Data Science Final Project Report

Jacob Brown* and Michael Zoorob†

December 2017

1 Problem Statement and Motivation

Homicide is a matter of life or death. How does political context structure levels of homicide? Does the political influence of police unions appear to shape violence in metropolitan areas? In the aftermath of several well-publicized police shootings,some scholars have characterized a "Ferguson Effect", whereby low police-morale discourages police officers from engaging in proactive police work that could prevent crimes (Wolfe and Nix, 2016). This has coincided with sharp increases in homicides, after decades of declines, in some cities. Strikingly, this recent increase has been concentrated among a handful of cities, including St. Louis, Baltimore and Chicago, while other cities (such as Boston and New York City) have not experienced a recent uptick in homicides (Towers and White, 2017). We hope to shed insight into these recent developments by examining the predictive power of Fraternal Order of Police (FOP) lodge density – the very conservative police unions that represents the most police officers in the United States – and is the collective bargaining unit in several large cities like Chicago, Baltimore, Pittsburgh, Philadelphia, Nashville, Atlanta, and El Paso.

2 Introduction and Description of Data

We compiled and combined data from a variety of sources to shed insight into the phenomena driving patterns of homicide in the United States. Along the way, we encountered many of the usual problems of wide-ranging analysis – missing data, absence of common codes to link data together, and imperfect mapping of measurement of data onto concepts of interest. Nevertheless, the problem of homicide is so important – literally life or death – that it is important to build systematic knowledge about its determinants. Through the process of exploratory data analysis, we learned how to merge, cleanse, and navigate our data to start getting a handle on the potential determinants of homicide.

*jrbrown@g.harvard.edu; Department of Government, Harvard University
†mzoorob@g.harvard.edu; Department of Government, Harvard University

2.1 Data Sources

The data for our project comes from the FBI database on city-level crime rates, the United States Census and American Community Survey. Our outcome variable, Murder Rate, was gathered by scraping the FBI website. We merged this data with MSA Census data downloaded from National Historical Geographic Systems (NHGIS) and American Fact Finder. Our preliminary models rely solely on these two data-sources. We then supplemented this data with county-level presidential election returns for the 2008, 2012, and 2016 presidential election, and MSA-level counts of the number of Fraternal Order of Police lodges (achieved through geolocation of addresses obtained through Google Maps API).

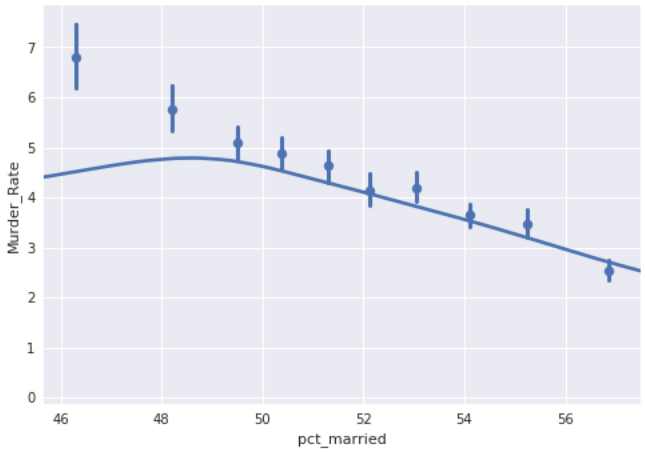
2.2 Data Collection and Wrangling

We first scraped the FBI murder rate data by MSA from the FBI website for 2006 through 2016. This data was subsequently cleaned and set into a data-frame. Next, we downloaded and merged a variety of US census and American Community Survey data-sets. These data-sets include information on total population, racial demographics, median household income, age demographics, percent receiving food stamps or other public assistance and marital status. After cleaning and merging the US Census data, we fuzzy string merged these data with the scraped FBI data using the *difflib* library, matching to MSA by name. This results in final data-set where each row is an MSA in a given year with matched variables from the Census and the FBI data.

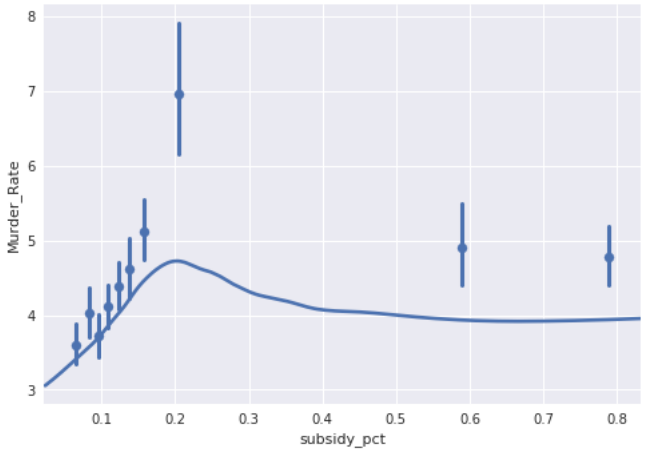
Next, we analyzed the data for mis-codings and measurement errors. We foundbsome variables with implausible values. We double checked our merging codebto make sure the error was not a result of our merging process. Given that the error is derived from the US Census coding process, these respective rows will be dropped from analysis. The total number of rows in question (n=4) is sufficiently small that it should not impact the predictive quality of our eventual model nor our overall results.

2.3 EDA

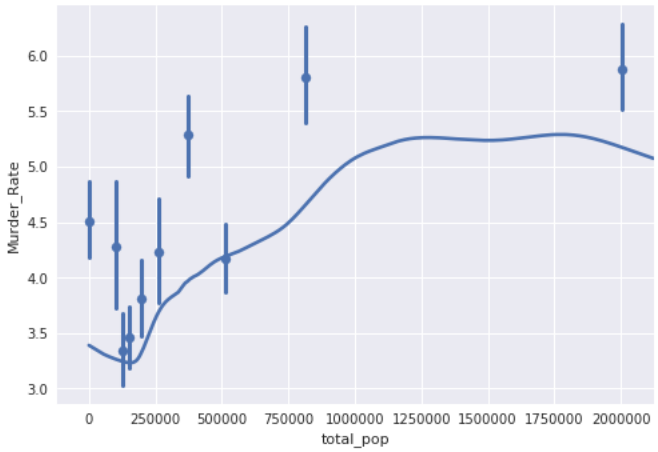
Our preliminary exploratory data analysis (EDA) focused on exploring the relationship between Census variables and FBI crime data. We used our EDA to examine what we had available to us via the Census. We were particularly interested in racial demographics, economic indicators, age, levels of public assistance, and familial data such as percent of population that was married. We wanted to establish a baseline of predictors that focus on individual-level characteristics of the population, before adding in our outside data-sources on policing and political context. The plots below depict the first pass relationships between Census predictors and murder rate.



This graph (a binned scatter plot with a loess line of best fit) shows a monotonic declining relationship between the percentage of married adults in an MSA and the murder rate. This graph suggests that marriage rate is a reasonably informative predictor of murder rate.



This graph shows a non-monotonic relationship between subsidy percentage and murder rate. While MSAs with very low percentages of public welfare provision have lowest murder rates, murder rates only increase with percent subsidy until about relationship 0.2% and changes sign to become negative for higher levels.



This figure suggests that MSAs with larger absolute populations tend to have higher murder *rates*.

Next, we explored other outside data sources, adding in Democratic vote share in the most recent presidential election to get a sense of political contexts effects on murder rates. We have research interests in the effects of policing and police unions on a variety of outcomes, so we included data from a previous project measuring the number of Fraternal Order of Police lodges (a proxy for measure of police union strength and police presence) to our data. This variable emerged as our main predictor of interest, because we believe it holds substantive political relevance related to our work.

3 Literature Review/Related Work

Ample research has demonstrated the role played by "social disorganization" – such as declining marriage, poverty, unemployment, and low education – on crime rates.(İmrohoroğlu, Merlo and Rupert, 2006) Recent scholarship attempts to empirically examine the Ferguson Effect by looking at police officer willingness to engage in community relations activities after negative publicity about the police department (Wolfe and Nix, 2016). Fear of crime is also an important cleavage in American politics (Enns, 2014).

4 Modeling Approach and Project Trajectory

We have estimated dozens of model specifications on these data, including OLS, Bayesian Automatic Relevance Determination models, Ridge Regressions, Lassos, and Elastic Net. For each model, we used k-fold cross-validation to select the model with the best test prediction. Here, we show primarily the model estimates and diagnostics of OLS Models (which are easy to interpret, but have relatively poor test prediction accuracy) and Elastic Nets (which are complex to interpret, but performed the best in our analyses). In this report, we outline our results in three stages:

1. Estimate models of first-order covariates on MSA murder rate.

- Use OLS to get a first-cut at the relationships between conventional demograhic variables and homicide.
- Utilize an Elastic Net procedure to shrink covariates that are not informative of homicide rates.

2. Incorporate State and Time dummies and interactions.
3. Incorporate FOP Lodge Density and Democratic Vote share.
    - Include polynomial terms and interactions.
    - Use Elastic Net to prune model features.
    - report Final OLS model coefficients.

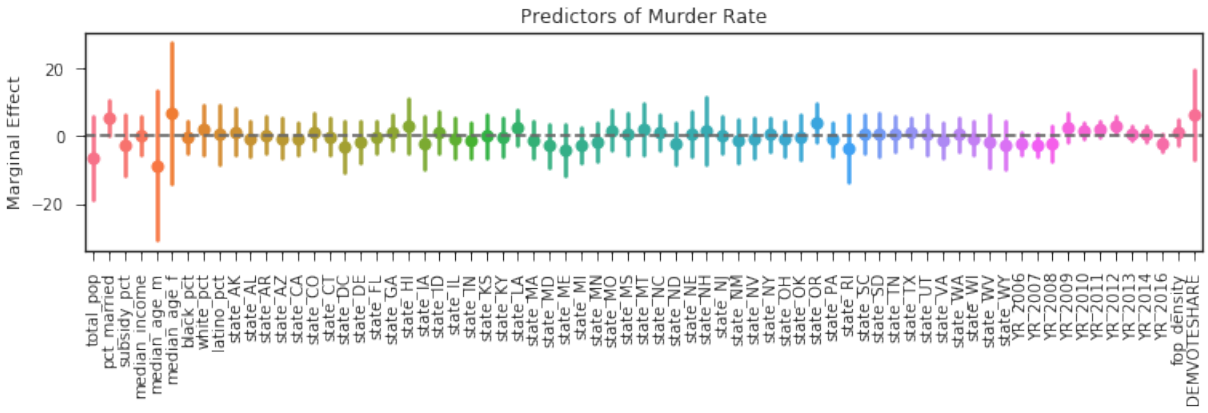


Figure 1: This shows the estimated marginal effects for the third set of model covariates (sans interactions) estimated through OLS.

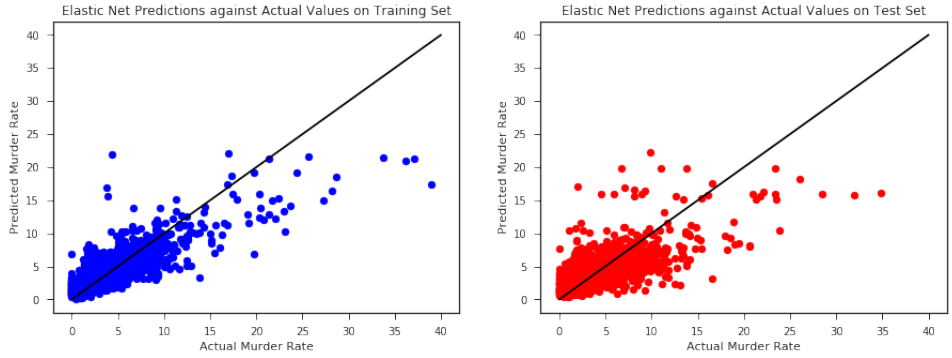


Figure 2: This shows the predicted values against the actual values for the elastic net model, for train (left) and test (right) data.

5 Results, Conclusions, and Future Work

5.1 Analyzing Our Results

Using an elastic net method, we were able to explain a substantial portion of variation in murder rates between MSAs ($r^2 = 0.44$). The Elastic Net consistently performed somewhat better than other model specifications (about 0.02 larger k-fold cross-validated r^2 on the test data compared to Lasso or Ridge. However, we did not find significant additional predictive value yielded by the variables we added in the third model, compared to the second set of models.

Here, we examine the model estimates from the final set of models. We first examine the strongest predictors of higher murder rates, then the strongest predictors of reduced murder rates, and finally, our coefficients of particular interest. These should be taken with considerable grains of salt. The methods of analysis we employed do not allow for causal interpretations. Furthermore, these are fully interacted models, so the first-order terms should be interpreted with that in mind – conditional

on other model parameters equal to zero.

Most Positive Coefficients

Coef	Predictors
5.100205	latino_pct
4.494669	black_pct state_MI
4.438997	white_pct YR_2008
3.770535	latino_pct YR_2011
3.473148	white_pct state_TX
3.441757	subsidy_pct latino_pct
3.285273	state_LA YR_2008
2.968451	state_MI DEMVOTESHARE
2.885476	latino_pct YR_2013
2.824805	median_age_m black_pct

Since the variables were standardized, we can compare them directly. This table shows the ten most substantial predictors of higher murder rates. The percent latino was the strongest predictor of higher murder rates, followed by the interaction between percent black and Michigan, the interaction term for percent white and year 2008, and interaction term for percent white and Texas. Then, the next highest coefficients were percent public assistance and percent latino, the state of Louisiana in the year 2008, the interaction between state of Michigan and Democratic Vote Share, and the percentage of Latinos multiplied by the year 2013. FInally, the 10th biggest predictor of higher murder rates was the median age of males interacted with the percent black.

Most Negative Coefficients

Coef	Predictors
-6.456212	latino_pct state_TX
-4.147626	white_pct
-3.828272	median_income
-3.672966	latino_pct state_CA
-3.348230	white_pct state_LA
-3.181992	black_pct YR_2008
-3.123856	white_pct latino_pct
-2.818669	state_IN fop_density
-2.685127	total_pop state_NY
-2.603210	state_NM YR_2011

This table shows the ten strongest predictors of ameliorated murder rates. The biggest predictor of reduced murder rates is percent Latino and state of Texas, then percent white and median income. Other substantively informative predictors of percent Latinos is Latino percent interacted

with California, white percent with Louisiana, and black percent times year 2008. Then, percent white interacted with percent Latino, the state of Indiana interacted with FOP density, and total population interacted with New York, and the state of New Mexico and Year 2011.

5.2 Analyzing Our Substantive Coefficients

Coef	Predictors
-0.346768	Fraternal Order of Police Density
2.690479	Percent Public Assistance
1.310973	Democratic Vote Share
-0.0	Percent Married
-3.828272	Median Income

Of the variables we included with particular substantive interest, we found several interesting findings. Since these values were standardized prior to estimating the model, we can compare them directly. Fraternal Order of Police Lodge Density was associated with significantly fewer murder rates. The percentage of people receiving public assistance was associated with more murder rates. The percentage of democratic vote share was associated with higher murder rates. Higher median income was associated with much lower murder rates. Interestingly, after including other model parameters, the percent of people married was shrunk to zero by the Elastic Net model.

5.3 Future Research

An interesting avenue for future work might be to study the relationship between police killings of citizens, police deaths on duty, and homicide rates. This could help gain insight into a related question of interest – what determines levels of police violence?

References

Enns, Peter K. 2014. "The Public's Increasing Punitiveness and Its Influence on Mass Incarceration in the United States." *American Journal of Political Science* 58(4).

İmrohoroğlu, Ayse, Antonio Merlo and Peter Rupert. 2006. "Understanding the determinants of crime." *Journal of Economics and Finance* 30(2):270–284.

Towers, Sherry and Michael D White. 2017. "The "Ferguson effect", or too many guns?" *Significance* 14(2):26–29.

Wolfe, Scott E and Justin Nix. 2016. "The alleged "Ferguson Effect" and police willingness to engage in community partnership." *Law and human behavior* 40(1):1.