final project

Athena Liu

11/22/2021

Summary

This analysis focus on exploring the geolocations and time related patterns of hate crimes in the United States. A hierarchical linear regresssion model with random intercepts was used to answer the questions. Fixed effects include hate crime types and Obama's presidency. Random effects include states and regions of the United States. The model indicated that there is no significant relationship between Obama's presidency and hate crime victim counts. However, race is a significant indicator in the severity of hate crime. The analysis also confirms the thoughts that some states has more occurrence of hate crimes leading to more victims. And, the analysis leads the surprising finding that the Southern regions of the United States has the smallest victim count ratio among the United States.

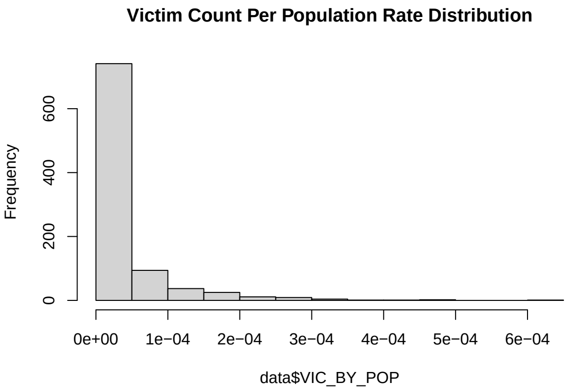
Introduction

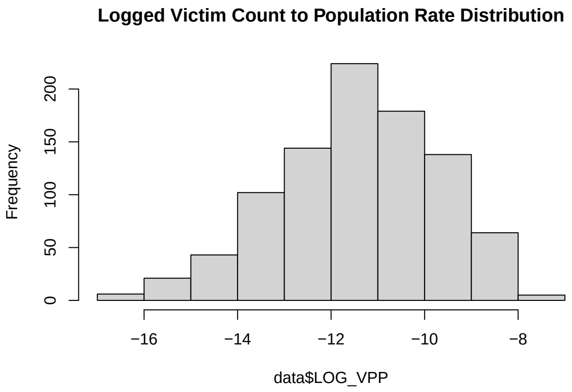
This analysis is inspired by the recent uprising of hate crimes targeting Black, African Americans, and Asians in the United States. It sparks my wonder that if there is correlation between hate crime and geographical locations and time periods. Therefore, in this analysis, the focus is the explore if hate crime severity, measured by victim count by state population ratio, would differ between the targeted groups (Black, Asians, etc.). Another topic of interest if some states have less occurrence of hate crime, or if states will have more occurrence of hate crime. Since the Southern regions of the states are constantly depicted as racist, are there more hate crimes in the South? Lastly, does the presidency of Obama correlates to hate crimes in the United States?

Data

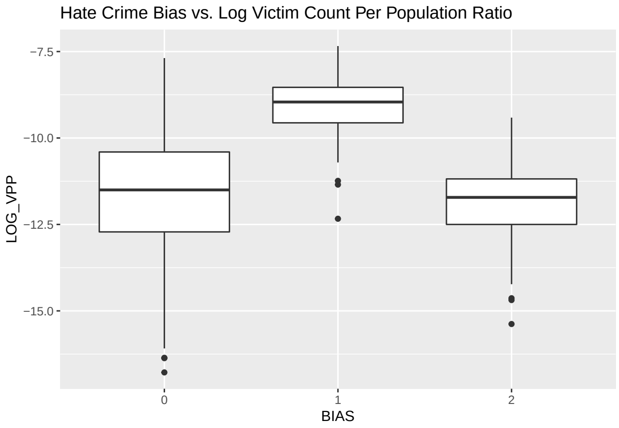
This hate crime analysis uses a combined and pre-modified dataset from two other sets. The main dataset consists of Hate Crime Data provided by the Federal Bureau of Investigation (FBI) public crime explorer database. It contains 219,073 observations ranging from hate crime records in 1990 to records in 2020. The secondary dataset is the United States Population from 2010-2020. The two datasets were combined by states name so that the final dataset could contain both crime data and population information. The final dataset contains the following fields: state name (STATE_NAME), region name (REGION_NAME), hate crime description (BIAS_DESC), numericalized hate crime description (BIAS), Obama's Presidency (POST_Obama), and population by state (POPULATION). The original dataset was first grouped by state, Obama presidency, and bias type, and the total victim count was calculated. Then, the response variable VIC_VPP is calculated by dividing the population of the corresponding state. The variable BIAS was also converted to numerical values the following condition: the victim is black (value = 1), the victim is Asian (value = 2), and others (value = 0). And since this analysis is limited to hate crimes related to race, hate crimes motivated by sexual orientation, gender, and religion were removed from the dataset. The generation of the field POST_OBAMA could be confusing to the reader. If the year of the incident is before 2008, before Obama was elected president, value will be 0. If the incident happened after 2008, during and after his presidency, the value will be 1. Thankfully, there is not missing values or erroneous values, and this concludes the data cleaning process.

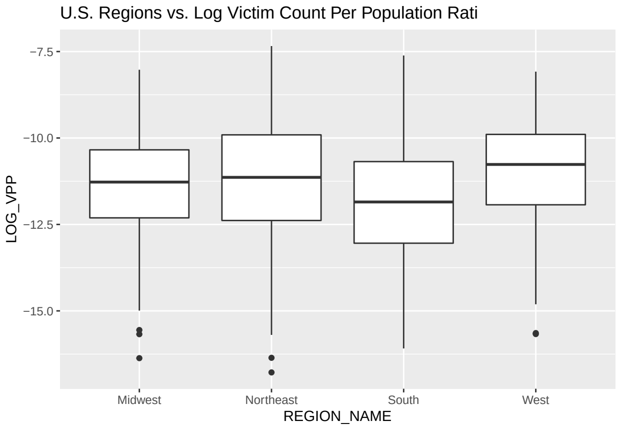
The plots below displays the distribution of the response variable victim count per population. The original distribution of the variable is obviously right skewed. Later in the modeling process, it is proven that this distribution would violate the normality assumption of linear regression. Thus, the response variable is log transformed.



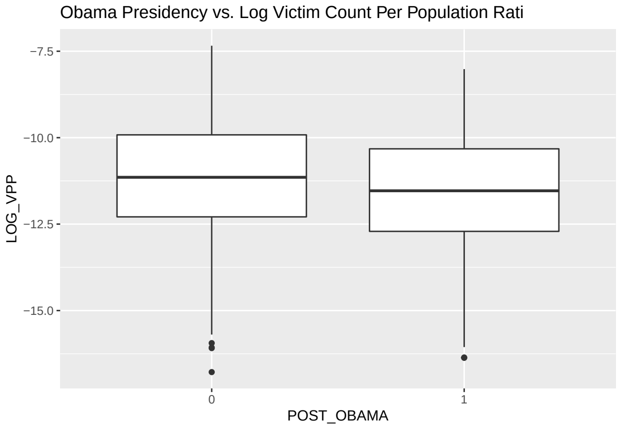


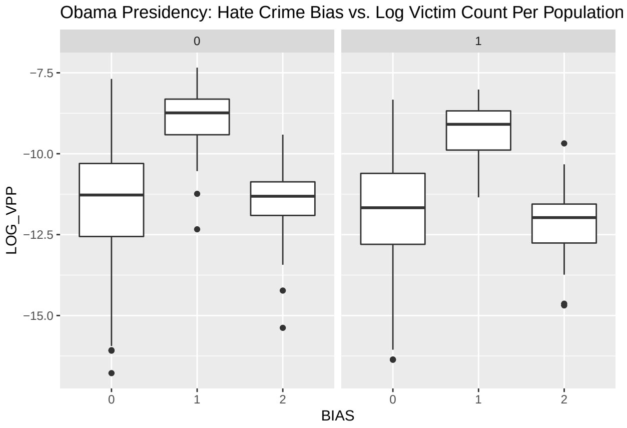
The plots below displays the exploratory data analysis for relationships that we would like to examine in the model. The victim count per population for anti-Black and African American motivated hate crime is higher than Anti-Asian or other types of hate crime. The model should indicate if this is significant. The other plot shows how victim count per population distribution for different regions. Thus far, there's no observable differences.





The images below displays the relationship between Obama's presidency and victim count by population. Again, the left plot indicated no obvious difference between the Pre-Obama and Post Obama period. Similar patterns are displayed in the second plot, where the distribition of victim count per population divided by hate crime types has no obvious difference between the Pre-Obama and Post-Obama Period.





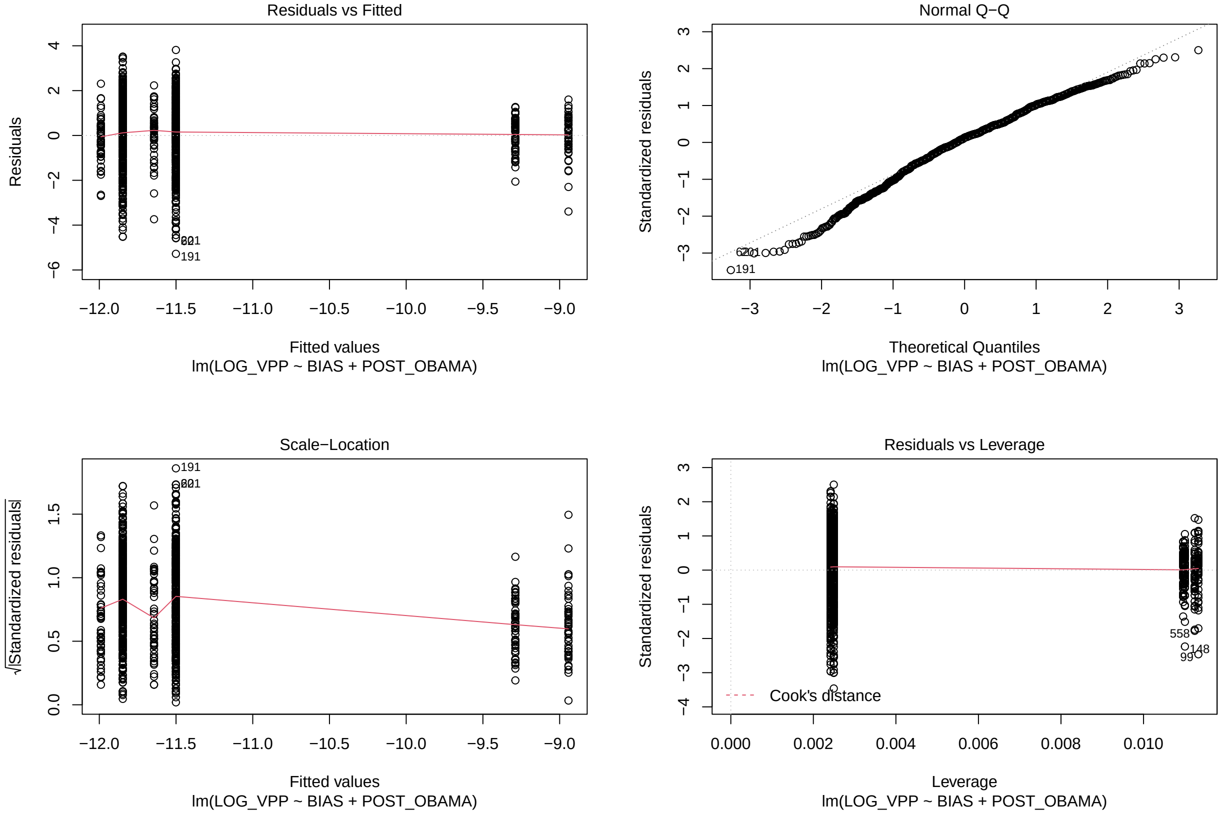
Model

In this analysis, a hierarchical linear regression model is used to analyze questions relating to hate crimes in the United States. The response variable is the logged victim counts per population field (LOG_VPP). The potential predictor variables include Obama Presidency (POST_OBAMA), States (STATE_NAME), Regions (REGION_NAME), and bias types (BIAS). Since all the predictor variables are categorical, they are factorized in the data cleaning process. Unlike a regular linear regression model, a hierarchical linear regression model contains both fixed and random effects. Before constructing a full hierarchical model, it is crucial to determine the appropriate fixed effect variables in which the current candidates are POST_OBAMA, BIAS, and POST_OBAMA:BIAS interactions. The AIC step-wise model selection process on a regular linear regression model determines the selection of the fixed effects variables. The null model is the response variable with a constant as the predictor; meanwhile, the full model contains all the fixed effects candidates. The table displays the final linear regression model for the fixed effects variables, suggesting the removal of the POST_OBAMA:BIAS interaction.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.5008	0.0763	-150.83	0.0000
BIAS1	2.5584	0.1621	15.78	0.0000
BIAS2	-0.1425	0.1643	-0.87	0.3858
POST_OBAMA1	-0.3462	0.1004	-3.45	0.0006

Table 1: Linear Regression Model on Fixed Effect Variables

The following plots assess the validity of the linear regression model. Points fall almost precisely on the 45 degrees line in the Q-Q plot. This assures us that the logged victim count per population ratio satisfies regression's normality assumptions. Meanwhile, the other three plots might be more disturbing but still acceptable. The patterns displayed in the Residuals vs. Fitted, Scale-Location, and Residuals vs. Leverage plots are inevitably caused by the fact that there are only categorical predictor variables in the model. Thus, it is appropriate to construct the hierarchical linear regression model.



The fixed effect variables excluded REGION_NAME and STATE_NAME because they are naturally hierarchical in this model as the observations are nested in the states and region. The ideal hierarchical model would contain both the regions and states as the random effects because both are needed to answer the questions of interest. However, three hierarchical models are fitted to compare the model performance between random effects of states only, regions only, and states and regions combined. The hierarchical model with states only has an AIC value of 3333.194. The hierarchical model with regions only has an AIC value of 3390.587. Finally, the hierarchical model with states and regions has an AIC value of 3329.466. While the model using regions only has the best AIC values among the three models, all the AIC values are similar to each other. This is an assuring sign that including both states and regions are random effects would not greatly reduce the performance of the hierarchical model. In addition, the ANOVA test between the three models suggests the model with both variables has the best performance.

The final model is a random-intercept hierarchical linear regression model containing state and region as random effects, hate crime bias, and Obama presidency as fixed effects. Among all fixed effects, only BIAS2 is insignificant. POST_OBAMA1 represents the period after Obama was elected president. The negative intercept indicated that the victim count per population ratio decreased, which suggests a correlation between his presidency and the decrease in hate crime. On the other hand, there is a significant positive intercept for BIAS1, which it represents that the victim is Black or African American. This suggests a positive correlation between the victim being black and the increase of victim count per population, suggesting more hate crimes in the United States targeting Black or African Americans.

The random effects of the model provide some insight into how geographical locations could correlate with hating crimes in the United States. The random intercepts presented in the plot in the appendix show the victim count per population ratio when other variables are constant. It allows us to see which states have higher hate crime victims ratio.

The second iamge in the appendix showed the random intercept for the regions within the United States. The only region with a negative interept is the South, indicating there is an association of South and relatively less hate crime incident. This is the opposite of the common sterotype of the Southerns states having more hatreds towards minority groups.

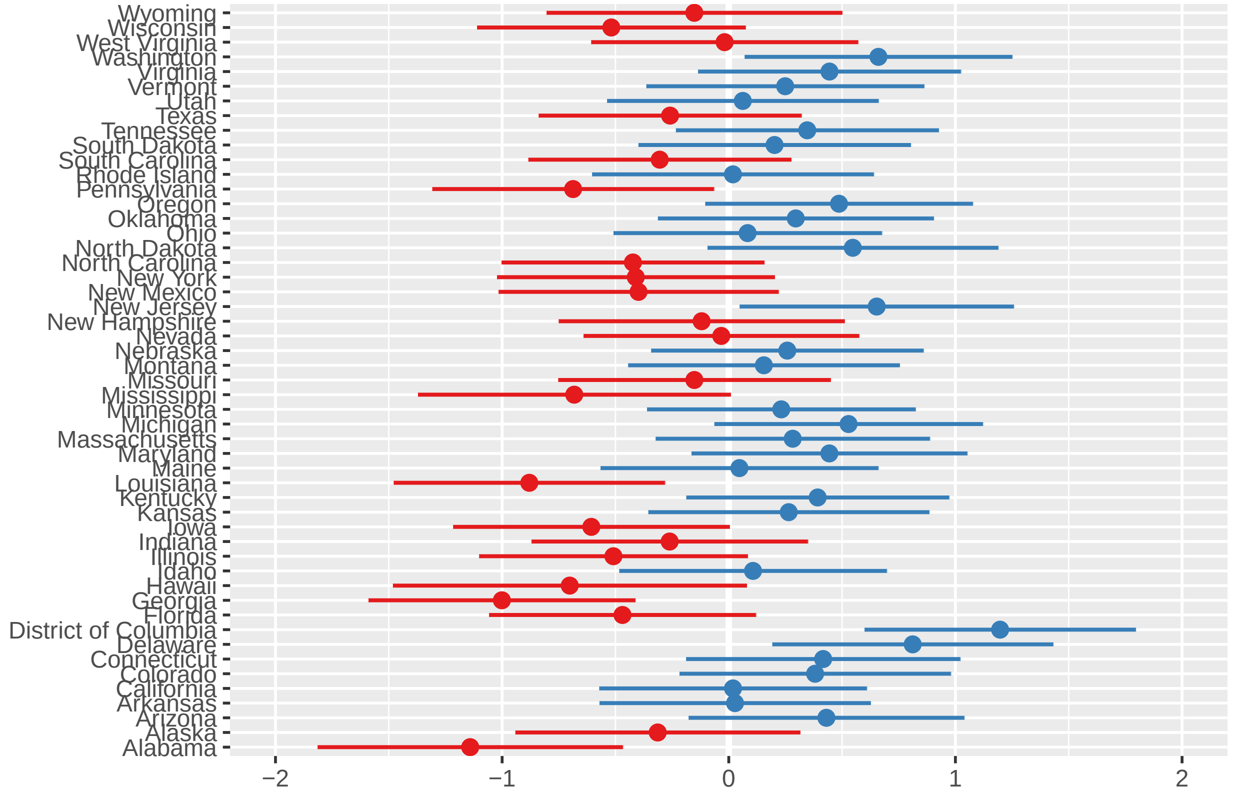
```
## Linear mixed model fit by REML. t-tests use Satterthwaite’s method [
## lmerModLmerTest]
## Formula: LOG_VPP ~ POST_OBAMA + BIAS + (1 | STATE_NAME) + (1 | REGION_NAME)
##    Data: data
##
## REML criterion at convergence: 3315.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6186 -0.5973  0.0979  0.6434  2.5420
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  STATE_NAME  (Intercept) 0.3325   0.5767
##  REGION_NAME (Intercept) 0.1157   0.3402
##  Residual                1.9226   1.3866
## Number of obs: 926, groups:  STATE_NAME, 51; REGION_NAME, 4
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) -11.50795    0.20170   3.71457 -57.054 1.32e-06 ***
## POST_OBAMA1  -0.29916    0.09161 878.01979  -3.266  0.00113 **
## BIAS1         2.58375    0.14736 870.73862  17.534  < 2e-16 ***
## BIAS2        -0.12068    0.14930 870.49979  -0.808  0.41914
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##             (Intr) POST_O BIAS1
## POST_OBAMA1 -0.233
## BIAS1       -0.091  0.005
## BIAS2       -0.088  0.001  0.121
```

Conclusion

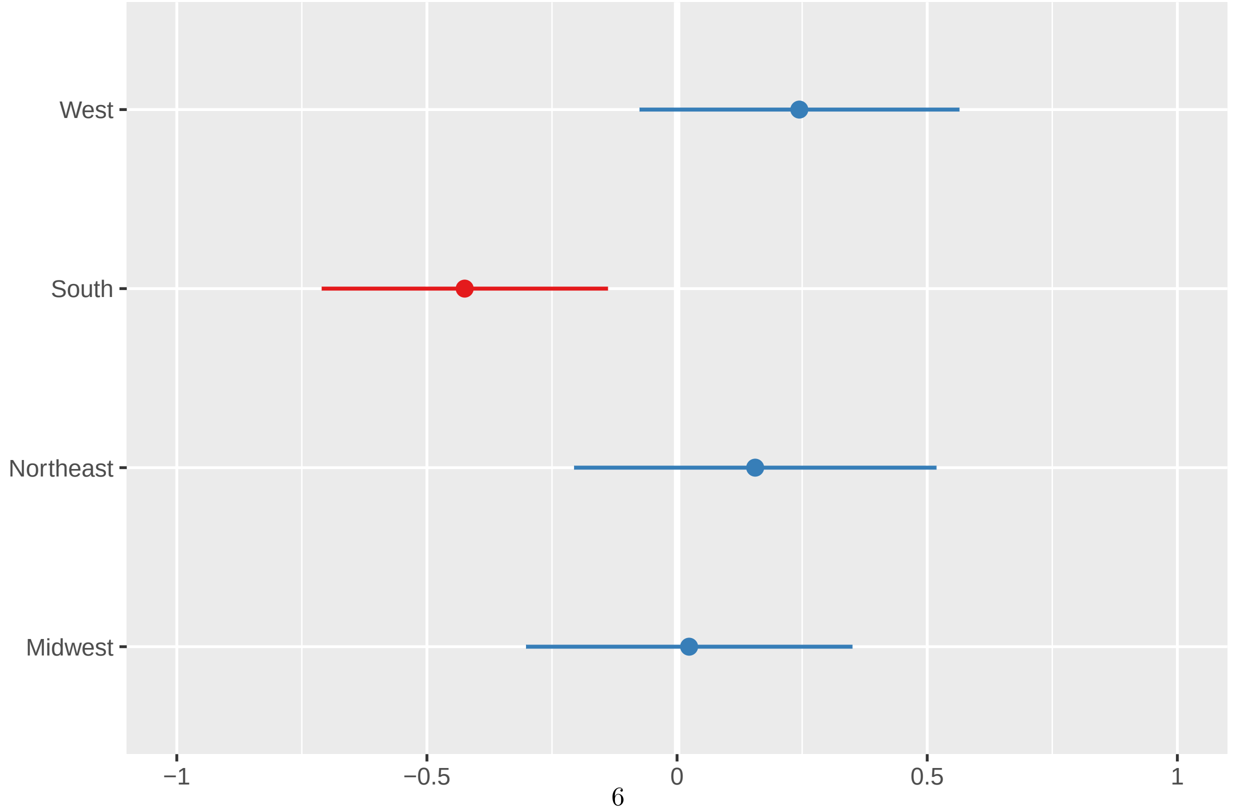
The hierarchial model provided great insight into the questions of hate crime in the United States. However, it is worth mentioning the limiation of this analysis. The FBI discourages using the dataset to evaluate the justice system’s effiency by state or region. Thus, all the results found in this paper should not be public distributed.

Appendix

Random effects



Random effects



Github: https://github.com/Athena112233/hatecrime_analysis