2020 Fall CS534 Final Project Proposal

Howard Huang, Sean O'Connell, and Sagar Soni

Introduction

An inverse relationship between education and crime seems straightforward and intuitive, but this effect needs to be supported across a range of peer-reviewed studies before substantial legislative action will be taken. Fortunately, studies with evidence for this relationship have grown over the past several decades and show a predictive inverse relationship in both directions[1-4]. These studies used a range of approaches and were performed across several geographical and demographic conditions, but all corroborate that increased education can reduce the probability of criminal activity.

Previous Studies

- Lochner and Moretti performed a study in 2004 that looked at the relationship between crime rates and education spending across the US. They categorized between whites and blacks, levels of compulsory attendance, and years of schooling over two separate models where years of schooling were the main independent variable of one and a dummy for high school graduation was the main independent variable of the other. All evaluation was done looking at aggregate statistics and all correlations came from linear regressions.[2]
- Gonzalez in 2015 published a similar study that looked across world education and crime rates where the main differentiator to previous studies was its differences between developing and developed countries. The range of study spanned from elementary through college graduation. Independent variables were graduation rate, youth unemployment rate, and level of economic development. Similar to Lochner and Moretti, all models were based on aggregate statistics and linear regressions to determine which variables were statistically significant.[3]
- McClendon and Meghanathan in 2015 focused on crime detection and prevention instead of relationship to education but used linear regression, additive regression, and decision stump algorithms in their evaluation. Linear regression came away as the best predictor of future violent crime patterns. They used mean absolute error, RMSE, relative absolute error, and root relative squared error as measures of predictive power.[4]

Motivation for Further Analysis

Together, the above studies indicate that education decreases the probability of incarceration and violent crimes on several scales: global, national, and individual reports. Although these findings can inform national budget legislation, they do not inform optimal budgets at the level of individual schools. For this, it is necessary to investigate these effects at the school district and county level and to identify educational spending patterns that produce these beneficial effects most reliably. Below, we describe our strategy for organizing the data and propose our approach for identifying spending patterns that are most impactful against future violent crime (i.e., show the most reliable negative correlation). We will use regularized linear regression methods to identify the most relevant features of spending and finally, compare across different time delays in the datasets to identify the most probable timeframe of these effects.

Data Cleaning and Organization

The two main datasets we will be using for this analysis are the "Annual Survey of School System Finances Tables" from the U.S. Census Bureau and a range of publicly available local crime statistics

recorded by the Georgia Bureau of Investigation. Fortunately, both datasets are subdivided into localities (counties and districts) that will enable matching of Georgia sub-regions across the two datasets. The education spending data, acting as our predictor features, span 1992-2018, while the crime data, acting as our target variable, spans 2009-2017. We will arrange all school budget data from several Georgia districts (predictor features) into a Pandas DataFrame in the following five-year increments: 2002, 2007, and 2012. The DataFrame will also contain all violent crime data from matched counties (target variable) from 2017. This will set us up for the regularized regression analysis described in the next section. We will perform the analysis three separate times, for each of these five-year delay increments to assess the effect of the time delay on the results and predictive quality.

Approach and Evaluation Metrics

We plan to use the data from the U.S. Census survey data for educational spending by district by year. This will then be used to find patterns within a time-delayed crime rate data for districts in Georgia. The aim is to predict the level of crime in a specific year due to the categorical education spending by a specific district. Since we have more features than we do data points due to the various types and amounts of spend and transformation that are possible for the feature compared to crime data being reported from 2009 to 2017 by the district in Georgia, we will be using a Lasso based linear regression on the base feature sets to set a baseline for the most important spend categories that are able to predict the crime rate in a given district for a given time delay. Once the accuracy, defined by R square and MSE, is benchmarked for a specific feature set, we will be conducting dimensional reduction to see if there are spending in multiple dimensions that can correlate to the crime rate target variable more accurately. We believe that in general, while 1 specific type of educational spending will not lower the crime rate. A mixture of spending across various categories that are linked (i.e. Teaching salary and yearly classroom supplies) and can show a correlation between increased spending and decreased crime rates.

To normalize the impact of population changes impacting the crime rate, we will be determining the crime rate per 1000 people. Additionally, the education spend will be normalized against the nominal GDP for Georgia which should eliminate some of the variability in the changing demographics of Georgians across the timeframe that we are observing.

While the changing demographics are a concern for the time-series data. The data related concern is that for the limited data, we could face an overfitting scenario with many of the techniques we will try due to the features >> data set. A success for the model will be to explain in *plain English* where and how districts should spend their education budget to decrease crime rates in a specified time delay.

Conclusion

In summary, we propose to address a gap in the literature that urgently needs to be filled: “How do spending patterns across local school districts influence local violent crime rates?” We propose a holistic analysis of spending patterns across several localities to identify the budget balances that produce the most reliable negative correlations with crime rates in the same locality. This analysis could produce valuable insights into which patterns of spending are most impactful on local crime rates, and if these effects are substantial, how long it generally takes for changes in the school budget to improve future social climate.

References

1. (n.d.). Education and Crime - Criminal Justice - IResearchNet. Retrieved November 03, 2020. https://criminal-justice.iresearchnet.com/crime/education-and-crime/
2. Lochner, L., & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review*, *94*(1), 155-189. https://eml.berkeley.edu/~moretti/lm46.pdf
3. Gonzalez, A. (2015). Education: the secret to crime reduction. *Unpublished thesis draft, New York University. politics. as. nyu. edu/docs/IO/5628/Gonzalez. pdf. Accessed December*, *22*, 2015. https://as.nyu.edu/content/dam/nyu-as/politics/documents/Gonzalez.pdf
4. McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1), 1-12. https://www.researchgate.net/publication/275220711_Using_Machine_Learning_Algorithms_to_Analyze_Crime_Data
5. Hastie, T., & Qian, J. (2016, September 13). Glmnet Vignette. Retrieved November 3, 2020. https://web.stanford.edu/~hastie/glmnet/glmnet_beta.html
6. School Financial Data. https://www.census.gov/programs-surveys/school-finances/data/tables.html
7. Georgia Crime Data. https://gbi.georgia.gov/services/crime-statistics