuming everything else stays constant. Similarly, this interpretation can be extrapolated to other variables in the univariate regression case.

Moreover, we see that the coefficients are significant for unemployment rate and percent graduated college or higher, but not significant for total housing units as individual predictors, assuming our $\alpha = 0.05$. Finally, we can also interpret the R^2 as follows. That is, for unemployment rate, percent graduating college or higher, and total housing units respectively, 35.25%, 72.39%, and 5.608% of the variation in the median housing dollars for owner-occupied housing units can be explained by the variation in the individual covariates themselves.

Comparing the slope values year-to-year, we can look at the individual covariates. For unemployment rate, 2016 had the steepest slope in magnitude in the negative direction. For percent graduating college or higher, 2017 had the steepest slope in magnitude in the positive direction. Finally, for total housing units, 2016 had the steepest slope in magnitude in the positive direction. Overall, this does not suggest anything on the causal relationship between the individual covariates and the response since there are a lot of additional market forces that cause year-to-year instability.

3.3 Multiple Regression

If we select all covariates together and perform multiple regression, we see that percent graduated from college or higher has the only significant p-value in 2013, 2015, 2016, and 2017, whereas in 2014, both unemployment rate and percent graduating from college or higher had significant p-values if we assume $\alpha = 0.05$. Furthermore, we can interpret the R^2 and the coefficients similarly as in the univariate regression case.

3.4 LASSO

In this section, we train/test using a 75/25 split. We run LASSO by finding the optimal coefficients for all covariates on the training set and choose an optimal λ using cross validation. We evaluate our results by computing the test set MSE. This is computed as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2,$$

where n is the number of observations in our training set, $\hat{y}_i$ are our fitted values based on the coefficients fitted from the training set, and y_i indicates our observed response for all $i \in \{1, ..., n\}$.

Overall, we see that 2013 has the lowest test MSE and 2017 has the highest. However, this does not suggest that our LASSO performs better for 2013 compared to 2017. It may be due to the way we train/test split or that we simply have left out other important factors that would have been more insightful for modeling the response. Moreover, using the LASSO may involve some statistical bias that could lead to misleading results.

4 Conclusion

Overall, our results seem to suggest that unemployment rate is highly predictive of median housing dollars for owner-occupied housing units in the Bay Area, whereas education level can be considered moderately predictive, whereas total housing units may not be so.

However, further research is warranted since we could extend our research by adding additional factors since one of the main assumptions of this project was that we ignored lots of additional market forces that cause year-to-year instability.

5 References

[1] https://abc7news.com/tech-bubble-housing-crisis-bay-area/5419967/

[2] https://lao.ca.gov/reports/2015/finance/housing-costs/housing-costs.aspx

[3] https://pubs.aeaweb.org/doi/pdfplus/10.1257/000282805774670293