Hate Crimes in America

(R-Script included at end of report)

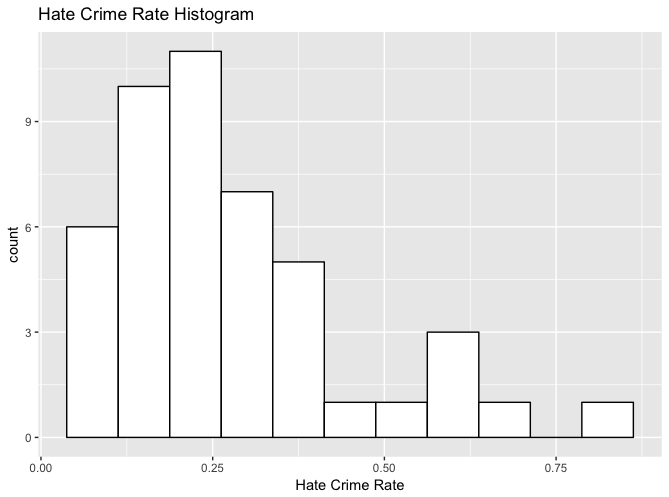
By: Nick Bruno

Hate crimes have been an issue in the United States since its inception, and continue to be an issue today. According to voanews.com, a hate crime is a violent crime motivated by hate based on race, color, or national origin, among other factors. Although it may seem that society is taking steps in the right direction concerning social tolerance, hate crimes in the United States continue to be committed. Many media sources claim that hate crimes have increased significantly in 2016 and 2017, and many of these media outlets blame the 2016 election of Donald Trump as U.S. president as the cause of this increase in hate crimes. One website argued, “Since the election of President Donald Trump, news outlets and social media accounts have swelled with reports of hate-fueled attacks and acts of intimidation” (slate.com). This is true, as evidence of more hate crimes are being uploaded and shared through social media platforms, such as Facebook and Twitter. However, social media was not nearly as large of a platform it was five years ago compared to how large it is today. So, did the number of hate crimes in the United States increase immediately after the 2016 presidential election, or are citizens more aware of it now because of its increased reporting and awareness spread through social media platforms? Using a data set containing hate crime data from all 50 states from fivethirtyeight.com, I hope to examine the effect that Donald Trump’s presidential election might have had on US hate crime rates, and conclude if the data supports the slate.com report.

Before I begin the analysis, I want to point out that the dataset contains data on all 50 U.S. states and Washington, DC. However, the hate crime rate in Washington D.C. was disproportionately high, creating a massive outlier in the data. The high hate crime rate in the District of Columbia could be explained by many different factors, but I think one of the largest characteristics that leads to the high hate crime rate is the density of the city, as well as the small population, skewing the hate crimes. Comparing the data given and the Washington D.C. population, around 10 hate crimes occurred in D.C. from November 9 to November 18, 2016, which compared to other states seems like a small number, but the small D.C. population exaggerates the hate crime rate statistic (United States Census Bureau). For these reasons, Washington D.C. was omitted from the analysis, as comparing data from the 50 states should give a more accurate depiction of the hate crime rates across the United States.

The dataset analyzed in this project contains statistics concerning the U.S. state, hate crimes per 100,000 population from November 9 - 18, 2016 (according to Southern Poverty Law Center), average annual hate crimes per 100,000 from 2010 - 2015 (according to the FBI), 2016 median household income, share of the population that is unemployed in September 2016, share of the population that lives in metropolitan areas in 2015, share of adults 25 or older with a high-school degree, share of the population that are not U.S. citizens in 2015, and the share of 2016 U.S. presidential voters who voted for Donald Trump.

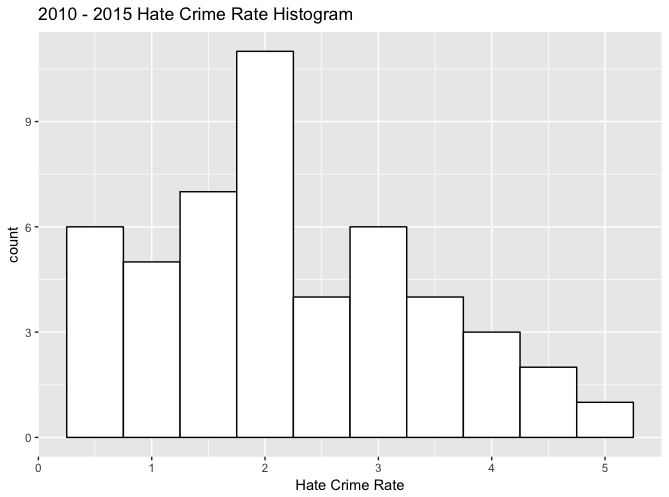
Before running any regressions, I started by comparing the number of hate crimes from November 9-18, 2016 compared to earlier years. The main variable from the data I wish to examine is the share of hate crimes per 100,000 members of the population the ten days following the election of Donald Trump. I will refer to this variable as the post-election hate crime rate throughout the project, and I will take a closer look at the data across each state. First, Figure 1 shows a histogram comparing the post-election hate crime rates.

Figure 1

Here, the hate crime rates seem to be skewed to the right, and it appears that there is one state that is a slight outlier in the data, but does not seem substantial enough to largely affect the data analysis. To confirm the right skew of the data, I compared the mean and median of the hate crime rates. The mean of hate crime rates across the 50 states is 0.2776, but the median is only 0.2258. This confirms that the data is slightly skewed to the right.

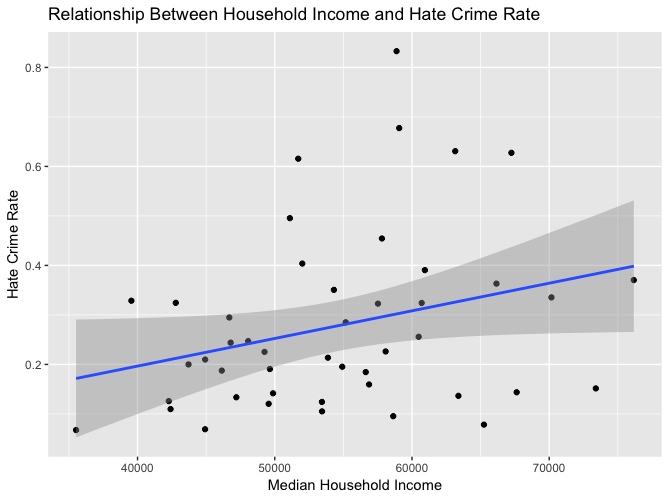
To further study this hate crime rate data, I created a 95% confidence interval based off of the 46 data points given in the dataset concerning the post-election hate crime rate. First, taking the mean across all of the states, 0.2776 of 100,000 American citizens were a victim of a hate crime. To put this into perspective, around three out of one million Americans fell victim to a hate crime during the ten-day stretch after the 2016 election. With the American population around 323.1 million, that works out to around 897 hate crimes (United States Census Bureau). The confidence interval using the mean as the xbar value and a t-statistic with 45 degrees of freedom is (0.224833, 0.330373). This means that I am 95% confident that the population mean of post-election hate crime rates per state in the United States falls between 0.224833 and 0.330373 hate crimes per 100,000 citizens. Now that there is a basic background and a greater understanding of what the post-election hate crime rate data entails, I will compare these results to the average hate crime rate amongst the United States from 2010 – 2015.

I ran the same calculations concerning hate crimes in the US from 2010 - 2015 as I did the post-election hate crimes, and found that the average number of hate crimes per 100,000 citizens was 2.192. After finding the average United States population from 2010 to 2015 and careful calculation, an average of around 6,907 hate crimes occurred per year during that time (United States Census Bureau). The histogram in Figure 2 looks similar to the post-election hate crime rate histogram, but is slightly less skewed.

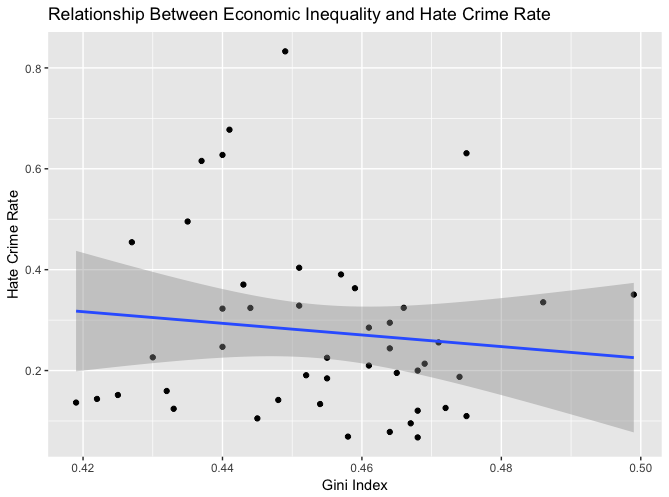
Figure 2

So, putting the absurd number of hate crimes that occurred during the ten day span after the 2016 presidential election into perspective, if the hate crime rate persisted at the rate at which it did in the ten post-election days, then roughly 32,738 hate crimes would have been committed in 2016, which would be five times larger than the average hate crime rate from 2010 to 2015. Here, it is obvious that the hate crime rate substantially rose during the ten-day period after Donald Trump was elected president.

After determining that hate crimes increased after the election of Trump in 2016, I delved deeper into the difference between states with high hate crime rates and those that had low rates during the post-election time frame. Finding what characteristics of states impact hate crimes can allow lawmakers and citizens to learn from the past and understand the necessary steps to decrease post-election hate crime rates in the future. The first relationship I wanted to study was the relationship between median income and post-election hate crime rate. From 2008 to 2012, the Bureau of Justice Statistics found that poorer households had a higher rate of crime, specifically violent crime. Across the 50 US states, the median income was $54,613 in 2016. Taking a quick glance at the two states with the highest post-election hate crime rate, Oregon (0.83284961 per 100,000) and Washington (0.67748765 per 100,000), both had median household incomes greater than the $54,916 median found earlier, as Oregon’s median household income was $58,875 and Washington’s was $59,068. The two states with the lowest rate of post-election hate crimes were Mississippi (0.0674468 per 100,000) and Arkansas (0.06906077 per 100,000). From these numbers, it seems that the states with lower median household incomes led to less hate crimes, but taking a further look at the scatter plot comparing household income and crime rates (Figure 3), there is actually a slight positive correlation between household income and the post-election hate crime rate.

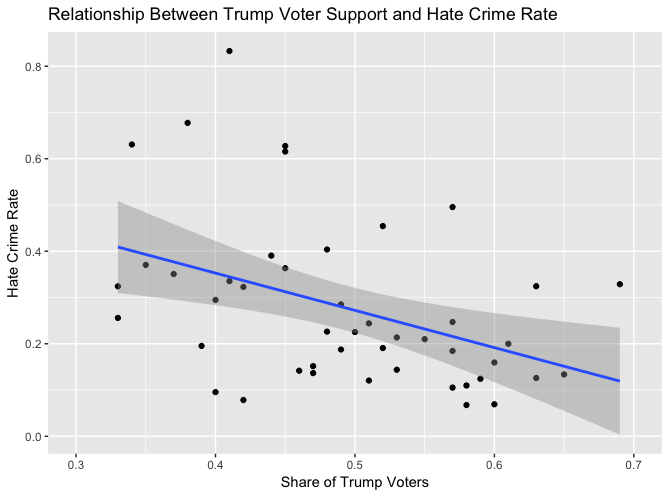
Figure 3

According to ‘The Atlantic,’ income inequality causes increases in crime, specifically when the affluent express their wealth ostentatiously, so I did a similar analysis comparing the 2015 gini index of each state with the state’s hate crime rate. The gini index is a statistical measure used to gauge economic inequality by measuring wealth and income distribution (investopedia.com). A gini coefficient of zero represents perfect economic equality, while a coefficient of one represents perfect inequality (investopedia.com). Plotting the relationship results in Figure 4 below.

Figure 4

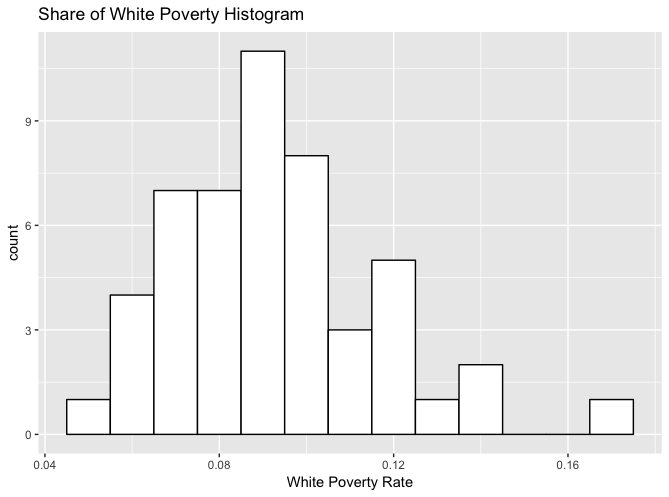
Here, the data points seem to be scattered, and there is a slight negative correlation between gini index and post-election hate crime rate, but there is not a very obvious relationship between these two variables. So, these two variables do not seem to be very correlated, and this will be important to note later in the analysis when I compare regressions of different variables.

Another variable I want to isolate and take a longer look at is the share of voters per state that voted for Donald Trump in 2016 and compare that to the post-election hate crime rate. If the slate.com commentary is accurate, one would expect that post-election hate crimes would occur in states where Donald Trump received the largest share of votes. The best way to visually understand the relationship is through scatterplots, like the ones above, so I will repeat the process once more. Figure 5 below represents this relationship.

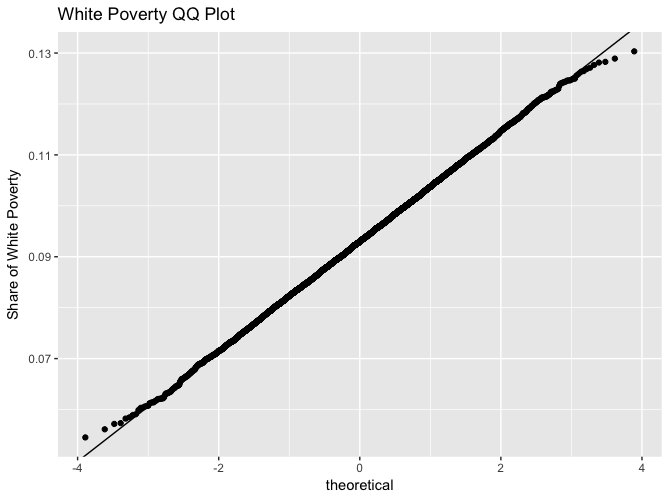
Figure 5

In this scatterplot, there seems to be a relatively strong negative correlation between the share of citizens who voted for Donald Trump in the 2016 presidential election and the number of post-election hate crimes per 100,000 citizens per state. From these results, one could conclude that the states with the greatest voting support for Donald Trump actually had lower post-election hate crime rates compared to states where Donald Trump received a smaller share of the votes. Of course, there are many other underlying variables that could lead to this result, but of the three variables investigated so far; this scatterplot seems to have the strongest relationship to the post-election hate crime rate.

The last variable I wanted to look further into was the share of white residents living in poverty in 2015, per state. After comparing histograms of each variable, this variable seemed to have a sample that most closely resembled a normal distribution, so I decided to use a sampling method and compare the results to the results to the actual mean and standard deviation of the sample from the dataset. The white poverty histogram is printed below.

Figure 6

To find this sample, I ran a Monte Carlo simulation that replicated five samples 10,000 times with the mean and standard deviation parameters that are consistent with the actual data found in the problem set. After running the Monte Carlo sampling simulation, the mean of the samples was very similar to the actual mean of the data, as the real data had a mean of 0.0928, and the mean of the simulated samples was 0.09297471. The actual mean translates to a mean of white poverty per states is around 9.28%. The same is true when comparing standard deviations, as the sample standard deviation for the data was 0.02382247 while the standard deviation of the replicated simulation was 0.02392001. This is not unexpected, as the simulation sampling should provide a sample with a mean and standard deviation as close to the actual data as possible, and the Monte Carlo simulation replicating five samples 10,000 times is basically replicating the data 50,000 times, and should give a precise number near the actual data.

Figure 7

The QQ plot in Figure 7 above verifies that the Monte Carlo simulated data is normally distributed, as it almost perfectly matches the straight reference line, disregarding a couple of outlying simulations towards the start and end of the reference line.

Rather than including a scatterplot or running a Monte Carlo simulation for every variable, I ran multiple regressions instead, taking into account each of the variables in the dataset, and continued this process until a clear and statistically sound regression was found. These regressions will give a better understanding of which variables are statistically significant and are relevant when studying the hate crime rate. However, before running regressions, I ran a correlation test between each variable in the data set with the post-election hate crime rate, and found that the only statistically significant variables at the 5% significance level were the share of Trump voters and the share of adults with a high school degree, although the 2016 median household incomes and the 2015 share of the nonwhite population per state were significant at a 10% significance level. These correlation tests are important and share a lot of information about the significance of each variable, but should not be used as the sole test for validity. So, it is important to run multiple regressions, including regressions with variables that might not seem significant, to see if they might actually be important when trying to determine the post-election hate crime rate of a state.

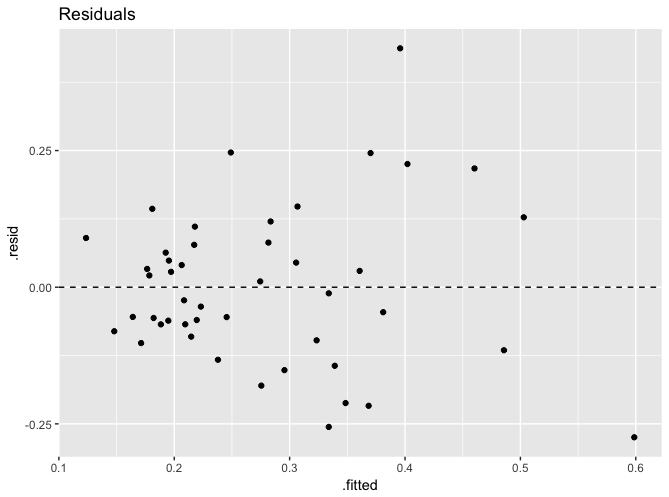
The first regression I ran only took into account the three explanatory variables studied above; the share of Trump voters, median household income and the gini index of each state, all regressed on post-election hate crime rate. The regression looks like this: After running the regression, it appears that the only statistically significant variable is the share of Trump voters in 2016, as this variable had a p-value of 0.0141. Each variable also had a negative coefficient, implying that an increase in any of the variables would lead to a decrease in post-election hate crime rates. Although one expected the Trump voters and gini index to be negatively correlated, it is odd that the median household income variable is also negative, as the scatterplot shows a positive correlation between post-election hate crime rate and median household income. The next regression statistic I evaluated was the R-squared value, which was quite low at 0.2067, meaning that the combination of these three variables does not seem to have a strong relationship with the post-election hate crime rate. This analysis is mostly consistent with the scatterplots found earlier in the report, although I thought that the median household income would have a positive coefficient and also have a greater impact on post-election hate crime rates. Because of this, I tested to see if multicollinearity existed between the median household income variable and the gini index. After running a variance inflation factor test, all of the resulting vifs were below 2.4, meaning that there was not strong evidence for multicollinearity. Although multicollinearity does not seem to be an issue, the general F-statistic for the regression is significant, meaning that the current model is significant, so at least one of the variables should be useful when trying to determine which variables affect hate crime rates. After this conclusion, I decided to move on and run more regressions to try to find a better fitting model.

Next, I ran a regression including all of the other variables in the dataset that I did not get to analyze as deeply as the three variables in the first regression model. The explanatory variables of this model included the share of the population that was unemployed in September 2016, the share of the population that lives in metropolitan areas in 2016, the share of adults 25 and older with a high-school degree in 2009, the share of the population that are not U.S. citizens in 2015, and the share of the population that is nonwhite in 2015. Despite having six explanatory variables, the R-squared was still low at 0.2747, and the adjusted R-squared was significantly lower than that (0.1571). The only variable that was significant at a 5% significance level was the share of adults 25 or older with a high-school degree in 2009. Because there were so many variables in the model, a test for multicollinearity was necessary, as the insignificance of some of the variables could be caused by multicollinearity. However, to my surprise, none of the variables seemed to be multicollinear, as none of the variance inflation factors of the explanatory variables was greater than five, with the nonwhite population variable having the highest vif at 4.024919. Rather than scrapping all of the variables, I decided to omit the highly insignificant variables, but include the variables that were almost statistically significant.

The third model regressed the September 2016 unemployment, share of adults with a high school diploma, and the share of the population that are not U.S. citizens in 2015 on the post-election hate crime rate. The regression gave similar results to the model 3 regression, as the education variable was significant, but the other two variables were statistically insignificant. In this case, again, there is no multicollinearity present. With this result, I concluded that the only two variables that seemed statistically relevant to the post-election hate crime rate were the Trump voter share and the 2009 share of adults with a high school degree.

I decided to first run a simple regression regressing those two variables on the hate crime rate, resulted in this regression equation:. This resulted in an adjusted R-squared value that was low at 0.2367, but relatively high compared to the previous models. Also, both variables are statistically significant, and both keep their sign consistent with the previous models, as the share of Trump voters variables had a negative coefficient (-0.6168) and the share of the population with high school educated adults had a positive coefficient (1.6992). The education variable means that if 100% of the state population had a high school diploma, then the post-election hate crime rate would increase by 1.6992 hate crimes per 100,000 population members, holding the Trump voter variable constant. The Trump voter’s variable implies that if 100% of the state’s population voted for Donald Trump, then 0.6168 less post-election hate crimes per 100,000 citizens would occur, holding the education variable constant. Although this model is an improvement compared to the previous models, the low R-squared is concerning, so I tried to implement quadratic regressions and instrumental variables to try to create a more accurate model.

After running many models that created instrumental variables and comparing quadratic regressions, the most successful model was a regression on the 2016 hate crime rate with the share of Trump voters, share of adults with a high school diploma, a quadratic term for the Trump variable, and an instrumental variable multiplying the Trump variable with the education variable as explanatory variables. This is the equation: . This regression resulted in an R-squared of 0.3431 and adjusted R-squared of 0.279, the highest of any models. It seems that the addition of a quadratic term and an instrumental variable created a stronger model, but the R-squared is still rather low. There does seem to be a little bit of heteroskedasticity in the residauls, but it is not very obvious. The residuals are shown below in Figure 8.

Figure 8

Although many different regressions were run, there did not seem to be a model that could accurately find a relationship between the variables in the data set and their impact on the 2016 post-election hate crime rate. The models were most likely a result of omitted variable bias, as there was most likely a variable that had a stronger correlation with the post-election hate crime rate that was not included in the data set. However, throughout the tested models, the two variables that were the most significant were the share of voters per state who voted for Donald Trump in the 2016 presidential election and the share of adults 25 or older with a high school degree. In each linear model, the Trump variable had a negative coefficient, implying that a larger share of Trump voters in a state led to less post-election hate crimes. The education variable had positive coefficients, meaning that an increase in the percentage of a state’s population that held a high school degree in 2009 increased post-election hate crime rates, a result that seemed surprising to me, as many studies have linked higher education attainment with lower crime rates (phys.org).

All in all, more past data concerning hate crimes could give a better idea of what causes them and have those results compared to the findings of this report. It would have been interesting to compare the 2016 post-election hate crime rate with other hate crime rates immediately after elections, to see if 2016 is an outlier of the data or if this behavior has been consistent throughout U.S. history. 2008 post-election hate crime rates especially would have been interesting to compare, as that was the election in which Barack Obama became the first African American United States president. Other variables, such as more accurate demographic data, could also create a clearer picture on what caused high hate crime rates shortly after the 2016 Presidential election. However, based off of the information given in the data set, it does seem that the slate.com’s comments about increased hate crimes after the 2016 presidential election are true, and social media is just capturing this fact and spreading hate related incidents, not exaggerating them. Hopefully in the near future, the U.S. can come to a point where the hate crime rate is zero. Until then, stricter punishments and other methods must be implemented to deter citizens from committing hate crimes in order to decrease the disturbingly high hate crime rates present in the United States.

References:

<https://www.criminaldefenselawyer.com/resources/hate-crimes-laws-and-penalties.htm>

<https://data.worldbank.org/indicator/SP.POP.TOTL?locations=US>

<https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5137>

<https://www.theatlantic.com/business/archive/2014/10/does-inequality-cause-crime/381748/>

<https://www.investopedia.com/terms/g/gini-index.asp>

<https://github.com/fivethirtyeight/data/tree/master/hate-crimes>

<http://www.cyclismo.org/tutorial/R/confidence.html>

<https://www.voanews.com/a/hate-crimes-rising-in-us/4034719.html>

<http://www.slate.com/articles/news_and_politics/politics/2016/12/hate_in_america_a_list_of_racism_bigotry_and_abuse_since_the_election.html>

<https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/>

<https://phys.org/news/2011-12-lowers-crime.html>

Professor Martinet’s “Regression in R,” “Graphics in R,” “Introduction to R,” “Simulations in R,” and “Hypothesis testing in R”

R Script

## Final STAT 3080 Project ##

## By: Nick Bruno ##

library(car)

library(leaps)

library(lmtest)

library(boot)

library(ggplot2)

# Hate crimes

crime <- read.csv("hate\_crimes.csv")

# Reads in the dataset I will analyze

# Before I start I will take out Washington DC, for reasons mentioned in the paper

crime = crime[-9,]

# ONLY HIT ONCE OR ELSE IT MESSES UP THE OTHER RESULTS

## Comparing hate crimes in 2016 compared to 2010 - 2015 ##

## Post-election crime rate data

summary(crime$hate\_crimes\_per\_100k\_splc)

# Takes a look at the statistics of the hate crime rate in 2016 (numbers per 100,000)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 0.06745 0.14220 0.22580 0.27760 0.34680 0.83280

0.2776 \* 10

# Number of hate crime victims per 1 million US citizens

# [1] 2.776

2.776 \* 323.1

# The US population in 2016 was 323.1 million people, so multiplying that by the number of hate crime

# gives around the number of hate crime victims in the US during the post-election time frame

# [1] 896.9256

# Confidence Interval

# 2016 Hate crime rate confidence interval

crimemean <- mean(crime$hate\_crimes\_per\_100k\_splc, na.rm=T)

# [1] 0.2776102

crimesd <- sd(crime$hate\_crimes\_per\_100k\_splc, na.rm=T)

# [1] 0.1777228

length(crime$hate\_crimes\_per\_100k\_splc, na.rm=T)

# [1] 50

summary(crime$hate\_crimes\_per\_100k\_splc)

# 4 null values, so n = 50-4

n <- length(which(crime$hate\_crimes\_per\_100k\_splc > 0))

# Because all of the values are greater than zero, this gives the total number of data points

# within this variable, taking into account the null values. In this case, there are

# 4 null values out of 50 states, so n = 46

# [1] 46

tstat <- qt(0.975, df=n-1)

# [1] 2.014103

error <- tstat\*crimesd/sqrt(n)

# [1] 0.05104793

lower <- crimemean-error

# [1] 0.2265622

upper <- crimemean+error

# [1] 0.3286581

# CI: (0.224833, 0.3303873)

# 2010 - 2015 hate crime rate data

summary(crime$avg\_hatecrimes\_per\_100k\_fbi)

# Same as above. Divided by 6 because it was the sum of the averages over 6 years (2010-2015)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 0.2669 1.2830 1.9370 2.1920 3.1360 4.8020

hatepermil <- 2.192 \* 10

# Number of hate crime victims per 1 million US citizens

# [1] 21.92

# I will now find the average number of hate crimes per year from 2010-2015. I will do this by first

# finding the US population per year during that time frame.

US15 <- 320.9

US14 <- 318.6

US13 <- 316.2

US12 <- 314

US11 <- 311.7

US10 <- 309.3

avgpop10to15 <- (US15 + US14 + US13 + US12 + US11 + US10)/6

# Average US population from 2010-2015

# [1] 315.1167

tot1015 <- hatepermil\*avgpop10to15

# [1] 6907.357

# This number shows the average numbe of victims who fell to a hate crime per million per year from 2010-2015

per1000 <- 0.2776 \* 36.5

# [1] 10.1324

permilnew <- per1000\*10

# [1] 101.324

US16 <- 323.1

tot16 <- US16\*permilnew

# [1] 32737.78

tot16/tot1015

# [1] 4.739553

## Taking a deeper look into the variables of the data set ##

# Find median income per household in US across states

median(crime$median\_household\_income)

# [1] 54613

# Highest and lowest crime rates

sort(crime$avg\_hatecrimes\_per\_100k\_fbi)

# Highest:

# Oregon = 0.83284961

# Washington = 0.67748765

# Lowest:

# Mississippi = 0.06744680

# Arkansas = 0.06906077

# Median Household Incomes of these four states

# Oregon

crime[37,2]

# [1] 58875

# Washington

crime[47,2]

# [1] 59068

# Mississippi

crime[24,2]

# [1] 35521

# Arkansas

crime[4,2]

# [1] 44922

## Histograms of hate crime rates ##

# Post-election hate crime histogram

crimeratehist <- ggplot(crime, aes(x=crime$hate\_crimes\_per\_100k\_splc))

crimeratehist + geom\_histogram(binwidth=0.075, fill='white', colour='black') + labs(title='Hate Crime Rate Histogram', x = 'Hate Crime Rate')

# post-election hate crime rate histogram

# 2010-2015 hate crime histogram

crimeratehist2010 <- ggplot(crime, aes(x=crime$avg\_hatecrimes\_per\_100k\_fbi))

crimeratehist2010 + geom\_histogram(binwidth=0.5, fill='white', colour='black') + labs(title='2010 - 2015 Hate Crime Rate Histogram', x = 'Hate Crime Rate')

# 2010 - 2015 hate crime rate histogram

## Scatterplots of each column in the data set ##

# Household Income Scatterplot

ggplot(crime, aes(x=crime$median\_household\_income, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title="Relationship Between Household Income and Hate Crime Rate",

x="Median Household Income", y="Hate Crime Rate")

# somewhat strong positive correlation

# Gini Index plot

ggplot(crime, aes(x=crime$gini\_index, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title="Relationship Between Economic Inequality and Hate Crime Rate", x= "Gini Index", y="Hate Crime Rate")

# Very weak negative correlation

# Trump plot

ggplot(crime, aes(x=crime$share\_voters\_voted\_trump, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between Trump Voter Support and Hate Crime Rate', x='Share of Trump Voters', y='Hate Crime Rate')

# substantial negative correlation

# Unemployment plot

ggplot(crime, aes(x=crime$share\_unemployed\_seasonal, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between Unemployment and Hate Crime Rate', x='Unemployemnt', y='Hate Crime Rate')

# slight negative correlation

# Metro Area plot

ggplot(crime, aes(x=crime$share\_population\_in\_metro\_areas, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between Metro Area Habitation and Hate Crime Rate', x='Metro Habitation', y='Hate Crime Rate')

# almost absolutely no correlation

# Education plot

ggplot(crime, aes(x=crime$share\_population\_with\_high\_school\_degree, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between Education and Hate Crime Rate', x = 'High School Degree Owners', y='Hate Crime Rate')

# strange positive correlation (possible outlier though)

# Noncitizens plot

ggplot(crime, aes(x=crime$share\_non\_citizen, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between Share of Population that are not US Citizens and Hate Crime Rate', x='Non US Citizens', y='Hate Crime Rate')

# very small positive correlation

# White Poverty Plot

ggplot(crime, aes(x=crime$share\_white\_poverty, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between White Poverty and Hate Crime Rate', x='White Residents Living in Poverty', y='Hate Crime Rate')

# slightly negative correlation (outlier pulling the line in a negative direction)

# Nonwhite population plot

ggplot(crime, aes(x=crime$share\_non\_white, y=crime$hate\_crimes\_per\_100k\_splc)) + geom\_point() + geom\_smooth(method=lm) + labs(title='Relationship Between the Non-White Population and Hate Crime Rate', x='Non-White Population', y='Hate Crime Rate')

# pretty significant negative correlation

## White poverty sampling

histpov <- ggplot(crime, aes(x=crime$share\_white\_poverty))

histpov + geom\_histogram(binwidth = .01, fill='white', colour='black') + labs(title='Share of White Poverty Histogram', x = 'White Poverty Rate')

# will do normal sampling (most normal)

## Trump histogram

histpov <- ggplot(crime, aes(x=crime$share\_voters\_voted\_trump))

histpov + geom\_histogram(bins = 10, fill='white', colour='black') + labs(title='Share of Trump Voters Histogram', x = 'Share of Voters for Trump')

# almost normal (kind of trimodal)

## White poverty Monte Carlo simulation sampling

meanpov <- mean(crime$share\_white\_poverty)

# [1] 0.0928

sdpov <- sd(crime$share\_white\_poverty)

# [1] 0.02382247

K <- 10000

# number of simulations that will be run

a <- 4

## Draw 10,000 samples of size 5 from the population distribution

set.seed(21)

samps <- replicate(K, rnorm(5, mean=meanpov, sd=sdpov))

mean(samps)

# [1] 0.09297471

sd(samps)

# [1] 0.02392001

## Determine the sample mean from each random sample

means5 <- apply(samps,2,mean)

## QQ Plot of Monte Carolo

mean\_data <- data.frame(X=means5)

y <- quantile(mean\_data$X, c(0.25, 0.75))

# 25% 75%

# 0.08565365 0.10020053

x <- qnorm(c(0.25, 0.75))

# [1] -0.6744898 0.6744898

slope <- diff(y)/diff(x)

# 75%

# 0.01078362

int <- y[1] - slope\*x[1]

# 25%

# 0.09292709

# Actual QQ Plot

ggplot(mean\_data, aes(sample = X)) + stat\_qq() +

geom\_abline(intercept=int, slope=slope) + labs(title="n=5") + labs(title='White Poverty QQ Plot', y='Share of White Poverty')

## Correlation tests testing the correlation between each variable and the post-election hate crime

# rate

cor.test(hatecrimes, trump, method="pearson", alternative="two.sided")

# p-value = 0.003873

cor.test(hatecrimes, education, method="pearson", alternative="two.sided")

# p-value = 0.003532

cor.test(hatecrimes, metro, method="pearson", alternative="two.sided")

# p-value = 0.873

cor.test(hatecrimes, householdincome, method="pearson", alternative="two.sided")

# p-value = 0.0532

cor.test(hatecrimes, gini\_index, method="pearson", alternative="two.sided")

# p-value = 0.4531

cor.test(hatecrimes, unemployment, method="pearson", alternative="two.sided")

# p-value = 0.3837

cor.test(hatecrimes, nonuscitizen, method="pearson", alternative="two.sided")

# p-value = 0.6459

cor.test(hatecrimes, poverty, method="pearson", alternative="two.sided")

# p-value = 0.5134

cor.test(hatecrimes, nonwhite, method="pearson", alternative="two.sided")

# p-value = 0.07261

## Linear Regressions ##

# First will set variables to shorten the plot equations

householdincome <- crime$median\_household\_income

hatecrimes <- crime$hate\_crimes\_per\_100k\_splc

gini\_index <- crime$gini\_index

# Model 1 (scrapped from report)

model1 <- lm(hatecrimes ~ median\_household\_income + gini\_index, data=crime)

summary(model1)

# Low r-squared (0.08284)

# Neither variables are significant

# Household income has a positive coefficient, gini index is a negative coefficient

# Gini index small, but substantially larger than the household income

# Residual info:

# Residuals:

# Min 1Q Median 3Q Max

# -0.25516 -0.10431 -0.02547 0.07459 0.53033

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 9.195e-02 7.797e-01 0.118 0.9067

# median\_household\_income 5.433e-06 3.000e-06 1.811 0.0771 .

# gini\_index -2.434e-01 1.567e+00 -0.155 0.8773

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1741 on 43 degrees of freedom

# (4 observations deleted due to missingness)

# Multiple R-squared: 0.08284, Adjusted R-squared: 0.04018

# F-statistic: 1.942 on 2 and 43 DF, p-value: 0.1558

model1res <- residuals(model1)

# Creates a list of residual points

plot(model1res)

# residuals look good

# Model 1 (actual)

trump <- crime$share\_voters\_voted\_trump

model2 <- lm(hatecrimes ~ median\_household\_income + gini\_index + trump, data=crime)

summary(model2)

# R-squared still low, but increased significantly from Model 1, is now 0.2067

# All coefficients are negative now

# Only the Trump voter share variable is significant

# F-statistic is significant, meaning that at least one of the variables is important

# Residuals:

# Min 1Q Median 3Q Max

# -0.23775 -0.11698 -0.02146 0.07808 0.47173

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 1.848e+00 1.005e+00 1.840 0.0728 .

# median\_household\_income -2.118e-06 4.083e-06 -0.519 0.6066

# gini\_index -2.132e+00 1.649e+00 -1.293 0.2032

# trump -9.888e-01 3.862e-01 -2.560 0.0141 \*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1639 on 42 degrees of freedom

# (4 observations deleted due to missingness)

# Multiple R-squared: 0.2067, Adjusted R-squared: 0.15

# F-statistic: 3.647 on 3 and 42 DF, p-value: 0.02002

resid2 <- residuals(model2)

# creates a list of residuals

plot(resid2)

# Residuals look good

vif(model2)

# tests for multicollinearity

# median\_household\_income gini\_index trump

# 2.330944 1.392941 2.122251

anova(model2)

# Response: hatecrimes

# Df Sum Sq Mean Sq F value Pr(>F)

# median\_household\_income 1 0.11701 0.117010 4.3583 0.04293 \*

# gini\_index 1 0.00073 0.000731 0.0272 0.86969

# trump 1 0.17599 0.175989 6.5551 0.01414 \*

# Residuals 42 1.12761 0.026848

# Model 3 (but second one mentioned in the paper)

unemployment <- crime$share\_unemployed\_seasonal

metro <- crime$share\_population\_in\_metro\_areas

education <- crime$share\_population\_with\_high\_school\_degree

nonuscitizen <- crime$share\_non\_citizen

poverty <- crime$share\_white\_poverty

nonwhite <- crime$share\_non\_white

model3 <- lm(hatecrimes ~ unemployment + metro + education + nonuscitizen + poverty + nonwhite, data=crime)

summary(model3)

# R-squared higher, but still low (0.2747) (adjusted R-squared = 0.1571)

# F-statistic is statistically insignificant

# Education is the only variable that is statistically significant

# Nonuscitizen is almost significant, so I do not want to throw that one out

# metro is incredibly insignificant

# unemployment, education, and nonuscitizen are all positive coefficients

# Residuals:

# Min 1Q Median 3Q Max

# -0.31247 -0.10031 0.00521 0.09004 0.35086

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -2.94891 1.38283 -2.133 0.0397 \*

# unemployment 4.29035 3.25890 1.317 0.1961

# metro 0.06149 0.23743 0.259 0.7971

# education 3.19754 1.33301 2.399 0.0216 \*

# nonuscitizen 2.54654 1.45182 1.754 0.0877 .

# poverty 1.80619 1.68995 1.069 0.2921

# nonwhite -0.37105 0.33779 -1.098 0.2791

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1573 on 37 degrees of freedom

# (6 observations deleted due to missingness)

# Multiple R-squared: 0.2747, Adjusted R-squared: 0.1571

# F-statistic: 2.336 on 6 and 37 DF, p-value: 0.0517

resid3 <- residuals(model3)

plot(resid3)

# Residuals more spread out, but still look random and good

vif(model3)

# unemployment metro education nonuscitizen poverty nonwhite

# 1.807864 2.519537 3.367868 3.395102 2.782949 4.024919

# Model 4 (third mentioned in the paper)

model4 <- lm(hatecrimes ~ unemployment + education + nonuscitizen, data=crime)

summary(model4)

# Adjusted R-squared even lower than model 3 (0.1293)

# All coefficients are positive

# Education is the only significant variable

# Maybe some multicollinearity between unemployment and nonuscitizen?

# F-statistic is significant

# Residuals:

# Min 1Q Median 3Q Max

# -0.28957 -0.09306 0.00074 0.08488 0.43912

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -2.2927 0.9036 -2.537 0.01517 \*

# unemployment 3.1816 3.0421 1.046 0.30191

# education 2.7163 0.9244 2.938 0.00545 \*\*

# nonuscitizen 0.9852 0.8365 1.178 0.24586

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1599 on 40 degrees of freedom

# (6 observations deleted due to missingness)

# Multiple R-squared: 0.19, Adjusted R-squared: 0.1293

# F-statistic: 3.128 on 3 and 40 DF, p-value: 0.0362

resid4 <- residuals(model4)

plot(resid4)

# residuals may be a little heterodastic

vif(model4)

# unemployment education nonuscitizen

# 1.524902 1.567763 1.091009

# Model 5 (fourth mentioned in the paper)

model5 <- lm(hatecrimes ~ trump + education, data=crime)

summary(model5)

# Residuals:

# Min 1Q Median 3Q Max

# -0.25824 -0.10538 -0.01873 0.05752 0.46125

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -0.8895 0.6690 -1.330 0.1907

# trump -0.6168 0.2634 -2.342 0.0239 \*

# education 1.6992 0.7136 2.381 0.0218 \*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1553 on 43 degrees of freedom

# (4 observations deleted due to missingness)

# Multiple R-squared: 0.2706, Adjusted R-squared: 0.2367

# F-statistic: 7.978 on 2 and 43 DF, p-value: 0.00113

vif(model5)

# trump education

# 1.099679 1.099679

resid5 <- residuals(model5)

plot(resid5)

# Model 6 (not mentioned in paper)

trump2 <- trump^2

education2 <- education^2

model6 <- lm(hatecrimes ~ trump + education + trump2, data=crime)

summary(model6)

# Residuals:

# Min 1Q Median 3Q Max

# -0.24815 -0.08992 -0.02347 0.07548 0.46124

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -0.2451 0.8012 -0.306 0.7611

# trump -4.2851 2.5887 -1.655 0.1053

# education 1.9593 0.7285 2.689 0.0102 \*

# trump2 3.7352 2.6226 1.424 0.1618

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1534 on 42 degrees of freedom

# (4 observations deleted due to missingness)

# Multiple R-squared: 0.3042, Adjusted R-squared: 0.2545

# F-statistic: 6.122 on 3 and 42 DF, p-value: 0.001498

vif(model6)

# going to obviously be very high

# trump education trump2

# 108.727287 1.173404 110.397494

resid6 <- residuals(model6)

plot(resid6)

ggplot(model6, aes(x=.fitted, y=.resid)) + geom\_point() +

geom\_hline(yintercept=0, linetype="dashed")

new <- predict.lm(model5, interval="confidence")

plot(new)

# Model 7 (used in paper)

model7 <- lm(hatecrimes ~ trump + education + trump2 + trump\*education, data=crime)

summary(model7)

# best adjusted R-squared model so far (0.3431) and adjusted r-squared (0.279)

# Residuals:

# Min 1Q Median 3Q Max

# -0.27455 -0.08798 -0.01757 0.08061 0.43708

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -5.637 3.551 -1.587 0.1201

# trump 7.323 7.878 0.930 0.3580

# education 7.881 3.870 2.037 0.0482 \*

# trump2 2.435 2.711 0.898 0.3743

# trump:education -12.048 7.737 -1.557 0.1271

#---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1509 on 41 degrees of freedom

# (4 observations deleted due to missingness)

# Multiple R-squared: 0.3431, Adjusted R-squared: 0.279

# F-statistic: 5.353 on 4 and 41 DF, p-value: 0.001458

resid7 <- residuals(model7)

plot(resid7)

ggplot(model7, aes(x=.fitted, y=.resid)) + geom\_point() +

geom\_hline(yintercept=0, linetype="dashed") + labs(title='Residuals')

# Maybe a littler heteroskedastic

# Model 9 (not used in paper)

model9 <- lm(hatecrimes ~ trump + education + trump\*education, data=crime)

summary(model9)

# education is the only statistically significant value (but the other ones are very close)

# R-squared = 0.3302 and adjusted R-squared = 0.2823

# Residuals:

# Min 1Q Median 3Q Max

# -0.26289 -0.09102 -0.01923 0.07622 0.43280

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) -6.975 3.216 -2.169 0.0358 \*

# trump 11.550 6.303 1.833 0.0740 .

# education 8.779 3.729 2.354 0.0233 \*

# trump:education -14.188 7.344 -1.932 0.0601 .

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.1506 on 42 degrees of freedom

# (4 observations deleted due to missingness)

# Multiple R-squared: 0.3302, Adjusted R-squared: 0.2823

# F-statistic: 6.901 on 3 and 42 DF, p-value: 0.0006994

residfit3 <- residuals(model9)

plot(residfit3)

# residuals look fine. Maybe heteroskedasticity?

# F-statistic for the model is significant

# education is significant, but Trump is not

vif(dtime.fit3.1)

# there is obviously multicollinearity because of the interaction term