STA 4102 Project: A Linear Model to Predict Reported Hate Crimes

Elaine Ng, Roshnaey Khattak, Weijun Huang

Final Project for STA 4102
Dr. Liu
Due: 2 December 2019

1. Description of Data

The purpose of this project was to use the Statistical Analysis System (SAS) software to analyze a set of data regarding hate crimes reported to both the Federal Bureau of Intelligence (FBI) and the Southern Poverty Law Center (SPLC). This dataset was obtained from github.com and was originally collected by news source FiveThirtyEight. The dataset consisted of twelve variables: State (identifying variable), Median Household Income, Unemployment Rate, Share of Population in Metro Areas, Share of Population with High School Degree, Share of Population Non-US Citizen, Level of White Poverty, GINI Index, Share of Non-White Population, Share of Population that Voted for Trump, Hate Crimes per 100k reported to SPLC, Hate Crimes per 100k reported to FBI. The dataset had 51 observations, consisting of the 50 states within the United States and the District of Columbia.

2. Objective

We desired to determine if a linear model could be built to determine if the number of hate crimes reported to the FBI or CPLC could be estimated by the aforementioned variables. Because of poor R-Squared values and difficulty forming a proper model, we focused on estimating the number of hate crimes reported to the Southern Poverty Law Center, as opposed to estimating the number of hate crimes reported to the FBI. We were determined to ascertain the best linear model for our given dataset. To do this, many analytical methods were used to obtain and determine the best fit model for our dataset.

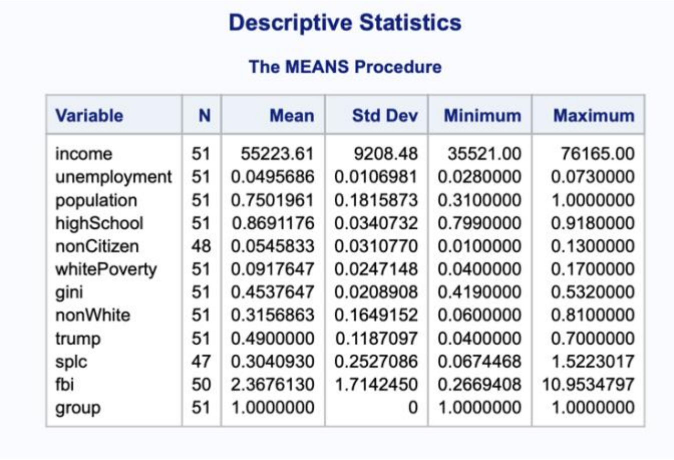
3. Methodology

Boxplot Analysis

Various methods of statistical analysis were used to examine the dataset. First, boxplots of each variable were viewed to preliminarily determine if there were any possible outliers or influential observations within our dataset. For our dataset, five variables had outliers. The boxplot matrix for our dataset can be viewed in the additional graphs section of our paper (last page). The distribution of the number of hate crimes reported to the SPLC per 100k members of a population had two upper outliers. The distribution of the share of white poverty for each given population had one upper outlier. The distribution of the GINI index had one upper outlier. The distribution of the share of non-white members of each given population had one upper outlier. The distribution of the share of the population that voted for Trump in the 2016 presidential election had one lower outlier.

Model Building

Means Table. Following this, a means table was created for every variable in our model. The means table can be viewed in Figure 2. It is important to note that according to our means table, some of the observations had missing values. Specifically, the variables for the share of non-citizens per state, the number of hate crimes reported to the SPLC per 100k members of the state's population, and the number of hate crimes reported to the FBI per 100k members of the state's population had observations of 48, 47, and 50, respectively. These discrepancies were noted, and future code was written to take into account these missing values.



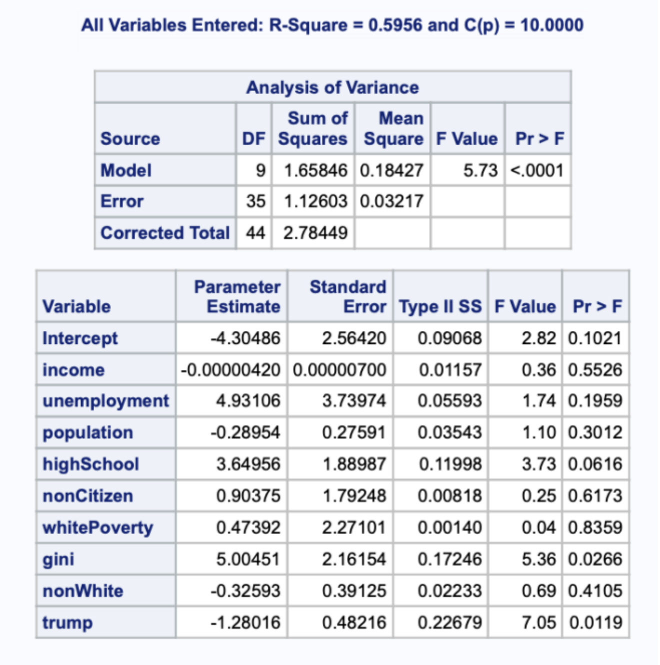
Descriptive Statistics

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
income	51	55223.61	9208.48	35521.00	76165.00
unemployment	51	0.0495686	0.0106981	0.0280000	0.0730000
population	51	0.7501961	0.1815873	0.3100000	1.0000000
highSchool	51	0.8691176	0.0340732	0.7990000	0.9180000
nonCitizen	48	0.0545833	0.0310770	0.0100000	0.1300000
whitePoverty	51	0.0917647	0.0247148	0.0400000	0.1700000
gini	51	0.4537647	0.0208908	0.4190000	0.5320000
nonWhite	51	0.3156863	0.1649152	0.0600000	0.8100000
trump	51	0.4900000	0.1187097	0.0400000	0.7000000
splc	47	0.3040930	0.2527086	0.0674468	1.5223017
fbi	50	2.3676130	1.7142450	0.2669408	10.9534797
group	51	1.0000000	0	1.0000000	1.0000000

Figure 1: Means Table

Backward Elimination Model. After these two steps, we were ready to begin building a linear model to fit our dataset. We began by building a base model with the number of hate crimes reported to the SPLC per 100k members of the population as the dependent variable, and nine independent variables (Median Household Income, Unemployment Rate, Share of Population in Metro Areas, Share of Population with High School Degree, Share of Population Non-US Citizen, Level of White Poverty, GINI Index, Share of Non-White Population, and Share of Population that Voted for Trump). This model had an R-Squared value of 0.5956, meaning that the model explains approximately 59.56 percent of the variance within the dataset. The model also has a C(p) value of 10.000, which is ideal given that the model contains nine variables. While this model is significant with a global F-test (F value of 5.73 and p-value of <0.0001), however, only two of the variables were significant within the model. Because of this, it was determined that more models needed to be created to best determine the best fit for our dataset.



All Variables Entered: R-Square = 0.5956 and C(p) = 10.0000

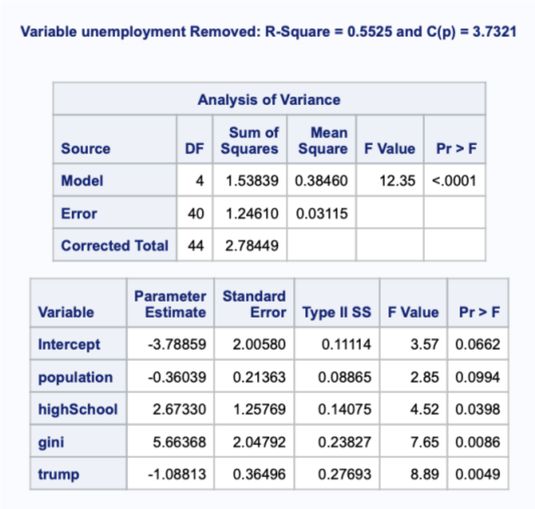
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1.65846	0.18427	5.73	<.0001
Error	35	1.12603	0.03217		
Corrected Total	44	2.78449			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-4.30486	2.56420	0.09068	2.82	0.1021
income	-0.00000420	0.00000700	0.01157	0.36	0.5526
unemployment	4.93106	3.73974	0.05593	1.74	0.1959
population	-0.28954	0.27591	0.03543	1.10	0.3012
highSchool	3.64956	1.88987	0.11998	3.73	0.0616
nonCitizen	0.90375	1.79248	0.00818	0.25	0.6173
whitePoverty	0.47392	2.27101	0.00140	0.04	0.8359
gini	5.00451	2.16154	0.17246	5.36	0.0266
nonWhite	-0.32593	0.39125	0.02233	0.69	0.4105
trump	-1.28016	0.48216	0.22679	7.05	0.0119

Figure 2: Model Containing all Variables

First, the backward elimination method was used to decrease the model. This was done in five steps, starting with the aforementioned model with nine variables, then reducing by removing the variables for the share of noncitizens, median household income, the share of non-white population, and the unemployment rate. The final model obtained through backward elimination contained four variables: share of the population in metro areas, the share of the population that graduated from high school, GINI index, and share of the population that voted for Trump. The model is $y = 3.78859 - 0.36039x_1 + 2.6733x_2 + 5.66368x_3 - 1.08813x_4$ where y is the predicted number of hate crimes per 100k in population reported to the SPLC, x_1 is the share of the population in metropolitan areas, x_2 is the share of the population that graduated from high school, x_3 is the GINI index, and x_4 is the share of the population that voted for Trump in the 2016 election. The model has an R-Squared value of 0.5525, meaning that 55.25 percent of the variance in the dataset can be explained by the model. The model has a C(p) value of 3.7321. There is significant evidence to indicate that the model is statistically significant by the F test, with an F-value of 12.35 and a p-value of <0.0001. Additionally, all variables within this model are significant at the $\alpha = 0.05$ significance level. Ideally, we would want the R-squared of our dataset to be around 0.9, and the C(p) of 5 for this given model. After this, the fit diagnostics for the model were viewed and analyzed. Because of this, more methods must be used to build additional models to determine what the best fit for this dataset is.



Variable unemployment Removed: R-Square = 0.5525 and C(p) = 3.7321

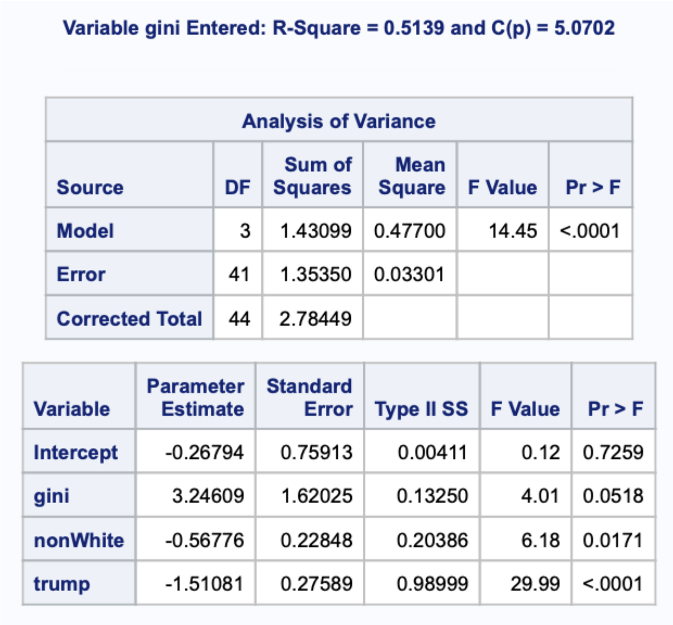
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.53839	0.38460	12.35	<.0001
Error	40	1.24610	0.03115		
Corrected Total	44	2.78449			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-3.78859	2.00580	0.11114	3.57	0.0662
population	-0.36039	0.21363	0.08865	2.85	0.0994
highSchool	2.67330	1.25769	0.14075	4.52	0.0398
gini	5.66368	2.04792	0.23827	7.65	0.0086
trump	-1.08813	0.36496	0.27693	8.89	0.0049

Figure 3: Final Model with the Backwards Elimination Method

Forward Selection and Stepwise Model. Next, the forward selection method was utilized to create a model. With this method, the final model obtained contained the GINI index, the share of the population that is non-white, and the share of the population that voted for Trump. The final model as determined through forward selection is $y = -0.26794 + 3.24609x_1 - 0.56776x_2 - 1.51081x_3$ where y is the predicted number of hate crimes per 100k in population reported to the SPLC, x_1 is the GINI index, x_2 is the share of the population that is non-white, and x_3 is the share of the population that voted for Trump. The model has an R-Squared of 0.5139, meaning that 51.39 percent of the variance within the dataset is explained by the model. The model also had a C(p) value of 5.0702. There is sufficient evidence to indicate that the model is statistically significant as determined by an F test. The model has an F-value of 14.45 and a p-value of <0.0001. All of the variables are statistically significant at the $\alpha = 0.05$ significance level. For this model, we would want a C(p) value of 4 and an R-squared value of around 0.9. It was decided that we should use the stepwise method to build another model, but the model determined was the same as the model determined with the forward selection method.

Subsequently, we had to decide which model building method would create the best model. Both of the models found with the backward elimination and forward selection methods were considered statistically significant with the global F test. Since none of the models had exceptionally high R-squared and similar statistics otherwise, we decided to choose the model building technique that yielded the result with the highest R-squared. The backward elimination method had the highest R-squared, so we have determined that for our dataset that the backward selection method is the best model estimation method.



Variable gini Entered: R-Square = 0.5139 and C(p) = 5.0702

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.43099	0.47700	14.45	<.0001
Error	41	1.35350	0.03301		
Corrected Total	44	2.78449			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.26794	0.75913	0.00411	0.12	0.7259
gini	3.24609	1.62025	0.13250	4.01	0.0518
nonWhite	-0.56776	0.22848	0.20386	6.18	0.0171
trump	-1.51081	0.27589	0.98999	29.99	<.0001

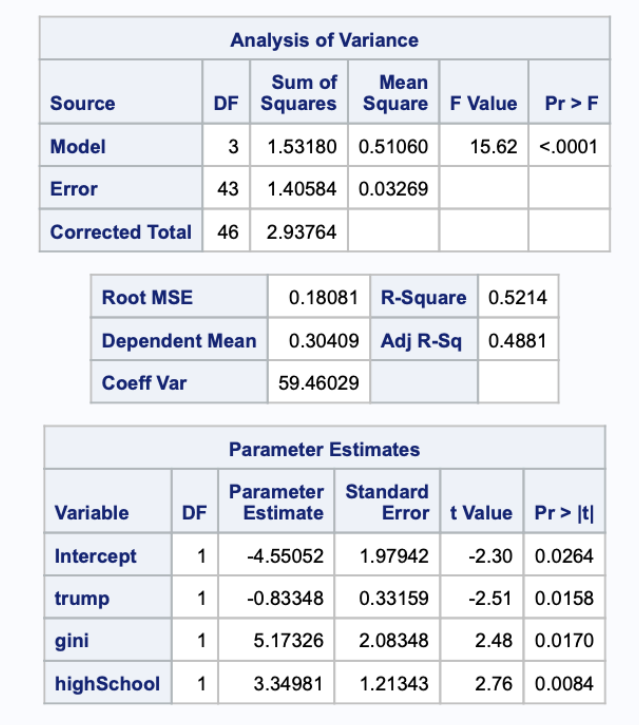
Figure 4: Final Forward Selection and Stepwise Elimination Model

Regression Assumptions

Linearity. Since the models we created did not have high R-squared values but were all significant, we decided to test our variables for linearity. To do this, we used SAS to create a scatter plot of each variable against the number of hate crimes per 100k in population reported to the SPLC. The scatterplot matrix for our dataset can be viewed in the additional graphs section of our paper (last page). From this, we were able to determine that in our dataset, there are five linear variables and four non-linear variables. The five linear variables are: Median Household Income, the share of Population with High School Degree, the GINI Index, Level of White Poverty, and the share of Population that Voted for Trump. Conversely, the non-linear variables in our dataset are: Unemployment Rate, share of Population in Metro Areas, share of Population Non-US Citizen, GINI Index, share of Non-White Population, share of Population that

Voted for Trump, the share of the Non-White Population.

Due to the inclusion of non-linear variables in our dataset, it was determined that another model should be built with only linear variables. Once again, the backward elimination method was used to create the best fit linear model for this dataset. This new model contained three variables, the share of the population that voted for Trump, the GINI index, and the share of the population that graduated from high school. The equation for this model is $y = -4.55052 - 0.83348x_1 + 5.17326x_2 + 3.34981x_3$ where y is the predicted number of hate crimes per 100k in population reported to the SPLC, x_1 is the share of the population that voted for Trump, x_2 is the GINI index, and $x_3$is the share of the population that graduated from high school. This model has an R-squared value of 0.5214. 52.14 percent of the variation within the dataset could be explained with the model. All of the variables in this model are significant at the $\alpha = 0.05$ level. Overall, there is sufficient evidence to indicate that the model is significant as determined by the F test. The model has an F-value of 15.62 and a p-value of <0.0001. From this, we were able to determine that this is the best model for our given dataset.



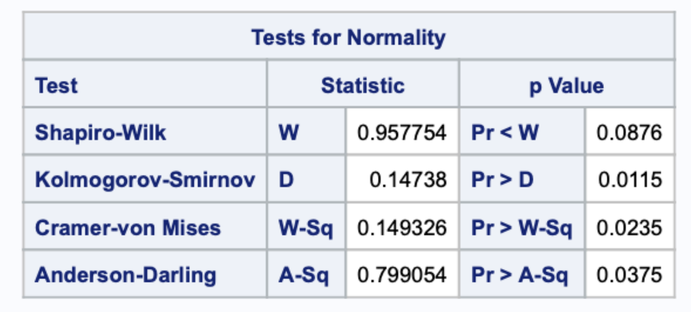
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.53180	0.51060	15.62	<.0001
Error	43	1.40584	0.03269		
Corrected Total	46	2.93764			

Root MSE	0.18081	R-Square	0.5214
Dependent Mean	0.30409	Adj R-Sq	0.4881
Coeff Var	59.46029		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.55052	1.97942	-2.30	0.0264
trump	1	-0.83348	0.33159	-2.51	0.0158
gini	1	5.17326	2.08348	2.48	0.0170
highSchool	1	3.34981	1.21343	2.76	0.0084

Figure 5: Final Model without Nonlinear Variables

Test for Normality and Constant Variance. Once we determined the best linear regression model for our dataset, we had to confirm that our linear assumptions were sound. First, we had to conduct a test for normality, followed by a test for homogeneity, test for collinearity, and finally a test for outliers, leverages, and influential points. First, to test the normality of the best fit model of the dataset, a Shapiro-Wilks test of Normality was conducted. Since the P-value of the W statistic is 0.0876, there is insufficient evidence at the =0.05significance level to indicate that the model is not normally distributed. Therefore, we can conclude that the model is normally distributed. Following this, the Chi-Squared test for homogeneity (constant variance) was conducted. Once again, with a p-value of 0.4830, there is insufficient evidence to indicate that the model does not have constant variance.



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.957754	Pr < W	0.0876
Kolmogorov-Smirnov	D	0.14738	Pr > D	0.0115
Cramer-von Mises	W-Sq	0.149326	Pr > W-Sq	0.0235
Anderson-Darling	A-Sq	0.799054	Pr > A-Sq	0.0375

Figure 6: Summary of Normality Statistics

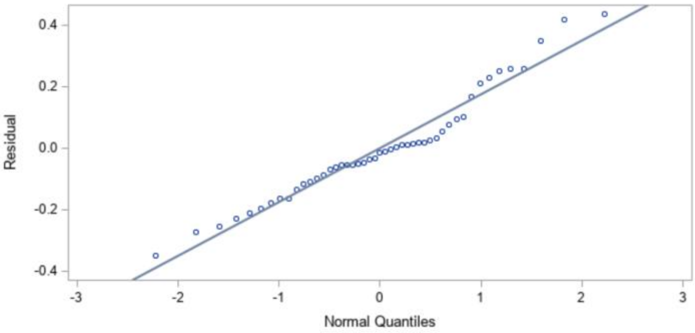
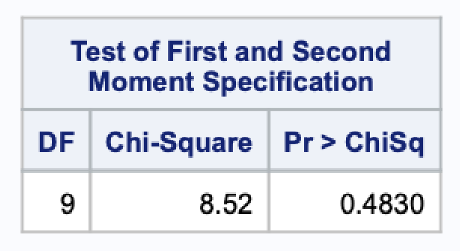


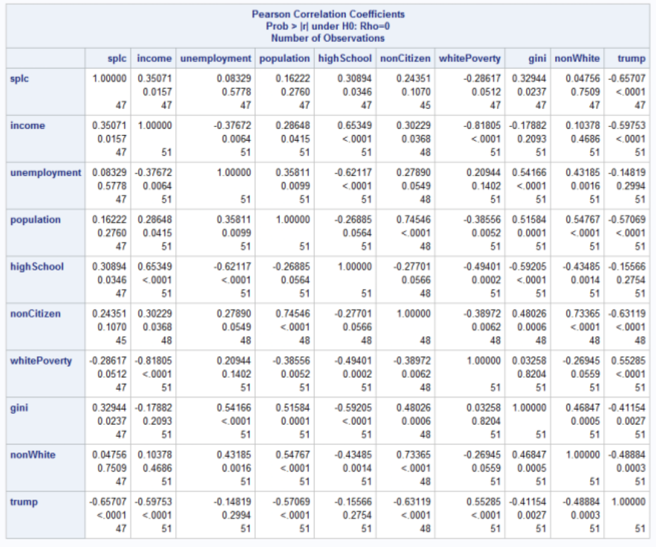
Figure 7: QQ Plot of the Data



Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
9	8.52	0.4830

Figure 8: Summary of Chi Squared Statistics

Test for Collinearity. After this, the model was tested for collinearity by using the correlation matrix. As a basic rule of thumb, any combination of variables with an absolute value of a Pearson Correlation Coefficient of 0.8 should be evaluated for collinearity issues. In our dataset, the variables for the level of white poverty and median household income were issues. But since neither were in our model, we determined there would not be any collinearity issues with our model.



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	splc	income	unemployment	population	highSchool	nonCitizen	whitePoverty	gini	nonWhite	trump
splc	1.00000 47	0.35071 0.0157 47	0.08329 0.5778 47	0.16222 0.2760 47	0.30894 0.0346 47	0.24351 0.1070 45	-0.28617 0.0512 47	0.32944 0.0237 47	0.04756 0.7509 47	-0.65707 <.0001 47
income	0.35071 0.0157 47	1.00000 51	-0.37672 0.0064 51	0.28648 0.0415 51	0.65349 <.0001 51	0.30229 0.0368 48	-0.81805 <.0001 51	-0.17882 0.2093 51	0.10378 0.4686 51	-0.59753 <.0001 51
unemployment	0.08329 0.5778 47	-0.37672 0.0064 51	1.00000 51	0.35811 0.0099 51	-0.62117 <.0001 51	0.27890 0.0549 48	0.20944 0.1402 51	0.54166 <.0001 51	0.43185 0.0016 51	-0.14819 0.2994 51
population	0.16222 0.2760 47	0.28648 0.0415 51	0.35811 0.0099 51	1.00000 51	-0.26885 0.0564 51	0.74546 <.0001 48	-0.38556 0.0052 51	0.51584 0.0001 51	0.54767 <.0001 51	-0.57069 <.0001 51
highSchool	0.30894 0.0346 47	0.65349 <.0001 51	-0.62117 <.0001 51	-0.26885 0.0564 51	1.00000 51	-0.27701 0.0566 48	-0.49401 0.0002 51	-0.59205 <.0001 51	-0.43485 0.0014 51	-0.15566 0.2754 51
nonCitizen	0.24351 0.1070 45	0.30229 0.0368 48	0.27890 0.0549 48	0.74546 <.0001 48	-0.27701 0.0566 48	1.00000 48	-0.38972 0.0062 48	0.48026 0.0006 48	0.73365 <.0001 48	-0.63119 <.0001 48
whitePoverty	-0.28617 0.0512 47	-0.81805 <.0001 51	0.20944 0.1402 51	-0.38556 0.0052 51	-0.49401 0.0002 51	-0.38972 0.0062 48	1.00000 51	0.03258 0.8204 51	-0.26945 0.0559 51	0.55285 <.0001 51
gini	0.32944 0.0237 47	-0.17882 0.2093 51	0.54166 <.0001 51	0.51584 0.0001 51	-0.59205 <.0001 51	0.48026 0.0006 48	0.03258 0.8204 51	1.00000 51	0.46847 0.0005 51	-0.41154 0.0027 51
nonWhite	0.04756 0.7509 47	0.10378 0.4686 51	0.43185 0.0016 51	0.54767 <.0001 51	-0.43485 0.0014 51	0.73365 <.0001 48	-0.26945 0.0559 51	0.46847 0.0005 51	1.00000 51	-0.48884 0.0003 51
trump	-0.65707 <.0001 47	-0.59753 <.0001 51	-0.14819 0.2994 51	-0.57069 <.0001 51	-0.15566 0.2754 51	-0.63119 <.0001 48	0.55285 <.0001 51	-0.41154 0.0027 51	-0.48884 0.0003 51	1.00000 51

Figure 9: Correlation Matrix

Analysis of Outliers, Leverage, and Influential Observations. Finally, an analysis of outliers, leverages, and influential points was conducted. From this, it was concluded that there were three outliers, West Virginia, Oregon and New Jersey. It was also determined that there was one leverage point, California. There was also one point that was an outlier and a leverage point, District of Columbia.

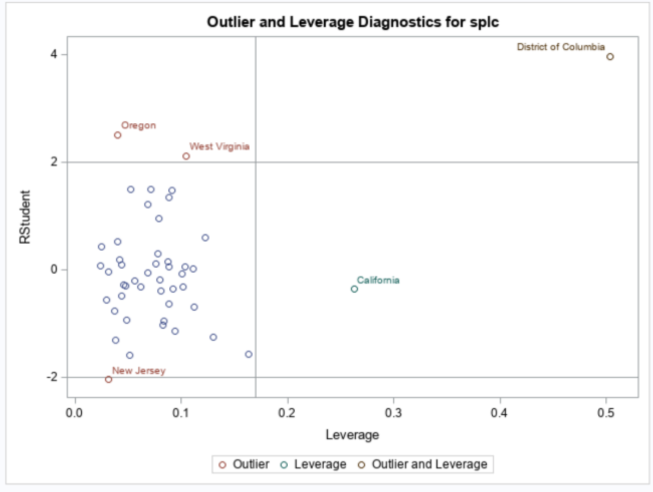


Figure 10: Outlier and Leverage Diagnostics

To determine if these outliers significantly affected our model, we conducted backward elimination and forward selection again without the outliers, and again without the leverage points and found a small shift in R-squared. However, when we conducted these methods again without the point that is both an outlier and leverage point (the District of Columbia), the R-squared of the backward elimination model fell to 0.3733, and the R-squared of the forward selection model dropped to 0.3316. Because of this drastic change in R-squared when the District of Columbia is not included in our dataset, we are certain that the District of Columbia is an influential point. Additionally, because of this stark change in R-squared, we cannot justify the removal of the outliers, leverages, or influential points within our dataset.

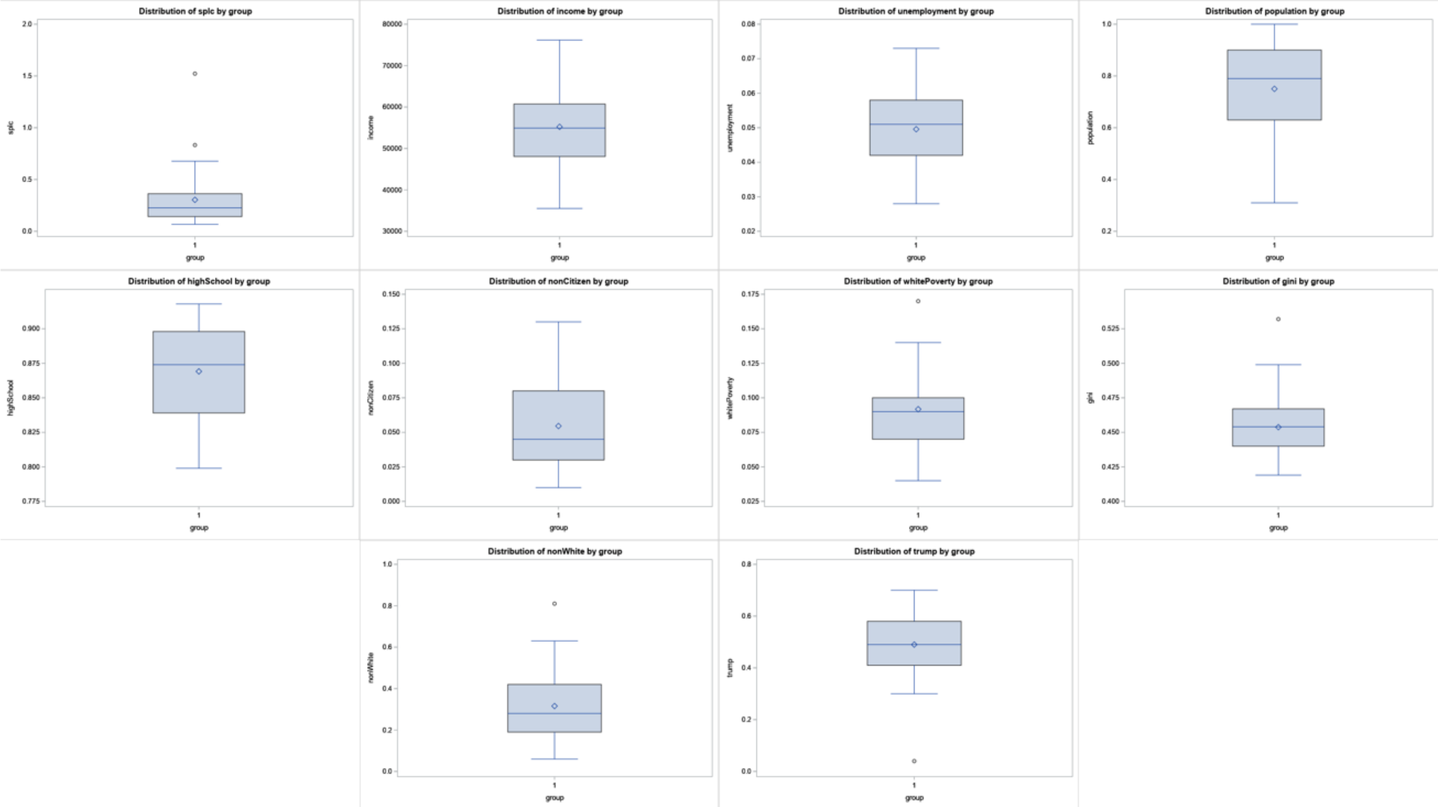
4. Conclusion

Since our model met all of the linear assumptions, we can conclude that this is the best linear model for our given dataset. Our final model is $y = -4.55052 - 0.83348x_1 + 5.17326x_2 + 3.34981x_3$ where y is the predicted number of hate crimes per 100k in population reported to the SPLC, x_1 is the share of the population that voted for Trump, x_2 is the GINI index, and x_3 is the share of the population that graduated from high school. As previously stated, this model was generated by SAS using all of the various analytical processes. Our model had an R-Squared value of 0.5214. This model does not have the highest R-squared value of the models generated by SAS. We could not justify using the model with the highest R-squared because it had all of the variables and a difference in R-squared of less than 0.05. We cannot justify increasing the complexity of the model for such a small difference in R-squared.

Overall the model states that the number of hate crimes reported to the Southern Poverty Law Center per 100k residents in a state can be estimated using a model relating the share of the population that voted for Trump, the GINI index, and the share of the population that graduated from high school. For every one unit increase in share of the population that voted for Trump, it is estimated that the number of hate crimes reported to the Southern Poverty Law Center per 100k residents in a state will decrease by 0.83348, holding all other variables constant. For every one unit increase in the GINI index, it is estimated that the number of hate crimes reported to the Southern Poverty Law Center per 100k residents in a state will increase by 5.17326, holding all other variables constant. For every one unit increase in the share of the population that graduated from high school, it is estimated that the number of hate crimes reported to the Southern Poverty Law Center per 100k residents in a state will increase by 3.34981, holding all other variables constant.

We are a bit disappointed in our final model’s utility because of the low R-squared value. It is more likely that our dataset does not follow a proper linear pattern, and instead, the best fit model for our dataset would follow some other relationship. However, because of the extent of our knowledge given our previous coursework we are unable to properly find the correct model for our dataset. Because of this, we are happy with the overall results of our project and know we tried our best with the knowledge we have at this point in our education.

4. Additional Graphs



Distribution of splc by group

Distribution of income by group

Distribution of unemployment by group

Distribution of population by group

Distribution of highSchool by group

Distribution of nonCitizen by group

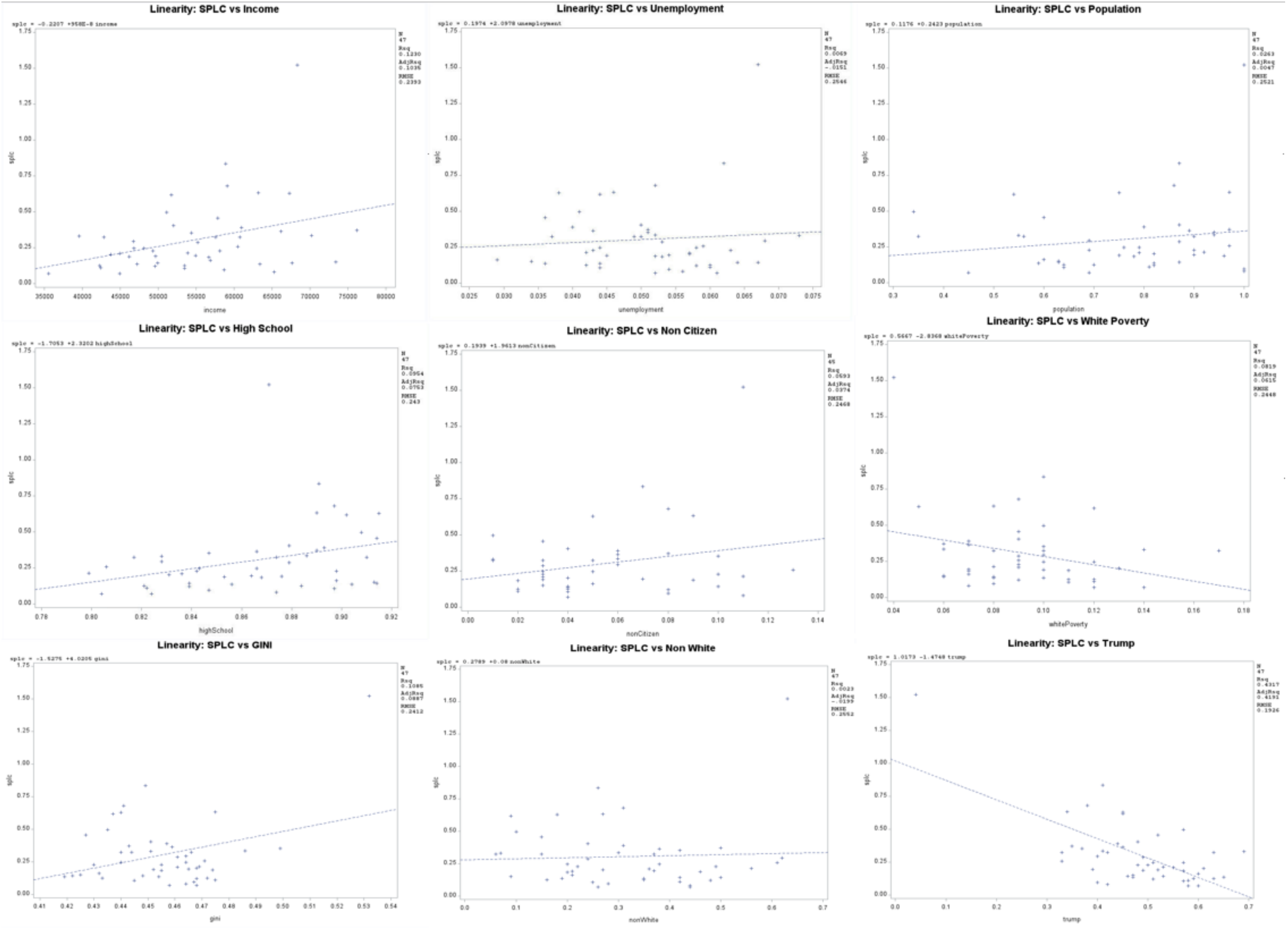
Distribution of whitePoverty by group

Distribution of gini by group

Distribution of nonWhite by group

Distribution of trump by group

Boxplot Matrix



Linearity: SPLC vs Income

Linearity: SPLC vs Unemployment

Linearity: SPLC vs Population

Linearity: SPLC vs High School

Linearity: SPLC vs Non Citizen

Linearity: SPLC vs White Poverty

Linearity: SPLC vs GINI

Linearity: SPLC vs Non White

Linearity: SPLC vs Trump

Scatterplot Matrix