Project Report

Caroline Andy, Vasili Fokaidis, Stella Li, Tessa Senders, Lily Wang

12/11/2020

Abstract

Introduction

Since 2014, the national rate of hate crimes has been steadily increasing in the United States [CITE]. In the days following the 2016 US Presidential election, an average of 90 hate crimes per day were reported to the Southern Poverty Law Center [CITE].

Existing research suggests that community-level demographic variables such as racial breakdown, population density, level of educational attainment, and economic considerations (median income, poverty level, job availability) may be significant predictors of regional and state-level rates of hate crimes. [CITE] A 2017 FivethirtyEight article titled, "Higher Rates Of Hate Crimes Are Tied To Income Inequality," used 2016 FBI and Southern Poverty Law Center data to assess the association between hate crime rate and select community-level variables [CITE].

For this project, we used this dataset to critically analyze this research team's findings to identify state-level variables associated with rates of hate crimes, and to generate a high performing predictive model for population-adjusted hate incidents in the United States.

Methods

The data used for this project included state-level hate crime rates (hate crimes per 100,000 population), as reported by the Southern Poverty Law Center during the first weeks of November, 2016. Collected state-level demographic variables include:

- Unemployment rate (high vs low) (as of 2016)
- Urbanization (high vs low) (as of 2015)
- Median household income (as of 2016)
- Percent of residents with a high school degree (as of 2009)
- Percent of residents who are non-citizens (as of 2015)
- Income Gini coefficient (a measure of the extent to which the distribution of income among individuals within an economy deviates from a perfectly equal distribution; as of 2015)
- Percent of residents who are non-White (as of 2015)

First, we looked for missing data points. Four states–Hawaii, North Dakota, South Dakota, and Wyoming–did not report hate crime rate data, and thus were excluded from subsequent analyses. One additional state, Maine, did not report its percent of residents who were non-citizens. Washington DC was included as a state for the purposes of these analyses.

Using this smaller dataset, our goal was to generate a multivariable linear regression model to assess which of these variables, if any, are associated with population-adjusted hate incidents in the United States. To do so, we first generated descriptive statistics and plotted the distribution of the outcome (population-adjusted hate incidents per 100,000 population) to determine whether any data transformations would be necessary, and to assess whether any outliers exist within the data.

To test for any multi-collinearity between the continuous variables, we calculated a correlation matrix. We decided that any correlation coefficient of 0.6 and above may suggest multi-collinearity, thus, one of those correlated variables should then be dropped from subsequent analyses.

Results

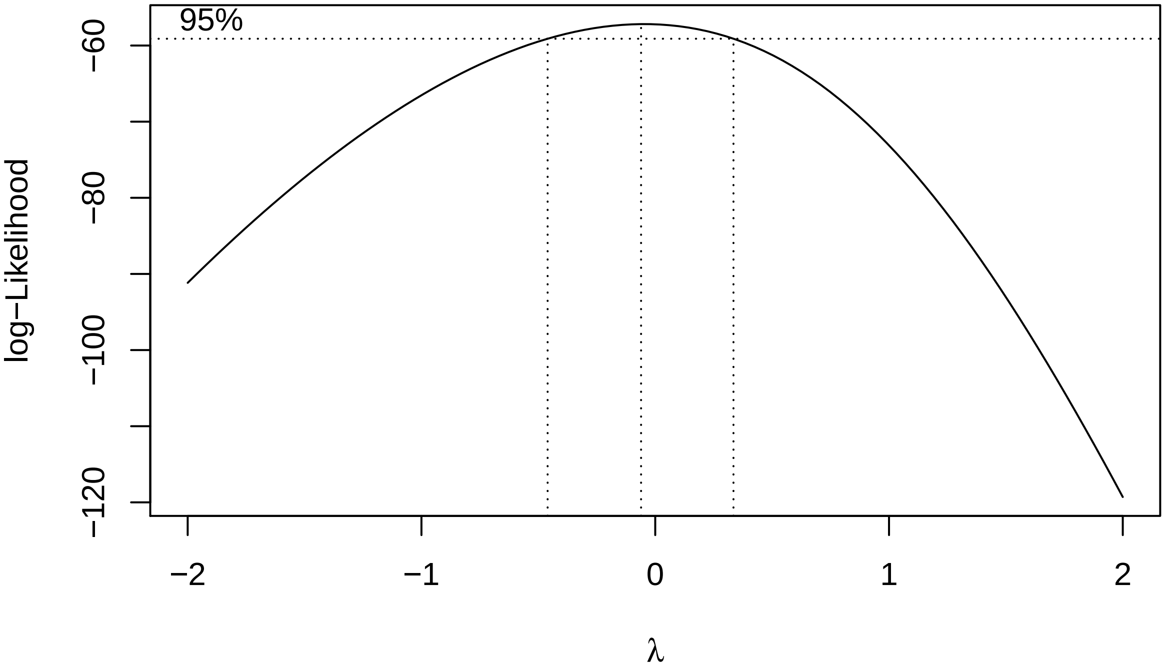
Descriptive Statistics and Data Distribution We first generated descriptive statistics for all collected variables.

Table 1: Table 1: Descriptive Statistics of Tidied Dataset

	Overall (N=47)
Unemployment	
high	24 (51.1%)
low	23 (48.9%)
Missing	0
Urbanization	
low	23 (48.9%)
high	24 (51.1%)
Missing	0
Median Household Income	
Mean (SD)	54802.298 (9255.117)
Median (Q1, Q3)	54310.000 (47629.500, 60597.500)
Min - Max	35521.000 - 76165.000
Missing	0
% Adults >25yrs With HS Degree	
Mean (SD)	0.866 (0.034)
Median (Q1, Q3)	0.871 (0.839, 0.895)
Min - Max	0.799 - 0.915
Missing	0
% of Population Not U.S. Citizens	
Mean (SD)	0.055 (0.031)
Median (Q1, Q3)	0.050 (0.030, 0.080)
Min - Max	0.010 - 0.130
Missing	2
Gini Index	
Mean (SD)	0.456 (0.021)
Median (Q1, Q3)	0.455 (0.441, 0.468)
Min - Max	0.419 - 0.532
Missing	0
% of Population Not White	
Mean (SD)	0.315 (0.150)
Median (Q1, Q3)	0.300 (0.205, 0.420)
Min - Max	0.060 - 0.630
Missing	0
Hate Crime Rate Per 100k	
Mean (SD)	0.304 (0.253)
Median (Q1, Q3)	0.226 (0.143, 0.357)
Min - Max	0.067 - 1.522
Missing	0

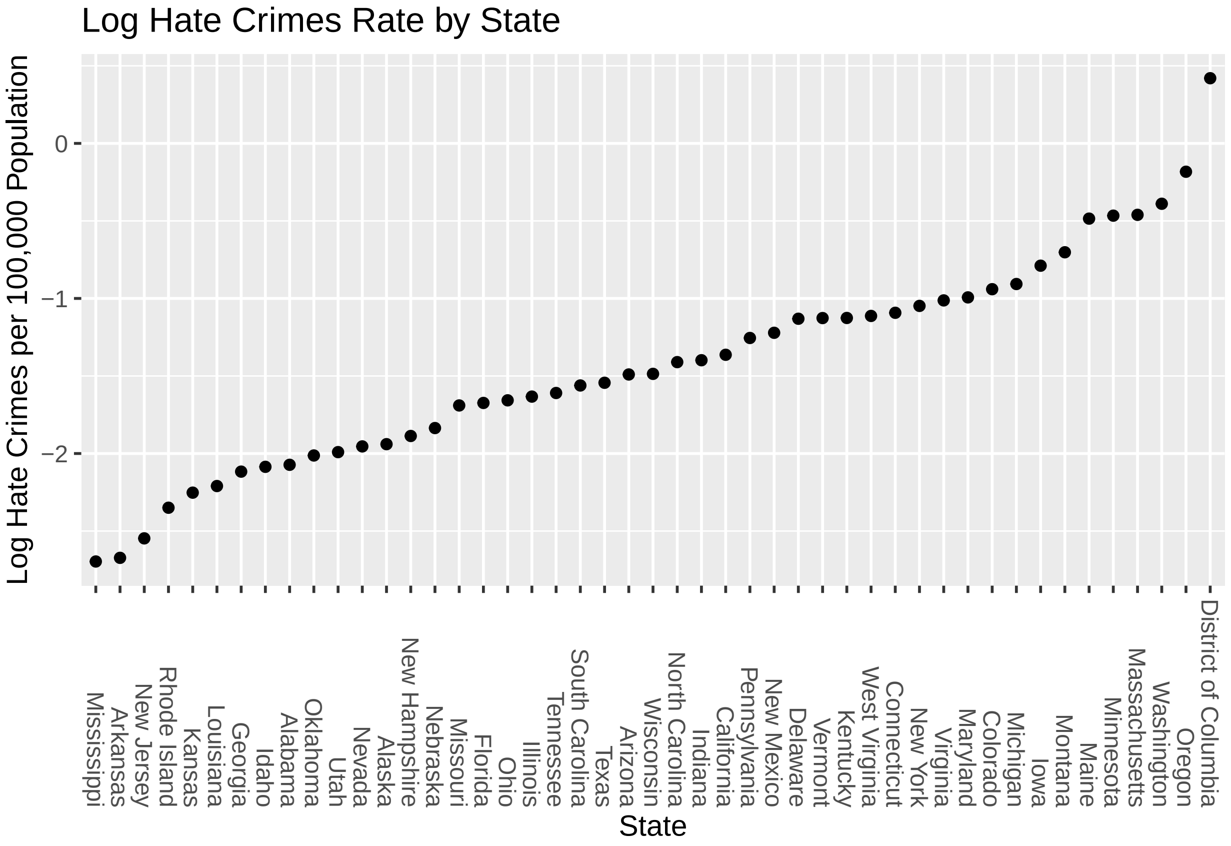
Multivariable linear regression modeling operates under several assumptions, which include residual homoscedasticity (constant variance) and normality. Initial exploration of the distribution of the hate crimes rate data showed a strong departure from standard normal distribution. Thus, we performed a Box Cox test to isolate the 'best' power transformation on the hate crimes rate variable to achieve normal residuals.

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ unemployment + urbanization +
##     median_household_income + perc_pop_hs + perc_non_citizen +
##     gini_index + perc_non_white, data = hate_crimes_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36552 -0.10314 -0.01316  0.09731  0.51389
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -8.263e+00  1.897e+00  -4.356 0.000101 ***
## unemploymentlow          1.307e-02  7.173e-02   0.182 0.856425
## urbanizationhigh        -3.309e-02  8.475e-02  -0.390 0.698475
## median_household_income -1.504e-06  5.961e-06  -0.252 0.802193
## perc_pop_hs              5.382e+00  1.835e+00   2.933 0.005735 **
## perc_non_citizen         1.233e+00  1.877e+00   0.657 0.515332
## gini_index               8.624e+00  1.973e+00   4.370 9.67e-05 ***
## perc_non_white          -5.842e-03  3.673e-01  -0.016 0.987396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2014 on 37 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.461,  Adjusted R-squared:  0.3591
## F-statistic: 4.521 on 7 and 37 DF,  p-value: 0.001007
```

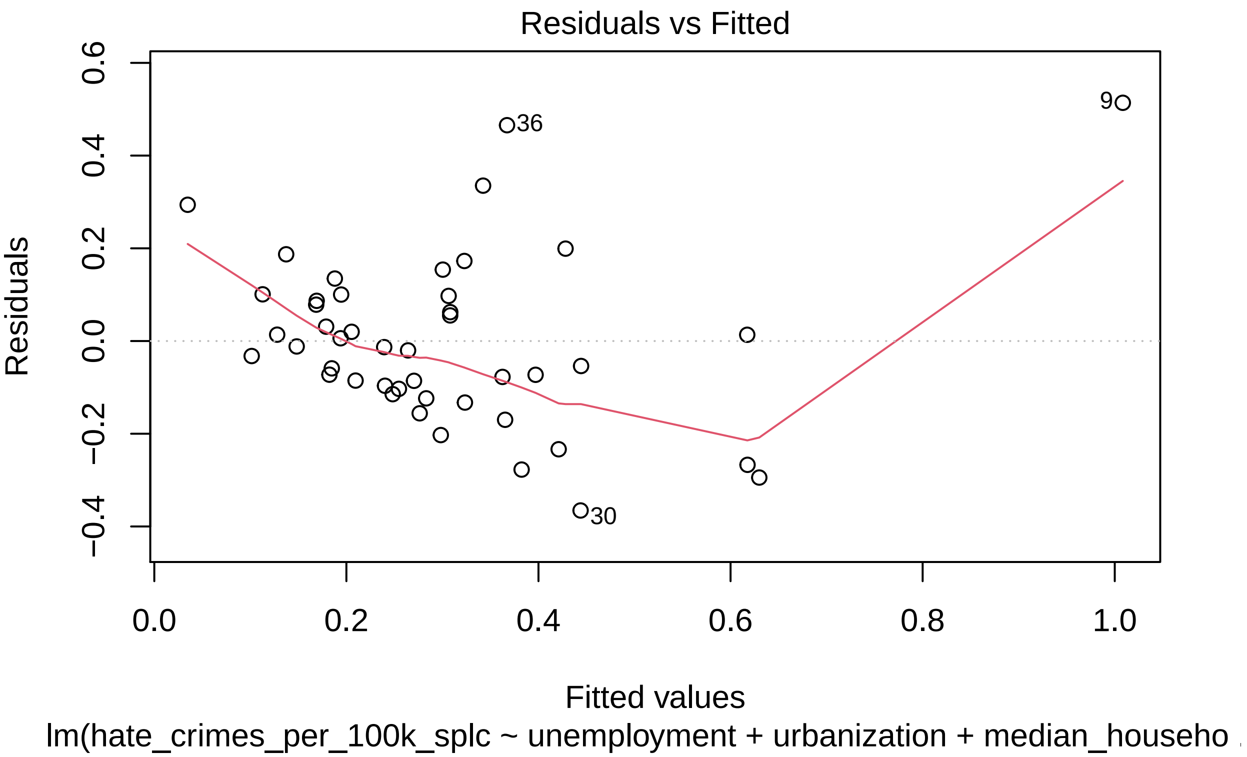


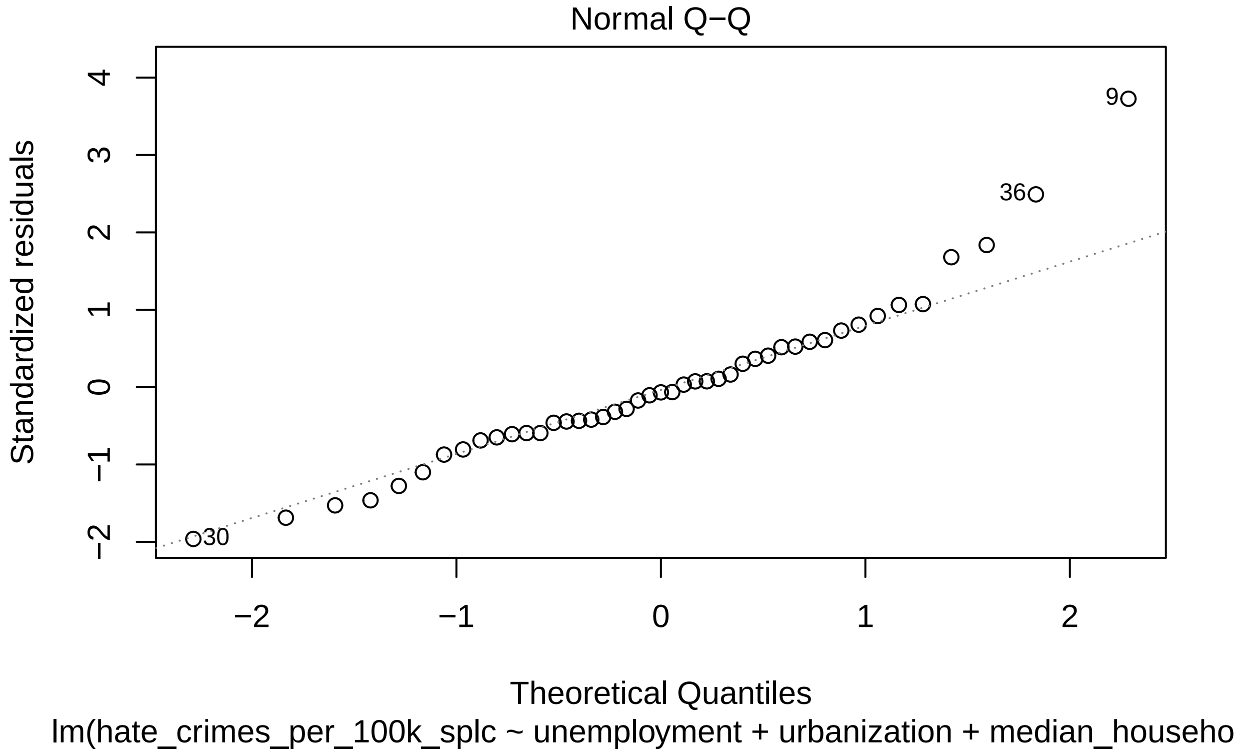
These results suggest that a logarithmic transformation of the hate crimes rate data would most closely approximate normal distribution. Thus, we operate moving forward in model development using the log hate crimes rate as the outcome of interest.

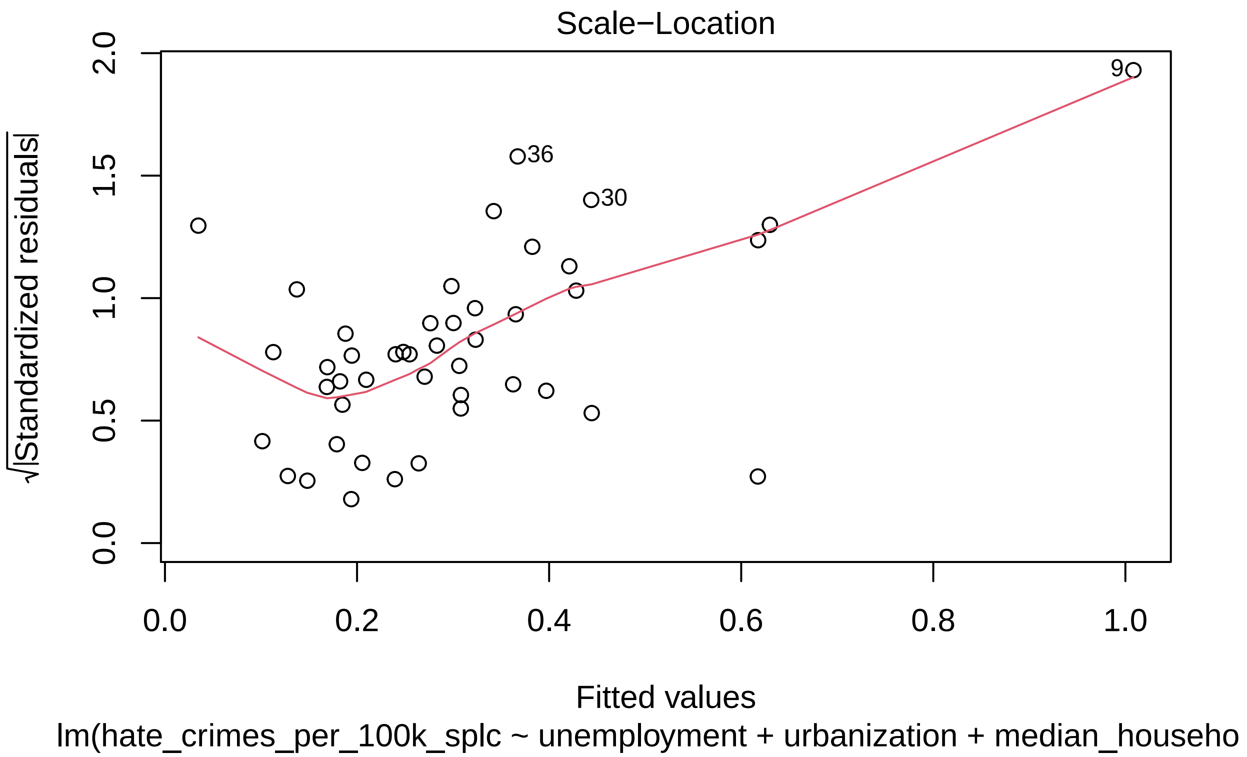
To determine whether any outliers exist within the data, we first visualized the log hate crimes rate of each state.

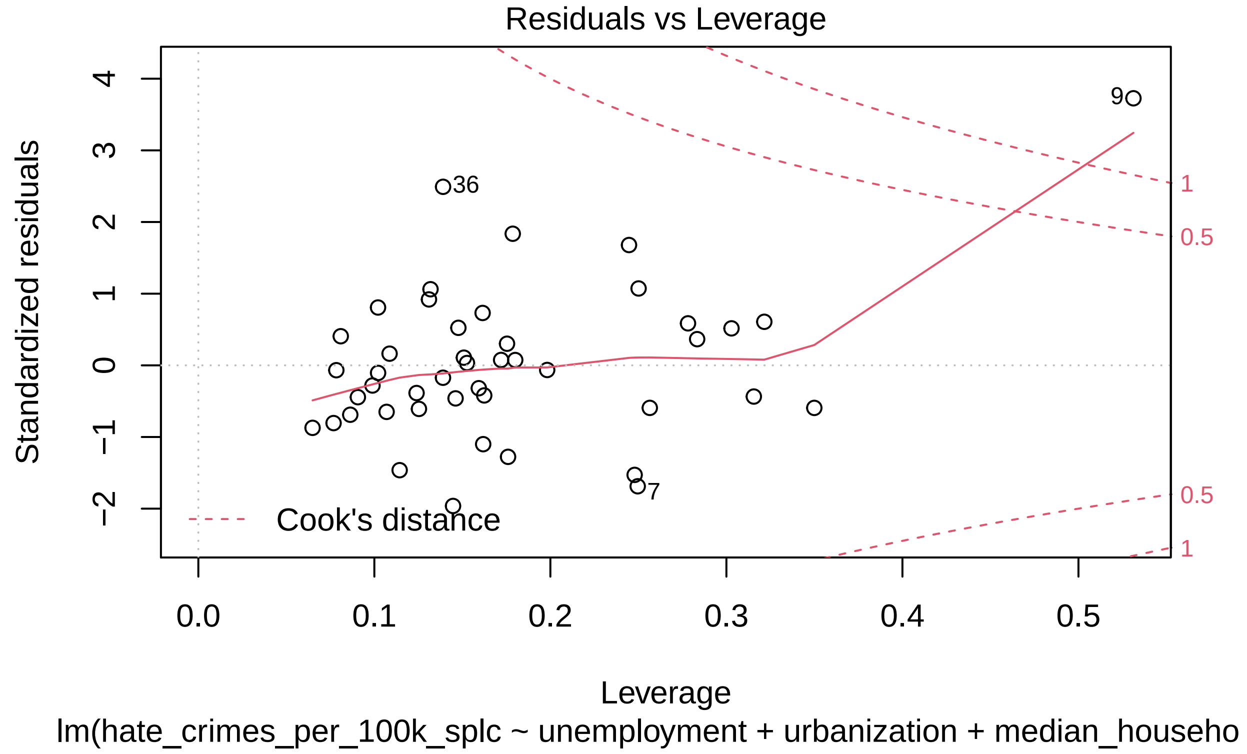


Results show that Washington, DC may be an outlier and/or influential point in the data. To further investigate this, we generated the residuals vs fitted, normal Q-Q, scale-location, and residuals vs leverage plots.


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.2
−0.4
36
9
30
0.0
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(hate_crimes_per_100k_splc ~ unemployment + urbanization + median_househo


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
9
36
30
−2
−1
0
1
2
Theoretical Quantiles
lm(hate_crimes_per_100k_splc ~ unemployment + urbanization + median_househo





As Washington DC is clearly outside of the Cook's distance line, we conclude that it is an outlier.

Tests for Association and Modeling We began our regression analyses by running linear regression models containing all possible predictors for both the untransformed hate crime rate and the log hate crime rate. Our results support the conclusions drawn in the FiveThirtyEight article: that the Gini index (income inequality) is the most significant predictor, and that percent population with a high school degree is the only other significant predictor.

In the interest of maximizing model parsimony and predictive performance, we then performed several automated procedures for variable selection. Specifically, we used backward and forward selection. These results suggest that the Gini index and percent population with a high school degree variables are the only significant predictors included in the final model.

We then employed criterion based approaches in variable selection,

- We wanted to enhance model parsimony and improve performance so we first started performing automated procedures for variable selection
- We also used criterion based approaches in variable selection, including running tests to generate the 2 best models for various criteria: Cp, adjusted R squared and SSE/RSS.
- We performed a lit review to see what variables are practically important to include for interpretability and application purposes

Discussion/Conclusion

Figures and Tables

References