Project Report

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

12/11/2020

Abstract

Background: Hate crimes are a growing public health threat in the United States and are the highest priority of the FBI's civil rights program (Miller et al., 2016). 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 rate of population-adjusted hate incident using data reported to the Southern Poverty Law Center and analyzed in a 2017 FiveThrityEight article (Majumder, 2017).

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 natural log transformed to adhere to the linear regression assumption of normal distribution. After identifying the statistically 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 model that fit our data well.

Multicollinearity of covariates, outliers, and variable interaction were also tested for and considered in model development.

Finally, we assessed model goodness of fit and predictive performance through model validation. All steps of model development and validation were performed before and after removing the District of Columbia, an outlier in the data.

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 high school 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 a regression model which optimizes parsimony and goodness of fit. This model contains only the Gini Index and percent population with a high school degree as predictors.

All steps of model development and validation were performed before and after removing the District of Columbia, an outlier in the data. After removing the District of Columbia, the Gini Index was no longer a

statistically significant predictor of hate crime rate; instead, the percentage of the population with a high school degree was the only significant variable, making the District of Columbia an influential point.

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 statistically significantly and independently associated with state-level hate crime rate.

Introduction

As of 2020, the national rate of hate crimes in the United States is at its highest level in over a decade ["FBI Incidents and Offenses," 2019]. 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 (Miller et al., 2016).

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 ("FBI: Variables Affecting Crime," 2012, Shively, 2005, Van Dyke et al., 2014). 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 (Majumder, 2017).

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. Three states–Maine, Mississippi, and South Dakota did not report their percent of residents who were non-citizens. The District of Columbia was included as a state for the purposes of these analyses.

Using these data, our goal was to construct our own multivariable linear regression models to assess which of the collected variables, if any, are statistically significantly associated (at a 0.05 significance level) 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 that we reported in 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). The variable Gini Index was particularly interesting as it indicated income inequality in each state. The mean Gini Index value was 0.456 with a standard deviation of 0.021, and the range of these values was from 0.419 to 0.532. Together, these values indicated great income inequality in the overall dataset. Additionally, hate crime rate per 100,000 individuals had a mean and standard deviation of 0.315 and 0.150, respectively. The range of rates was from 0.067 to 1.522 indicating considerable variation in hate crime crimes per 100,000 individuals on a state-level. For categorical variables, the data showed that majority of states had high unemployment and high urbanization at 51.1% each.

In order to determine whether any data transformations would be necessary we visualized the distribution of the outcome (population-adjusted hate incidents per 100,000 population) by generating 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 transformation to isolate the 'best' power transformation on the hate crimes rate variable to achieve normal residuals.

To elucidate the existence of outliers and thus potential influential points, we visualized the hate crime rate by state (Figure 3). We then generated residuals vs fitted, normal Q-Q, scale-location, and residuals vs leverage plots for a linear regression model that regressed the transformed hate crime rate onto all possible covariates (Figure 4). Specifically, we used the Residuals vs Leverage plot to identify any outliers. Any points outside of Cook's Distance were considered potential influential points. Any states that were deemed to be outliers were included in the subsequent analyses, but the same analyses were also run a second time on the dataset excluding the outliers for comparison in order to determine if the outlier was indeed influencing the results of our analyses.

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. Additionally, we calculated the variance inflation factors (VIFs) for all variables, which quantify the degree of multicollinearity between the given predictor all other remaining covariates.

Next, we plotted potential interactions between the continuous variables and each categorical variable: urbanization and unemployment. These plots were created once including outliers and once excluding outliers. We then formally tested for significant interactions between any variable pairs with intersecting lines. All interaction tests were performed on datasets containing and not containing any observed outliers.

Model Development and Validation We began by performing two automated procedures for model variable selection (forward and backward stepwise selection). We then used two criterion based approaches (Mallow's Cp and adjusted r-squared). These two criterion based approaches in R each generated the best performing model which optimizes the given criterion for each possible number of predictors. The first of these approaches used Mallow's 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). The next approach used adjusted r-squared which favors models with a smaller SSE but also penalizes for additional predictors. We 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).

Using these results, we then identified the two best models which maximize parsimony, interpretability and practical application, which takes into account variables deemed to be significant and practically important in existing literature. We then compared these two models using an analysis of variance (a partial F-test for nested models).

We repeated this process of model selection on the dataset excluding any observed outliers.

To assess the two models' predictive performance, we performed 5-fold Cross Validation on each including and excluding outliers.

Results

Data Exploration The generated Box Cox transformation plot indicates that a lambda of 0 (i.e. a natural logarithmic transformation of the hate crimes rate outcome) would most closely approximate normal distribution (Figure 2). Thus, we operated moving forward in model development using the natural log of hate crimes rate as the outcome of interest.

Examining the Residuals vs Leverage plot generated from the regression model including all possible covariates, we see that the District of Columbia is clearly outside of the Cook's Distance line; thus we concluded that this point is an outlier and a potential influential point.

A table of basic descriptive statistics for the collected socioeconomic variables is included in the appendix (Table 1). Continuous variables show their mean, standard deviation, median, quartiles 1 and 3, their minimum to maximum range of values, and number of missing values. Categorical variables show their factor levels (i.e. high and low), percents of each factor level within the data set, and number of missing values as well. Percent of population that are not U.S. citizens is the only variable containing missing values (2 missing). FIX THIS PARAGRAPH!

Multicollinearity and Interactions Out of the continuous variables, percent non-citizen and percent non-white have a correlation coefficient of 0.753, and median household income and percentage of population with a high school degree have 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) have 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.

For the dataset containing the District of Columbia, the interaction plots suggest potential interaction between unemployment and median household income, as well as between unemployment and percent population of high school graduates, as seen by the differing directions in slope (crossing lines). The urbanization interaction plots for these data suggest that there may be potential interaction between urbanization and each of the following variables: Gini Index, median household income, percent non-citizen, and percent population of high school graduates.

For the dataset excluding the District of Columbia, the interaction plots show potential interaction between unemployment and Gini Index as well as between unemployment and percent population of high school graduates. The urbanization interaction plots for these data suggest potential interaction between urbanization and median household income as well as between urbanization and percent population of high school graduates.

When performing regression analyses for all of the above interactions for each dataset, no interactions were found to be statistically significant at a 0.05 significance level.

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 natural log transformed 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 and that the percent high school graduates variable was the only other statistically and independently significant variable (Majumder, 2017).

The model proposed during forward stepwise selection contained all variables provided in the original dataset (Adjusted R-squared = 0.1849). The model generated through backward stepwise 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, both of which were significant. When the District of Columbia is removed from the dataset, however, the Gini Index is no longer statistically significant at a 0.05 significance level, indicating that the District of Columbia is an influential point. The adjusted R-squared of the two-predictor model decreases as well when the District of Columbia is removed (from 0.2541 to 0.1185).

Both Mallow's Cp and the adjusted r-squared criterion confirmed that the best model that includes two variables contains Gini Index and the percent of high school graduates. Further, the plot for Mallow's Cp showed that the model with only two predictors had the lowest Cp value as compared to the best performing model of all other possible numbers of predictors. The plot for the adjusted r-squared criterion showed that the model with 3 predictors (Gini Index, the percent of high school graduates, and unemployment) had the highest adjusted r-squared value. However, the adjusted r-squared for this model was less than 6% greater than the adjusted r-squared for the model with only 2 covariates, suggesting that their performance is not significantly different with regard to this criterion. Further the partial F-test for nested models shows that adding the unemployment variable does not significantly improve the model. All of these results hold true when the District of Columbia is removed from the dataset.

Through cross-validation, we determined that the aforementioned two-predictor model also had better predictive performance than the three-predictor model when including the District of Columbia. The CV root mean squared error (RMSE) values are essentially equivalent (0.5948853 for 2 covariates vs 0.6038494 for 3 covariates) and the CV adjusted r-squared is slightly higher for the two-predictor model (0.2943289 vs 0.2783783). When excluding the District of Columbia, the two different models have virtually the same predictive performance. Both have similar CV RMSE values and the CV adjusted r-squared for the three-predictor model is only slightly higher than the two-predictor model(0.1730294 vs 0.1530310). This indicates that the addition of unemployment as a third predictor did not significantly add to model performance whether the District of Columbia is included or not (Table 4). The two-predictor model, whether the District of Columbia is included or not, performs well as a predictive model since the model's RMSE and adjusted r-squared are very similar to the CV average RMSE and adjusted r-squared.

[ADD FIGURE NUMBER AND THEN CITE SOMEWHERE ABOVE]

Discussion/Conclusion

After extensive analyses and modeling, we found that the optimal model in terms of goodness of fit and predictive value contained the Gini Index and percent population with a high school degree, with the latter variable being a more significant predictor of hate crime in this two-parameter model. However, these two predictors only accounted for 25.4% of the variability in the data.

The change in predictor significance of Gini Index and the adjusted r-squared between the model including and excluding the District of Columbia suggests it is indeed an influential point. The District of Columbia has the highest hate crime rate and the highest income inequality out of all the states.

It is possible that other factors not included in our dataset, such as percentage of the population who are LGBTQ+ or religious minorities, may also be important predictors of hate crime.

ADD PARAGRAPH-WHY IS THIS STUDY IMPORTANT???

Unfortunately the data for hate crime rates was missing for four states. These four states may have drastically different hate crime rates than the rest of the states and could potentially change the significance of the predictors and impact the predictive ability of our models.

It is important to note that a natural logarithmic transformation of the outcome was performed in order to satisfy the assumptions of linear regression (residual homoscedasticity and normality) which impacts the modeling results.

Our final model which included significant predictors did not account for much of the variance in the data which implies other variables not in the data should be included in future studies of hate crime rates. Such variables may include variables related to the politics of people in the state or the percent of the population that are LGBTQ+ or percent of the population that belongs to a religious minority as one study showed that religious hate crimes are on the rise(“Incidents and Offenses.” FBI, FBI, 29 Oct. 2019).

-why is this important? - limitations -assumptions - suggestions for possible future studies

“Each figure and table should be of publishable quality and well notated, i.e., labeled and/or captioned”

Figures and Tables

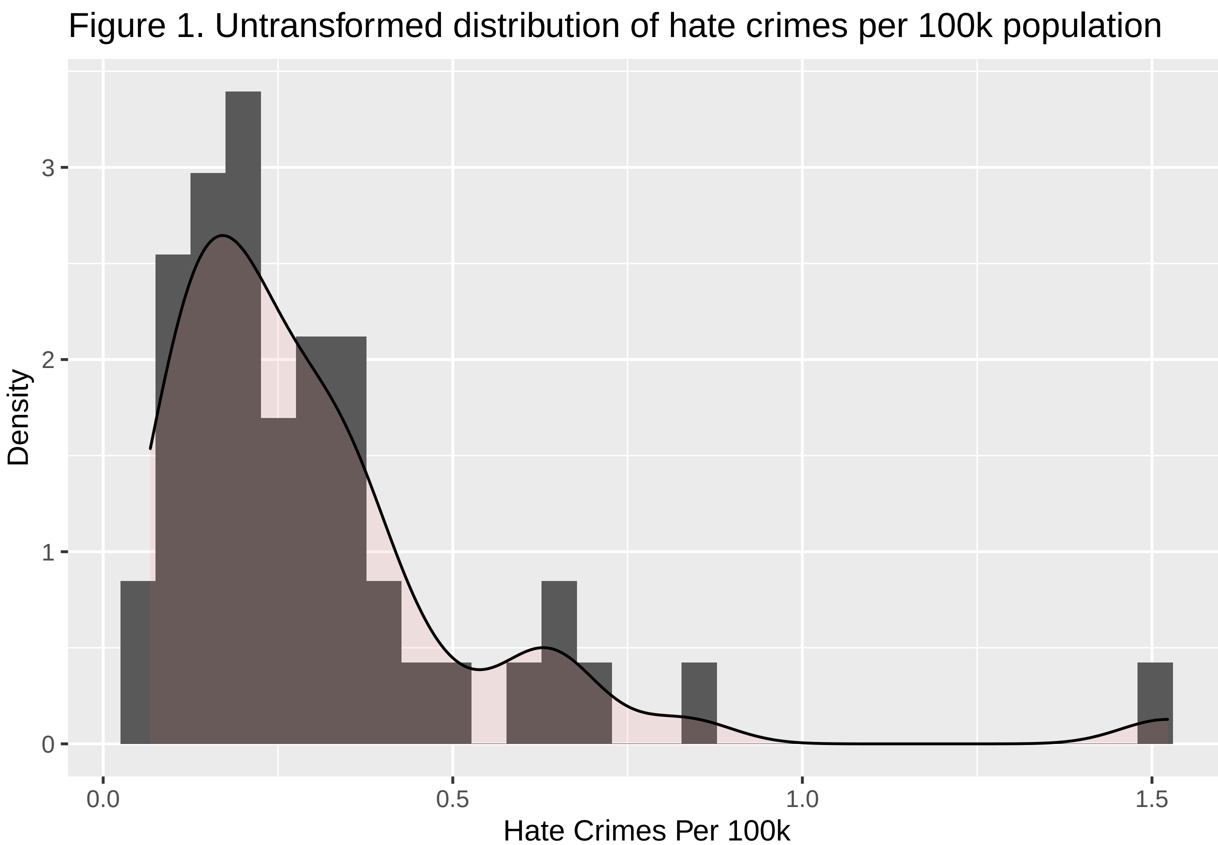

Figure 2. BoxCox Transformation Plot

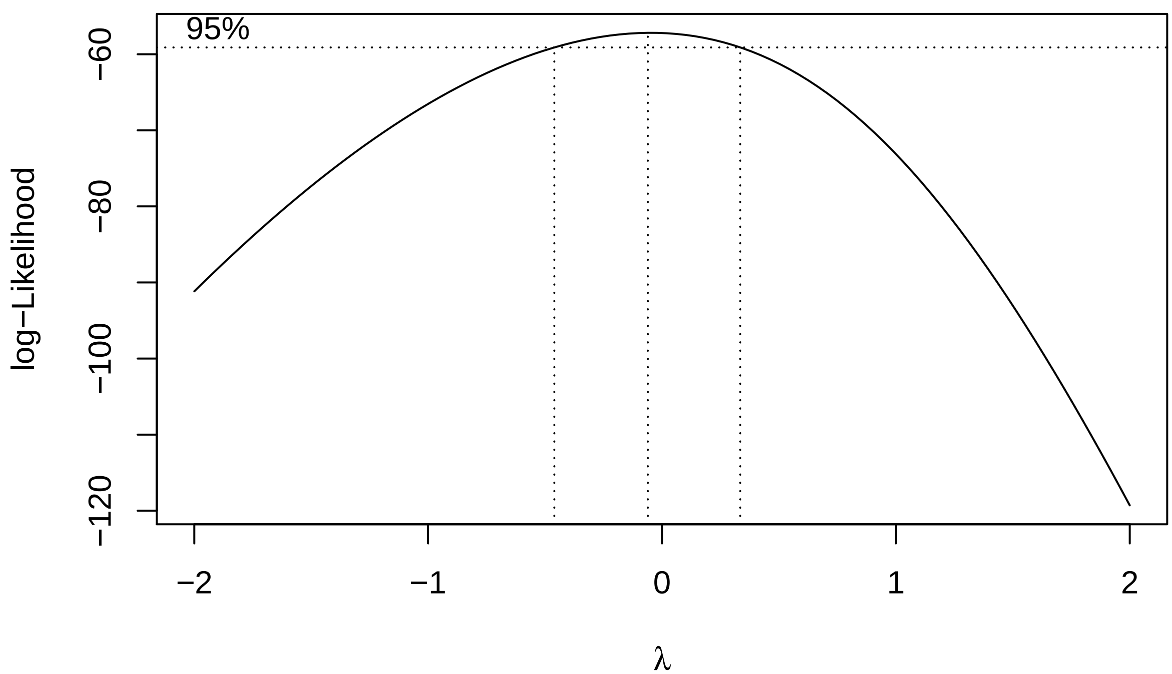

Figure 3. Log Hate Crimes Rate by State

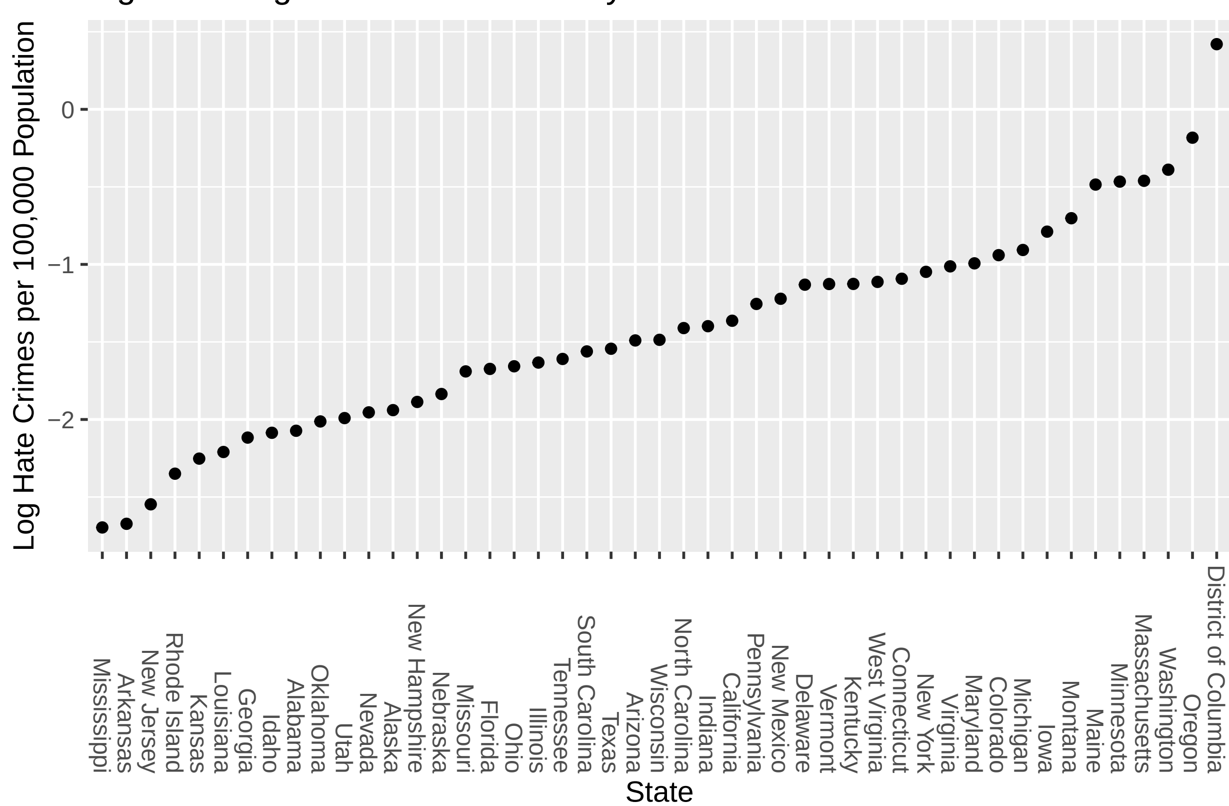

Figure 4. Full Model Hate Crime Diagnostic Plots

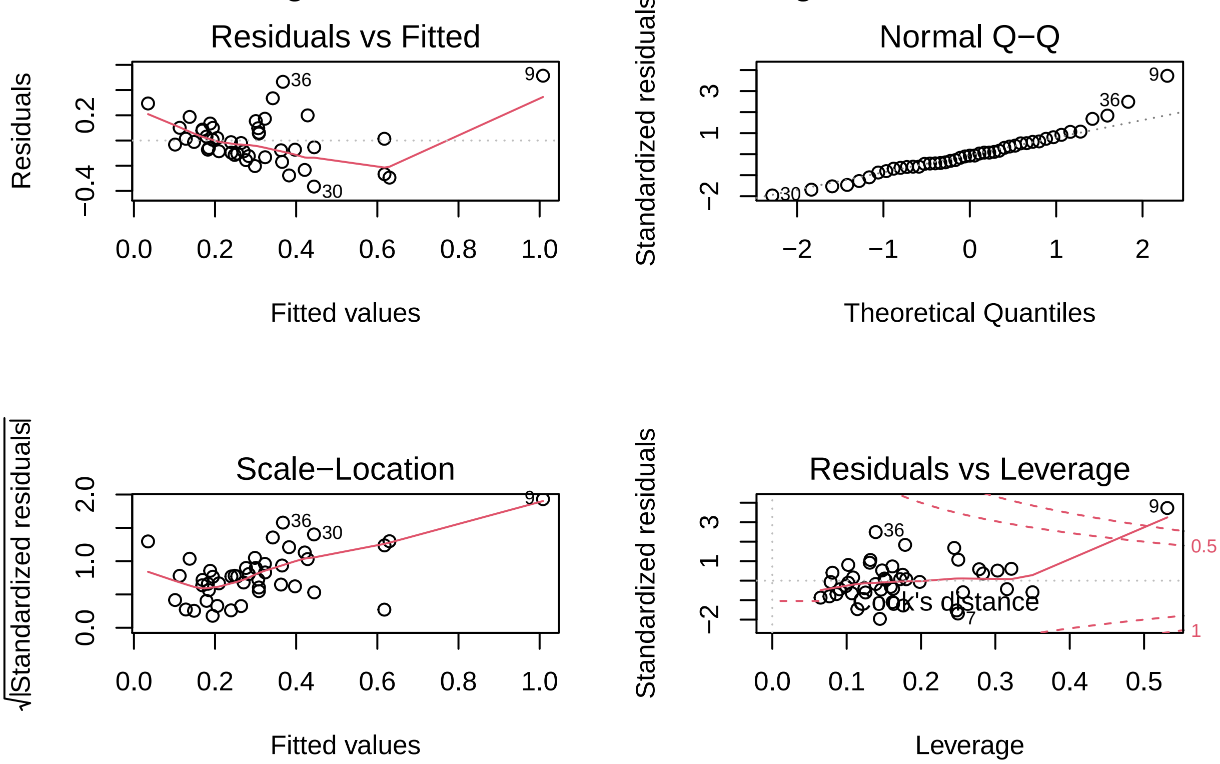

Figure #. Two predictors without DC model Hate Crime Diagnostic Plots

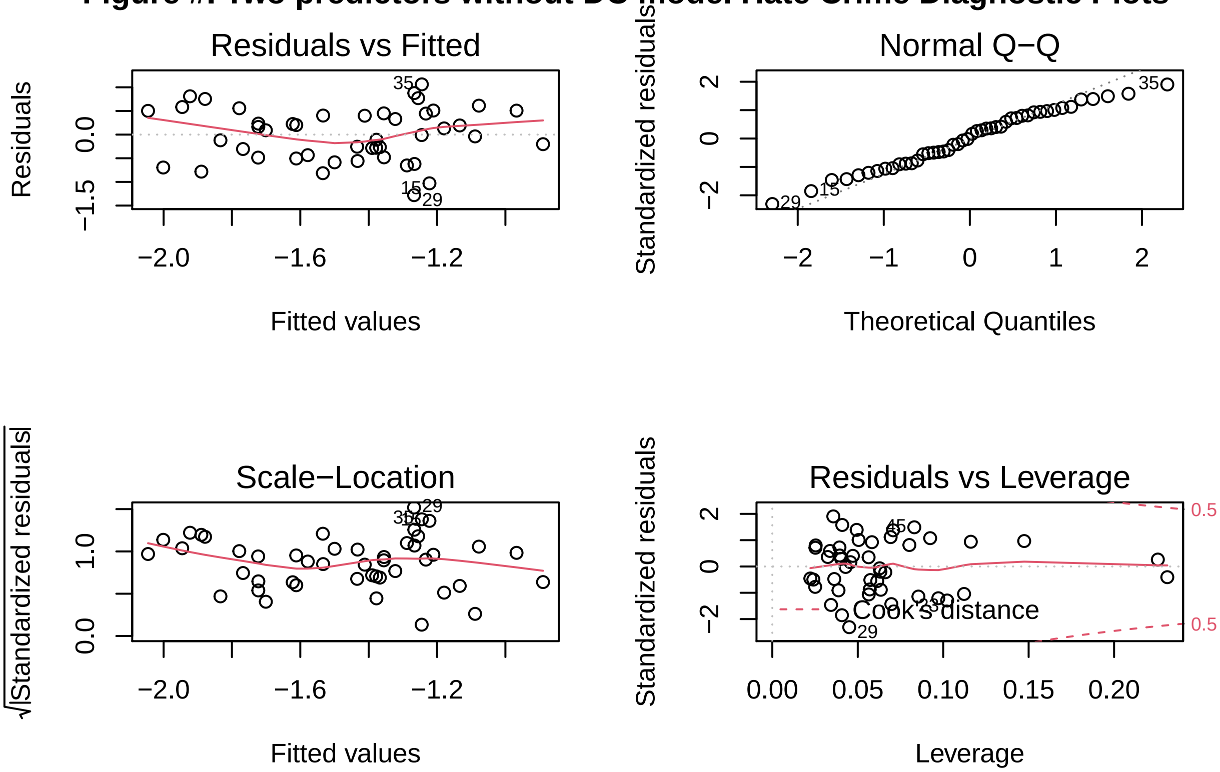

Table 1: Descriptive Statistics of States Reporting Hate Crimes

	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)
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

Figure 5. Correlation Matrix

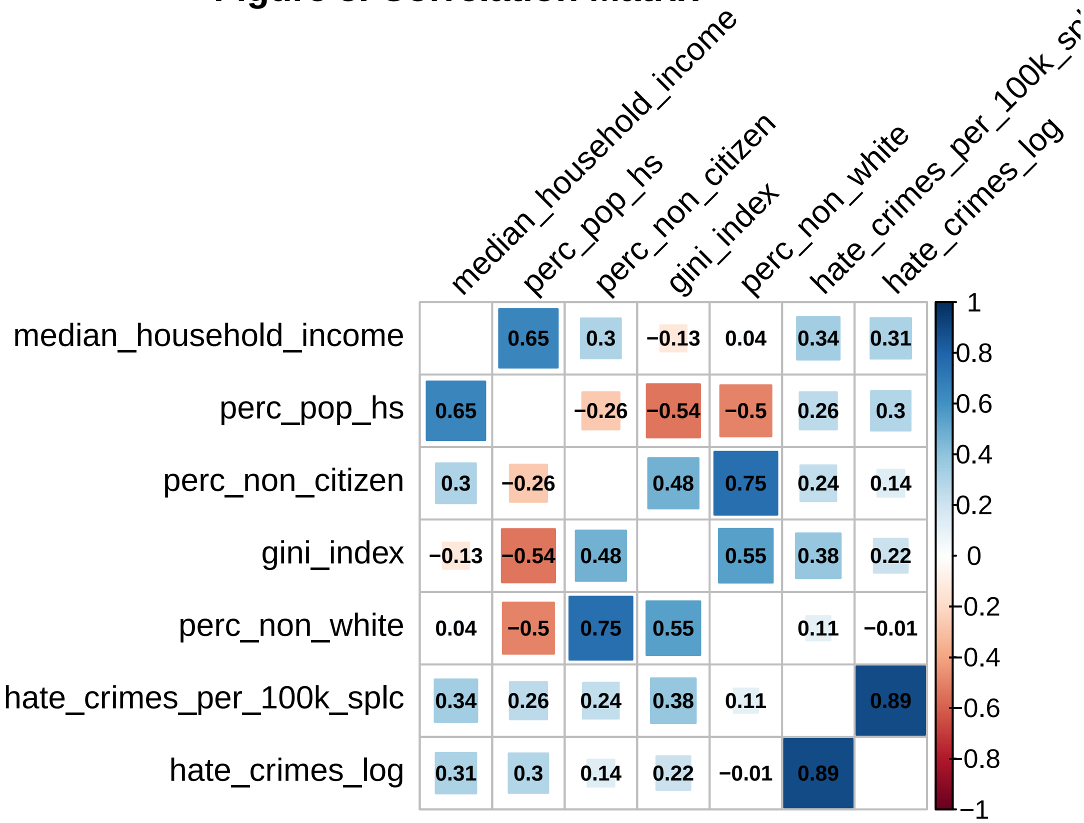

Table 2: VIF Values

	VIF
unemployment	1.426
urbanization	1.983
median_household_income	3.108
perc_pop_hs	3.895
perc_non_citizen	3.728
gini_index	1.845
perc_non_white	3.236

Figure 6. Interaction Plots Between Hate Crime and All Continuous Variable

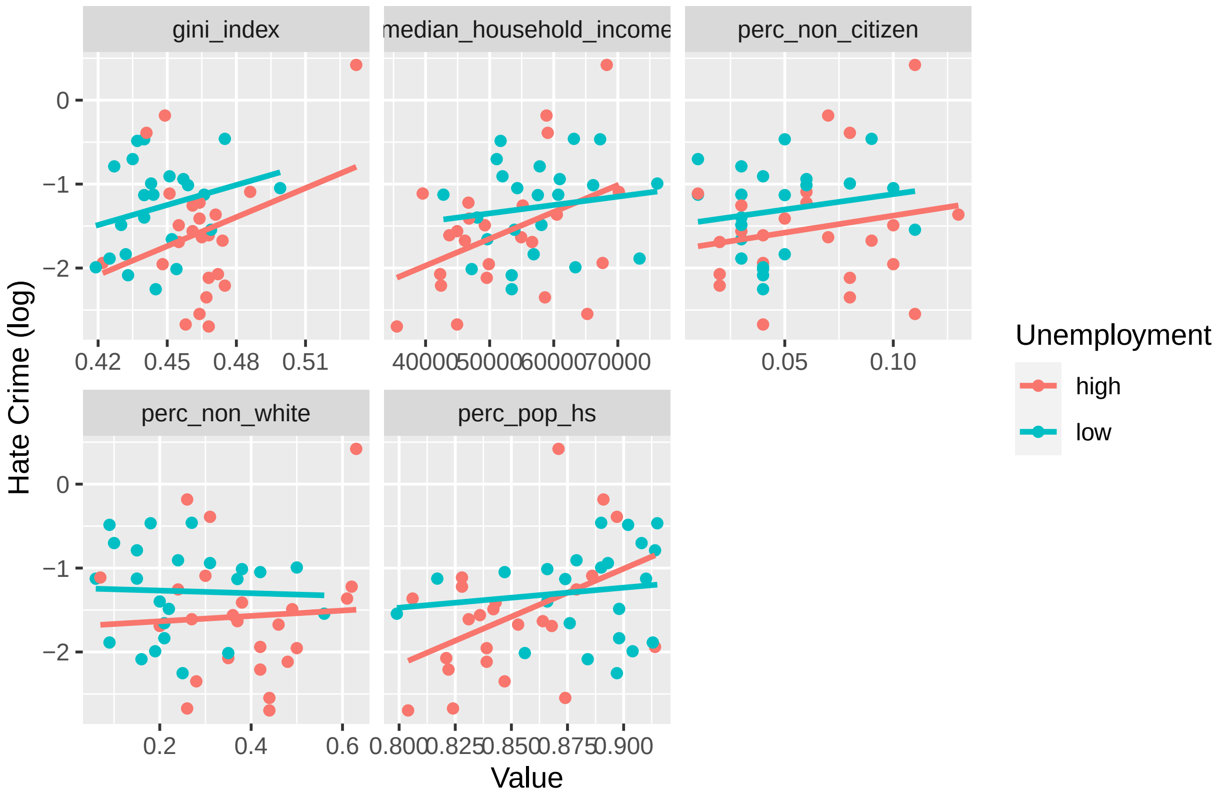

Figure 7. Interaction Plots Between Hate Crime and All Continuous Variable

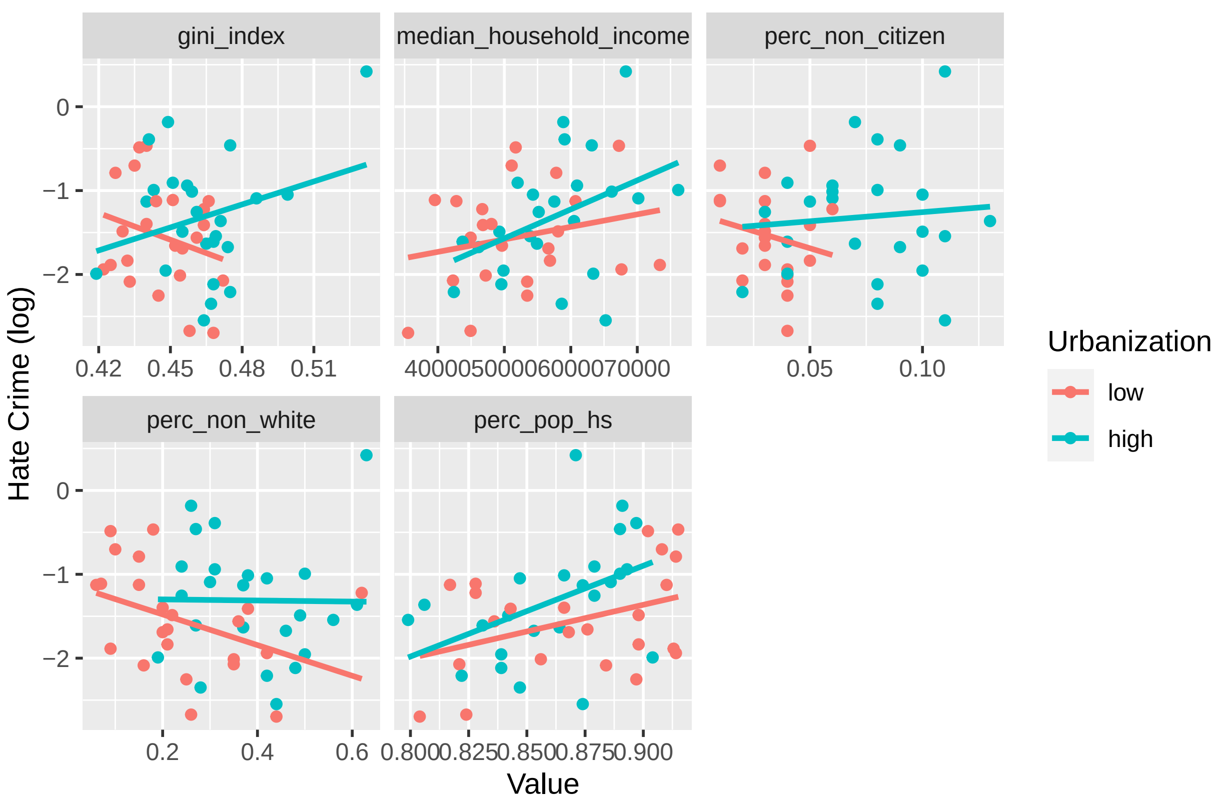

Figure 8. CP Statistic and Adjusted R−Squared Plots

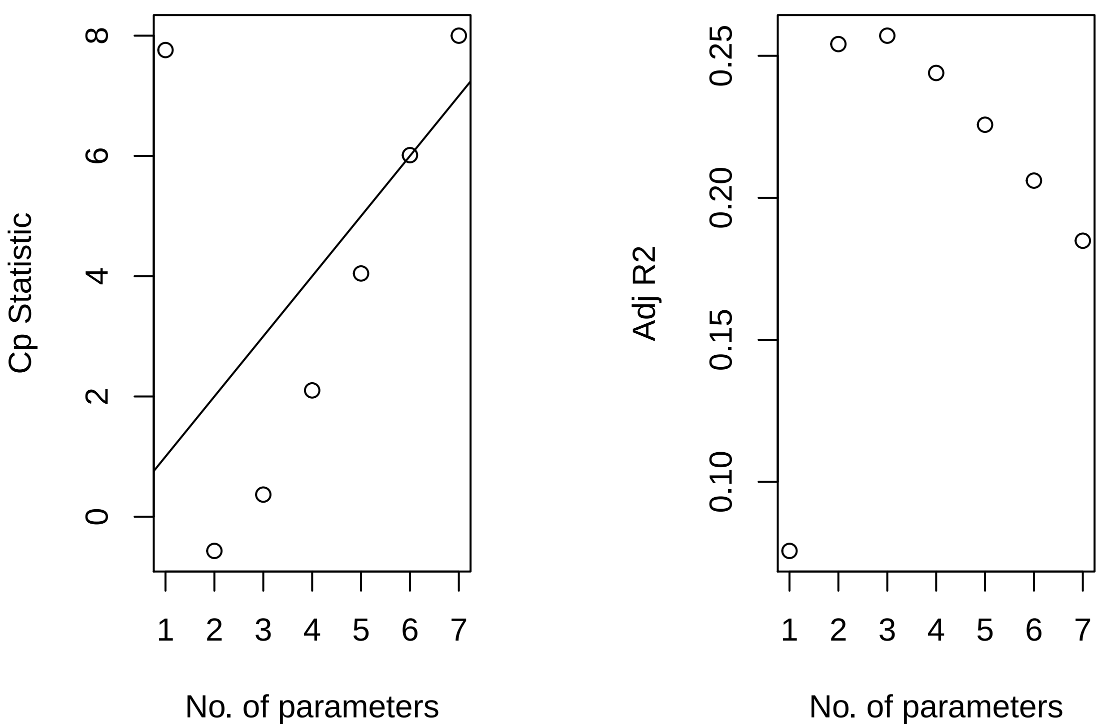

Table 3: Comparison of adjusted R^2 and RMSE with all models

	Model adjusted R^2	Model RMSE	CV adjusted R^2	CV RMSE
Two predictors with DC	0.2541	0.5417445	0.2943289	0.5948853
Three predictors with DC	0.2571	0.5341956	0.2783783	0.6038494
Two predictors without DC	0.1185	0.5367246	0.1530310	0.5554347
Three predictors without DC	0.1250	0.5281813	0.1730294	0.5600140

^^^ check these plots since the numbers seem weird

References

- "Incidents and Offenses." FBI, FBI, 29 Oct. 2019, ucr.fbi.gov/hate-crime/2019/topic-pages/incidents-and-offenses.
- Majumder, M. "Higher Rates Of Hate Crimes Are Tied To Income Inequality." FiveThirtyEight, FiveThirtyEight, 23 Jan. 2017, fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.
- Miller, C., Werner-Winslow, A. "Ten Days After: Harassment and Intimidation in the Aftermath of the Election." Southern Poverty Law Center, 29 Nov. 2016, www.splcenter.org/20161129/ten-days-after-harassment-and-intimidation-aftermath-election.

- Shively, Michael. Abt Associates Inc, 2005, Study of Literature and Legislation on Hate Crime in America.
- Van Dyke, N., and Tester, G. "Dangerous Climates." Journal of Contemporary Criminal Justice, vol. 30, no. 3, 2014, pp. 290–309., doi:10.1177/1043986214536666.

- “Variables Affecting Crime.” FBI, FBI, 5 Nov. 2012, ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime.