Methods

Data Cleaning

```
hate_df =
  read_csv("./data/HateCrimes.csv") %>%
  mutate(
    state = as.factor(state),
    unemployment = as.factor(unemployment),
    urbanization = as.factor(urbanization),
    hate_crimes_per_100k_splc = as.numeric(hate_crimes_per_100k_splc)
  )
```

Descriptive Statistics

```
# Table labels
my_labels =
  list(
    unemployment = "Unemployment",
    urbanization = "Urbanization",
    median_household_income = "Median Household Income",
    perc_population_with_high_school_degree = "Percent with HS Degree",
    perc_non_citizen = "Percent Non-Citizen",
    gini_index = "Gini Index",
    perc_non_white = "Percent Non-White",
    hate_crimes_per_100k_splc = "Hate Crimes per 100k"
)

# Table controls
my_controls = tableby.control(
  total = F,
  test = F,
  numeric.stats = c("N", "meansd", "medianq1q3", "range", "Nmiss2"),
  cat.stats = c("N", "countpct"),
  stats.labels = list(
    meansd = "Mean (SD)",
    medianq1q3 = "Median (Q1, Q3)",
    range = "Min - Max",
    Nmiss2 = "Missing",
    countpct = "N (%)",
    N = "N"
    )
  )
```

```r
# Generate table
descriptive_tab =
  tableby( ~ unemployment +
           urbanization +
           median_household_income +
           perc_population_with_high_school_degree +
           perc_non_citizen +
           gini_index +
           perc_non_white +
           hate_crimes_per_100k_splc,
           data = hate_df,
           control = my_controls)

summary(
  descriptive_tab,
  title = "Descriptive Statistics: Hate Crimes Data",
  labelTranslations = my_labels,
  text = T)
```

```
##
## Table: Descriptive Statistics: Hate Crimes Data
##
## |                         |          Overall (N=51)          |
## |:------------------------|:--------------------------------:|
## |Unemployment             |                                  |
## |-  N                     |                51                |
## |-  high                  |           24 (47.1%)             |
## |-  low                   |           27 (52.9%)             |
## |Urbanization             |                                  |
## |-  N                     |                51                |
## |-  high                  |           24 (47.1%)             |
## |-  low                   |           27 (52.9%)             |
## |Median Household Income  |                                  |
## |-  N                     |                51                |
## |-  Mean (SD)             |       55223.608 (9208.478)       |
## |-  Median (Q1, Q3)       | 54916.000 (48657.000, 60719.000) |
## |-  Min - Max             |      35521.000 - 76165.000       |
## |-  Missing               |                0                 |
## |Percent with HS Degree   |                                  |
## |-  N                     |                51                |
## |-  Mean (SD)             |          0.869 (0.034)           |
## |-  Median (Q1, Q3)       |       0.874 (0.841, 0.898)       |
## |-  Min - Max             |          0.799 - 0.918           |
## |-  Missing               |                0                 |
## |Percent Non-Citizen      |                                  |
## |-  N                     |                48                |
## |-  Mean (SD)             |          0.055 (0.031)           |
## |-  Median (Q1, Q3)       |       0.045 (0.030, 0.080)       |
## |-  Min - Max             |          0.010 - 0.130           |
## |-  Missing               |                3                 |
## |Gini Index               |                                  |
## |-  N                     |                51                |
## |-  Mean (SD)             |          0.454 (0.021)           |
```

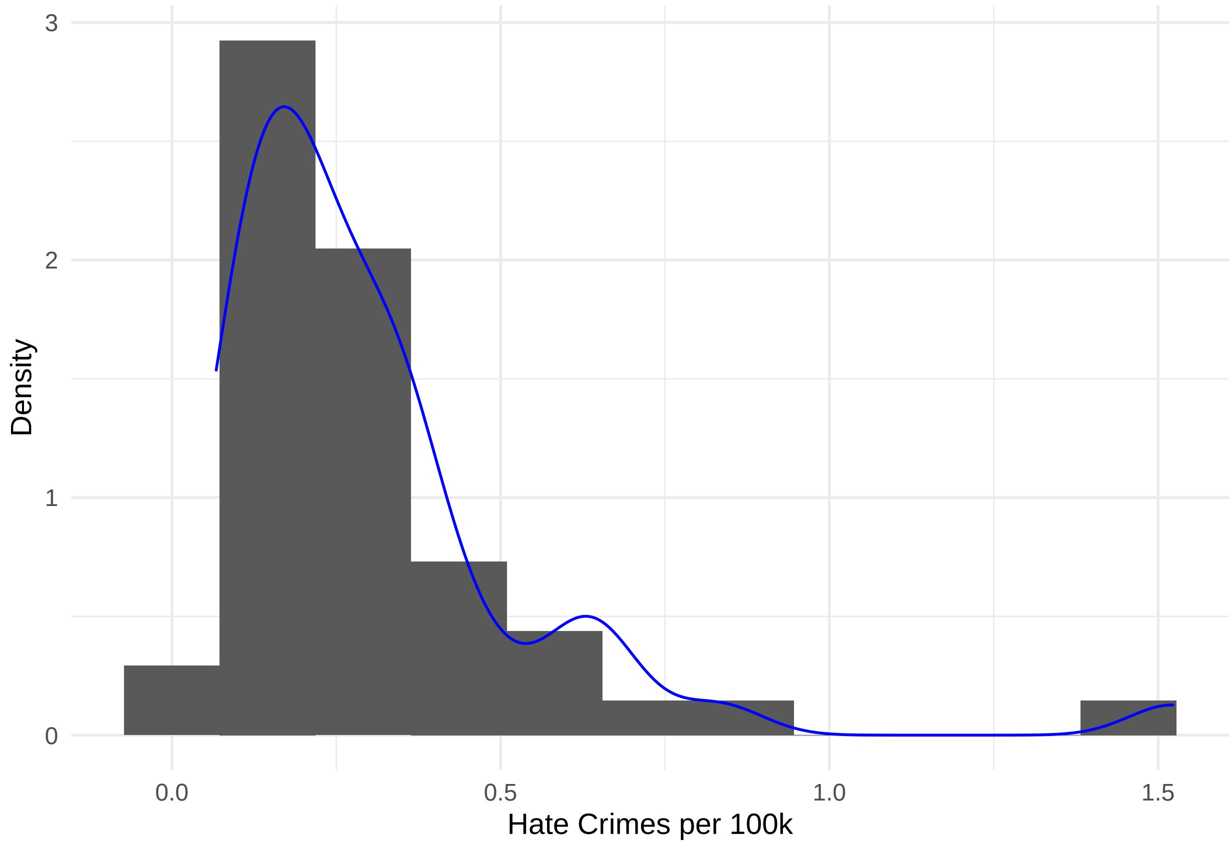
```
## |-  Median (Q1, Q3)      |       0.454 (0.440, 0.467)       |
## |-  Min - Max            |          0.419 - 0.532           |
## |-  Missing              |                0                 |
## |Percent Non-White       |                                  |
## |-  N                    |                51                |
## |-  Mean (SD)            |          0.316 (0.165)           |
## |-  Median (Q1, Q3)      |       0.280 (0.195, 0.420)       |
## |-  Min - Max            |          0.060 - 0.810           |
## |-  Missing              |                0                 |
## |Hate Crimes per 100k    |                                  |
## |-  N                    |                47                |
## |-  Mean (SD)            |          0.304 (0.253)           |
## |-  Median (Q1, Q3)      |       0.226 (0.143, 0.357)       |
## |-  Min - Max            |          0.067 - 1.522           |
## |-  Missing              |                4                 |
```

As a note, I didn't include the "states" variable as the output was huge and not that helpful. Suggest we include a note somewhere that data from 50 states + Washington, DC.

Distribution of Outcome Data

Histogram of raw outcome data (hate crimes per 100k).

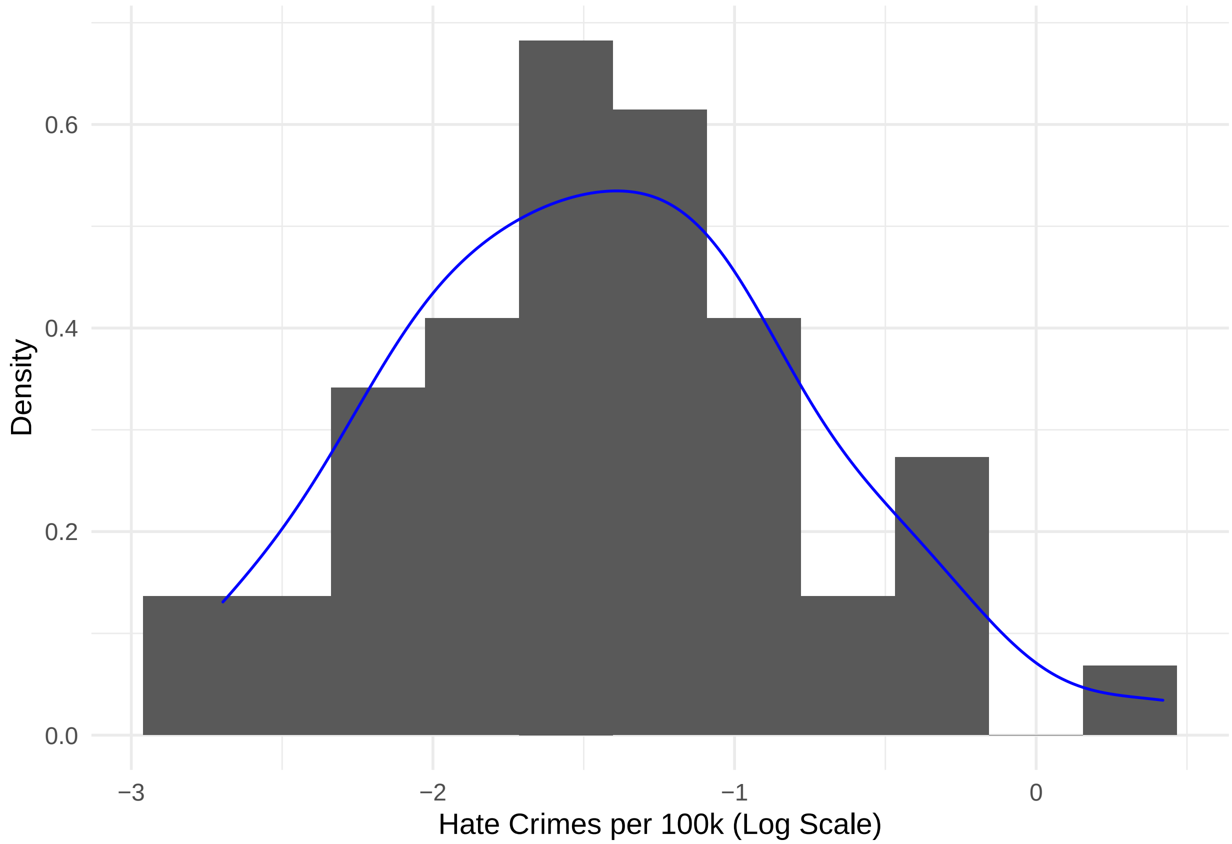
```r
hate_df %>%
  ggplot(aes(x = hate_crimes_per_100k_splc, y = ..density..)) +
  geom_histogram(bins = 11) +
  geom_density(alpha = 0.2, color = "blue") +
  labs(
    x = "Hate Crimes per 100k",
    y = "Density"
  )
```



These data look skewed :(

Histogram of log-transformed outcome data (hate crimes per 100k).

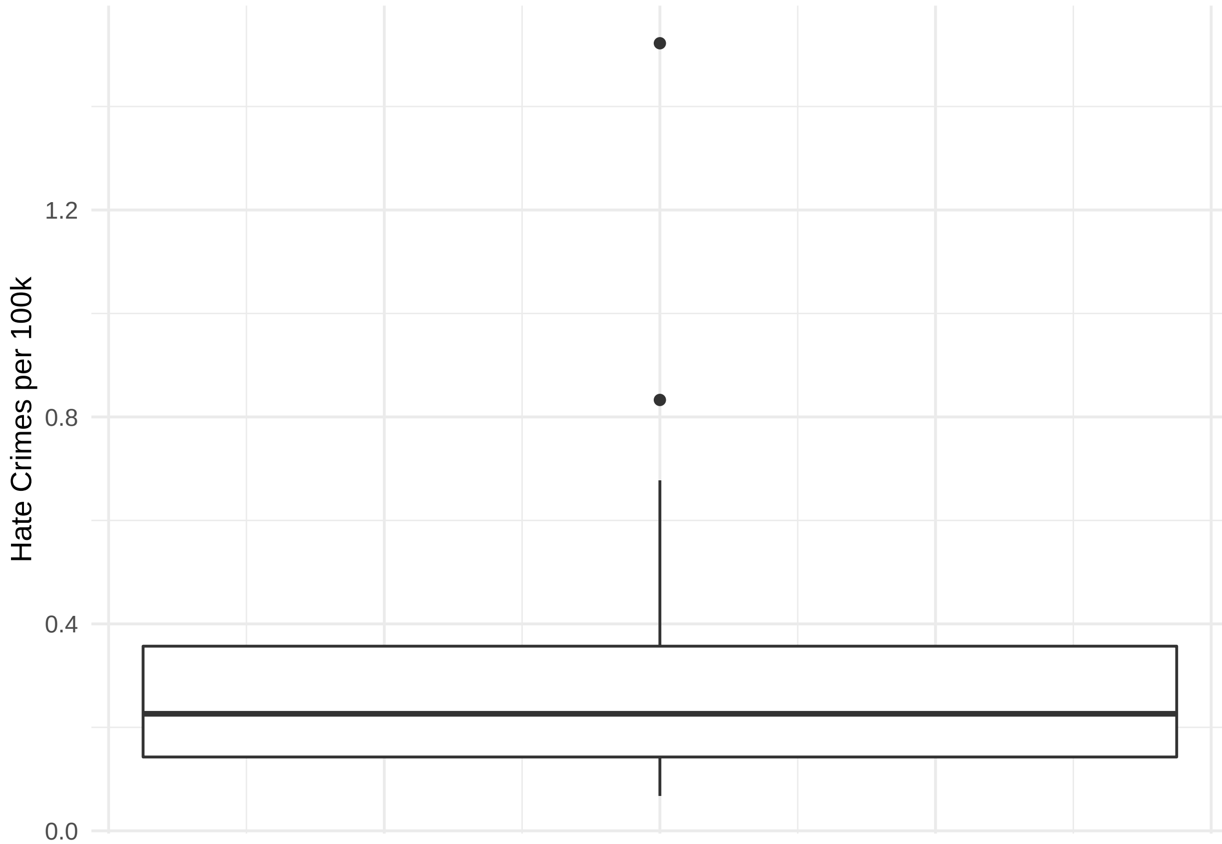
```
hate_df %>%
  ggplot(aes(x = log(hate_crimes_per_100k_splc), y = ..density..)) +
  geom_histogram(bins = 11) +
  geom_density(alpha = 0.2, color = "blue") +
  labs(
    x = "Hate Crimes per 100k (Log Scale)",
    y = "Density"
  )
```



Looks better!

Box plot of the (raw) outcome data.

```
hate_df %>%
  ggplot(aes(y = hate_crimes_per_100k_splc)) +
  geom_boxplot() +
  labs(
    y = "Hate Crimes per 100k"
  ) +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
    )
```



Just based on the box plot, it looks like there are two states with potential usually high rates (Washington, DC and Oregon).

Examining Potential Multicollinearity

```
hate_df %>%
  select(
    hate_crimes_per_100k_splc,
    median_household_income,
    perc_population_with_high_school_degree,
    perc_non_citizen,
    gini_index,
    perc_non_white
    ) %>%
  cor(use = "complete.obs") %>% # Ignoring NA values
  round(., 2)
```

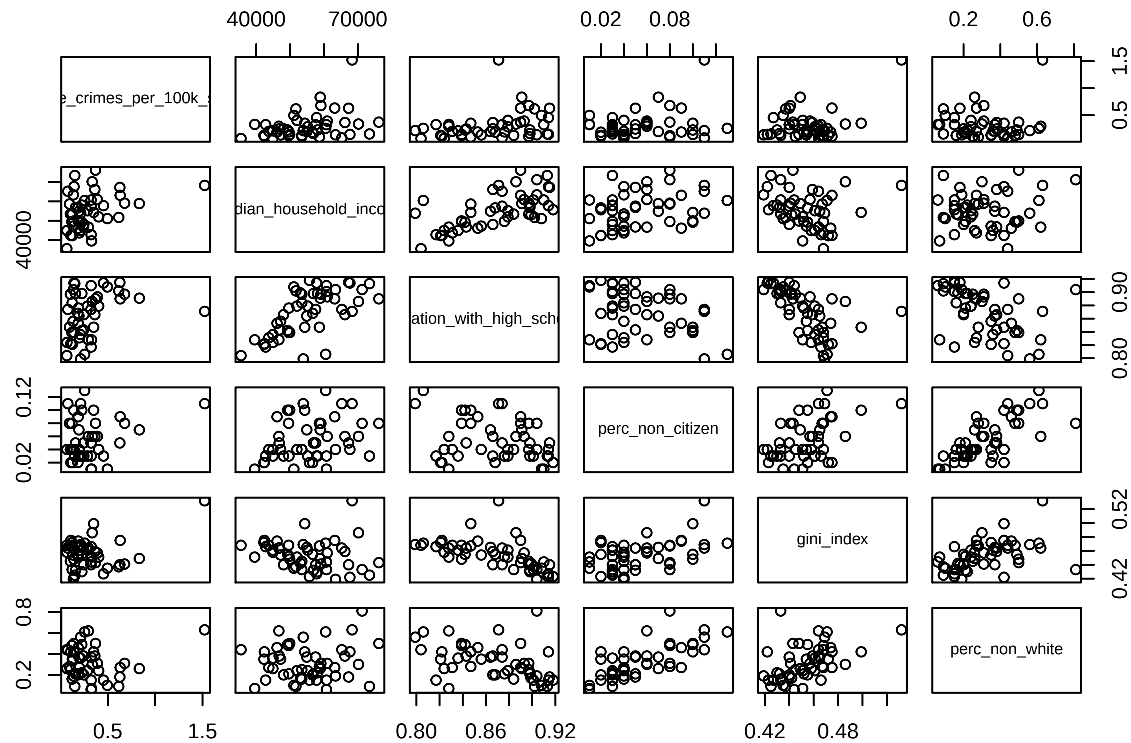
```
##                                         hate_crimes_per_100k_splc
## hate_crimes_per_100k_splc                                    1.00
## median_household_income                                      0.34
## perc_population_with_high_school_degree                      0.26
## perc_non_citizen                                             0.24
## gini_index                                                   0.38
```

```
## perc_non_white                                                      0.11
##                                          median_household_income
## hate_crimes_per_100k_splc                                   0.34
## median_household_income                                     1.00
## perc_population_with_high_school_degree                     0.65
## perc_non_citizen                                            0.30
## gini_index                                                 -0.13
## perc_non_white                                              0.04
##                                          perc_population_with_high_school_degree
## hate_crimes_per_100k_splc                                                   0.26
## median_household_income                                                     0.65
## perc_population_with_high_school_degree                                     1.00
## perc_non_citizen                                                           -0.26
## gini_index                                                                 -0.54
## perc_non_white                                                             -0.50
##                                          perc_non_citizen gini_index
## hate_crimes_per_100k_splc                            0.24       0.38
## median_household_income                              0.30      -0.13
## perc_population_with_high_school_degree             -0.26      -0.54
## perc_non_citizen                                     1.00       0.48
## gini_index                                           0.48       1.00
## perc_non_white                                       0.75       0.55
##                                          perc_non_white
## hate_crimes_per_100k_splc                          0.11
## median_household_income                            0.04
## perc_population_with_high_school_degree           -0.50
## perc_non_citizen                                   0.75
## gini_index                                         0.55
## perc_non_white                                     1.00
```

Based on this output, the following pairs of variables have a correlation of 60% or higher:

- Percentage non-citizens & percentage non-white (0.75)
- Median household income & percentage of population with a high school degree (0.65)

```r
hate_df %>%
  select(
    hate_crimes_per_100k_splc,
    median_household_income,
    perc_population_with_high_school_degree,
    perc_non_citizen,
    gini_index,
    perc_non_white
    ) %>%
  pairs()
```



Simple Linear Regression Using Income Inequality (Per FiveThirtyEight)

Fitting SLR using income inequality (measured by Gini index) per FiveThirtyEight findings.

```
slr_gini_lm = lm(hate_crimes_per_100k_splc ~ gini_index, data = hate_df)
slr_gini_log_lm = lm(log(hate_crimes_per_100k_splc) ~ gini_index, data = hate_df)

summary(slr_gini_lm)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ gini_index, data = hate_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28669 -0.14565 -0.04991  0.07356  0.91085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5275     0.7833  -1.950   0.0574 .
## gini_index    4.0205     1.7177   2.341   0.0237 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2412 on 45 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.08872
## F-statistic: 5.478 on 1 and 45 DF,  p-value: 0.02374
```

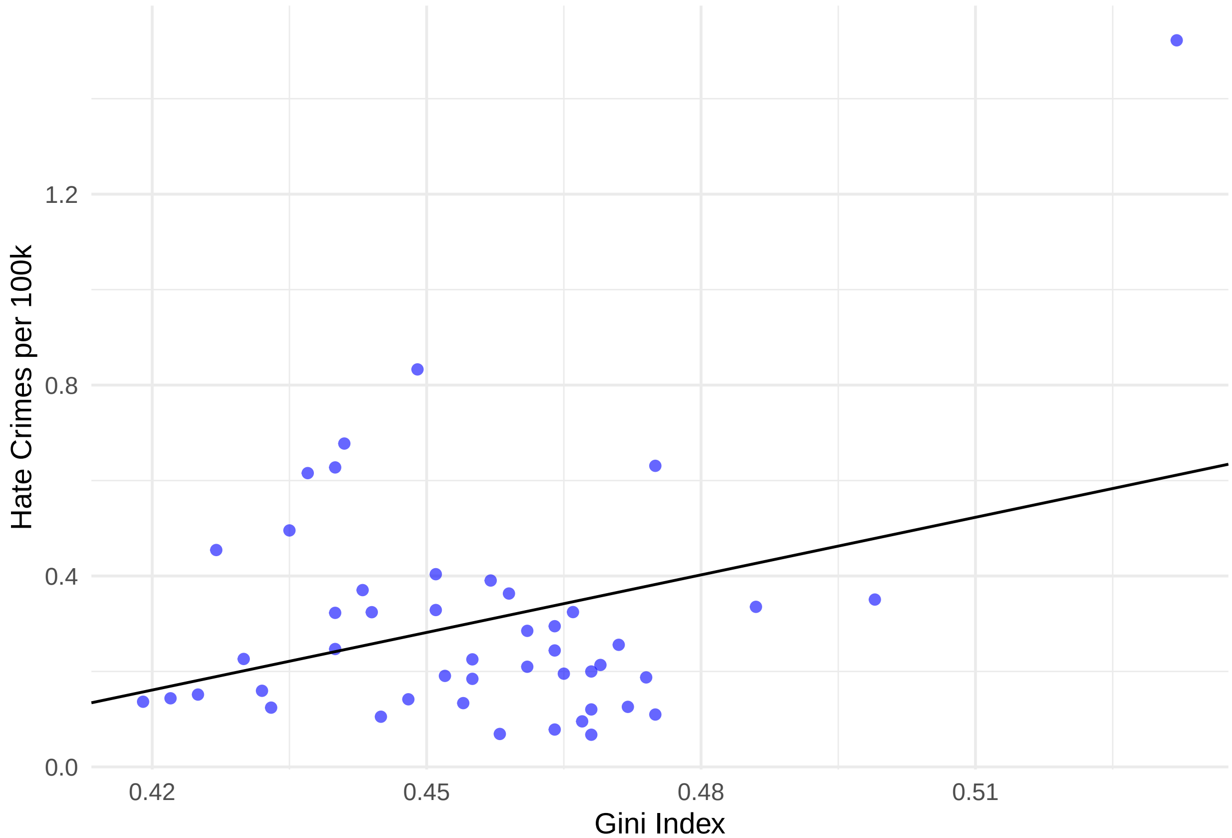
```r
summary(slr_gini_log_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ gini_index, data = hate_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32883 -0.36358 -0.02325  0.38705  1.47219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.676      2.195  -1.674    0.101
## gini_index     4.932      4.814   1.024    0.311
##
## Residual standard error: 0.6761 on 45 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.02279,    Adjusted R-squared:  0.001073
## F-statistic: 1.049 on 1 and 45 DF,  p-value: 0.3111
```

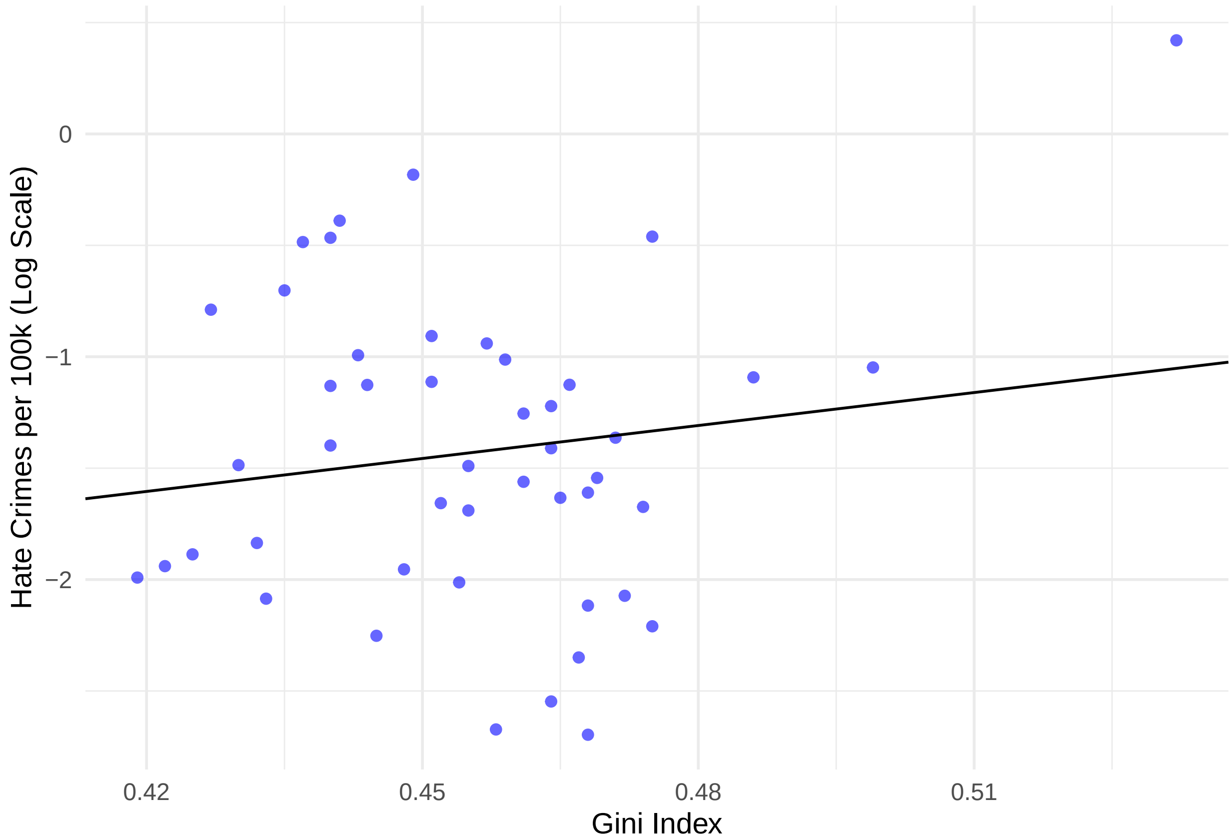
Gini index appears to be a significant predictor only when using the raw outcome data (not the log-transformed outcome data).

Scatter plots associated with these simple linear regression models.

```r
hate_df %>%
  ggplot(aes(x = gini_index, y = hate_crimes_per_100k_splc)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    x = "Gini Index",
    y = "Hate Crimes per 100k"
    ) +
  geom_abline(intercept = -1.5275, slope = 4.0205)
```

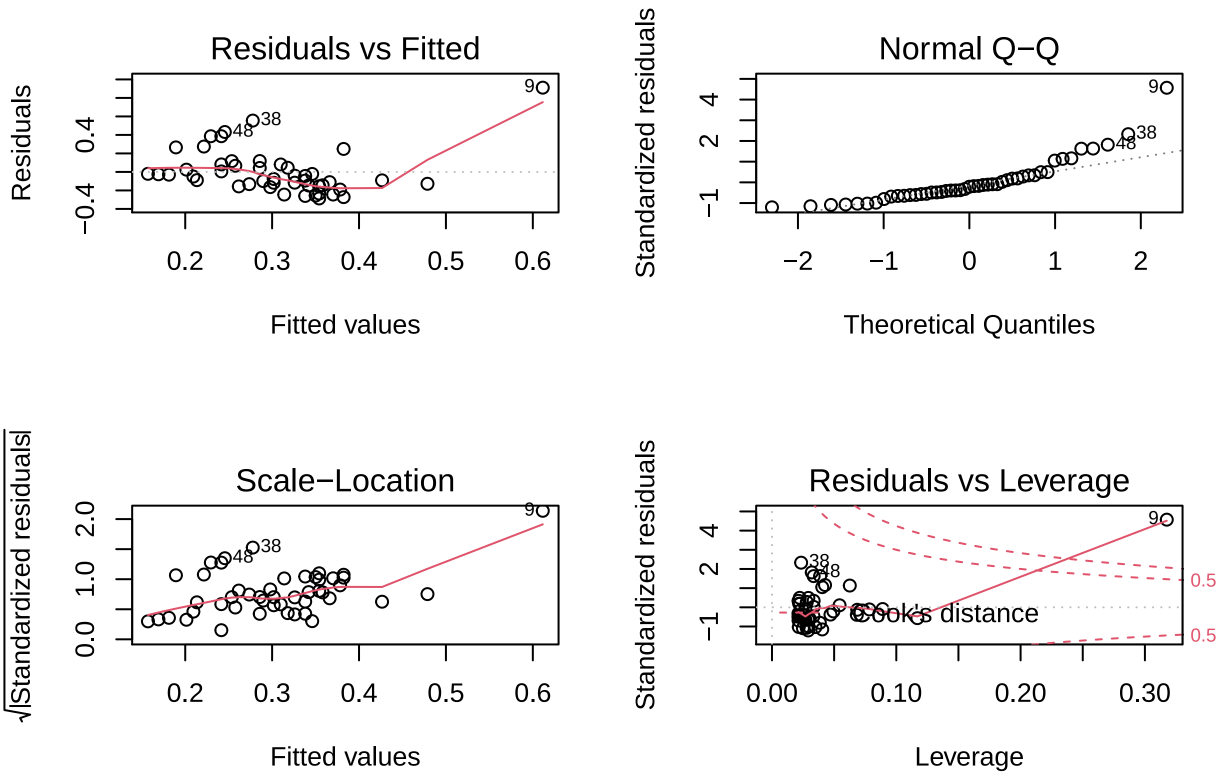



```r
hate_df %>%
  ggplot(aes(x = gini_index, y = log(hate_crimes_per_100k_splc))) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    x = "Gini Index",
    y = "Hate Crimes per 100k (Log Scale)"
    ) +
  geom_abline(intercept = -3.676, slope = 4.932)
```

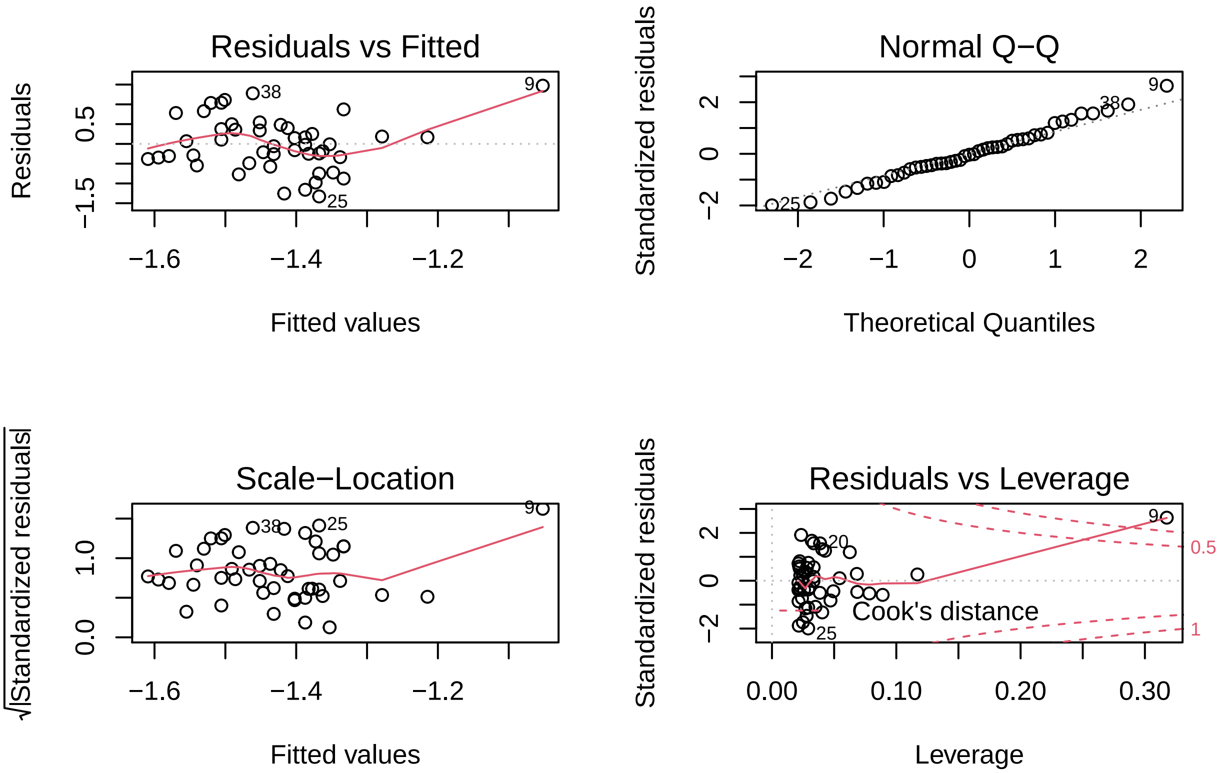


Diagnostic plots of these two simple linear regression models.

```
par(mfrow = c(2, 2))
plot(slr_gini_lm)
```




```r
plot(slr_gini_log_lm)
```



Normality looks better when we perform the log transformation on the outcome data. However, for both versions of this model, we can see an outlying value in the upper right corner (this corresponds to Washington, DC).

Trying Stepwise Approach

First, looking at the full model (with and without log transformation of outcome).

```
hate_nona_df = # Removing rows with missing values from the dataset
  hate_df %>%
  drop_na()

full_lm = lm(
  hate_crimes_per_100k_splc
  ~ unemployment +
    urbanization +
    median_household_income +
    perc_population_with_high_school_degree +
    perc_non_citizen +
    gini_index +
    perc_non_white,
  data = hate_nona_df
)

full_log_lm = lm(
```

```r
  log(hate_crimes_per_100k_splc)
  ~ unemployment +
    urbanization +
    median_household_income +
    perc_population_with_high_school_degree +
    perc_non_citizen +
    gini_index +
    perc_non_white,
  data = hate_nona_df
)

summary(full_lm)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ unemployment + urbanization +
##     median_household_income + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index + perc_non_white, data = hate_nona_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)                             -8.296e+00  1.908e+00  -4.349 0.000103
## unemploymentlow                          1.307e-02  7.173e-02   0.182 0.856425
## urbanizationlow                          3.309e-02  8.475e-02   0.390 0.698475
## median_household_income                 -1.504e-06  5.961e-06  -0.252 0.802193
## perc_population_with_high_school_degree  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
##
## (Intercept)                             ***
## unemploymentlow
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree **
## perc_non_citizen
## gini_index                              ***
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2014 on 37 degrees of freedom
## Multiple R-squared:  0.461,  Adjusted R-squared:  0.3591
## F-statistic: 4.521 on 7 and 37 DF,  p-value: 0.001007
```

```r
summary(full_log_lm)
```

```
##
```

```
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ unemployment +
##     urbanization + median_household_income + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index + perc_non_white, data = hate_nona_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28845 -0.41144  0.01898  0.31334  1.13022
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -1.857e+01  5.553e+00  -3.344  0.00190
## unemploymentlow                          2.179e-01  2.088e-01   1.043  0.30353
## urbanizationlow                         -9.885e-02  2.467e-01  -0.401  0.69092
## median_household_income                 -4.732e-06  1.735e-05  -0.273  0.78658
## perc_population_with_high_school_degree  1.121e+01  5.341e+00   2.098  0.04275
## perc_non_citizen                         1.168e+00  5.464e+00   0.214  0.83189
## gini_index                               1.670e+01  5.744e+00   2.908  0.00611
## perc_non_white                          -1.232e-01  1.069e+00  -0.115  0.90887
##
## (Intercept)                             **
## unemploymentlow
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree *
## perc_non_citizen
## gini_index                              **
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5862 on 37 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.1849
## F-statistic: 2.426 on 7 and 37 DF,  p-value: 0.03768
```

Trying stepwise approach.

```
step(full_lm, direction = "both") # Trying "both" directions
```

```
## Start:  AIC=-137.03
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + perc_non_white
##
##                                           Df Sum of Sq    RSS     AIC
## - perc_non_white                           1   0.00001 1.5008 -139.03
## - unemployment                             1   0.00135 1.5021 -138.99
## - median_household_income                  1   0.00258 1.5034 -138.95
## - urbanization                             1   0.00618 1.5070 -138.85
## - perc_non_citizen                         1   0.01750 1.5183 -138.51
## <none>                                                 1.5008 -137.03
## - perc_population_with_high_school_degree  1   0.34889 1.8497 -129.62
## - gini_index                               1   0.77465 2.2754 -120.30
```

```
##
## Step:  AIC=-139.03
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - unemployment                             1   0.00148 1.5023 -140.99
## - median_household_income                  1   0.00269 1.5035 -140.95
## - urbanization                             1   0.00617 1.5070 -140.85
## - perc_non_citizen                         1   0.02422 1.5250 -140.31
## <none>                                                 1.5008 -139.03
## + perc_non_white                           1   0.00001 1.5008 -137.03
## - perc_population_with_high_school_degree  1   0.38759 1.8884 -130.69
## - gini_index                               1   0.77888 2.2797 -122.22
##
## Step:  AIC=-140.99
## hate_crimes_per_100k_splc ~ urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - median_household_income                  1   0.00243 1.5047 -142.91
## - urbanization                             1   0.00693 1.5092 -142.78
## - perc_non_citizen                         1   0.02401 1.5263 -142.27
## <none>                                                 1.5023 -140.99
## + unemployment                             1   0.00148 1.5008 -139.03
## + perc_non_white                           1   0.00015 1.5021 -138.99
## - perc_population_with_high_school_degree  1   0.40517 1.9074 -132.24
## - gini_index                               1   0.78876 2.2910 -124.00
##
## Step:  AIC=-142.91
## hate_crimes_per_100k_splc ~ urbanization + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - urbanization                             1   0.00762 1.5123 -144.69
## - perc_non_citizen                         1   0.02232 1.5270 -144.25
## <none>                                                 1.5047 -142.91
## + median_household_income                  1   0.00243 1.5023 -140.99
## + unemployment                             1   0.00122 1.5035 -140.95
## + perc_non_white                           1   0.00034 1.5044 -140.92
## - gini_index                               1   0.78737 2.2921 -125.97
## - perc_population_with_high_school_degree  1   0.86254 2.3672 -124.52
##
## Step:  AIC=-144.69
## hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - perc_non_citizen                         1   0.01471 1.5270 -146.25
## <none>                                                 1.5123 -144.69
## + urbanization                             1   0.00762 1.5047 -142.91
## + median_household_income                  1   0.00311 1.5092 -142.78
```

```
## + unemployment                           1   0.00192 1.5104 -142.74
## + perc_non_white                         1   0.00028 1.5120 -142.69
## - gini_index                             1   0.78804 2.3004 -127.81
## - perc_population_with_high_school_degree  1   0.85561 2.3679 -126.51
##
## Step:  AIC=-146.25
## hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## <none>                                                1.5270 -146.25
## + perc_non_citizen                         1   0.01471 1.5123 -144.69
## + perc_non_white                           1   0.00522 1.5218 -144.40
## + unemployment                             1   0.00136 1.5257 -144.29
## + median_household_income                  1   0.00068 1.5263 -144.27
## + urbanization                             1   0.00001 1.5270 -144.25
## - perc_population_with_high_school_degree  1   0.85432 2.3813 -128.25
## - gini_index                               1   1.06513 2.5922 -124.44
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index, data = hate_nona_df)
##
## Coefficients:
##                             (Intercept)
##                                  -8.103
## perc_population_with_high_school_degree
##                                   5.059
##                              gini_index
##                                   8.825
```

```r
step(full_log_lm, direction = "both") # Trying "both" directions
```

```
## Start:  AIC=-40.88
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
##     median_household_income + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index + perc_non_white
##
##                                           Df Sum of Sq    RSS     AIC
## - perc_non_white                           1   0.00456 12.719 -42.859
## - perc_non_citizen                         1   0.01570 12.730 -42.820
## - median_household_income                  1   0.02556 12.740 -42.785
## - urbanization                             1   0.05519 12.770 -42.680
## - unemployment                             1   0.37413 13.089 -41.570
## <none>                                                12.715 -40.875
## - perc_population_with_high_school_degree  1   1.51318 14.228 -37.815
## - gini_index                               1   2.90660 15.621 -33.611
##
## Step:  AIC=-42.86
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
##     median_household_income + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index
```

```
##
##                                         Df Sum of Sq    RSS     AIC
## - perc_non_citizen                       1   0.01114 12.730 -44.820
## - median_household_income                1   0.02946 12.749 -44.755
## - urbanization                           1   0.05718 12.777 -44.657
## - unemployment                           1   0.41699 13.136 -43.408
## <none>                                               12.719 -42.859
## + perc_non_white                         1   0.00456 12.715 -40.875
## - perc_population_with_high_school_degree  1   1.73309 14.452 -39.111
## - gini_index                             1   2.90620 15.626 -35.599
##
## Step:  AIC=-44.82
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
##     median_household_income + perc_population_with_high_school_degree +
##     gini_index
##
##                                         Df Sum of Sq    RSS     AIC
## - median_household_income                1   0.01910 12.750 -46.752
## - urbanization                           1   0.11092 12.841 -46.429
## - unemployment                           1   0.41466 13.145 -45.377
## <none>                                               12.730 -44.820
## + perc_non_citizen                       1   0.01114 12.719 -42.859
## + perc_non_white                         1   0.00000 12.730 -42.820
## - perc_population_with_high_school_degree  1   1.92883 14.659 -40.471
## - gini_index                             1   3.00737 15.738 -37.277
##
## Step:  AIC=-46.75
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
##     perc_population_with_high_school_degree + gini_index
##
##                                         Df Sum of Sq    RSS     AIC
## - urbanization                           1   0.09183 12.841 -48.429
## - unemployment                           1   0.40424 13.154 -47.348
## <none>                                               12.750 -46.752
## + median_household_income                1   0.01910 12.730 -44.820
## + perc_non_white                         1   0.00293 12.747 -44.763
## + perc_non_citizen                       1   0.00077 12.749 -44.755
## - gini_index                             1   3.02492 15.774 -39.172
## - perc_population_with_high_school_degree  1   3.15688 15.906 -38.797
##
## Step:  AIC=-48.43
## log(hate_crimes_per_100k_splc) ~ unemployment + perc_population_with_high_school_degree +
##     gini_index
##
##                                         Df Sum of Sq    RSS     AIC
## - unemployment                           1    0.3655 13.207 -49.166
## <none>                                               12.841 -48.429
## + urbanization                           1    0.0918 12.750 -46.752
## + perc_non_citizen                       1    0.0417 12.800 -46.576
## + perc_non_white                         1    0.0049 12.837 -46.446
## + median_household_income                1    0.0000 12.841 -46.429
## - perc_population_with_high_school_degree  1    3.3459 16.187 -40.010
## - gini_index                             1    4.0555 16.897 -38.079
##
```

```
## Step:  AIC=-49.17
## log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##     gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## <none>                                                 13.207 -49.166
## + unemployment                             1    0.3655 12.841 -48.429
## + urbanization                             1    0.0531 13.154 -47.348
## + perc_non_citizen                         1    0.0288 13.178 -47.265
## + perc_non_white                           1    0.0026 13.204 -47.175
## + median_household_income                  1    0.0001 13.207 -47.167
## - gini_index                               1    3.7171 16.924 -40.007
## - perc_population_with_high_school_degree  1    4.4569 17.664 -38.081
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##     gini_index, data = hate_nona_df)
##
## Coefficients:
##                             (Intercept)
##                                  -18.95
## perc_population_with_high_school_degree
##                                   11.55
##                              gini_index
##                                   16.49
```

This procedure is retaining the following two predictors:

- Precent population with high school degree
- Gini index

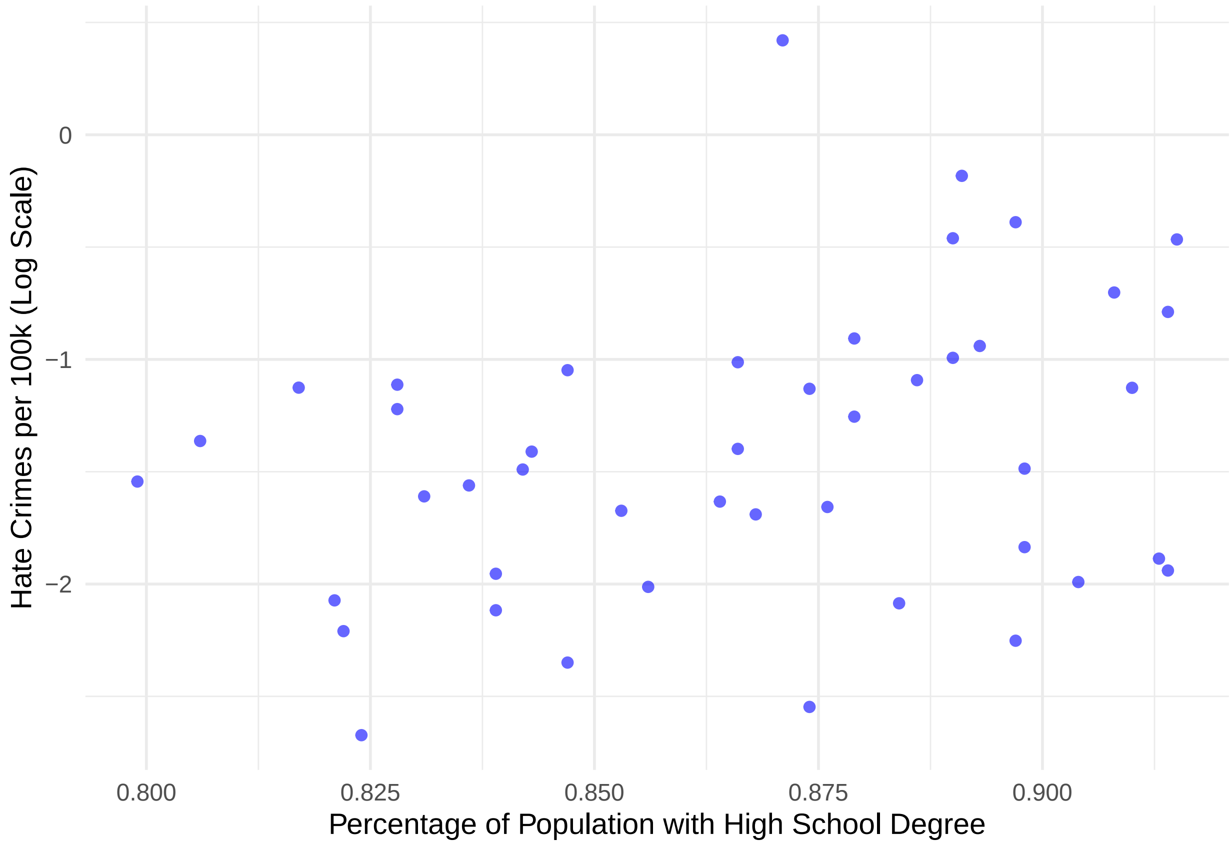
Jacy's Ideas

```r
##Project ideas
hate = read.csv("/Users/jacysparks/Downloads/HateCrimes.csv")
head(hate)
dim(hate)
hate$hate_crimes_per_100k_splc = as.character(hate$hate_crimes_per_100k_splc)
hate$hate_crimes_per_100k_splc = as.numeric(hate$hate_crimes_per_100k_splc)
summary(hate)
##Four NA's for outcome
##NA for Wyoming, South Dakota, North Dakota, and Idaho
hate[,c(1,9)]
##Could remove
hate = na.omit(hate)

##3 NA's for non citizen
```

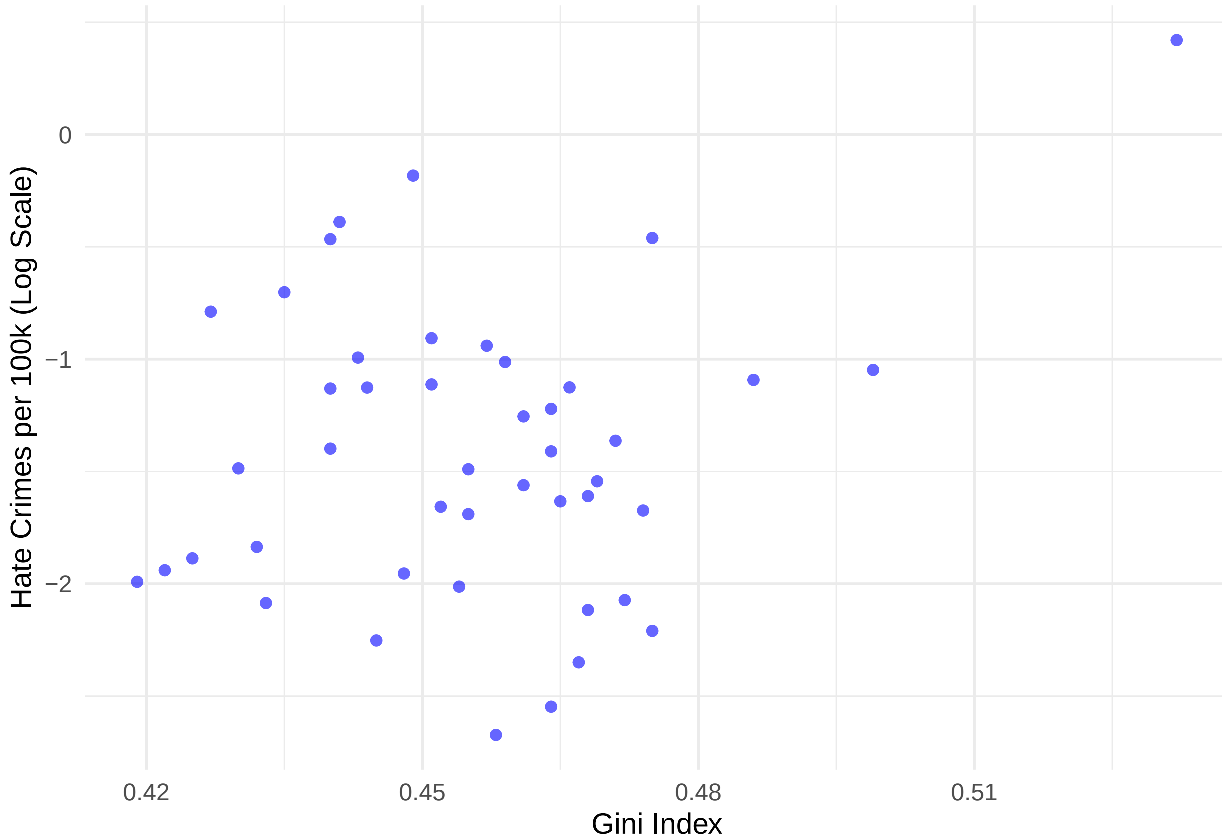
```r
##Create indicators
names(hate)[names(hate)=="unemployment"] = "High.Unemployment"
names(hate)[names(hate)=="urbanization"] = "High.Urban"
names(hate)[names(hate)=="median_household_income"] = "Med.Income"
names(hate)[names(hate)=="perc_population_with_high_school_degree"] = "HS.Degree"
names(hate)[names(hate)=="perc_non_citizen"] = "Non.Citizen"
names(hate)[names(hate)=="perc_non_white"] = "Non.White"
names(hate)[names(hate)=="hate_crimes_per_100k_splc"] = "Hate.Crime"
hate$High.Unemployment = ifelse(hate$High.Unemployment=="high",1,0)
hate$High.Urban = ifelse(hate$High.Urban=="high",1,0)

##Outcome var is skewed
hate$Hate.Crime = log(hate$Hate.Crime)
hist(hate$Hate.Crime)
##Much better

reg = lm(Hate.Crime~.-state,data=hate)
summary(reg)

pairs(hate[,4:9],lower.panel=NULL)
cor(hate[,4:9])
#Percent white and percent non-white highly correlated


##Check linearity
for(i in 4:8){
  plot(hate[,i],hate$Hate.Crime,main=colnames(hate)[i])
}
plot(reg)
```

Fit model based on stepwise (log transformed outcome data) and make a scatterplot with regression line, regenerate diagnostic plots using this model

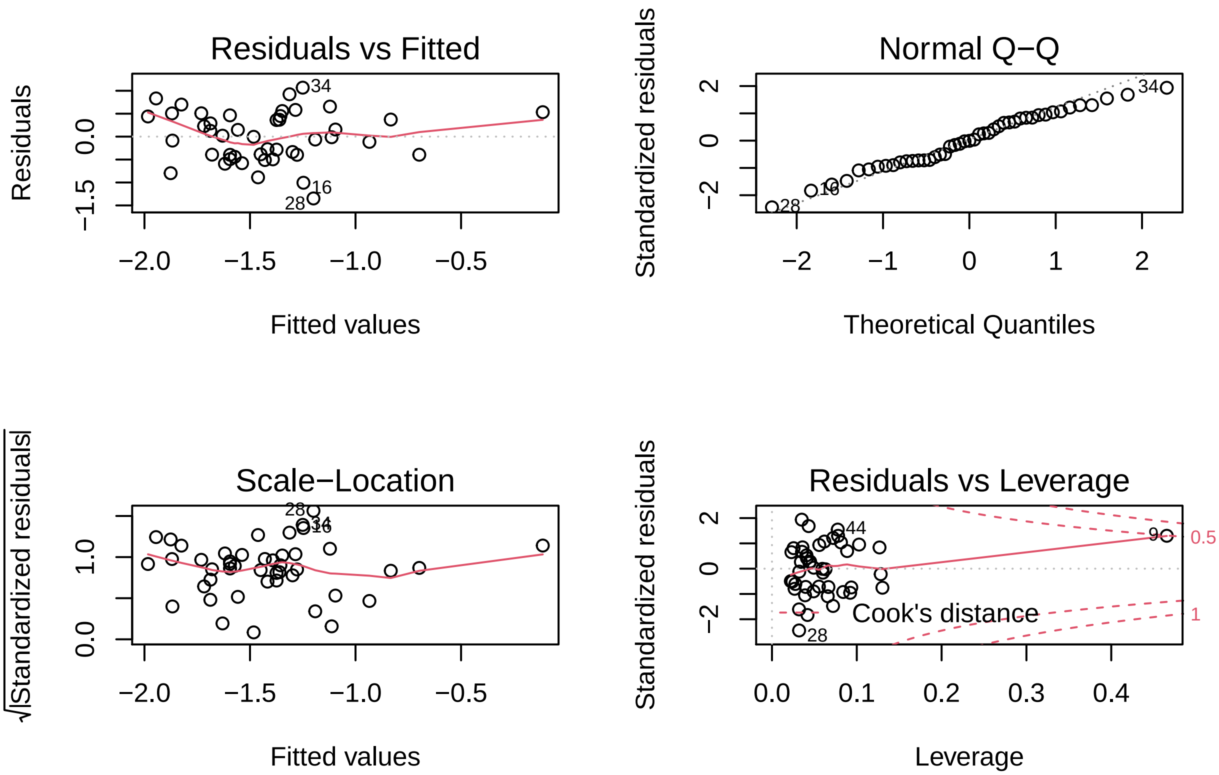
```r
stepwise_log_lm = lm(
  log(hate_crimes_per_100k_splc)
  ~ perc_population_with_high_school_degree + # Question - which is the main predictor we're putting on
    gini_index,
  data = hate_nona_df)

hate_nona_df %>%
  ggplot(aes(x = perc_population_with_high_school_degree, y = log(hate_crimes_per_100k_splc))) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    x = "Percentage of Population with High School Degree",
    y = "Hate Crimes per 100k (Log Scale)"
    ) +
  geom_abline(intercept = -18.947, slope = 11.554) # ab line not showing up...
```




```
hate_nona_df %>%
  ggplot(aes(x = gini_index, y = log(hate_crimes_per_100k_splc))) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    x = "Gini Index",
    y = "Hate Crimes per 100k (Log Scale)"
    ) +
  geom_abline(intercept = -18.947, slope = 16.486) # ab line not showing up...
```



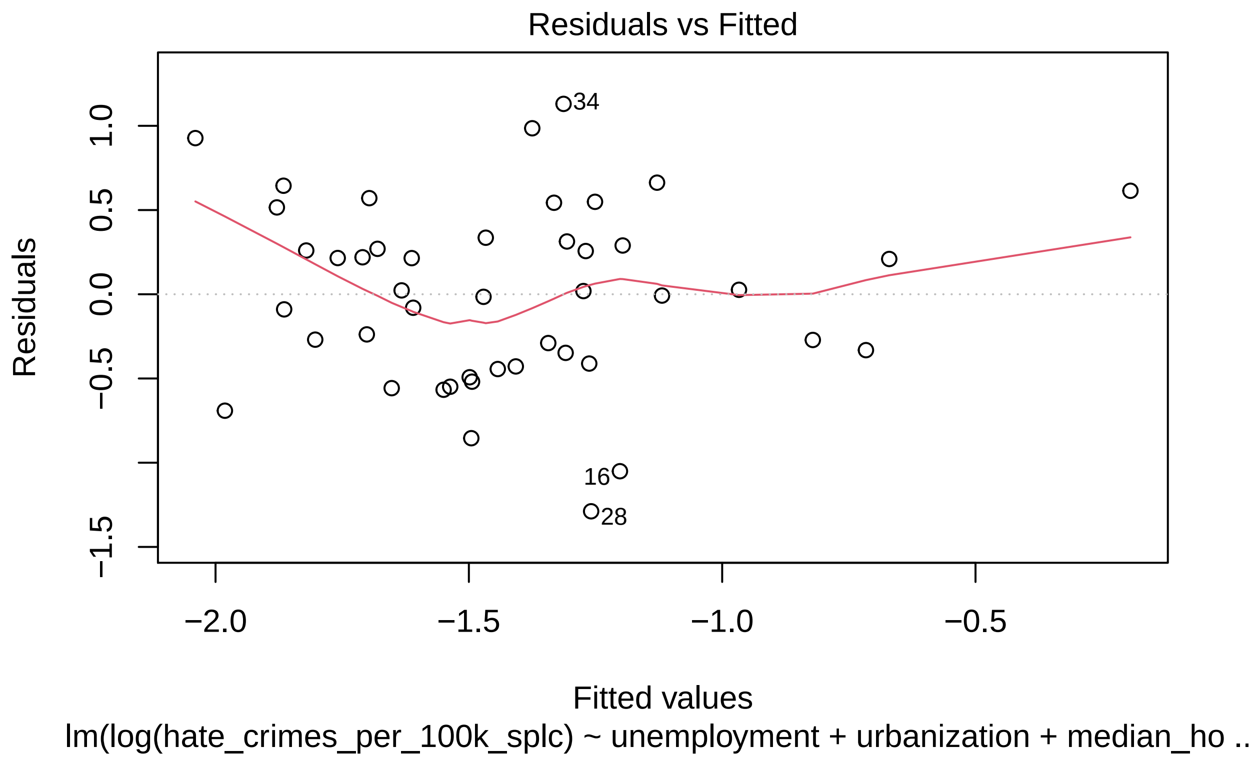

```
par(mfrow = c(2, 2))
plot(stepwise_log_lm)
```

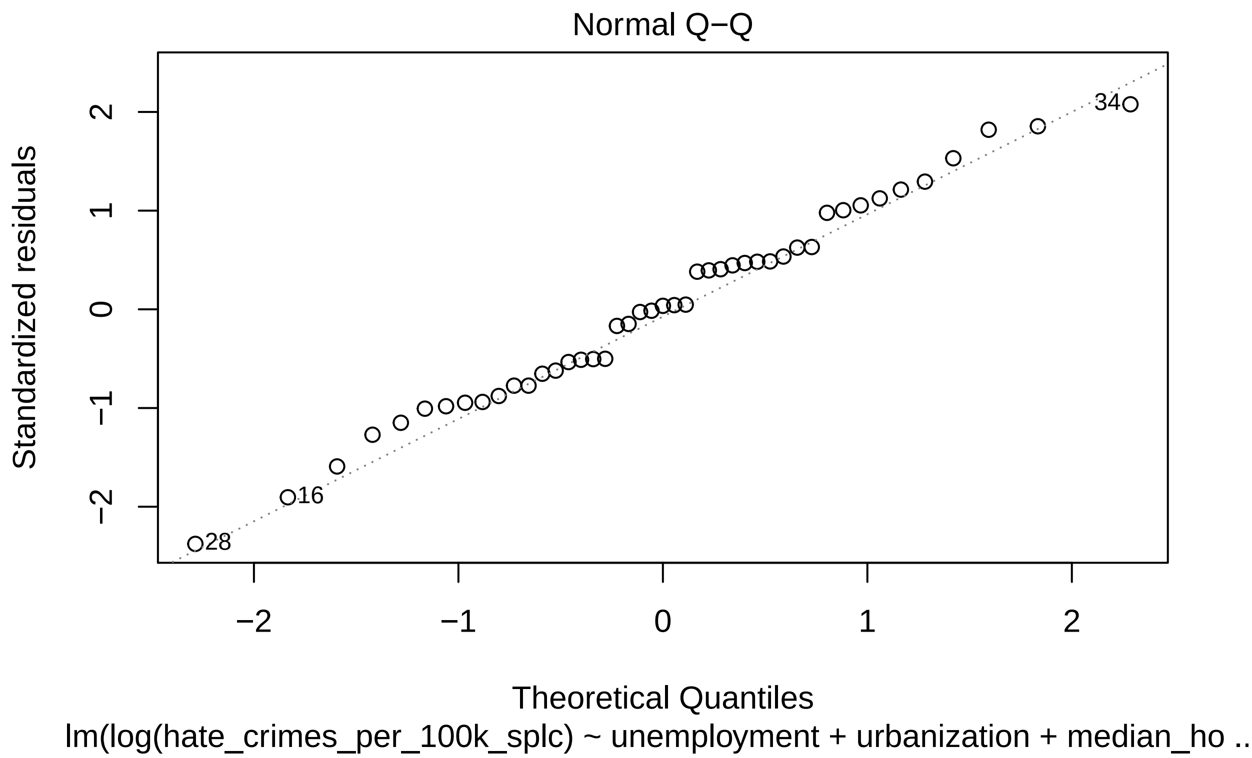


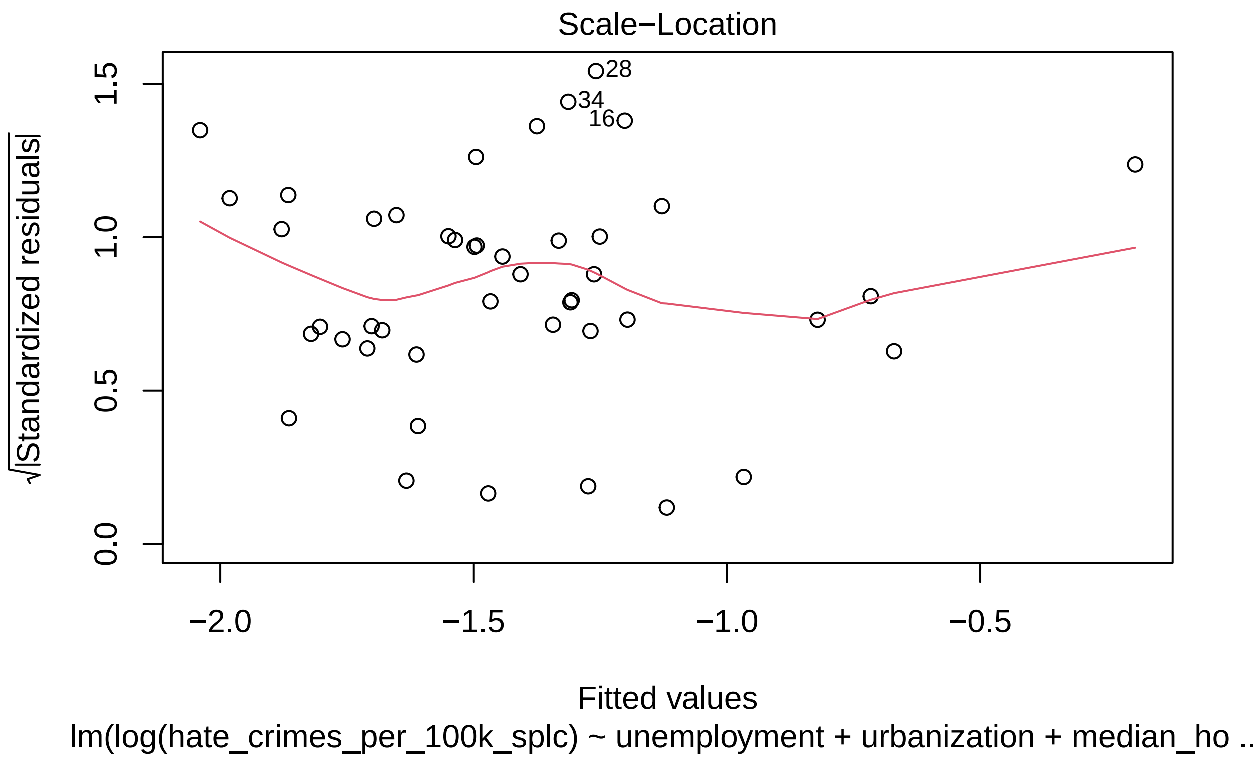
Formalize if DC is an influential point quantitatively using the Cook's value

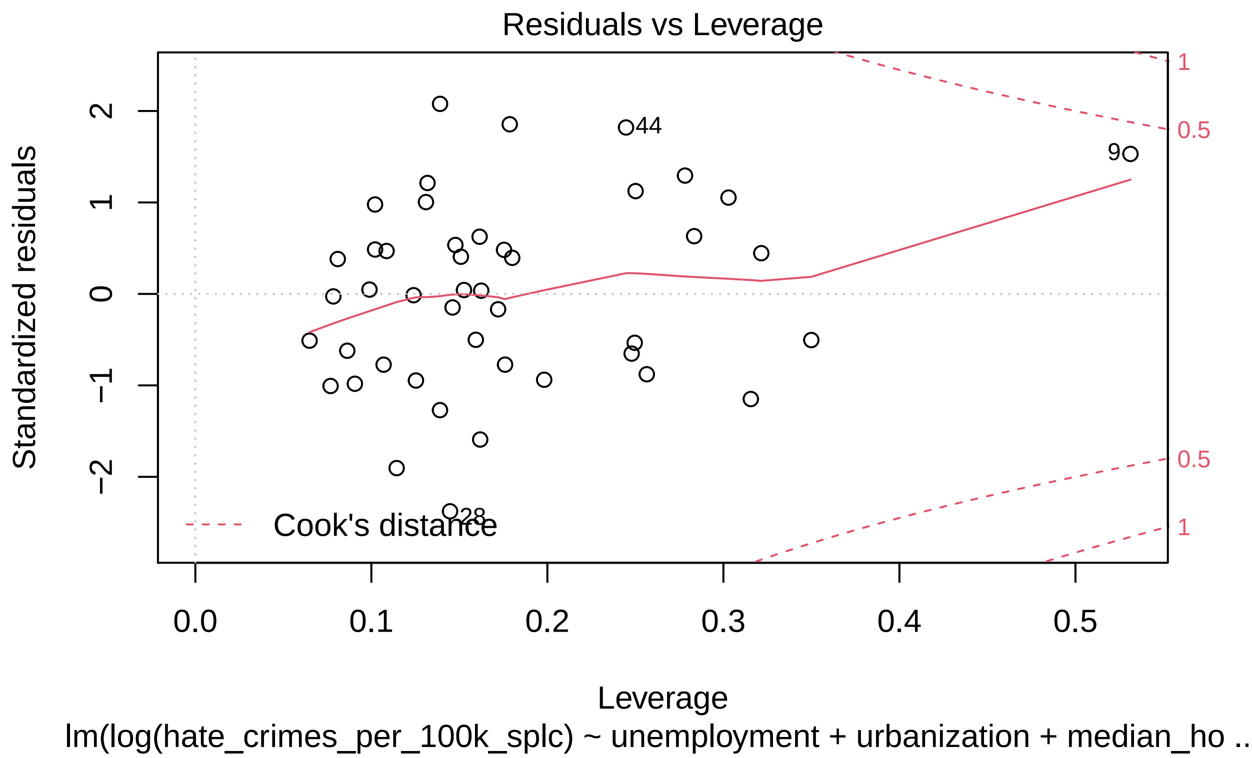
Using Cook's D and studentized residuals, DC is not a cause for concern. Cook's D is 0.332, not high enough for concern. The studentized residual is not greater than 2.5 for any variable. However, DFFIT is greater than 1. Could be cause for concern. Regression analysis shows it is not influential, so no need to delete.

```
plot(full_log_lm)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.5
−2.0
−1.5
−1.0
−0.5
34
16
28
Fitted values
lm(log(hate_crimes_per_100k_splc) ~ unemployment + urbanization + median_ho ..


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(hate_crimes_per_100k_splc) ~ unemployment + urbanization + median_ho ..
−2
−1
0
1
2
28
16
34


Scale−Location
28
34
16
1.5
1.0
0.5
0.0
√|Standardized residuals|
−2.0
−1.5
−1.0
−0.5
Fitted values
lm(log(hate_crimes_per_100k_splc) ~ unemployment + urbanization + median_ho ..




```r
influence.measures(full_log_lm)
```

```
## Influence measures of
##   lm(formula = log(hate_crimes_per_100k_splc) ~ unemployment +      urbanization + median_household_i
##
##        dfb.1_  dfb.unmp  dfb.urbn  dfb.md__  dfb.p___ dfb.prc_nn_c  dfb.gn_n
## 1  -0.016646  0.031350 -0.044614 -2.24e-02  0.050408     0.105642 -0.052786
## 2   0.012063  0.157918 -0.096174 -7.97e-02 -0.069039     0.095033  0.135200
## 3   0.019023 -0.039429 -0.013808 -8.19e-02  0.032744     0.087990 -0.074941
## 4  -0.236010  0.202864 -0.217649 -1.07e-01  0.262713    -0.096238  0.052876
## 5   0.406792 -0.074778  0.164944  3.22e-01 -0.432077     0.251954 -0.231014
## 6  -0.005176  0.008166 -0.008759 -3.02e-03  0.005263    -0.003645  0.003092
## 7   0.062223  0.109660  0.036905 -1.67e-01  0.031895     0.083860 -0.155210
## 8   0.064271  0.135859 -0.179319 -3.85e-03 -0.015767    -0.141394 -0.102337
## 9  -1.143310 -0.027740  0.412176  1.03e-01  0.566576    -0.074652  1.341169
## 10  0.174253  0.054632  0.037309  2.83e-01 -0.229455    -0.143129 -0.054502
## 11  0.002427  0.057339  0.094572  1.14e-01 -0.049844    -0.010129  0.041344
## 12 -0.034208 -0.069040 -0.118893  6.65e-02 -0.010997    -0.151748  0.073962
## 13  0.024622  0.067036  0.055788  3.06e-02 -0.037741    -0.006166 -0.005136
## 14  0.013964  0.052122  0.025554 -3.53e-02 -0.001867     0.013865 -0.015468
## 15 -0.090715  0.033274  0.084445 -1.05e-01  0.154886     0.070292 -0.050871
## 16  0.404672 -0.238811 -0.270244  3.93e-01 -0.455252    -0.141580 -0.146744
## 17  0.243226  0.253611  0.120183  9.99e-02 -0.370562     0.065359  0.127139
## 18 -0.052246 -0.035314  0.474494  2.97e-02  0.063115     0.630160 -0.065402
## 19  0.115926  0.119698 -0.072093  2.32e-01 -0.120884    -0.155918 -0.120610
```

```
## 20 -0.074770  0.058433 -0.005861 -2.90e-02  0.044595     0.090134  0.090794
## 21 -0.025739  0.105452 -0.165108 -8.52e-02  0.051425    -0.070406  0.005828
## 22 -0.088909  0.016417  0.272933  1.23e-01  0.047022     0.170697  0.051064
## 23 -0.008785  0.034709 -0.018169 -3.20e-02  0.018543     0.016433 -0.007498
## 24 -0.177524  0.072327 -0.000410 -2.10e-01  0.227952    -0.021745  0.053487
## 25  0.037991 -0.046432 -0.131087  8.09e-02 -0.082865    -0.145516  0.052510
## 26 -0.022190  0.017429  0.007901  2.66e-02 -0.003362    -0.034055  0.045934
## 27 -0.199744  0.043506 -0.128355 -3.90e-01  0.245592     0.016377  0.105863
## 28  0.045555  0.494479 -0.214028 -1.25e-01 -0.076388    -0.644096  0.094849
## 29 -0.062518 -0.020112  0.335046 -8.79e-02  0.088396    -0.178746 -0.056078
## 30  0.150384 -0.207005 -0.016418  9.56e-02 -0.058343    -0.120280 -0.238712
## 31 -0.026104 -0.058698  0.103613 -3.17e-02  0.021337     0.024352  0.023356
## 32  0.069832 -0.091385 -0.058018  7.97e-02 -0.060397    -0.010311 -0.065070
## 33  0.065648 -0.216958 -0.111445  1.30e-01 -0.065271     0.035793 -0.044660
## 34 -0.104315 -0.533391 -0.252593 -1.90e-01  0.281798     0.309129 -0.153623
## 35 -0.002137 -0.006017 -0.009950  7.89e-05  0.002374    -0.008066  0.002156
## 36 -0.290768  0.310644  0.081153 -2.84e-01  0.347888    -0.241430  0.039548
## 37  0.008720 -0.047766  0.055715 -1.94e-03 -0.014201    -0.053516  0.008566
## 38  0.003915 -0.003639 -0.011387 -2.52e-03 -0.003241    -0.005090  0.000393
## 39  0.171198  0.170417 -0.005849  7.68e-02 -0.186068     0.040223 -0.072333
## 40 -0.181138 -0.070810  0.316215 -6.30e-02  0.061426     0.128540  0.267284
## 41  0.001197 -0.000474 -0.000672 -1.17e-03 -0.000346     0.001080 -0.001844
## 42  0.077414  0.104719 -0.089658  1.34e-01 -0.109670    -0.120165 -0.009522
## 43 -0.128796 -0.490745 -0.177487 -3.20e-01  0.398684     0.395975 -0.298387
## 44  0.442715 -0.417120  0.061820  7.70e-02 -0.445905     0.083008 -0.076390
## 45  0.000181 -0.002123 -0.002174  1.79e-04 -0.001320     0.000763  0.002115
##    dfb.prc_nn_w    dffit cov.r   cook.d    hat inf
## 1     -0.053942 -0.21605 1.402 5.96e-03 0.1593
## 2     -0.196402 -0.36623 1.812 1.71e-02 0.3499   *
## 3      0.006586  0.16940 1.415 3.67e-03 0.1509
## 4      0.218692 -0.51497 1.012 3.26e-02 0.1390
## 5     -0.102595  0.69544 1.400 6.03e-02 0.3029
## 6      0.003267  0.01560 1.381 3.13e-05 0.0989
## 7      0.072184 -0.30517 1.560 1.19e-02 0.2496
## 8      0.135419  0.27222 1.364 9.42e-03 0.1615
## 9      0.366668  1.66099 1.574 3.32e-01 0.5312   *
## 10    -0.035083 -0.35530 1.326 1.60e-02 0.1760
## 11    -0.082401 -0.29041 1.080 1.05e-02 0.0768
## 12     0.107859 -0.30987 1.108 1.20e-02 0.0906
## 13     0.011207 -0.13327 1.258 2.27e-03 0.0649
## 14    -0.005532  0.11190 1.313 1.60e-03 0.0809
## 15    -0.001769  0.32955 1.125 1.36e-02 0.1021
## 16    -0.183044 -0.71073 0.616 5.85e-02 0.1144
## 17    -0.313981  0.65170 1.257 5.27e-02 0.2501
## 18    -0.350455 -0.78393 1.360 7.61e-02 0.3156
## 19     0.182647  0.39378 1.593 1.97e-02 0.2834
## 20    -0.082304  0.18282 1.468 4.28e-03 0.1800
## 21     0.024866  0.22055 1.373 6.20e-03 0.1477
## 22    -0.126418  0.47595 1.037 2.79e-02 0.1319
## 23     0.014236 -0.06034 1.451 4.68e-04 0.1461
## 24     0.015751  0.39001 1.149 1.90e-02 0.1310
## 25     0.031568 -0.26617 1.224 8.95e-03 0.1070
## 26    -0.005003 -0.07561 1.495 7.34e-04 0.1720
## 27     0.182952 -0.51415 1.415 3.33e-02 0.2565
```

```
## 28     0.309571 -1.04749 0.387 1.19e-01 0.1447
## 29     0.608783  0.81088 1.191 8.07e-02 0.2782
## 30     0.052125 -0.37171 1.509 1.75e-02 0.2479
## 31     0.025886  0.16206 1.318 3.35e-03 0.1021
## 32    -0.010658 -0.18927 1.253 4.55e-03 0.0863
## 33    -0.175438 -0.35778 1.170 1.60e-02 0.1253
## 34    -0.353941  0.87633 0.536 8.71e-02 0.1390
## 35    -0.000427  0.01537 1.486 3.04e-05 0.1624
## 36     0.537018 -0.71493 0.842 6.12e-02 0.1618
## 37     0.052399  0.16215 1.332 3.36e-03 0.1086
## 38    -0.004509  0.01783 1.469 4.08e-05 0.1526
## 39     0.012507  0.30335 1.758 1.18e-02 0.3215   *
## 40     0.002291 -0.46583 1.281 2.72e-02 0.1982
## 41     0.001507 -0.00524 1.421 3.53e-06 0.1240
## 42     0.048769  0.22000 1.436 6.18e-03 0.1754
## 43    -0.186325  0.89594 0.694 9.35e-02 0.1786
## 44    -0.676928  1.07075 0.778 1.34e-01 0.2447
## 45    -0.002153 -0.00781 1.351 7.84e-06 0.0784
```

```r
stu_res<-rstandard(full_log_lm)
outliers_y<-stu_res[abs(stu_res)>2.5]
outliers_y
```

```
## named numeric(0)
```

```r
summary(full_log_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ unemployment +
##     urbanization + median_household_income + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index + perc_non_white, data = hate_nona_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28845 -0.41144  0.01898  0.31334  1.13022
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -1.857e+01  5.553e+00  -3.344  0.00190
## unemploymentlow                         2.179e-01  2.088e-01   1.043  0.30353
## urbanizationlow                        -9.885e-02  2.467e-01  -0.401  0.69092
## median_household_income                -4.732e-06  1.735e-05  -0.273  0.78658
## perc_population_with_high_school_degree  1.121e+01  5.341e+00   2.098  0.04275
## perc_non_citizen                        1.168e+00  5.464e+00   0.214  0.83189
## gini_index                              1.670e+01  5.744e+00   2.908  0.00611
## perc_non_white                         -1.232e-01  1.069e+00  -0.115  0.90887
##
## (Intercept)                            **
## unemploymentlow
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree *
```

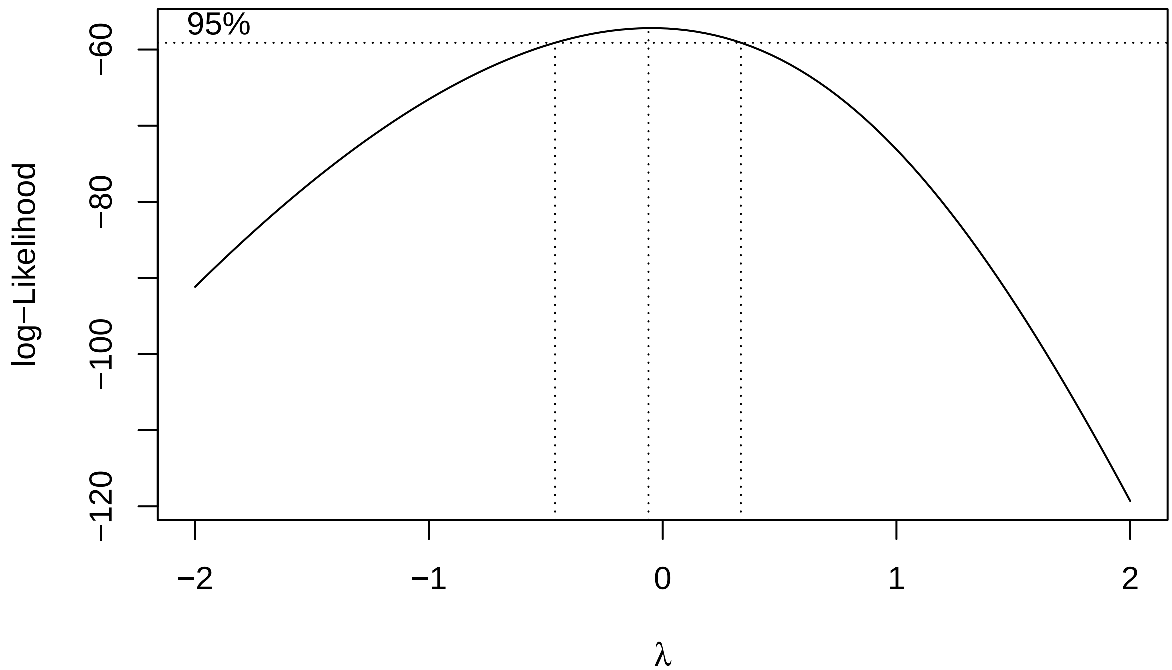
```
## perc_non_citizen
## gini_index                               **
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5862 on 37 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.1849
## F-statistic: 2.426 on 7 and 37 DF,  p-value: 0.03768
```

```
full_nodc = lm(log(hate_crimes_per_100k_splc)~.-state,data=hate_nona_df)
summary(full_nodc)
```

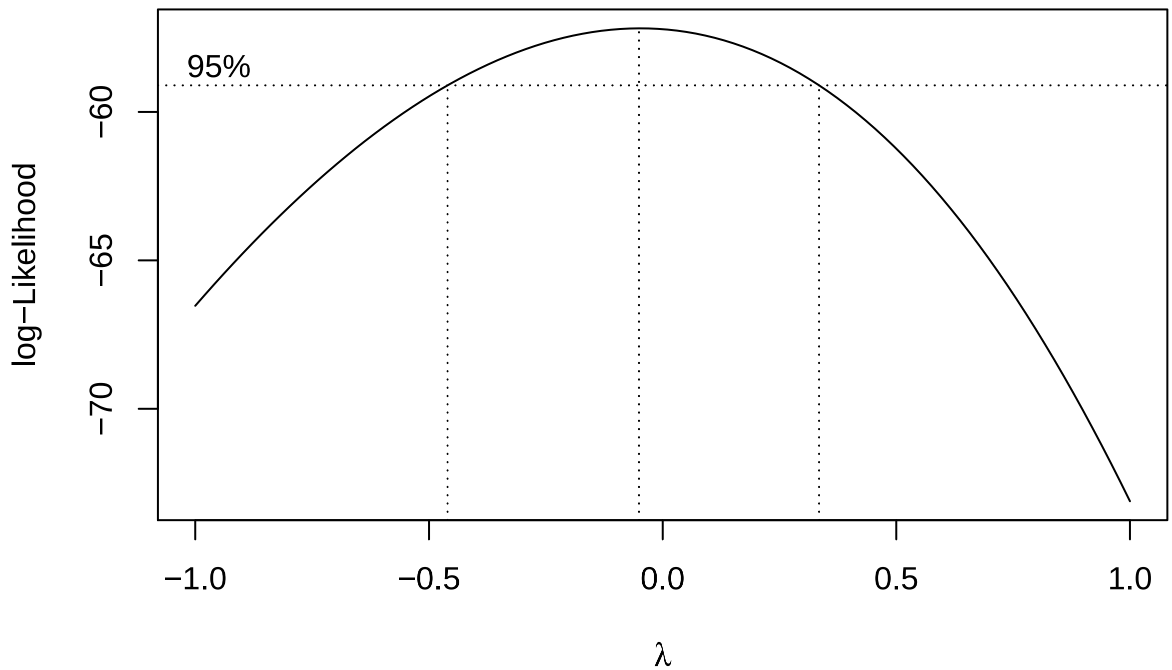
```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ . - state, data = hate_nona_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28845 -0.41144  0.01898  0.31334  1.13022
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -1.857e+01  5.553e+00  -3.344  0.00190
## unemploymentlow                        2.179e-01  2.088e-01   1.043  0.30353
## urbanizationlow                       -9.885e-02  2.467e-01  -0.401  0.69092
## median_household_income               -4.732e-06  1.735e-05  -0.273  0.78658
## perc_population_with_high_school_degree  1.121e+01  5.341e+00   2.098  0.04275
## perc_non_citizen                       1.168e+00  5.464e+00   0.214  0.83189
## gini_index                             1.670e+01  5.744e+00   2.908  0.00611
## perc_non_white                        -1.232e-01  1.069e+00  -0.115  0.90887
##
## (Intercept)                            **
## unemploymentlow
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree *
## perc_non_citizen
## gini_index                             **
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5862 on 37 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.1849
## F-statistic: 2.426 on 7 and 37 DF,  p-value: 0.03768
```

Use Box-Cox method

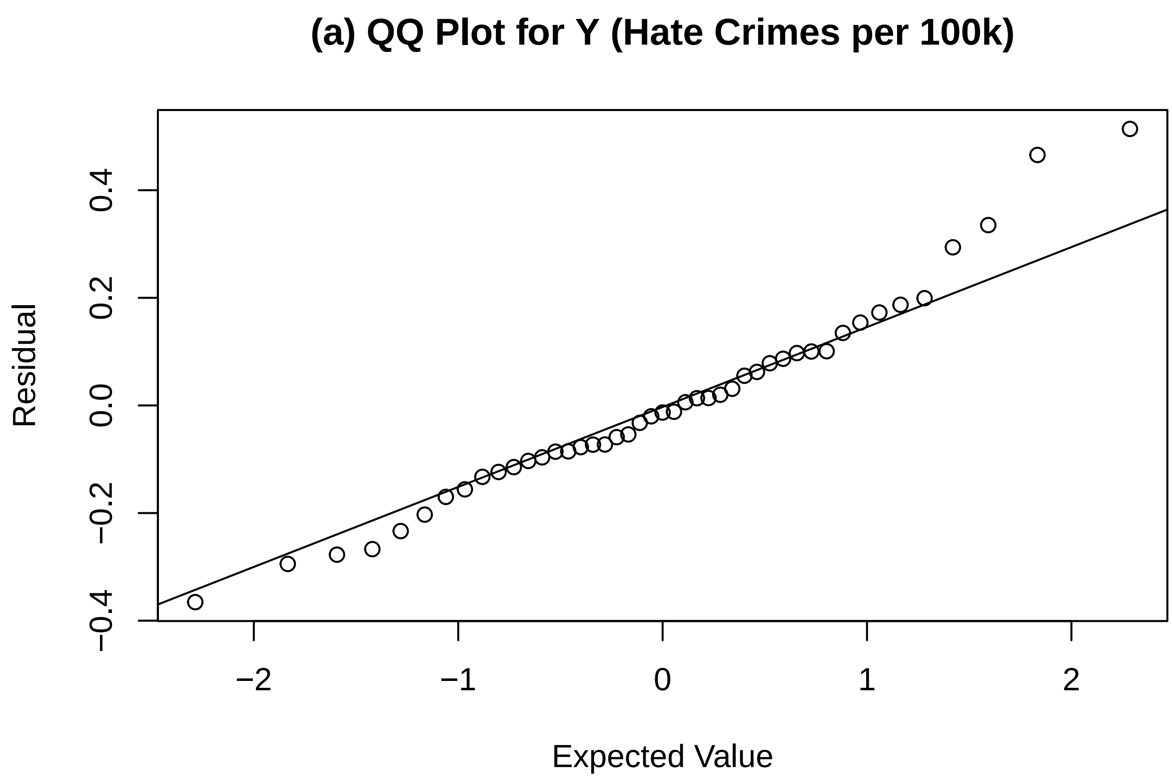
```
MASS::boxcox(full_lm) # default grid of lambdas is -2 to 2 by 0.1
```



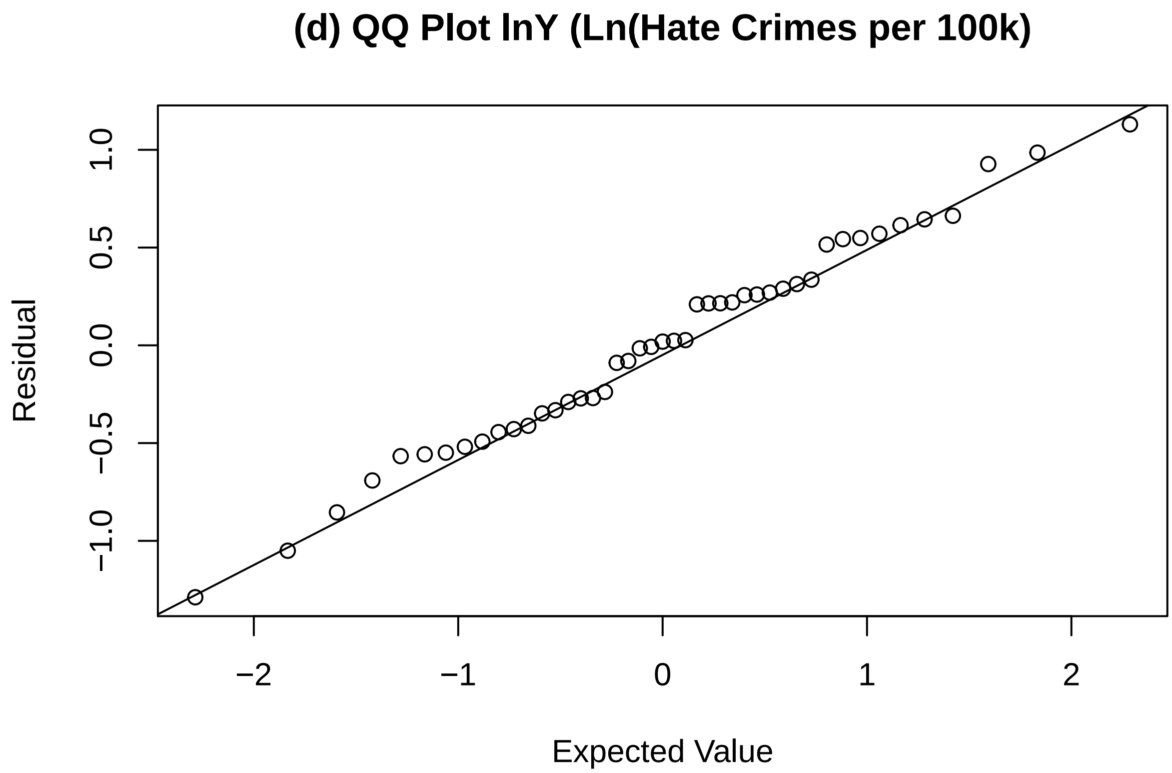

```r
# Could change grid of lambda values just to zoom in, get more precise
MASS::boxcox(full_lm, lambda = seq(-1, 1, by=0.05) )
```




```r
## i'll just QQ plot these as she does in lecture

qqnorm(resid(full_lm), xlab = "Expected Value", ylab = "Residual", main = "")
qqline(resid(full_lm))
title("(a) QQ Plot for Y (Hate Crimes per 100k)")
```




```r
qqnorm(resid(full_log_lm), xlab = "Expected Value", ylab = "Residual", main = "")
qqline(resid(full_log_lm))
title("(d) QQ Plot lnY (Ln(Hate Crimes per 100k)")
```




```
#Note that this is not strictly better. It's up for interpretation.
```

The optimal value of Yˆa is near 0, indicating that a natural log transformation of the outcome for all practical intents and purposes is optimal. We proceed with the log transformation.

Check for interactions