Analyzing the relationship between English church appointments and population of English settlements in 1575-1800

Iana Fedorchuk *

Data Analysis course paper

Volkswirtschaft Fakultät

Ludwig-Maximilians-Universität München

Betreuer: Mathias Bühler, PhD

München, 26.08.2022

*Matrikelnummer: 12148235; Iana.Fedorchuk@campus.lmu.de

Contents

Introduction

The purpose of this paper is to describe and present results of the project: ”Analyzing the relationship between English church appointments and population of settlements in 1575-1800.”. Also to define, is the relationship exist at all, and what else does one need to do to find it out. The goal of the project is to analyze the relationship between English church appointments and population of the English settlements in the past, to see if any conclusions about significance of relationship can be made, to provide ideas for extension of the project and discussion of how the results of this or extended project can be helpful for other historical data based papers, or even for research that is based on the recent data observations. Also, to propose further research ideas that somehow correlate with the topic. To understand how to look at the bigger picture and to see more potential effects inside and outside this project.

The H0 hypothesis of this paper is: ”The English church appointments effected on population of English settlements in 1575-1800”.

To test the H0 hypothesis was created a data set that included: 970 cities in England, English church appointments for each city where two or more people were present. The time frame is 10 periods between 1575 and 1800 with a step of 25 years. The database stores data up to year 1835, but after working with the data (grouping and cleaning) the time frame was fixed to 1575-1800. The total amount of appointments counted in the final dataset is 45777. Average yearly count of appointments is 10.2. Not all the appointments were selected for the further estimation. Only appointments that had multiple people present, got in to the final dataset and statistics. I explain it the following way: as meeting a person that serves in church for most of the visitors is not the first experience of meeting them, thus this kind of appointments can not tell us anything, but also potentially add noise to the data.

The data source for appointments is The Clergy Database (CCED or Clergy of the Church of England database). The database includes English church appointments for each unique location that is indicated as CCED id and year of the appointment. CCED id indicates a part of an English settlement. Appointments in CCED are classified by the type of appointment. And for each appointment one can find a unique person id that was present.

The paper is organized as follows. Section 1 presents the overview data preparation process for further estimation and testing of the H0 hypothesis. Section 2 describes the results of estimated panel OLS models and regression models estimated with python machine learning library scikit-learn.

1 Data

This section describes the steps that were made to prepare the data set for testing H0 hypothesis.

The initial data contains two parts:

1) CCED appointments data. It includes unique location id (CCED id), year of appointment, type of appointment and person id and name that attended the appointment.
2) Population data. It includes unique location (CCED id), city id (C ID), year, latitude, longitude and population.

At first as data with appointments and complete location for an appointment list were in different files, I created a loop that merges each pair of appointment list and location.

Next, as the goal for the time frame was to have data grouped by year groups 1575-1825 with a step 25 years, at first I assigned to all appointments and population records years a year group. The algorithm was created the way that each year gets the closest year group assigned. For example, year 1823 was mapped to year group with value 1825 and year 1861 to year group with value 1850.

The next step was to fill missing values in population records table. At this step the missing values were filled with an average value of population between the preceding period and later period (for which data was available).

After, I merged the population records with appointments and their locations, based on the unique location id (CCED id), and later grouped the final table based on the year group, city, count of appointments with multiple people present. Only appointments that had multiple people present, got in to the final data set and statistics. As not meeting anybody but the person who serves in the church would hardly affect population rate. And likely the attendee of the appointment already familiar with the person from the church. Thus, in the data set were only included the appointments where at least two people were present (not counting the people, who serve in the church).

With all the described steps and more minor ones that can be found in the code, I received the final data set, which I already could use in empirical estimation.

The final data set included in total: 970 cities in England and 45777 appointments observed in 10 periods between 1575-1800 with step 25 a years. Average amount of appointments per year was 10.2.

On the Figure 1, is shown the relationship between population and English church appointments. The relationship is barely seen.

The correlation between church appointments and population overall is 0.0085, which is very low.

The annual correlation provides a slightly different results, and also shows that the corre-

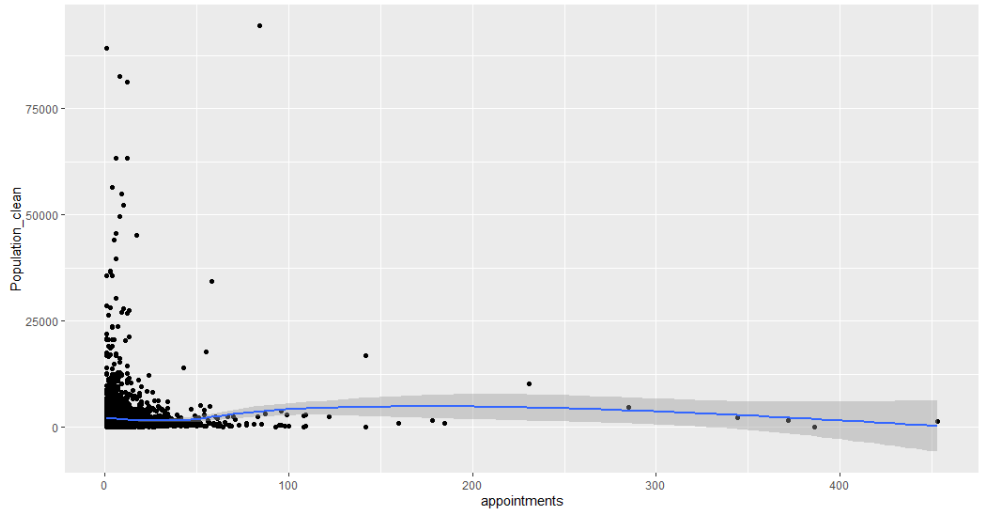

Figure 1 — Relationship between Population and English church appointments

lation change by time, which could be true, or caused by the a lot of missing data. Starting at the beginning of observed period with low negative correlation, and later changing to a small positive correlation.

year	correlation between population and appointments
1575	-0.06
1600	-0.1
1625	-1.0
1650	-0.07
1675	-0.05
1700	-0.06
1725	0.03
1750	-0.07
1775	0.2
1800	0.01

2 Models and Results

In this section presented the results of empirical estimation for testing the H0 hypothesis of the project for panel OLS estimation and scikit-learn regressor models.

For the first step of estimation were panel OLS models with Fixed Effects clustering and without FE clustering. The models were estimated witha help fro module PanelOLS from Python library linearmodels. Two specification of panel models were tested: normal regression and regression in logarithms

For estimating linear panel OLS in Python, the data should have MultiIndex that is created by combining city id and year group as a tuple for each row.

2.1 Results of Panel OLS models

The first model specification equation looks as follows:

$$population_{i,t} = appointments_{i,t} + city_i + \epsilon_{i,t} \quad (1)$$

The second model specification equation looks as follows:

$$ln(population_{i,t}) = ln(appointments_{i,t}) + city_i + \epsilon_{i,t} \quad (2)$$

The only difference between the first and second specification is that the second equation has log-log form.

The results of the first model specification showed: for the first model - with clustering city FE, the coefficient in front of appointments is statistically significant at less then 10 percent level. And for the second model without clustering city fixed effects, appointments coefficient has p-value=0.0003 which means that coefficient in front of appointments is statistically significant on less then 1 percent level. The point estimate of the coefficient in front of appointments for clustering and not-clustering models is the same 16.4.

The second specification model result showed that coefficient in front of ln(appointments) is statistically significant at less then 1 percent level for both clustering and not-clustering methods (p-value=0.00). The point estimate for both methods is also the same and equal to around 0.11.

Based on this result I can not deny that there might be effect of English church appointments on population, I would say this matter need a deeper look, and digging dipper into historical data.

Here I included analysis with clustering Fixed Effects in case the data heteroscedasticity or has correlation inside of the clusters.

As one can see there were only difference in p-values for the first model specification. And estimating the model in log-log format, almost didn't have difference in p-values and in point estimate, which mean that using log-log form for equation, helped with the heteroscedasticity problem within the clusters.

That is a not surprising result, even though the city fixed effects were included, the results show that most probably estimate has a big positive bias, because we have a problem of omitted variable. Population is a very complex, and a lot of different factors can affect it. And of course quite a number of factors were not included into the equation.

2.2 Results of Regressor models from Scikit-learn Python library

Next part of the empirical analysis was focused on predicting inside the data frame by using different methods from scikit-learn, machine learning library, in Python. Results from the four following regression models were estimated: Linear Regressor, Random Forest Regressor, Lasso regressor, and LassoLars Regressor. Only regression models were used, as both population and appointments data represents continues quantity values. Thus, classification models would be irrelevant in this case, as they are used for predicting categorical (or class variables).

The results of the models were rated between each other based on typical metrics that is used for comparing regression models:

1) Mean Absolute Error (MAE)
2) Mean Squared Error(MSE)
3) Root Mean Squared Error (RSME)

After receiving the metrics, the easiest metrics to compare results with was Mean Absolute Error. As the numbers are shorter from other metrics.

Listing of the MAE for the estimated models, sorted in ascending order. The lower the metrics the better. The same order of the scoring if one compares by MSE or RMSE in this specific case.

1) LassoLars Regressor, LARS stands for Least-Angle-Regression (MAE: 1601.82)
2) Lasso regressor (MAE: 1601.83)
3) Linear Regressor (MAE: 1602.86)
4) Random Forest Regressor (MAE: 1658.95)

The results show that the LassoLars model performed the best on predicting inside the data set. And Random Forest performed the worst from all tested models. Lasso and LassoLars MAE are very close. Scikit-learn library recommends to use linear or Lasso regression when the amount of rows in the data frame are lower than 100 thousand, which is the case for the data set from this research. The amount of rows in final dataset is 4394. Which pretty low for using advance methods like Random Forest. And the simpler methods like linear model and lasso model should manage to perform good on the final dataset from this research. LassoLars model is just a variation of Lasso model, that is why the results of Lasso and LassoLard are very close.

Conclusion

In conclusion, it would be good to say that for now with the current data available, one can not completely define the relationship between population and appointments. One can also argue that the appointments for the final data set should have been picked differently. But, the biggest challenge of this research goal that was just slightly tackled in this paper is the data has

too many missing values already, but also some explanatory variables that should be included in estimation equation is not available and never will be available, as the time period in the observed data was a long time ago. But still some of the explanatory variables can be added to the analysis. As a recommendation of expanding and advancing the analysis of church appointments effects on population, would be nice to do a deeper research of the literature and agglomeration and urban theory. New Economic Geography (NEG) is a theory that emerged in literature relatively recently suggests that agglomeration forces affect the growth of population in settlements. And one of the variables that is suggested by New Economic Geography is market potential of the cities. Also, to increase the precision of the analysis, one can look at the major events during the observable time period to confirm or deny if those events might affected the population in England. Also preceding the observable time period effect might have caused structural shift that still have effect that decreasing over time, for example. In this case even adding year fixed effects would not help. Another factors that can affect population are policy laws from the country or state level laws. Again, before doing any conclusions about relationship between English church appointments and population deeper and broader research and analysis should be made.

I understand the importance of the historical data analysis and making research about relationship between data that were might be present in the past. But as the last observation in CCED database is from year 1835. I can not help but wonder, how the current project can help us explain some relationship (if extended) now and in the future. What could be the equivalent or proxy to church appointments almost 2 centuries ago? Some counties still have a high portion of religious people, some countries have a large proportion of atheist people due to their different history. What nowadays the 'new religion' that would make people meet, start conversations and make new friends from it? How it will look like in future? And how often would people meet in future in close geographical proximity. What would be the next space where we are going to calculate distance? It would be an interesting research to find a similar kind of relationship but now, especially when more and more data is available every day. But interestingly before when doing research (not only scientific) people would carefully think about all the variables, data points, methods that they want to use, as each step of the research would take much longer time than now. While now people maybe put less thought into their data and methods, but the availability allows them to make mistakes and solve them faster.

References

1. Arts, Council) H. R. Clergy of the Church of England database. — URL: https://theclergydatabase.org.uk/jsp/search/index.jsp.
2. Hassink R., Gong H. New economic geography // The Wiley Blackwell Encyclopedia of Urban and Regional Studies. — 2016. — P. 513.