Final Project

```r
hc_df = read_csv("./data/HateCrimes.csv") %>%
  janitor::clean_names()
```

```
## Parsed with column specification:
## cols(
##   state = col_character(),
##   unemployment = col_character(),
##   urbanization = col_character(),
##   median_household_income = col_double(),
##   perc_population_with_high_school_degree = col_double(),
##   perc_non_citizen = col_double(),
##   gini_index = col_double(),
##   perc_non_white = col_double(),
##   hate_crimes_per_100k_splc = col_character()
## )
```

```r
hc_df[hc_df == "N/A"] = NA

hc_df = hc_df %>%
  mutate(hate_crimes_per_100k_splc = as.numeric(hate_crimes_per_100k_splc),
         urbanization = as.factor(urbanization),
         unemployment = as.factor(unemployment)) %>%
  na.omit()

fit1 = lm(hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income + perc_popul

summary(fit1)
```

```
##
## 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 + gini_index *
##     urbanization, data = hc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37526 -0.09740 -0.03516  0.09331  0.52712
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -8.412e+00  1.946e+00  -4.322 0.000117
## unemploymentlow                         1.432e-02  7.258e-02   0.197 0.844673
## urbanizationlow                        -8.459e-01  1.978e+00  -0.428 0.671479
## median_household_income                -1.318e-06  6.041e-06  -0.218 0.828546
## perc_population_with_high_school_degree  5.661e+00  1.958e+00   2.890 0.006482
## perc_non_citizen                        1.397e+00  1.933e+00   0.723 0.474501
## gini_index                               8.311e+00  2.115e+00   3.929 0.000370
## perc_non_white                          -1.344e-02  3.718e-01  -0.036 0.971353
## urbanizationlow:gini_index               1.965e+00  4.419e+00   0.445 0.659157
##
## (Intercept)                             ***
## unemploymentlow
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree **
## perc_non_citizen
## gini_index                              ***
## perc_non_white
## urbanizationlow:gini_index
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2036 on 36 degrees of freedom
## Multiple R-squared:  0.464,  Adjusted R-squared:  0.3448
## F-statistic: 3.895 on 8 and 36 DF,  p-value: 0.002136
```

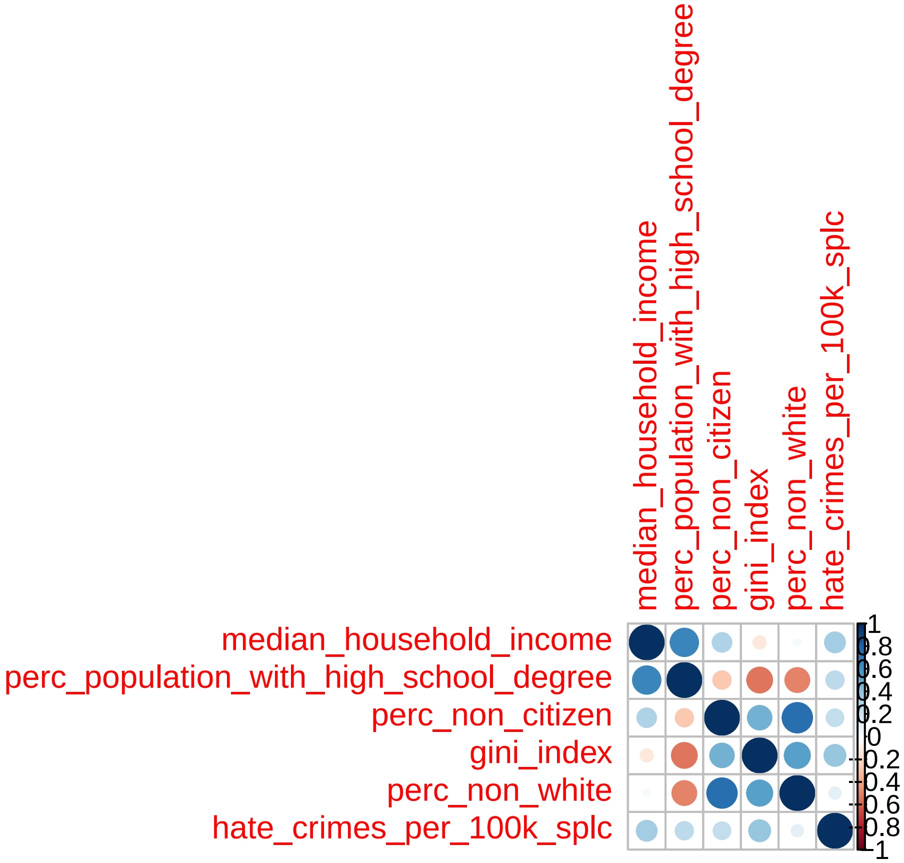
```
vif(fit1)
```

```
##                         unemploymentlow                         urbanizationlow
##                                1.428645                             1057.044019
##                 median_household_income perc_population_with_high_school_degree
##                                3.123157                                4.341070
##                        perc_non_citizen                              gini_index
##                                3.869728                                2.074331
##                          perc_non_white              urbanizationlow:gini_index
##                                3.243274                             1052.161495
```

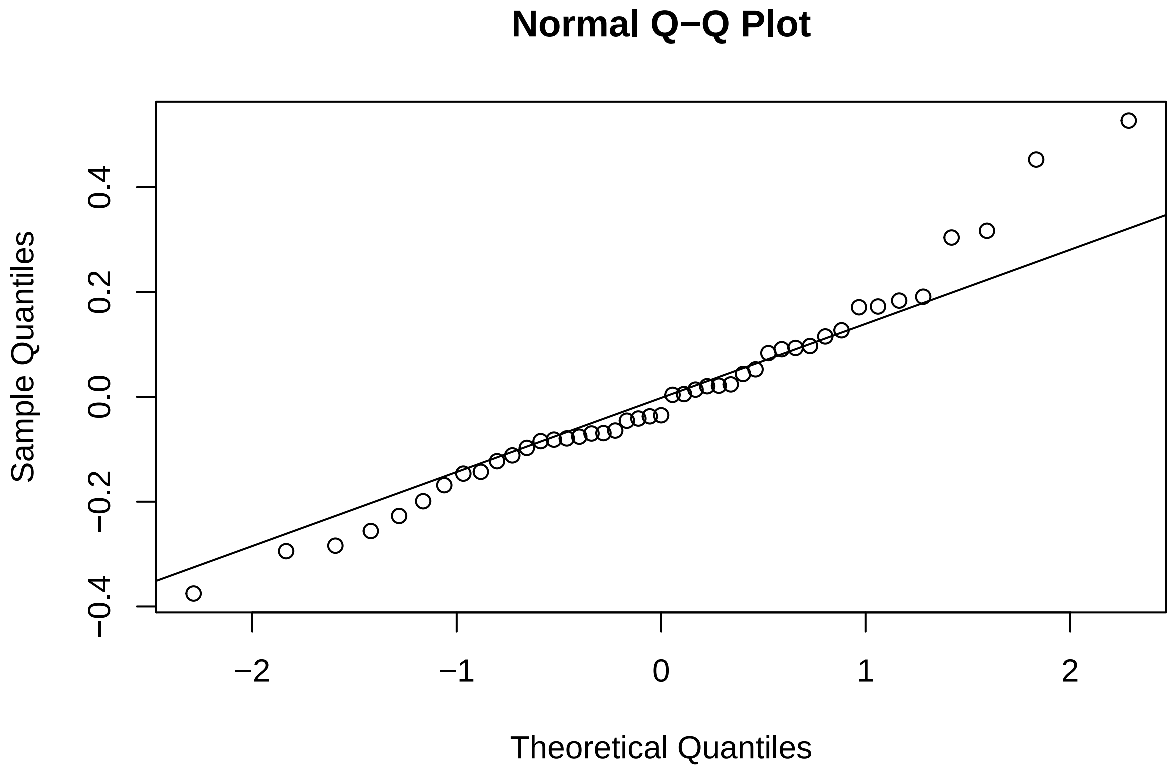
```
correlation_matrix =
cor(hc_df[, sapply(hc_df, is.numeric)],
    use = "complete.obs", method = "pearson")

correlation_matrix
```

```
##                                         median_household_income
## median_household_income                              1.00000000
## perc_population_with_high_school_degree              0.65113832
## perc_non_citizen                                     0.30173941
## gini_index                                          -0.12952158
## perc_non_white                                       0.03905399
## hate_crimes_per_100k_splc                            0.34378921
##                                         perc_population_with_high_school_degree
## median_household_income                                               0.6511383
## perc_population_with_high_school_degree                               1.0000000
## perc_non_citizen                                                     -0.2621288
## gini_index                                                           -0.5371591
## perc_non_white                                                       -0.4958932
## hate_crimes_per_100k_splc                                             0.2628198
##                                         perc_non_citizen gini_index
## median_household_income                        0.3017394 -0.1295216
## perc_population_with_high_school_degree       -0.2621288 -0.5371591
## perc_non_citizen                               1.0000000  0.4798976
## gini_index                                     0.4798976  1.0000000
## perc_non_white                                 0.7526102  0.5484035
## hate_crimes_per_100k_splc                      0.2435066  0.3805028
##                                         perc_non_white
## median_household_income                     0.03905399
## perc_population_with_high_school_degree    -0.49589321
## perc_non_citizen                            0.75261020
## gini_index                                  0.54840351
## perc_non_white                              1.00000000
## hate_crimes_per_100k_splc                   0.11116503
##                                         hate_crimes_per_100k_splc
## median_household_income                                 0.3437892
## perc_population_with_high_school_degree                 0.2628198
## perc_non_citizen                                        0.2435066
## gini_index                                              0.3805028
## perc_non_white                                          0.1111650
## hate_crimes_per_100k_splc                               1.0000000
```

```
correlation_plt =
  corrplot(correlation_matrix)
```



```
qqnorm(resid(fit1))
qqline(resid(fit1))
```



```
step(fit1, direction = "backward")
```

```
## Start:  AIC=-135.28
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + perc_non_white + gini_index * urbanization
##
##                                           Df Sum of Sq    RSS     AIC
## - perc_non_white                           1   0.00005 1.4926 -137.28
## - unemployment                             1   0.00161 1.4942 -137.23
## - median_household_income                  1   0.00197 1.4946 -137.22
## - urbanization:gini_index                  1   0.00820 1.5008 -137.03
## - perc_non_citizen                         1   0.02166 1.5142 -136.63
## <none>                                                 1.4926 -135.28
## - perc_population_with_high_school_degree  1   0.34638 1.8390 -127.89
##
## Step:  AIC=-137.28
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + urbanization:gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - unemployment                             1   0.00186 1.4945 -139.22
## - median_household_income                  1   0.00212 1.4948 -139.21
## - urbanization:gini_index                  1   0.00816 1.5008 -139.03
## - perc_non_citizen                         1   0.02904 1.5217 -138.41
## <none>                                                 1.4926 -137.28
## - perc_population_with_high_school_degree  1   0.37841 1.8710 -129.11
##
## Step:  AIC=-139.22
## hate_crimes_per_100k_splc ~ urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + urbanization:gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - median_household_income                  1   0.00187 1.4964 -141.16
## - urbanization:gini_index                  1   0.00779 1.5023 -140.99
## - perc_non_citizen                         1   0.02866 1.5232 -140.37
## <none>                                                 1.4945 -139.22
## - perc_population_with_high_school_degree  1   0.39187 1.8863 -130.74
##
## Step:  AIC=-141.16
## hate_crimes_per_100k_splc ~ urbanization + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index + urbanization:gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - urbanization:gini_index                  1   0.00834 1.5047 -142.91
## - perc_non_citizen                         1   0.02813 1.5245 -142.32
## <none>                                                 1.4964 -141.16
## - perc_population_with_high_school_degree  1   0.67549 2.1719 -126.40
##
## 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
## - 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
## - 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_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 = hc_df)
##
## Coefficients:
##                             (Intercept)
##                                  -8.103
## perc_population_with_high_school_degree
##                                   5.059
##                              gini_index
##                                   8.825
```

```r
step(fit1, direction = "both")
```

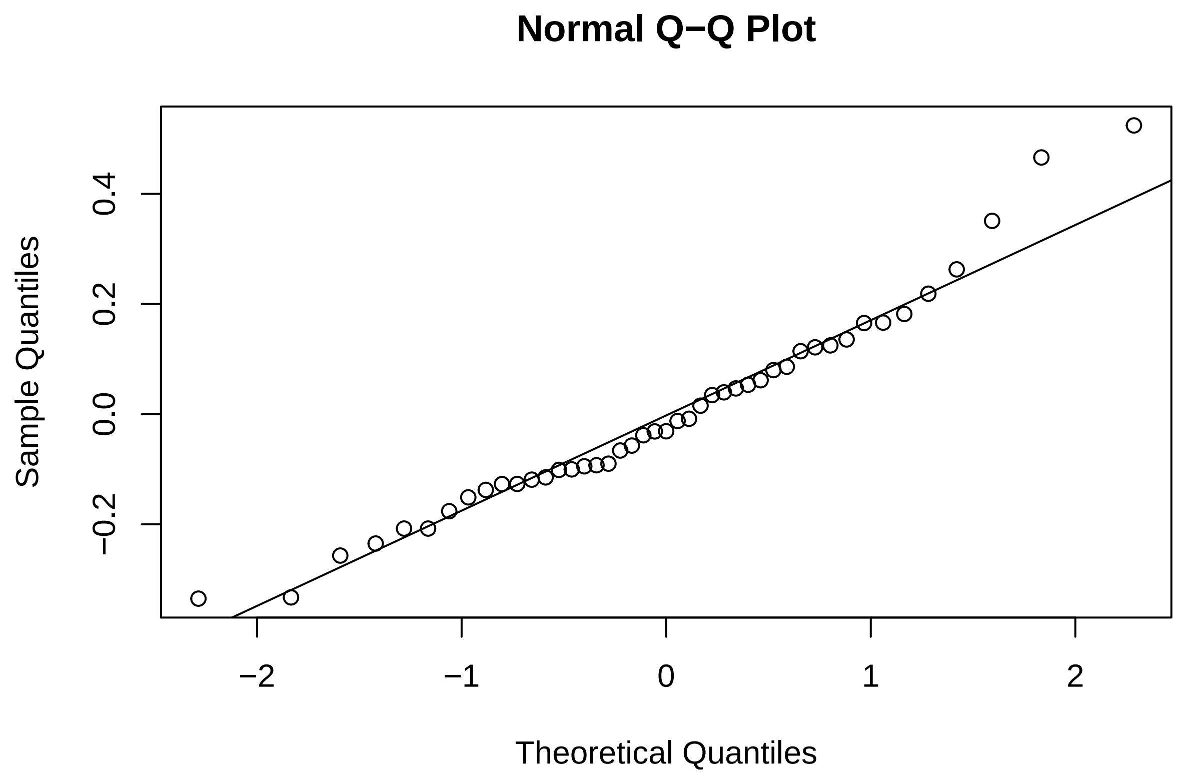
```
## Start:  AIC=-135.28
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + perc_non_white + gini_index * urbanization
##
##                                           Df Sum of Sq    RSS     AIC
## - perc_non_white                           1   0.00005 1.4926 -137.28
## - unemployment                             1   0.00161 1.4942 -137.23
## - median_household_income                  1   0.00197 1.4946 -137.22
## - urbanization:gini_index                  1   0.00820 1.5008 -137.03
## - perc_non_citizen                         1   0.02166 1.5142 -136.63
## <none>                                                 1.4926 -135.28
## - perc_population_with_high_school_degree  1   0.34638 1.8390 -127.89
##
## Step:  AIC=-137.28
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + urbanization:gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - unemployment                             1   0.00186 1.4945 -139.22
## - median_household_income                  1   0.00212 1.4948 -139.21
## - urbanization:gini_index                  1   0.00816 1.5008 -139.03
## - perc_non_citizen                         1   0.02904 1.5217 -138.41
## <none>                                                 1.4926 -137.28
## + perc_non_white                           1   0.00005 1.4926 -135.28
## - perc_population_with_high_school_degree  1   0.37841 1.8710 -129.11
##
## Step:  AIC=-139.22
## hate_crimes_per_100k_splc ~ urbanization + median_household_income +
##     perc_population_with_high_school_degree + perc_non_citizen +
##     gini_index + urbanization:gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - median_household_income                  1   0.00187 1.4964 -141.16
## - urbanization:gini_index                  1   0.00779 1.5023 -140.99
## - perc_non_citizen                         1   0.02866 1.5232 -140.37
## <none>                                                 1.4945 -139.22
## + unemployment                             1   0.00186 1.4926 -137.28
## + perc_non_white                           1   0.00030 1.4942 -137.23
## - perc_population_with_high_school_degree  1   0.39187 1.8863 -130.74
##
## Step:  AIC=-141.16
## hate_crimes_per_100k_splc ~ urbanization + perc_population_with_high_school_degree +
##     perc_non_citizen + gini_index + urbanization:gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - urbanization:gini_index                  1   0.00834 1.5047 -142.91
## - perc_non_citizen                         1   0.02813 1.5245 -142.32
## <none>                                                 1.4964 -141.16
## + median_household_income                  1   0.00187 1.4945 -139.22
## + unemployment                             1   0.00161 1.4948 -139.21
## + perc_non_white                           1   0.00052 1.4958 -139.18
## - perc_population_with_high_school_degree  1   0.67549 2.1719 -126.40
##
## 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
## + urbanization:gini_index                  1   0.00834 1.4964 -141.16
## + 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 = hc_df)
##
## Coefficients:
##                              (Intercept)
##                                   -8.103
## perc_population_with_high_school_degree
##                                    5.059
##                               gini_index
##                                    8.825
```

```
fit_after_step =
  lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
    gini_index, data = hc_df)

summary(fit_after_step)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index, data = hc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33490 -0.11891 -0.03105  0.11430  0.52418
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               -8.103      1.447  -5.601 1.48e-06
## perc_population_with_high_school_degree    5.059      1.044   4.847 1.74e-05
## gini_index                                 8.825      1.630   5.413 2.76e-06
##
## (Intercept)                             ***
## perc_population_with_high_school_degree ***
## gini_index                              ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1907 on 42 degrees of freedom
## Multiple R-squared:  0.4516, Adjusted R-squared:  0.4255
## F-statistic: 17.29 on 2 and 42 DF,  p-value: 3.32e-06
```

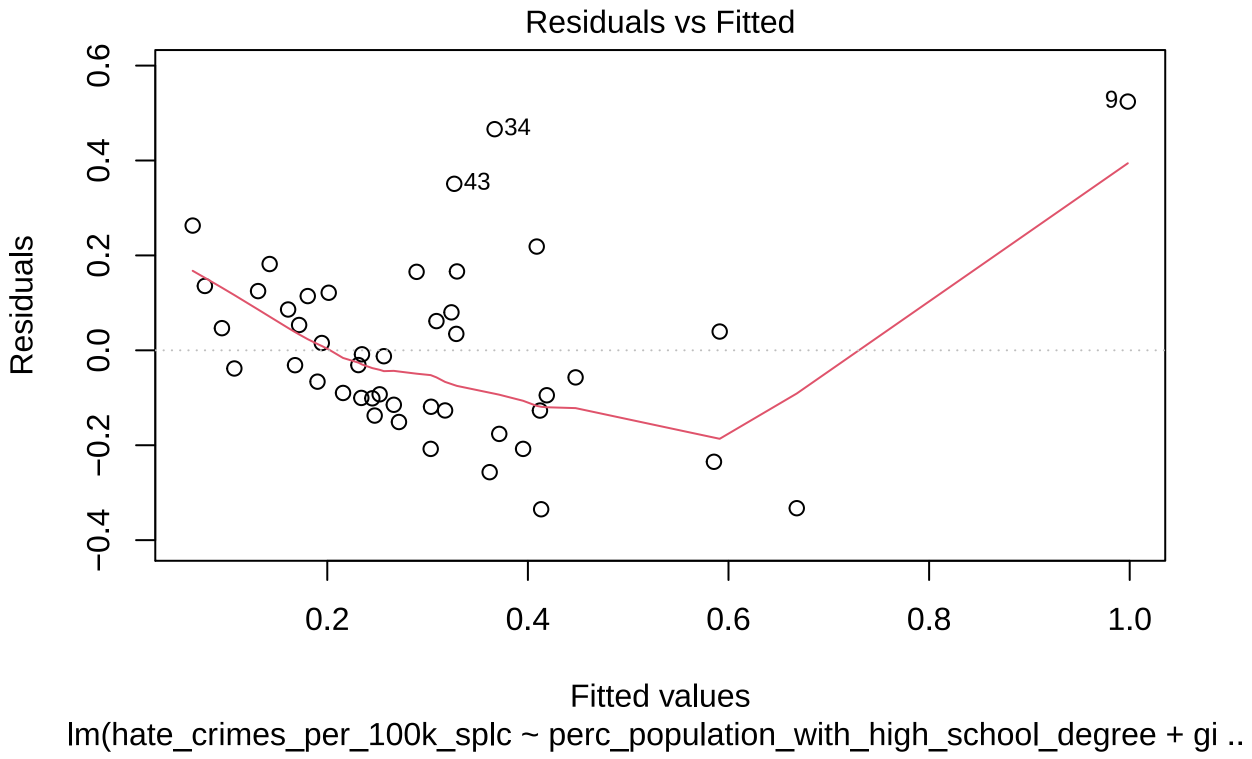
```
qqnorm(resid(fit_after_step))
qqline(resid(fit_after_step))
```

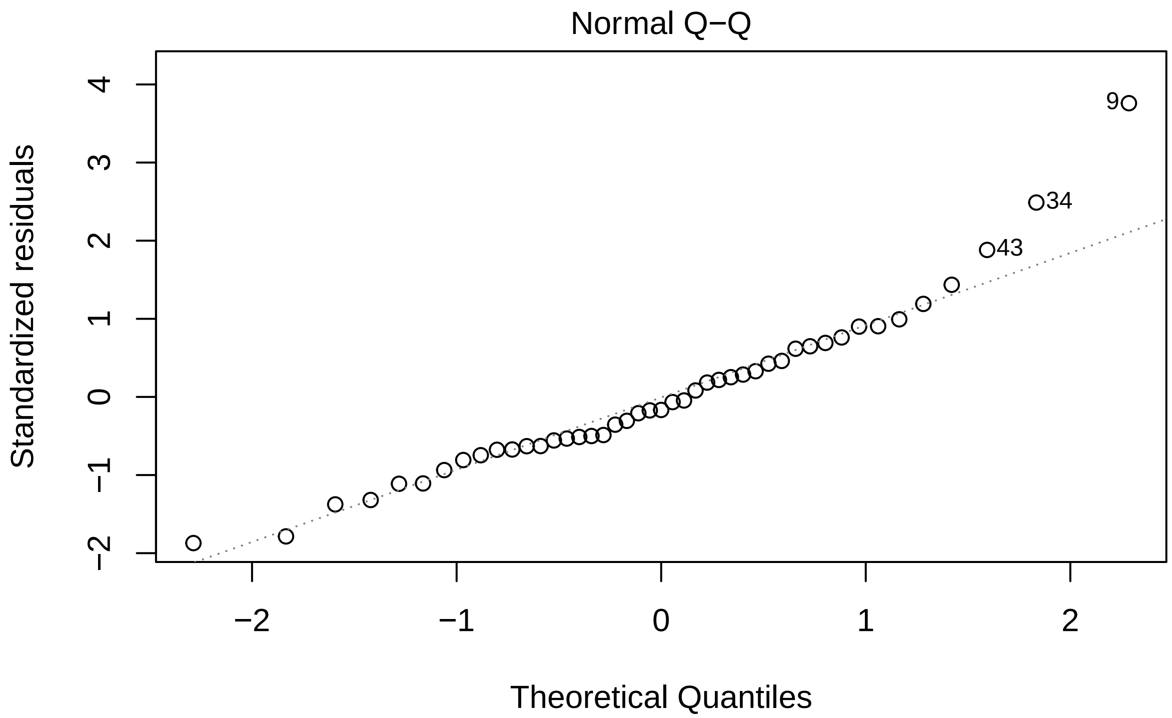


```r
vif(fit_after_step)
```

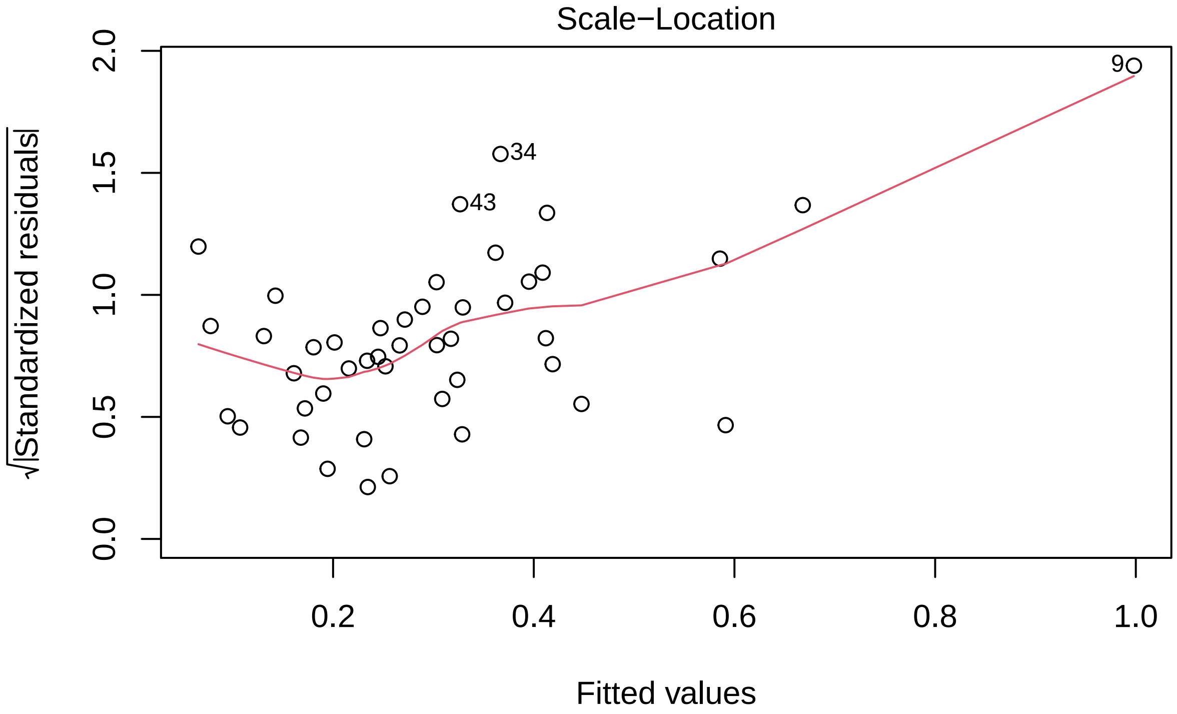
```
## perc_population_with_high_school_degree                              gini_index
##                                 1.40556                                 1.40556
```

```r
plot(fit_after_step)
```

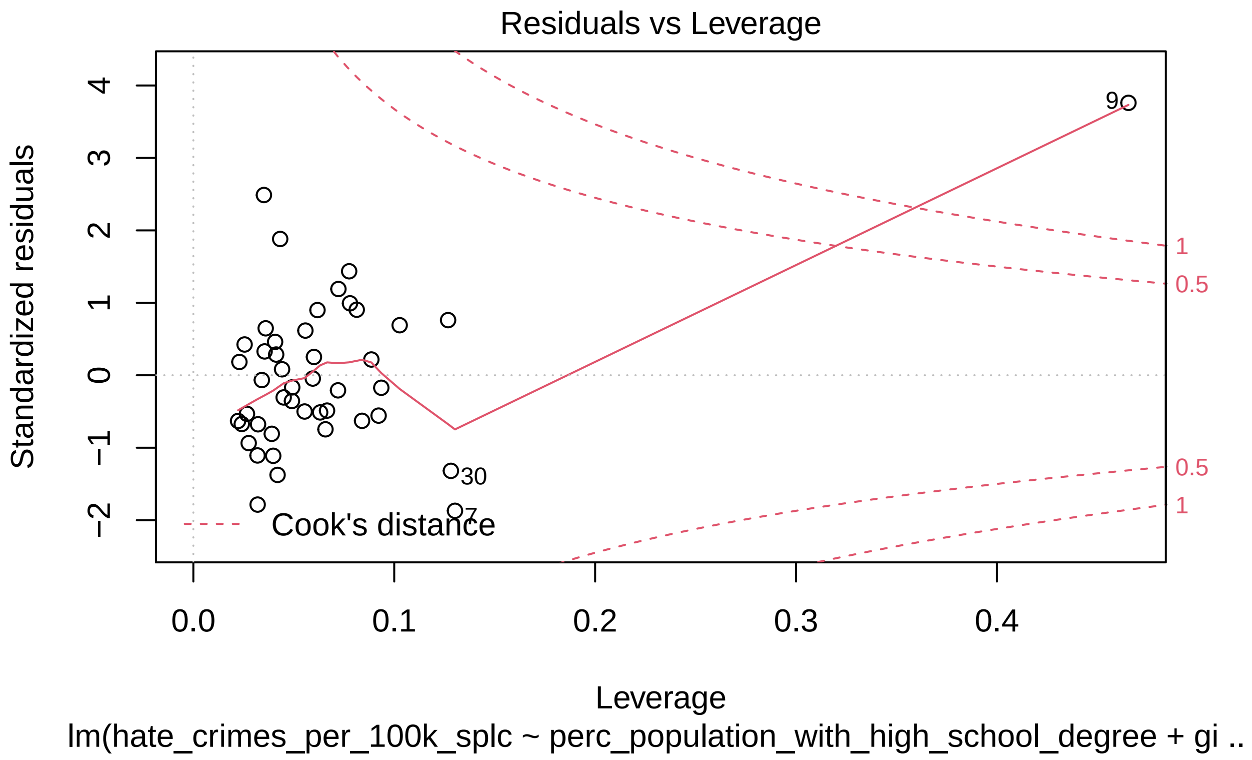




lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..



lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..



Exclude outliers

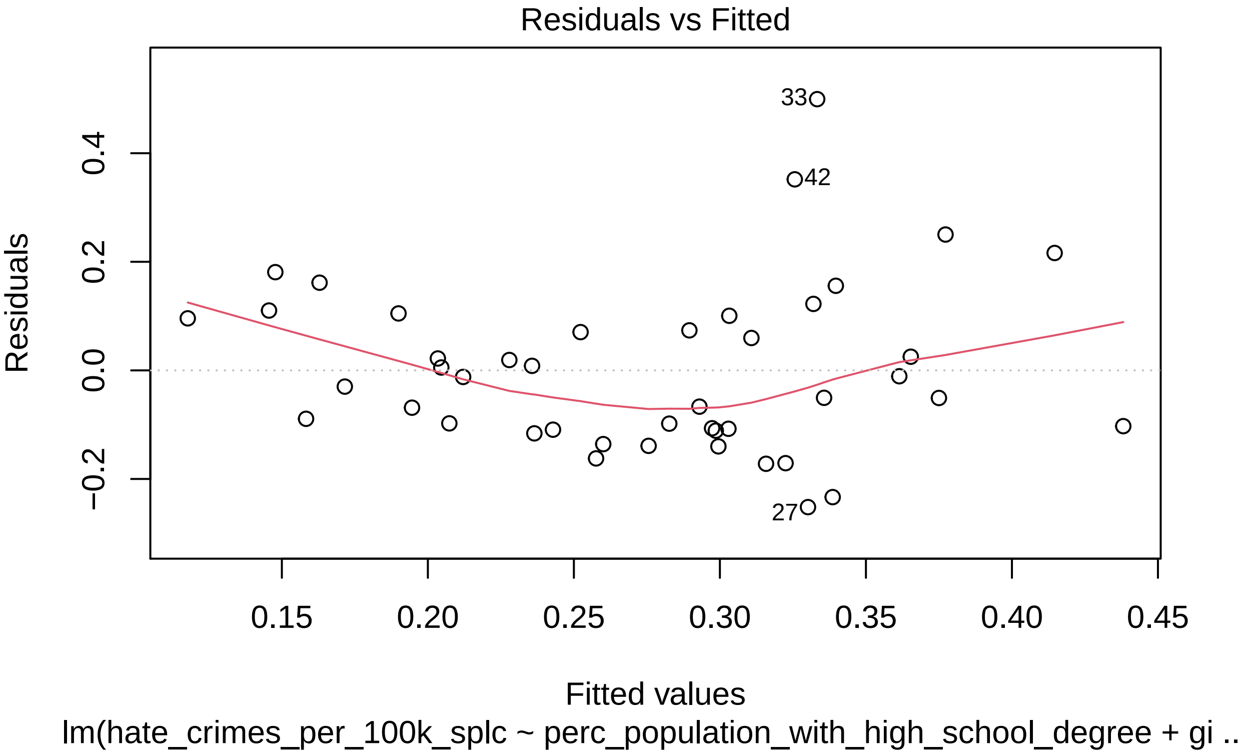
```
# take a look at the outlier
hc_df_only_9 = hc_df[c(9),]
hc_df_only_9
```

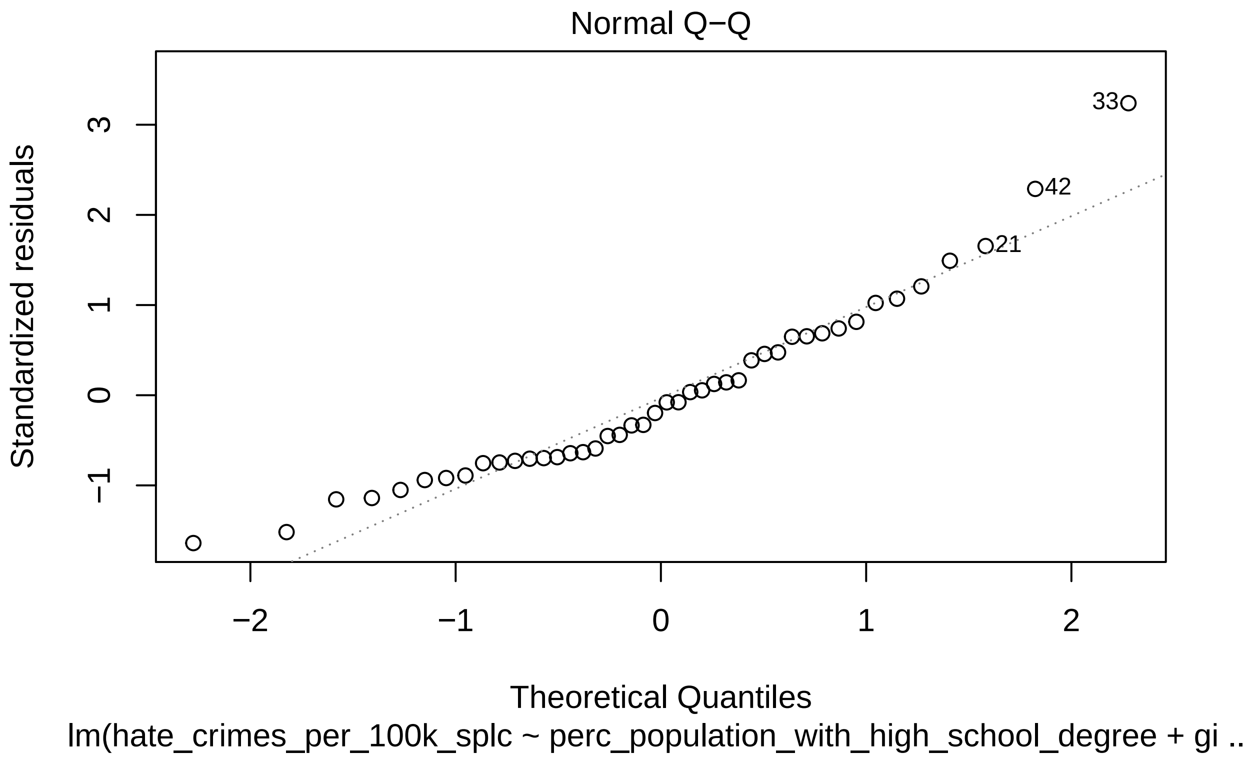
```
## # A tibble: 1 x 9
##   state unemployment urbanization median_househol~ perc_population~
##   <chr> <fct>        <fct>                   <dbl>            <dbl>
## 1 Dist~ high         high                    68277            0.871
## # ... with 4 more variables: perc_non_citizen <dbl>, gini_index <dbl>,
## #   perc_non_white <dbl>, hate_crimes_per_100k_splc <dbl>
```

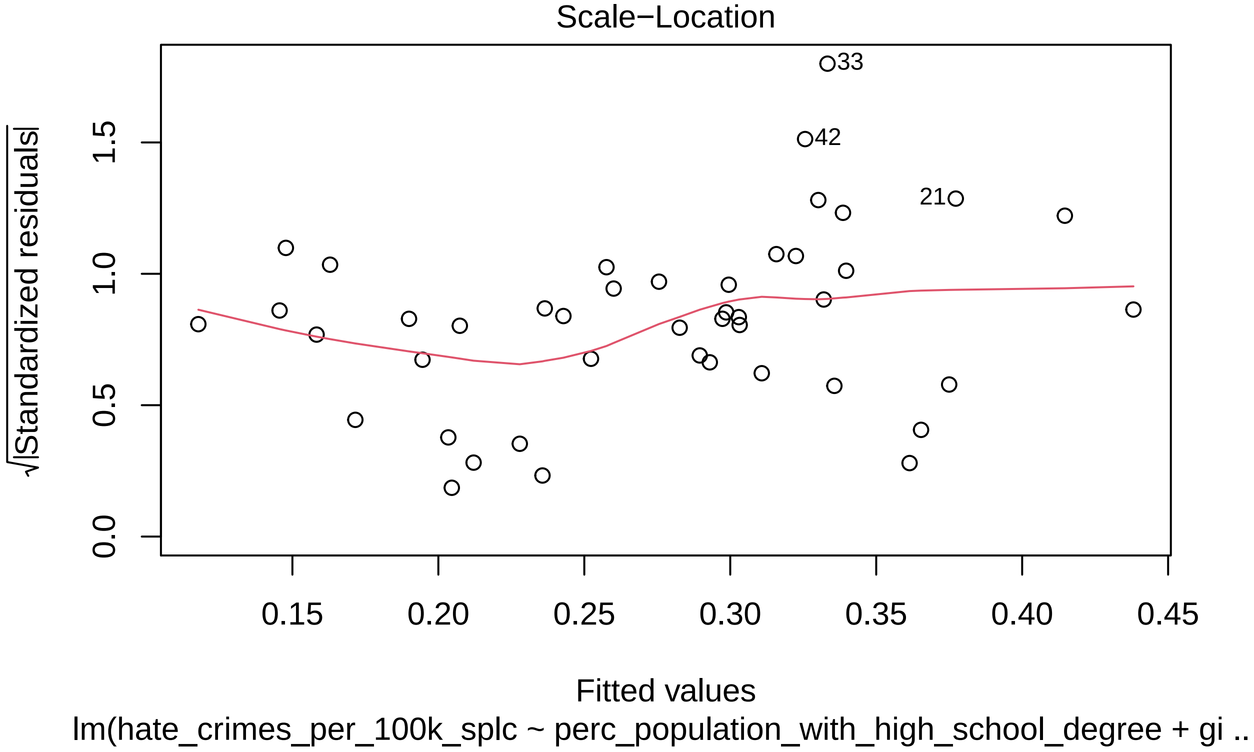
```
# data frame without the outlier
hc_df_no_outliers = hc_df[c(-9),]

fit_no_9 = lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
    gini_index, data = hc_df_no_outliers)

summary(fit_no_9)
```

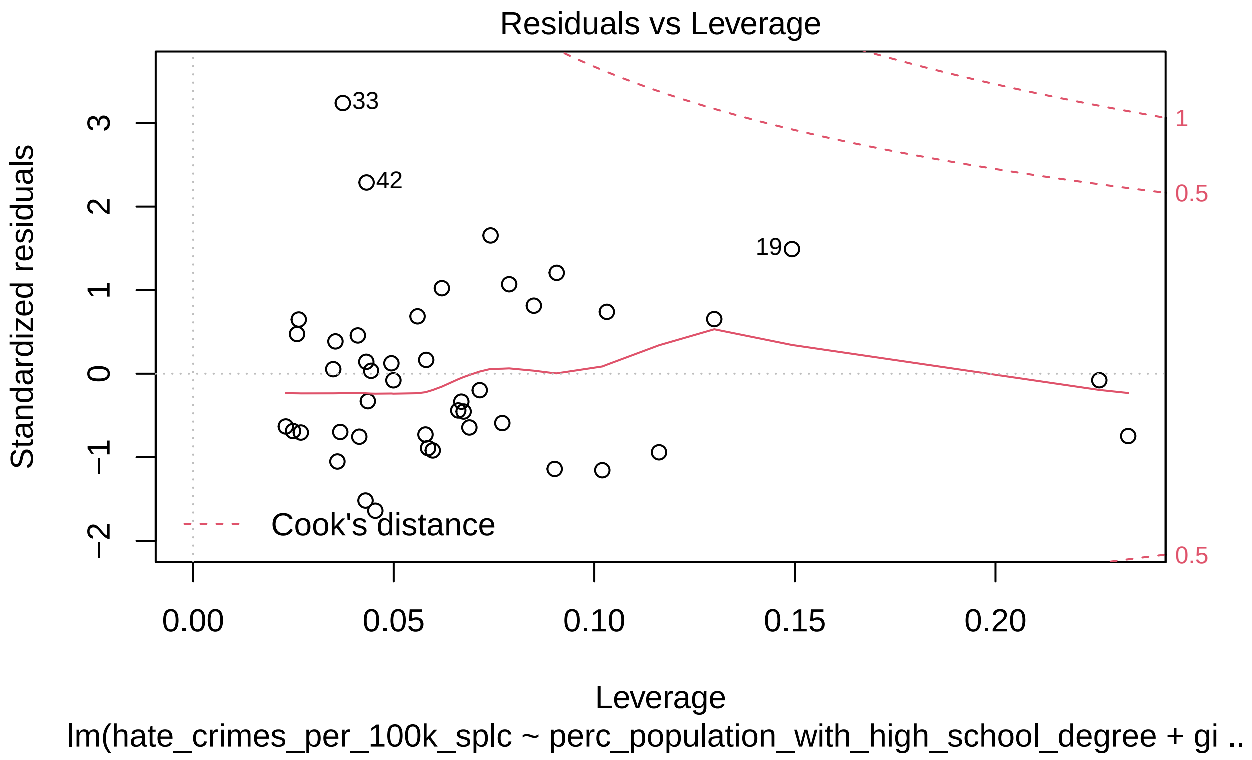
```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index, data = hc_df_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25186 -0.10799 -0.02101  0.09700  0.49954
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              -3.8396     1.5151  -2.534  0.01519 *
## perc_population_with_high_school_degree   3.0482     0.9666   3.154  0.00302 **
## gini_index                                3.2449     1.8174   1.785  0.08159 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1572 on 41 degrees of freedom
## Multiple R-squared:  0.1977, Adjusted R-squared:  0.1585
## F-statistic: 5.051 on 2 and 41 DF,  p-value: 0.01094
```

```
plot(fit_no_9)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
33
42
21
−2
−1
0
1
2
Theoretical Quantiles
lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
33
42
21
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Fitted values
lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..



```
# 95% confidence intervals
confint(fit_no_9, "perc_population_with_high_school_degree", 0.95)
```

```
##                                            2.5 %  97.5 %
## perc_population_with_high_school_degree 1.096097 5.00024
```

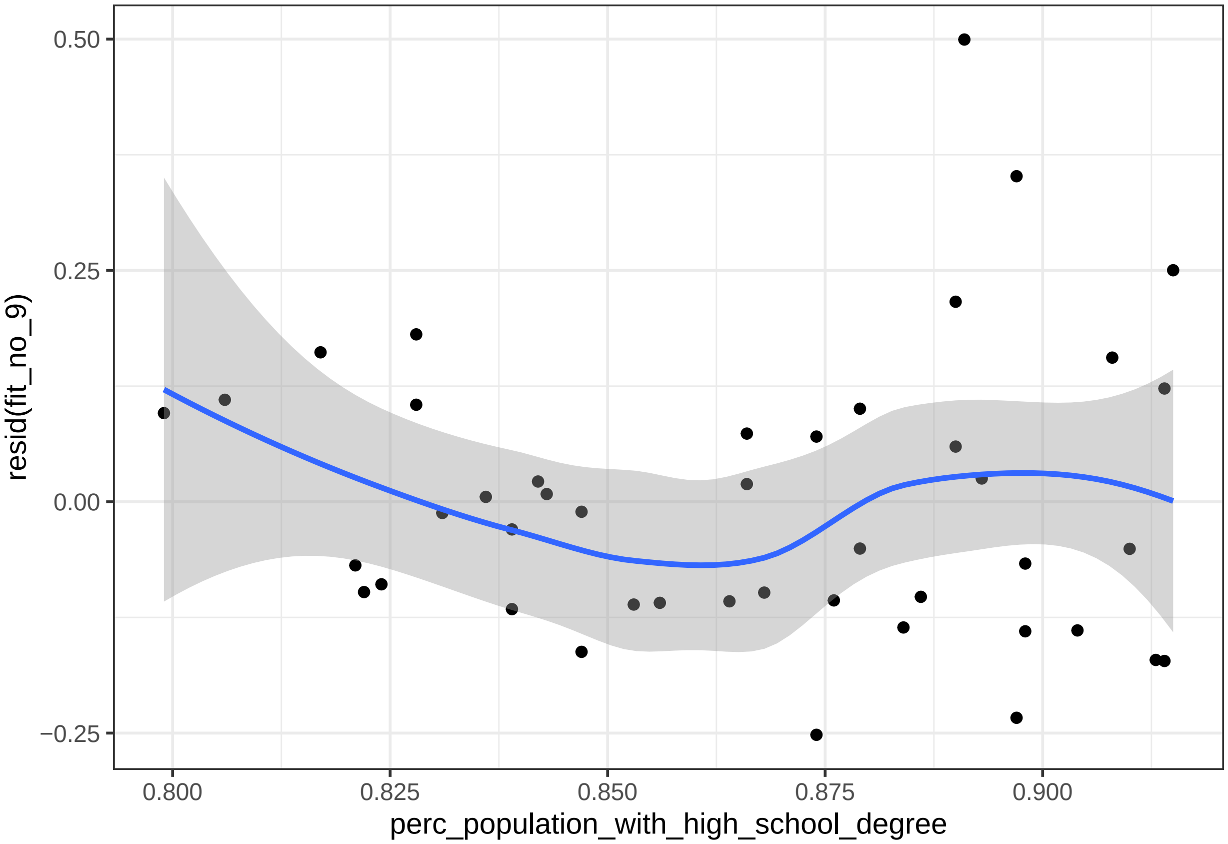
```
confint(fit_no_9, "gini_index", 0.95)
```

```
##                  2.5 %   97.5 %
## gini_index -0.4255136 6.915284
```

This line of observation has a hate_crime_per_100k_splc greater than 100%, which is absurd. There was probably a mistake. Excluding this observation makes gini_index an insignificant predictor.

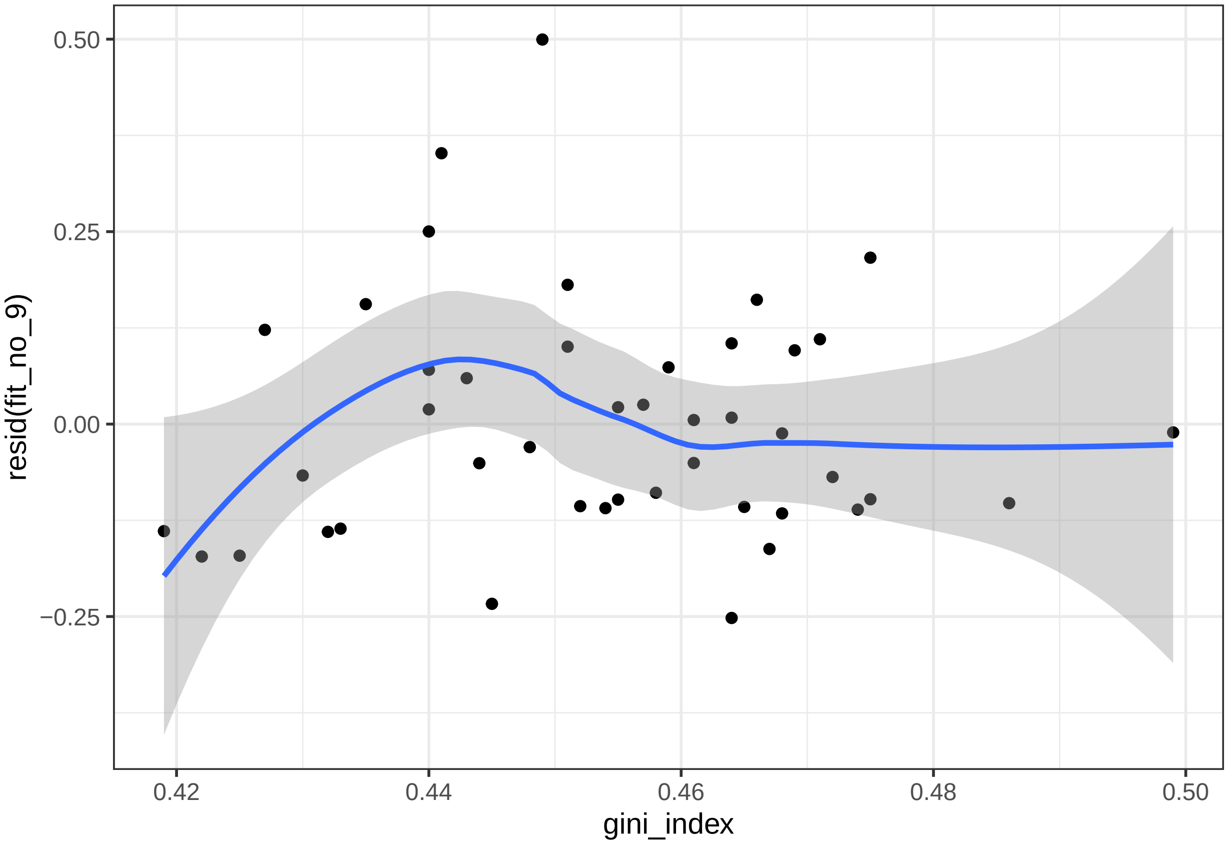
```
# Residuals vs. Covariates plots

hc_df_no_outliers %>%
  ggplot(aes(x = perc_population_with_high_school_degree,
             y = resid(fit_no_9))) +
  geom_point() +
  geom_smooth() +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
hc_df_no_outliers %>%
  ggplot(aes(x = gini_index,
             y = resid(fit_no_9))) +
  geom_point() +
  geom_smooth() +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Deal with collinearity

```
vif(fit_no_9)
```

```
## perc_population_with_high_school_degree                                                gini_index
##                                1.773775                                                  1.773775
```

No multicollinearity issues.

```
fit = lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
    gini_index * urbanization + gini_index, data = hc_df)
summary(fit)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index * urbanization + gini_index, data = hc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33743 -0.12559 -0.02552  0.10649  0.53288
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              -8.1782     1.5853  -5.159 7.13e-06
## perc_population_with_high_school_degree   5.2314     1.2693   4.121 0.000184
## gini_index                                8.6674     1.9173   4.521 5.38e-05
## urbanizationlow                          -0.4743     1.8377  -0.258 0.797673
## gini_index:urbanizationlow                1.0506     4.0781   0.258 0.798014
##
## (Intercept)                             ***
## perc_population_with_high_school_degree ***
## gini_index                              ***
## urbanizationlow
## gini_index:urbanizationlow
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1952 on 40 degrees of freedom
## Multiple R-squared:  0.4525, Adjusted R-squared:  0.3978
## F-statistic: 8.265 on 4 and 40 DF,  p-value: 5.885e-05
```

```
step(fit, direction = "both")
```

```
## Start:  AIC=-142.33
## hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index * urbanization + gini_index
##
##                                           Df Sum of Sq    RSS     AIC
## - gini_index:urbanization                  1   0.00253 1.5270 -144.25
## <none>                                                 1.5245 -142.32
## - perc_population_with_high_school_degree  1   0.64736 2.1719 -128.40
##
## Step:  AIC=-144.25
## hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index + urbanization
##
##                                           Df Sum of Sq    RSS     AIC
## - urbanization                             1   0.00001 1.5270 -146.25
## <none>                                                 1.5270 -144.25
## + gini_index:urbanization                  1   0.00253 1.5245 -142.32
## - perc_population_with_high_school_degree  1   0.84616 2.3732 -126.41
## - gini_index                               1   0.88037 2.4074 -125.77
##
## 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
## + 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 = hc_df)
##
## Coefficients:
##                             (Intercept)
##                                  -8.103
## perc_population_with_high_school_degree
##                                    5.059
##                               gini_index
##                                    8.825
```