Evaluating Modeling Housing Construction by region

Naomi Giertych

Housing construction in the US depends on a variety of factors including economic growth, population growth, and the wealth of residents in an area. In this project, I explore how incorporating these factors affects the modeling decisions for estimating the number of houses built in a US county (FIPS region). I am predominantly interested in determining how accounting for average income and state-level effects affect the model chosen and what can be gleaned about impacts on housing construction predictions.

I use US housing deed and tax data obtained from CoreLogic, Inc. The deed data consists of a record for each time a property was sold with information on the parcel ID, US FIPS code, state in which the property is located, sale date, and the amount for which the property was sold. The tax data contains information on the parcel ID, FIPS code, and the year in which the building was constructed. For the purposes of this analysis I examine residential properties with a sale date after 1970 and remove maritime FIPS codes 57 through 98.[[1]](#footnote-0)

In my final dataset, I want the number of houses built, the average selling price of a house, and the county population by FIPS code and year. The average selling price of a house acts as a proxy for the wealth of an area. I merge the deed and tax data together by FIPS code and parcel ID. I then obtain the count the number of houses built by FIPS code and year and augment the dataset with zeros to account for FIPS code-year pairs in which a house was not built. Next, I calculate the mean selling price of houses within a FIPS-year pair and merge this with the number of houses built. One limitation of using the average selling price of a house as a proxy for the income of an area is two-fold. One, many areas do not have a house sold in a year or have very few, and two, I am assuming that the house sold was an average home for an area. To account for the first of the limitations, I linearly-interpolate the missing values between years that did have a house sold, and any additional missing values are borrowed from the first year in which a sell was observed. Finally, I merge county population data obtained from the Census from 1970 forward. The final dataset contains 80495 observations for 49 states, including DC.

Finally I explore modeling the number of houses built using a negative binomial model. I fit three models. The first accounts for the years with 6 splines, the sale amount, and population; and the second, adds an offset for the state the property is located. Both of these models use FIPS as the grouping variable for the covariance structure. The third model accounts for years with 6 splines, the sale amount, and the population of a FIPS code by year but uses a nested covariance structure between the states and the FIPS codes.

First, I determine the shape parameter, alpha, of the Negative-Binomial model. I only need to choose alpha levels for the first two models described above. The third model uses the same alpha as the first model.

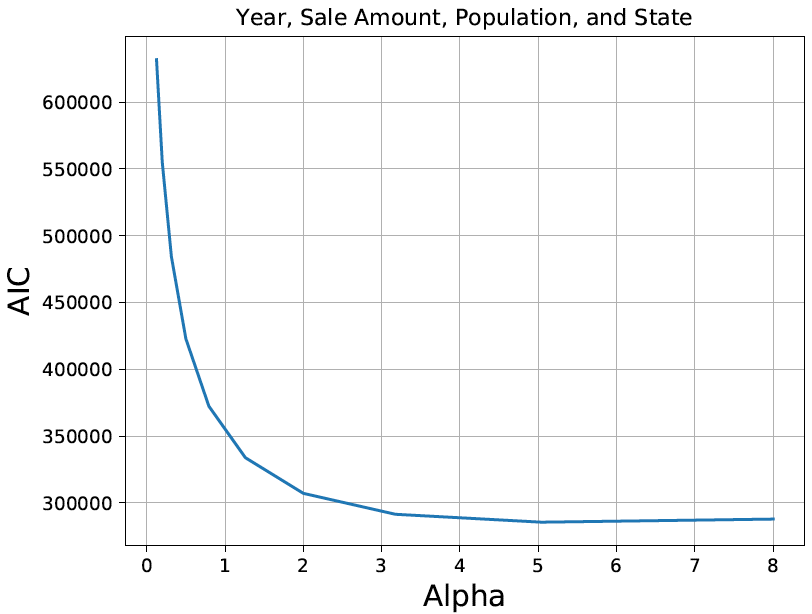
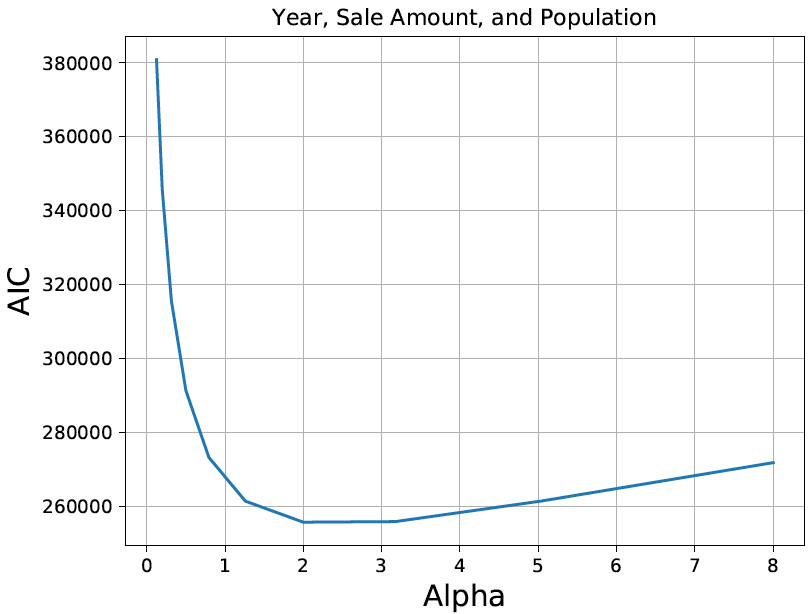


Figure 1: Comparing Alpha Values for the Models

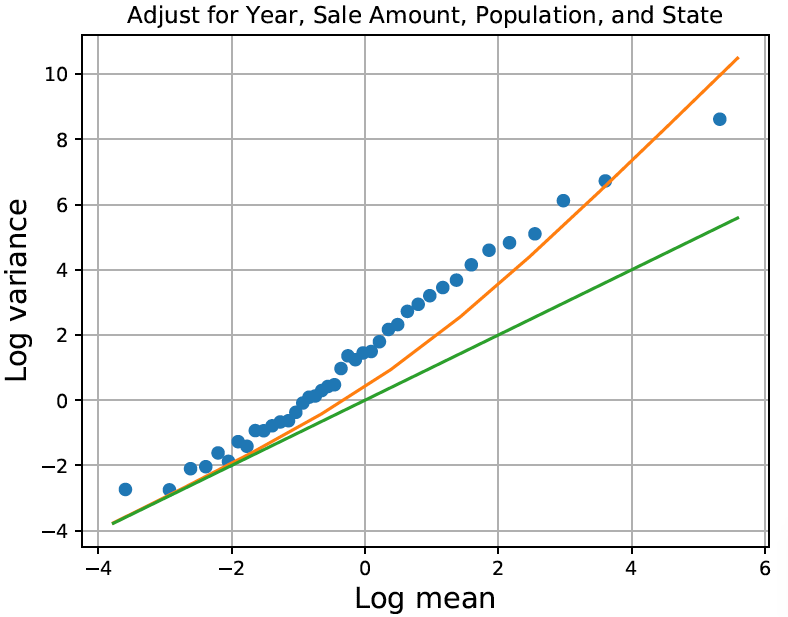
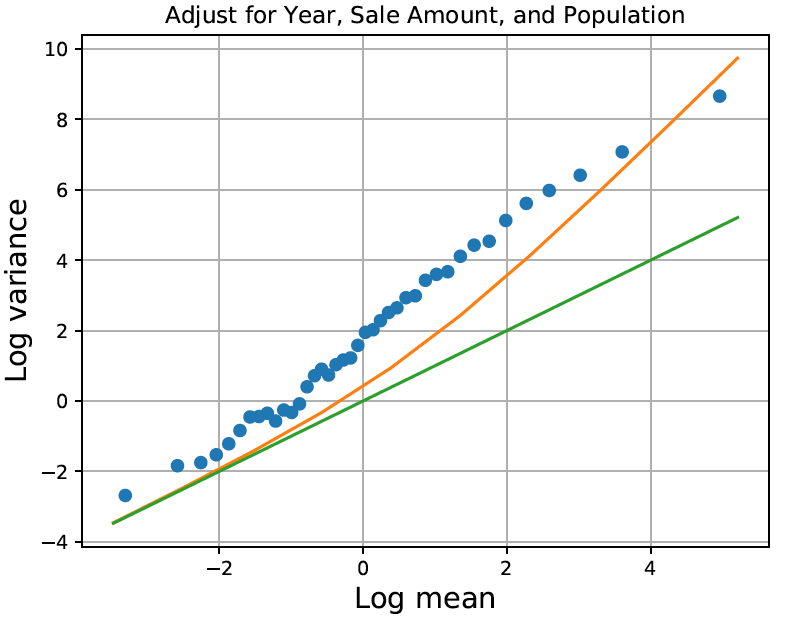
Based on the graphs above, I chose an alpha of 2 for the first model and last model and an alpha of 5 for the second model. Interestingly, when the state variable is added to the model, the alpha increases suggesting the mean-variance relationship is more quadratic. Below are the corresponding graphs of the mean-variance relationship.

Figure 2: Mean-Variance Relationships of models with sale amount, population, and state

Figures 2 shows that including a state offset does not change the mean-variance relationship very much despite what the alpha value suggested. Interestingly, the mean-variance relationship appears to be fairly linear which is also counter to the alpha values selected above.

Next, I examine the estimated autocorrelations from the models; below are the corresponding graphs.

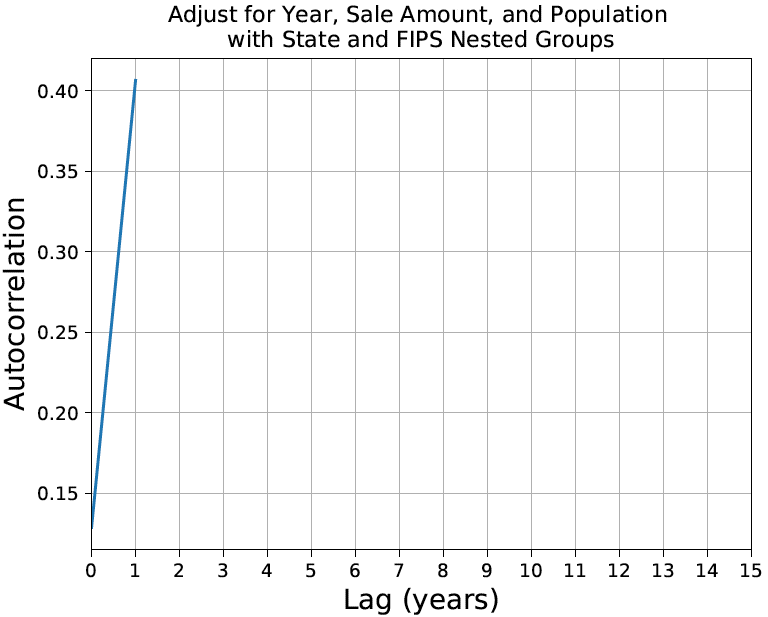
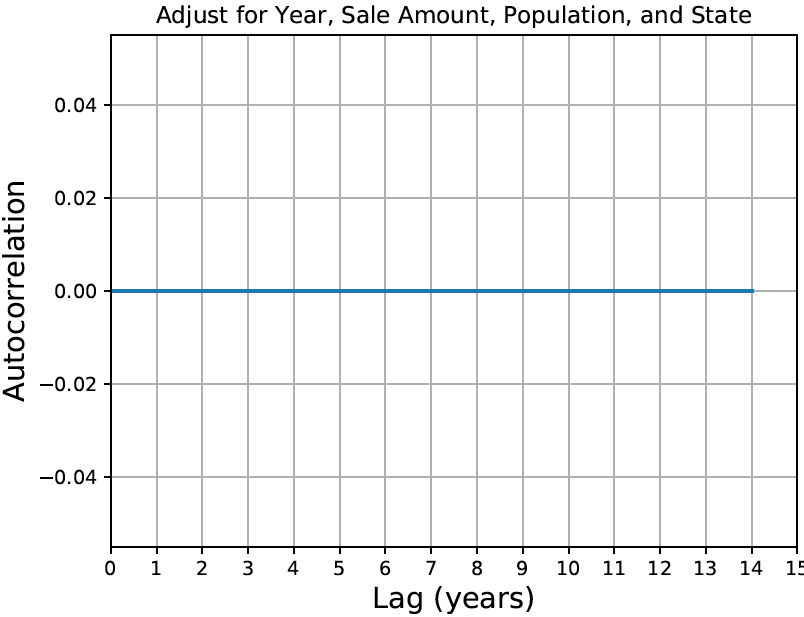
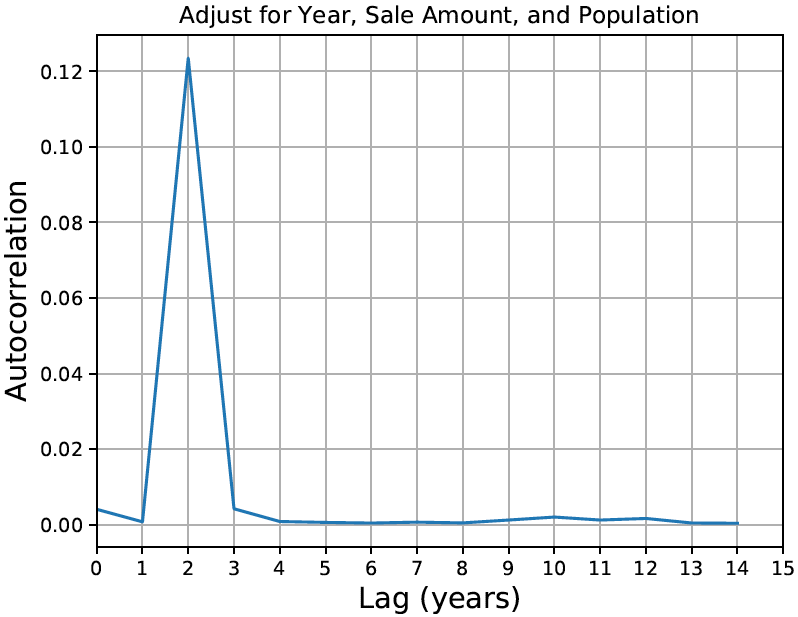


Figure 3: Comparison of Estimated Autocorrelations between the models.

As one would expect, including state in the model does not appear to change the autocorrelation between observations. However, it does drop to zero completely. The bottom graph in the figure plots the autocorrelation between the values for the model with the nested state and FIPS groups which shows a clear problem. Based on these graphs, it is appropriate to use a independence covariance structure for my two models with FIPS group effects.

Finally, I predict the housing construction for a three different population levels and sell prices for Michigan.

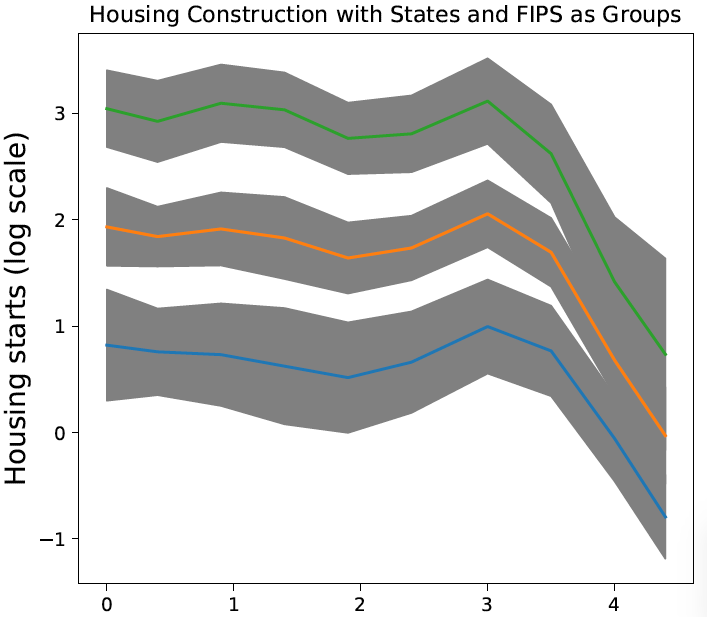
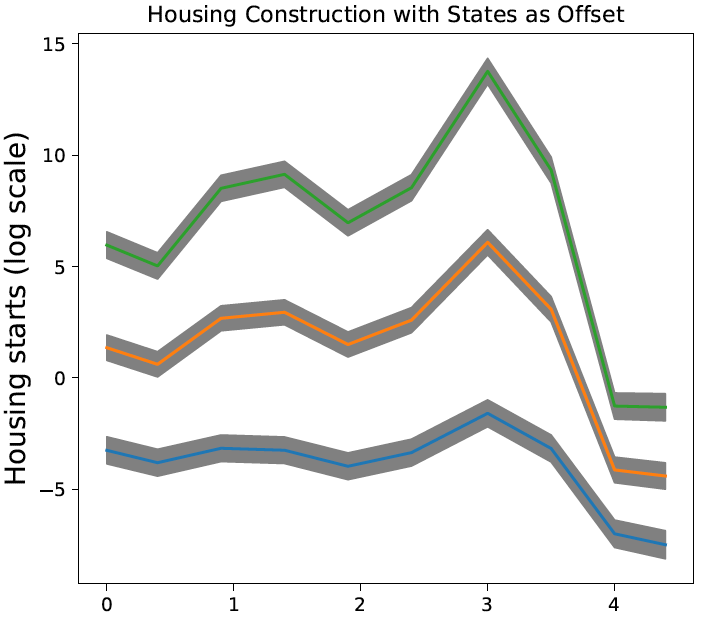
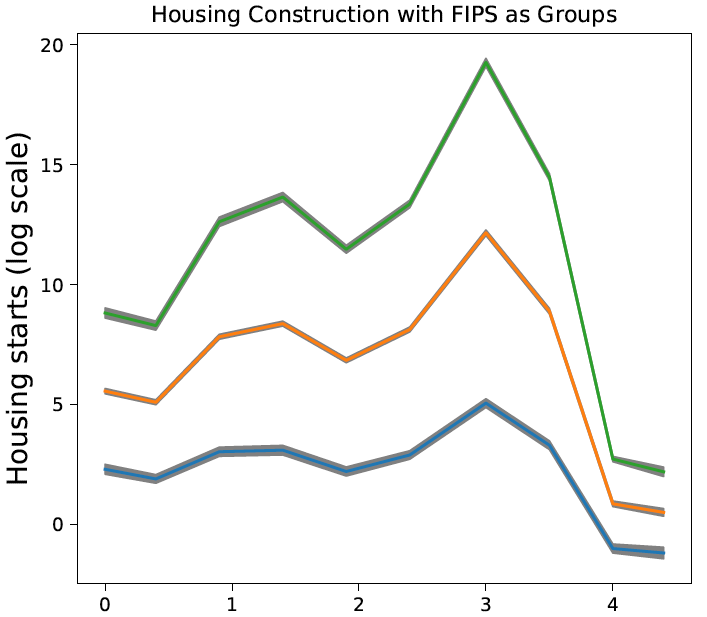


Figure 3: Predicted Housing Construction for Model 7 and Model 10, respectively.

As shown in the graphs above, the predicted number of houses built changes drastically between the models. The models with states as an offset or as a group also predict negative housing construction which is not reliable. Additionally, the uncertainty increases when states are included as an offset or as nested groups states with FIPS as a lower group. This indicates that any state-level macroeconomics that one might think drives housing construction is over-powered by FIPS-driven economics and population.

Sources:

Census Bureau: <http://www.nber.org/data/census-intercensal-county-population.html>

1. Missing state observations were filled based on the first two digits of the FIPS code and a FIPS code to state mapping created in the data. [↑](#footnote-ref-0)