

Project Report

Caroline Andy, Vasili Fokaidis, Stella Li, Tessa Senders, Lily Wang

12/11/2020

Abstract

Background: Hate crimes are currently the highest priority of the FBI's civil rights program. Existing research suggests that various community-level socioeconomic factors, such as income inequality, are associated with hate crime rate.

Objectives:

We aimed to analyze the association between community-level variables and state-level hate crime rates using data reported to the Southern Poverty Law Center and analyzed in a 2017 FiveThirtyEight article.

Methods:

The data used for this project included state-level hate crime rates per 100,000 individuals in the population, as reported by the Southern Poverty Law Center during the first weeks of November, 2016. Data elements also included socioeconomic factors that are hypothesized to be associated with hate crime.

The association between these socioeconomic factors and the hate crime rate outcome were examined using multivariable linear regression analysis. The outcome variable was first log transformed to adhere to the linear regression assumption of normal distribution. After identifying the significant predictors of state-level hate crime rate when controlling for all other covariates, we used automated and criterion-based approaches to generate a parsimonious high-performing predictive model. Multicollinearity of covariates and variable interaction were also tested for and considered in model development.

Results:

Gini Index (an indicator of wealth inequality) and percent population with a high school degree were both significant predictors of state-level hate crime rate when controlling for all other covariates.

Of the continuous variables, percent non-citizen and percent non-white were found to have a correlation coefficient of 0.753; median household income and percentage of population with a HS degree had a correlation coefficient of 0.651, both of which may suggest multicollinearity. No statistically significant interactions were identified between any variables in regression modeling.

Using automated and criterion-based approaches, and considering the body of scientific evidence with regard to significant and practically important socioeconomic factors associated with hate crimes, we identified one regression model which maximizes parsimony and predictive performance. This model includes the following predictors: [ADD FINAL PREDICTORS]

Conclusions:

Based on the November 2016 Southern Poverty Law Center data, Gini Index (an indicator of wealth inequality) and percent population with a high school degree were both significantly and independently associated of state-level hate crime rate.

Introduction

Since 2014, the national rate of hate crimes has been steadily increasing in the United States [CITE]. In the days following the 2016 US Presidential election, an average of 90 hate crimes per day were reported to the

Southern Poverty Law Center [CITE].

Existing research suggests that community-level socioeconomic factors such as racial breakdown, population density, level of educational attainment, and economic considerations (median income, poverty level, job availability) may be significant predictors of regional and state-level rates of hate crimes. [CITE] A 2017 FiveThirtyEight article titled, “Higher Rates Of Hate Crimes Are Tied To Income Inequality,” used 2016 FBI and Southern Poverty Law Center data to assess the association between hate crime rate and select community-level variables [CITE].

For this project, we used this dataset to critically analyze this research team’s findings to identify state-level variables associated with rates of hate crimes, and to generate a high performing predictive model for population-adjusted hate incidents in the United States.

Methods

Data Exploration

The data used for this project included state-level hate crime rates (hate crimes per 100,000 population), as reported by the Southern Poverty Law Center during the first weeks of November, 2016. Collected state-level demographic variables include:

- Unemployment rate (high vs low) (as of 2016)
- Urbanization (high vs low) (as of 2015)
- Median household income (as of 2016)
- Percent of residents with a high school degree (as of 2009)
- Percent of residents who are non-citizens (as of 2015)
- Income Gini coefficient (a measure of the extent to which the distribution of income among individuals within an economy deviates from a perfectly equal distribution; as of 2015)
- Percent of residents who are non-White (as of 2015)

First, we investigated the extent of missing data in our dataset. Four states—Hawaii, North Dakota, South Dakota, and Wyoming—did not report hate crime rate data, and thus were excluded from subsequent analyses. One additional state, Maine, did not report its percent of residents who were non-citizens. Washington DC was included as a state for the purposes of these analyses.

Using these data, our goal was to construct our own multivariable linear regression model to assess which of the collected variables, if any, are significantly associated with population-adjusted hate incidents in the United States, to altogether critically examine the article’s findings. To do so, we first generated descriptive statistics and plotted the distribution of the outcome (population-adjusted hate incidents per 100,000 population) to determine whether any data transformations would be necessary, and to assess whether any outliers exist within the data.

To further elucidate the existence of outliers/influential points, we visualized the hate crime rate by state, and generated residuals vs fitted, normal Q-Q, scale-location, and residuals vs leverage plots. Any states that were deemed to be an outlier were included in subsequent analyses, but the same analyses were also run a second time on the dataset excluding the outliers for comparison.

To report basic descriptive statistics, we generated a table which includes the mean, median, range (min-max), interquartile range, and count of missing entries for each numeric variable, and category percent breakdowns and count of missing entries for categorical variables (Table 1).

Multicollinearity and Interactions In order to investigate the existence of multicollinearity between each of the continuous variables, we generated a correlation matrix. We decided that any correlation coefficient of 0.6 and above may suggest multicollinearity; thus, in these instances, at least one of those correlated variables were dropped during subsequent model development. We also calculated the variance inflation

factors (VIFs) for all variables, which quantify the degree of multicollinearity between the given predictor and all other remaining covariates.

Next, we tested for significant interactions between each of the two categorical variables, unemployment and urbanization, and all other continuous variables. Any variables with significant interactions were then stratified and analyzed again.

Model Development and Validation To generate a high performing predictive model, we then performed two automated procedures for model variable selection (forward and backward selection) and several criterion based approaches (CP, r-squared and SSE/RSS). We also ran an analysis of variance (ANOVA) test to identify the most optimal model subset.

Using these results, we then identified a model which maximizes parsimony, predictive performance, interpretability and practical application, which takes into account variables deemed to be significant and practically important in existing literature.

Model goodness of fit was also assessed on a dataset excluding Washington DC, a previously noted outlier.

For validation, [ADD VALIDATION STUFF HERE]

Results

Data Exploration To visualize the distribution of the hate crimes rate data, we first generated an overlaid density and histogram plot (Figure 1). Multivariable linear regression modeling operates under several assumptions, which include residual homoscedasticity (constant variance) and normality. Initial exploration of the distribution of the hate crimes rate data showed a strong departure from standard normal distribution with significant right skewness. Thus, we performed a Box Cox test to isolate the ‘best’ power transformation on the hate crimes rate variable to achieve normal residuals.

The generated Box Cox plot results suggest that a logarithmic transformation of the hate crimes rate data would most closely approximate normal distribution (Figure 2). Thus, we operated moving forward in model development using the log hate crimes rate as the outcome of interest.

To investigate for any outliers in the data, we then plotted the hate crimes rate by state, and generated residuals vs fitted, normal Q-Q, scale-location, and residuals vs leverage plots (Figures 3-7). As Washington DC is clearly outside of the Cook’s distance line in the residuals vs leverage plot, we concluded that this point is an outlier. This data point was included in any subsequent analyses, though all analyses were also rerun on this dataset excluding Washington DC for comparison.

A table of basic descriptive statistics for the collected socioeconomic variables is included in the appendix (Table 1).

Multicollinearity and Interactions Out of the continuous variables, percent non-citizen and percent non-white has a correlation coefficient of 0.753, median household income and percentage of population with a HS degree has a correlation coefficient of 0.651, both of which may suggest multicollinearity. All other correlation coefficients do not suggest multi-collinearity (Table 2).

Use of variance inflation factors (VIFs) showed that the percent population without a high school degree (3.895), percent non-citizen (3.728), percent non-white (3.236) and median household income (3.108) had the highest degrees of multicollinearity between all other predictors. These values approach but do not exceed 4, the generally accepted value which denotes the need for further investigation and/or consideration of multicollinearity corrections.

Model Development and Validation We began our regression analyses by running linear regression models containing all possible predictors for both the untransformed hate crime rate and the log hate crime rate. Our results support the conclusions drawn in the FiveThirtyEight article: that the Gini Index was the most significant independent predictor of state hate crime rate when controlling for all other covariates; percent high school graduates variables was the only other statistically and independently significant variable.

To generate our own predictive model maximizing model parsimony and predictive performance, we employed two automated procedures: backward and forward selection. While the model proposed during forward selection contained all provided variables (Adjusted R-squared = 0.1849), the model generated through backwards selection was much more parsimonious and had a substantially higher adjusted R squared (Adjusted R-squared = 0.2868). The only included predictors were the Gini Index, and the percent of high school graduates variables.

We then employed three criterion based approaches in R, each of which generated the two best performing models which optimize the given criterion for each possible number of predictors. The first of these approaches used the Cp criterion, which compares the predictive ability of model subsets to the full model. To visualize these results, we generated a plot containing the Cp criterion distribution for the top performing model for each number of predictors (Figure 8). This shows that the best performing model with 3 covariates (Gini Index, percent high school graduates, and unemployment) had the best predictive performance as compared to the top performing models of all other possible sizes.

We also isolated the top two best performing models per number of predictors using the adjusted R squared criterion. We then generated a plot of the distribution of the adjusted R squared statistic for the top performing model for each number of included predictors (Figure 9). Results show that the model with three predictors performed the best while maximizing parsimony.

We generated the same outputs using the SSE/RSS criterion. The plot can be found in Figure 10 (Figure 10). Results show that [INPUT RESULTS HERE FOR SSE/RSS CRITERION].

Lastly, to help distinguish the most optimal model between one with three predictors (Gini Index, percent high school graduates, and unemployment), and one with two (Gini Index and percent high school graduates), we ran an analysis of variance (ANOVA) test. Results suggest that the model containing two parameters is most optimal.

[VALIDATION, CHANGES IN PERFORMANCE WHEN INCLUDING DC]

Discussion/Conclusion

Figures and Tables

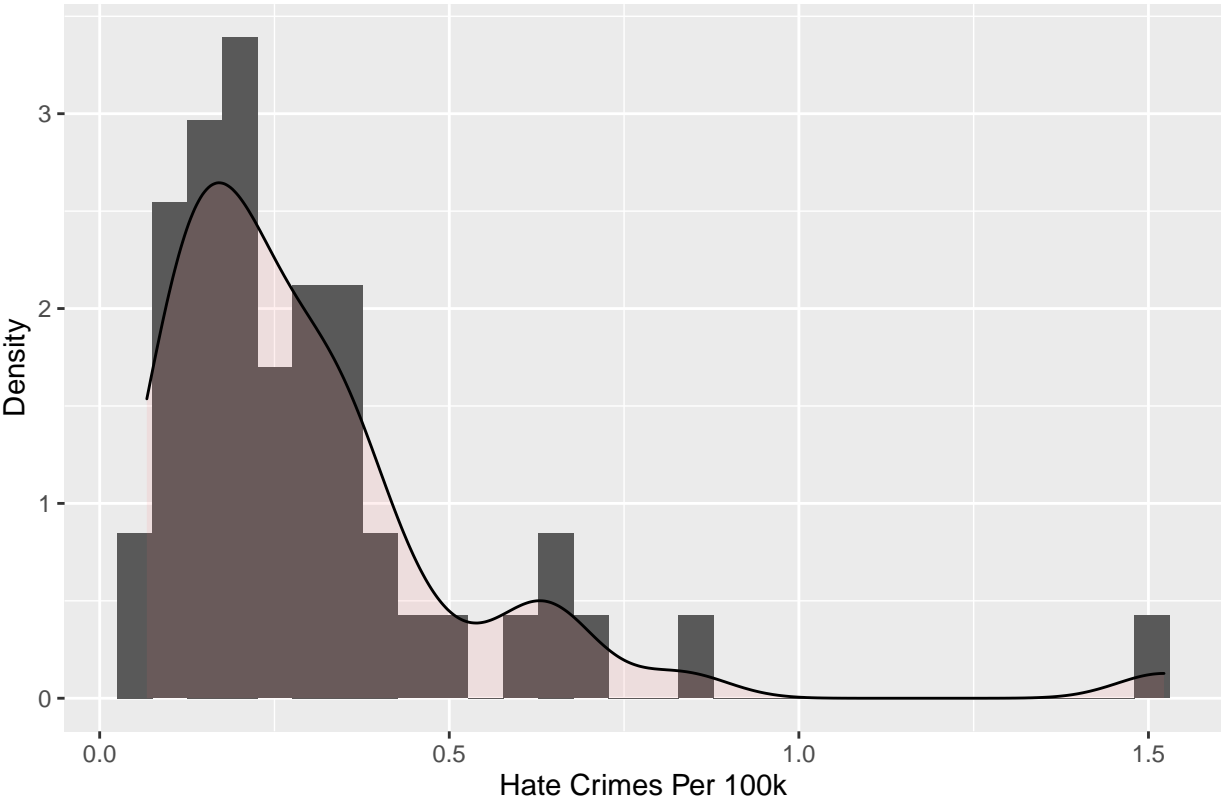


Figure 1: Untransformed distribution of hate crimes per 100k

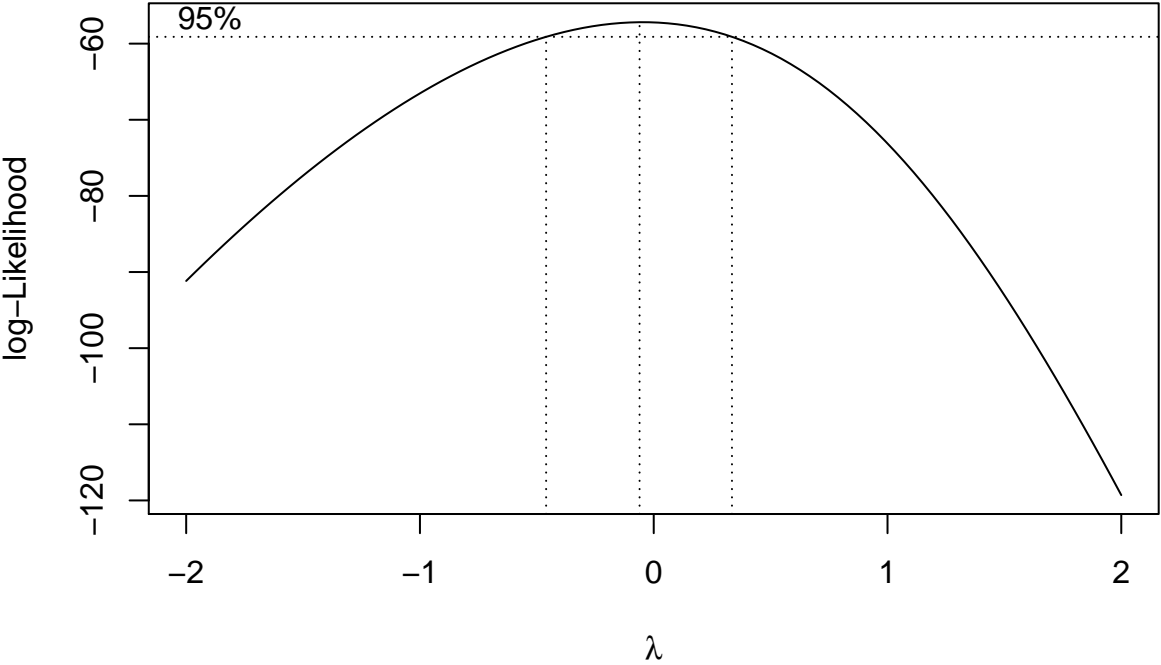
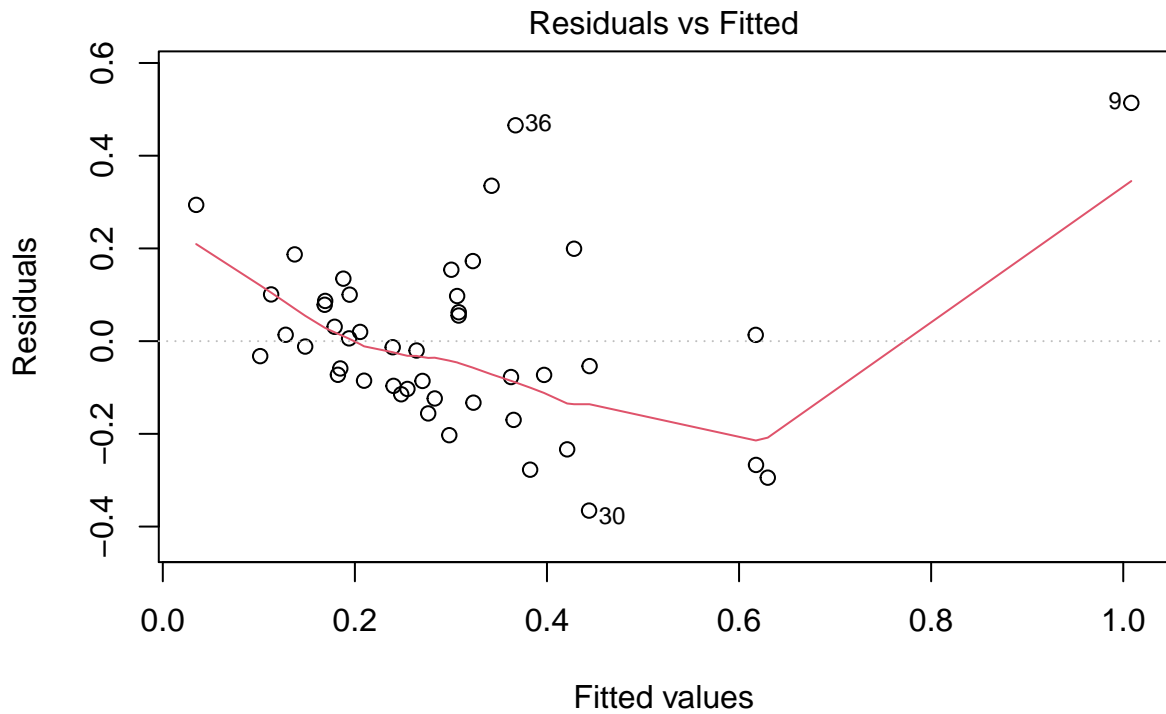
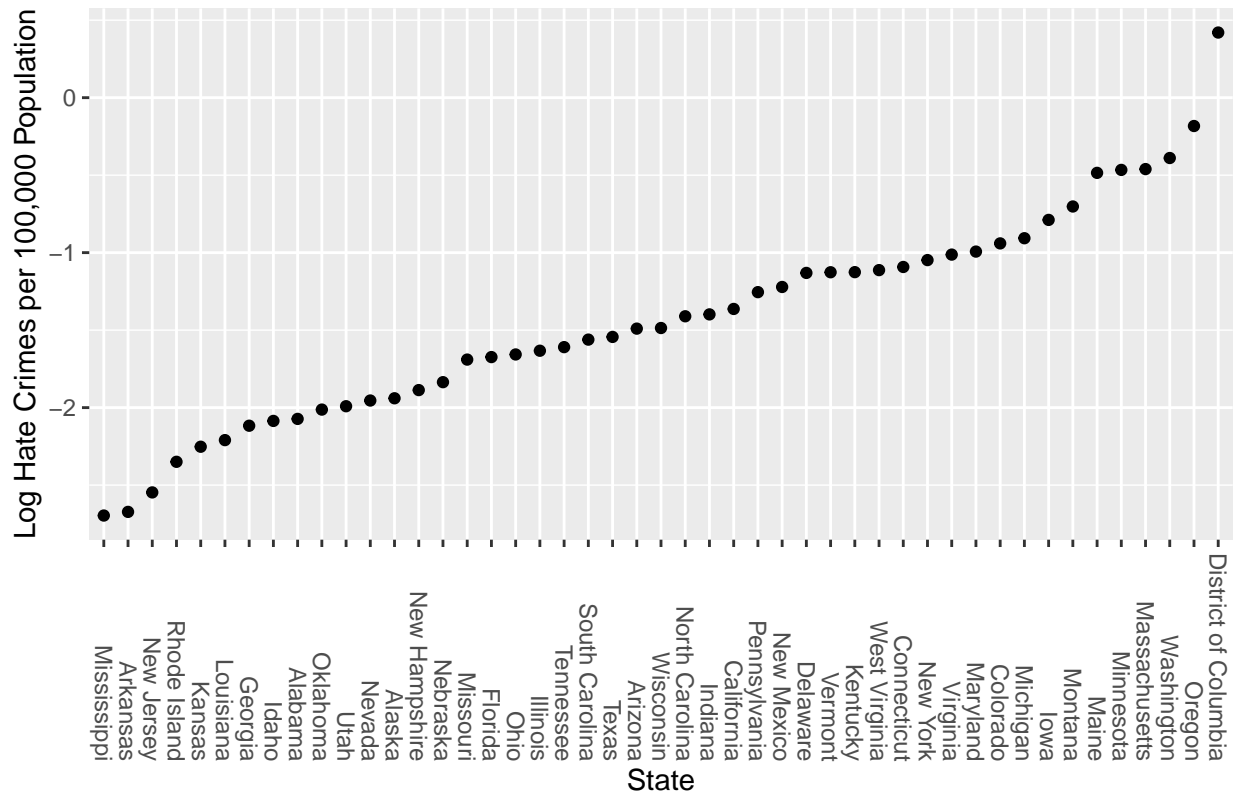
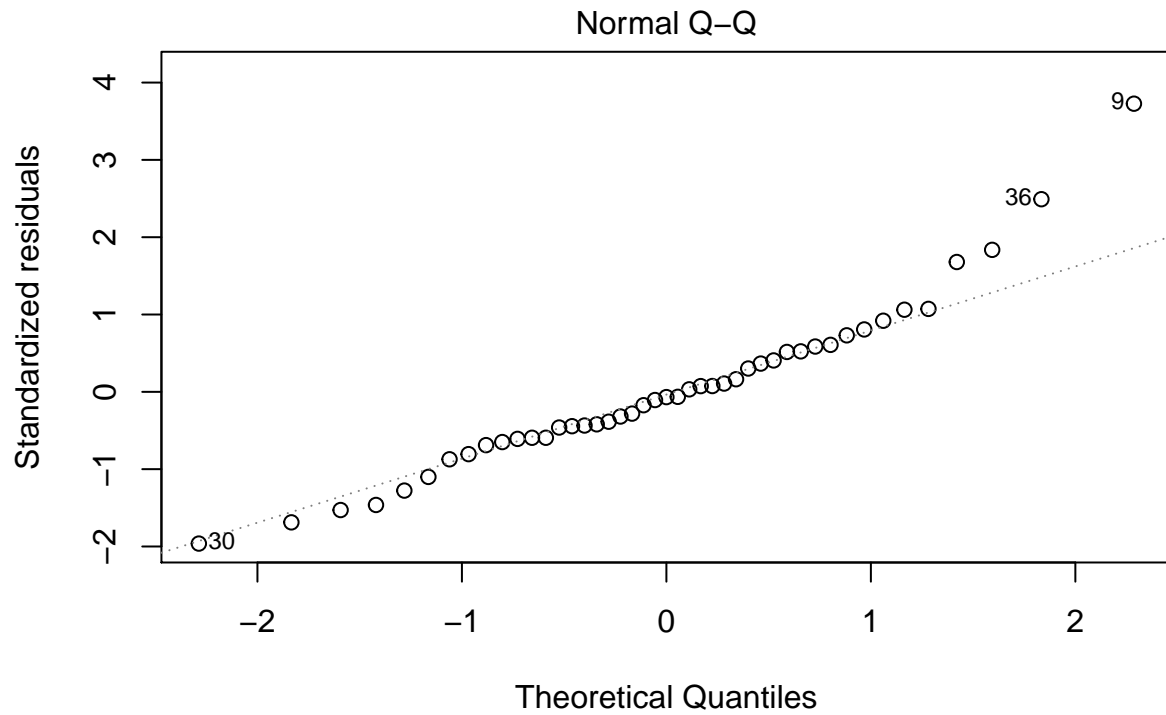


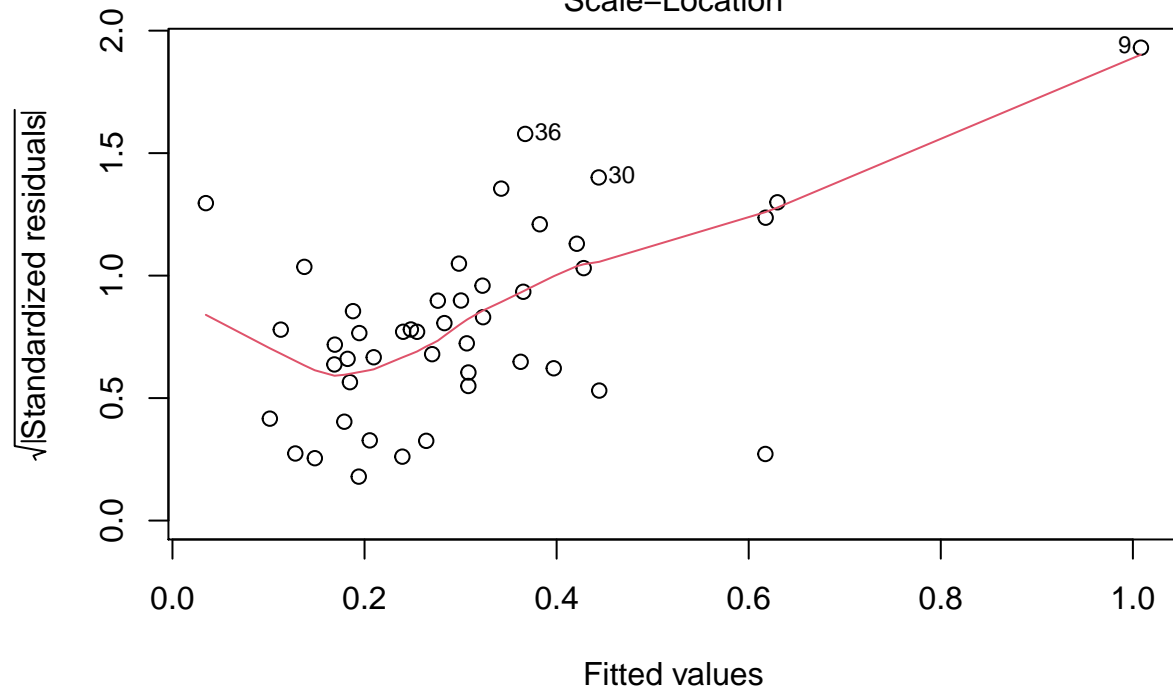
Figure 3. Log Hate Crimes Rate by State



$\text{lm}(\text{hate_crimes_per_100k_splc} \sim \text{unemployment} + \text{urbanization} + \text{median_househo})$



Scale-Location



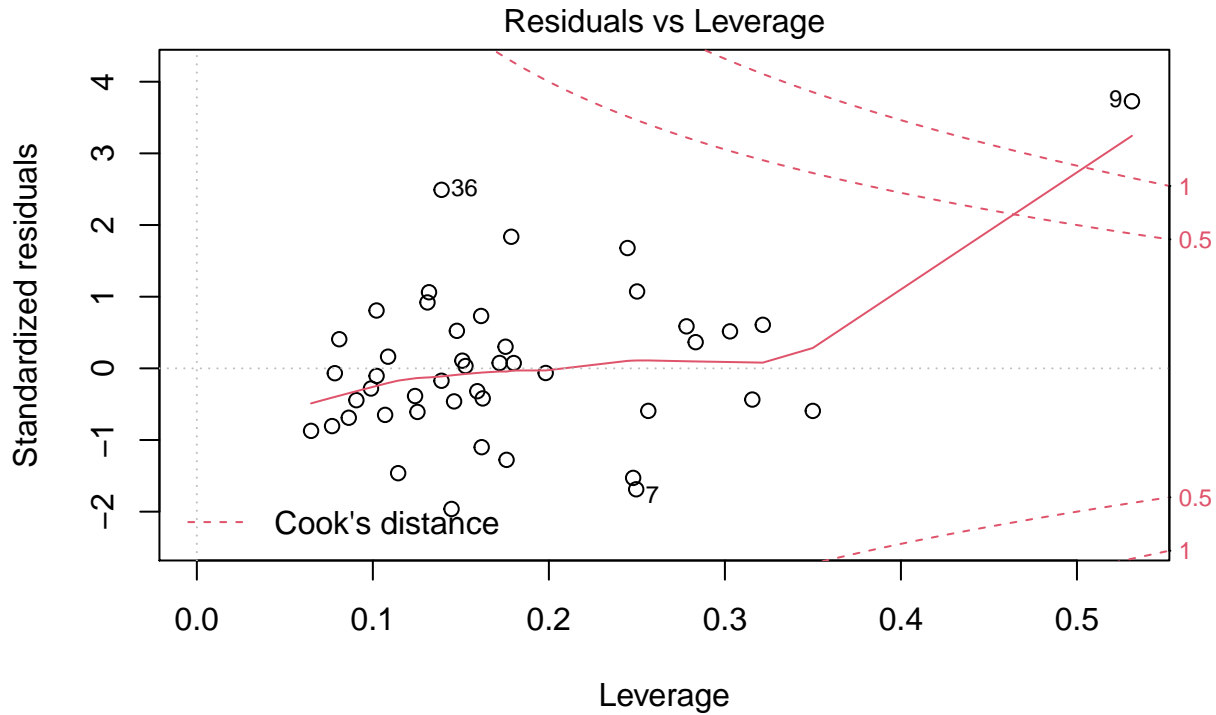


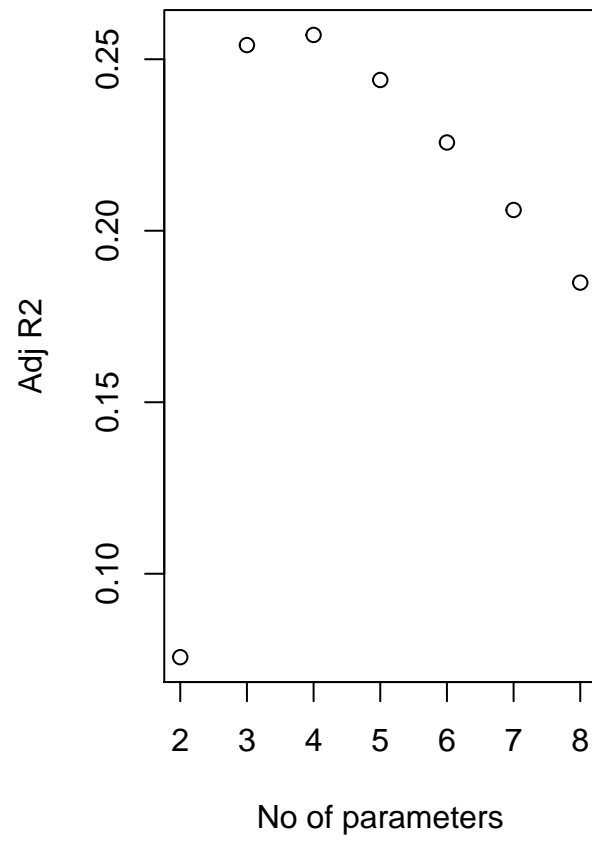
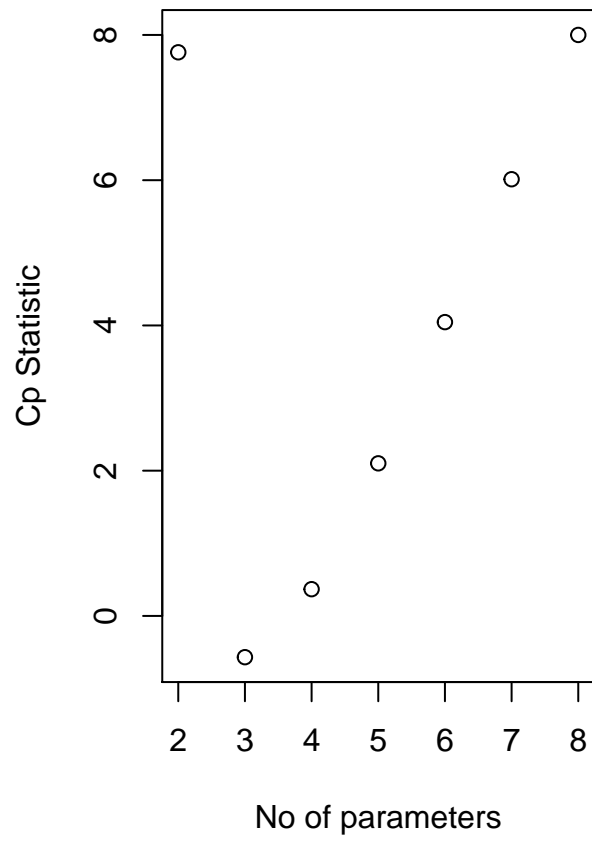
Table 1: Table 1: Descriptive Statistics of Tidied Dataset

	Overall (N=47)
Unemployment	
high	24 (51.1%)
low	23 (48.9%)
Missing	0
Urbanization	
low	23 (48.9%)
high	24 (51.1%)
Missing	0
Median Household Income	
Mean (SD)	54802.298 (9255.117)
Median (Q1, Q3)	54310.000 (47629.500, 60597.500)
Min - Max	35521.000 - 76165.000
Missing	0
% Adults >25yrs With HS Degree	
Mean (SD)	0.866 (0.034)
Median (Q1, Q3)	0.871 (0.839, 0.895)
Min - Max	0.799 - 0.915
Missing	0
% of Population Not U.S. Citizens	
Mean (SD)	0.055 (0.031)
Median (Q1, Q3)	0.050 (0.030, 0.080)
Min - Max	0.010 - 0.130
Missing	2
Gini Index	
Mean (SD)	0.456 (0.021)
Median (Q1, Q3)	0.455 (0.441, 0.468)

	Overall (N=47)
Min - Max	0.419 - 0.532
Missing	0
% of Population Not White	
Mean (SD)	0.315 (0.150)
Median (Q1, Q3)	0.300 (0.205, 0.420)
Min - Max	0.060 - 0.630
Missing	0
Hate Crime Rate Per 100k	
Mean (SD)	0.304 (0.253)
Median (Q1, Q3)	0.226 (0.143, 0.357)
Min - Max	0.067 - 1.522
Missing	0
hate_crimes_log	
Mean (SD)	-1.429 (0.676)
Median (Q1, Q3)	-1.486 (-1.947, -1.030)
Min - Max	-2.696 - 0.420
Missing	0

Table 2: Table 2: Correlation Matrix of All Continuous Variables

	median_household_income	perc_pop_hs	perc_non_citizen	gini_index	perc_non_white	hate_crimes_per_100k	hate_crimes_log
median_household_income	1.000	0.651	0.302	-0.130	0.039	0.344	0.311
perc_pop_hs	0.651	1.000	-0.262	-0.537	-0.496	0.263	0.296
perc_non_citizen	0.302	-0.262	1.000	0.480	0.753	0.244	0.137
gini_index	-0.130	-0.537	0.480	1.000	0.548	0.381	0.219
perc_non_white	0.039	-0.496	0.753	0.548	1.000	0.111	-0.007
hate_crimes_per_100k	0.344	0.263	0.244	0.381	0.111	1.000	0.892
hate_crimes_log	0.311	0.296	0.137	0.219	-0.007	0.892	1.000



References