REPORT

Group Members: Hanfei Qi, Anmol Singh, Wuraola Olawole, Ting Lian

Abstract

In recent years, America has seen an increase in the occurrences of hate crimes. These crimes are often based on biases such as religious, racial, sexual orientations etc. and have become a source of concern. In the initial analysis published by FiveThirtyEight, it was shown that income inequality was the main predictor of hate crimes. We worked with a dataset containing details on the incidence of hate crimes in all states to explore and identify other variables that were associated with hate crimes and predict its outcome. We conducted a thorough analysis of all the predictors including performing multicollinearity analysis, stepwise regression procedures, and checking model diagnostics to come up with the best possible model to explain the incidence of these crimes. Our findings show that income inequality is not the only predictor that is pertinent in determining hate crime incidence and that other factors play a role as well.

Introduction

Hate crimes in the United States have become a severe problem and their occurrence has been rising in recent years [1]. According to the FBI, a hate crime is a "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."[2] These types of crimes can have lasting impact and cause devastating effects to people due to the horrific nature of the crimes, which is why they are the highest priority of the FBI's civil rights program [2].

A previous study used data from FBI and a self-reported survey to analyze the association between hate crime rates (outcome) and different variables (potential predictors) [3]. The author concluded that income inequality was the most significant predictor of hate crime [3]. In this project, our goal is to use the author's data to build our own model, check if the author's conclusion is correct, as well as assess if any other factors may play a role in hate crime occurrence. Potential predictors include level of state unemployment (low/high), level of state urbanization (high/low), median household income per state, percentage of adults (>25 yrs.) with a high school degree, percentage of population that are not US citizens, percentage of population that are non-white, Gini index that measuring income inequality (range

0-100). Finding out what factors play a role in the occurrence of hate crimes may be able to help us curb the incidence of these horrible crimes in the United States.

Methods

Data Exploration

There are 8 variables in the dataset. Numerical variables are: hate_crimes_per_100k_splc, median_household_income, perc_population_with_high_school_degree, perc_non_citizen, gini_index, and perc_non_white, while categorical variables include: unemployment, and urbanization. Both categorical variables contain two levels: low and high. All coding processes were done by using RStudio.

First, we generated a descriptive statistics table (Table 1). This included the mean, standard deviation (SD), median, 25% quantile (Q1), 75% quantile (Q3), minimum value, maximum value, and count of missing values for each numerical variable. For categorical variables, we obtained counts of each level, and a count of missing values. Secondly, we generated a density plot of outcome to show its distribution by using the ggplot function (Figure 1). We, furthermore, used the boxcox function to find the optimal transformation of the outcome and then double-checked the distribution of transformed outcome (Figure 2,3). Finally, we generated a scatter plot of hate_crimes_per_100k_splc versus state, from low hate crime rate to high crime rate to observe any potential outliers (Figure 4).

Testing Original Association

We first wanted to check if the original association presented in the prior study on hate crimes [3] holds true. We thus performed a linear regression analysis with hate_crimes_per_100k_splc as the response variable and gini_index as the predictor. For the original data we found that the association was significant at a threshold of 0.05. However, for the log transformed data we found that the significance decreased as the relationship was not significant at a threshold of 0.05. We thought this could be a problem with model diagnostics, so we checked them using graphical displays (data not shown). Doing this we found definite outliers that are affecting the association from the leverage plots (data not shown). We then identified outliers and classified them as any value not within ± 1.5*IQR (Interquartile Range). Doing this we confirmed that the District of Columbia and Oregon were outliers. We then removed these outliers and tested the original association again. The model assumptions were met when the outliers were deleted and thus, we proceeded to check for multicollinearity in the model.

Multicollinearity

To check for multicollinearity in the model we started by creating a correlation matrix of all the variables that could be used in the model and then isolating the pairs of variables that had a correlation above 0.6. Furthermore, we conducted some research to figure out why these variables are so highly correlated and found sources that support the relationships we saw. Now that we know that there are highly correlated variable pairs in the model, we need to perform a stepwise regression procedure to eliminate variables that do not contribute to our model.

Stepwise Regression Procedure

For this procedure we started out with a model with all possible predictor variables and used the R function step in the backward direction to eliminate non-essential predictors one by one based on their AIC value.

Interactions

After obtaining our final model we thought to check for interactions between the predictors in our model to see how that affects the associations between the predictors and the outcome. We performed a 2-way interaction among all variables, 3-way interaction (between our predictors) of interest and an ANOVA.

Model Diagnostics of Final Model

For model diagnosis, we plotted four diagnosis plots of our final model to check the assumptions. We also checked the existence of outliers using the studentized residuals. Since there is no candidate for outlier, we did not further explore the existence of the influential point. Finally, the existence of multicollinearity was checked by the variance inflation factor (VIF).

Results

Data Exploration

Table 1 showed that there were 4 NA's in the variable hate_crimes_per_100k_splc and 3 NA's in the variable perc_non_citizen. The low level of unemployment and urbanization were similar, about 50% across 51 states. The distribution of the outcome variable was highly skewed to right (Figure 1). The box-cox transformation indicated that a natural logarithm transformation should be applied to the outcome

variable (Figure 2). The distribution of transformed outcome variable was approximately normal (Figure 3). The scatter plot indicated that data from District of Columbia and Oregon could be potential outliers (Figure 4).

Checking Original Association

When checking the association between income inequality and hate crime rate using the original data, we found that there is a significant linear relationship exists with p-value equals 0.024 (Table 2). However, in the data exploration part, we found that the original data is heavily left skewed. We also checked the association using the log transformed data. The p-value for the association using transformed data was 0.311 (Table 3) indicating that the linear association between income inequality and hate crime rate is not significant when income inequality is the sole predictor.

Multicollinearity

There were three pairs of variables that were highly correlated (correlation > 0.6) in this analysis (Table 4). We found that studies have shown that high income is correlated with increased education and thus it would make sense that the median income and percentage of high school diploma holders are highly correlated [4]. Furthermore, a study conducted by the pew research center found that only 17.7% of immigrants are white non-hispanic which makes sense why the percentage of non citizens and the percentage of white people are very highly correlated as well [5]. To address the problem of the high correlation between these variables as well as eliminate any other insignificant predictors we ran the model through a stepwise regression procedure which narrowed the final model down to unemployment, perc_population_with_high_school_degree, and gini_index as predictors.

Interactions

We explored all 2-way interactions among all the variables. Here, we observed that there were three different interactions (Figure 5). Next, we did a stratified analysis to further explore these interactions. From these we found that two of the interactions were not significant. The only significant interaction upon stratification was between urbanization and perc_population_with_high_school_degree: high urbanization affects the percentage of the population with high school degree. We further carried out an

ANOVA test and obtained that there was no statistical evidence of interaction . We also checked for 3-way interactions between our predictors of interest but found no significant interaction.

Model Diagnostics of Final Model

We plotted the four diagnosis graphs after we got the final model (Figure 6). Points in the Residuals vs. Fitted plot show a random pattern and are evenly spread above and below the line of 0. The red line is approximately horizontal and is bouncing around the line of 0. This graph shows that this model fit the assumption of homoscedasiticity. Similar pattern also shows in the Scale-Location plot indicating that the variance of the model is equal. No point in the Residual vs. Leverage plot is beyond the boundary of Cook's distance so we could assume that there is no significant influence point. In the Normal Q-Q Plot, all points align in an approximately straight line with no significant departure which indicates that the residuals are normal. Overall, these four graphs show that this fitted model is good for represent the data. There is no observation has absolute studentized residual value greater than 2.5. We conclude that there is no outlier in Y in the data and assume that there is no influential point. The VIF values for the three predictors are 1.335 for unemploymentlow, 1.806 for gini_index and, 1.971 for perc_population_with_high_school_degree (Table 5). None of the predictor has variance inflation factor (VIF) value greater than 5 indicating that there is no multicollinearity in the final model.

Conclusion/Discussion

The final model included unemployment, perc_population_with_high_school_degree, and gini_index as predictors, without any interactions. We concluded that gini_index was not the only main predictor. Adjusting for other predictors made the association between gini_index and hate crimes more significant than when gini_index was the only predictor based on the comparison of p-values between the two models. Furthermore, our final model had the highest adjusted R-squared value when comparing to all the other models we tried thus indicating that it is the best fit model. Thus, from our analysis we conclude that it is important to account for unemployment, the percentage of population with a high school degree, and the income inequality of an area when assessing the incidence of hate crimes.

References

1. BBC NEWS (2020, November 17). US hate crime highest in more than a decade - FBI. Retrieved December 17, 2020, from https://www.bbc.com/news/world-us-canada-54968498

2. FBI (2016, May 03). Hate Crimes. Retrieved December 17, 2020, from https://www.fbi.gov/investigate/civil-rights/hate-crimes

3. Majumder, M. (2017, January 23). Higher Rates Of Hate Crimes Are Tied To Income Inequality. Retrieved December 17, 2020, from https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/

4. School, W. (2016, June 27). Education and Income Growth. Retrieved December 17, 2020, from https://budgetmodel.wharton.upenn.edu/issues/2016/2/22/education-and-income-growth

5. Budiman, A. et. al. (2020, August 20). Facts on U.S. immigrants, 2018. Retrieved December 17, 2020, from https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants-current-data/

Appendix:

Descriptive Statistics

Table 1: Descriptive Statistics

	Overall (N=51)
hate_crimes_per_100k_splc	
-Mean (SD)	0.304 (0.253)
-Median (Q1, Q3)	0.226 (0.143, 0.357)
- Min - Max	0.067 - 1.522
- Missing	4
unemployment	
- high	24 (47.1%)
- low	27 (52.9%)
- Missing	0
urbanization	
- high	24 (47.1%)
- low	27 (52.9%)
- Missing	0
median_household_income	
- Mean (SD)	55223.608 (9208.478)
- Median (Q1, Q3)	54916.000 (48657.000, 60719.000)
- Min - Max	35521.000 - 76165.000
- Missing	0
perc_population_with_high_school_degree	
- Mean (SD)	0.869 (0.034)
- Median (Q1, Q3)	0.874 (0.841, 0.898)
- Min - Max	0.799 - 0.918
- Missing	0
perc_non_citizen	
- Mean (SD)	0.055 (0.031)
- Median (Q1, Q3)	0.045 (0.030, 0.080)
- Min - Max	0.010 - 0.130
- Missing	3
gini_index	

- Mean (SD)	0.454 (0.021)
- Median (Q1, Q3)	0.454 (0.440, 0.467)
- Min - Max	0.419 - 0.532
- Missing	0
perc_non_white	
- Mean (SD)	0.316 (0.165)
- Median (Q1, Q3)	0.280 (0.195, 0.420)
- Min - Max	0.060 - 0.810
- Missing	0

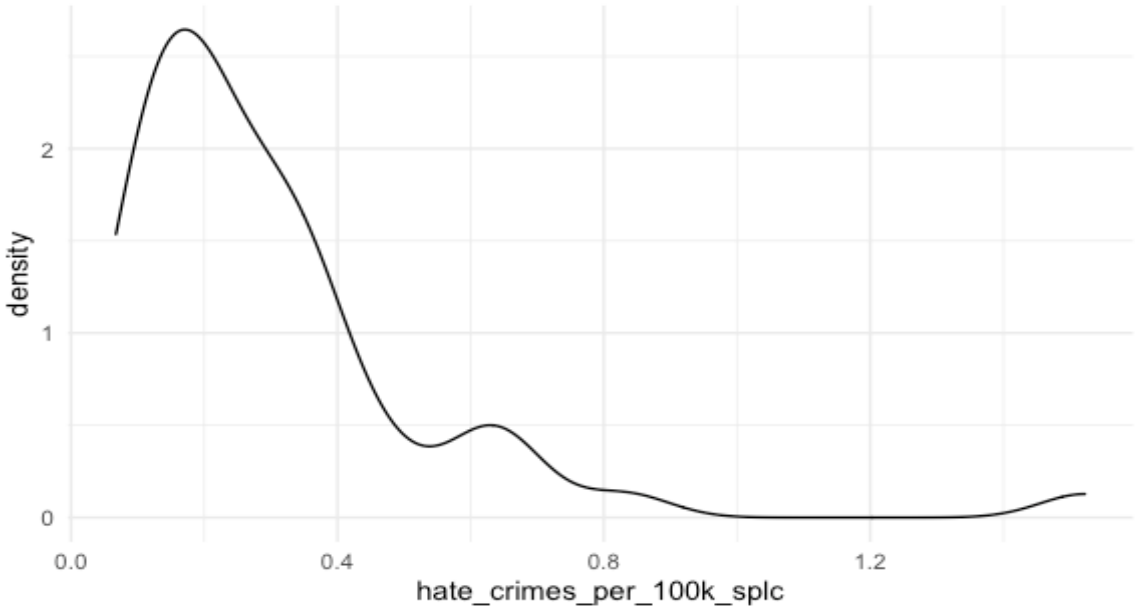


Figure 1: Distribution of hate_crimes_per_100k_splc.

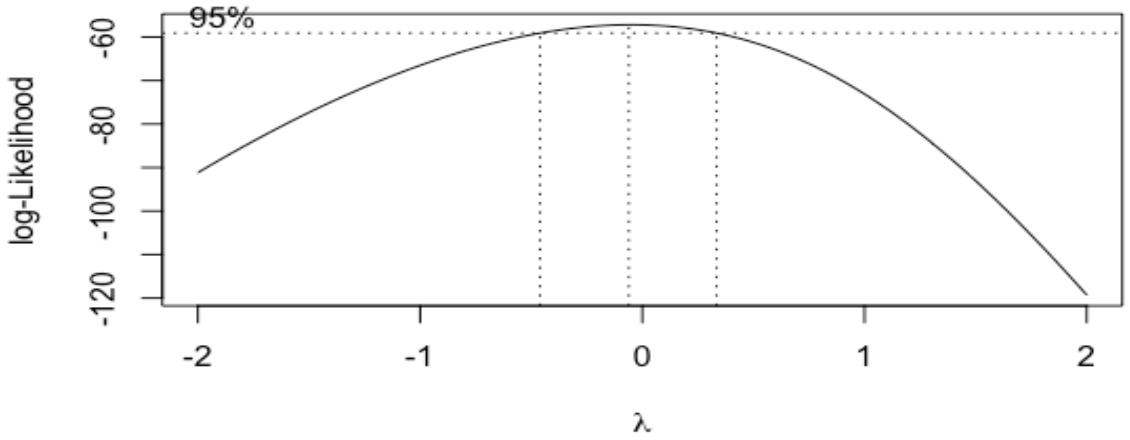


Figure 2: Box-Cox transformation.

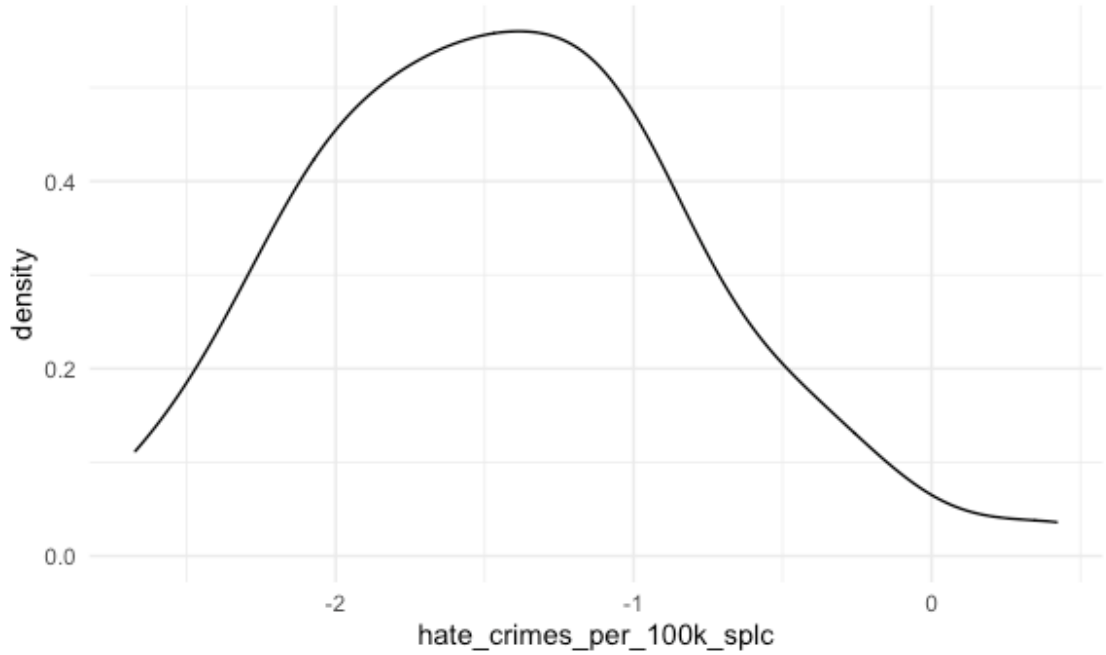


Figure 3: Distribution of hate_crimes_per_100k_splc after applied logarithm transformation.

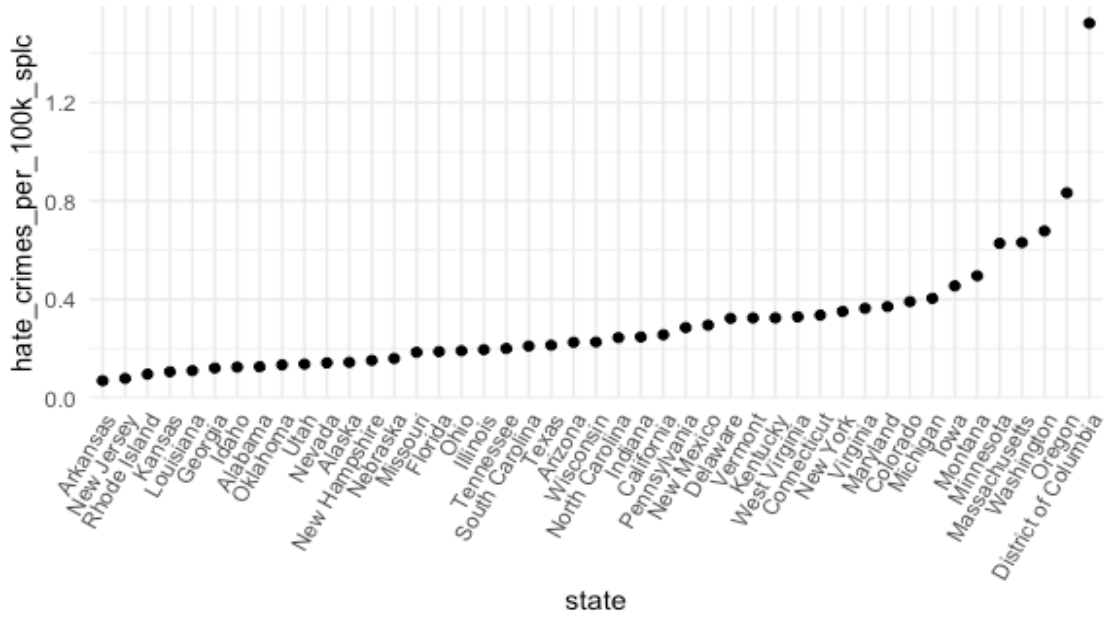


Figure 4: Scatter plot of hate_crimes_per_100k_splc of each state, from lowest hate crime rate to highest crime rate.

Modeling

Table 2: Testing Association Between Income Inequality & Hate Crime Rate Using Original Data

Term	Estimate	std.error	Statistic	p.value
(Intercept)	-1.527463	0.7833043	-1.950025	0.0574197
gini_index	4.020510	1.7177215	2.340606	0.0237445

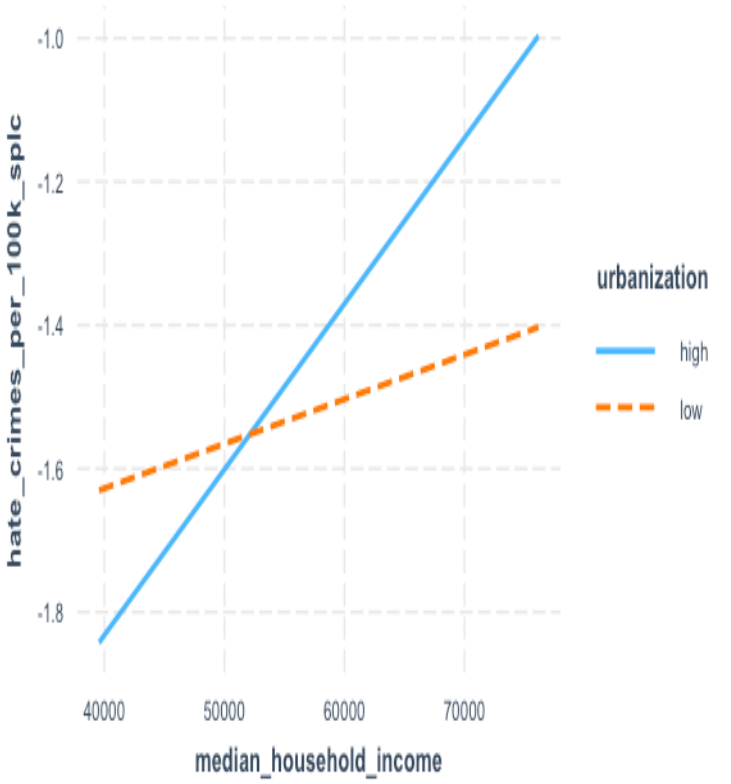
Table 3: Testing Association Between Income Inequality & Hate Crime Rate Using Transformed Data.

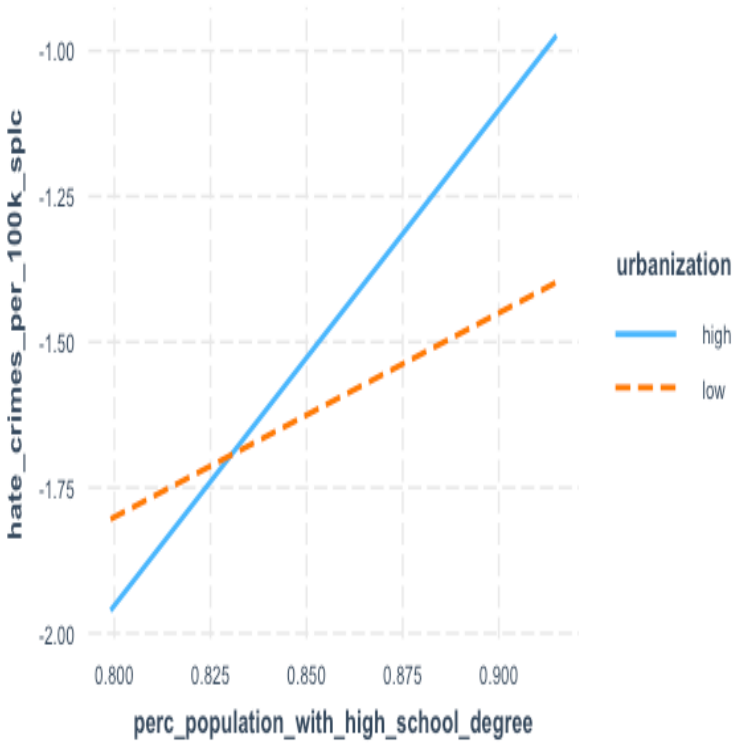
Term	Estimate	std.error	Statistic	p.value
(Intercept)	-3.675547	2.195289	-1.674288	0.1010115
gini_index	4.931538	4.814087	1.024398	0.3111231

Table 4: Variables that are Highly Correlated from Correlation Matrix (Correlation ≥ |0.6|)

Correlation	Variable_1	Variable_2
0.6807743	urbanization	perc_non_citizen
0.6511383	median_household_income	perc_population_with_high_school_degree
0.7526102	perc_non_citizen	perc_non_white

Interactions





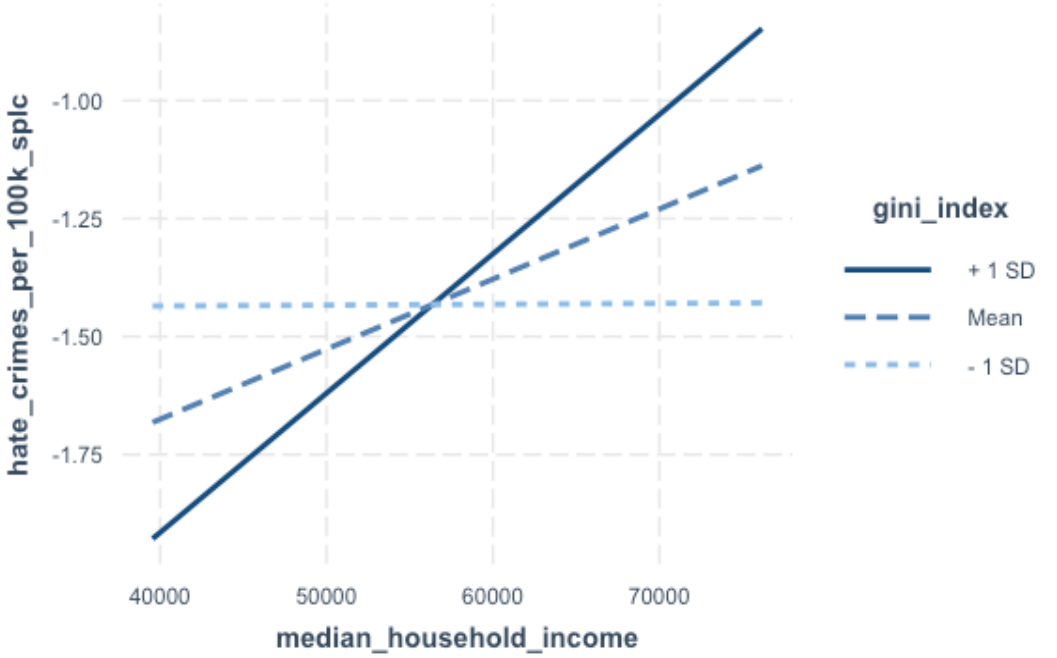


Figure 5: Two-way Interactions Among All Predictors

Model Diagnostics for Final Model

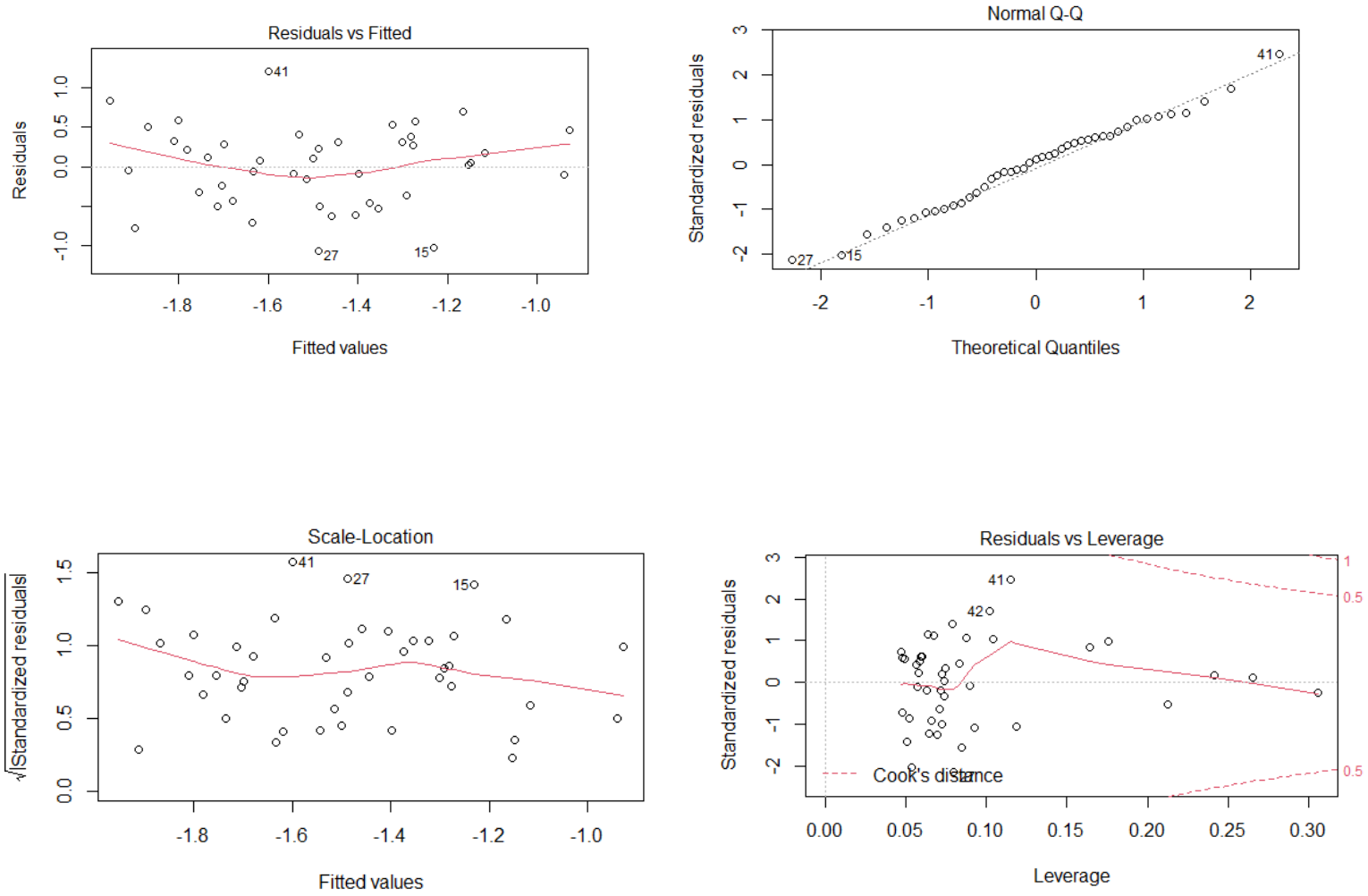


Figure 6: Four Diagnosis Graphs for the Final Model

Table 5: The variance inflation factor Values for the Final Model

	x
unemploymentlow	1.335118
perc_population_with_high_school_degree	1.970627
gini_index	1.806199