Turbulent Times and Hate Crimes in the United States

Sarah Hafez

Decemeber 17, 2021

In an article titled "Higher Rates of Hate Crimes are tied to Income Inequality", FiveThirtyEight claim - as the title alludes to - that hate crimes in the United States are correlated to income inequality. However, through a huge misstep on their part, they miss the fact that the data includes an outlier. In doing so, they reach a conclusion that is heavily influenced by it. In this paper, I show that a more accurate conclusion of their data is to imply that hate crimes are tied to the share of non-citizens and non-white population. I also show that during the 10 day-period after the 2016 U.S. elections, hate crime rates were higher in states where the majority did **not** vote for Trump.

During the 2016 elections, the United States faced a higher than usual number of *reported* hate crimes. In order to investigate why, the authors collected data for the country's fifty states, measuring various socioeconomic factors and two measures for hate crime:

1. Annual Average Hate Crimes Per 100,000 People (2010-2015) [FBI]
2. Hate Crimes Per 100,000 People (10 days post elections) [Southern Poverty Law Center (SPLC)]

They then built two models: one with the pre-elections (FBI) data as the response variable and another with the post-elections (SPLC) data as the response. The goal was to compare both models to determine whether the elections have influenced the relationship between hate crime and socioeconomic factors. Table 1 shows the explanatory variables found in the data.

Table 1: Explanatory Variables found in the Data

Percent Adults 25 and older with at least a High School Degree (2009)	Percent of non-white Population and non-citizen Population (2015)
Median Household Income (2016)	Seasonally Adjusted Unemployment (2016)
Gini-Index (measure of Income Inequality) (2015)	Percentage of Population Voting for Trump (2016)
Percent Population in Metropolitan Areas (2015)	Poverty Among White People (2015)

States with more Trump Voters faced less Hate Crimes

A stand out explanatory variable in the data is the percentage of Trump voters per state. Given the context of the article and U.S. politics, it seems like the authors hoped to link the percentage of Trump voters to the incidence of hate crime. Surprisingly, they did not elaborate further on that variable. In this section, we investigate whether the percentage of Trump voters had any effect on hate crime rates post the elections.

Trump won the majority of votes in twenty-five states - labeled Trump majority states, and he lost the majority in the other twenty-five states - labeled Trump minority states. Before the elections, there were no significant differences between the Trump majority and the minority states when it comes to hate crime rates. From 2010 to 2015, the Trump majority states had on average of 2 hate crimes while the minority states had an average of 2.4. As shown in Figure 1 (left), both type of states had a similar distribution of hate crimes with slightly different means. However, during the 10-days post the elections, the story turned different. The Trump majority states had significantly less hate crimes. The majority states faced a rate of 0.2 hate crimes in the 10 day period while the minority states faced a rate of 0.35. Figure 1 (right) emphasizes this story well. Most of the Trump minority states faced a hate crime rate larger than 0.3 while most Trump majority states enjoyed a rate of 0.2. This suggests that the trump minority states faced ~50% higher hate crime rates in the 10 day period post-elections.

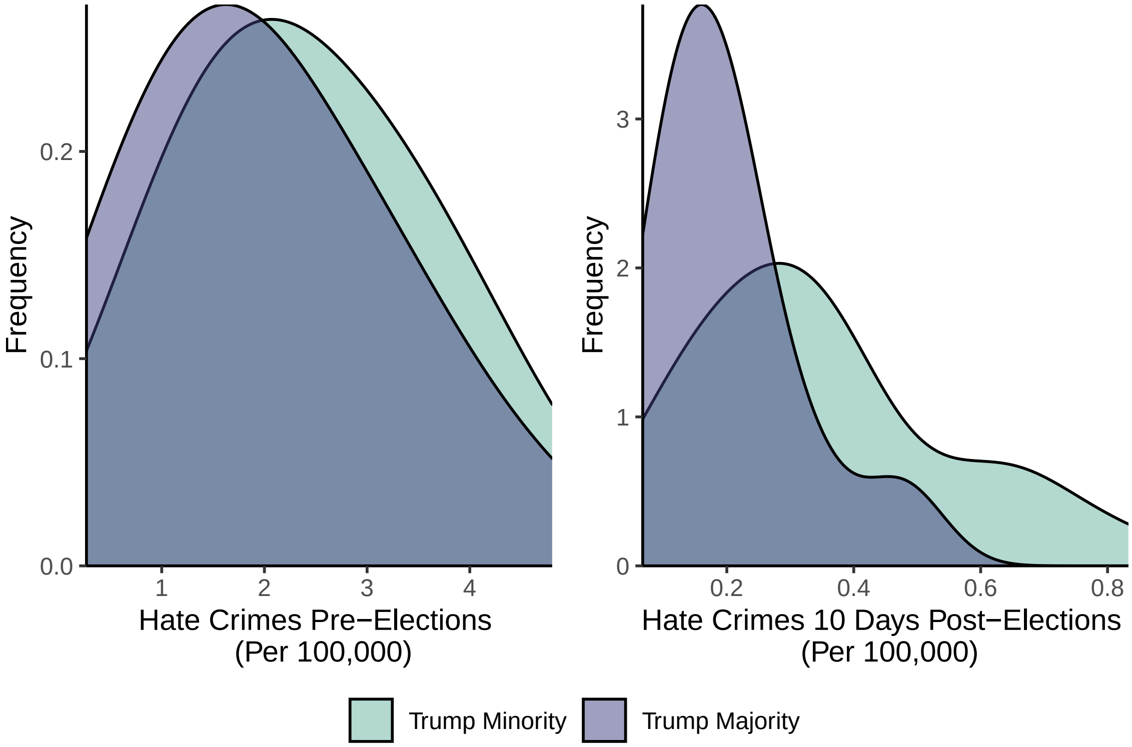

Figure 1: Hate Crimes Distribution in Trump Majority vs Minority States

Figure 2 also shows that there was a negative relationship between hate crimes 10-days post the elections and the percentage of Trump voters. As the latter increased, the former decreased. The plot also emphasis that the Trump majority states faced a more stable rate of hate crime, while many minority states faced an irregular spike in hate crimes. Finally, it is important to note that this is merely an observation in the data and that it provides us no clues on why this is the case. For example, we have no information on the perpetrators of such crimes. The only explanation the data can give us is on the relationship between hate crimes and socioeconomic factors.

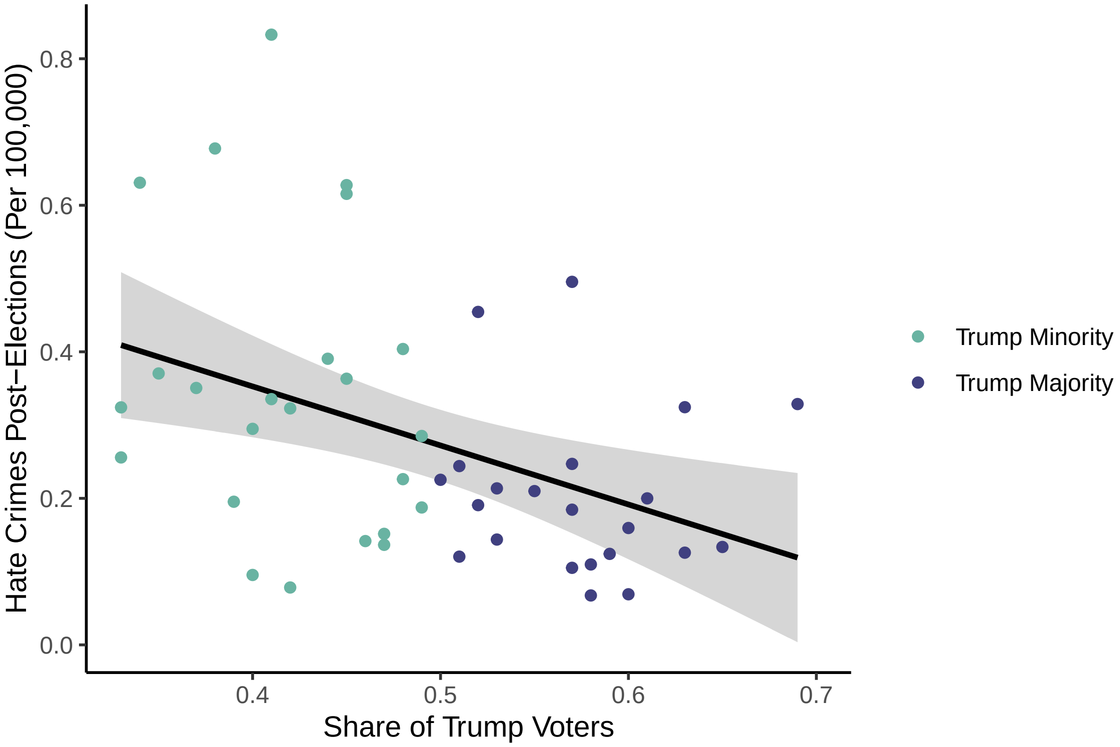

Figure 2: Hate Crimes Post-Elections and Trump Voters

The Relationship Between Hate Crimes and Socioeconomic Factors

To be able to concisely study this relationship, this part of the report will only focus on the pre-elections data. Firstly, this is to avoid repetition. Secondly, the post-elections data was collected over a ten-day period that was highly sensitive and turbulent. It would be inaccurate to claim any relationship between hate crimes and socioeconomic factors when our data itself could be an exception in comparison to other times of the year.

The Influence of the District of Columbia The data included fifty-one observations, one for each state and another for the District of Columbia, more famously known as Washington, D.C. - the capital of the United States. Not only is D.C. the only non-state observation, but it is also the only observation with an abnormally high hate crime rates, standing at a rate of 10 per 100,000 people.

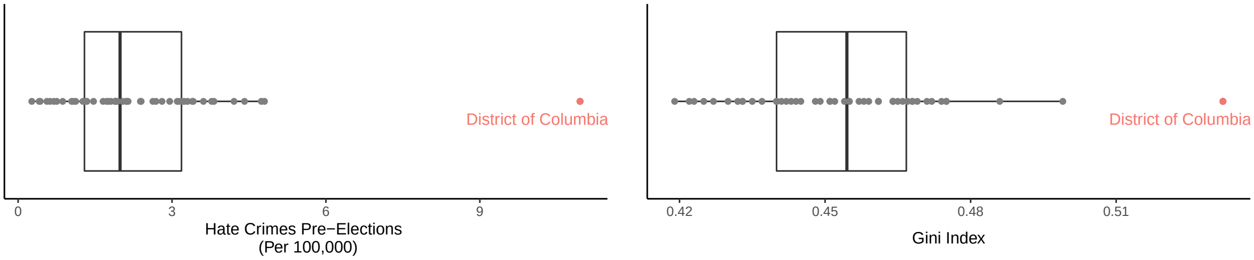

Figure 3: Boxplots for Hate Crime and Income Inequality Values

The Gini Index measures the level of income inequality with it ranging from 0 (perfect equality) and 1 (maximum inequality). D.C. also stands out in this case, with it being the only observation to exceed a Gini Index of 0.5 (at 0.532). Figure 3 shows the box plots for the pre-elections hate crime data (left) and the Gini Index (right). In both we can see the extent in which D.C. can only be labeled as an "outlier" to the data.

The author's model heavily picked up these drastic D.C. values, thus choosing to emphasize the relationship between income inequality and Gini Index. If we remove D.C. from our dataset, this relationship becomes less meaningful. Figure 4 further illustrates this notion. When we model the relationship between hate crime and the gini index, while including the District of Columbia, the best-fit line is positive (in red), indicating that higher levels of income inequality lead to more hate crimes. However, when we exclude the United States capital, the best-fit line falls flat (in black), indicating that higher levels of income inequality may lead to *insignificant* changes to hate crime rates.

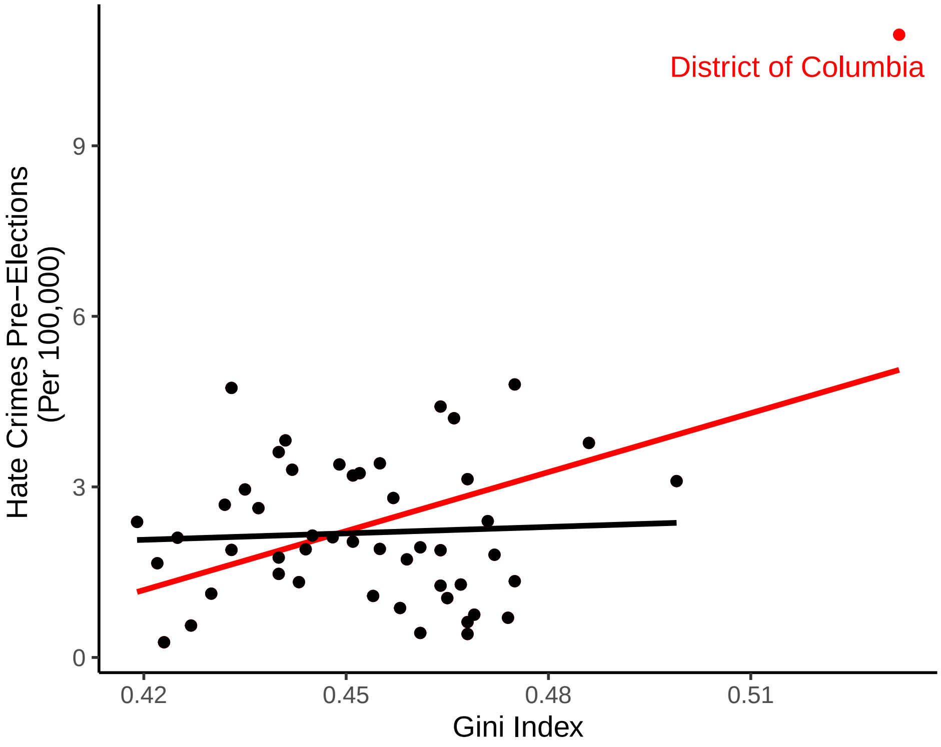

Figure 4: The Relationship between Income Inequality and Hate Crime

Share of Non-Citizen and Non-White Population After removing The District of Columbia from the dataset, the new model shows that hate crimes are instead tied to the share of non-citizens and non-white populations. Figure 5 shows that there is a positive relationship between the percentage of non-citizens and hate crime rates. If the share of non-citizens increases by 1%, hate crime rates will increase by a factor of 1.13. In other words, hate crime rates will increase by 13%. The figure also shows that there is a negative relationship between the percentage of non-white residents and hate crime rates. If the share of non-white population increases by 1%, hate crime rates will decrease by a factor of 0.97, meaning that they will decrease by 3%.

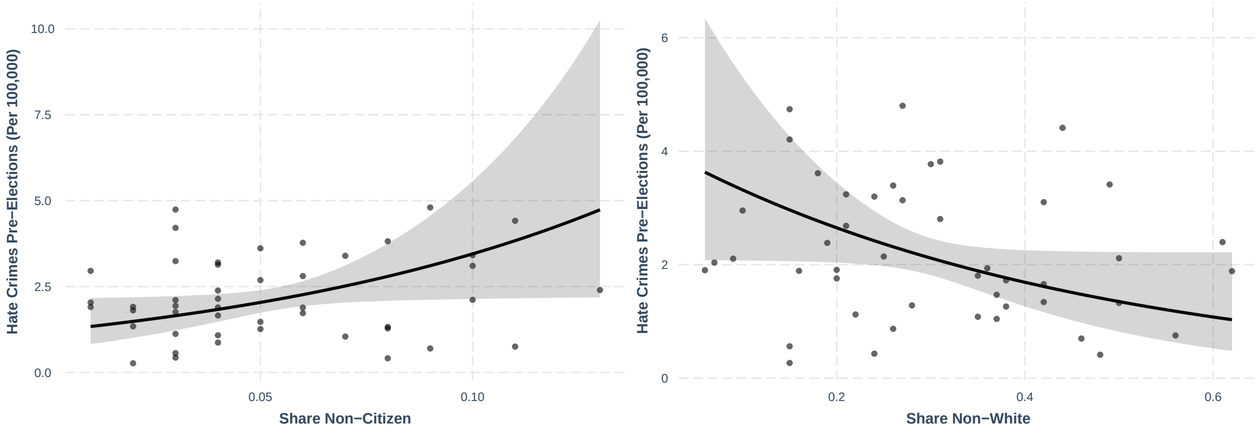

Figure 5: Effect of the Share of Non-Citizen and Non-White Population on Hate Crimes

The model suggests that when there is more diversity in a state, hate crimes decrease. It also suggests that the when there are more immigrants, hate crimes increase. Overall, the model shows that hate crimes are tied to superficial differences and the fear of "the other" - fear of those with different backgrounds or different color. It gives explanations to hate crimes that are more intuitive and more inherent to its definition. Finally, it is important to note what this model can and can not do. This model suggests that hate crime rates are correlated with the share of non-white and non-citizen population. However, it is not able to accurately capture or predict true hate crime rates. The model also does not invalidate that income inequality may have caused for higher hate crime rates in the District of Columbia. Rather, it merely tries to focus on the bigger picture.

Technical Analysis

Mean Comparison of Trump Majority States and Trump Minority States

Trump Majority states where defined as the states where more than 50% of the voters voted for Trump. I also dropped D.C. as it was an outlier.

	Mean	**Variance**
Pre-Elections Trump Majority States (25)	2	1.45
Pre-Elections Trump Minority States (24)	2.39	1.4
Post-Elections Trump Majority States (22)	0.203	0.013
Post-Elections Trump Minority States (24)	0.346	0.04

For both the pre-elections and post-election data, the variance and the sample size for the two groups was unequal. Welch's t-test was used to determine the significance of the differences in the mean hate crime rate. In the case of pre-elections, the difference is ~0.4 (in favor of trump majority states) and was found to be insignificant (p-value: 0.25). In the case of post-elections, the difference is ~0.14 (also in favor of trump majority states) which was significant with a p-value of ~0.005.

Linear Regression Model

Model 1: Linear Regression With an Outlier In the article, the authors used a multivariate linear regression model. If we run the same model, then we can indeed see that the Gini Index is the only significant factor, alongside the constant, see appendix A for the summary table. The model has a R_α^2 of 0.38, indicating that it has a poor fit to the data. The diagnostics plots also raised red flags about the District of Columbia - the 9th observation. Figure 6 shows some of the model's diagnostic plots, the Residuals vs Leverage plot shows that observation 9 has high leverage and residuals. The figure also shows that it has Cook's distance higher than one, which indicates that this observation is 100% an outlier. The Residuals vs Leverage plot also signals out other observations: 10 and 17, Vermont and Kentucky respectively. However, their effect on the model isn't nearly as large as the effect of D.C.. In fact, once we remove D.C. from the data, the diagnostic plots no longer signal them out (as seen in Figure 7)

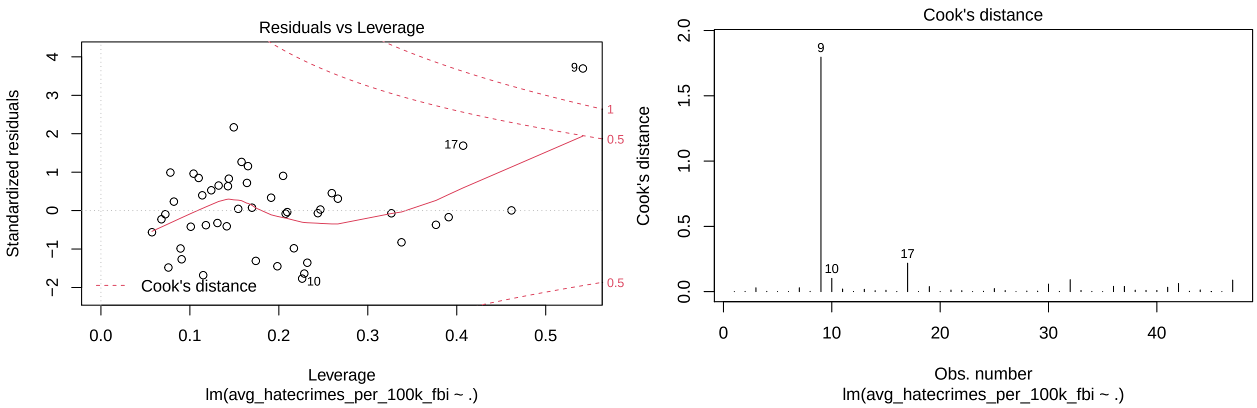

Figure 6: Diagnostic Plots: FiveThirtyEight's Linear Regression Model

Model 2: Linear Regression With No Outliers After removing D.C. from the data, the Gini Index no longer becomes significant and instead we have the share of non-white (at the 7% significance level) and the share non-citizens (at the 5% significance level). Refer to Appendix B for the summary table. However, the model still has a very poor fit with an R_α^2 of 0.15. It is interesting to see how much the R_α^2 fell due to the removal of one observation. This also speaks to how much the previous model was influenced by D.C..

Figure 7 shows the diagnostic plots for the new linear regression model. In a linear regression model, we expect the residuals to be evenly spread through the zero-line in the Residuals vs Fitted plot. Instead, we see a parabola-like shape. This indicates that there is a non-linear relationship in the data that we are not capturing with the linearity assumption. We also see that the Residuals vs Leverage plot no longer shows any worrying signs of an outlier.

Finally, I used step AIC to determine the best model. It narrowed down to the share of non-citizens and non-white population as the best covariates. As seen in appendix C, the Residuals vs Fitted plot also resembled a parabolic shape - even more so this time. The model was also unable to achieve a significantly higher R_α^2. At this point, we must look further than the linear assumption as it clearly is not a good fit to the data.

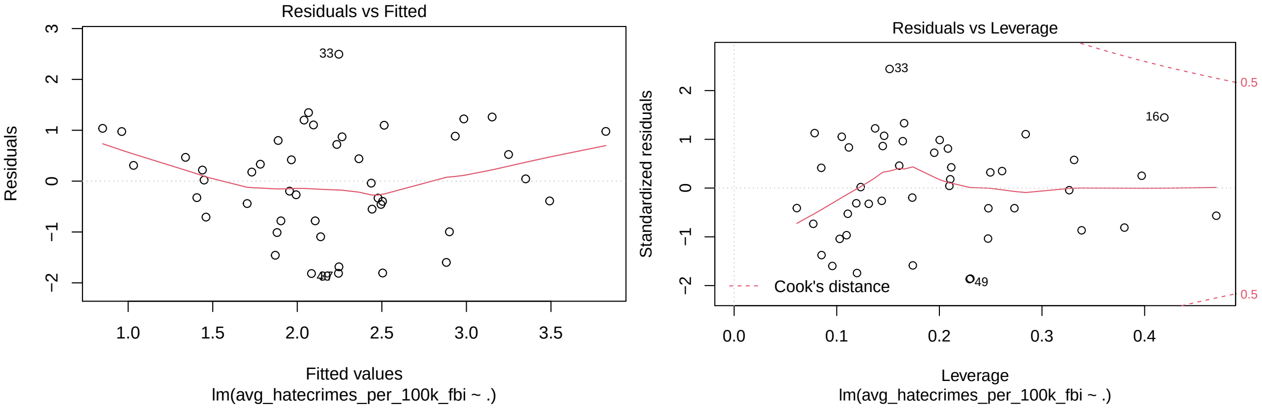

Figure 7: Diagnostic Plots: Linear Regression Model with No Outlier Variables

Poisson Regression

The data that we have is in root count data. However, since it is adjusted for the population size and is an average over a 5 year period, we have it in form of decimals instead of integers. While it is not ideal to use poisson regression with decimals, it is still worth giving it a try.

Table 5 shows that mean and variance for the Pre-Elections data. The variance is much smaller than the mean, indicating that the data is underdispersed. This suggests that hate crimes in the country are in a way correlated. In other words, outside of D.C., each state suffers from *comparable* and *nearly equal* levels of hate crime.

	Mean	**Variance**
Pre-Elections Hate Crimes	2.2	1.43

In this case, the negative binomial and the quasi poisson models can both be used. I will show the results for the quasi poisson model. Since we saw in the linear regression model that the relationship between the counts and the explanatory variables are likely not linear, I opted to use the log link.

Appendix D shows the summary from the quasi poisson regression model. The overdispersion parameter is approximately 0.54. It again shows that the only significant predictors of hate crimes in this case are the share of non-citizens (at the 5% significance level) and non-white population (at the 6% significance level). The Residual Deviance of the Quasi-Poisson model is approximately 22 with 37 degrees of freedom while the null deviance was approximately 30.8. The decrease in deviance suggests that the model is a good fit to the data. Additionally, the goodness of fit test with the residual deviance yields a p-value of 0.98, indicating that there is no evidence of a lack of fit.

In figure 8, we see an indication that the Quasi-Poisson model is a better fit than the linear regression. The "Residual vs Fitted" plot no longer resembles a parabola. There is no distinctive pattern and instead we mostly see a line hovering around the zero mean. Clearly, the variance is not homogeneous, which is also what we expect using the quasi poisson model.

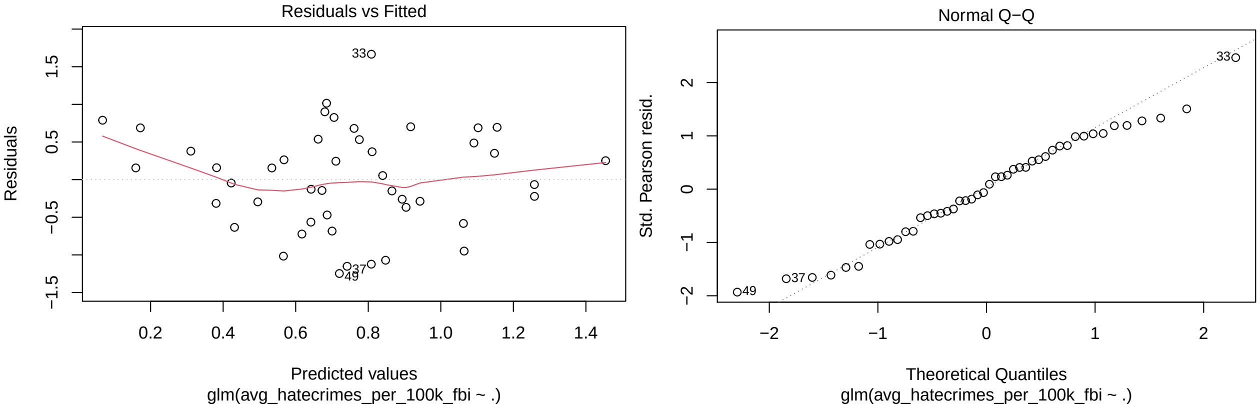

Figure 8: Diagnostic Plots: Quasipoisson Model

The “Normal Q-Q” plot shows that the pearson residuals approximately follow the normal’s theoretical quantiles (indicating that they have an equal variance). This is a good sign as the Pearson residuals correct for the unequal variance in the residuals. This again indicates that our model is a good fit for the data. The plots signal out some variables. However, they are not influential enough to be outliers.

Finally, while our model seems to be a great fit to the data, it might be worth it to consider the quasi poisson model using only the share of non-white population and the share of non-citizens as explanatory variables. Table 4 shows the resulting analysis of deviance table when we compare both models. By considering this new model, we gain an extra 1.9 in deviance while freeing up 6 degrees of freedom. The resulting p-value is 0.925, indicating that we can indeed consider the smaller model with little sacrifices to the goodness of fit.

Table 4: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance
43	23.95	NA	NA
37	22.01	6	1.942

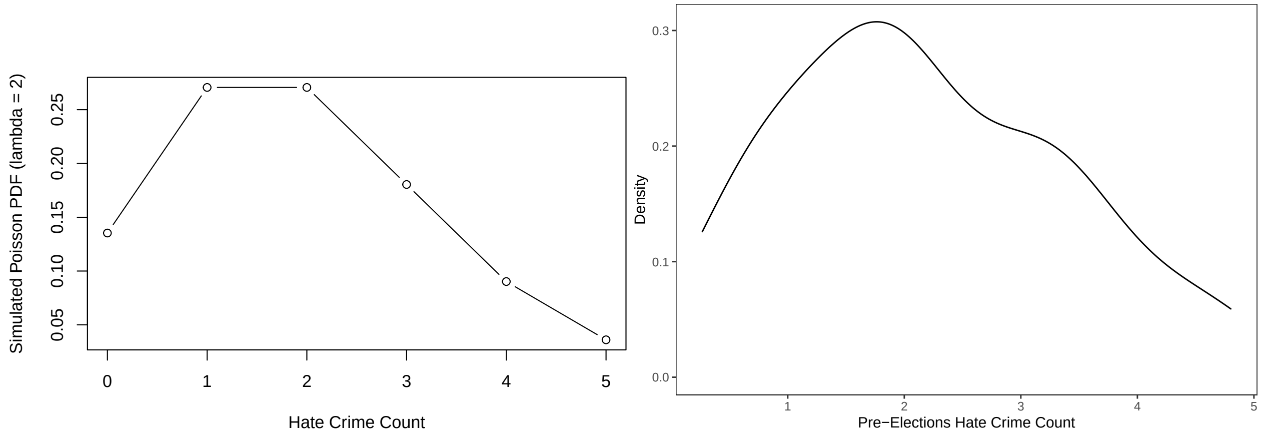

Figure 9: Distribution of Hate Crime Data and Poisson

Conclusion: Which Model is better?

Statistically, the quasi poisson is the more appropriate model to capture the structure of the dataset. As we can see in figure 9, the distribution of the hate crime data does resemble the poisson distribution with lambda 2 (our hate crime mean was 2.2) . Moreover, when we compare the effect plots as seen in figures 10 and 11, the confidence intervals of the poisson regression model do capture more data points. However, it seems like both models agree on one thing: hate crimes are strongly correlated to the share of non-citizens and non-white population. The caveat is that they are **not** perfect predictors of hate crime rates.

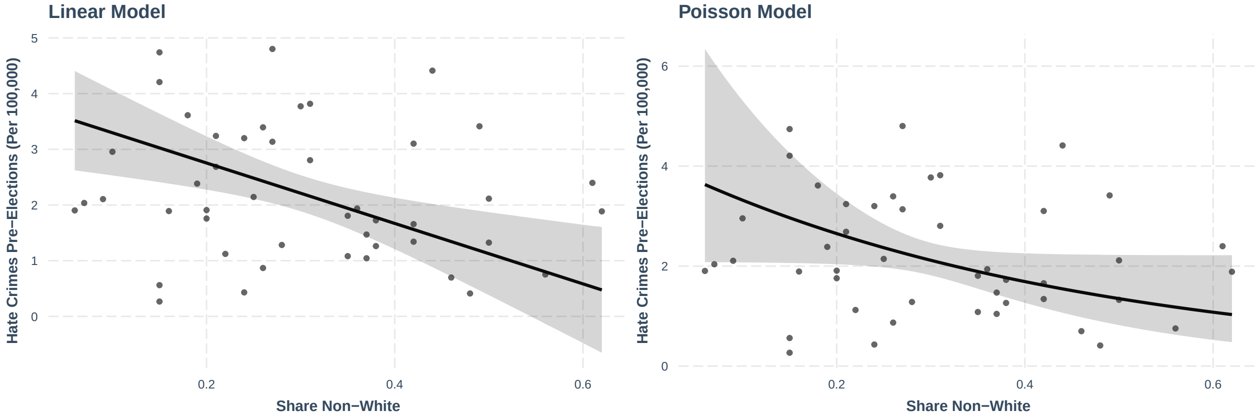

Figure 10: Effect Plots of Non-White Population (Both Models)

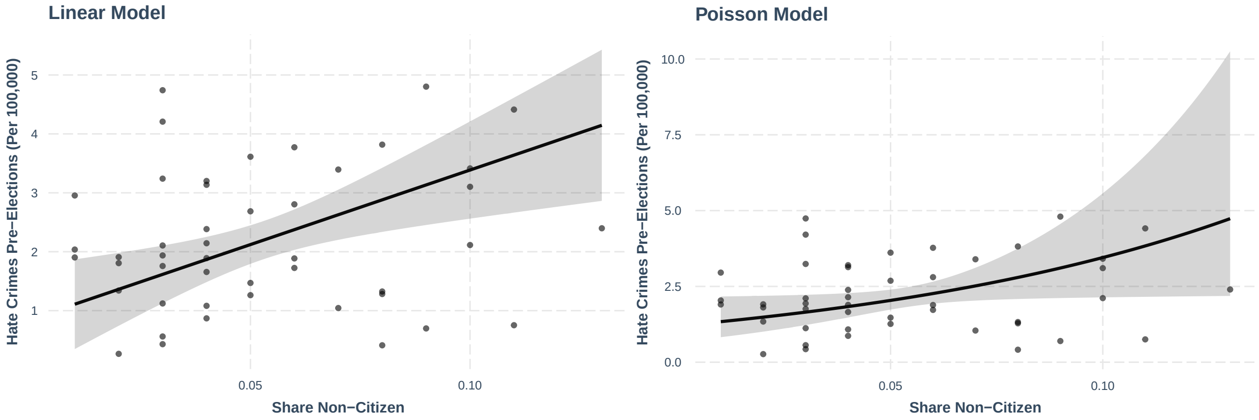

Figure 11: Effect Plots of Non-Citizens (Both Models)

Appendices

Appendix A: FiveThirtyEight's Linear Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-50.04	14.81	-3.378	0.001698
median_household_income	4.061e-05	4.929e-05	0.8238	0.4152
share_unemployed_seasonal	32.51	27.22	1.194	0.2397
share_population_in_metro_areas	-2.452	1.916	-1.28	0.2084
share_population_with_high_school_degree	24.45	13.12	1.864	0.07005
share_non_citizen	21.55	12.95	1.664	0.1044
share_white_poverty	7.114	17.23	0.413	0.682
gini_index	62.6	13.51	4.634	4.138e-05
share_non_white	-3.454	2.948	-1.172	0.2486

Table 6: Fitting linear model: avg_hatecrimes_per_100k_fbi ~ .

Observations	Residual Std. Error	R^2	Adjusted R^2
47	1.367	0.4926	0.3858

Appendix B: Linear Regression Model with No Outlier

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-20.36	13.66	-1.49	0.1448
median_household_income	3.545e-05	3.998e-05	0.8866	0.381
share_unemployed_seasonal	15.98	22.37	0.7144	0.4795
share_population_in_metro_areas	-0.3836	1.619	-0.237	0.814
share_population_with_high_school_degree	10.79	11.05	0.9767	0.335
share_non_citizen	22.19	10.5	2.112	0.04146
share_white_poverty	15.77	14.1	1.119	0.2704
gini_index	20.95	14.26	1.469	0.1504
share_non_white	-4.575	2.403	-1.904	0.06473

Table 8: Fitting linear model: avg_hatecrimes_per_100k_fbi ~ .

Observations	Residual Std. Error	R^2	Adjusted R^2
46	1.109	0.2991	0.1475

Appendix C: Best Linear Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.503	0.382	6.553	5.74e-08
share_non_citizen	25.3	7.975	3.173	0.002787
share_non_white	-5.426	1.692	-3.206	0.002539

Table 10: Fitting linear model: avg_hatecrimes_per_100k_fbi ~ share_non_citizen + share_non_white

Observations	Residual Std. Error	R^2	Adjusted R^2
46	1.089	0.2136	0.177

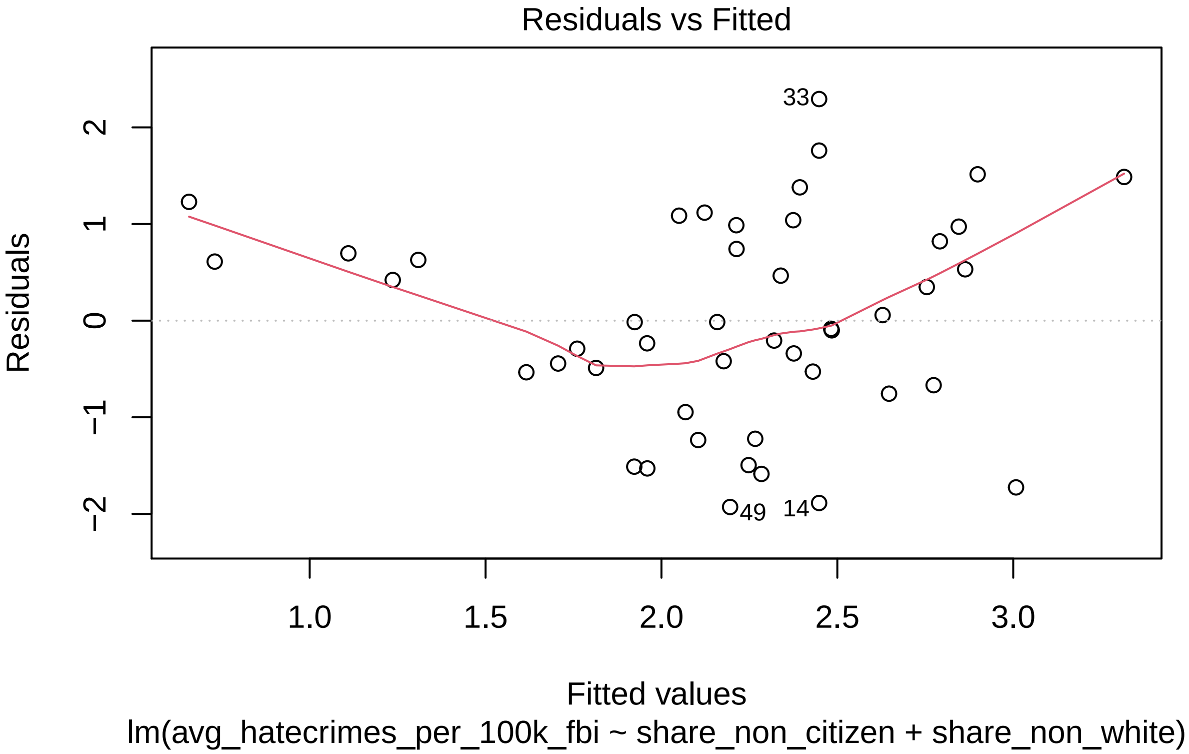

Figure 12: Residuals vs Fitted Plot for The Best Linear Regression Model

Appendix D: Quasi-Poisson Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.773	5.877	-1.323	0.194
median_household_income	1.532e-05	1.79e-05	0.8559	0.3976

	Estimate	Std. Error	t value	Pr(>|t|)
share_unemployed_seasonal	6.096	10.14	0.6009	0.5516
share_population_in_metro_areas	-0.1333	0.7452	-0.1789	0.859
share_population_with_high_school_degree	4.139	4.826	0.8576	0.3966
share_non_citizen	10.53	5	2.107	0.04199
share_white_poverty	6.809	6.264	1.087	0.2841
gini_index	7.453	5.892	1.265	0.2138
share_non_white	-2.25	1.135	-1.982	0.05494

(Dispersion parameter for quasipoisson family taken to be 0.5371281)

Null deviance:	30.75 on 45 degrees of freedom
Residual deviance:	22.01 on 37 degrees of freedom