Project Report

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

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 state-level variables and rate of population-adjusted hate incidents using data reported to the Southern Poverty Law Center and analyzed in a 2017 FiveThrityEight article (Majumder, 2017).

Methods: We used data containing state-level hate crime rates per 100,000 individuals and several socioeconomic factors 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. We then used automated and criterion-based approaches to identify a regression model which optimizes parsimony and goodness of fit.

Multicollinearity of covariates, outliers, and variable interaction were also tested for and considered in model development. Model goodness of fit and predictive performance were assessed through model diagnostics and validation. All abovementioned steps were performed before and after removing any outliers in the data.

Results and Conclusions: 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. When these two predictors were the only parameters included in the model they were both positively associated with the outcome.

Multicollinearity tests showed possible correlation between median household income and percentage of population with a high school degree, as well as the percent non-citizen and percent non-white variables.

Our identified model optimizing goodness of fit and predictive performance contained only the Gini Index and percent population with a high school degree as parameters, both of which were significant. Washington DC, the single outlier in the data, was found to be an influential point, as its absence from the data nullified the significance of the Gini Index predictor in the two-parameter model.

Introduction

Hate crimes are a growing public health threat in the United States; 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]. 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 state-level hate crime rate and select socioeconomic variables (Majumder, 2017).

For this project, we used this dataset to critically analyze this article's findings to identify state-level socioeconomic variables significantly associated with hate crime rate, 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 (DC) was included as a state for the purposes of these analyses.

We then generated a table of descriptive statistics, including 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).

Using these data, our goal was 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 do so, we employed multivariable linear regression modeling, which operates under several assumptions, which include residual homoscedasticity (constant variance) and normality. To test whether these assumption were met, we first generated an overlaid density and histogram plot of population-adjusted hate incidents per 100,000 population, which showed a strong departure from standard normal distribution and significant right skewness (Figure 1). Thus, we performed a Box Cox transformation to isolate the 'best' power transformation on the hate crimes rate variable to achieve normal residuals (Figure 2).

To identify any outliers, and thus potential influential points, we plotted 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 outliers. Points outside of Cook's Distance on the Residuals vs Leverage plot 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 To investigate the existence of multicollinearity between each of the continuous variables, we generated a correlation matrix (Figure 5). 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 (Table 2).

Next, we plotted potential interactions between the continuous variables and each categorical variable, urbanization and unemployment, to see if the effect of each continuous variable on the log rate of hate crimes is different with respect to the levels of either urbanization or unemployment (Figures 6 and 7). 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 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. The most ideal model will have a Cp value less than or equal to the number of predictors in the 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 sum of squares error (the sum of the differences between each observed value and predicted value) but also penalizes for additional predictors. By this criterion, the best model will have the highest adjusted r-squared value. 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 8).

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 ensured that the model selected after this test did not violate any of the assumptions of linear regression by considering four main diagnostic plots: residuals vs fitted, normal Q-Q, scale-location, and residuals vs leverage.

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

To assess the predictive performance of the top two identified models, we performed 5-fold Cross Validation (CV) on each including and excluding outliers. We evaluated the model performance based on their root mean square errors (RMSE) and adjusted r-squared values generated from CV (Table 3). A lower RMSE is desirable as it indicates a larger proportion of variance explained by the model. We also compared the values generated from CV to the values from the original regression model to ensure the model performs well on new data.

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 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 (Figure 4).

A table of basic descriptive statistics for the collected socioeconomic variables is included in the appendix (Table 1). Notably, the hate crime rate per 100,000 individuals had a mean of 0.315 and range of 0.067 to 1.522, indicating considerable variation in state-level hate crime rates per 100,000 individuals. The Gini Index values, ranging from 0.419 to 0.532 across the included states, indicating substantial state-level income inequality in the overall dataset. A majority of states had both high urbanization and unemployment (51.1% each) (Table 1).

Multicollinearity and Interactions The correlation matrix and variance inflation factors (VIFs) show potential multicollinearity between percent non-citizen and percent non-white, and median household income and percentage of population with a high school degree. All other correlation coefficients do not suggest multi-collinearity. Correlation matrix and VIF results are visualized in Table 2 and Figure 5, respectively (Table 2, Figure 5).

For the dataset containing DC, 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 intersecting slope lines (Figure 6). 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 (Figure 7).

For the dataset excluding DC, the interaction plots show potential interaction between unemployment and Gini Index as well as between unemployment and percent population of high school graduates (Figure 8). 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 (Figure 8).

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). All subsequent model development used only the natural log transformed outcome.

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 in this model were the Gini Index, and the percent of high school graduates, both of which were significant. When DC was removed from the dataset, however, the Gini Index was no longer statistically significant at a 0.05 significance level in the two-parameter model, indicating that DC is an influential point. The adjusted R-squared of the two-predictor model decreases as well when DC 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 (Figure 8).

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 (Figure 8). 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 showed that adding the unemployment variable did not significantly improve the model. All of these results hold true when DC is removed from the dataset.

The diagnostic plots for the two-covariate model including DC showed no obvious violation of the linear regression assumptions (Figure 9). The points in the Residuals vs Fitted plot were randomly scattered around the line at 0 indicating the assumption of homoscedasticity was met. The points followed a straight line in the Q-Q plot indicating the assumption of normality was met. The Residuals vs Leverage plot, however, indicated that point 9 (DC) may be an outlier. When DC was removed, all diagnostic plots remained similar except that there were no longer any outliers indicated in the Residuals vs Leverage plot.

The association between each of the two included variables and the natural log of hate crime rates was positive whether DC was included in the data or not. For a one unit increase in the Gini Index, there was a 16.486 expected increase in the natural log of hate crime rates, holding the percent of high school graduates constant (when excluding DC this expected increase is 10.811). For a one unit increase in the percent of high school graduates, there was a 11.554 expected increase in the natural log of hate crime rates holding the gini index constant (when excluding DC the expected increase is 9.509).

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

The two-predictor model, whether DC is included or not, performs well as a predictive model since the model's RMSE and adjusted r-squared were very similar to the CV average RMSE and adjusted r-squared.

Discussion/Conclusion

Interpretation of findings After extensive analyses and modeling, the Gini Index and percent population with a high school degree variables were the only covariates found to be significantly and independently associated with hate crime rate. However, the model containing only these parameters, despite having the most optimal goodness of fit and predictive performance, accounted for only 25.4% of the variability in the data.

This implies that there may be other variables yet unaccounted for in model development which may predict hate crime rate. For example, existing research has found sexuality, gender identity, and religion to be factors significantly associated with hate crime rate ("FBI: Incidents and Offenses," 2019).

The positive association between Gini Index and log transformed hate crime rate indicates that high wealth inequality is associated with high hate crime rate. Interestingly, the association between percent high school graduates and log transformed hate crime was also positive. This result may be a reflection of the presence of an unmeasured confounding variable causing a spurious association or a more complex phenomenon occurring altogether. For example, the US Department of Justice has reported dramatic increases in school-based hate crimes in recent years, which could serve as an explanation for higher hate crime rates among states with high proportions of high school graduates (Keierleber, 2018).

Limitations 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 country and could potentially influence the significance of the predictors and/or the predictive ability of our models. As a result, these findings may not be generalizable to the entire United States.

While testing for variable interactions, we also only considered interactions between the categorical variables and each continuous covariate due to their practical interpretability. However, it is possible that other interactions may exist between the continuous variables, and even three-way interactions that could be incorporated in model development to explain more variability in the data.

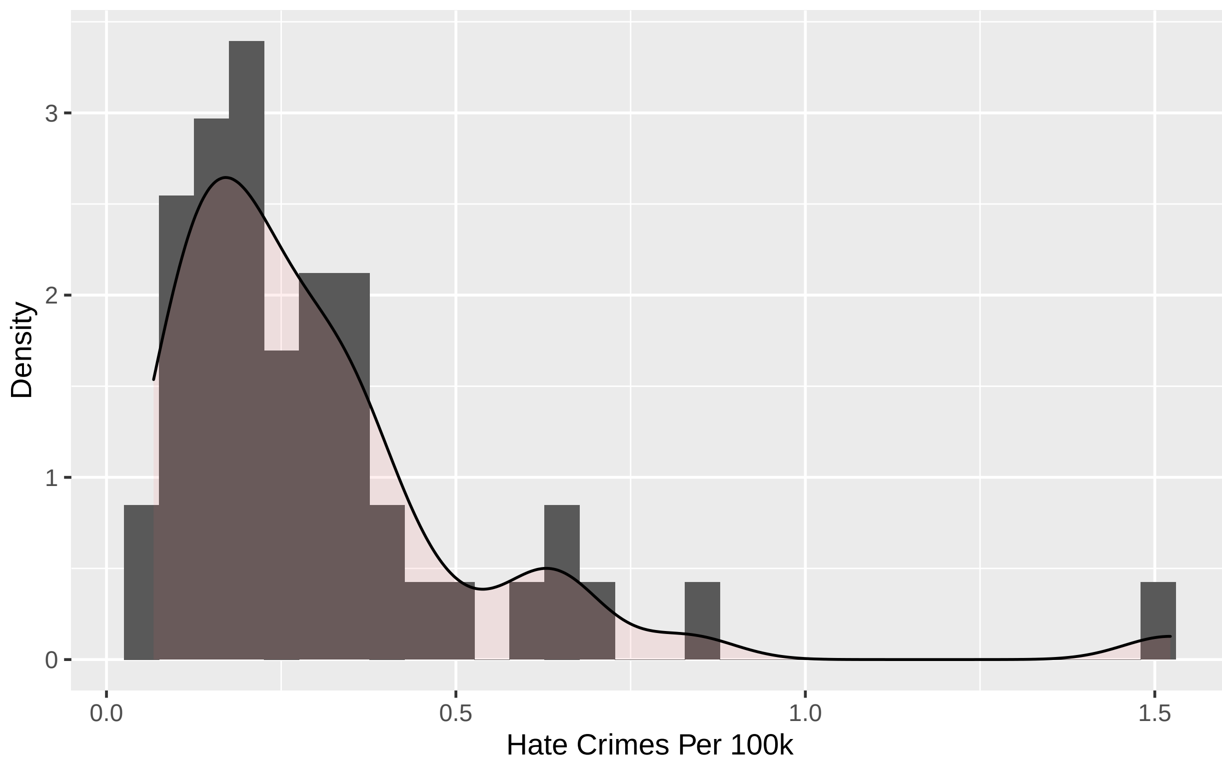
Take-aways Overall, this project identified several significant state-level demographic variables that are associated with hate crime rates. This is of practical importance for health and crime policy implementation and resource allocation. Future work might seek to hone model goodness of fit and predictive performance through consideration of additional variables, such as percent religious minorities.

Figures and Tables

Table 1: Descriptive Statistics of States Reporting Hate Crimes

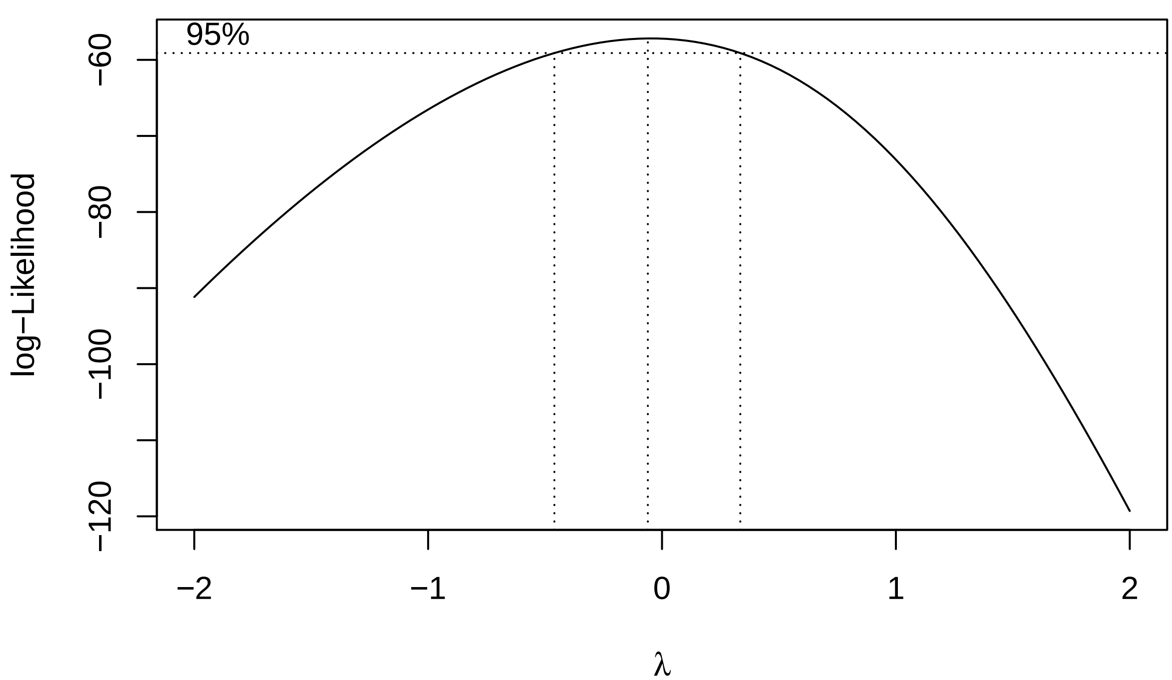
	Overall (N=47)
Unemployment	
high	24 (51.1%)
low	23 (48.9%)
Missing	0
Urbanization	
low	23 (48.9%)
high	24 (51.1%)
Missing	0
Median Household Income	
Mean (SD)	54802.298 (9255.117)
Median (Q1, Q3)	54310.000 (47629.500, 60597.500)
Min - Max	35521.000 - 76165.000
Missing	0
% Adults >25yrs With HS Degree	
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

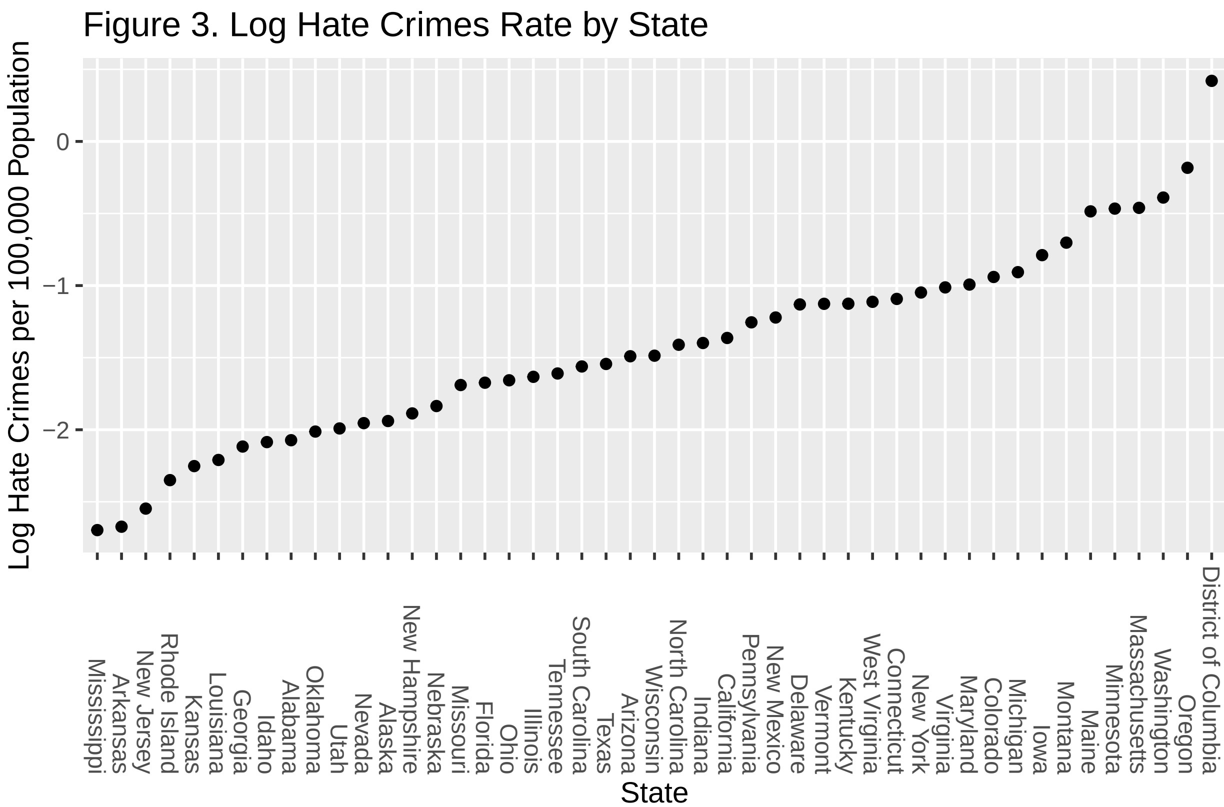
Figure 1. Untransformed distribution of hate crimes per 100k population



ate crimes across the states is clearly right−skewed, with a majority of states between 0.1−0.5 hate crimes per 100k.

Figure 2. BoxCox Transformation Plot





States are arranged in increasing order of log hate crime. DC's hate crime rate seems abnormally high

Figure 4. Full Model Hate Crime Diagnostic Plots

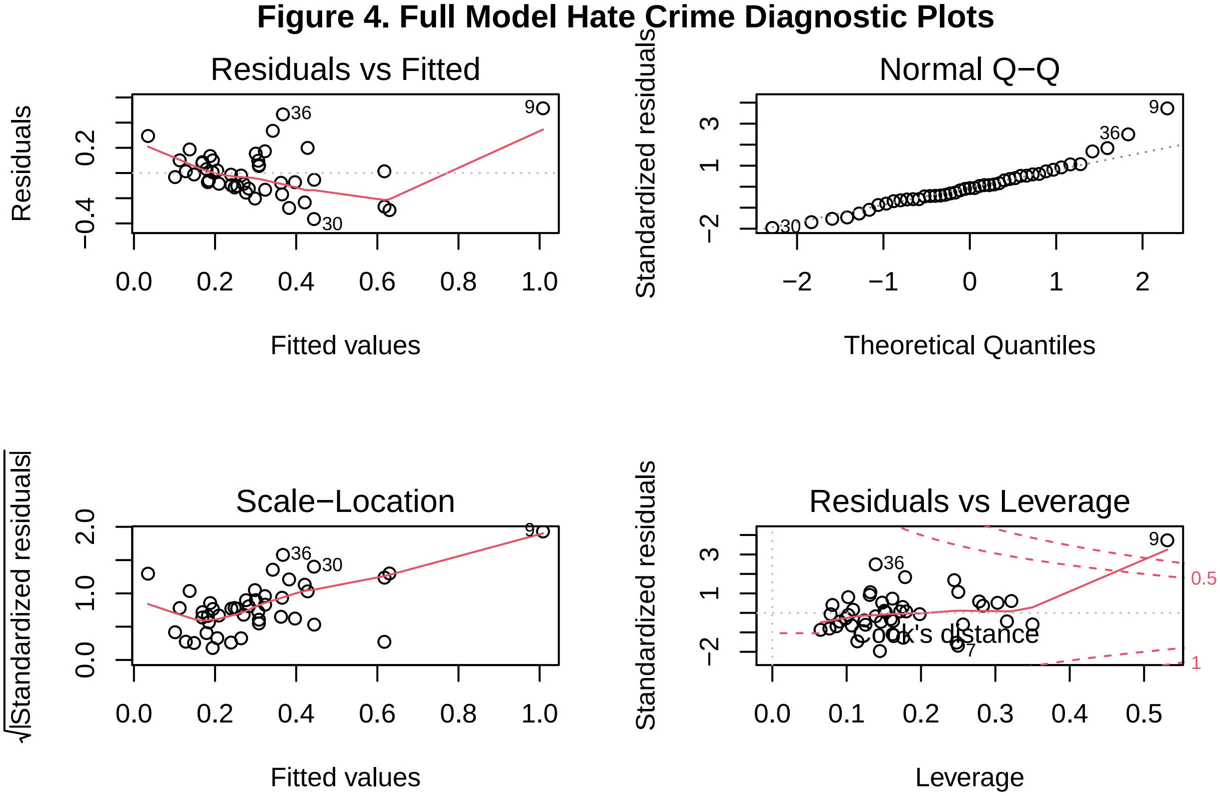


Figure #. Two predictors without DC model Hate Crime Diagnostic Plots

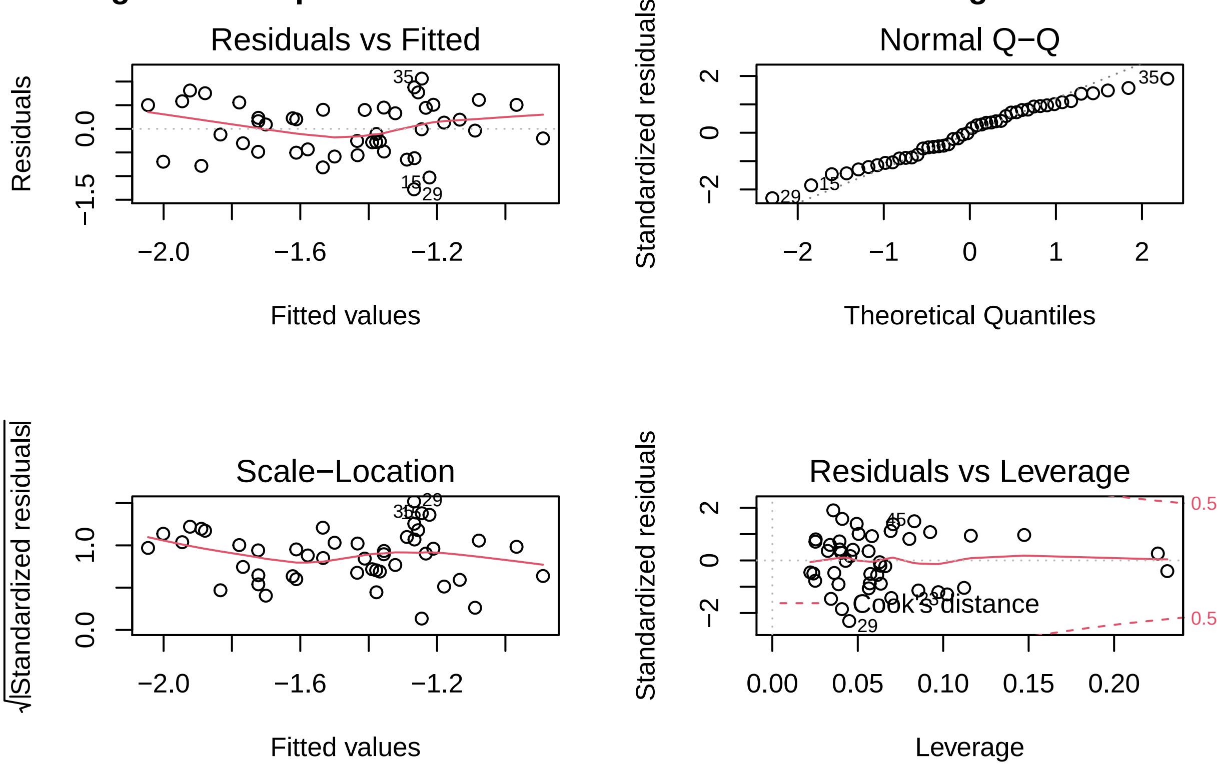


Figure 5. Correlation Matrix

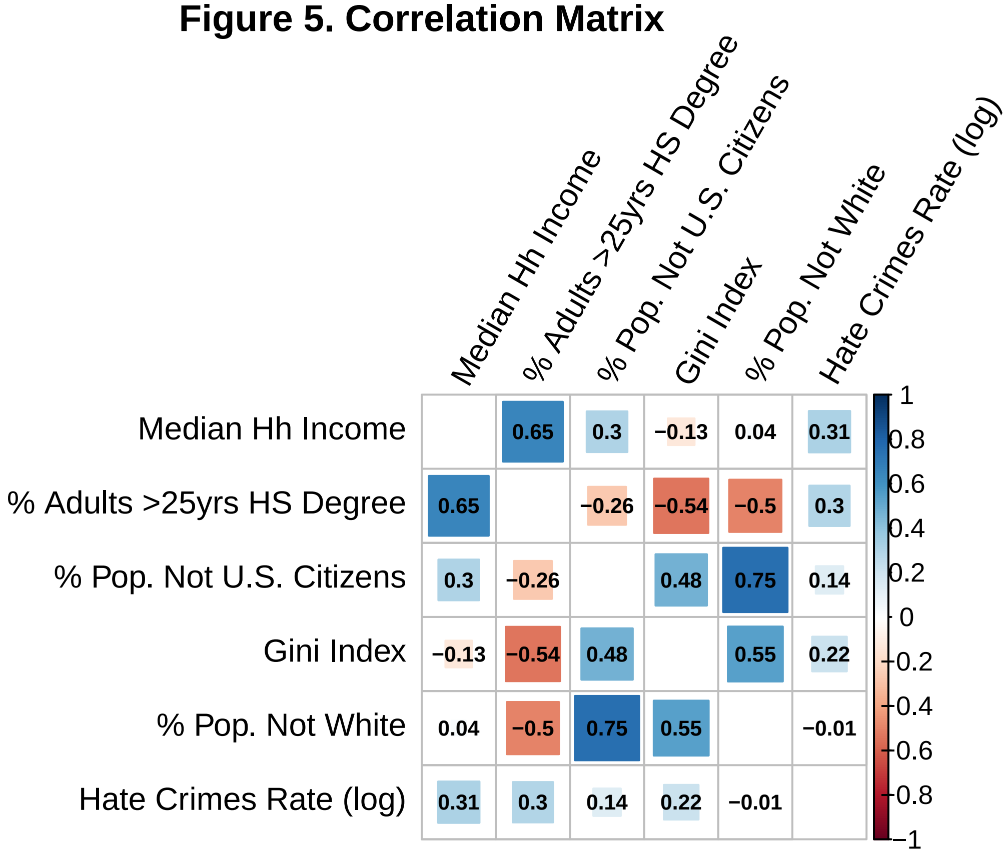
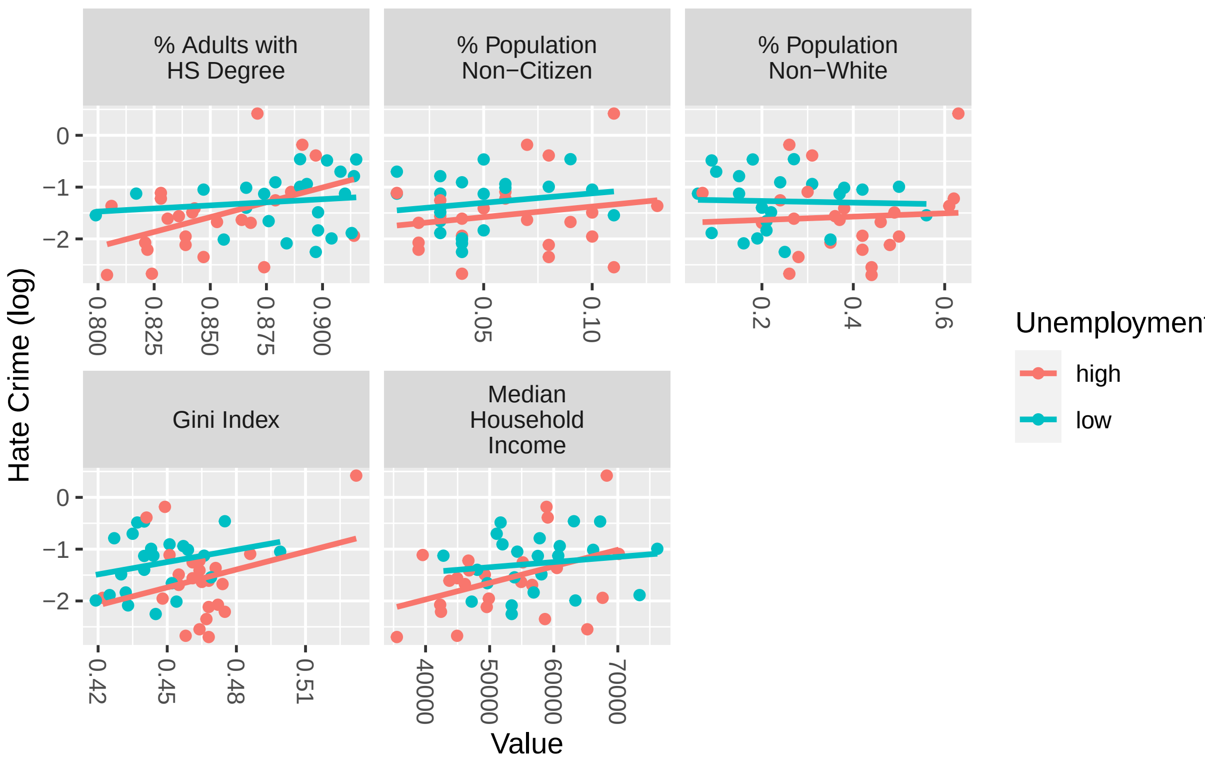


Table 2: Table of variance inflation factors for each variable.

	VIF
‘Unemployment ‘	1.426
‘Urbanization ‘	1.983
‘Median Household Income‘	3.108
‘% Adults >25yrs With HS Degree‘	3.895
‘% of Population Not U.S. Citizens‘	3.728
‘Gini Index‘	1.845
‘% of Population Not White‘	3.236

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



Data comes from FiveThirtyEight Hate Crimes data

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

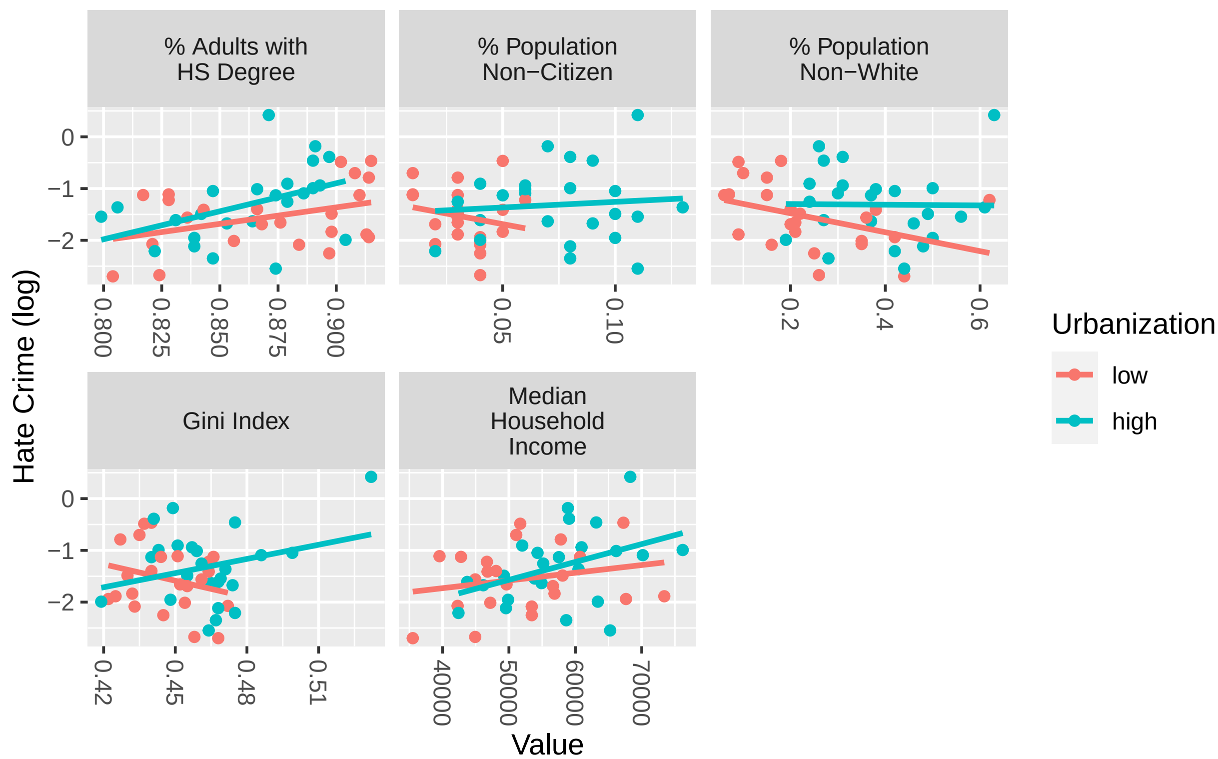


Figure 8. Mallow's Cp and Adjusted R−Squared by # Parameters

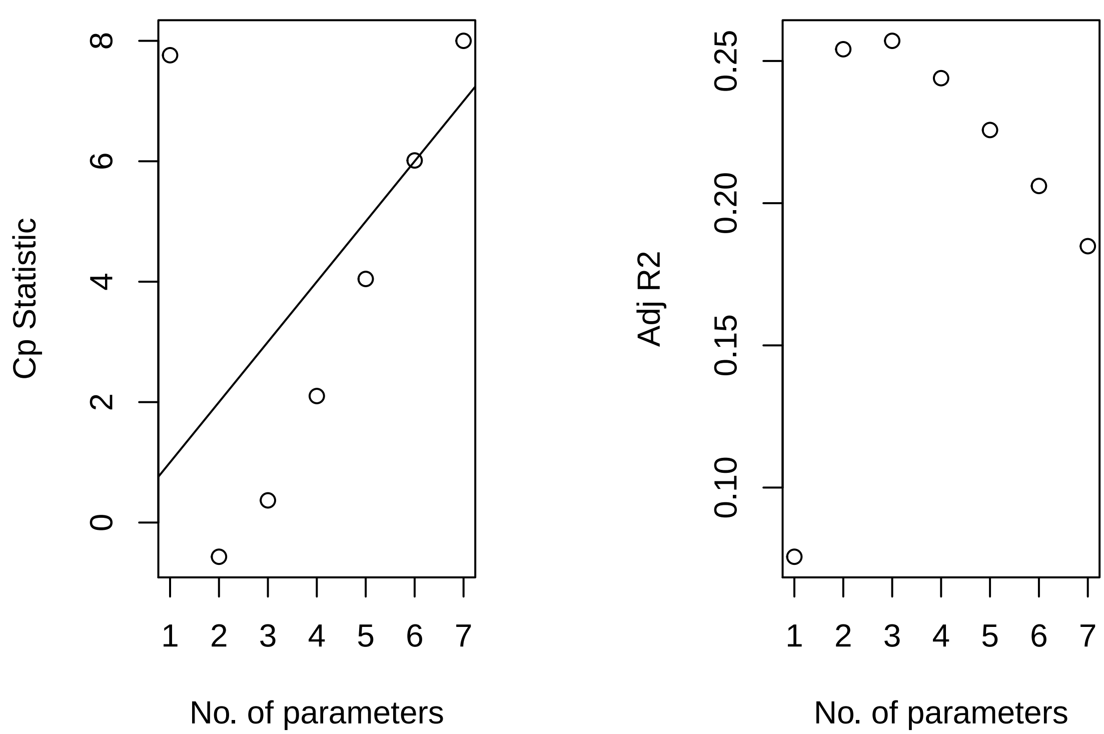


Table 3: Comparison of Adjusted R^2 and RMSE for All Models

	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

References

- "Incidents and Offenses." FBI, FBI, 29 Oct. 2019, ucr.fbi.gov/hate-crime/2019/topic-pages/incidents-and-offenses.
- Keierleber, M. "New FBI Data: School-Based Hate Crimes Jumped 25 Percent Last Year - for the Second Year in a Row." 2018, www.the74million.org/new-fbi-data-school-based-hate-crimes-jumped-25-percent-last-year-for-the-second-year-in-a-row/.
- 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.