Biostatistical Methods I
Final Project – Group #22
Daniel Ojeranti, David Nemirovsky, Ford Holland, Jared Klug, Justin Vargas
12/18/20

Abstract

In the United States (U.S.), hate crimes are defined as actions of violence that are caused by bias and discrimination with the purpose of inflicting pain or terrorizing someone based on social characteristics, such as sexual orientation, national origin, race, ethnicity, or disability (*Hate Crimes Bulletin*, 2001). The Southern Poverty Law Center (SPLC) and other organizations collected information on the number of hate crimes per population of 100,000 individuals, unemployment, urbanization, income, education level, ethnicity, race, and income inequality in every state in the U.S. in 2009, 2015, or 2016 (*Fivethirtyeight/Data*, 2020). The goal of this study is to assess the variables that are mostly associated with the hate crimes rate throughout the U.S. in 2016. According to a FiveThirtyEight model which reported that income inequality was the most significant predictor of population-adjusted hate crimes. An alternative model has been developed to validate their model. Model diagnostics have shown that FiveThirtyEight's model does not hold the assumption of homoscedasticity. Analysis of model diagnostics determined that a natural-log transformed model best fits the data, according to the Box-Cox transformation. Transforming the dataset rectifies heteroscedasticity, as well as changes variable significance. Variables were selected using backwards elimination and the resulting predictors chosen were unemployment, income inequality index, and the percent populate with a highschool degree. Model diagnostic plots highlighted the presence of influential points associated with DC and Oregon. After adjusting for influential points, our final model concludes that it is not only income inequality that is mostly associated with hate crimes, but also the percentage of adults with a highschool degree along with unemployment.

Introduction

Hate crimes are defined by the U.S. Department of Justice as violent acts of narrow-mindedness and prejudice that are performed with the intention of hurting and frightening an individual due to particular social factors, such as disability, sexual orientation, national origin, ethnicity, or race (*Hate Crimes Bulletin*, 2001). In 1990, Congress passed the Hate Crime Statistics Act that required the collection of information about hate crimes that took place throughout the U.S. (*FBI, 2001*). In 2016, the Uniform Crime Reporting (UCR) Program, which is a part of the FBI that provides dependable information for law enforcement management, reported that 6,121 hate crimes occurred throughout the U.S., which exhibited the highest number in the previous five years, from 15,254 law enforcement agencies, which voluntarily submitted the data (*Incidents and Offenses*, 2016)(*FBI*, 2017)(*Uniform Crime Reporting (UCR) Program*, n.d.). More than half of these hate crimes were prompted by bias related to race and ethnicity (*Incidents and Offenses*, 2016). A correlational study conducted throughout the state of New Jersey found that ethnic diversity was not a significant predictor of hate crimes (Ciobanu, 2019). The authors also found that residential mobility and population density had a positive association with hate crimes, while unemployment rates were negatively correlated with hate crimes (Ciobanu, 2019).

The SPLC, which is a non-profit organization aimed at the guarantee of civil rights for all, collected data on the number of hate crimes per population of 100,000 people in every state of the U.S. in 2016 (*About Us*, n.d.)(*Fivethirtyeight/Data*, 2020). Other organizations, such as the Kaiser Family Foundation and U.S. Census Bureau, had collected information on covariates related to income, income inequality, education, unemployment, race, ethnicity, and urbanization in each of the states of the U.S. in 2009, 2015, or 2016 (*Fivethirtyeight/Data*, 2020). Through the use of this data, the purpose of this study is to identify the variables that are associated with the hate crime rate per population of 100,000 individuals throughout the U.S. in 2016.

Methods

To assess the associations between socioeconomic factors and hate crimes, we used data collected from the SPLC, Kaiser Family Foundation, and U.S. Census Bureau that were aggregated and made publicly available by FiveThirtyEight (*Fivethirtyeight/Data*, 2020). The data set contains 51 records, representing observations on the number of hate crimes committed per 100,000 people between November 9th and 18th, 2016 and several socioeconomic indicators for all 50 states and the District of Columbia (*Fivethirtyeight/Data*, 2020). Four states, which were Hawaii, North Dakota, South Dakota, and Wyoming, did not report hate crimes in 2016. The socioeconomic markers include education as the percent of adults aged 25 and older with at least a high school degree as of 2009, seasonally adjusted unemployment in 2016, Gini index of income inequality in 2015, median household income in 2016, the percent non-white individuals, the percent of the non-citizen population in 2015, and the percent of population living in metropolitan areas in 2015.

We developed a multiple linear regression (MLR) model to measure the association between these factors and the outcome, hate crimes. We first examined the outcome distribution graphically. Using a Box-Cox transformation, we determined that logarithmic transformation of hate crimes data was appropriate based on the log-likelihood curve having a maximum value at a lambda of 0 (Figure 2). We considered both untransformed and log-transformed data in model development. Based on the residuals versus fitted and the quantile-quantile (QQ) plots, it was determined that the log-transformed model consisted of residuals that were more homoscedastic, but slightly less normally distributed than those in the untransformed model (Figures 1 and 3). Overall, the transformed model was chosen to continue with the development of the MLR model.

To determine our model specification, we applied backward elimination to identify variables having significant linear associations with hate crimes. Each model iteration we considered was evaluated for performance and validity using adjusted-R^2, Akaike information criteria (AIC), residual plots, and their respective QQ plots. The model with the highest adjusted-R^2, lowest AIC, most randomly scattered and centered at 0 plot of residuals vs fitted values, and most normalized QQ plot of standardized residuals was chosen. Using residuals vs. leverage plots, we identified Washington D.C. as an influential observation and Oregon was identified as an outlier using studentized residual values. Models created with and without these data were evaluated and it was determined that removing these two extreme observations resulted in normality of the residuals and best model performance, according to diagnostic plots and adjusted R^2. After removing Washington D.C. and Oregon from the training set, a correlation matrix was made to examine possible correlation between covariates in the model. The matrix showed that percent of the population with at least a high school degree is moderately correlated with Gini index (R = -0.66). Testing for multicollinearity, the variance inflation factor (VIF) was found for each covariate to be less than 2, indicating no multicollinearity in the model.

Results

Our final model specification applied a logarithmic transformation to hate crimes per 100,000 population and included the Gini index of income inequality, percent of population with at least high school education, and unemployment as predictors. The final model had an adjusted-R^2 value of 0.15 and an AIC value of 71.73. The plot of residuals vs fitted values for our final model showed randomly scattered residuals centered at the residual value of 0, holding the assumption of homoscedasticity to be true. The normal Q-Q plot showed a normal distribution of standardized residuals. The plot of residuals vs leverage does not show any more influential points in our model. Overall, this model illustrated the covariates most associated with hate crimes, all while having little bias and maintaining the assumptions required for building a regression model.

Discussion

In this project, we propose a parsimonious model isolating the most associated variables to predict the rate of hate crimes per population of 100,000 individuals in the U.S. by state. Our model includes 3 covariates, which are unemployment rate, percentage of adults that are aged 25 years or older with a high school degree, and the index measuring income inequality. According to the article published by FiveThirtyEight, they claim that income inequality was the main predictor of hate crimes rates (Majumder, 2017). However, according to our model, the percentage of adults with a high school degree was the most significant variable.

The initial difference between our model and that of FiveThirtyEight is that their model was not transformed. In our assessment of a non-transformed model controlled for all other covariates, similar to the model FiveThirtyEight employed, our results agreed with theirs in that income inequality was the most significant determinant of population-adjusted hate crimes followed by percent population with a high school degree. However, when looking at the Residuals versus Fitted Values plot and Scale-Location plot, we observe the assumption of homoscedasticity is not met in this model (Figure 1).

This led us to our natural log transformed model (Figure 3). The transformed model becomes more homoscedastic. When controlled for all other variables, the most significant predictors remain the same as the non-transformed model. After further analysis, the model was reduced to the two significant variables: Gini index and percentage of the population with a high school degree. The removal of unemployment, the third most significant variable, only slightly decreased the AIC of the model. Both models were interpreted and concluded adding unemployment back to our model allowed us to have better normality of the residuals for our transformed model (Figure 4). Observing the plot of residuals vs leverage, we identified a problematic point associated with the data related to the District of Columbia. According to the DC Fiscal Institute, DC had the highest income inequality in 2016 according to the Gini Index (DC Fiscal Policy Institute, 2017). Also, both poverty and income inequality in the District differed greatly along racial lines (DC Fiscal Policy Institute, 2017). This raises a major point of concern for the model as it is likely influencing the significance of the Gini index predictor. We fit the transformed model excluding the District of Columbia. As expected, the Gini index was no longer a significant predictor for both our transformed and untransformed models. Removing this observation also altered our model diagnostics (Figure 5). Oregon, having a studentized residual value of 2.3, still stood out as an outlier after removing DC, so it was also removed and the new fitted model retained homoscedasticity (Figure 6). The significance of interaction in our final model by unemployment and urbanization groups was checked by fitting models with the respective interaction terms and no significant interactions between subgroups for both variables were observed (Figures 7 & 8). We cross validated the data and tested our models to observe RMSE distributions for the transformed model versus the non-transformed model. It was concluded that the untransformed model did a better job in predicting the outcome in comparison to our final, log-transformed model.

Ultimately, our final model identified the Gini index of income inequality, percent of population with at least a high school education, and unemployment as predictors that are the most associated with the hate crime rate per population of 100,000 people throughout the U.S. in 2016.

References

2001.pdf. (2001). FBI. Retrieved December 12, 2020, from https://ucr.fbi.gov/hate-crime/2001

About Us. (n.d.). Southern Poverty Law Center. Retrieved December 13, 2020, from https://www.splcenter.org/about

Ciobanu, D. M. (2019). Social Disorganization Theory: The Role of Diversity in New Jersey's Hate Crimes Based on Race and Ethnicity. Journal of Social, Behavioral, and Health Sciences, 13(1). https://doi.org/10.5590/JSBHS.2019.13.1.02

FBI: Hate crimes reach 5-year high in 2016, jumped as Trump rolled toward presidency. (2017). Southern Poverty Law Center. Retrieved December 12, 2020, from https://www.splcenter.org/hatewatch/2017/11/13/fbi-hate-crimes-reach-5-year-high-2016-jumped-trump-rolled-toward-presidency-0

Fivethirtyeight/data. (2020). GitHub. Retrieved December 13, 2020, from https://github.com/fivethirtyeight/data

Hate Crimes Bulletin. (2001). Retrieved December 12, 2020, from https://www.justice.gov/archive/crs/pubs/crs_pub_hate_crime_bulletin_1201.htm

Incidents and Offenses. (2016). FBI. Retrieved December 12, 2020, from https://ucr.fbi.gov/hate-crime/2016/topic-pages/incidentsandoffenses

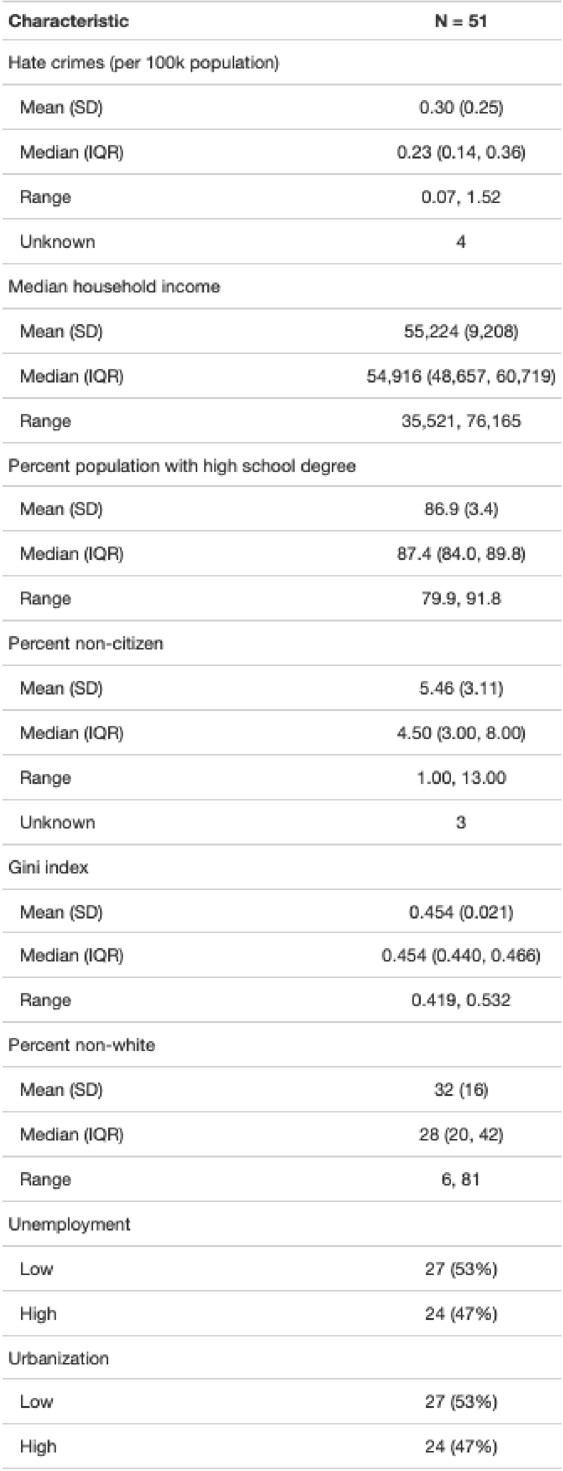
Income Inequality in DC Highest in the Country. (2017, December 15). *DC Fiscal Policy Institute*. https://www.dcfpi.org/all/income-inequality-dc-highest-country/

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

Uniform Crime Reporting (UCR) Program. (n.d.). [Folder]. Federal Bureau of Investigation. Retrieved December 12, 2020, from https://www.fbi.gov/services/cjis/ucr

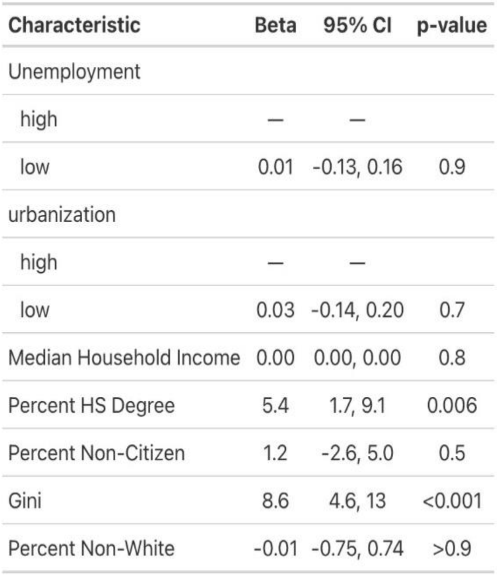
Supplemental Figures

Table 1: Descriptive statistics for hate crime and socioeconomic data.



Characteristic	**N = 51**
Hate crimes (per 100k population)	
Mean (SD)	0.30 (0.25)
Median (IQR)	0.23 (0.14, 0.36)
Range	0.07, 1.52
Unknown	4
Median household income	
Mean (SD)	55,224 (9,208)
Median (IQR)	54,916 (48,657, 60,719)
Range	35,521, 76,165
Percent population with high school degree	
Mean (SD)	86.9 (3.4)
Median (IQR)	87.4 (84.0, 89.8)
Range	79.9, 91.8
Percent non-citizen	
Mean (SD)	5.46 (3.11)
Median (IQR)	4.50 (3.00, 8.00)
Range	1.00, 13.00
Unknown	3
Gini index	
Mean (SD)	0.454 (0.021)
Median (IQR)	0.454 (0.440, 0.466)
Range	0.419, 0.532
Percent non-white	
Mean (SD)	32 (16)
Median (IQR)	28 (20, 42)
Range	6, 81
Unemployment	
Low	27 (53%)
High	24 (47%)
Urbanization	
Low	27 (53%)
High	24 (47%)

Figure 1: Regression summary table and residuals with Q-Q plots for the non-transformed model using all covariates.



Characteristic	Beta	95% CI	p-value
Unemployment			
high	—	—	
low	0.01	-0.13, 0.16	0.9
urbanization			
high	—	—	
low	0.03	-0.14, 0.20	0.7
Median Household Income	0.00	0.00, 0.00	0.8
Percent HS Degree	5.4	1.7, 9.1	0.006
Percent Non-Citizen	1.2	-2.6, 5.0	0.5
Gini	8.6	4.6, 13	<0.001
Percent Non-White	-0.01	-0.75, 0.74	>0.9

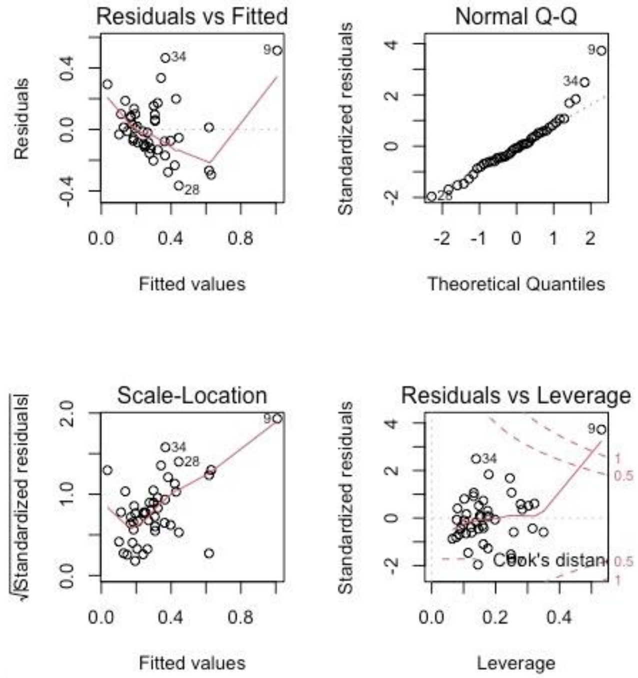


Figure 2: Box-Cox plot of non-transformed model using all covariates.

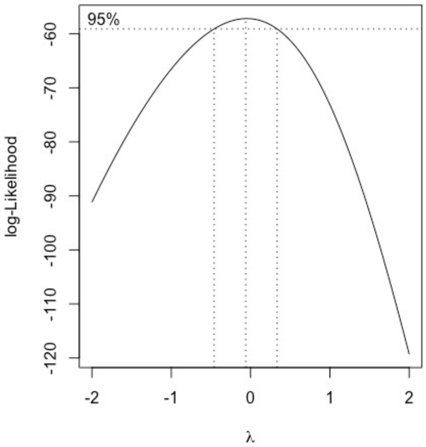
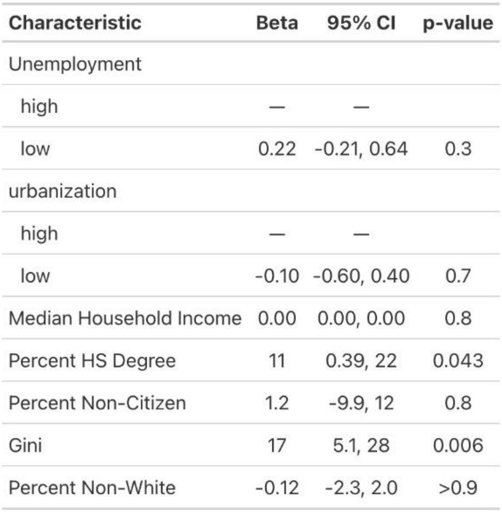


Figure 3: Regression summary table and residuals with Q-Q plots for the log-transformed model using all covariates.



Characteristic	Beta	95% CI	p-value
Unemployment			
high	—	—	
low	0.22	-0.21, 0.64	0.3
urbanization			
high	—	—	
low	-0.10	-0.60, 0.40	0.7
Median Household Income	0.00	0.00, 0.00	0.8
Percent HS Degree	11	0.39, 22	0.043
Percent Non-Citizen	1.2	-9.9, 12	0.8
Gini	17	5.1, 28	0.006
Percent Non-White	-0.12	-2.3, 2.0	>0.9

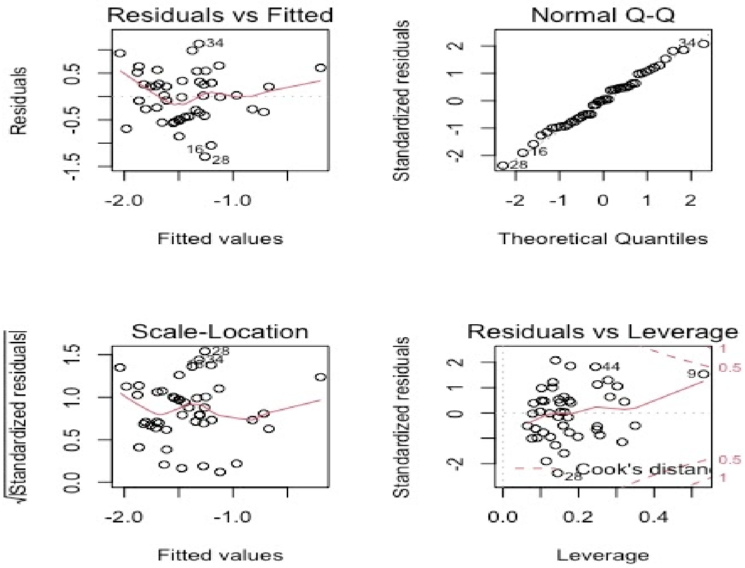
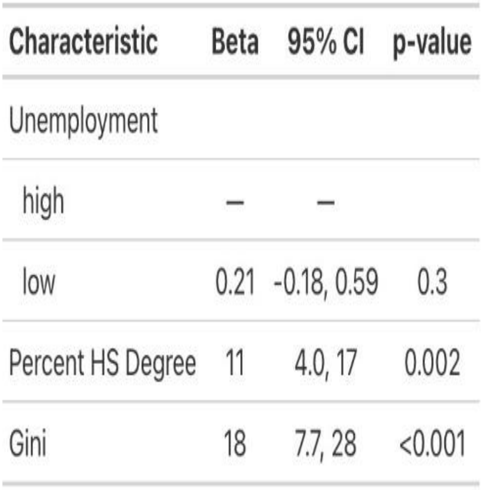
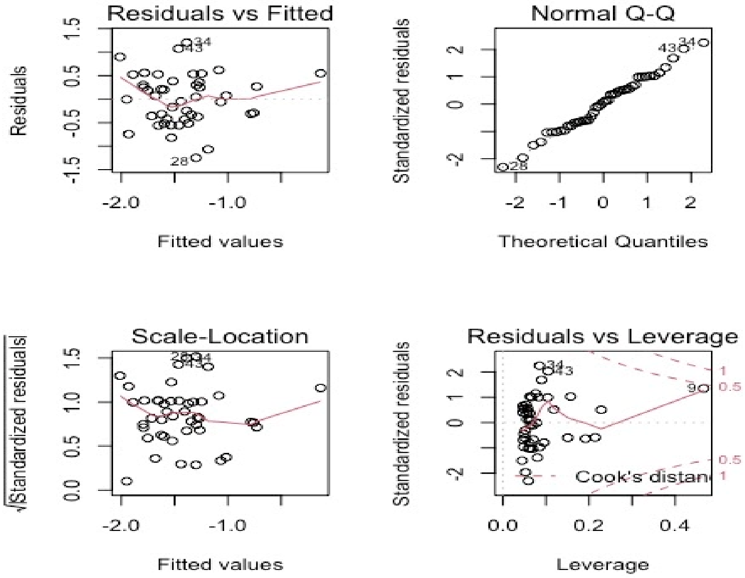
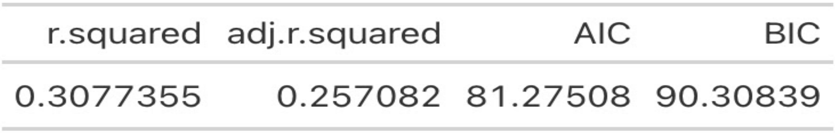


Figure 4: Regression summary table and residuals with Q-Q plots for the reduced log-transformed model using unemployment, percentage of population with HS degree, and Gini index as the three covariates after backward-step variable selection.



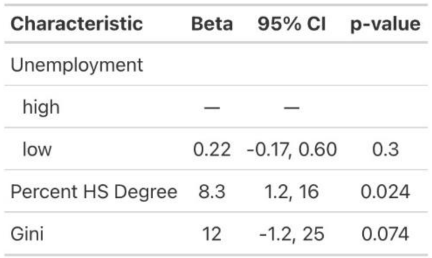
Characteristic	Beta	95% CI	p-value
Unemployment			
high	—	—	
low	0.21	-0.18, 0.59	0.3
Percent HS Degree	11	4.0, 17	0.002
Gini	18	7.7, 28	<0.001



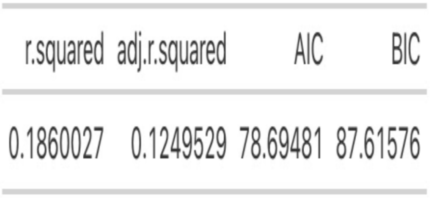


r.squared	adj.r.squared	AIC	BIC
0.3077355	0.257082	81.27508	90.30839

Figure 5: Regression summary table and residuals with Q-Q plots for the reduced log-transformed model using unemployment, percentage of population with HS degree, and Gini index as the three covariates after removing DC from the model.



Characteristic	Beta	95% CI	p-value
Unemployment			
high	—	—	
low	0.22	-0.17, 0.60	0.3
Percent HS Degree	8.3	1.2, 16	0.024
Gini	12	-1.2, 25	0.074



r.squared	adj.r.squared	AIC	BIC
0.1860027	0.1249529	78.69481	87.61576

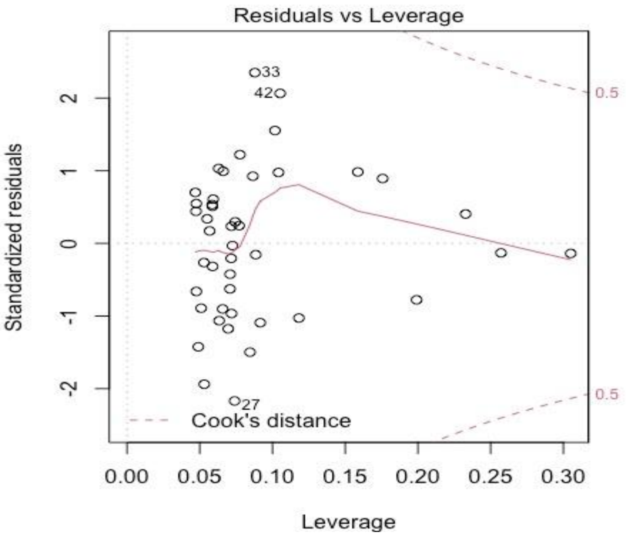
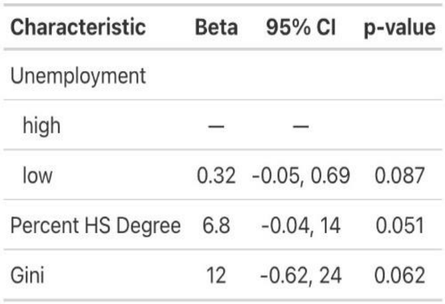
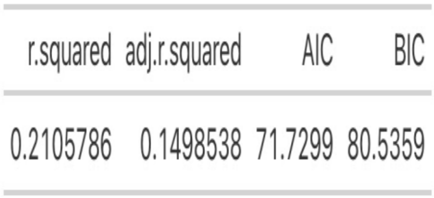


Figure 6: Regression summary table and residuals with Q-Q plots for the reduced log-transformed model using unemployment, percentage of population with HS degree, and Gini index as the three covariates after removing Oregon from the model that already had DC removed.

.



Characteristic	Beta	95% CI	p-value
Unemployment			
high	—	—	
low	0.32	-0.05, 0.69	0.087
Percent HS Degree	6.8	-0.04, 14	0.051
Gini	12	-0.62, 24	0.062



r.squared	adj.r.squared	AIC	BIC
0.2105786	0.1498538	71.7299	80.5359

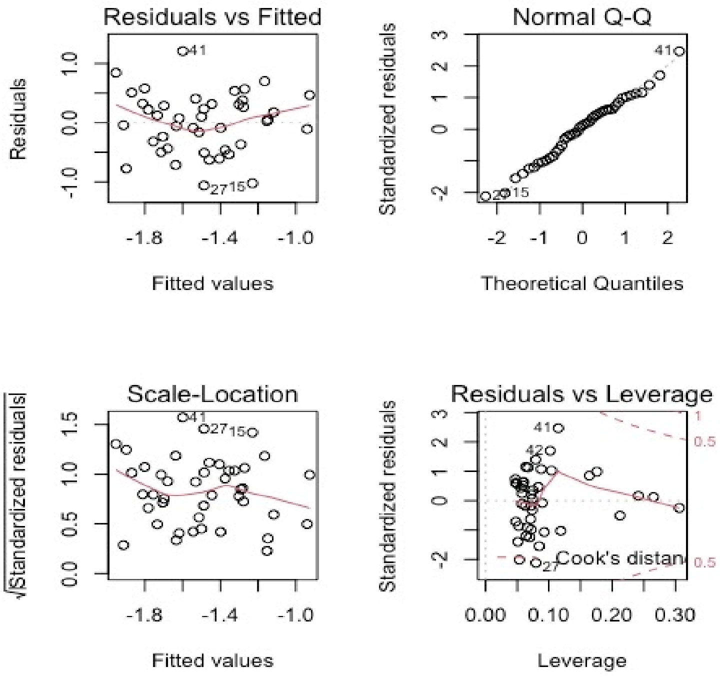
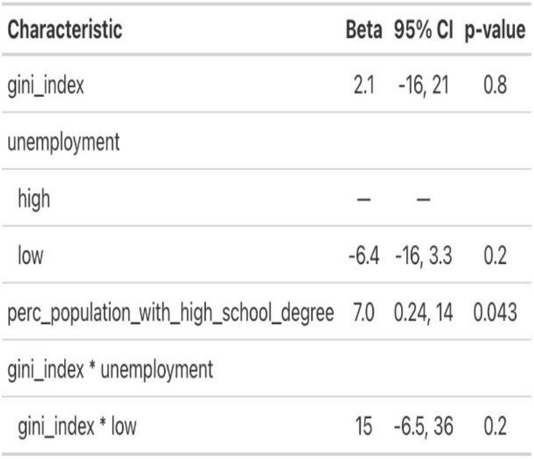


Figure 7: Scatterplot showing possible interaction between Gini index and unemployment with regression summary table of log-transformed model using interaction between Gini index and unemployment, their main terms, and percentage of population with HS degree.



Characteristic	Beta	95% CI	p-value
gini_index	2.1	-16, 21	0.8
unemployment			
high	—	—	
low	-6.4	-16, 3.3	0.2
perc_population_with_high_school_degree	7.0	0.24, 14	0.043
gini_index * unemployment			
gini_index * low	15	-6.5, 36	0.2

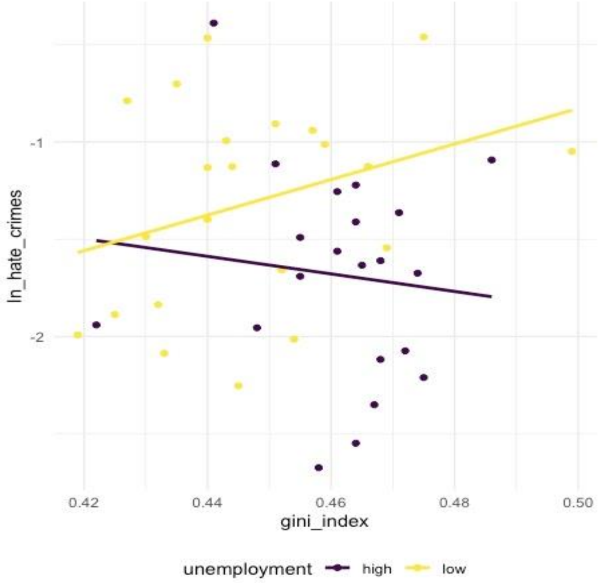
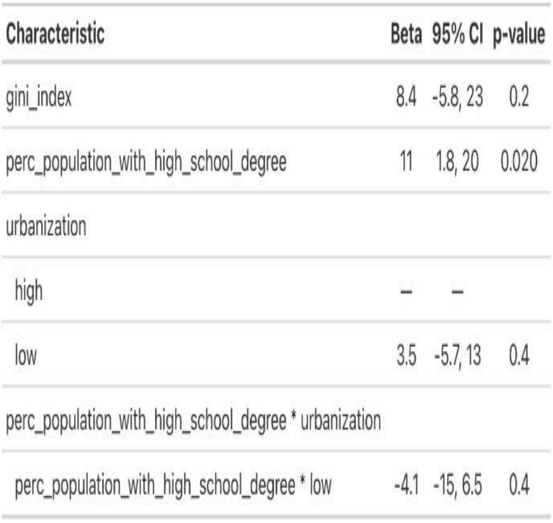
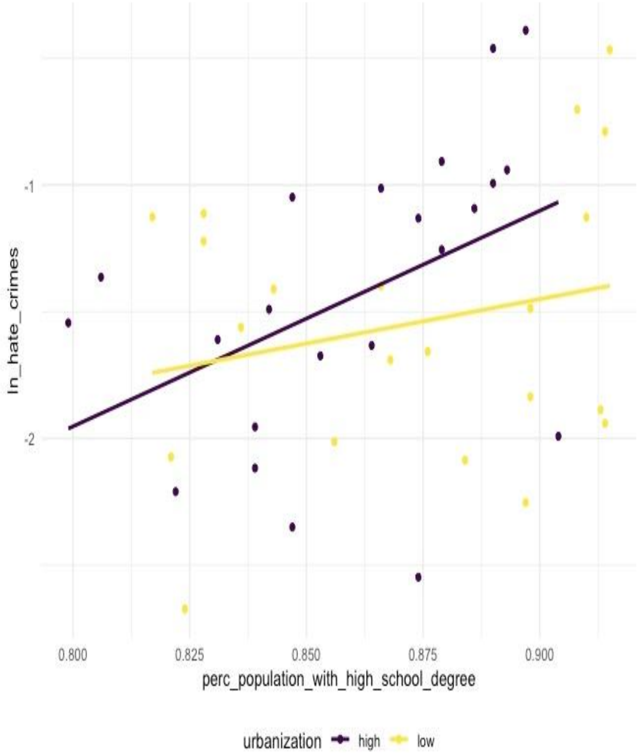


Figure 8: Scatterplot showing possible interaction between percentage population with HS degree and urbanization with regression summary table of log-transformed model using interaction between percentage population with HS degree and urbanization, their main terms, and Gini index.



Characteristic	Beta	95% CI	p-value
gini_index	8.4	-5.8, 23	0.2
perc_population_with_high_school_degree	11	1.8, 20	0.020
urbanization			
high	—	—	
low	3.5	-5.7, 13	0.4
perc_population_with_high_school_degree * urbanization			
perc_population_with_high_school_degree * low	-4.1	-15, 6.5	0.4


ln_hate_crimes
perc_population_with_high_school_degree
urbanization
high
low