Stat 21 Final Project Proposal + Analysis

Xinxin Li, Christopher Meng, Jessica Sang, Shikha Shrestha

April 28, 2022

Proposal

Section 1: Introduction

Who?

List all group members and the responsibilities of each group member. Verify that each group member has read these instructions carefully.

Group 1 consists of the following members:

- Xinxin Li, who will be responsible for data collection, cleaning, and importing into R, analyzing data in R, report write-up
- Christopher Meng, who will be responsible for analyzing data in R, report write-up, poster presentation creation
- Jessica Sang, who will be responsible for analyzing data in R, group meeting organizing (scheduling), report write-up
- Shikha Shrestha, who will be responsible for analyzing data in R, poster presentation creation, report write-up

We confirm that we have read the instructions detailed on this proposal carefully.

What?

```
# Our data
homicide_data <- read_excel("group1_data.xlsx", sheet = "Standardized")
homicide_data$Political_side <- as.factor(homicide_data$Political_side)
homicide_data$State <- as.factor(homicide_data$State)
homicide_data %>% head
```

```
## # A tibble: 6 x 11
##   State      Abbreviation Political_side Gini_Index Uninsured_perce~ Median_househol~
##   <fct>      <chr>        <fct>               <dbl>            <dbl>            <dbl>
## 1 Alabama    AL           red                  0.47              9.7             56.2
## 2 Alaska     AK           red                  0.44             11.5             78.4
## 3 Arizona    AZ           red                  0.46             11.1             70.7
```

```
## 4 Arkansas   AR          red                 0.48            9.1            54.5
## 5 California CA          blue                0.49            7.8            78.1
## 6 Colorado   CO          blue                0.45            7.8            72.5
## # ... with 5 more variables: Poverty_percentage <dbl>,
## #   Unemployment_percentage <dbl>, Homicide_rate <dbl>,
## #   High_school_degree_percentage <dbl>, Bachelor_degree_percentage <dbl>
```

Research question What variables have relatively significant effects on the homicide rate in the US in 2019?

Motivation We were inspired to explore this research question because of the findings in the FiveThirtyEight article "Higher Rates of Hate Crimes are Tied to Income Inequality." They found that states with more inequality were more likely to have higher rates of hate incidents per capita (Majumder 2017). Instead of using hate crimes as our response variable, we decided to look at the homicide rate per 100,000 people. We look at more recent data from 2019 and look at similar variables, in addition to a few different ones.

According to John R. Hipp's paper "Income Inequality, Race, and Place: Does the Distribution of Race and Class within Neighborhoods Affect Crime Rates?" income inequality has been linked to violence possibly for the reason that viewing your situation relative to others around you, as opposed to your own situation in isolation, and seeing a very large gap can make you angry or resentful towards the other group (Majumder 2017). In addition to income inequality, other variables like education and the level of social support have also been found to affect crime (Bell, Costa, and Machin 2018) (Kort-Butler 2018). Looking at which variables are significantly related to crime more generally are important because crime has negative effects on everyone, and figuring out which of these predictors has the strongest influence on crime can help us develop targeted solutions for crime reduction.

Statistical questions related to the larger research question

1. Which predictor variable has the strongest influence on the homicide rate in the US in 2019?
2. Are there any multicollinearity problems that are relatively significant? Which variables are positively correlated with the response variable (the homicide rate)?
3. Which subset of predictors can best predict the response variable?
4. Were there any data points that were influential?

Description for the data

Observational units Our observational units are state data.

Response variable and its variable type The response variable is the homicide rate per 100,000 and it is a numerical variable.

Predictor variables Below are the population coefficients we wish to understand using statistical inference.

Median_household_income (numerical)

Measured in US dollars. Using statistical inference to determine the significance of this variable's effect on the homicide rate could be interesting because there are multiple ways that median income can affect the homicide rate. One possibility is that states with higher median household incomes could be more well-resourced, which improves many qualities of life from education, healthcare access, state and local welfare

support, and etc, which could lower crime. However, median household income is not adjusted for cost of living, which means that it is also possible that states that have higher median incomes might be ones with larger cities, which could mean that having a higher median income could also lead to a higher homicide rate ("Release Tables: Median Household Income by State, Annual | FRED | St. Louis Fed" 2019).

Poverty_rate - (numerical)

Measured in the average number of people that live below the poverty line per 1,000 people. This data is a two-year average for 2018-2019. Lower income might be associated with higher crime rates as people resort to illegal methods to earn more money. Likewise, it might also be correlated with the happiness index (Bureau n.d.).

Unemployment_percentage (numerical)

Using statistical inference to determine the significance of this variable's effect on the homicide rate could be interesting because people tend to commit crimes when there is no better alternative. Therefore, if there is a higher unemployment rate, there might be a higher rate of homicide ("Unemployment Rates for States" n.d.).

Gini_Index - (numerical)

Gini_Index (Measure of socioeconomic inequality) (numerical): using statistical inference to determine the significance of this variable's effect on the homicide rate could be interesting because states with greater socioeconomic inequality might have higher homicide rates due to greater wealth gap. Gini index should lie between 0 and 1, where 0 means perfectly equal and 1 means totally unequal ("Gap Between Rich and Poor, by State in the U.S. 2019" n.d.).

Uninsured_percentage - (numerical)

Measures the percentage of the population with no coverage of health insurance in any form. For states with on average lower coverage rate for health insurance, the insecurities could lead to higher homicide rate.

Political_side - (categorical)

This would be coded for two different levels, "blue" or "red," where blue means the state in the 2016 election supported the Democratic Party, and red means the state supported the Republican Party. Whether a state is blue or red can have implications for state elected leaders that have influence over state support in terms of unemployment, social services, public education spending, and etc. We decided to use the political side from the 2016 election because that was the most recent election before 2019 ("2016 Presidential Election Results" 2017).

High_school_degree_percentage (numerical)

The percentage of individuals that have a high school degree among the 25-44-year-old population. Using statistical inference to determine the significance of this variable's effect on the homicide rate could be interesting because there have been associations between education and crime ("Individuals with High School or Higher Level Degree Among 25–44-Year-Old Population | State Indicators | National Science Foundation - State Indicators" n.d.).

Bachelor_degree_percentage (numerical)

The percentage of individuals that have a bachelor degree among the 25-44-year-old population. Using statistical inference to determine the significance of this variable's effect on the homicide rate could be interesting because there have been associations between education and crime ("Bachelor's Degree Holders Among Individuals 25–44 Years Old | State Indicators | National Science Foundation - State Indicators" n.d.).

Note: all the data except the political side are 2019 data.

Other relevant variables

Quality of social services provided by the state

Homelessness rate

Divorce rate

% of population in urban areas

Urban Areas tend to have higher crime rates. This variable might also be collinear to population.

Happiness index

Mental well-being could play a role in people choosing to engage in criminal activities.

Section 2: Regression Analysis

How?

Generalized Steps

- Choose
  - Based on the data that we found, we would first use all the predictor variables for our first rough model.
  - Combining with the result from the Assessing stage, we would use the following method to choose predictors for the model:
    - Exploratory data analysis (scatter plots)
    - Forward selection/backward elimination
    - Added variable plots
    - Mallow's C_p
    - Adjusted R^2
    - Identifying moderately or extremely unusual data points (Standardized/studentized residuals, Leverage, Cook's Distance)
- Fit
  - We will use lm() with all predictors and their interaction terms in for our first rough model and then make edits based on our results from assess and choose.
  - Check betas and the intercept.
- Assess
  - Residual vs fitted value plots for linearity and constant variance
  - Normal quantile plots for normality
  - Variance inflation factor and scatterplot matrix for multicollinearity
  - Added Variable Plots
  - ANOVA F-test for overall model fit
  - Nested F-test for a reduced model
  - T-test for the significance of a specific predictor term
  - Check adjusted R^2 for the amount of variability of response variable that could be explained by the model
  - Identifying moderately or extremely unusual data points:
    - Standardized/studentized residuals
    - Leverage
    - Cook's Distance
- Use

- We would use confidence interval and prediction interval to predict homicide rate for specific predictor values
- We would use hypothesis test to make inference.

Detailed decription First, we will try to understand and summarize the data using exploratory data analysis. We will make scatterplots to visualize the relationships between the quantitative predictors, as well as the predictors with the response; we will make boxplots for any categorical predictors. A correlation matrix will tell us which variables are most strongly related to the response variable; the strongest correlations are a starting point for us to think about which variables might be most important for modeling the variability in the response.

We could also employ stepwise regression to choose predictor variables, starting with forward selection and choosing variables that give the biggest statistically significant increase in adjusted R2 and then using backward elimination as needed. We can also use the method of best subsets with the goals of maximizing adjusted R2 and minimizing Mallows Cp and compare these models to each other to help with our selection process.

As we add more predictors to our model, we also have to consider multicollinearity, drawing from the scatterplot matrix and the VIF values. If variables are highly multicollinear (VIF > 5), then we should be careful interpreting those individual coefficients/t-tests, consider combining predictors, and/or possibly dropping some predictors. Additionally, we should consider an interaction model, and which predictors might possibly interact with each other, and use nested F-tests/t-tests to determine if an interaction is significant to include in the model.

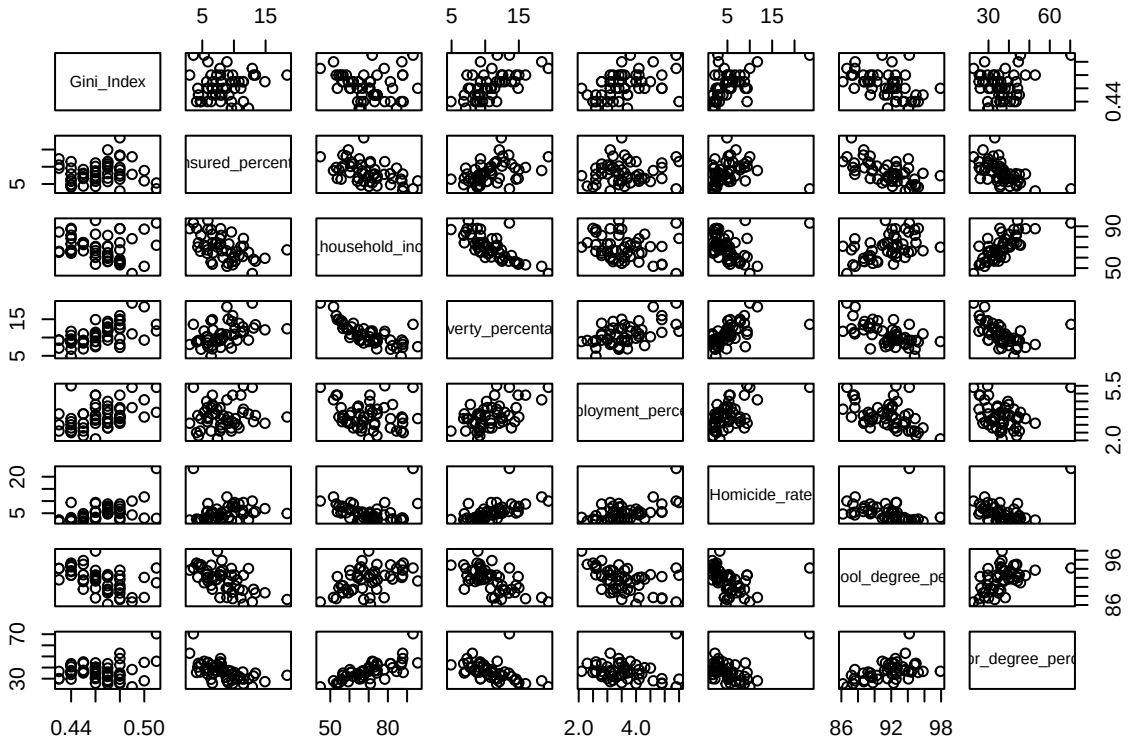
Finally, we also have to ensure that our regression assumptions are met, namely linearity, constant variance, zero mean of error, independence, and for inference, normality and randomness. Looking at residual plots and normal quantile plots, we may need to transform the predictors and/or the response to better meet the assumptions. We may also want to isolate one variable while accounting for the contributions of the other predictors we are considering, using added variable plots. AVPs can help us see if any regression assumptions are violated with respect to a particular variable, or if there are any interesting patterns in unusual points that would help guide our exploration. Similarly, calculating residuals, leverage, and Cook's distance will help us consider the generalizability of our model and any patterns within unusual points that we might be able to account for in our variable selection process.

Once we have undergone this iterative selection process, we can fit our model, double checking the regression assumptions and examining summary statistics like adjusted R2 and the overall F-test. We may want to see the performance of our model on similar data from other years, or we can perform cross-validation to better estimate our performance on data that our model was not fitted to. We can also calculate intervals and conduct hypothesis tests to answer our research questions, keeping in mind the limitations of our model based on our sample of data, generalizability, and the regression assumptions.

Deriving the best model

Exploring the scatterplot and correlation matrices for the (numeric) predictor variables and the response:

```
pairs(homicide_data[,c(4:11)])
```




```
cor(homicide_data[,c(4:11)])
```

```
##                                  Gini_Index Uninsured_percentage
## Gini_Index                       1.00000000           0.04492758
## Uninsured_percentage             0.04492758           1.00000000
## Median_household_income_1k      -0.13908682          -0.49912814
## Poverty_percentage               0.57318141           0.41792626
## Unemployment_percentage          0.44594217           0.09842216
## Homicide_rate                    0.53648027           0.11266603
## High_school_degree_percentage   -0.48855943          -0.59250996
## Bachelor_degree_percentage       0.18499325          -0.59267516
##                                 Median_household_income_1k Poverty_percentage
## Gini_Index                                       -0.1390868          0.5731814
## Uninsured_percentage                             -0.4991281          0.4179263
## Median_household_income_1k                        1.0000000         -0.7428485
## Poverty_percentage                               -0.7428485          1.0000000
## Unemployment_percentage                          -0.1975750          0.5423449
## Homicide_rate                                    -0.1166421          0.5812251
## High_school_degree_percentage                     0.4675901         -0.6114788
## Bachelor_degree_percentage                        0.7440450         -0.4843645
##                                 Unemployment_percentage Homicide_rate
## Gini_Index                                   0.44594217     0.5364803
## Uninsured_percentage                         0.09842216     0.1126660
## Median_household_income_1k                  -0.19757495    -0.1166421
## Poverty_percentage                           0.54234491     0.5812251
```

```
## Unemployment_percentage                    1.00000000     0.5686683
## Homicide_rate                              0.56866828     1.0000000
## High_school_degree_percentage             -0.46227067    -0.3137342
## Bachelor_degree_percentage                -0.13648970     0.1352436
##                               High_school_degree_percentage
## Gini_Index                                       -0.4885594
## Uninsured_percentage                             -0.5925100
## Median_household_income_1k                        0.4675901
## Poverty_percentage                               -0.6114788
## Unemployment_percentage                          -0.4622707
## Homicide_rate                                    -0.3137342
## High_school_degree_percentage                     1.0000000
## Bachelor_degree_percentage                        0.5098739
##                               Bachelor_degree_percentage
## Gini_Index                                     0.1849932
## Uninsured_percentage                          -0.5926752
## Median_household_income_1k                     0.7440450
## Poverty_percentage                            -0.4843645
## Unemployment_percentage                       -0.1364897
## Homicide_rate                                  0.1352436
## High_school_degree_percentage                  0.5098739
## Bachelor_degree_percentage                     1.0000000
```

Determine the predictors using adjusted R^2 and Mallow's C_p

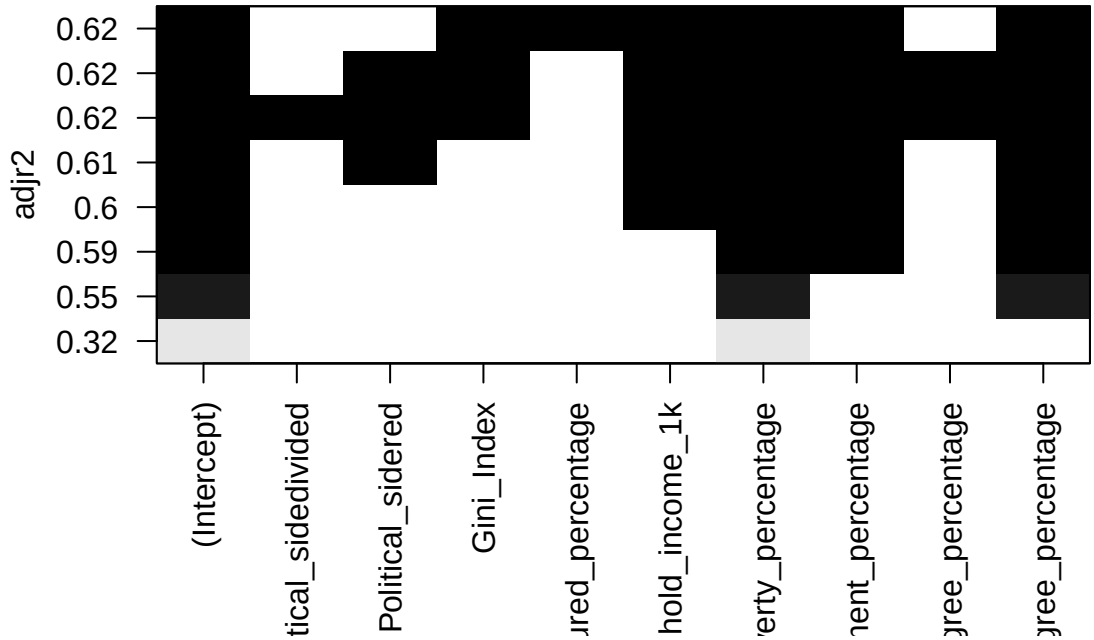
```
homicide_data_noState <- select(homicide_data, select = -c(State, Abbreviation))
all <- regsubsets(Homicide_rate ~., data = homicide_data_noState, really.big=T)

summary(all)$adjr2
```

```
## [1] 0.3243088 0.5466252 0.5899691 0.6030597 0.6119173 0.6221338 0.6181943
## [8] 0.6160540
```

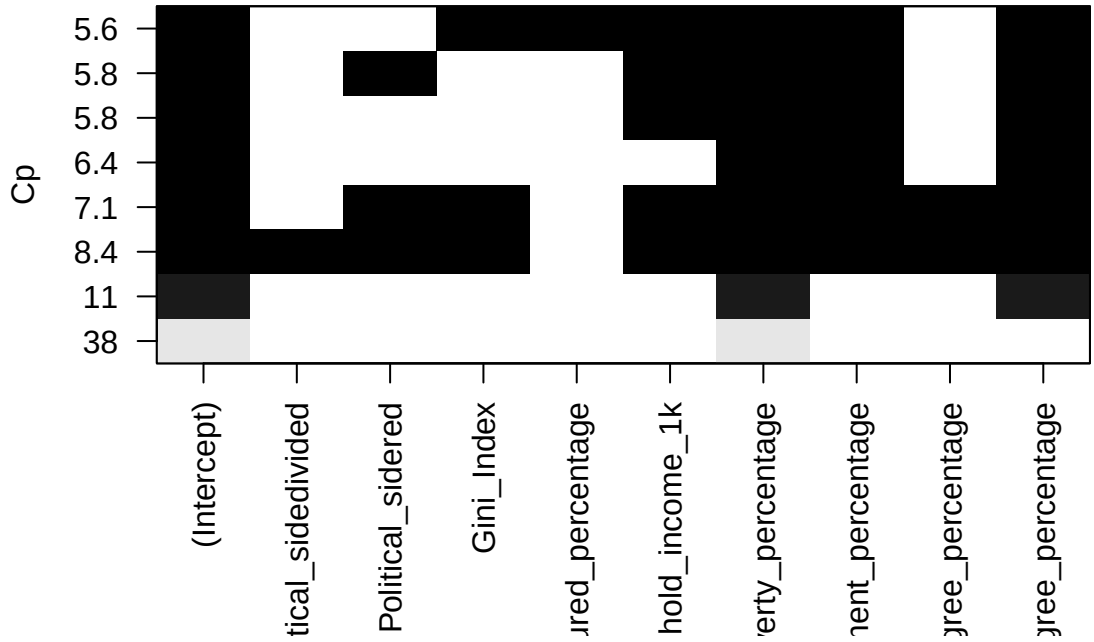
```
summary(all)$cp
```

```
## [1] 37.904500 10.806527  6.419781  5.824094  5.784033  5.636068  7.101463
## [8]  8.352882
```

```
plot(all, scale = "adjr2")
```




```
plot(all, scale = "Cp")
```




```
# the predictor names are cut off
# we tried different things but unable to fix it
```

Best subset of predictors from both adjusted R^2 and Mallow's C_p:

Gini_Index, Uninsured_percentage, Median_household_income_1k, Poverty_percentage, Unemployment_percentage, Bachelor_degree_percentage.

Forward selection

```
# Poverty rate has the highest R^2 for the model with one predictor
mod <- lm(Homicide_rate ~ Poverty_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7074 -1.4105 -0.4087  0.6838 16.2067
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         -2.6715     1.6281  -1.641    0.107
## Poverty_percentage   0.7253     0.1451   5.000 7.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.006 on 49 degrees of freedom
## Multiple R-squared:  0.3378, Adjusted R-squared:  0.3243
## F-statistic:    25 on 1 and 49 DF,  p-value: 7.741e-06
```

Add bachelor degree

```
mod <- lm(Homicide_rate ~ Poverty_percentage + Bachelor_degree_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Bachelor_degree_percentage,
##     data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5385 -1.3873 -0.1202  1.1755  7.0136
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -15.16147    2.83051  -5.356 2.37e-06 ***
## Poverty_percentage           1.05449    0.13583   7.763 5.03e-10 ***
## Bachelor_degree_percentage   0.24441    0.04885   5.003 7.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.463 on 48 degrees of freedom
## Multiple R-squared:  0.5648, Adjusted R-squared:  0.5466
## F-statistic: 31.14 on 2 and 48 DF,  p-value: 2.135e-09
```

See whether the interaction between bachelor degree and poverty rate is significant

```
mod <- lm(Homicide_rate ~ Poverty_percentage * Bachelor_degree_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage * Bachelor_degree_percentage,
##     data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8039 -1.4445  0.1085  1.1943  6.2768
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                 -2.85413    6.90066  -0.414
```

```
## Poverty_percentage                                0.07336    0.52148   0.141
## Bachelor_degree_percentage                       -0.09661    0.18166  -0.532
## Poverty_percentage:Bachelor_degree_percentage     0.02804    0.01442   1.945
##                                                 Pr(>|t|)
## (Intercept)                                       0.6810
## Poverty_percentage                                0.8887
## Bachelor_degree_percentage                        0.5974
## Poverty_percentage:Bachelor_degree_percentage     0.0578 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.394 on 47 degrees of freedom
## Multiple R-squared:  0.5972, Adjusted R-squared:  0.5715
## F-statistic: 23.23 on 3 and 47 DF,  p-value: 2.285e-09
```

The interaction term above is not significant, so we discard the interaction term and add unemployment rate.

```
mod <- lm(Homicide_rate ~ Poverty_percentage + Bachelor_degree_percentage + Unemployment_percentage,
          data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Bachelor_degree_percentage +
##     Unemployment_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4283 -1.5793 -0.3139  1.3508  6.0571
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -16.49180    2.74540  -6.007 2.63e-07 ***
## Poverty_percentage           0.84511    0.15461   5.466 1.71e-06 ***
## Bachelor_degree_percentage   0.22445    0.04716   4.759 1.89e-05 ***
## Unemployment_percentage      1.21101    0.49137   2.465   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.342 on 47 degrees of freedom
## Multiple R-squared:  0.6146, Adjusted R-squared:   0.59
## F-statistic: 24.98 on 3 and 47 DF,  p-value: 8.208e-10
```

All terms are significant and the adjusted R^2 is getting large, so we try to add an interaction term

```
mod <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage * Bachelor_degree_percentage,
          data = homicide_data)
mod %>% summary
```

```
##
## Call:
```

```
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage *
##     Bachelor_degree_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4948 -1.4327 -0.0895  1.2301  5.7492
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      8.40771    6.68474   1.258
## Poverty_percentage                               0.70320    0.13933   5.047
## Unemployment_percentage                         -3.91610    1.35475  -2.891
## Bachelor_degree_percentage                      -0.41779    0.16617  -2.514
## Unemployment_percentage:Bachelor_degree_percentage  0.14206    0.03561   3.989
##                                                 Pr(>|t|)
## (Intercept)                                     0.214831
## Poverty_percentage                               7.5e-06 ***
## Unemployment_percentage                         0.005850 **
## Bachelor_degree_percentage                      0.015488 *
## Unemployment_percentage:Bachelor_degree_percentage 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.04 on 46 degrees of freedom
## Multiple R-squared:  0.7136, Adjusted R-squared:  0.6887
## F-statistic: 28.66 on 4 and 46 DF,  p-value: 5.627e-12
```

Adding the interaction between unemployment and bachelor degree raises adjusted R^2 and brings into no insignificant terms, so we keep it. Then add another interaction term

```
mod <- lm(Homicide_rate ~ Poverty_percentage * Bachelor_degree_percentage + Unemployment_percentage
          * Bachelor_degree_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage * Bachelor_degree_percentage +
##     Unemployment_percentage * Bachelor_degree_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6734 -1.3809 -0.0199  1.0364  5.1065
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      7.34276    6.59735   1.113
## Poverty_percentage                               1.78262    0.67234   2.651
## Bachelor_degree_percentage                      -0.36329    0.16655  -2.181
## Unemployment_percentage                         -6.67769    2.14631  -3.111
## Poverty_percentage:Bachelor_degree_percentage   -0.03313    0.02021  -1.640
## Bachelor_degree_percentage:Unemployment_percentage  0.21908    0.05856   3.741
##                                                 Pr(>|t|)
## (Intercept)                                     0.271624
```

```
## Poverty_percentage                                    0.011030 *
## Bachelor_degree_percentage                            0.034427 *
## Unemployment_percentage                               0.003231 **
## Poverty_percentage:Bachelor_degree_percentage         0.108025
## Bachelor_degree_percentage:Unemployment_percentage 0.000517 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.004 on 45 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.6998
## F-statistic: 24.31 on 5 and 45 DF,  p-value: 9.091e-12
```

It seems the additional interaction term we added between bachelor and poverty rate is insignificant, and it does not change adjusted R^2 much, so we discard it and choose another interaction term.

```
mod <- lm(Homicide_rate ~ Poverty_percentage * Unemployment_percentage + Unemployment_percentage
          * Bachelor_degree_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage * Unemployment_percentage +
##     Unemployment_percentage * Bachelor_degree_percentage, data = homicide_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.873 -1.347 -0.349  1.038  5.785
##
## Coefficients:
##                                                    Estimate Std. Error t value
## (Intercept)                                        24.10409   10.46036   2.304
## Poverty_percentage                                 -0.23620    0.50890  -0.464
## Unemployment_percentage                            -7.65611    2.35562  -3.250
## Bachelor_degree_percentage                         -0.56376    0.17863  -3.156
## Poverty_percentage:Unemployment_percentage          0.22990    0.12005   1.915
## Unemployment_percentage:Bachelor_degree_percentage  0.17280    0.03816   4.528
##                                                    Pr(>|t|)
## (Intercept)                                         0.02587 *
## Poverty_percentage                                  0.64479
## Unemployment_percentage                             0.00219 **
## Bachelor_degree_percentage                          0.00285 **
## Poverty_percentage:Unemployment_percentage          0.06186 .
## Unemployment_percentage:Bachelor_degree_percentage 4.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.984 on 45 degrees of freedom
## Multiple R-squared:  0.7352, Adjusted R-squared:  0.7058
## F-statistic: 24.99 on 5 and 45 DF,  p-value: 5.817e-12
```

After we bring into the interaction term between poverty and unemployment, adjusted R^2 increases, but this interaction term is insignificant.

```r
mod <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage +
            Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentag
##     data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5080 -1.3246 -0.3035  1.3365  6.2877
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                       11.019282
## Poverty_percentage                                                -0.235635
## Unemployment_percentage                                           -1.756269
## Bachelor_degree_percentage                                        -0.198567
## Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentage  0.007283
##                                                                   Std. Error
## (Intercept)                                                         7.220025
## Poverty_percentage                                                  0.299509
## Unemployment_percentage                                             0.849961
## Bachelor_degree_percentage                                          0.112510
## Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentage   0.001804
##                                                                   t value
## (Intercept)                                                         1.526
## Poverty_percentage                                                 -0.787
## Unemployment_percentage                                            -2.066
## Bachelor_degree_percentage                                         -1.765
## Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentage   4.037
##                                                                   Pr(>|t|)
## (Intercept)                                                       0.133804
## Poverty_percentage                                                0.435470
## Unemployment_percentage                                           0.044458
## Bachelor_degree_percentage                                        0.084222
## Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentage 0.000203
##
## (Intercept)
## Poverty_percentage
## Unemployment_percentage                                           *
## Bachelor_degree_percentage                                        .
## Poverty_percentage:Unemployment_percentage:Bachelor_degree_percentage ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.034 on 46 degrees of freedom
## Multiple R-squared:  0.7154, Adjusted R-squared:  0.6907
## F-statistic: 28.91 on 4 and 46 DF,  p-value: 4.89e-12
```

This model decreases adjusted R^2 and increases the number of insignificant terms, so we decide to go back

to the model with poverty, unemployment, bachelor degree, and the interaction between unemployment and bachelor degree. We then add household income.

```
mod <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage +
            Unemployment_percentage:Bachelor_degree_percentage + Median_household_income_1k, data = homi
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Unemployment_percentage:Bachelor_degree_percentage +
##     Median_household_income_1k, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2518 -1.2644 -0.3017  1.2607  5.3532
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       2.60764    7.10137   0.367
## Poverty_percentage                                0.95282    0.18412   5.175
## Unemployment_percentage                          -4.29479    1.32660  -3.237
## Bachelor_degree_percentage                       -0.49522    0.16565  -2.990
## Median_household_income_1k                        0.10021    0.05025   1.994
## Unemployment_percentage:Bachelor_degree_percentage  0.14427    0.03453   4.178
##                                                  Pr(>|t|)
## (Intercept)                                      0.715189
## Poverty_percentage                               5.13e-06 ***
## Unemployment_percentage                          0.002267 **
## Bachelor_degree_percentage                       0.004517 **
## Median_household_income_1k                       0.052215 .
## Unemployment_percentage:Bachelor_degree_percentage 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.977 on 45 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7076
## F-statistic: 25.21 on 5 and 45 DF,  p-value: 5.06e-12
```

It seems the household income is insignificant, but since its p-value does not exceed much from 0.05, and it does raise the adjusted R^2, we decide to keep it can see whether the interaction terms can have any effects.

```
mod <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage +
            Unemployment_percentage:Bachelor_degree_percentage + Median_household_income_1k +
            Median_household_income_1k:Poverty_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Unemployment_percentage:Bachelor_degree_percentage +
##     Median_household_income_1k + Median_household_income_1k:Poverty_percentage,
##     data = homicide_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6984 -1.0769 -0.2342  1.2020  5.4805
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     -0.345898   7.117869  -0.049
## Poverty_percentage                               1.886711   0.546331   3.453
## Unemployment_percentage                         -5.821451   1.544841  -3.768
## Bachelor_degree_percentage                      -0.658418   0.185065  -3.558
## Median_household_income_1k                       0.245106   0.093872   2.611
## Unemployment_percentage:Bachelor_degree_percentage  0.188261   0.041540   4.532
## Poverty_percentage:Median_household_income_1k    -0.015816   0.008738  -1.810
##                                                 Pr(>|t|)
## (Intercept)                                     0.961461
## Poverty_percentage                              0.001236 **
## Unemployment_percentage                         0.000485 ***
## Bachelor_degree_percentage                      0.000910 ***
## Median_household_income_1k                      0.012296 *
## Unemployment_percentage:Bachelor_degree_percentage 4.44e-05 ***
## Poverty_percentage:Median_household_income_1k    0.077124 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.929 on 44 degrees of freedom
## Multiple R-squared:  0.7551, Adjusted R-squared:  0.7217
## F-statistic: 22.61 on 6 and 44 DF,  p-value: 5.839e-12
```

Median_household_income_1k:Poverty_rate is ok though has one insignificant term. We then try to include Median_household_income_1k:Bachelor_degree_rate but no Median_household_income_1k:Poverty_rate. This gives us a higher adjusted R^2 and the terms are relatively significant (though median household income is insignificant, its interaction term is, so we keep it).

```
mod <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage +
            Median_household_income_1k + Unemployment_percentage:Bachelor_degree_percentage +
            Median_household_income_1k:Bachelor_degree_percentage, data = homicide_data)
mod %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Median_household_income_1k +
##     Unemployment_percentage:Bachelor_degree_percentage + Median_household_income_1k:Bachelor_degree_
##     data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9188 -1.1582 -0.3306  0.9887  5.4699
##
## Coefficients:
##                                                   Estimate Std. Error
## (Intercept)                                     26.964009  13.059148
```

```
## Poverty_percentage                                      0.695926   0.212295
## Unemployment_percentage                                -3.945945   1.284067
## Bachelor_degree_percentage                             -1.035677   0.293902
## Median_household_income_1k                             -0.201234   0.146037
## Unemployment_percentage:Bachelor_degree_percentage      0.131269   0.033695
## Bachelor_degree_percentage:Median_household_income_1k   0.007501   0.003430
##                                                        t value Pr(>|t|)
## (Intercept)                                              2.065 0.044870 *
## Poverty_percentage                                       3.278 0.002046 **
## Unemployment_percentage                                 -3.073 0.003630 **
## Bachelor_degree_percentage                              -3.524 0.001006 **
## Median_household_income_1k                              -1.378 0.175182
## Unemployment_percentage:Bachelor_degree_percentage       3.896 0.000329 ***
## Bachelor_degree_percentage:Median_household_income_1k    2.187 0.034097 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.899 on 44 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:  0.7303
## F-statistic: 23.57 on 6 and 44 DF,  p-value: 2.983e-12
```

$$\text{Homicide rate} = 26.964 + 0.696 * \text{Poverty \%} - 3.946 * \text{Unemployment \%} - 1.036 * \text{Bachelor degree \%} - 0.201 * \text{Median household income (1k)} + 0.131 * \text{Unemployment \%} * \text{Bachelor degree \%} + 0.008 * \text{Bachelor degree \%} * \text{Median household income (1k)}$$

Conduct t-test to see the significance of each predictor

```
mod0 <- lm(Homicide_rate ~ Gini_Index + Uninsured_percentage + Median_household_income_1k +
             Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage, data = homicide_

mod1 <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage,
           data = homicide_data)

mod0 %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Gini_Index + Uninsured_percentage +
##     Median_household_income_1k + Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8143 -1.4156 -0.6073  1.5930  5.3091
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -11.63668    9.43152  -1.234  0.22382
## Gini_Index                 -40.43014   25.78895  -1.568  0.12411
## Uninsured_percentage         0.19559    0.13048   1.499  0.14103
## Median_household_income_1k   0.10174    0.05774   1.762  0.08504 .
```

```
## Poverty_percentage           1.27864    0.26481   4.829  1.7e-05 ***
## Unemployment_percentage      1.06309    0.50576   2.102  0.04131 *
## Bachelor_degree_percentage   0.25183    0.07359   3.422  0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.248 on 44 degrees of freedom
## Multiple R-squared:  0.6675, Adjusted R-squared:  0.6221
## F-statistic: 14.72 on 6 and 44 DF,  p-value: 3.878e-09
```

```
mod1 %>% summary
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4283 -1.5793 -0.3139  1.3508  6.0571
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -16.49180    2.74540  -6.007 2.63e-07 ***
## Poverty_percentage           0.84511    0.15461   5.466 1.71e-06 ***
## Unemployment_percentage      1.21101    0.49137   2.465   0.0174 *
## Bachelor_degree_percentage   0.22445    0.04716   4.759 1.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.342 on 47 degrees of freedom
## Multiple R-squared:  0.6146, Adjusted R-squared:   0.59
## F-statistic: 24.98 on 3 and 47 DF,  p-value: 8.208e-10
```

Checking the multicollinearity of the predictors in our final model, none of the VIF values are greater than 5.

```
vif(lm(Homicide_rate ~ Median_household_income_1k + Poverty_percentage + Unemployment_percentage +
         Bachelor_degree_percentage, data = homicide_data))
```

```
## Median_household_income_1k         Poverty_percentage
##                   4.427291                   3.621155
##    Unemployment_percentage Bachelor_degree_percentage
##                   1.649832                   2.315253
```

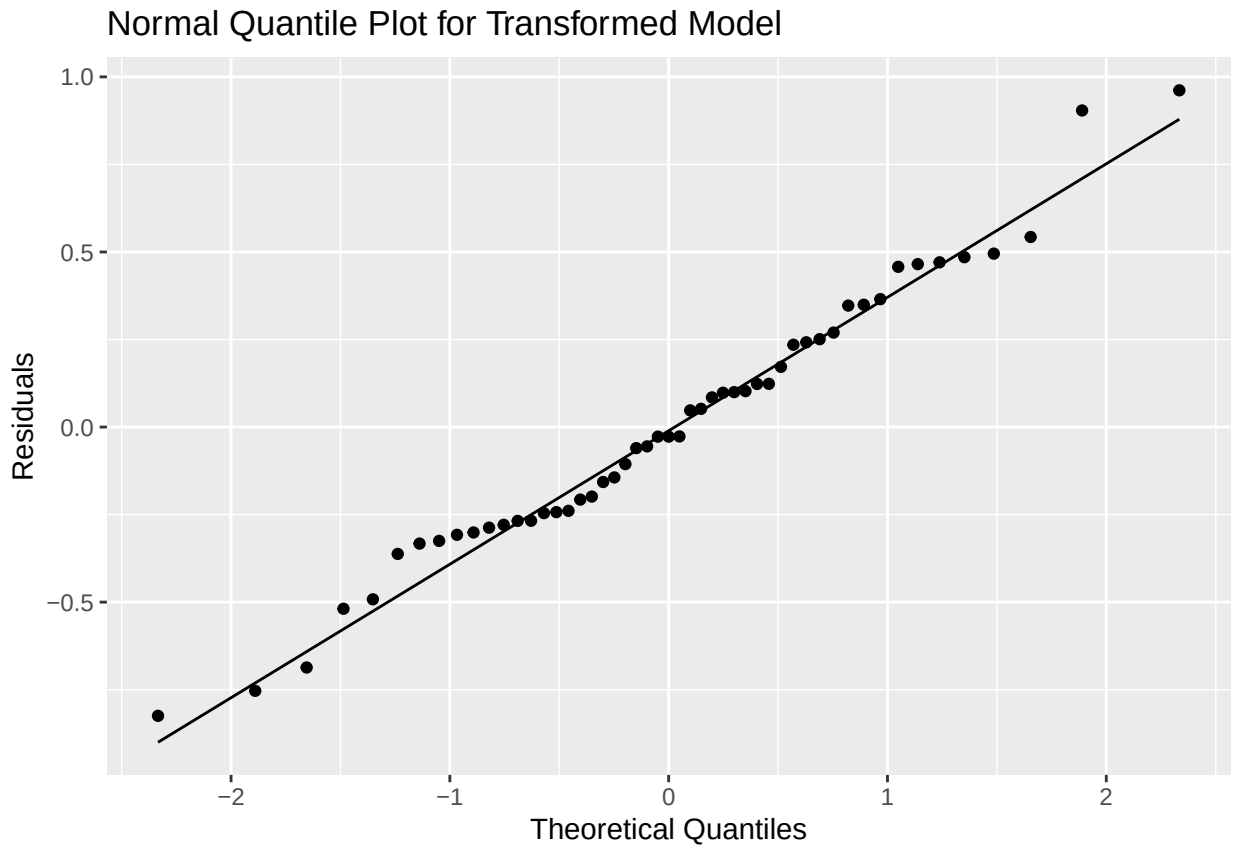
Assessing the Model

We see that the residual plot for the final model shows non-linearity and non-constant variance. We tested multiple different transformations. Below is a log transformation, which seems to marginally improve the non-linearity but not the non-constant variance. The quantile plots look similar.

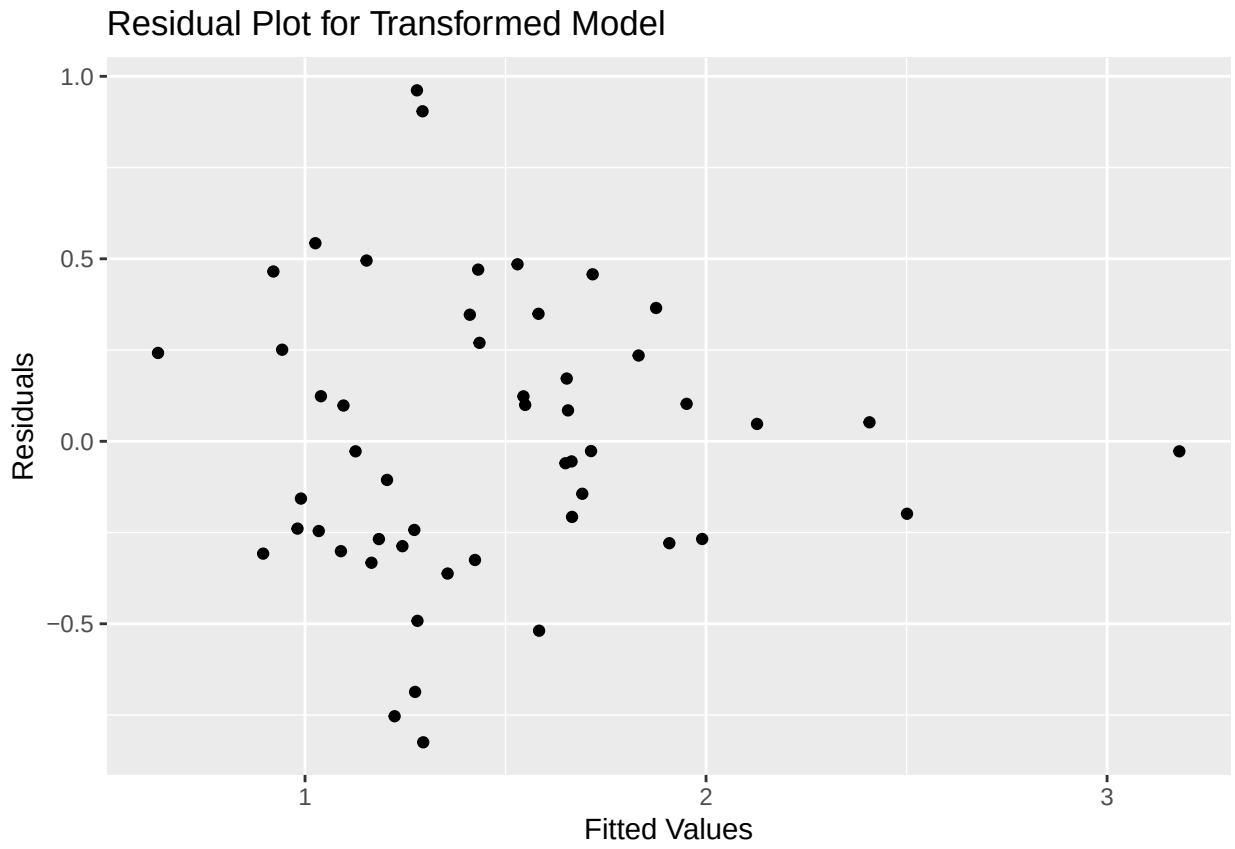
```r
mod_t <- lm(log(Homicide_rate) ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percenta
            Median_household_income_1k + Unemployment_percentage:Bachelor_degree_percentage +
            Median_household_income_1k:Bachelor_degree_percentage, data = homicide_data)
summary(mod_t)
```

```
##
## Call:
## lm(formula = log(Homicide_rate) ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Median_household_income_1k +
##     Unemployment_percentage:Bachelor_degree_percentage + Median_household_income_1k:Bachelor_degree_
##     data = homicide_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82471 -0.26775 -0.02743  0.24650  0.96159
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       1.5562263  2.7976093
## Poverty_percentage                                0.1293544  0.0454792
## Unemployment_percentage                          -0.2967074  0.2750805
## Bachelor_degree_percentage                       -0.0821703  0.0629615
## Median_household_income_1k                       -0.0013268  0.0312850
## Unemployment_percentage:Bachelor_degree_percentage 0.0128666  0.0072183
## Bachelor_degree_percentage:Median_household_income_1k 0.0003789  0.0007347
##                                                   t value Pr(>|t|)
## (Intercept)                                         0.556  0.58084
## Poverty_percentage                                  2.844  0.00673 **
## Unemployment_percentage                            -1.079  0.28663
## Bachelor_degree_percentage                         -1.305  0.19865
## Median_household_income_1k                         -0.042  0.96636
## Unemployment_percentage:Bachelor_degree_percentage  1.783  0.08157 .
## Bachelor_degree_percentage:Median_household_income_1k  0.516  0.60862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4069 on 44 degrees of freedom
## Multiple R-squared:  0.5885, Adjusted R-squared:  0.5324
## F-statistic: 10.49 on 6 and 44 DF,  p-value: 3.328e-07
```

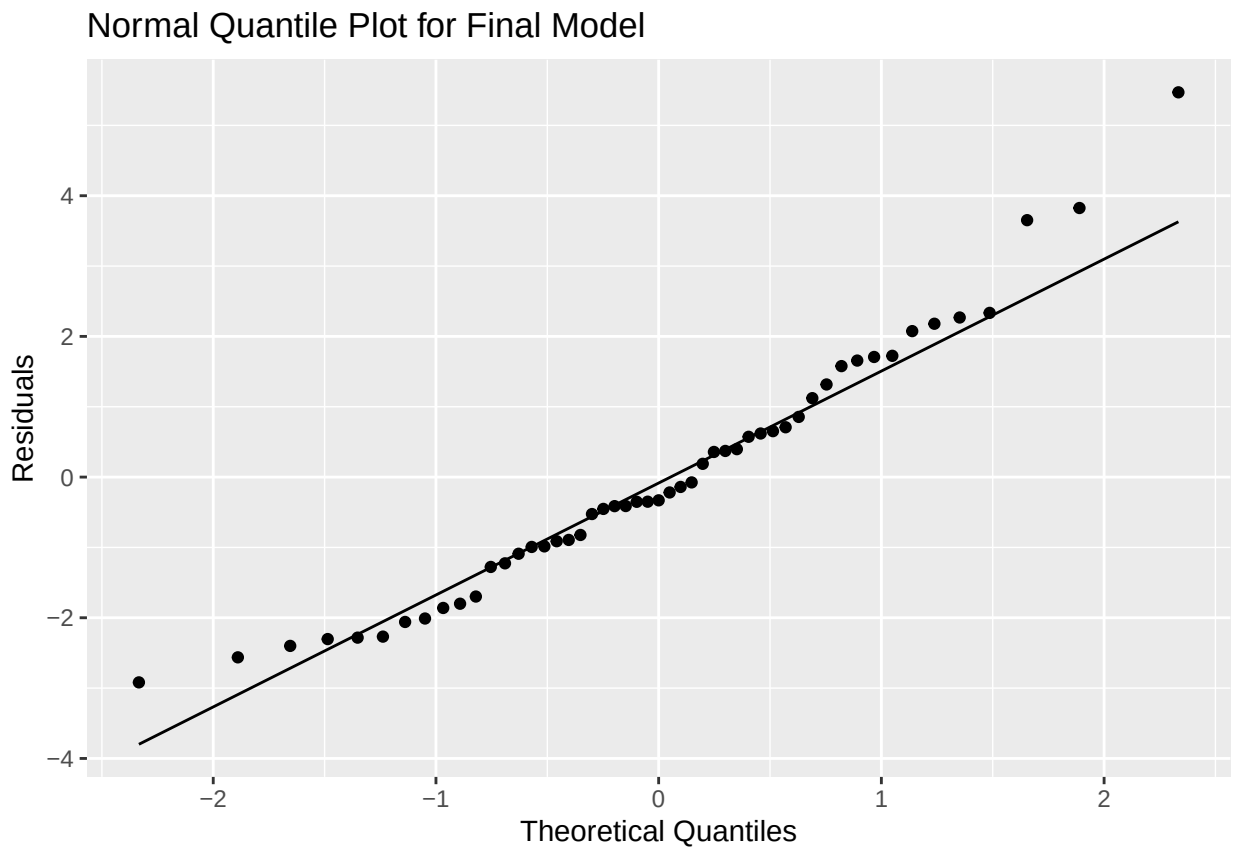
```r
homicide_data %>%
  ggplot(aes(sample = mod_t$residuals)) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("Normal Quantile Plot for Transformed Model") +
  xlab("Theoretical Quantiles") +
  ylab("Residuals")
```



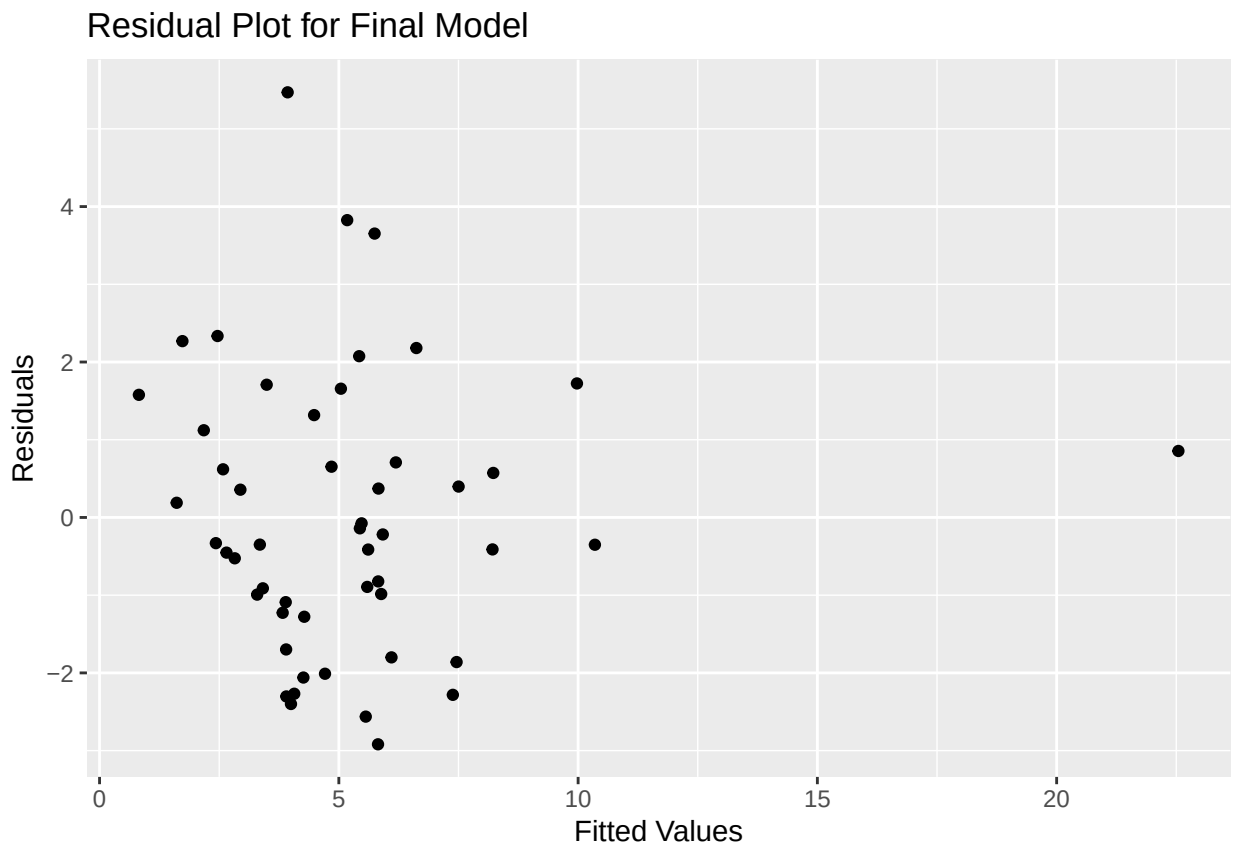

```
homicide_data %>%
  ggplot(aes(x = mod_t$fitted.values, y = mod_t$residuals)) +
  geom_point() +
  ggtitle("Residual Plot for Transformed Model") +
  xlab("Fitted Values") +
  ylab("Residuals")
```



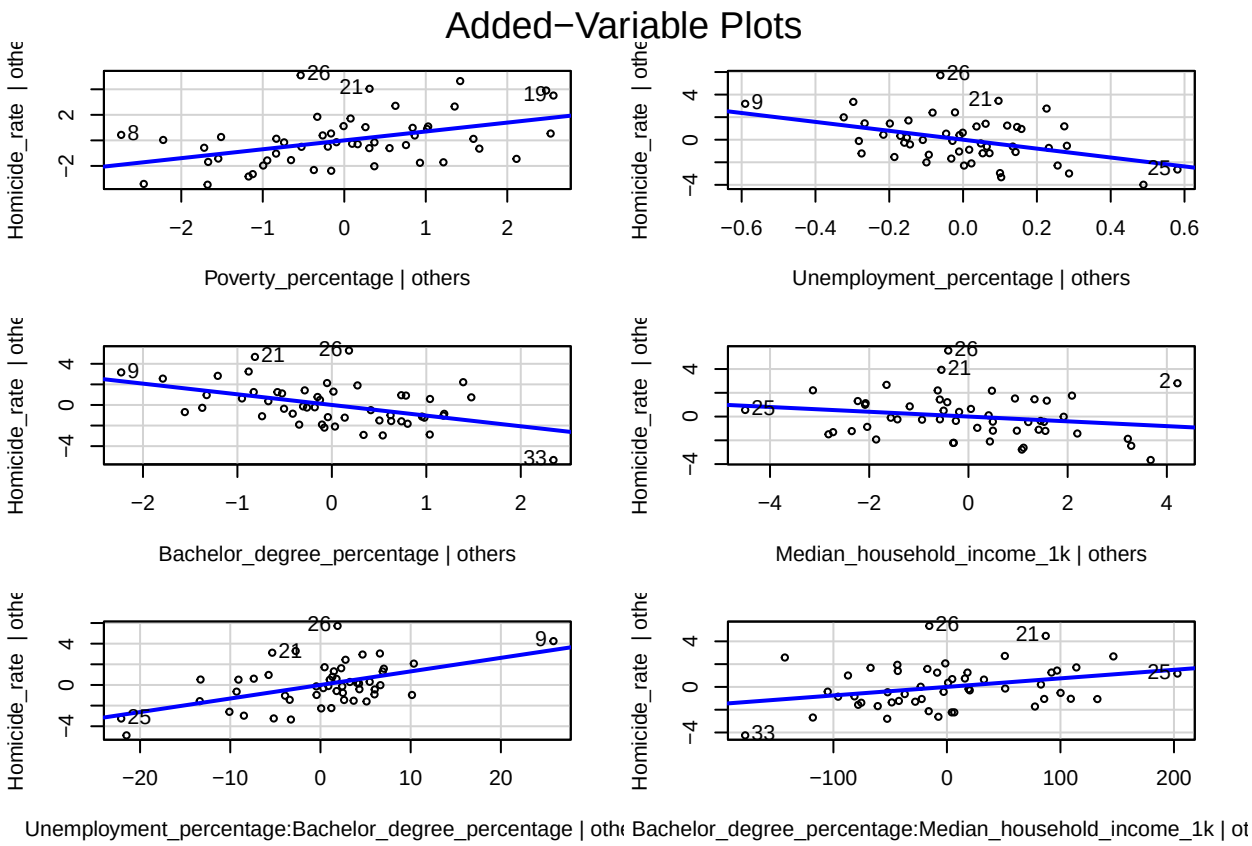

```
homicide_data %>%
  ggplot(aes(sample = mod$residuals)) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("Normal Quantile Plot for Final Model") +
  xlab("Theoretical Quantiles") +
  ylab("Residuals")
```




```
homicide_data %>%
  ggplot(aes(x = mod$fitted.values, y = mod$residuals)) +
  geom_point() +
  ggtitle("Residual Plot for Final Model") +
  xlab("Fitted Values") +
  ylab("Residuals")
```


Residual Plot for Final Model
Residuals
4
2
0
−2
0
5
10
15
20
Fitted Values

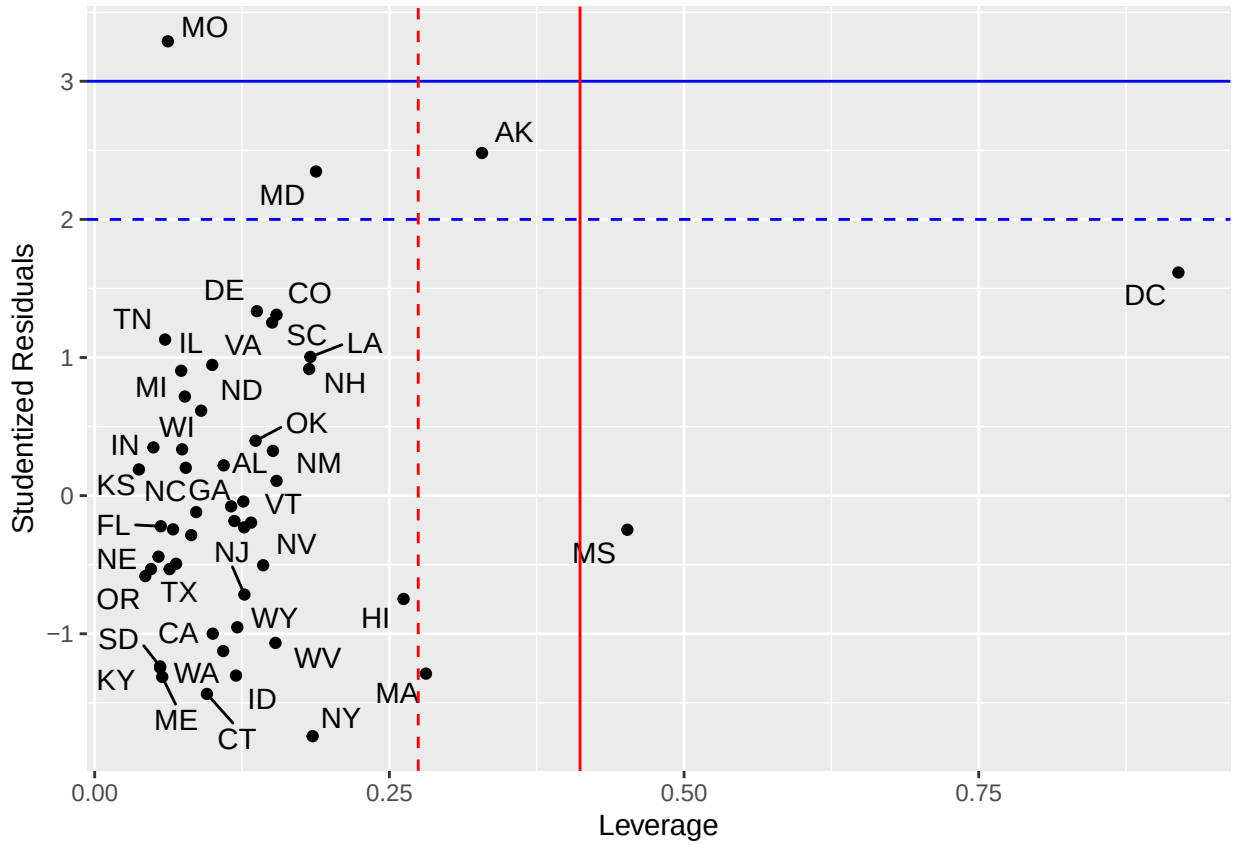
```
avPlots(mod)
```



Plot Resids, Leverage

D.C. has **extremely** unusual leverage, so this suggests we might try to fit a model without D.C., as it is heavily influencing the regression line.

```
homicide_data %>%
  ggplot(aes(x = hatvalues(mod), y = rstudent(mod))) +
  geom_point() +
  geom_text_repel(aes(label = Abbreviation)) +
  geom_hline(yintercept = 2, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 3, color = "blue") +
  geom_vline(xintercept = 2*7/51, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 3*7/51, color = "red") +
  xlab("Leverage") +
  ylab("Studentized Residuals")
```



Fit model without DC

We go through the same process as before. First checking the original final model without D.C., we see that our R^2 values have decreased drastically, and none of the predictors except for `Poverty_percentage` are significant anymore.

```
homicide_data_noDC <- homicide_data[-c(9),]
mod_noDC <- lm(Homicide_rate ~ Poverty_percentage + Unemployment_percentage + Bachelor_degree_percentage +
                Median_household_income_1k + Unemployment_percentage:Bachelor_degree_percentage +
                Median_household_income_1k:Bachelor_degree_percentage, data = homicide_data_noDC)
summary(mod_noDC)
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Median_household_income_1k +
##     Unemployment_percentage:Bachelor_degree_percentage + Median_household_income_1k:Bachelor_degree_p
##     data = homicide_data_noDC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6232 -1.3449 -0.3769  1.1109  5.3667
##
## Coefficients:
```

```
##                                                    Estimate Std. Error
## (Intercept)                                        7.933458  17.415134
## Poverty_percentage                                 0.661416   0.209606
## Unemployment_percentage                           -1.080861   2.176203
## Bachelor_degree_percentage                        -0.468625   0.454459
## Median_household_income_1k                        -0.061189   0.167597
## Unemployment_percentage:Bachelor_degree_percentage 0.044951   0.062850
## Bachelor_degree_percentage:Median_household_income_1k 0.003553 0.004162
##                                                   t value Pr(>|t|)
## (Intercept)                                         0.456  0.65101
## Poverty_percentage                                  3.156  0.00292 **
## Unemployment_percentage                            -0.497  0.62195
## Bachelor_degree_percentage                         -1.031  0.30823
## Median_household_income_1k                         -0.365  0.71683
## Unemployment_percentage:Bachelor_degree_percentage  0.715  0.47835
## Bachelor_degree_percentage:Median_household_income_1k 0.854 0.39792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.865 on 43 degrees of freedom
## Multiple R-squared:  0.5474, Adjusted R-squared:  0.4843
## F-statistic: 8.668 on 6 and 43 DF,  p-value: 3.376e-06
```

Looking at the best subsets for Mallow's C_p and adjusted R^2, no model appears to be able to exceed an adjusted R^2 of 0.54, which is significantly lower than the models with D.C. Likewise, none of the models from the Mallow's C_p output is below $m + 1$, where m is the number of predictor terms in the proposed model. This suggests that none of the models do very well in balancing simplicity and reducing SSE.

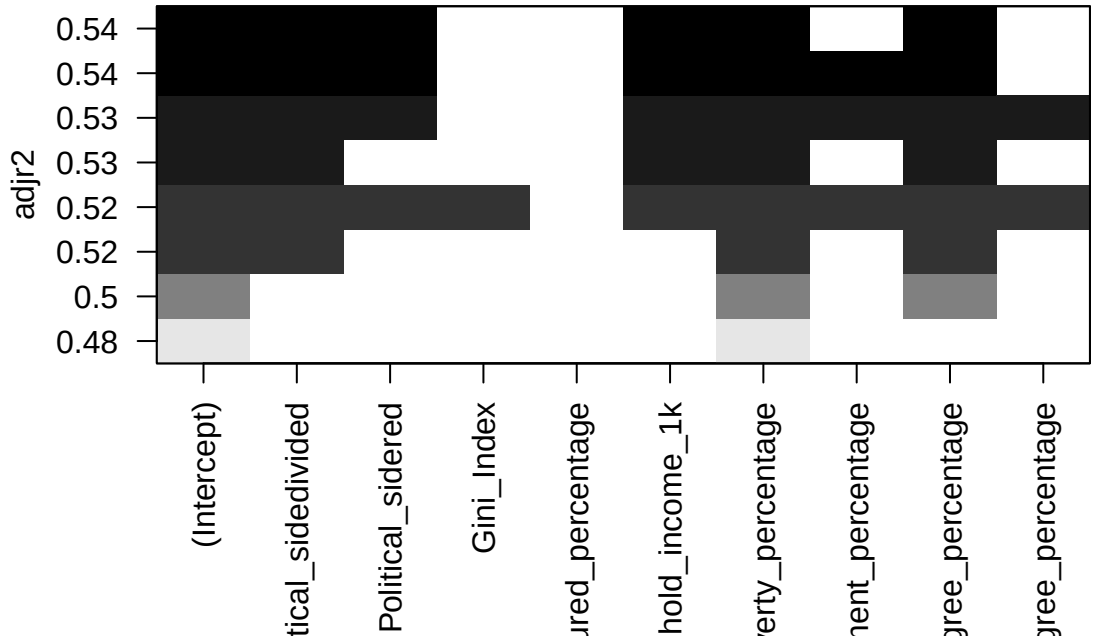
```
homicide_data_noDCstate <- homicide_data_noDC[, -c(1:2)]
all_noDC <- regsubsets(Homicide_rate ~ ., data = homicide_data_noDCstate, really.big=T)

summary(all_noDC)$adjr2
```

```
## [1] 0.4750989 0.5036968 0.5238541 0.5321018 0.5414628 0.5411607 0.5339160
## [8] 0.5244021
```

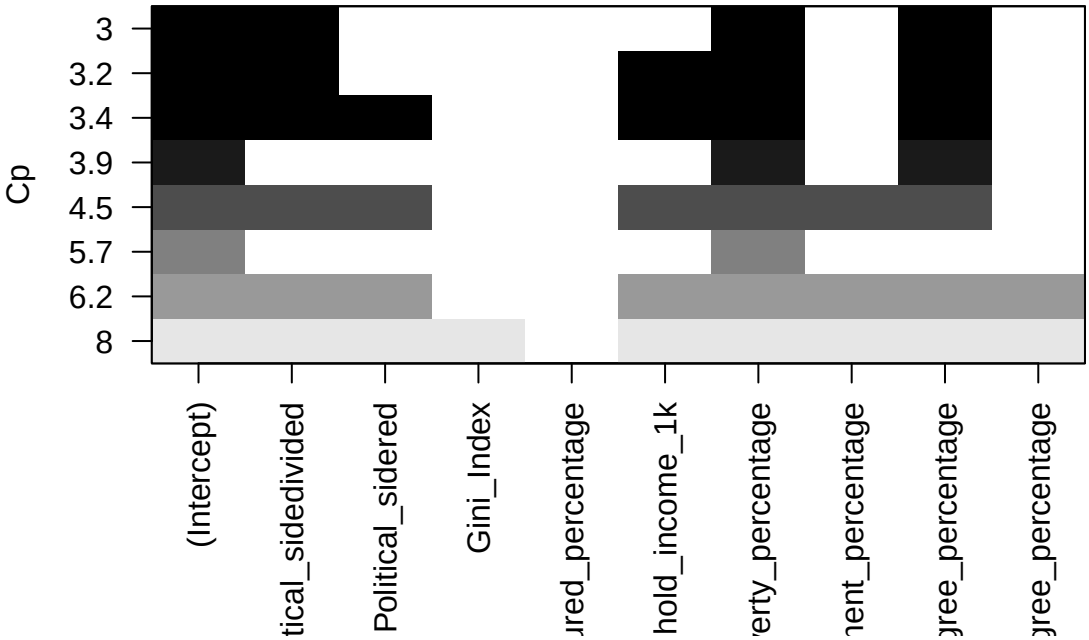
```
summary(all_noDC)$cp
```

```
## [1] 5.738214 3.900235 2.977013 3.237106 3.430478 4.515550 6.198162 8.042068
```

```
plot(all_noDC, scale = "adjr2")
```

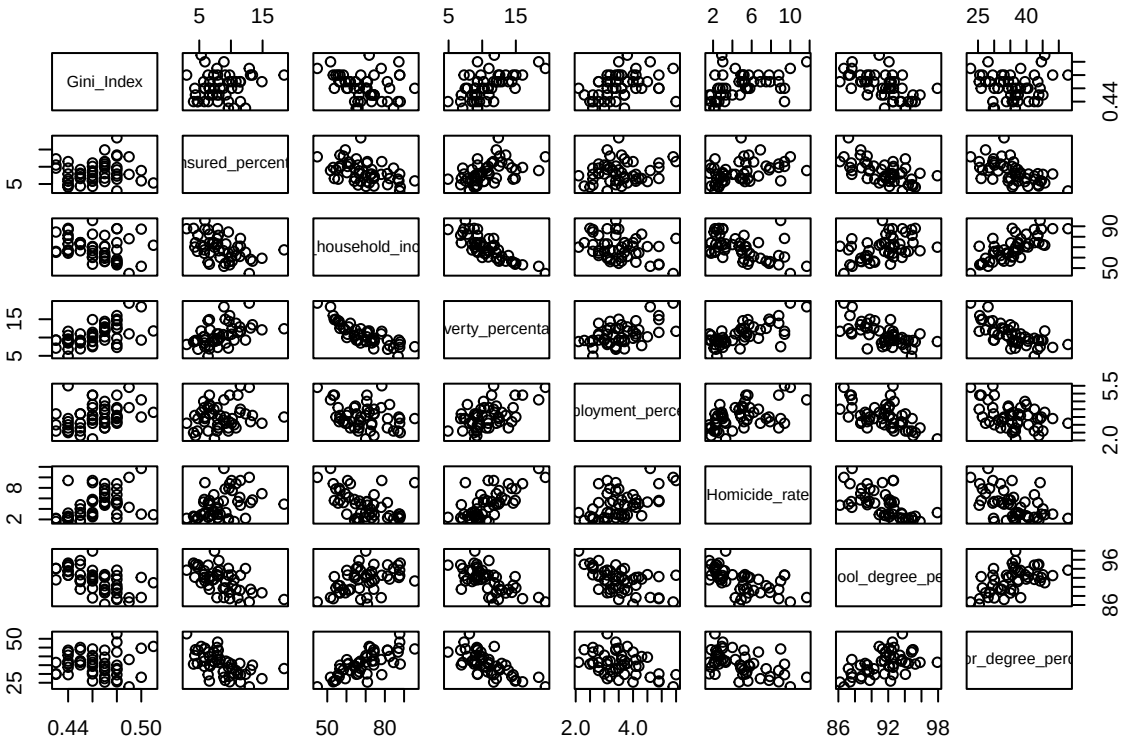



```
plot(all_noDC, scale = "Cp")
```



From the scatterplot matrix, we see that the outlying point on many of the scatterplots from before is gone. `Poverty_percentage` still has the highest correlation with `Homicide_rate` with `High_school_degree_percentage` with the second highest correlation.

```
pairs(homicide_data_noDC[,c(4:11)])
```




```
cor(homicide_data_noDC[,c(4:11)])
```

```
##                                  Gini_Index Uninsured_percentage
## Gini_Index                        1.00000000            0.1265366
## Uninsured_percentage              0.12653659            1.0000000
## Median_household_income_1k       -0.25663036           -0.4668964
## Poverty_percentage                0.56515082            0.4627442
## Unemployment_percentage           0.38109383            0.1823953
## Homicide_rate                     0.45717750            0.3914074
## High_school_degree_percentage    -0.57281724           -0.5810066
## Bachelor_degree_percentage       -0.01249957           -0.5892923
##                                  Median_household_income_1k Poverty_percentage
## Gini_Index                                       -0.2566304          0.5651508
## Uninsured_percentage                             -0.4668964          0.4627442
## Median_household_income_1k                        1.0000000         -0.8225052
## Poverty_percentage                               -0.8225052          1.0000000
## Unemployment_percentage                          -0.3181445          0.5318685
## Homicide_rate                                    -0.4739914          0.6970015
## High_school_degree_percentage                     0.4497290         -0.6432887
## Bachelor_degree_percentage                        0.7453458         -0.7079881
##                                  Unemployment_percentage Homicide_rate
## Gini_Index                                     0.3810938     0.4571775
## Uninsured_percentage                           0.1823953     0.3914074
## Median_household_income_1k                    -0.3181445    -0.4739914
## Poverty_percentage                             0.5318685     0.6970015
```

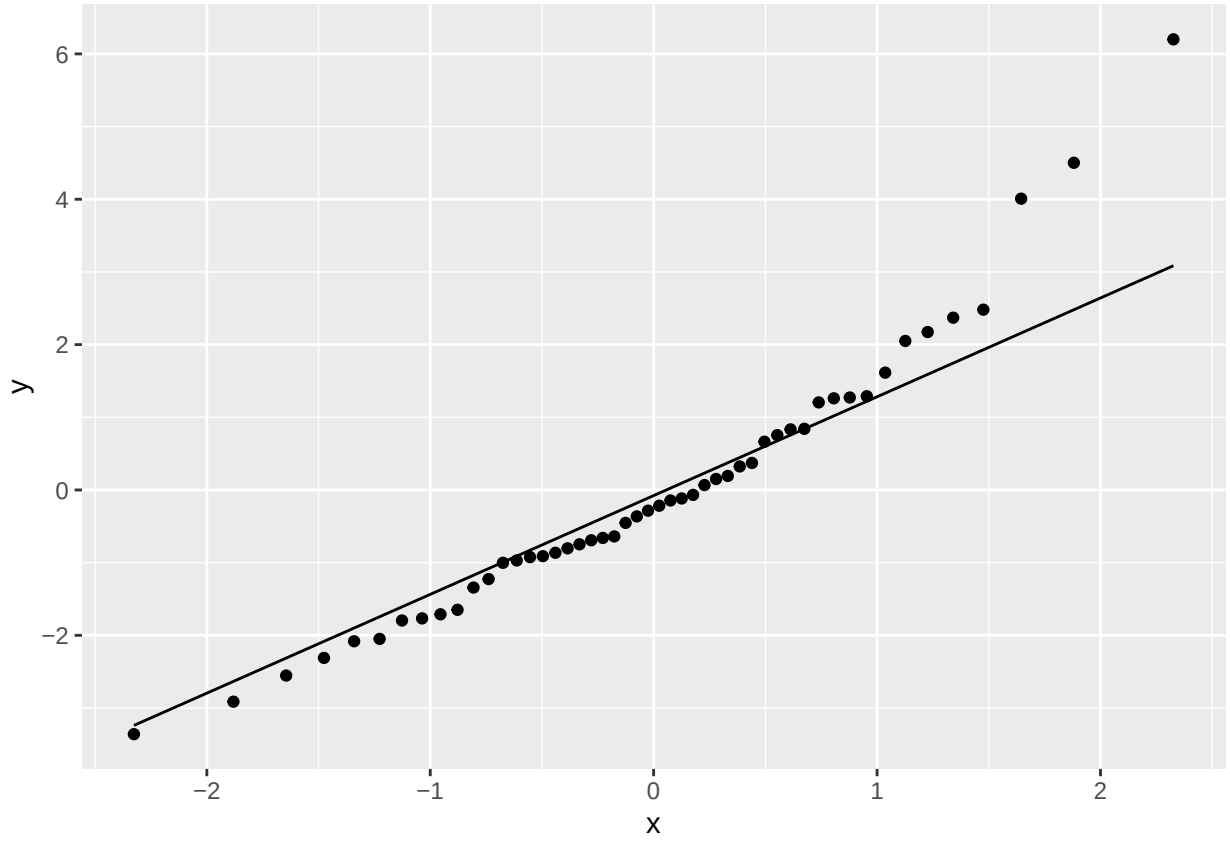
```
## Unemployment_percentage                     1.0000000     0.5119328
## Homicide_rate                               0.5119328     1.0000000
## High_school_degree_percentage              -0.5422082    -0.5979011
## Bachelor_degree_percentage                 -0.4288260    -0.5080477
##                               High_school_degree_percentage
## Gini_Index                                       -0.5728172
## Uninsured_percentage                             -0.5810066
## Median_household_income_1k                        0.4497290
## Poverty_percentage                               -0.6432887
## Unemployment_percentage                          -0.5422082
## Homicide_rate                                    -0.5979011
## High_school_degree_percentage                     1.0000000
## Bachelor_degree_percentage                        0.5332133
##                               Bachelor_degree_percentage
## Gini_Index                                    -0.01249957
## Uninsured_percentage                          -0.58929233
## Median_household_income_1k                     0.74534584
## Poverty_percentage                            -0.70798809
## Unemployment_percentage                       -0.42882597
## Homicide_rate                                 -0.50804773
## High_school_degree_percentage                  0.53321327
## Bachelor_degree_percentage                     1.00000000
```

Adding `Poverty_percentage` to a model, we see that it already explains almost 50% of the variability in the response. However, the residual plot still seems to show non-linearity with some right skew evidenced in the quantile plot. Trying to add each of the other predictors as a possibel second term with `Poverty_percentage`, no other predictor is significantly different from 0 when `Poverty_percentage` is in the model.

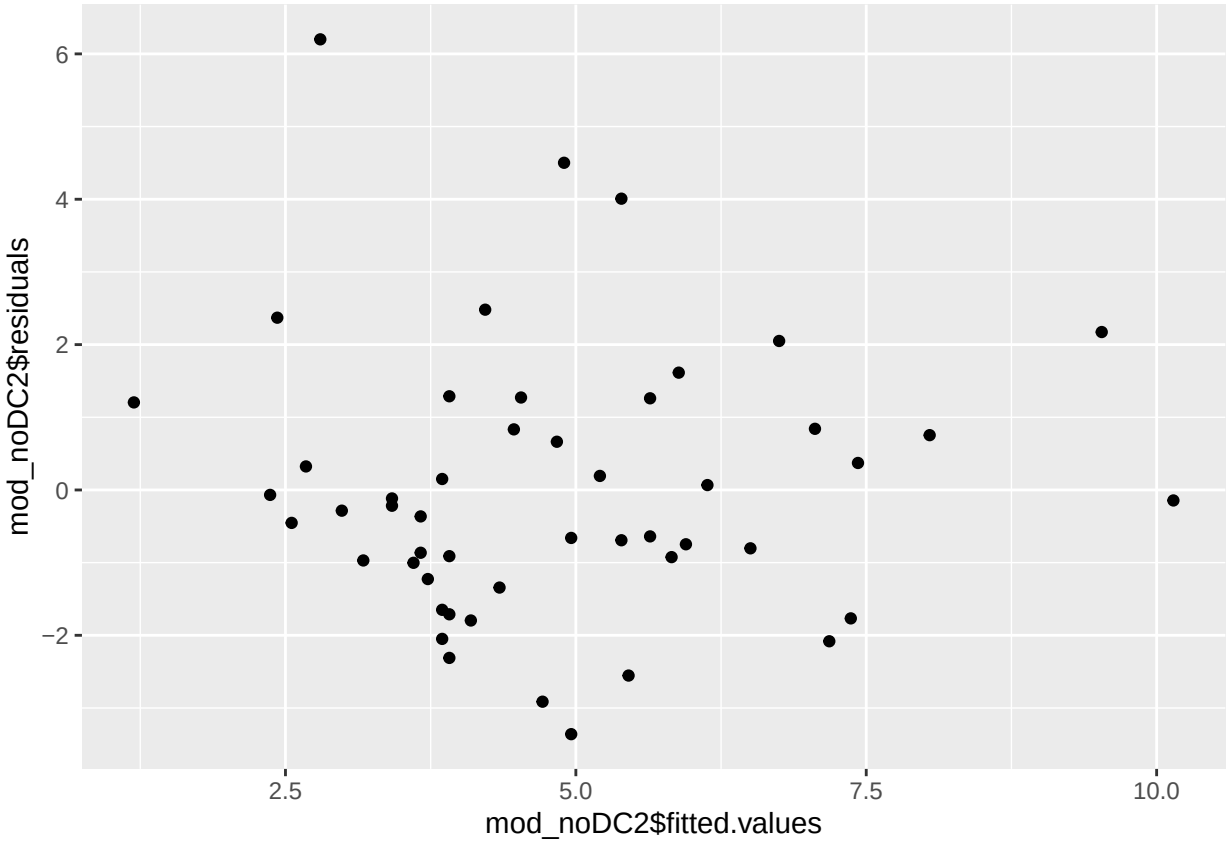
```
mod_noDC2 <- lm(Homicide_rate ~ Poverty_percentage, data = homicide_data_noDC)
summary(mod_noDC2)
```

```
##
## Call:
## lm(formula = Homicide_rate ~ Poverty_percentage, data = homicide_data_noDC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3601 -0.9943 -0.2511  0.8395  6.2001
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.82895    1.02371  -1.787   0.0803 .
## Poverty_percentage  0.61718    0.09165   6.734 1.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.882 on 48 degrees of freedom
## Multiple R-squared:  0.4858, Adjusted R-squared:  0.4751
## F-statistic: 45.35 on 1 and 48 DF,  p-value: 1.879e-08
```

```
homicide_data_noDC %>%
  ggplot(aes(sample = mod_noDC2$residuals)) +
  geom_qq() +
  geom_qq_line()
```



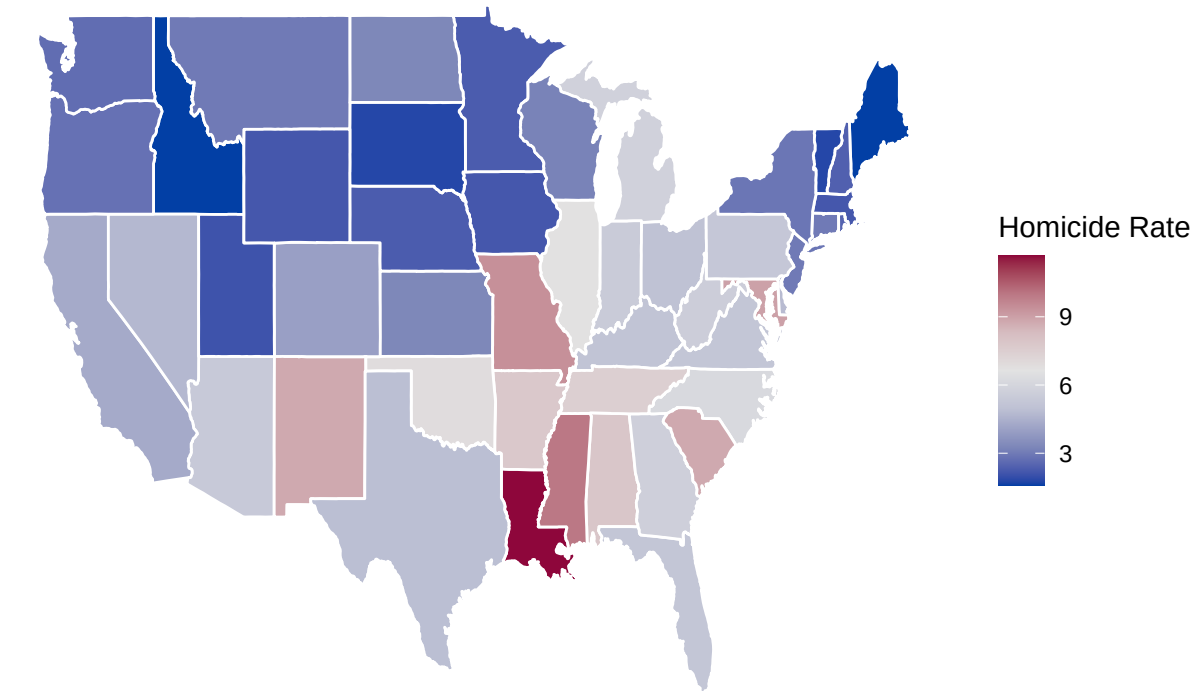

```
homicide_data_noDC %>%
  ggplot(aes(x = mod_noDC2$fitted.values, y = mod_noDC2$residuals)) +
  geom_point()
```



Visualization of Homicide Rate

Visualizing the response variable for a nice graphic on our poster. Done without D.C. so the variation in colors can be seen a bit more easily. D.C. has the highest homicide rate of 23.4 out of 100k people (nearly two times bigger than the next highest value).

```
homicide <- homicide_data_noDC %>%
  mutate("Homicide Rate" = Homicide_rate)
homicide$region <- tolower(homicide_data_noDC$State)

states_map <- map_data("state")
homicide_map <- left_join(states_map, homicide, by = "region")

ggplot(homicide_map, aes(long, lat, group = group)) +
  geom_polygon(aes(fill = `Homicide Rate`), color = "white") +
  scale_fill_gradientn(colours = colorspace::diverge_hcl(7)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), panel.background = eleme
  ggtitle("Homicide Rate per 100,000 in 2019")
```



Cross Validation

Trying out the original final model with 2020 data, we see a shrinkage in R^2 of 55.8%. According to the textbook, a model with $> 50\%$ shrinkage is concerning.

```
homicide_2020 <- read_excel("group1_data.xlsx", sheet = "2020")

homicide_2020$HomicideHat <- predict(mod, homicide_2020)
homicide_2020$Residuals <- homicide_2020$Homicide_rate - homicide_2020$HomicideHat

SSE <- sum(homicide_2020$Residuals^2)
MSEcv <- SSE/51
MSE <- anova(mod)$"Mean Sq"[7]
c(MSEcv, MSE)
```

```
## [1] 31.68860  3.60703
```

```
crossR <- cor(homicide_2020$Homicide_rate, homicide_2020$HomicideHat)
crossR2 <- crossR^2
shrinkage <- summary(mod)$r.squared-crossR2
c(crossR, crossR2, shrinkage)
```

```
## [1] 0.4519719 0.2042786 0.5584015
```

References

"2016 Presidential Election Results." 2017. *Nytimes.comv*. https://www.nytimes.com/elections/2016/results/president.

"Bachelor's Degree Holders Among Individuals 25–44 Years Old | State Indicators | National Science Foundation - State Indicators." n.d. Accessed April 3, 2022. https://ncses.nsf.gov/indicators/states/indicator/bachelors-degree-holders-per-25-44-year-olds.

Bell, Brian, Rui Costa, and Stephen Machin. 2018. "Why Does Education Reduce Crime?" Center for Economic Policy Research. https://cepr.org/active/publications/discussion_papers/dp.php?dpno=13162.

Bureau, US Census. n.d. "Income and Poverty in the United States: 2019." *Census.gov*. Accessed April 3, 2022. https://www.census.gov/library/publications/2020/demo/p60-270.html.

"Gap Between Rich and Poor, by State in the U.S. 2019." n.d. *Statista*. Accessed April 3, 2022. https://www.statista.com/statistics/227249/greatest-gap-between-rich-and-poor-by-us-state/.

"Individuals with High School or Higher Level Degree Among 25–44-Year-Old Population | State Indicators | National Science Foundation - State Indicators." n.d. Accessed April 3, 2022. https://ncses.nsf.gov/indicators/states/indicator/hs-graduates-per-25-44-year-olds.

Kort-Butler, Lisa. 2018. "Social Support Theory." In *The Encyclopedia of Juvenile Delinquency and Justice*, edited by Christopher Schreck, 819–23. Wiley-Blackwell. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1783&context=sociologyfacpub#:~:text=In%20general%2C%20the%20more%20social,is%20the%20risk%20for%20delinquency.

Majumder, Maimuna. 2017. "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/.

"Release Tables: Median Household Income by State, Annual | FRED | St. Louis Fed." 2019. https://fred.stlouisfed.org/release/tables?rid=249&eid=259462&od=2019-01-01#.

"Unemployment Rates for States." n.d. Accessed April 3, 2022. https://www.bls.gov/lau/lastrk19.htm.