Final Project

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(broom)
library(faraway)
library(arsenal)
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:faraway':
##
##     melanoma
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
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()
```

```
## )
```

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

hc_df = hc_df %>%
  mutate(hate_crimes_per_100k_splc = as.numeric(hate_crimes_per_100k_splc)) %>%
  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, data = hc_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
```

```
vif(fit1)
```

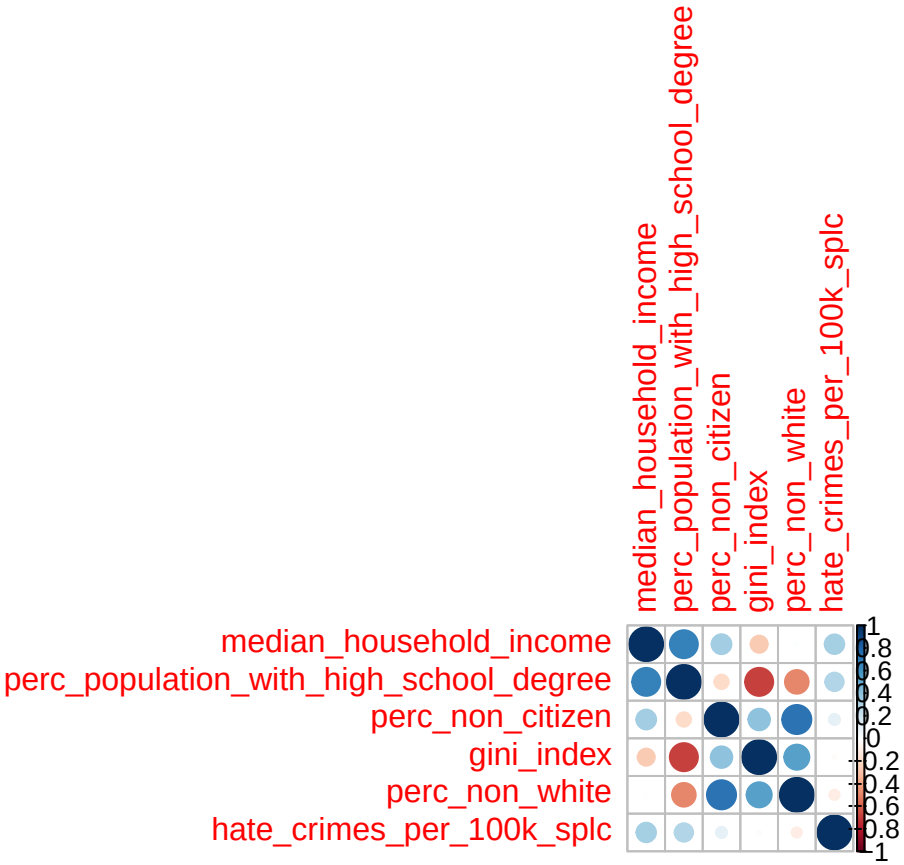
```
##                         unemploymentlow                         urbanizationlow
##                                1.426492                                1.983246
##                 median_household_income perc_population_with_high_school_degree
##                                3.108161                                3.895361
##                        perc_non_citizen                            gini_index
##                                3.728286                                1.845436
```

```
##                           perc_non_white
##                                 3.236419
```

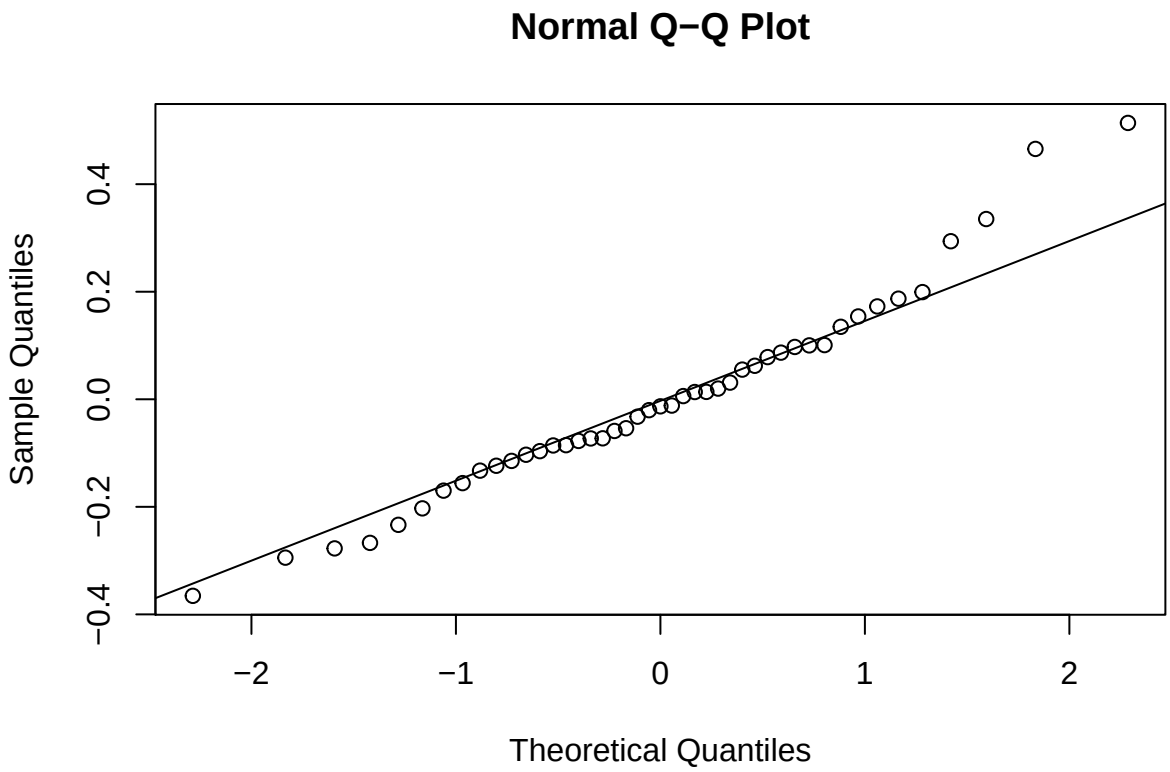
```
correlation_matrix =
cor(hc_df[, sapply(hc_df, is.numeric)],
    use = "complete.obs", method = "spearman")

correlation_matrix
```

```
##                                         median_household_income
## median_household_income                              1.000000000
## perc_population_with_high_school_degree              0.679209263
## perc_non_citizen                                     0.341436579
## gini_index                                          -0.254786328
## perc_non_white                                       0.009227221
## hate_crimes_per_100k_splc                            0.330830040
##                                         perc_population_with_high_school_degree
## median_household_income                                                  0.6792093
## perc_population_with_high_school_degree                                  1.0000000
## perc_non_citizen                                                        -0.1815268
## gini_index                                                              -0.6830933
## perc_non_white                                                          -0.4863359
## hate_crimes_per_100k_splc                                                0.2954201
##                                         perc_non_citizen  gini_index
## median_household_income                        0.3414366 -0.25478633
## perc_population_with_high_school_degree       -0.1815268 -0.68309326
## perc_non_citizen                               1.0000000  0.40251818
## gini_index                                     0.4025182  1.00000000
## perc_non_white                                 0.7356863  0.54585916
## hate_crimes_per_100k_splc                      0.1076195 -0.01080832
##                                         perc_non_white
## median_household_income                     0.009227221
## perc_population_with_high_school_degree    -0.486335922
## perc_non_citizen                            0.735686324
## gini_index                                  0.545859160
## perc_non_white                              1.000000000
## hate_crimes_per_100k_splc                  -0.090558587
##                                         hate_crimes_per_100k_splc
## median_household_income                                0.33083004
## perc_population_with_high_school_degree                0.29542011
## perc_non_citizen                                       0.10761954
## gini_index                                            -0.01080832
## perc_non_white                                        -0.09055859
## hate_crimes_per_100k_splc                              1.00000000
```

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




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



```r
fit2 = lm(hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income + gini_inde

summary(fit2)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ unemployment + urbanization +
##     median_household_income + gini_index, data = hc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34364 -0.12793 -0.03623  0.05166  0.64538
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.192e+00  9.537e-01  -3.347  0.00179 **
## unemploymentlow           5.023e-02  7.418e-02   0.677  0.50217
## urbanizationlow           4.508e-02  7.832e-02   0.576  0.56815
## median_household_income   1.144e-05  4.096e-06   2.793  0.00797 **
## gini_index                6.181e+00  1.887e+00   3.275  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2179 on 40 degrees of freedom
## Multiple R-squared:  0.3182, Adjusted R-squared:  0.2501
## F-statistic: 4.668 on 4 and 40 DF,  p-value: 0.003466
```

```r
fit3 = lm(hate_crimes_per_100k_splc ~ unemployment + median_household_income + gini_index, data = hc_df

summary(fit3)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ unemployment + median_household_income +
##     gini_index, data = hc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34833 -0.10906 -0.05311  0.06369  0.67359
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.923e+00  8.241e-01  -3.546 0.000994 ***
## unemploymentlow           5.776e-02  7.242e-02   0.798 0.429664
## median_household_income   1.054e-05  3.752e-06   2.808 0.007599 **
## gini_index                5.737e+00  1.708e+00   3.358 0.001704 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2161 on 41 degrees of freedom
## Multiple R-squared:  0.3126, Adjusted R-squared:  0.2623
## F-statistic: 6.215 on 3 and 41 DF,  p-value: 0.001404
```

```r
fit4 = lm(hate_crimes_per_100k_splc ~ median_household_income + gini_index, data = hc_df)
```

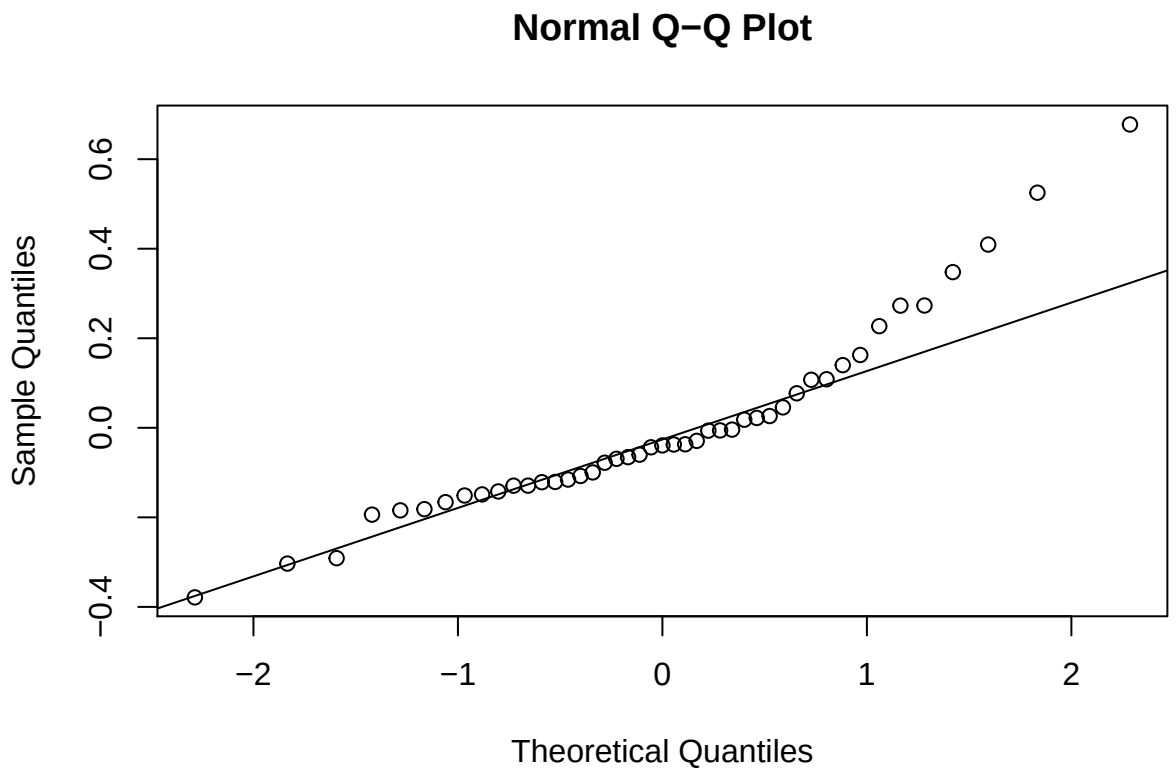
```r
summary(fit4)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ median_household_income +
##     gini_index, data = hc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37873 -0.12917 -0.03933  0.07706  0.67751
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.688e+00  7.663e-01  -3.507  0.00109 **
## median_household_income   1.120e-05  3.642e-06   3.075  0.00369 **
## gini_index                5.203e+00  1.565e+00   3.325  0.00184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2151 on 42 degrees of freedom
## Multiple R-squared:  0.3019, Adjusted R-squared:  0.2687
## F-statistic: 9.083 on 2 and 42 DF,  p-value: 0.0005272
```

```r
vif(fit4)
```

```
## median_household_income              gini_index
##                1.017062                1.017062
```

```r
qqnorm(resid(fit4))
qqline(resid(fit4))
```



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

```
## 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_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
## - 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
## - 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
```

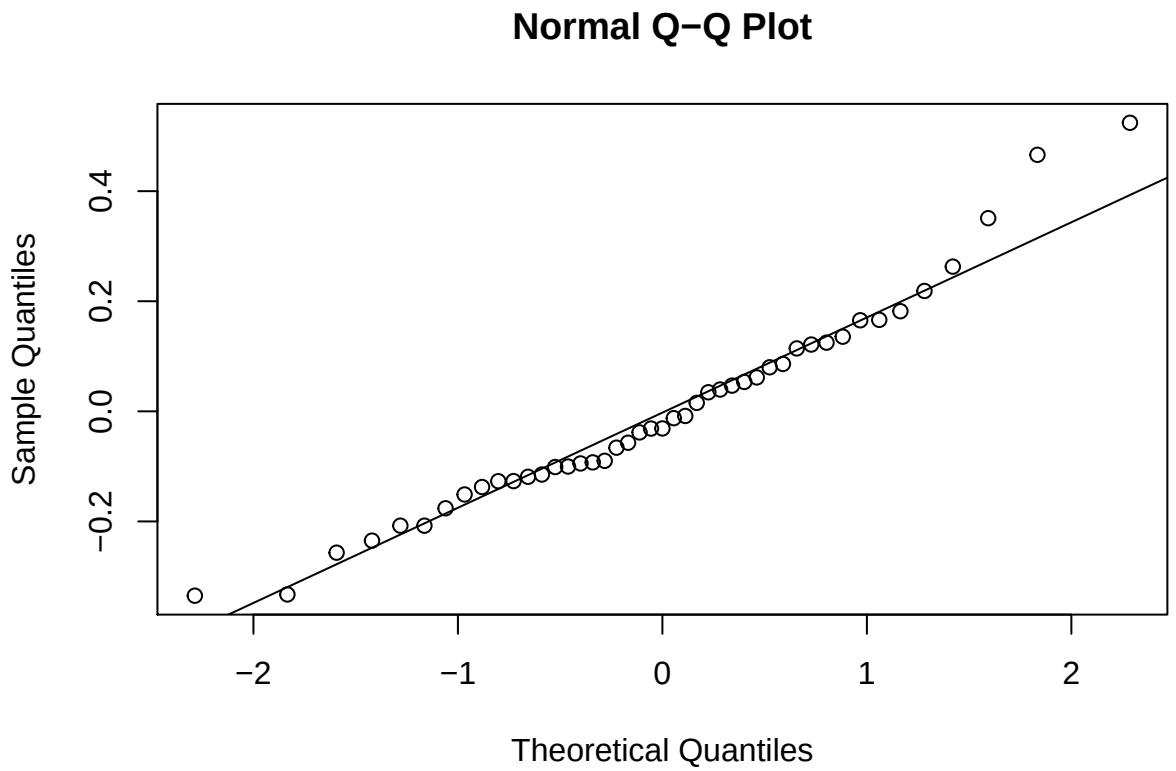
```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index, data = hc_df)
##
## Coefficients:
##                               (Intercept)
##                                    -8.103
## perc_population_with_high_school_degree
##          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

```
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
```

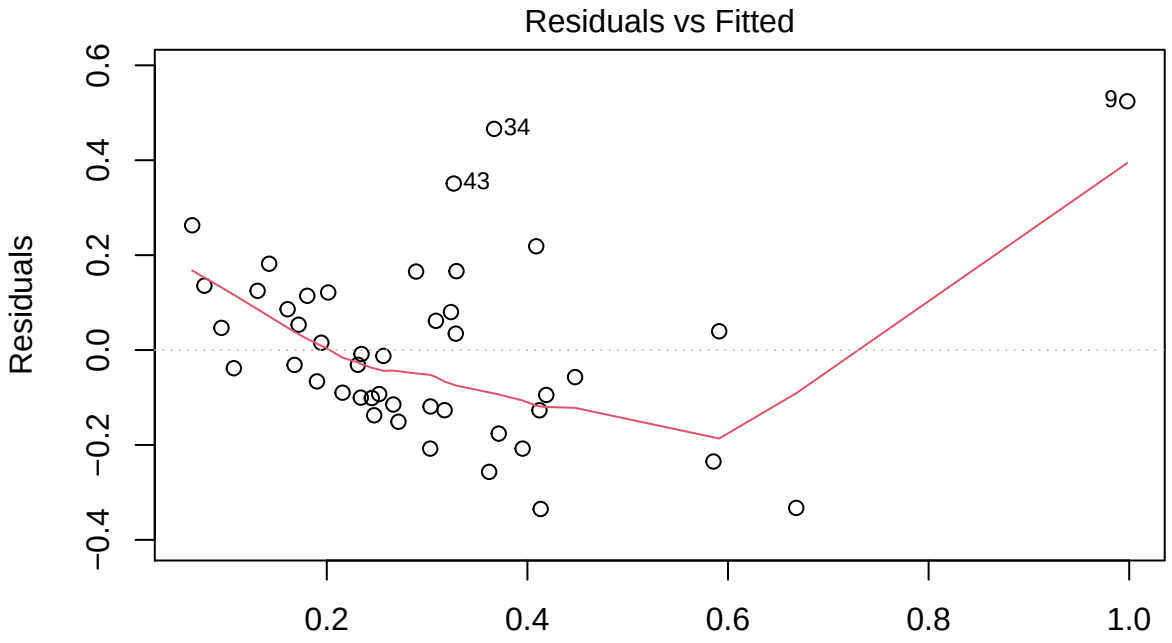
```
qqnorm(resid(fit_after_step))
qqline(resid(fit_after_step))
```

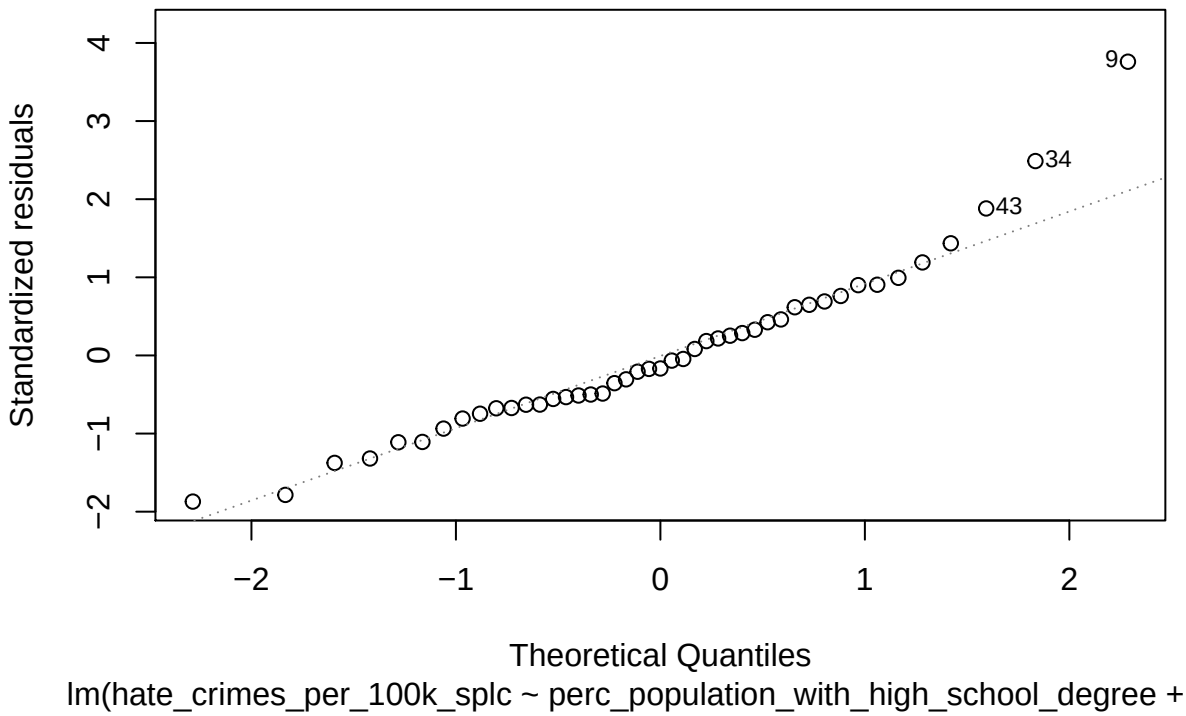


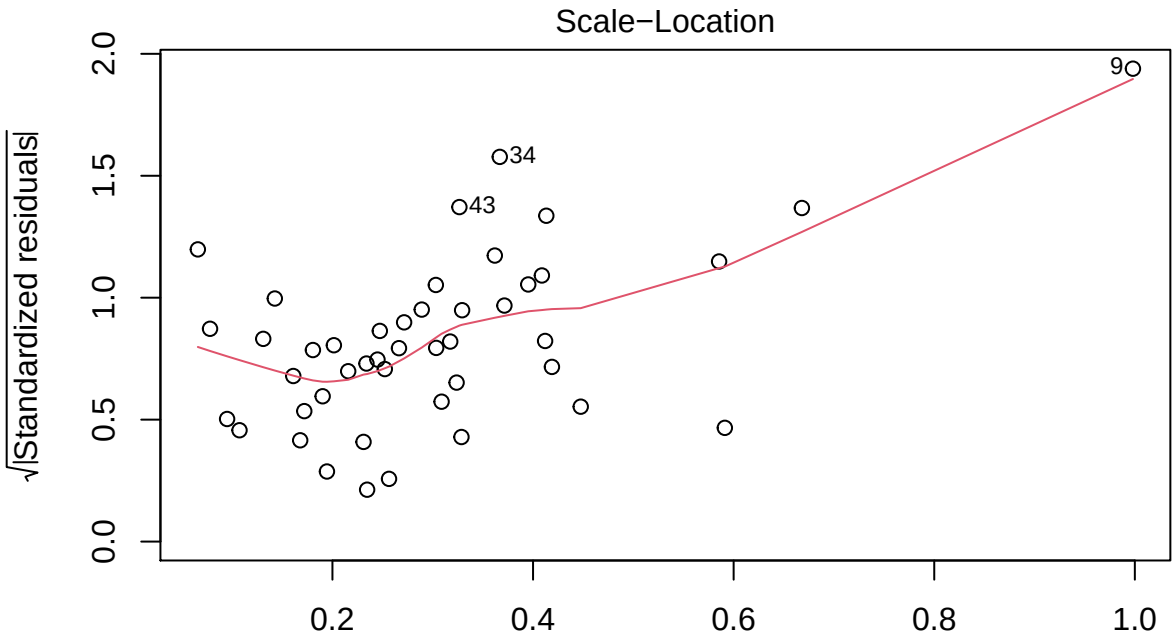

```
vif(fit_after_step)
```

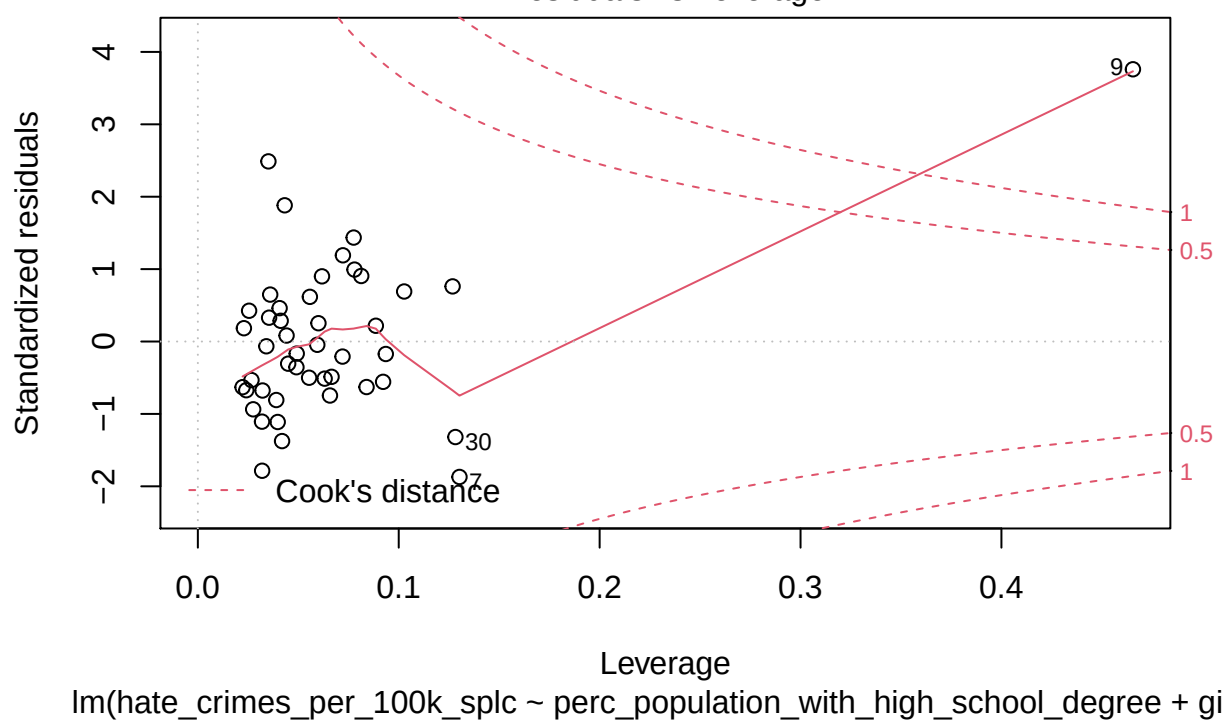
```
## perc_population_with_high_school_degree                           gini_index
##                                 1.40556                              1.40556
```

```
plot(fit_after_step)
```


Residuals vs Fitted
Residuals
Fitted values
lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..
9
34
43


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..
9
34
43





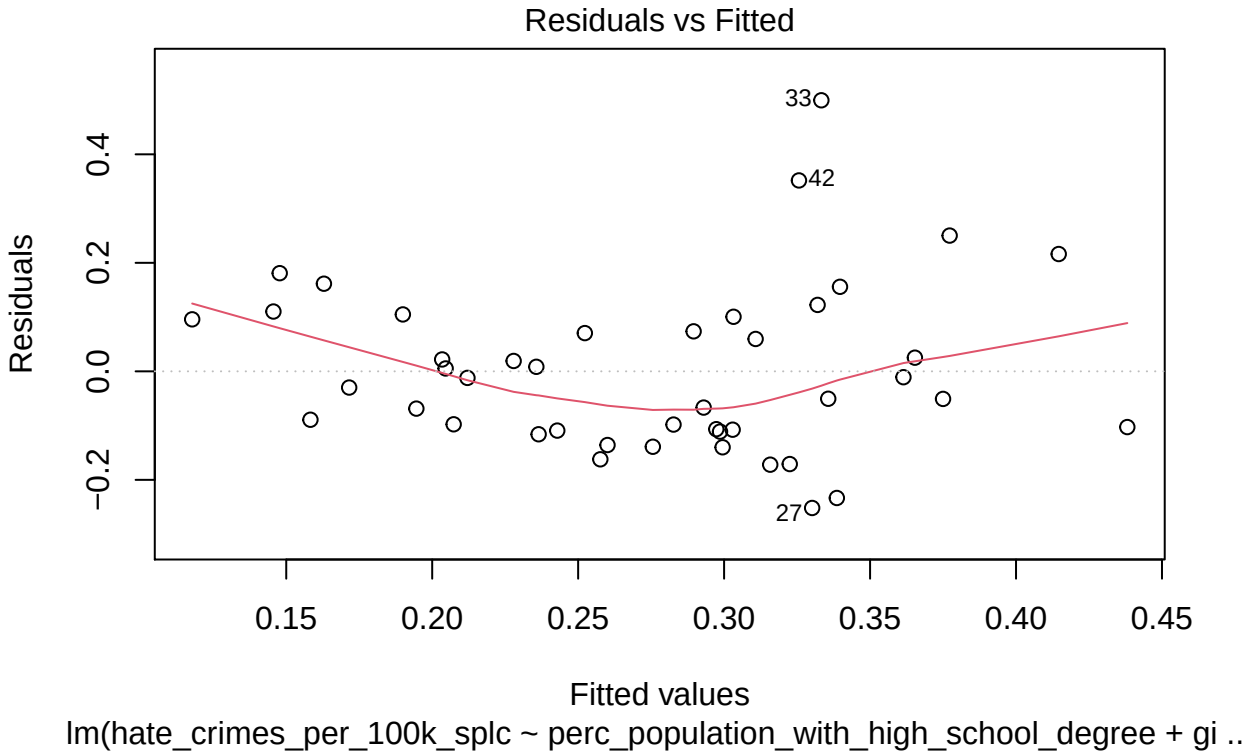
Exclude outliers

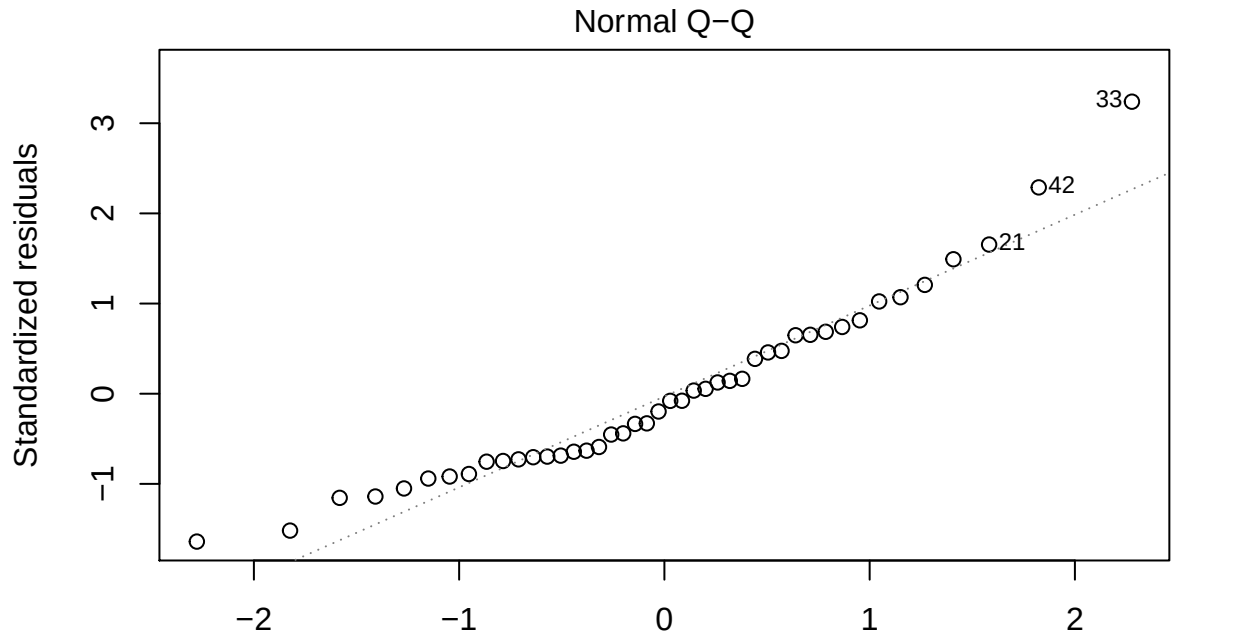
```
hc_df_no_outliers = hc_df[c(-9),]

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

```r
summary(fit_no_9)
```

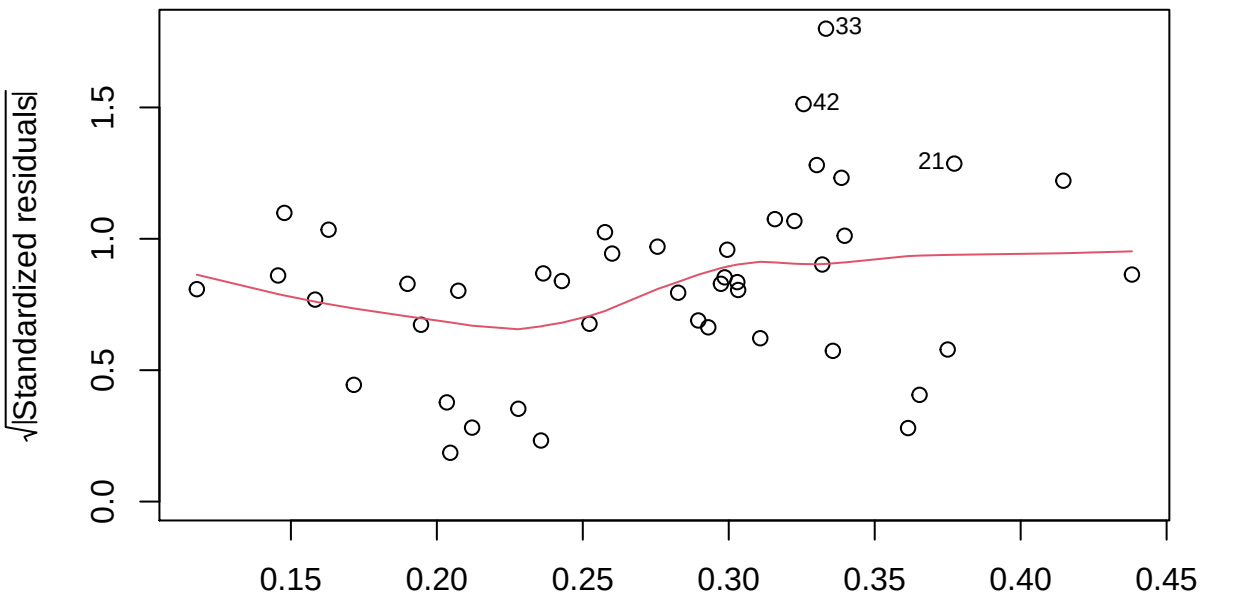
```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##     gini_index, data = hc_df_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25186 -0.10799 -0.02101  0.09700  0.49954
##
## Coefficients:
##                                       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
```

```r
plot(fit_no_9)
```

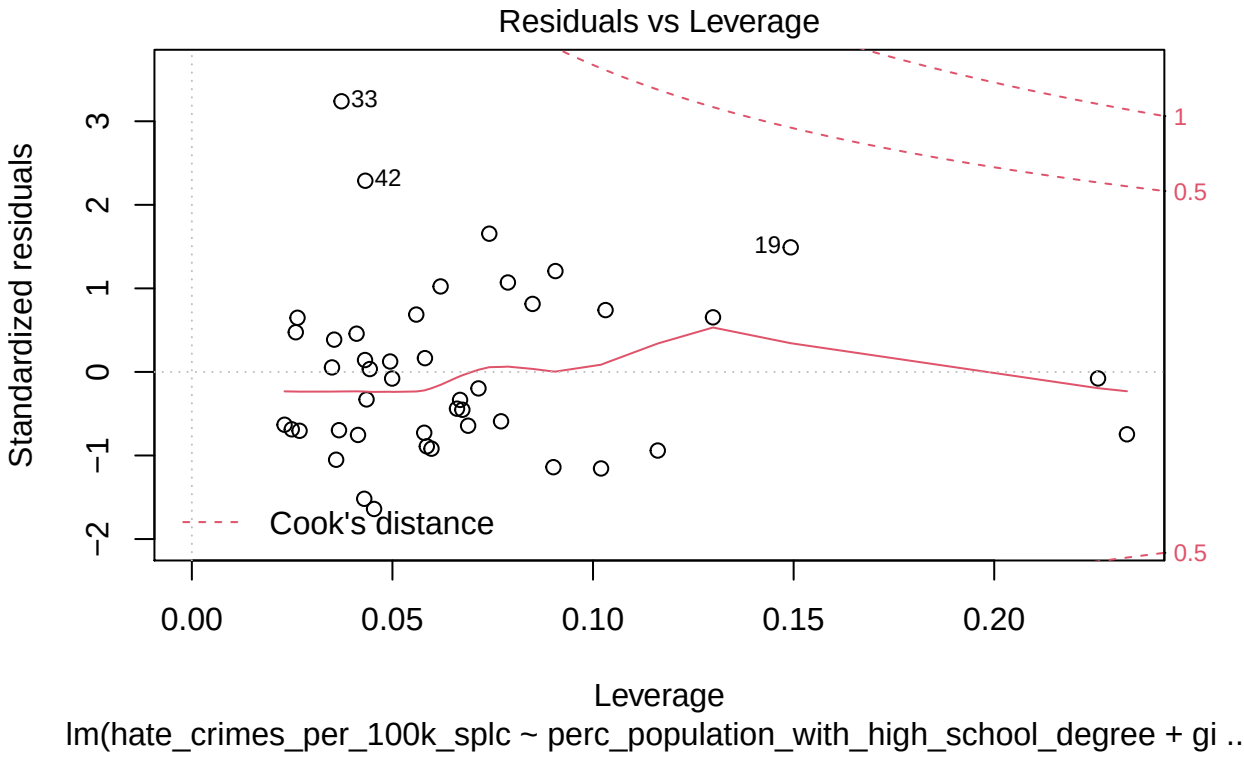




Theoretical Quantiles
lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..



Fitted values
lm(hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree + gi ..



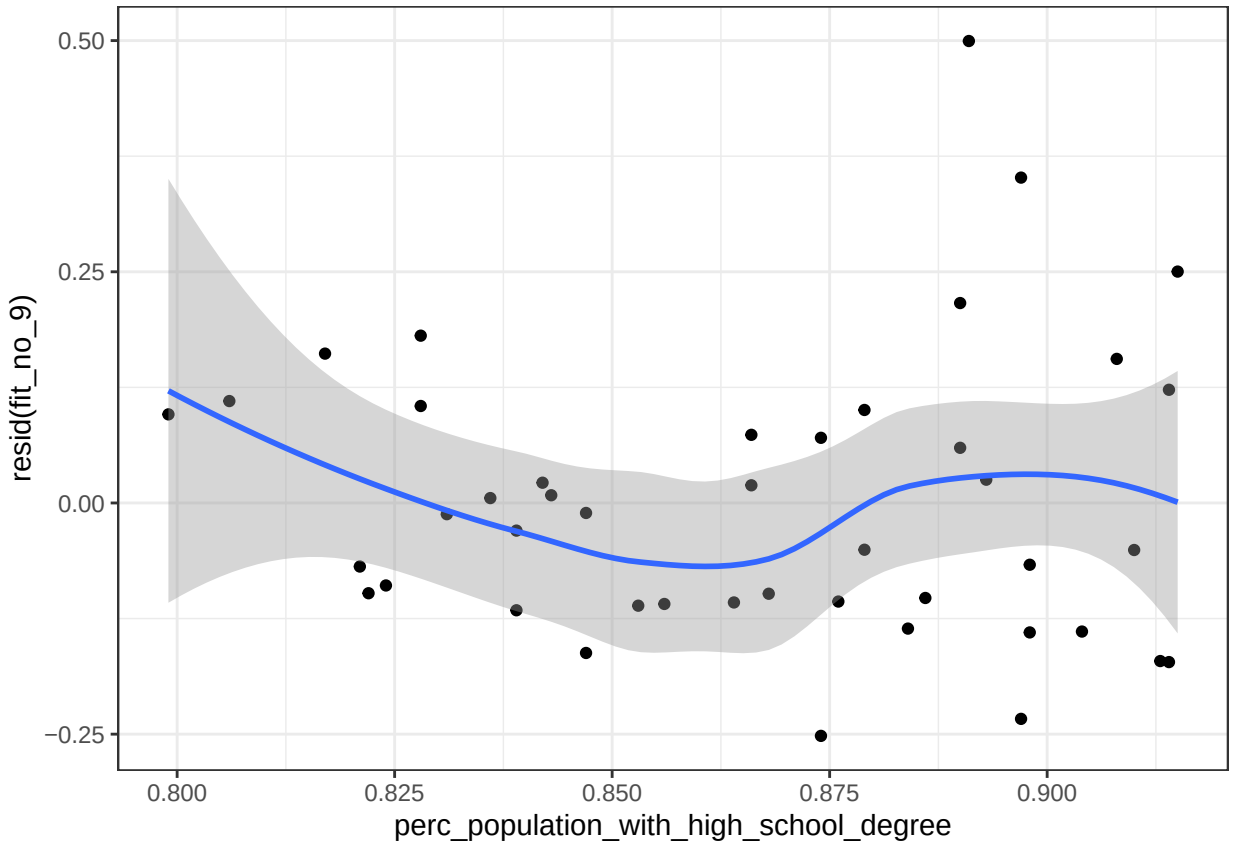

```
hc_df_only_9 = hc_df[c(9),]
hc_df_only_9
```

```
## # A tibble: 1 x 9
##   state unemployment urbanization median_househol~ perc_population~
##   <chr> <chr>        <chr>                  <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>
```

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.

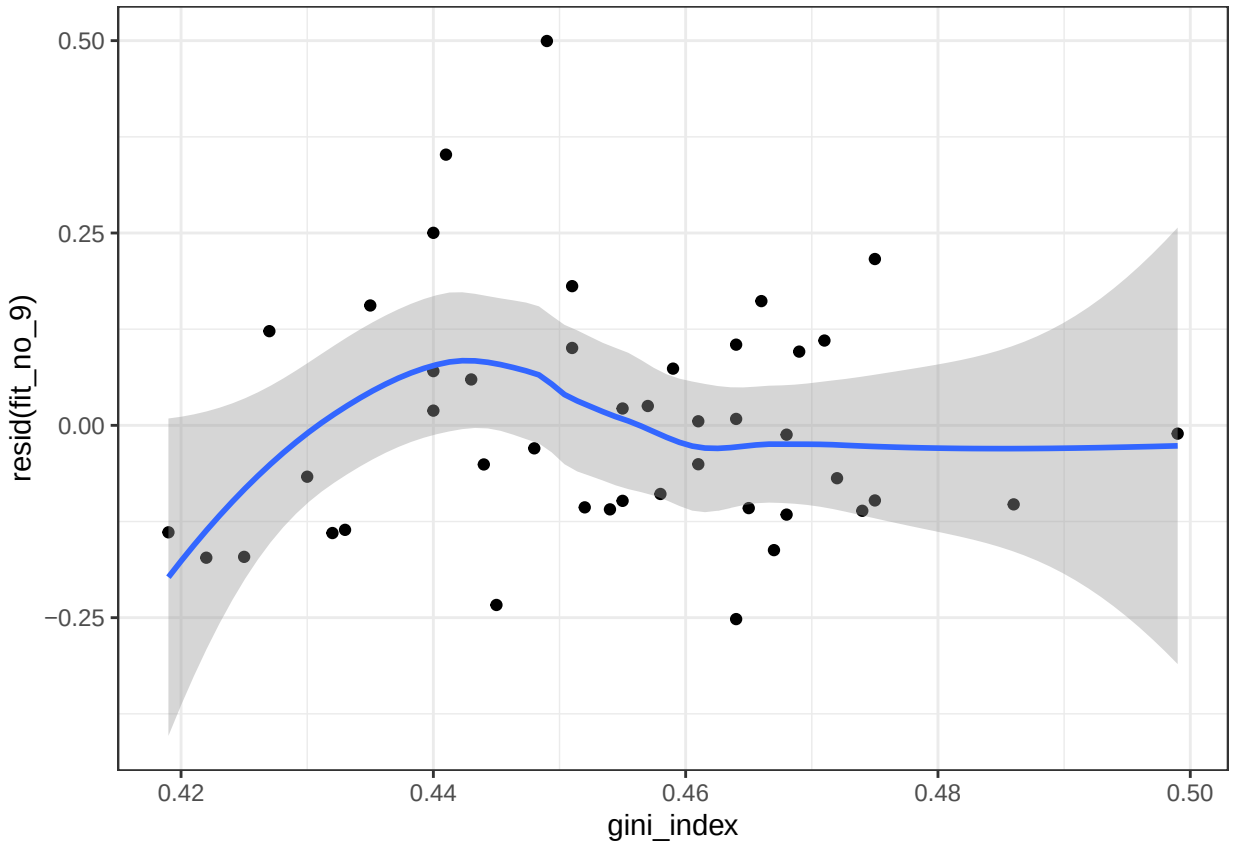
```
# Residuals vs. Covariates plots

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.