Report

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

## Abstract

## Introduction

Hate crimes in the United States has become a severe problem and their occurrence has been rising in recent years [[1]](https://www.bbc.com/news/world-us-canada-54968498). 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]](https://www.fbi.gov/investigate/civil-rights/hate-crimes) 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]](https://www.fbi.gov/investigate/civil-rights/hate-crimes).

A previous study used data from FBI and a self-reported survey to analysis the association between hate crime rates (outcome) and different variables (potential predictors) [[3]](https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/). The author concluded that the income inequality was the most significant predictor of hate crime [[3]](https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/). In this project, our goal is to use the author’s data to build our own model and 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 process was done by using RStudio.

First, we generated a descriptive statistics table to observe the data. This included obtaining 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. We, furthermore, used the boxcox function to find the optimal transformation of the outcome and then double-checked the distribution of transformed outcome.

Finally, we generated a scatter plot of hate\_crimes\_per\_100k\_splc versus state, from low hate crime rate to high crime rate, so we could observe any potential outliers.

### Modeling

#### 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. Doing this we found definite outliers that are affecting the association from the leverage plots. We confirmed these outliers by using the standard formula of a lower outlier being less than Q1 - 1.5(IQR) and an upper outlier being greater than Q3 - 1.5(IQR) where IQR is the interquartile range. Doing this we confirmed that the District of Columbia and Oregon were outliers. We then removed the outliers and tested the original association again. The model assumptions were met when the outliers were deleted and thus we proceeded to check for multicolliniarity in the model.

#### Checking Multicolliniarty

To check for multicolliniarity 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. This returned three pairs of variables that were highly correlated: . 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. The final model we ended up with used unemployment, perc\_population\_with\_high\_school\_degree, and gini\_index as predictors.

#### Checking Interactions:

After finding 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.

#### Model Diagnostics

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

The table 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. The box-cox transformation indicated that a natural logarithm transformation should be applied to the outcome variable. The distribution of transformed outcome variable was approximately normal. The scatter plot indicated that data from District of Columbia and Oregon could be potential outliers.

### Modeling

#### Checking Original Association

#### Checking Multicollinearity

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]](https://budgetmodel.wharton.upenn.edu/issues/2016/2/22/education-and-income-growth). 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]](https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants-current-data/).

#### Stepwise Regression Procedure

#### Checking Interactions

#### Model Diagnostics

## Conclusion/Discussion

## References

# use this code to do work count!  
# wordcountaddin::text\_stats("Report.Rmd")