Evaluation of Main Predictors for Hate Crimes

Ying Jin, Yining Xiang, Zhourong Li, Yu Liu

1.Abstract

In this project, we are identifying the meaningful predictors for hate crimes, and checking the associations between potential variables. Through the process of data visualization and descriptive statistics, reduction of explanatory variables, model selection, validation and diagnostics, we find the most significant predictors for hate crimes are percentage of adults(>25yrs) with a high school degree and Gini index (measuring income inequality) when adjusting for others. We also find there is a significant interaction between level of state unemployment and Gini index.

2. Introduction

Since 2014, the number of hate crimes has been increasing every year and reached a total of 7,314 [1] reported incidents and 59 deaths in 2019 [2]. Though the total number is not comparable to violent crimes and property crimes, hate crimes is under the attention of the Civil Rights Unit supported by the U.S. Attorney's Office, addressed as a violation of civil rights against the whole society [3]. By FBI's definition, hate crime is the "criminal offense against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender or gender identity." [4] It is highly related to the offender's personal will based on his/her bias but violating the boundary between freedom of speech and illegitimacy.

According to the nature of hate crime, it is highly randomized and unpredictable and has brought difficulty to the FBI's strategy to combat this problem. Unlike other crime types, like theft and robbery, primarily driven by poverty and economically inequality, the motivation of hate crimes is hard to be

elaborated. Though defined to be in relation with one's bias, the factors measuring bias are hard to be counted as statistics and used as prediction of future incidences. Does it mean hate crimes are not associated with any social measurements except bias? That is not the case because there has been analysis by FBI and Southern Poverty Law center data suggests that from the scope of the whole country, states with higher income inequality are likely to have higher hate crime incidents [5]. Such analysis could be a clue to the investigation of association between hate crimes and social measurements.

In this project, we are investigating the association between hate crimes rate per 100,000 population and social measurements including unemployment, urbanization, income inequality, education level, racial heterogeneity level etc. The data are describing the behavior of each states in the US aggregately, omitting the detailed performances of individuals. In the analysis, we are identifying the association of each variables with the incidences and the relation within the variables on this case.

3. Methods

The dataset that we work on contains detailed information on hate crime which happened in 51 states in the United States in 2016. This dataset was adapted from the one used by a FiveThirtyEight article to analyze the same topic. Our dataset contains 51 rows, corresponding to 51 states, and 9 columns. After dropping NAs, there are 45 rows in our dataset (See description for variables in table 1).

First, we do the data exploration. The distribution of hate crime rate is significantly right skewed (Fig. 1), which is also indicated by the Q-Q plot (Fig. 2). After taking the log transformation, the distribution is much more normal (Fig. 3 & Fig. 4). In this way, we decide to take the log of the hate crime rate as the outcome in our model.

Noticing that the hate crime rate is very high in the District of Columbia (1.522 per 100,000 population, while median = 0.226 per 100,000 population), we make the plot of residuals vs. leverage to find out

whether it's an influential observation (Fig. 5). As the plot shows, case 9, which is the observation for the District of Columbia, is close to the dashed lines. Moreover, before removing case 9, there is positive linear relationship between Gini index and hate crime rate (Fig. 6), while after removing it, there is slightly negative association between Gini index and hate crime rate (Fig. 7). So we consider the observation for the District of Columbia as an influential outlier and omit it when conducting further analysis.

Another issue to be addressed is the intercorrelation between potential predictors (Table 2). Fortunately, none of the VIF of these variables exceed 5. The results of the ANOVA test indicate that adding perc_population_with_high_school_degree, which has the highest VIF (4.41), is not redundant (p-value = 0.15) when setting α as 0.15 and improves Adj-R^2 from 0.027 to 0.057. However, the same procedure implies that perc_non_citizen should not be included into our model (p-value = 0.77).

Then we do model selection**.** To begin with, we use automatic procedure to identify different best models of different sizes and evaluate them based on multiple criteria: Cp, Adj-R^2 and BIC (Fig. 8). According to these plots, we think that the performance of models with two, three and four predictors (corresponding to models with three, four and five parameters) are similar and decide to further evaluate them with cross validation (Fig. 9). In this plot, we can see that the RMSE of the model with two predictors is a little smaller than the other two candidates, which indicates better fitness. Combined the rule of parsimony, the model with 2 predictors: gini_index and perc_population_with_high_school_degree is selected as our basic model (Table 3).

Next step is model modification. We further conduct a test to see if there exists interaction between Gini index and unemployment by drawing a plot of Gini index vs. hate crime rate stratified by unemployment (Fig. 10). The cross between the two slopes indicates likely interaction between Gini index and unemployment. We then add the interaction term into our basic model of which p-value is equal to

0.094 (α = 0.10). The Adj-R^2 increases from 0.12 to 0.17 (more than 6%), meaning adding this interaction term improved the model fitness.

Since there is an interaction, we perform the stratified analysis by different unemployment levels. Among states with low unemployment, there exists statistically significant positive association (19.37) between Gini index and hate crime rate (p-value = 0.018). While in states with high unemployment, insignificant negative association (-1.18) between Gini index and hate crime rate (p-value = 0.92) is observed.

We then conduct an ANOVA test to compare our new model (with the interaction term) to our basic model and result a p-value of 0.13 (Table 4). In consideration of the improvement in adjusted R-squared, we decide to keep the interaction term in the model and tolerate an α=0.15 for now.

Now we compare the predictive ability between the new model (with the interaction term between Gini index and unemployment) and the basic model. Though from figure 11, we find it is hard to see any appreciable difference between our new model and the basic one.

Last step, we do the model diagnostic and find that the observation 2 might be an influential point according to its Cook's distance (Fig.12). So here we fit another regression model without the observation 2. The result (Fig.13) (Table.5) suggests a great improvement of fitness of the large model (adj-R^2 increases from 0.16 to 0.25). We further conduct the cross-validation based on the filtered data (without observation 2), and the large model performs better in terms of RMSE.

4. Results

After automatic stepwise selection, evaluating through multiple criteria, further testing of interaction and analysis, and omitting two influential outliers, we find our overall best fitted model would be:

$$\ln(hate\ crimes) = \hat{\beta}_0 + \hat{\beta}_1 gini\ index + \hat{\beta}_2 perc\ high\ school\ degree + \hat{\beta}_3 gini\ index * unemployment$$

which contains three predictors, with p-value = 0.0042, and adjusted R-squared value of 0.25 (Table.5). In addition, there is significant negative association between Gini index and hate crime rate in states with low unemployment. While in those with high unemployment, the association doesn't exist.

5.Conclusion and Discussion

In this project, we start with exploring the distribution of outcome and making a logarithm transformation. Then we identify potential outliers and exclude one influential points. After that we study the correlation between all the predictors and figure out that per_non_citizen should be excluded as it has strong correlation with several other variables. Also including this predictor to the model doesn't improve the fitness of model. We identify the significant predictors associated with hate crime rate by fitting a linear regression model, and find that only percentage of adults (>25yrs) with a high school degree and Gini index are associated with hate crime rate while adjusting for other variables of interest (at 0.15 significance level). In order to study potential effect of other variables, we stratify the data by level of unemployment and level of urbanization respectively and find an interaction between unemployment and Gini index.

Still, there are several limitations of our model and findings. First, we do not verify the potential confounding (or interaction) of household income and percentage of population that are non-white with the two main predictors. In order to test the existence of the confounding effect, we need to find meaningful cutoff value for these variables. Second, aiming to increase the goodness of fitting of our final model, we sacrifice variance to some extent. Finally, the motivation of hate crimes is hard to be elaborated, so with the limited variables, our findings cannot explain the complex issue thoroughly.

Reference

[1] *Incidents and Offenses*. FBI. (2020). Retrieved 16 December 2020, from https://ucr.fbi.gov/hate-crime/2019/topic-pages/incidents-and-offenses.

[2] *US hate crime highest in more than a decade - FBI*. BBC News. (2020). Retrieved 16 December 2020, from https://www.bbc.com/news/world-us-canada-54968498.

[3] *The Civil Rights Program*. Justice.gov. (2020). Retrieved 16 December 2020, from https://www.justice.gov/usao-wdtn/civil-rights-program.

[4] *Hate Crimes | Federal Bureau of Investigation*. Federal Bureau of Investigation. (2020). Retrieved 16 December 2020, from https://www.fbi.gov/investigate/civil-rights/hate-crimes.

[5] Majumder, M. (2020). *Higher Rates Of Hate Crimes Are Tied To Income Inequality*. FiveThirtyEight. Retrieved 16 December 2020, from https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.

Appendix

Table 1: Variable Description

Notation	Meaning
hate_crimes_per_100k_splc	hate crime rate per 100,000 population
median_household_income	median household income per state
perc_population_with_high_school_degree	percentage of adults (>25 yrs.) with a high school degree
perc_non_citizen	percentage of population that are not US citizens
perc_non_white	percentage of population that are non-white
gini_index	index measuring income inequality
unemployment	level of state unemployment
urbanization	level of state urbanization

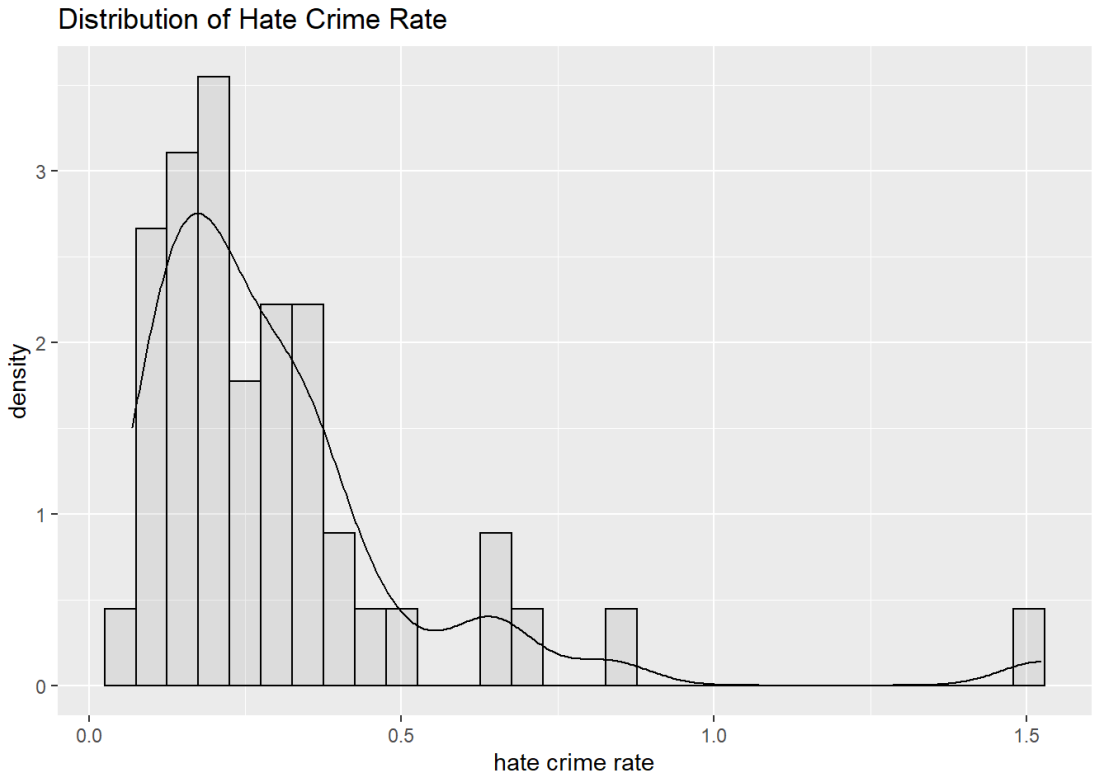
Distribution of Hate Crime Rate
density
hate crime rate
0
1
2
3
0.0
0.5
1.0
1.5

Fig. 1

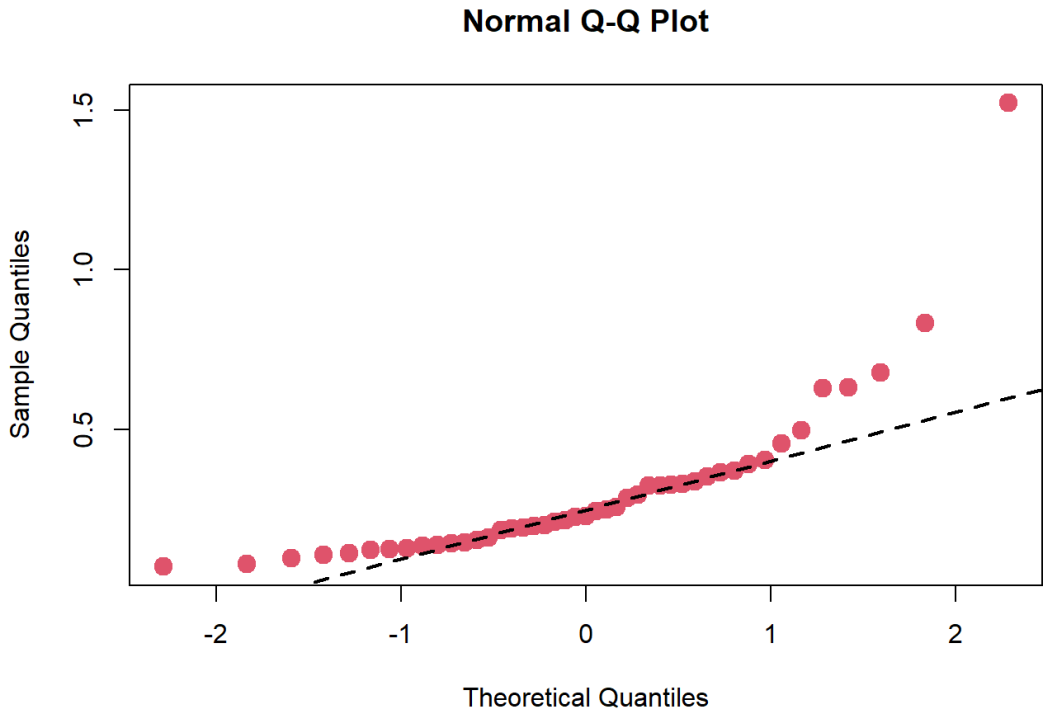
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0.5
1.0
1.5
-2
-1
0
1
2

Fig. 2

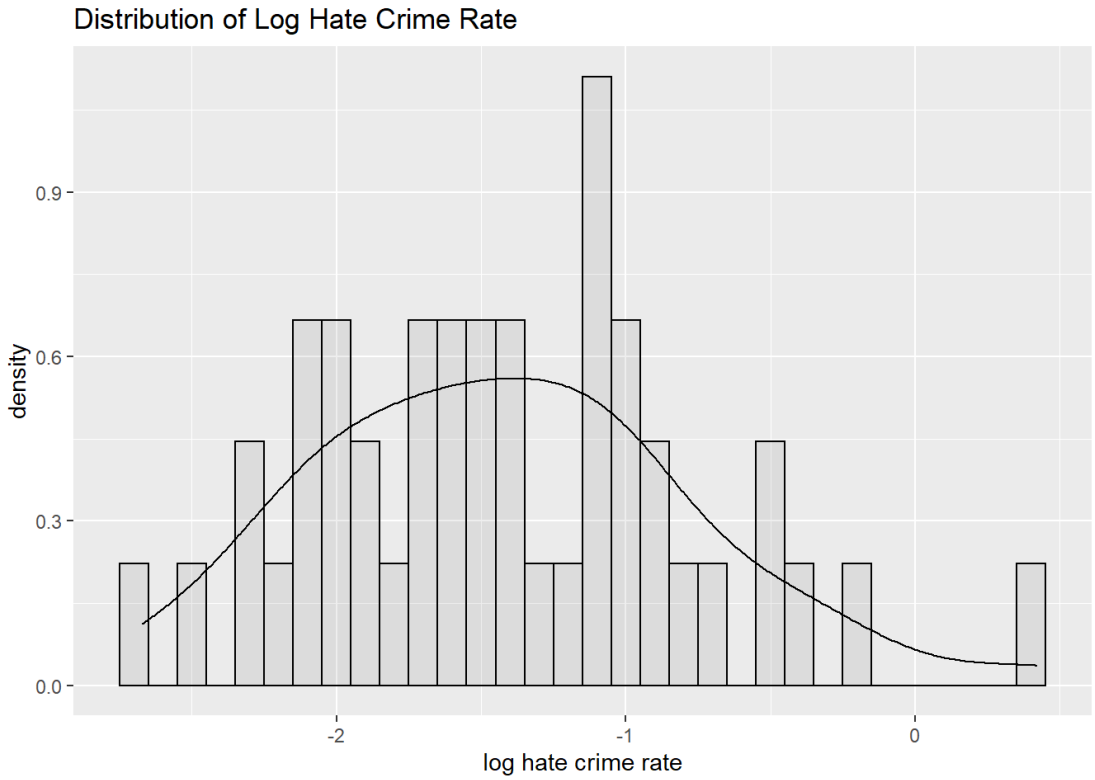
Distribution of Log Hate Crime Rate
density
0.0
0.3
0.6
0.9
-2
-1
0
log hate crime rate

Fig. 3

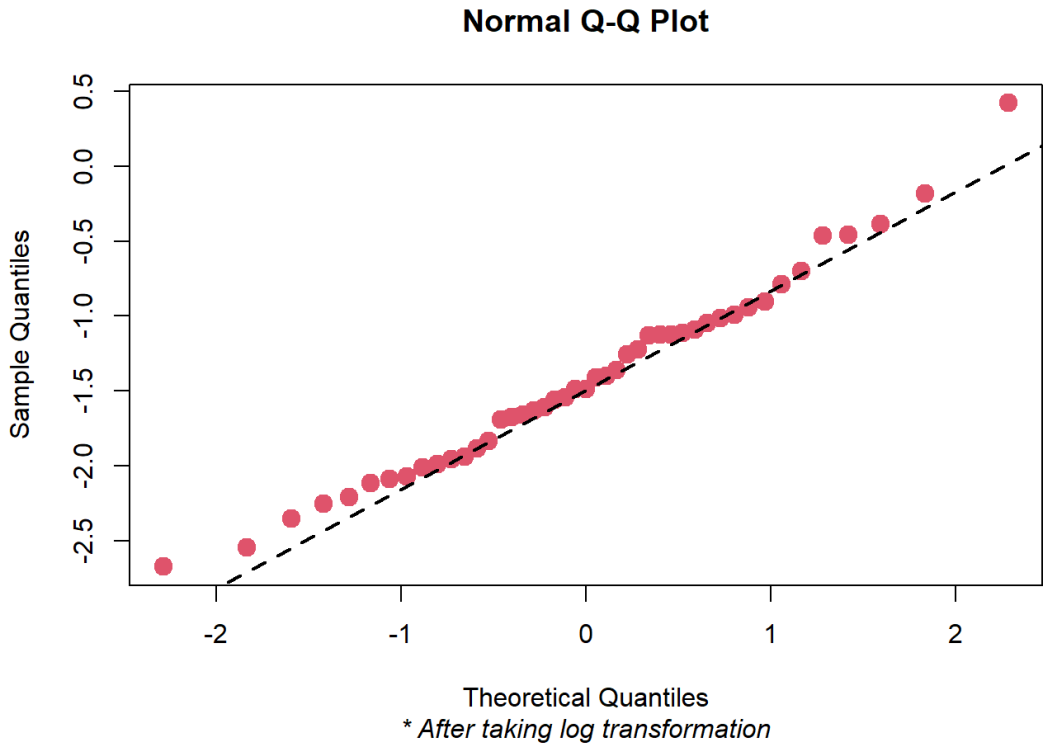
Normal Q-Q Plot
Sample Quantiles
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
-2
-1
0
1
2
Theoretical Quantiles
* After taking log transformation

Fig. 4

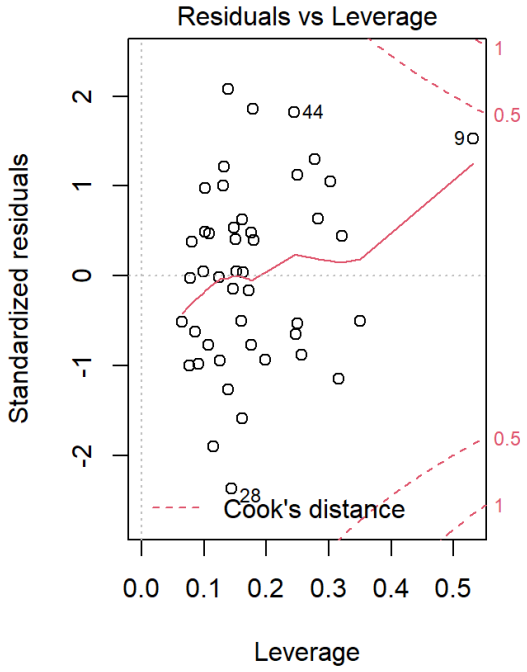
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
44
9
28
1
0.5
0.5
1
2
1
0
-1
-2
0.0
0.1
0.2
0.3
0.4
0.5

Fig. 5

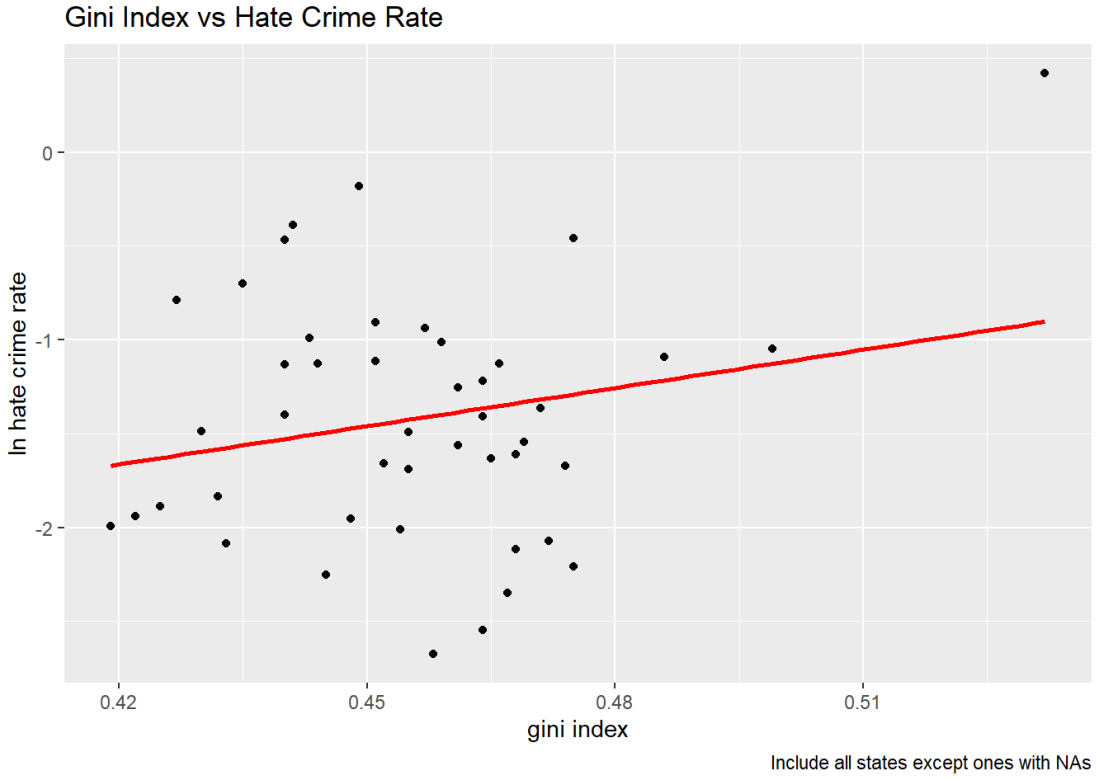
Gini Index vs Hate Crime Rate
ln hate crime rate
gini index
0
-1
-2
0.42
0.45
0.48
0.51
Include all states except ones with NAs

Fig. 6

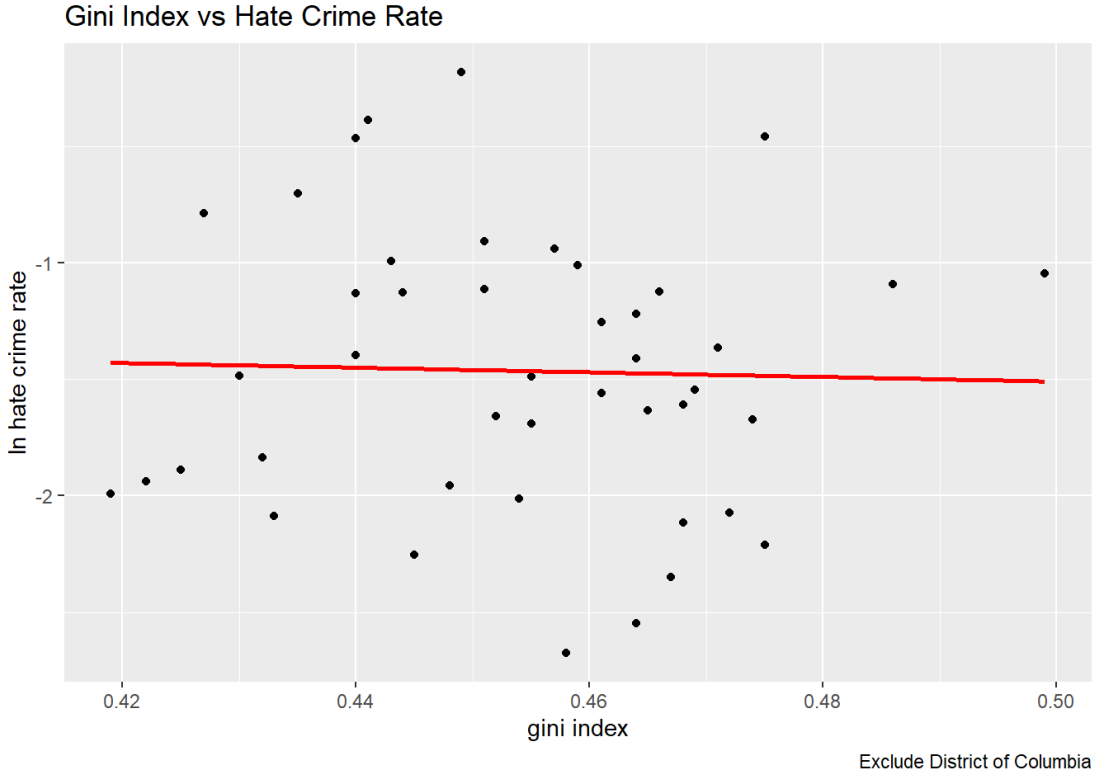

Fig. 7

Table 2: Correlation Matrix

	perc_non_citizen	median_household_income	gini_index
perc_non_white	.73	-	-
urbanization	.67	-	-
perc_population_with_high_school_degree	-	.66	-.66

* *Only contains variables between which* $\rho > 0.6$

Model Selection Based on Multiple Criteria

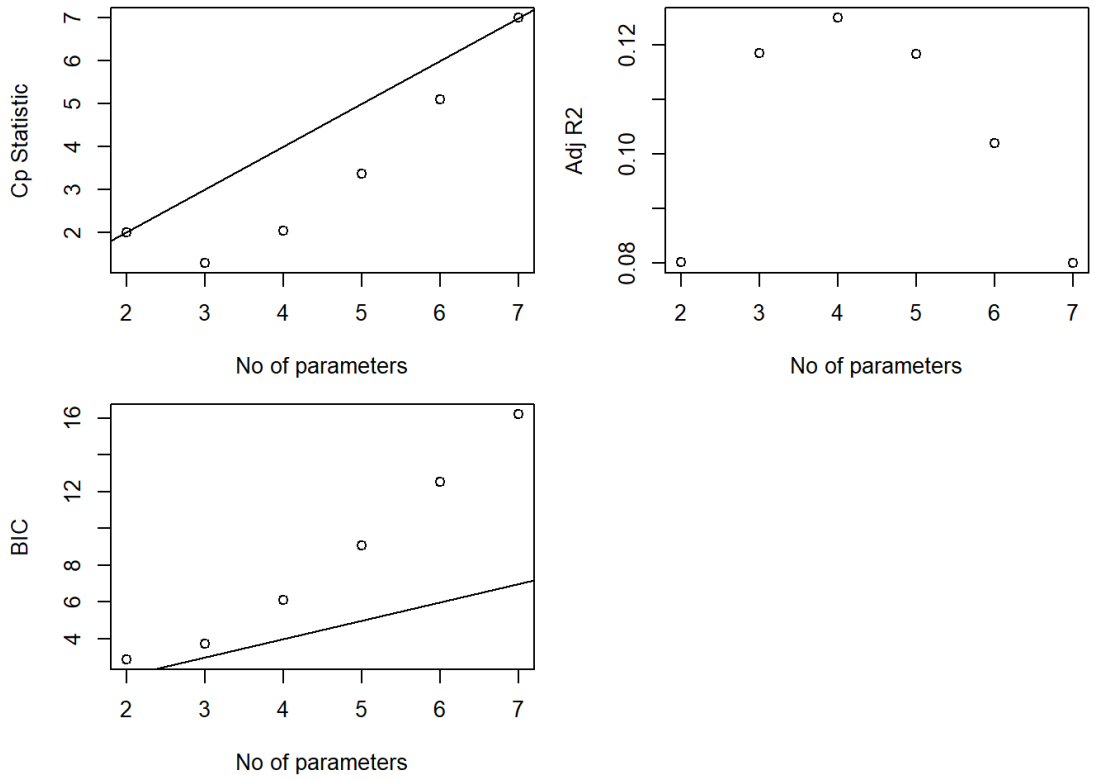

Fig. 8

Model Comparison for RMSE –Criteria Based Selection

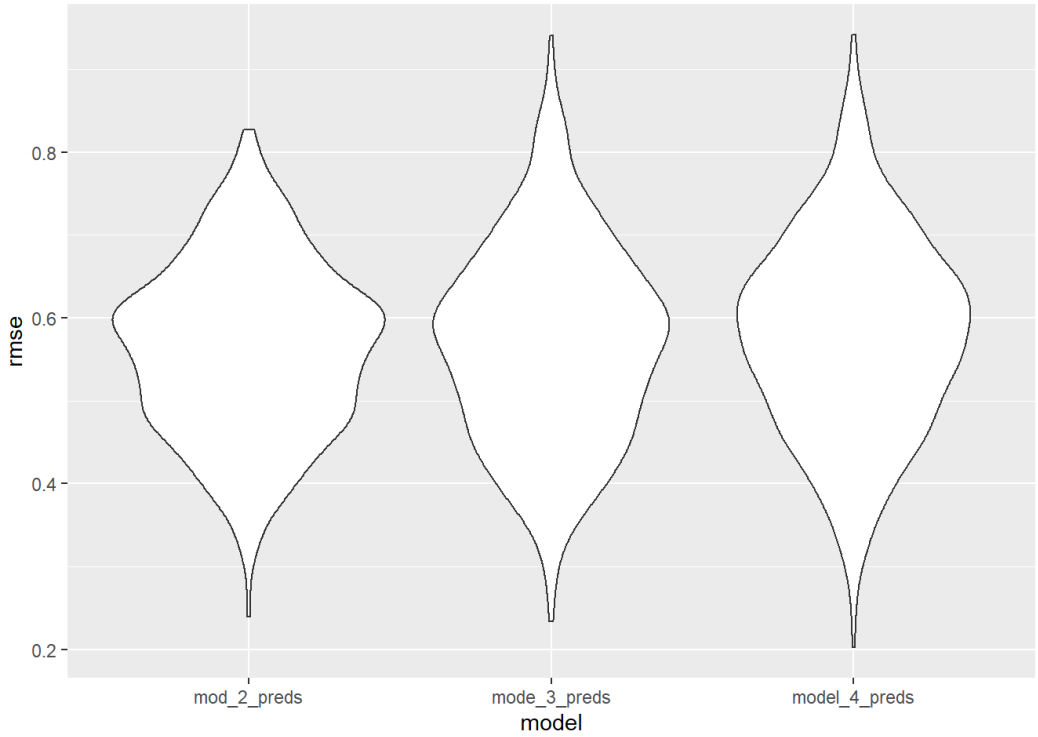

Fig. 9

Table 3: Regression Results for the Basic Model

Term	Estimate	Std.error	Statistic	P.value
(Intercept)	-14.611	5.359	-2.726	0.009
perc_population_with_high_school_degree	9.509	3.419	2.781	0.008
gini_index	10.811	6.429	1.682	0.100

Gini Index vs Hate Crime Rate Stratified by Unemployment

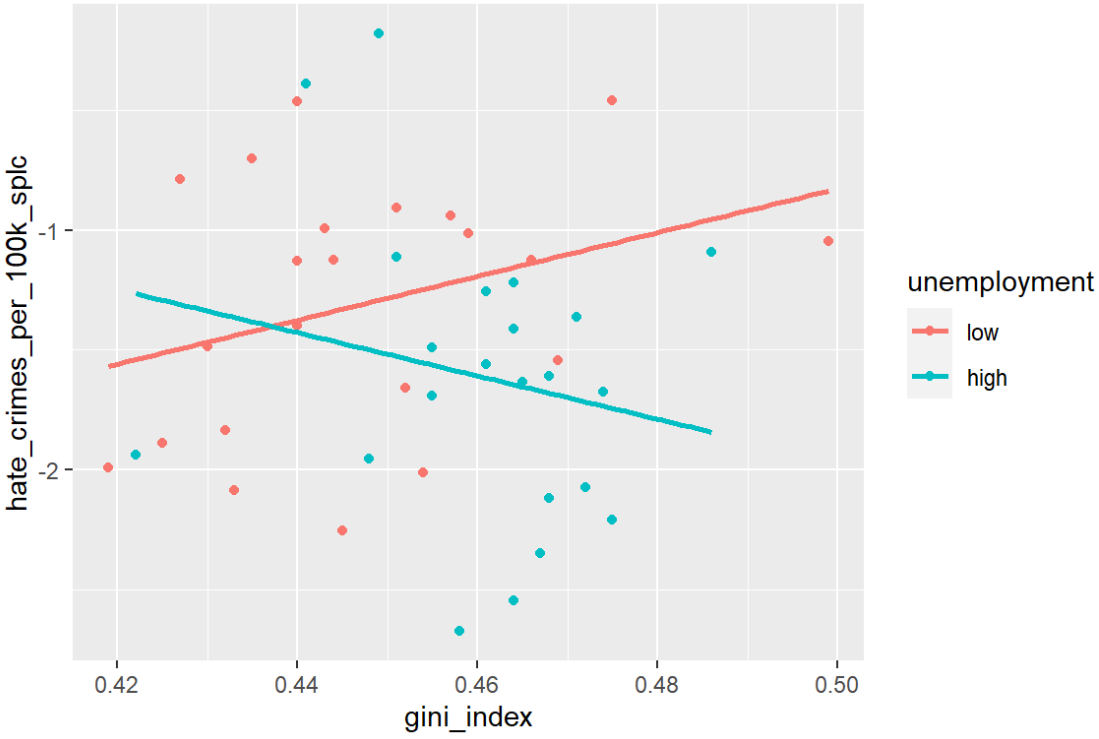

Fig. 10

Model Comparison for RMSE –Model Modification1

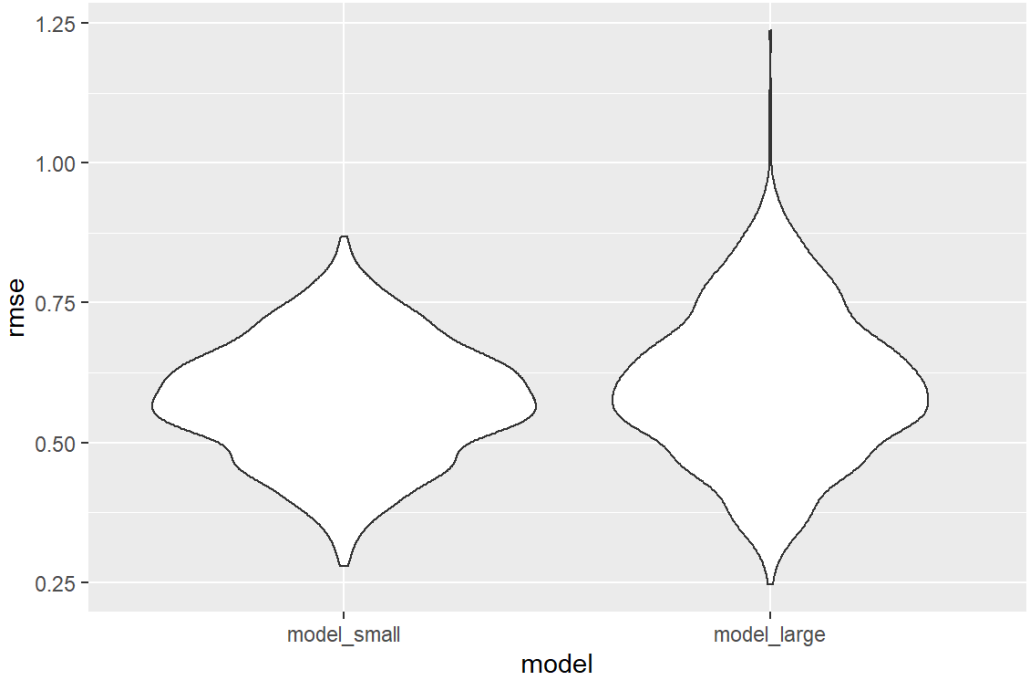

Fig. 11

Model Diagnostic 1

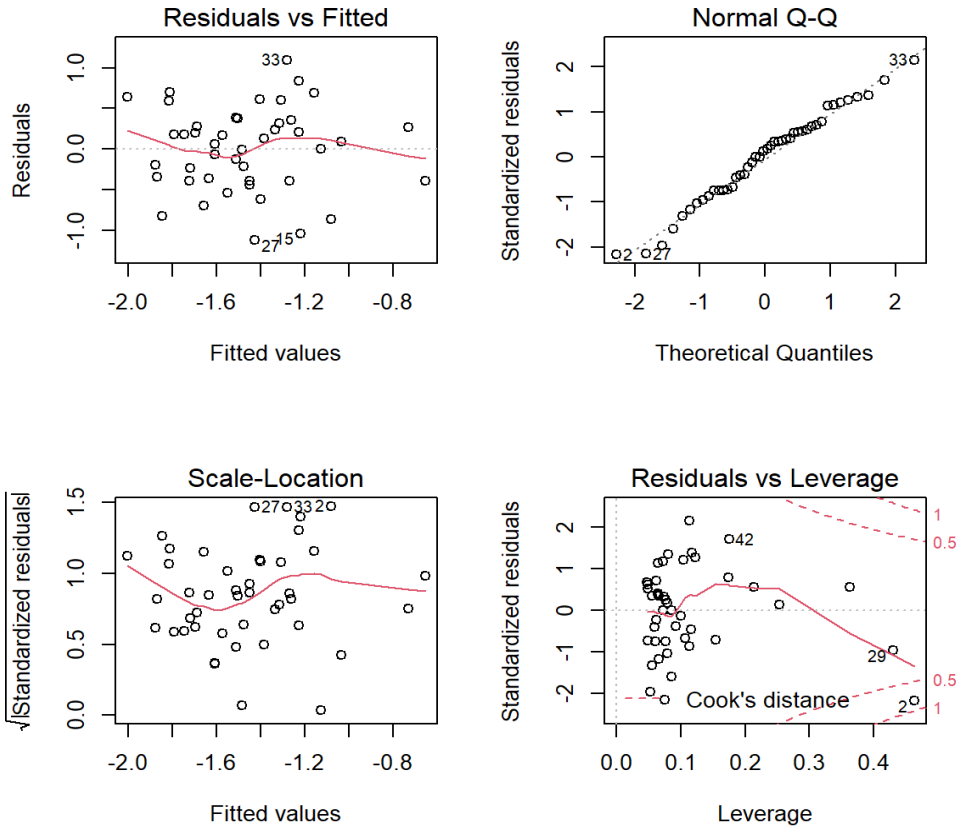

Fig. 12

Table 4: Regression Results for the Modified Model

Term	Estimate	Std.error	Statistic	p.value
(Intercept)	-16.897	5.517	-3.063	0.004
perc_population_with_high_school_degree	8.427	3.474	2.426	0.020
gini_index	18.247	7.338	2.487	0.017
unemploymenthigh	8.235	4.933	1.669	0.103
gini_index:unemploymenthigh	-18.526	10.804	-1.715	0.094

Model Diagnostic 2

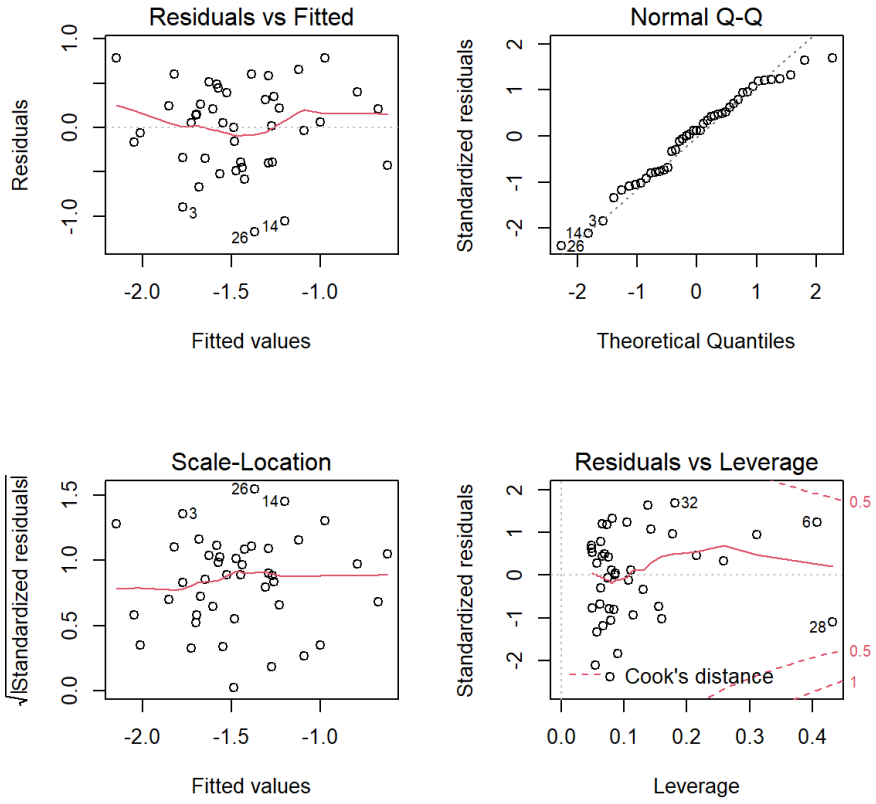

Fig. 13

Model Comparison for RMSE –Model Modification2

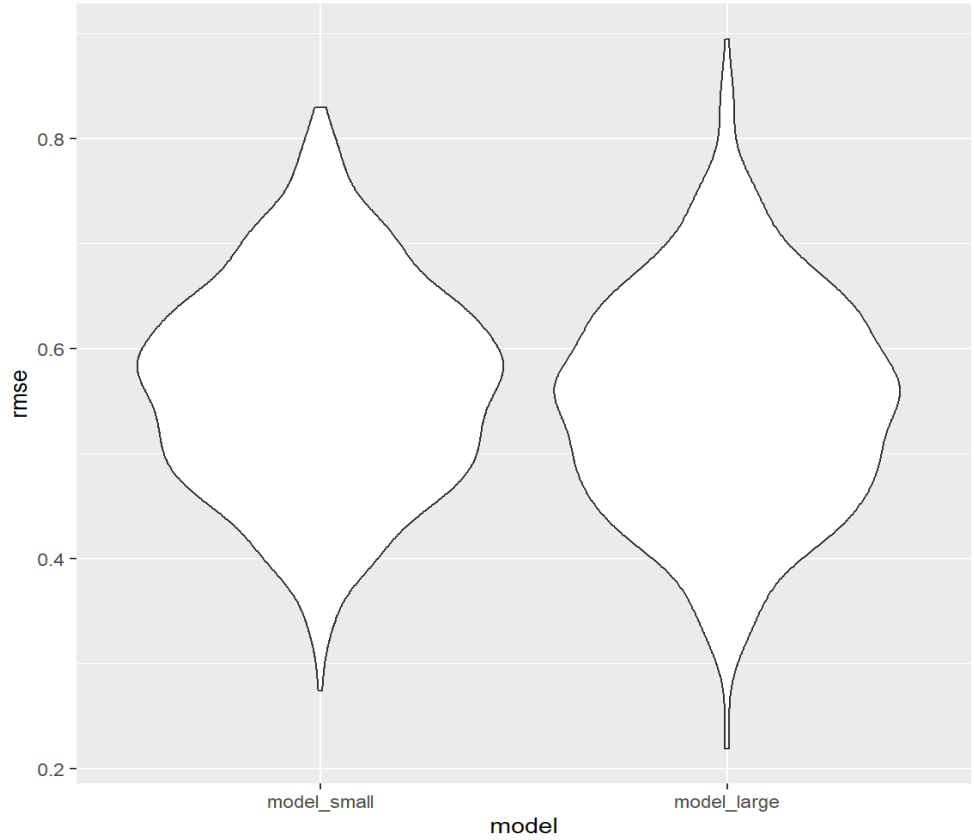

Fig. 14

Table 5: Regression Results for the Modified Model*

Term	Estimate	Std.error	Statistic	p.value
(Intercept)	-18.908	5.312	-3.559	0.001
perc_population_with_high_school_degree	9.900	3.361	2.945	0.005
gini_index	19.839	7.003	2.833	0.007
unemploymenthigh	15.921	5.763	2.762	0.009
gini_index:unemploymenthigh	-34.996	12.532	-2.793	0.008

* *Further omitted observation 2*