Lolakshi Rajlakshmi

26 August 2022

Project Results

Using church appointments to predict population in England (1500-1850)

This data project uses church appointment data to predict population in pre-industrial revolution England. The underlying idea was that there should be more church appointments in areas with increasing populations and so appointments can be used as predictors for population. This involved web scraping the data from the Clergy Church of England database (https://theclergydatabase.org.uk/) and merging it with the provided population data. The balanced panel data created after processing was used to predict population for the test data through regressions and prediction algorithms. The results show that the random forest prediction works best for this dataset but there is room for improvement.

The first step involved using python to scrape a large set of pages identified by the location id of each church entity and then to process the data to create datasets that could be used after cleaning. This was the most time-intensive part of the project. A large number of webpages in the pre-specified ranges could not be actually scraped and some more were lost during the initial processing to make datasets (see log.xlsx). However, the code produced a significant amount of data.

The second step was to clean and merge all the different files (two for each cced_id) that were produced by the processing. This was followed by merging this data with the population data to create a final dataset for analysis.

The third step was to use different methods to predict population on a subset of the final dataset (the test data). The first method was to use regressions - ordinary least squares, fixed effects and ordinary least squares with log transformed variables) to estimate the regressors and then to use these estimates to predict the population values. The results of the regressions were contrary to what was expected, with the regression coefficient of appointments being negative. Log transformed variables performed slightly better with a positive coefficient as expected in all three methods, but were no longer significant. The regression coefficient for the fixed effects estimation were extremely large and possibly point to an error within the data or specification. When used for prediction, the log transformed variable models continued to perform better, but there was a significant difference between the real and predicted values.

The second method was to use prediction algorithms. Here, the results from the random forest algorithm were much better, and the closest overall to the actual values. The ridge and lasso algorithms performed quite poorly in comparision and had almost the same predictive power as OLS (without log transformations). This might reflect the fact that the relationship between the predictors and population was quite complex and was definitely non linear.

In conclusion, the best method to predict the population values was the random forest prediction algorithm. Although it worked better than the other methods, the results seem to indicate that there are other factors which, if included, could improve prediction , specially in the later time period. When working with regressions, log transforming the data improved results in some ways but was outperformed by the weak predictors of lasso and ridge algorithms.