Investigating the relationship between socio-economic factors and hate incidents in the US

Mohsen Rahimi - 1073078

28/06/2021

```
## [1] "/Users/mohsenrahimi/Downloads/Data Management for Communication"
```

Introduction

Drawing from a FiveThirtyEight article from Maimuma Majumder, published on the 23rd of January 2017, we look for trends in our data to understand how hate crimes vary among States and what factors might be the strongest predictors for hate crimes.

Majunder's analysis used multivariate linear regression to understand which variables were significant determinants of population-adjusted hate incidents across the country. They used various socio-economic indicators to assess the independent impact of each one on hate crimes. They found that income inequality is the strongest predictor, and percentage of adults with a high school diploma was also significant (Majumder, 2017).

They explain this relationship with the fact that anger is generated from seeing your personal situation compared to that of others: when income inequality is high, it is easier to feel like others are doing better than you and this feeds a sentiment of hate towards those who are 'doing better'.

Hate might also depend on the level of education. Income inequality is tied to the fact that high-school-educated individuals are not able to earn as much as their college-educated neighbors. Unemployment, which is tied to level of education and contributes to income inequalities, could also be a determinant of hate crimes, as people who are unemployed might blame others for their situation - we often hear about immigrants being accused of 'stealing jobs'.

Hate crimes might also be higher where there is a high level of Trump supporter, due to his populist movement that fed anger towards ethnic minorities in the US. Generally, inequalities are exacerbated in metropolitan areas or where there is a high level or urbanization. Moreover, immigrants and ethnic minorities tend to live in the poorer districts of large metropolitan areas. Therefore we expect that when the share of the population living in metropolitan areas and the level of urbanization increase, hate crimes also increase.

Dataset description

The dataset includes 9 variables:

- 5 categorical:

  1. *state*: the name of the US State which the observation belongs to, 51 States in total;
  2. *median_house_inc*: median household income (categorized into low-high);
  3. *trump_support*: percent of the population who voted for Donald Trump (categorized into low-medium-high);
  4. *unemployment*: seasonally adjusted unemployment (categorized into low-high);
  5. *urbanization*: level of urbanization (categorized into low-high).

- 4 numeric:

1. *share_pop_metro*: percent population in metropolitan areas;
2. *hs*: percent of adults 25 and older with at least a high school degree;
3. *hate_crimes*: average annual hate crimes per 100,000 residents;
4. *income*: median household income.

Variables identification

Based on the previous considerations, we identify **hate_crimes** as the response variable and the following as potential explanatory variables:

- *income* (we include income and not median_house_income as the latter is based on a categorization of the first one);
- *share_pop_metro* (which we also use as a proxy for level of urbanization);
- *hs* ;
- *trump_support* ;
- *unemployment* ;

Exploratory data analysis

Summary statistics

Table 1 (sub-table variable type: factor) shows that the levels for *trump_support* and *unemployment* are evenly distributed across States, with no level being remarkably greater than the others. We can say we are dealing with a 'balanced' dataset.

The mean for the number of hate crimes is 0.304, with a remarkably lower median of 0.226. The mean family income is 55,224, with a slightly lower median of 54,916. The mean for the percentage of population with at least a high school diploma is 86.8%, with median at 87%. The mean percentage for the share of the population living in metropolitan areas is 75%, with slightly higher median at 79%.

Table 1 shows that there are missing values in the dataset for the variables *hs* and *hate_crimes*. We apply the imputation method that substitutes NAs with the median for that variable (alternatively we could have chosen to use the mean). We are aware that this might cause our results to be slightly biased, however we prefer this method rather than deletion of observations in order to prevent data loss in a dataset which is already small.

Table 1: Data summary

Name	Piped data
Number of rows	51
Number of columns	6
Column type frequency:	
factor	2
numeric	4
Group variables	None

Variable type: factor

skim_variable	n_missing	n_unique	top_counts
trump_support	0	3	low: 19, hig: 17, med: 15
unemployment	0	2	low: 27, hig: 24

Variable type: numeric

skim_variable	n_missing	mean	sd	p0	p25	p50	p75	p100
hate_crimes	4	0.30	0.25	0.07	0.14	0.23	0.36	1.52
income	0	55223.61	9208.48	35521.00	48657.00	54916.00	60719.00	76165.00
hs	3	86.79	3.44	80.00	84.00	87.00	90.00	92.00
share_pop_metro	0	75.02	18.16	31.00	63.00	79.00	89.50	100.00

Data visualization

Figure 1 shows that States with higher crimes are located in the North, such as Oregon (OR), Washington (WA), Minnesota (MN), Maine (ME) and Massachusetts (MA) - and if you zoom in you can see a dark red spot on the right side that corresponds to District of Columbia (DC). States with a lower number of crimes are spread evenly across the country. Some examples are Idaho (ID), Alaska (AK), Arkansas (AR) and New Jersey (NJ).

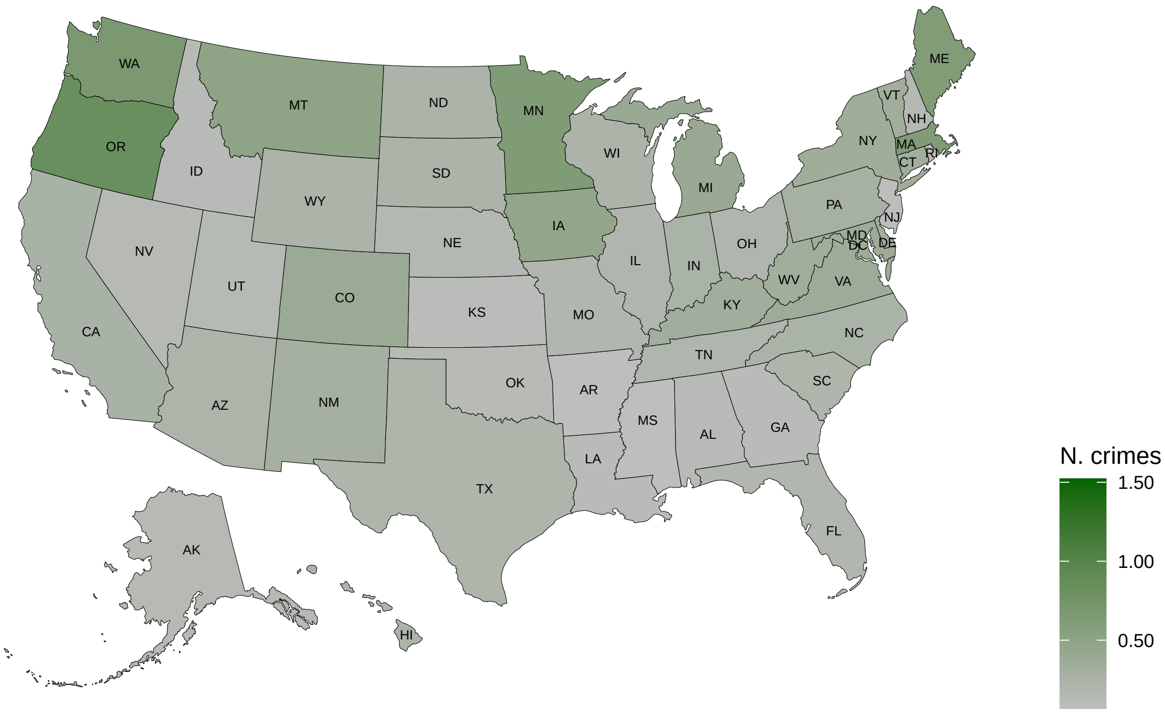


Figure 1: Average annual hate crimes per 100,000 residents by US State

Figure 2 shows that there are 3 values in the number of hate crimes that are particularly higher compared to the others. Table 4 shows that the 3 outliers correspond to Oregon, Washington and District of Columbia (DC), with the value for DC being remarkably higher than the others. We also note that these observations have the same levels for median house income, Trump support, unemployment and urbanization.

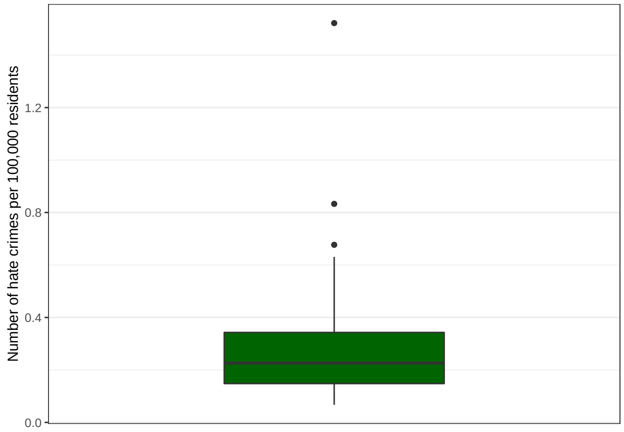


Figure 2: Distribution of number of hate crimes

Table 4: Observations classified as outliers based on number of hate crimes

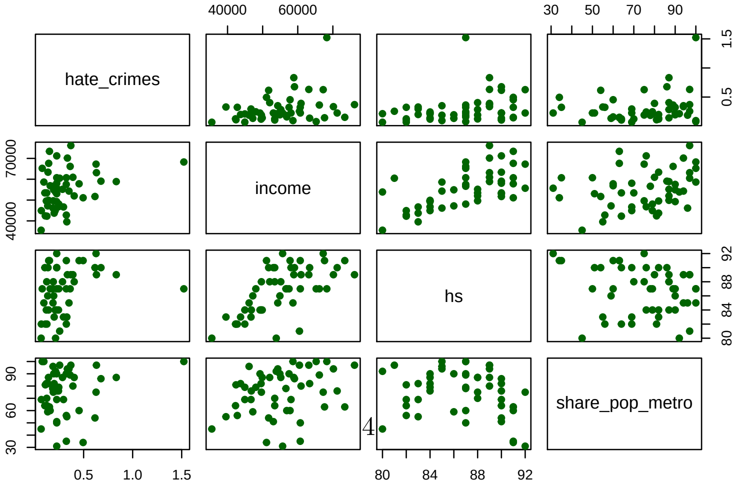
state	hate_crimes	median_house_inc	unemployment	trump_support	urbanization
Oregon	0.833	high	high	low	high
Washington	0.677	high	high	low	high
District of Columbia	1.522	high	high	low	high

Table 5 shows a positive correlation between the number of hate crimes and the other three numerical variables. This means that, as (i)median household income, (ii)the percentage of population with a high school diploma and (iii)the percentage of population living in metropolitan areas increase (independently from one another), so will the number of hate crimes. Specifically, the positive relationship is the highest for *income*, followed by *hs* and then *share_pop_metro*.

Table 5: Correlation between hate crimes and chosen numerical explanatory variables.

income	hs	pop_metro
0.3228828	0.2938603	0.1770854

Figure 3 shows the covariance between hate crimes and the numerical explanatory variables of our interest, as well as the pairwise covariance between all the numerical variables. We can note a slight positive covariance between the response variable and the explanatory variables. However, the relationship seems to be weak. There is no remarkable trend between the explanatory variables (pairwise), except for the positive covariance between *income* and *hs*. This makes sense as usually more educated people can get higher-paid jobs.



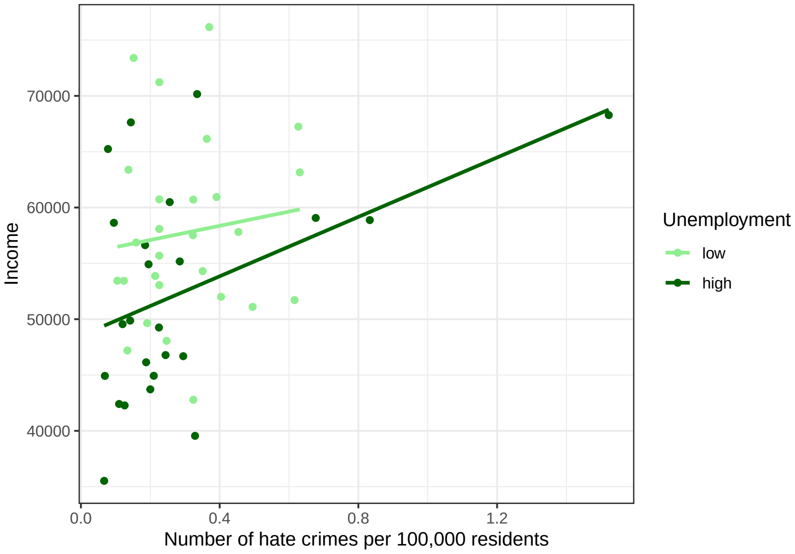


Figure 4: Hate crimes related to income by level of unemployment

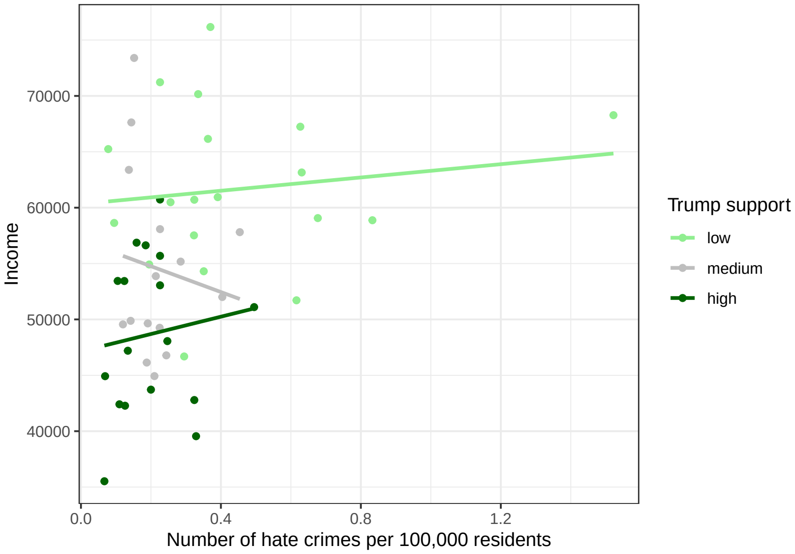


Figure 5: Hate crimes related to income by level of Trump support

Regression Model

After running an in-depth analysis of variance (ANOVA) to choose the regressors that better represent the relationship found in our dataset (see muted code chunk in Rmarkdown script), we chose the basic regression with *income* as the unique explanatory variable. We found that adding other regressors was not statistically significant to explain the observed variation.

The only exception was Trump support. We found that inserting level of Trump support (without interaction) to the basic model with income is statistically significant. However, only the coefficient for the 'low' level is statistically significant at alpha=0.05 when taking 'high' as baseline, while there is no significant difference between 'high' and 'medium'. Moreover, the coefficient for income in this case is not significant at alpha=0.05. Besides the statistical considerations made, we decide to reject this model mainly because of the theoretical setting of our problem. The model suggests that, ceteris paribus, hate crimes are higher when the level of Trump support is low. We would expect the opposite relationship as populist movements might feed more hate towards ethnic minorities, therefore increasing the number of hate crimes. It might be plausible that in States where the level of Trump support is higher there are also fewer hate crimes reported, possibly due to the fact that the police itself is also making discrimination towards minorities. The pattern found in this model clearly needs further investigation in the theoretical setting, therefore we proceed by considering the basic regression as the best model.

Residual analysis

After selecting the model that best explains the relationship found in our dataset, we run a residual analysis to check the robustness of our findings.

1. *Linearity*

Figure 6 shows that overall there is a linear relationship between *income* and *hate_crimes*, however there is one remarkable outlier which corresponds to the observation for DC and for which the linearity of the relationship might not hold. Therefore this assumption is not fully satisfied if the outlier is not removed.

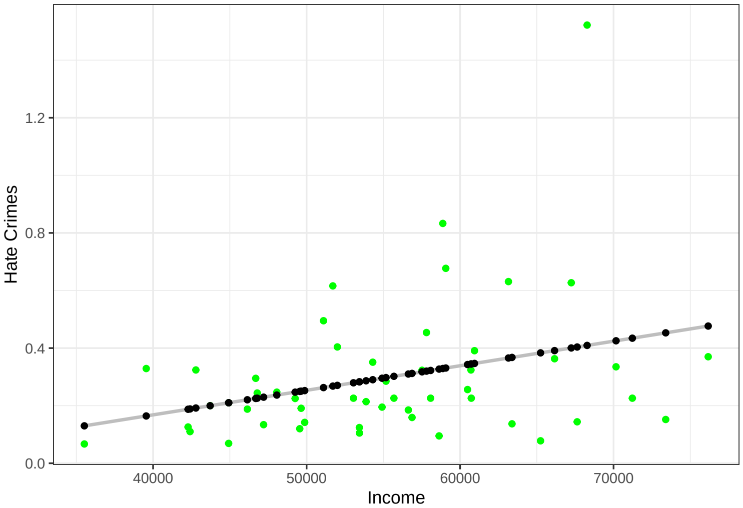


Figure 6: Relationship between Income and Hate Crimes with fitted points in black

2. *Independence*

Our observations are all independent from one another as measurements were taken independently for each state.

3. *Normality*

Figure 7 shows that the residuals are not normally distributed. Again, the outlier for DC seems to play an important role. The results from the Shapiro-Wilk test for normality (Table 6) show that indeed our model residuals are not normally distributed (p-value lower than 0.05). The normality assumption is not satisfied.

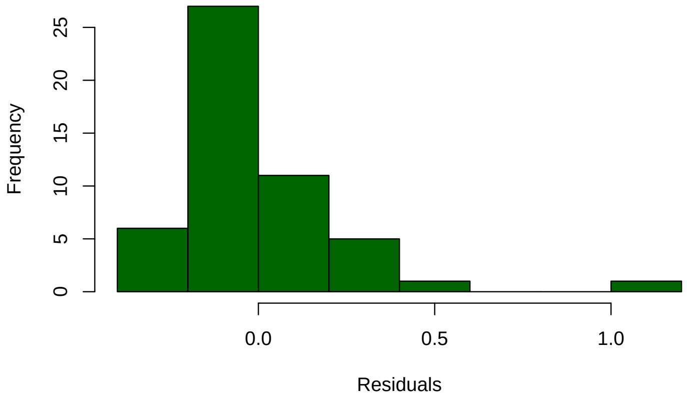


Figure 7: Distribution of residuals

Table 6: Results from Shapiro-Wilk normality test

statistic	value
W	0.7871820
p-value	0.0000004

4. *Equality of variance*

From Figure 8 we see that there is heteroskedasticity in the distribution of our residuals. Indeed, the magnitude of the residuals seems to increase as income increases.

Our findings suggest that a transformation of our variables might be better suited to plot the relationship between income and the number of hate crimes. Specifically, we apply a logarithmic transformation to account for the steep increase in the value of hate crime observed for DC. We then proceed to the residual analysis to check the robustness of our second model, where logarithm of hate crime is our response variable and income is our unique explanatory variable.

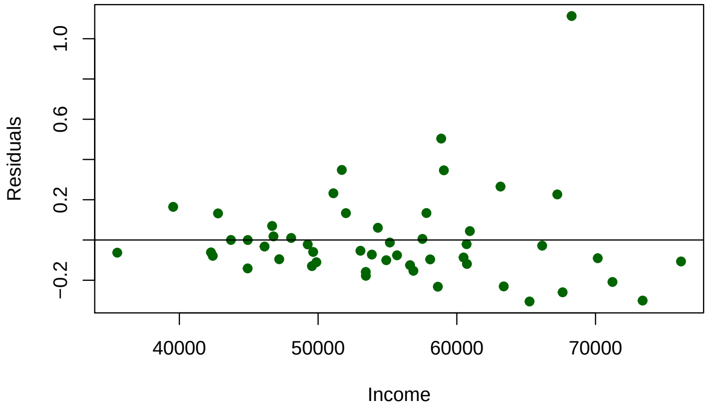


Figure 8: Variation of residuals according to income

Residual analysis for basic regression with log-transformed hate crime

1. *Linearity*

Figure 9 shows that there is an increasing linear relationship between Income and Log hate crimes, confirming that a linear regression can be appropriate to model this relationship.

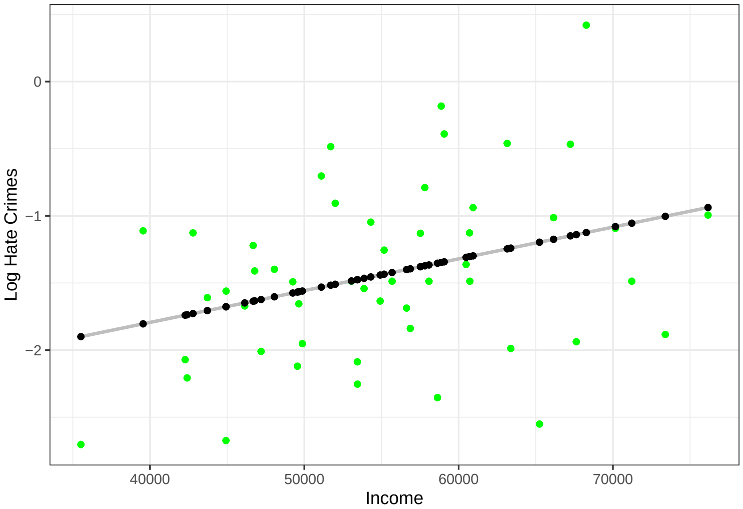


Figure 9: Relationship between log-tranformed hate crimes and income, with estimated regression points in black

2. *Independence*

The independence assumption is satisfied because each Sate represents an independent observation.

3. *Normality*

The histogram in Figure 10 shows that residuals are more normally distributed compared to the previous model. Results from the Shapiro-Wilk test reported in Table 7 confirm that residuals are normally distributed (as p>0.05).

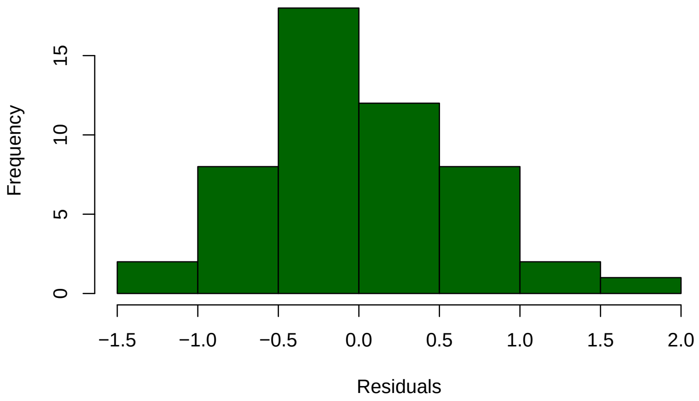


Figure 10: Distribution of residuals

Table 7: Results from Shapiro-Wilk normality test

statistic	value
W	0.9921977
p-value	0.9824532

4. *Equality of variance*

Figure 11 shows homoskedasticity in the distribution of residuals, as the number of positive residuals is similar to that of negative residuals and they are randomly distributed around the zero value.

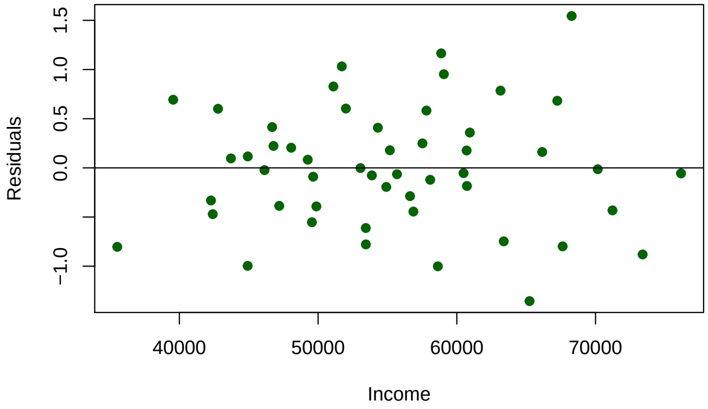


Figure 11: Variation of residuals according to income

As all four assumptions are satisfied, we can state that the basic regression model where the response variable has been log-transformed is the model that best represents the relationship found in our dataset.

Comment on estimated parameters

As shown in Table 8, our final model estimates a 2.4% increase in the average annual number of hate crimes per 100,000 residents for every $1,000 increase in the median household income. 95% confidence intervals for the estimated parameter are also reported in Table 8. We expect 95% of the reported interval to contain the true value of our estimated parameter. This means that, for every increase in $1,000 in income, the real increase in number of hate crimes could assume any value in between 0.5% and 4.2%. We can state that the change in income has an effect on the change in hate crimes as our CI does not contain the value 0.

A final important remark is that, although the parameters are significant at the 95% confidence level (p-value<0.05), the value for the adjusted R-squared is equal to 0.09. Hence the fit of our log-linear model is very poor. This means that income on its own might not be the best variable to explain the variation in hate crimes.

Table 8: Model summary

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-2.740423	0.531158	-5.159333	0.000004	-3.807826	-1.673020
income	0.000024	0.000009	2.493577	0.016071	0.000005	0.000043

Conclusion

The positive relationship we found between income and hate crime should be carefully interpreted. In fact our analysis is full of limitations. First of all, we highlight that the relationship we found does not imply causation, therefore the increase in income might not be itself the cause of the increase in hate incidents. Given the low fit of our model, other economic drivers that were excluded from our analysis should be identified to explain the variation in hate crimes across US States. More information on the data collection process should also be gathered to control for disturbance and allow for causal inferences (although synthetic experiments are usually needed).

Moreover, our results might be biased because of the imputation method used to substitute missing variables. Ideally, the analysis should be run again on a complete dataset. It is also possible that the dataset is not truly representative of hate incidents across the United States, as whether people report or do not report hate incidents might vary across States. This means that States with residents and law enforcement agencies that are more likely to report hate crimes might be overrepresented, and States which do not report as much will be underrepresented.

All things considered, the relationship we found is significant but not exhaustive and the problem should be further investigated considering the limitations that we reported.

Reference

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