Data Exploration

Hanfei Qi, Anmol Singh, Wuraola Olawole, Ting Lian

12/4/2020

### Data Exploration:

Read data and make sure factors are factors.

crime\_df =   
 read\_csv("data/HateCrimes.csv") %>%   
 mutate(  
 unemployment = factor(unemployment),  
 urbanization = factor(urbanization),  
 hate\_crimes\_per\_100k\_splc = as.numeric(hate\_crimes\_per\_100k\_splc)  
 )

Descriptive statistics

my\_controls = tableby.control(  
 total = F,  
 test = F,   
 numeric.stats = c("meansd", "medianq1q3", "range", "Nmiss2"),  
 cat.stats = c("countpct", "Nmiss2"),  
 stats.labels = list(  
 meansd = "Mean (SD)",  
 medianq1q3 = "Median (Q1, Q3)",  
 range = "Min - Max",  
 Nmiss2 = "Missing",  
 countpct = "N (%)"))  
  
tab = tableby( ~ 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 = crime\_df,   
 control = my\_controls)  
  
summary(tab, title = "Descriptive Statistics: Hate Crime Data", text = T) %>%  
 knitr::kable()

|  |  |
| --- | --- |
|  | Overall (N=51) |
| hate\_crimes\_per\_100k\_splc |  |
| - Mean (SD) | 0.304 (0.253) |
| - Median (Q1, Q3) | 0.226 (0.143, 0.357) |
| - Min - Max | 0.067 - 1.522 |
| - Missing | 4 |
| unemployment |  |
| - high | 24 (47.1%) |
| - low | 27 (52.9%) |
| - Missing | 0 |
| urbanization |  |
| - high | 24 (47.1%) |
| - low | 27 (52.9%) |
| - Missing | 0 |
| median\_household\_income |  |
| - Mean (SD) | 55223.608 (9208.478) |
| - Median (Q1, Q3) | 54916.000 (48657.000, 60719.000) |
| - Min - Max | 35521.000 - 76165.000 |
| - Missing | 0 |
| perc\_population\_with\_high\_school\_degree |  |
| - Mean (SD) | 0.869 (0.034) |
| - Median (Q1, Q3) | 0.874 (0.841, 0.898) |
| - Min - Max | 0.799 - 0.918 |
| - Missing | 0 |
| perc\_non\_citizen |  |
| - 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 |  |
| - Mean (SD) | 0.454 (0.021) |
| - Median (Q1, Q3) | 0.454 (0.440, 0.467) |
| - Min - Max | 0.419 - 0.532 |
| - Missing | 0 |
| perc\_non\_white |  |
| - Mean (SD) | 0.316 (0.165) |
| - Median (Q1, Q3) | 0.280 (0.195, 0.420) |
| - Min - Max | 0.060 - 0.810 |
| - Missing | 0 |

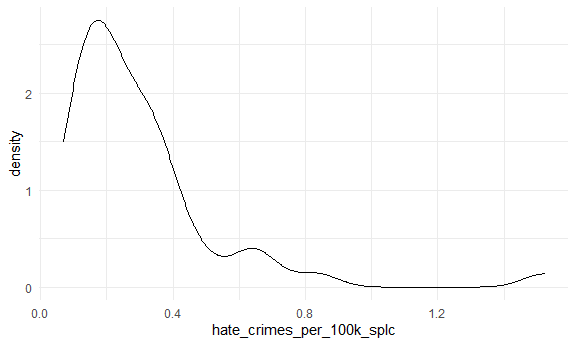
Comment: there are 4 NA’s in variable hate\_crimes\_per\_100k\_splc and 3 NA’s in variable perc\_non\_citizen.

Code to remove NA’s (not sure if we wanna do that).

crime\_df\_no\_na =   
 read\_csv("data/HateCrimes.csv", na = c("", "N/A")) %>%   
 mutate(  
 unemployment = factor(unemployment),  
 urbanization = factor(urbanization),  
 hate\_crimes\_per\_100k\_splc = as.numeric(hate\_crimes\_per\_100k\_splc)  
 ) %>%   
 na.omit()

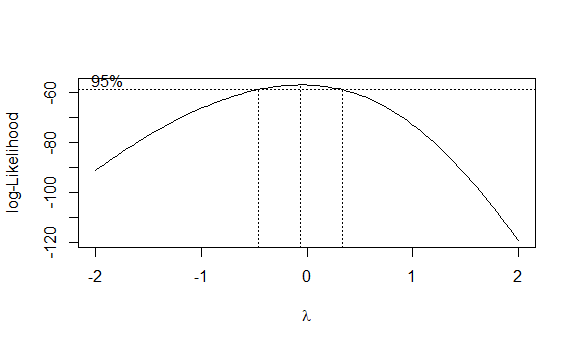
Distribution of the outcome

crime\_df\_no\_na %>%   
 ggplot(aes(x = hate\_crimes\_per\_100k\_splc)) + geom\_density()



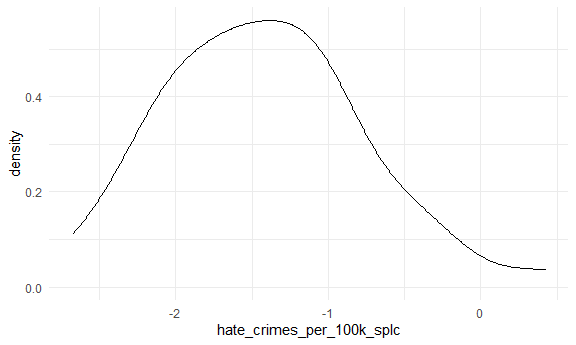
Comment: we need transformation

mod =   
 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 = crime\_df\_no\_na)  
  
boxcox(mod)



Perform the tranformation.

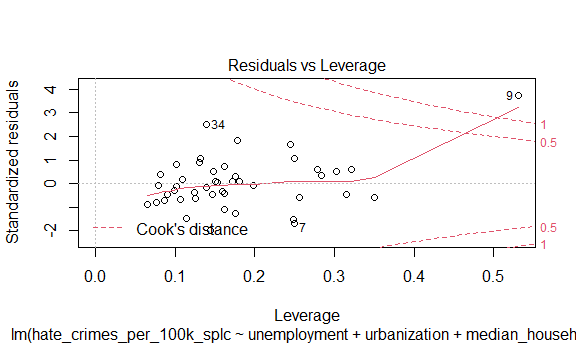
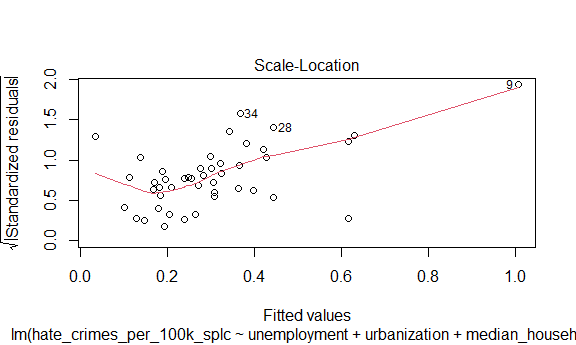
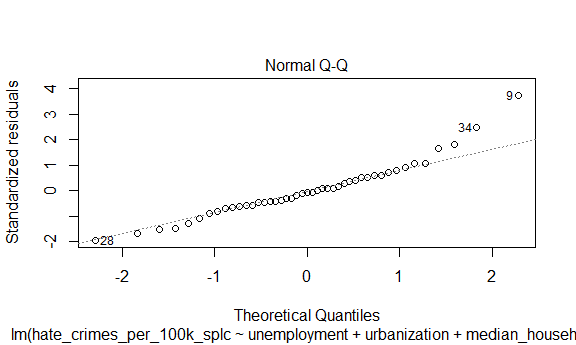
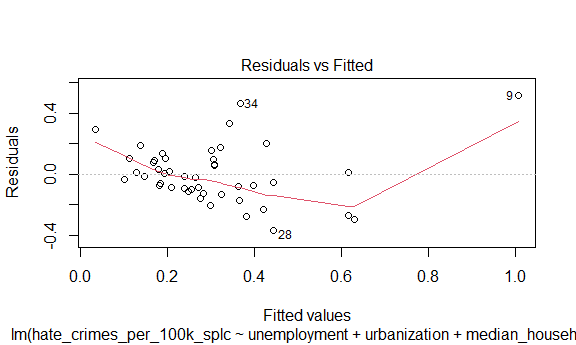
trans\_df =   
 crime\_df\_no\_na %>%   
 mutate(  
 hate\_crimes\_per\_100k\_splc = log(hate\_crimes\_per\_100k\_splc)  
 )  
  
trans\_df %>%   
 ggplot(aes(x = hate\_crimes\_per\_100k\_splc)) + geom\_density()



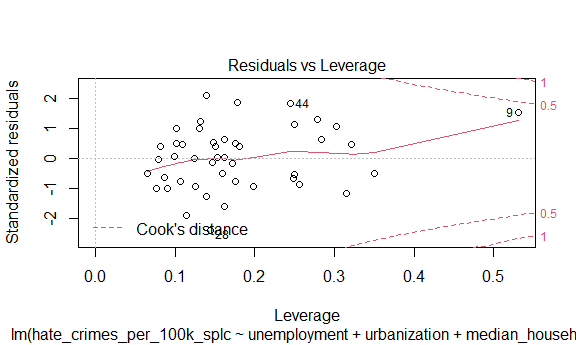
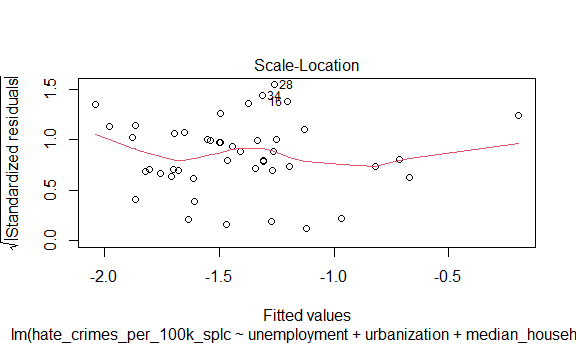
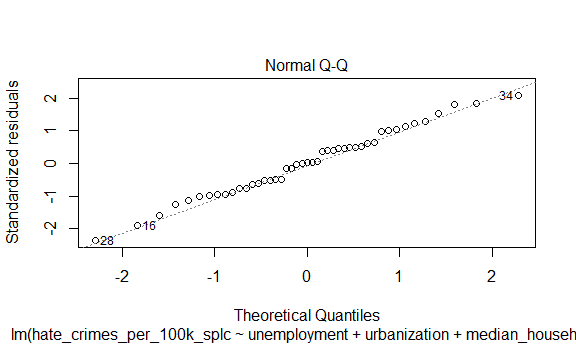
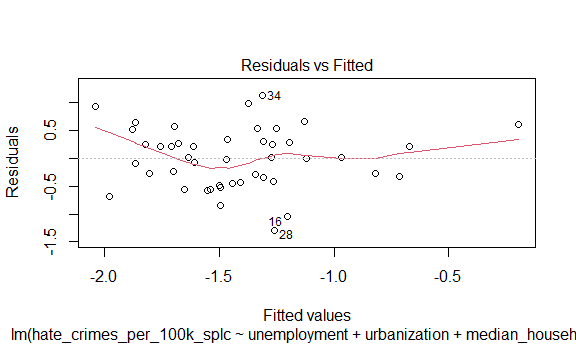
Looks good!

Plots before and after transformation.

plot(mod)



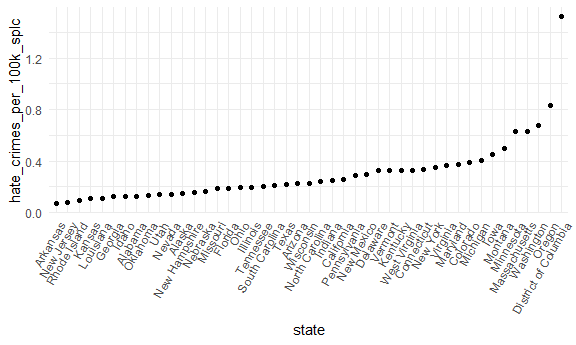
mod\_trans = lm(hate\_crimes\_per\_100k\_splc ~ unemployment + urbanization + median\_household\_income + perc\_population\_with\_high\_school\_degree + perc\_non\_citizen + gini\_index + perc\_non\_white,   
 data = trans\_df)  
  
plot(mod\_trans)



Plot crime rate by states.

crime\_df\_no\_na %>%   
 mutate(state = fct\_reorder(state, hate\_crimes\_per\_100k\_splc)) %>%   
 ggplot(aes(x = state, y = hate\_crimes\_per\_100k\_splc)) +   
 geom\_point() +   
 geom\_line() +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))

## geom\_path: Each group consists of only one observation. Do you need to adjust  
## the group aesthetic?



Comment: A wired point in District of Columbia. The rate of Oregon is slightly higher than other states. I will also consider Minnesota, Massachusetts, Washington as potential outliers.

### Modeling:

**Test if association between income inequality and hate crimes holds true:**

income\_hate\_model\_full\_data = crime\_df%>%  
 lm(hate\_crimes\_per\_100k\_splc~gini\_index,data=.)  
  
income\_hate\_model\_full\_data%>%  
 broom::tidy()%>%  
 knitr::kable(caption = "Testing Association between Income Inequality and Hate Crime using all the data")

Testing Association between Income Inequality and Hate Crime using all the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -1.527463 | 0.7833043 | -1.950025 | 0.0574197 |
| gini\_index | 4.020510 | 1.7177215 | 2.340606 | 0.0237445 |

income\_hate\_model\_trans = crime\_df%>%  
 mutate(  
 hate\_crimes\_per\_100k\_splc = log(hate\_crimes\_per\_100k\_splc)  
 )%>%  
 lm(hate\_crimes\_per\_100k\_splc~gini\_index,data=.)  
  
income\_hate\_model\_trans%>%  
 broom::tidy()%>%  
 knitr::kable(caption = "Testing Association between Income Inequality and Hate Crime using log transformed data")

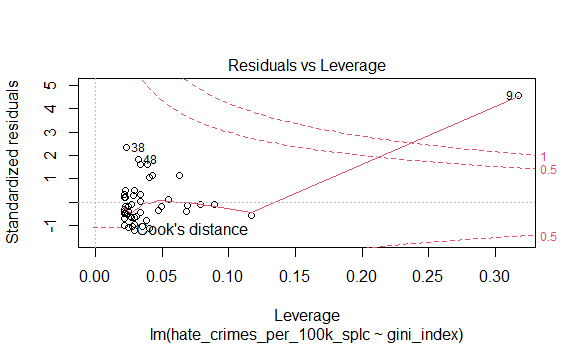
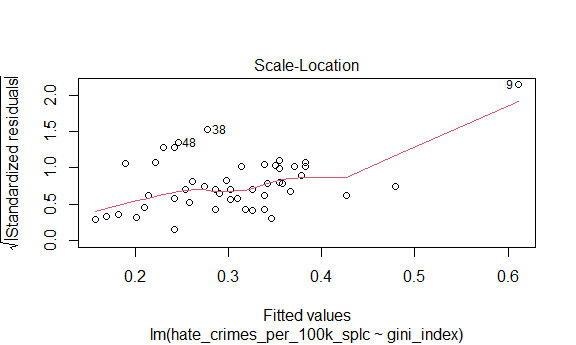
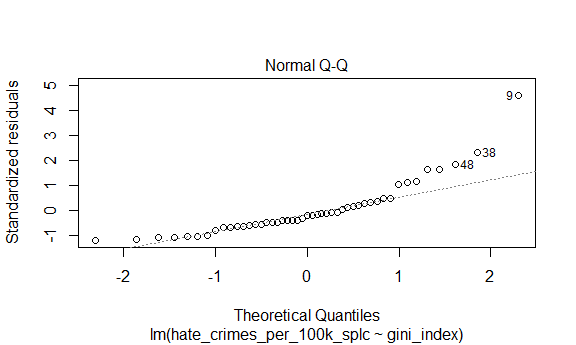
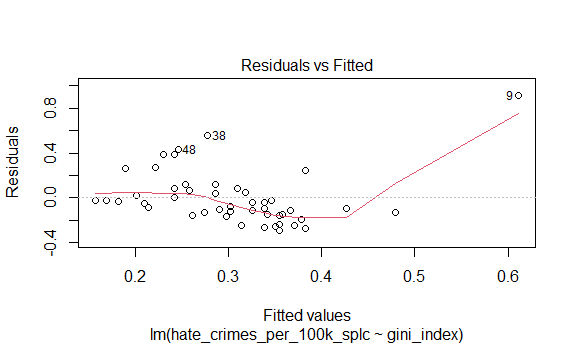
Testing Association between Income Inequality and Hate Crime using log transformed data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -3.675547 | 2.195289 | -1.674288 | 0.1010115 |
| gini\_index | 4.931538 | 4.814087 | 1.024398 | 0.3111231 |

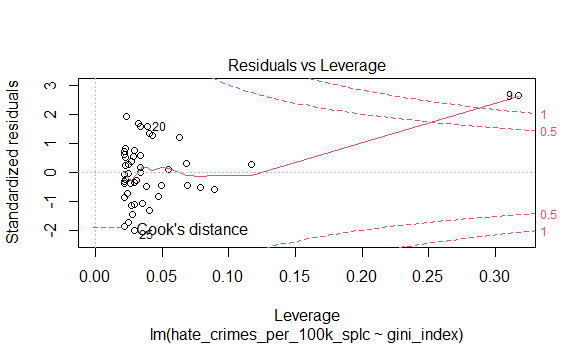
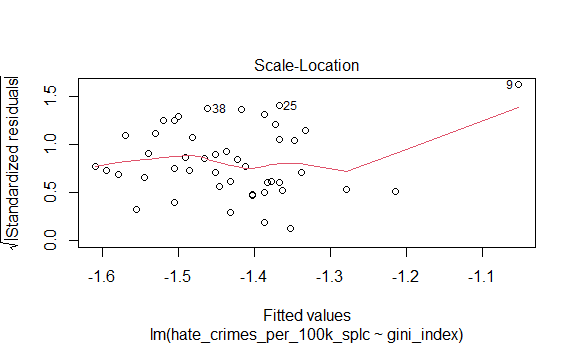
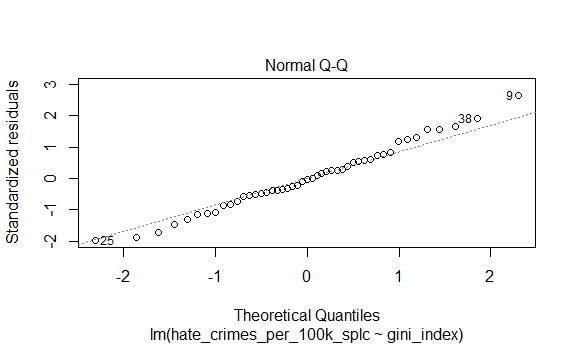
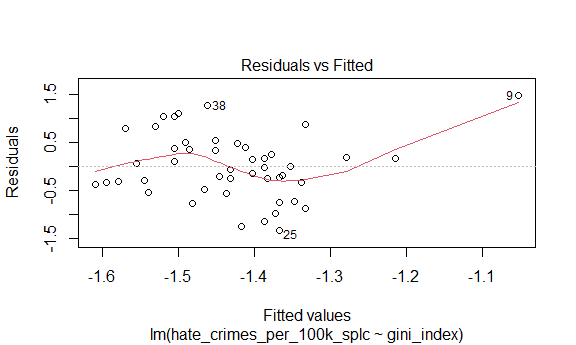
p-value is only significant for data which is not transformed or when outliers are not removed. Lets check model diagnostics to confirm though.

**Model Diagnostics:**

#For original data:  
  
plot(income\_hate\_model\_full\_data)



#For log-transformed data  
  
plot(income\_hate\_model\_trans)



Looking at leverage plots for both models there are definite outliers, will remove and check models again.

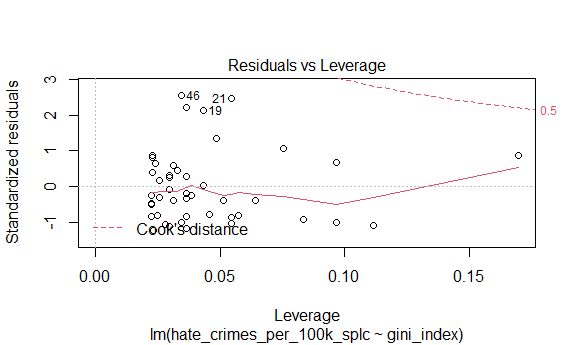
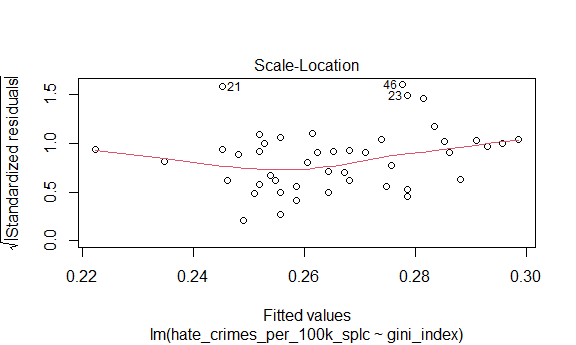
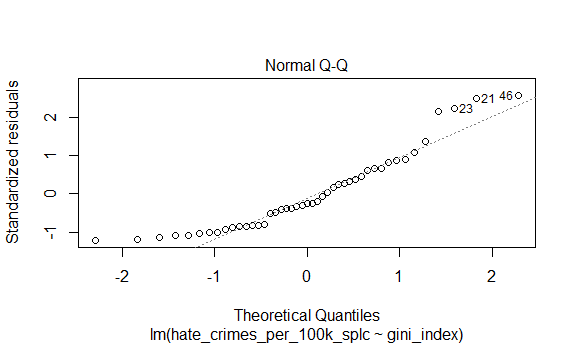
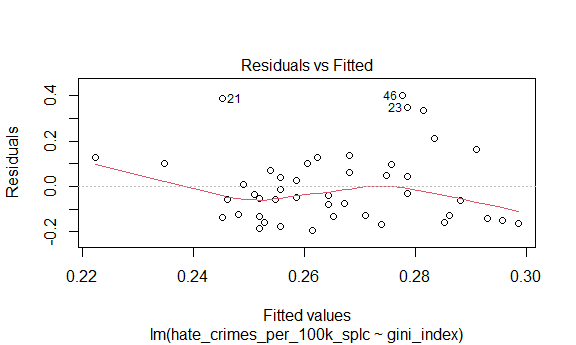
**Confirming Outliers:**

lower = quantile(crime\_df$hate\_crimes\_per\_100k\_splc,0.25,na.rm = T)-(1.5\*IQR(crime\_df$hate\_crimes\_per\_100k\_splc,na.rm=T)) #determining lower bound for outlier  
upper = quantile(crime\_df$hate\_crimes\_per\_100k\_splc,0.75, na.rm=T)+(1.5\*IQR(crime\_df$hate\_crimes\_per\_100k\_splc,na.rm=T)) #determining upper bound for outlier  
outliers = crime\_df$state[(crime\_df$hate\_crimes\_per\_100k\_splc>upper |crime\_df$hate\_crimes\_per\_100k\_splc<lower)] #finding states that are outliers

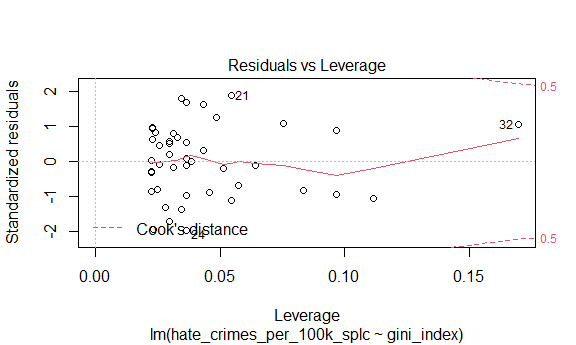
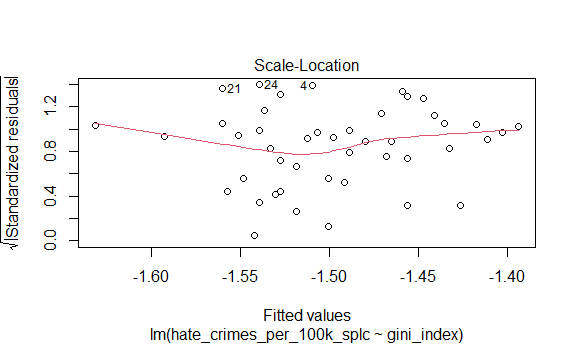
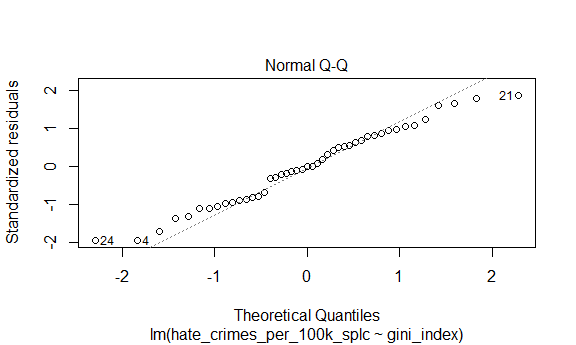
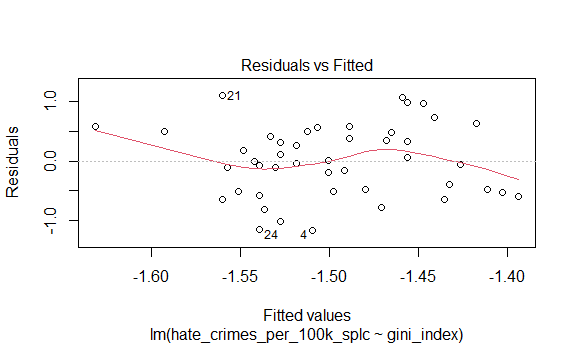
**Removing Outliers:**

crime\_df\_no\_outlier = crime\_df%>%  
 filter(!state %in% outliers)

#For original data without outliers:  
  
income\_hate\_model\_no\_outlier = lm(hate\_crimes\_per\_100k\_splc~gini\_index,data=crime\_df\_no\_outlier)  
  
plot(income\_hate\_model\_no\_outlier)



#For log-transformed data with no outliers  
  
income\_hate\_model\_trans\_no\_outlier = crime\_df\_no\_outlier%>%  
 mutate(  
 hate\_crimes\_per\_100k\_splc = log(hate\_crimes\_per\_100k\_splc)  
 )%>%  
 lm(hate\_crimes\_per\_100k\_splc~gini\_index,data=.)  
  
plot(income\_hate\_model\_trans\_no\_outlier)



Plots are closer to model assumptions without outliers present.

**Correlation Matrix of All Variables:**

corr\_matrix = trans\_df%>%  
 mutate(unemployment = ifelse(unemployment=="high",1,0),  
 urbanization = ifelse(urbanization=="high",1,0))%>%  
 dplyr::select(-state)%>%  
 cor()%>%  
 data.frame()  
  
corr\_matrix%>%  
 knitr::kable(caption = "Correlation Matrix for all variables except for State in Data Set", format = "html")

Correlation Matrix for all variables except for State in Data Set

unemployment

urbanization

median\_household\_income

perc\_population\_with\_high\_school\_degree

perc\_non\_citizen

gini\_index

perc\_non\_white

hate\_crimes\_per\_100k\_splc

unemployment

1.0000000

0.2435648

-0.2538367

-0.4526394

0.2408705

0.4089989

0.4281971

-0.1680668

urbanization

0.2435648

1.0000000

0.2613532

-0.1591792

0.6807743

0.4298763

0.5085536

0.1837247

median\_household\_income

-0.2538367

0.2613532

1.0000000

0.6511383

0.3017394

-0.1295216

0.0390540

0.3109375

perc\_population\_with\_high\_school\_degree

-0.4526394

-0.1591792

0.6511383

1.0000000

-0.2621288

-0.5371591

-0.4958932

0.2960528

perc\_non\_citizen

0.2408705

0.6807743

0.3017394

-0.2621288

1.0000000

0.4798976

0.7526102

0.1369321

gini\_index

0.4089989

0.4298763

-0.1295216

-0.5371591

0.4798976

1.0000000

0.5484035

0.2185514

perc\_non\_white

0.4281971

0.5085536

0.0390540

-0.4958932

0.7526102

0.5484035

1.0000000

-0.0066948

hate\_crimes\_per\_100k\_splc

-0.1680668

0.1837247

0.3109375

0.2960528

0.1369321

0.2185514

-0.0066948

1.0000000

highly\_correlated = data.frame(Correlation = corr\_matrix[corr\_matrix>=abs(0.6)&corr\_matrix<abs(1)])  
  
highly\_correlated = highly\_correlated[!duplicated(highly\_correlated),]%>%  
 data.frame()%>%  
 rename(Correlation = ".")%>%  
 mutate(Variable\_1 = c("urbanization","median\_household\_income","perc\_non\_citizen"),Variable\_2 = c("perc\_non\_citizen","perc\_population\_with\_high\_school\_degree","perc\_non\_white"))  
  
highly\_correlated%>%  
 knitr::kable(caption = "Variables that are Highly Correlated (Correlation >= absolute value (0.6))", format = "html")

Variables that are Highly Correlated (Correlation >= absolute value (0.6))

Correlation

Variable\_1

Variable\_2

0.6807743

urbanization

perc\_non\_citizen

0.6511383

median\_household\_income

perc\_population\_with\_high\_school\_degree

0.7526102

perc\_non\_citizen

perc\_non\_white

Studies have shown that high income is correlated with increased education and thus it would make sense that the median income and percentage of high school diploma holders are highly correlated [[1]](https://budgetmodel.wharton.upenn.edu/issues/2016/2/22/education-and-income-growth). Furthermore, a study conducted by the pew research center found that only 17.7% of immigrants are white non-hispanic which makes sense why the percentage of non citizens and the percentage of white people are very highly correlated as well [[2]](https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants-current-data/). Since these sets of variables are so highly correlated it would only be beneficial to adjust for one from each set in our model due to multicollinearity.

**Stepwise Regression procedure for Removing highly-correlated predictors:**

library(leaps)

## Warning: package 'leaps' was built under R version 4.0.3

#start with model using all predictors:   
crime\_trans = crime\_df\_no\_outlier%>%  
 mutate(  
 hate\_crimes\_per\_100k\_splc = log(hate\_crimes\_per\_100k\_splc)  
 )%>%  
 drop\_na()  
  
mod\_trans = 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=crime\_trans)  
  
summary(mod\_trans)

##   
## 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 = crime\_trans)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0548 -0.3860 0.0577 0.2990 1.1378   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.194e+01 6.423e+00 -1.859 0.0714  
## unemploymentlow 3.282e-01 2.005e-01 1.637 0.1106  
## urbanizationlow -1.377e-01 2.396e-01 -0.575 0.5692  
## median\_household\_income -3.352e-06 1.628e-05 -0.206 0.8381  
## perc\_population\_with\_high\_school\_degree 6.892e+00 5.335e+00 1.292 0.2049  
## perc\_non\_citizen -3.026e-02 5.153e+00 -0.006 0.9953  
## gini\_index 1.015e+01 7.079e+00 1.434 0.1604  
## perc\_non\_white -1.428e-01 1.038e+00 -0.138 0.8913  
##   
## (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: 0.5468 on 35 degrees of freedom  
## Multiple R-squared: 0.2192, Adjusted R-squared: 0.0631   
## F-statistic: 1.404 on 7 and 35 DF, p-value: 0.2348

#start stepwise regression procedure :  
  
mod\_tidy = mod\_trans%>%  
 broom::tidy()  
  
step(mod\_trans, direction='backward')

## Start: AIC=-44.77  
## 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\_citizen 1 0.00001 10.463 -46.774  
## - perc\_non\_white 1 0.00566 10.469 -46.750  
## - median\_household\_income 1 0.01267 10.476 -46.722  
## - urbanization 1 0.09870 10.562 -46.370  
## <none> 10.463 -44.774  
## - perc\_population\_with\_high\_school\_degree 1 0.49883 10.962 -44.771  
## - gini\_index 1 0.61487 11.078 -44.318  
## - unemployment 1 0.80098 11.264 -43.602  
##   
## Step: AIC=-46.77  
## hate\_crimes\_per\_100k\_splc ~ unemployment + urbanization + median\_household\_income +   
## perc\_population\_with\_high\_school\_degree + gini\_index + perc\_non\_white  
##   
## Df Sum of Sq RSS AIC  
## - perc\_non\_white 1 0.00851 10.472 -48.739  
## - median\_household\_income 1 0.01440 10.477 -48.715  
## - urbanization 1 0.11541 10.579 -48.302  
## <none> 10.463 -46.774  
## - perc\_population\_with\_high\_school\_degree 1 0.50889 10.972 -46.732  
## - gini\_index 1 0.61961 11.083 -46.300  
## - unemployment 1 0.81655 11.280 -45.542  
##   
## Step: AIC=-48.74  
## 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 0.02725 10.499 -50.627  
## - urbanization 1 0.10708 10.579 -50.301  
## <none> 10.472 -48.739  
## - gini\_index 1 0.63081 11.102 -48.223  
## - perc\_population\_with\_high\_school\_degree 1 0.76738 11.239 -47.698  
## - unemployment 1 0.89292 11.365 -47.220  
##   
## Step: AIC=-50.63  
## hate\_crimes\_per\_100k\_splc ~ unemployment + urbanization + perc\_population\_with\_high\_school\_degree +   
## gini\_index  
##   
## Df Sum of Sq RSS AIC  
## - urbanization 1 0.08043 10.579 -52.299  
## <none> 10.499 -50.627  
## - gini\_index 1 0.62337 11.122 -50.147  
## - unemployment 1 0.87727 11.376 -49.176  
## - perc\_population\_with\_high\_school\_degree 1 0.95942 11.458 -48.867  
##   
## Step: AIC=-52.3  
## hate\_crimes\_per\_100k\_splc ~ unemployment + perc\_population\_with\_high\_school\_degree +   
## gini\_index  
##   
## Df Sum of Sq RSS AIC  
## <none> 10.579 -52.299  
## - unemployment 1 0.83599 11.415 -51.028  
## - gini\_index 1 1.00103 11.580 -50.411  
## - perc\_population\_with\_high\_school\_degree 1 1.09580 11.675 -50.061

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ unemployment + perc\_population\_with\_high\_school\_degree +   
## gini\_index, data = crime\_trans)  
##   
## Coefficients:  
## (Intercept)   
## -12.8782   
## unemploymentlow   
## 0.3223   
## perc\_population\_with\_high\_school\_degree   
## 6.8299   
## gini\_index   
## 11.6847

#Final Recommended Model:  
  
final\_rec = lm(formula = hate\_crimes\_per\_100k\_splc ~ unemployment + perc\_population\_with\_high\_school\_degree + gini\_index, data = crime\_trans)  
  
summary(final\_rec)

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ unemployment + perc\_population\_with\_high\_school\_degree +   
## gini\_index, data = crime\_trans)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0599 -0.4014 0.0557 0.3175 1.2095   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12.8782 5.0597 -2.545 0.0150 \*  
## unemploymentlow 0.3223 0.1836 1.756 0.0870 .  
## perc\_population\_with\_high\_school\_degree 6.8299 3.3981 2.010 0.0514 .  
## gini\_index 11.6847 6.0826 1.921 0.0621 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5208 on 39 degrees of freedom  
## Multiple R-squared: 0.2106, Adjusted R-squared: 0.1499   
## F-statistic: 3.468 on 3 and 39 DF, p-value: 0.02514

final\_rec\_df = final\_rec%>%broom::tidy()  
  
#Comparing R-squared of final model vs model containing all variables:  
  
#Final Recommendation from Stepwise Regression:  
  
final\_rec %>% broom::glance()

## # A tibble: 1 x 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.211 0.150 0.521 3.47 0.0251 3 -30.9 71.7 80.5  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

#Model with all variables:  
  
mod\_trans %>% broom::glance()

## # A tibble: 1 x 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.219 0.0631 0.547 1.40 0.235 7 -30.6 79.3 95.1  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

R squared is around the same for both models, however final model has 138% improvement in adjusted R squared compared to model that contains all variables.

# **Explore interactions of all variables**

crime =  
 crime\_trans %>%  
 dplyr::select(-state)

Check for 2-way interactions between all predictors

lm.fit = lm(hate\_crimes\_per\_100k\_splc ~ (.)^2, data = crime)  
lm\_fit\_df = broom::tidy(lm.fit)  
lm\_fit\_df

## # A tibble: 29 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -447. 3.32e+2 -1.35 0.199   
## 2 unemploymentlow 18.4 1.94e+1 0.947 0.360   
## 3 urbanizationlow 28.7 2.92e+1 0.983 0.342   
## 4 median\_household\_income -0.00394 1.77e-3 -2.23 0.0427  
## 5 perc\_population\_with\_high\_school\_de~ 752. 3.96e+2 1.90 0.0786  
## 6 perc\_non\_citizen 83.2 8.58e+2 0.0970 0.924   
## 7 gini\_index 862. 6.57e+2 1.31 0.211   
## 8 perc\_non\_white 59.3 2.03e+2 0.293 0.774   
## 9 unemploymentlow:urbanizationlow 0.798 1.05e+0 0.764 0.458   
## 10 unemploymentlow:median\_household\_in~ 0.000258 9.72e-5 2.65 0.0188  
## # ... with 19 more rows

# obtain significant interactions  
all\_int =  
 broom::tidy(lm.fit) %>%  
 slice(9:29) %>%  
 filter(p.value < 0.05)  
  
all\_int # these are the significant interactions present in our data

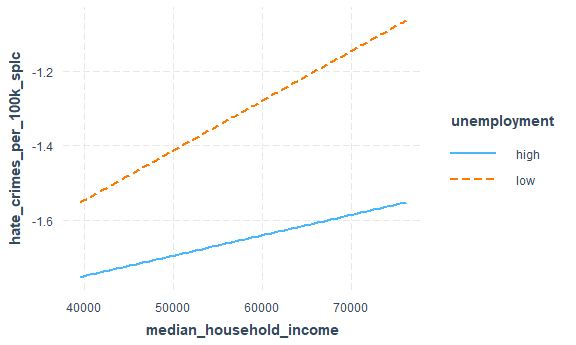
## # A tibble: 4 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 unemploymentlow:median\_household\_income 0.000258 9.72e-5 2.65 0.0188  
## 2 urbanizationlow:median\_household\_income 0.000270 1.21e-4 2.22 0.0431  
## 3 urbanizationlow:perc\_population\_with\_h~ -62.2 2.78e+1 -2.24 0.0421  
## 4 median\_household\_income:gini\_index 0.00769 3.02e-3 2.55 0.0232

From the analysis above, we see 4 significant interactions: unemployment and median household income, urbanization and median household income, urbanization and perc\_population with HS degree, and finally median\_household income and gini index.

# Unemployment and median household income  
reg\_med<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income \*unemployment, data = crime)  
 summary(reg\_med)

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ median\_household\_income \*   
## unemployment, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.94931 -0.37985 0.06987 0.35528 1.25644   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.970e+00 7.261e-01 -2.713 0.00987  
## median\_household\_income 5.490e-06 1.375e-05 0.399 0.69187  
## unemploymentlow -1.126e-01 1.103e+00 -0.102 0.91919  
## median\_household\_income:unemploymentlow 7.899e-06 1.981e-05 0.399 0.69234  
##   
## (Intercept) \*\*  
## median\_household\_income   
## unemploymentlow   
## median\_household\_income:unemploymentlow   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5454 on 39 degrees of freedom  
## Multiple R-squared: 0.1344, Adjusted R-squared: 0.06782   
## F-statistic: 2.018 on 3 and 39 DF, p-value: 0.1272

interact\_plot(reg\_med, pred = median\_household\_income, modx = unemployment ) ## interaction plot



# high unemployment  
 int\_1<-filter(crime, unemployment=="high")   
 reg\_1<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income, data=int\_1)  
 broom::tidy(reg\_1) # not significant

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -1.97 0.739 -2.66 0.0153  
## 2 median\_household\_income 0.00000549 0.0000140 0.392 0.699

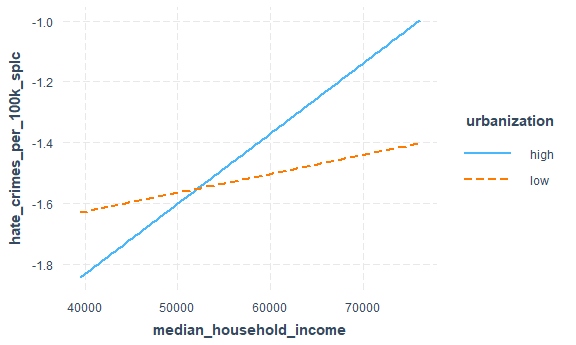
# low unemployment  
 int\_2<-filter(crime, unemployment=="low")   
 reg\_2<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income, data=int\_2)  
 broom::tidy(reg\_2) # not significant

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -2.08 0.815 -2.55 0.0189  
## 2 median\_household\_income 0.0000134 0.0000140 0.955 0.351

# Urbanization and median household income  
reg\_med\_2<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income\* urbanization, data = crime)  
summary(reg\_med\_2)

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ median\_household\_income \*   
## urbanization, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29800 -0.43155 0.02884 0.36844 1.00236   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.756e+00 8.296e-01 -3.322 0.00195  
## median\_household\_income 2.309e-05 1.442e-05 1.601 0.11740  
## urbanizationlow 8.797e-01 1.114e+00 0.790 0.43441  
## median\_household\_income:urbanizationlow -1.688e-05 2.001e-05 -0.843 0.40416  
##   
## (Intercept) \*\*  
## median\_household\_income   
## urbanizationlow   
## median\_household\_income:urbanizationlow   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5639 on 39 degrees of freedom  
## Multiple R-squared: 0.07467, Adjusted R-squared: 0.003492   
## F-statistic: 1.049 on 3 and 39 DF, p-value: 0.3817

interact\_plot(reg\_med\_2, pred = median\_household\_income, modx = urbanization )



# high urbanization  
 int\_3<-filter(crime, urbanization=="high")   
 reg\_3<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income, data=int\_3)  
 broom::tidy(reg\_3) # not significant

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -2.76 0.831 -3.32 0.00344  
## 2 median\_household\_income 0.0000231 0.0000144 1.60 0.126

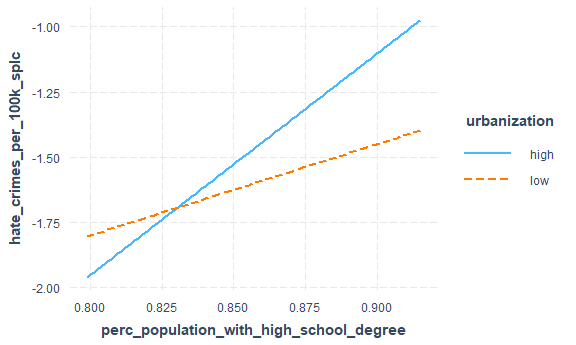
# low urbanization  
 int\_4<-filter(crime, urbanization=="low")   
 reg\_4<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income, data=int\_4)  
 broom::tidy(reg\_4) # not significant

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -1.88 0.742 -2.53 0.0205  
## 2 median\_household\_income 0.00000621 0.0000139 0.448 0.659

# Urbanization and perc\_population with HS degree  
reg\_per<-lm(hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree\* urbanization, data=crime)  
 summary(reg\_per)

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree \*   
## urbanization, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2244 -0.3935 0.1048 0.3976 0.9315   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -8.759 3.448  
## perc\_population\_with\_high\_school\_degree 8.509 4.007  
## urbanizationlow 4.155 4.534  
## perc\_population\_with\_high\_school\_degree:urbanizationlow -5.004 5.237  
## t value Pr(>|t|)   
## (Intercept) -2.540 0.0152 \*  
## perc\_population\_with\_high\_school\_degree 2.124 0.0401 \*  
## urbanizationlow 0.917 0.3650   
## perc\_population\_with\_high\_school\_degree:urbanizationlow -0.956 0.3452   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5457 on 39 degrees of freedom  
## Multiple R-squared: 0.1333, Adjusted R-squared: 0.06663   
## F-statistic: 1.999 on 3 and 39 DF, p-value: 0.13

interact\_plot(reg\_per, pred = perc\_population\_with\_high\_school\_degree, modx = urbanization )



# high urbanization  
reg\_5<-lm(hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree, data=int\_3)  
 summary(reg\_5) # significant!

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree,   
## data = int\_3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2244 -0.3066 0.1164 0.3769 0.7376   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.759 3.418 -2.563 0.0185 \*  
## perc\_population\_with\_high\_school\_degree 8.509 3.971 2.142 0.0446 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5409 on 20 degrees of freedom  
## Multiple R-squared: 0.1867, Adjusted R-squared: 0.146   
## F-statistic: 4.59 on 1 and 20 DF, p-value: 0.04464

# low urbanization  
reg\_6<-lm(hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree, data=int\_4)  
 summary(reg\_6) # not significant

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree,   
## data = int\_4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.95634 -0.40840 -0.02925 0.48099 0.93148   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -4.604 2.971 -1.55 0.138  
## perc\_population\_with\_high\_school\_degree 3.504 3.404 1.03 0.316  
##   
## Residual standard error: 0.5507 on 19 degrees of freedom  
## Multiple R-squared: 0.05284, Adjusted R-squared: 0.002994   
## F-statistic: 1.06 on 1 and 19 DF, p-value: 0.3161

# Median\_household income and gini index  
reg\_gin<-lm(hate\_crimes\_per\_100k\_splc ~ median\_household\_income\* gini\_index, data=crime)  
 summary(reg\_gin)

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ median\_household\_income \*   
## gini\_index, data = crime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.31961 -0.35251 0.05307 0.36962 1.03199   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.887e+01 1.587e+01 1.189 0.241  
## median\_household\_income -3.603e-04 2.662e-04 -1.353 0.184  
## gini\_index -4.657e+01 3.470e+01 -1.342 0.187  
## median\_household\_income:gini\_index 8.262e-04 5.838e-04 1.415 0.165  
##   
## Residual standard error: 0.5544 on 39 degrees of freedom  
## Multiple R-squared: 0.1056, Adjusted R-squared: 0.03681   
## F-statistic: 1.535 on 3 and 39 DF, p-value: 0.2207

interact\_plot(reg\_gin, pred = median\_household\_income, modx = gini\_index ) # not significant

 With the plots, we observe that there are actually just 3 interactions and there appears to be a significant interaction between urbanization and percent\_pop with high school degree.

Check 3\_way interactions between predictors of interest

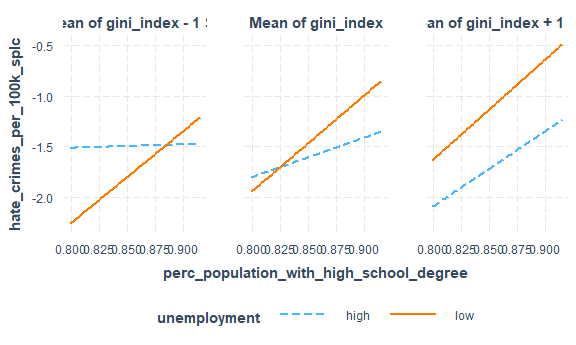
fit\_4 = lm(hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree \*unemployment\* gini\_index, data = crime\_trans)  
summary(fit\_4)

##   
## Call:  
## lm(formula = hate\_crimes\_per\_100k\_splc ~ perc\_population\_with\_high\_school\_degree \*   
## unemployment \* gini\_index, data = crime\_trans)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.05139 -0.36654 0.06147 0.32696 1.07430   
##   
## Coefficients:  
## Estimate  
## (Intercept) 74.19  
## perc\_population\_with\_high\_school\_degree -85.87  
## unemploymentlow -83.30  
## gini\_index -174.12  
## perc\_population\_with\_high\_school\_degree:unemploymentlow 84.89  
## perc\_population\_with\_high\_school\_degree:gini\_index 197.60  
## unemploymentlow:gini\_index 173.39  
## perc\_population\_with\_high\_school\_degree:unemploymentlow:gini\_index -174.76  
## Std. Error  
## (Intercept) 109.72  
## perc\_population\_with\_high\_school\_degree 124.18  
## unemploymentlow 148.62  
## gini\_index 238.43  
## perc\_population\_with\_high\_school\_degree:unemploymentlow 168.30  
## perc\_population\_with\_high\_school\_degree:gini\_index 270.19  
## unemploymentlow:gini\_index 323.43  
## perc\_population\_with\_high\_school\_degree:unemploymentlow:gini\_index 366.82  
## t value  
## (Intercept) 0.676  
## perc\_population\_with\_high\_school\_degree -0.691  
## unemploymentlow -0.560  
## gini\_index -0.730  
## perc\_population\_with\_high\_school\_degree:unemploymentlow 0.504  
## perc\_population\_with\_high\_school\_degree:gini\_index 0.731  
## unemploymentlow:gini\_index 0.536  
## perc\_population\_with\_high\_school\_degree:unemploymentlow:gini\_index -0.476  
## Pr(>|t|)  
## (Intercept) 0.503  
## perc\_population\_with\_high\_school\_degree 0.494  
## unemploymentlow 0.579  
## gini\_index 0.470  
## perc\_population\_with\_high\_school\_degree:unemploymentlow 0.617  
## perc\_population\_with\_high\_school\_degree:gini\_index 0.469  
## unemploymentlow:gini\_index 0.595  
## perc\_population\_with\_high\_school\_degree:unemploymentlow:gini\_index 0.637  
##   
## Residual standard error: 0.5287 on 35 degrees of freedom  
## Multiple R-squared: 0.2699, Adjusted R-squared: 0.1239   
## F-statistic: 1.848 on 7 and 35 DF, p-value: 0.1087

probe\_interaction(fit\_4, pred = perc\_population\_with\_high\_school\_degree, modx = unemployment, mod2 = gini\_index, alpha = .1)

## Warning: Johnson-Neyman intervals are not available for factor moderators.

## ¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦ While gini\_index (2nd moderator) = 0.44 (- 1 SD) ¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦   
##   
## SIMPLE SLOPES ANALYSIS   
##   
## Slope of perc\_population\_with\_high\_school\_degree when unemployment = high:   
##   
## Est. S.E. t val. p  
## ------ ------ -------- ------  
## 0.34 7.81 0.04 0.97  
##   
## Slope of perc\_population\_with\_high\_school\_degree when unemployment = low:   
##   
## Est. S.E. t val. p  
## ------ ------ -------- ------  
## 8.98 7.26 1.24 0.22  
##   
## ¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦ While gini\_index (2nd moderator) = 0.45 (Mean) ¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦   
##   
## SIMPLE SLOPES ANALYSIS   
##   
## Slope of perc\_population\_with\_high\_school\_degree when unemployment = high:   
##   
## Est. S.E. t val. p  
## ------ ------ -------- ------  
## 3.85 4.95 0.78 0.44  
##   
## Slope of perc\_population\_with\_high\_school\_degree when unemployment = low:   
##   
## Est. S.E. t val. p  
## ------ ------ -------- ------  
## 9.39 5.10 1.84 0.07  
##   
## ¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦ While gini\_index (2nd moderator) = 0.47 (+ 1 SD) ¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦   
##   
## SIMPLE SLOPES ANALYSIS   
##   
## Slope of perc\_population\_with\_high\_school\_degree when unemployment = high:   
##   
## Est. S.E. t val. p  
## ------ ------ -------- ------  
## 7.36 5.83 1.26 0.22  
##   
## Slope of perc\_population\_with\_high\_school\_degree when unemployment = low:   
##   
## Est. S.E. t val. p  
## ------ ------ -------- ------  
## 9.80 6.18 1.59 0.12



Conclusion: There are no significant interactions between our predictors of interest

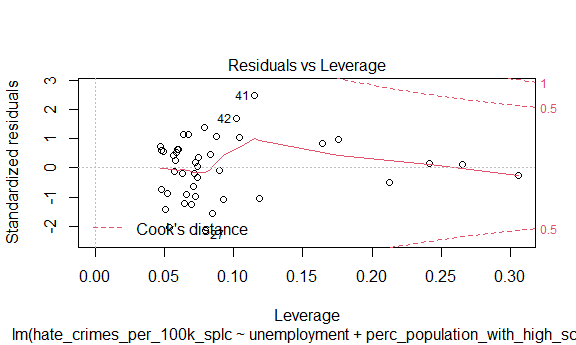
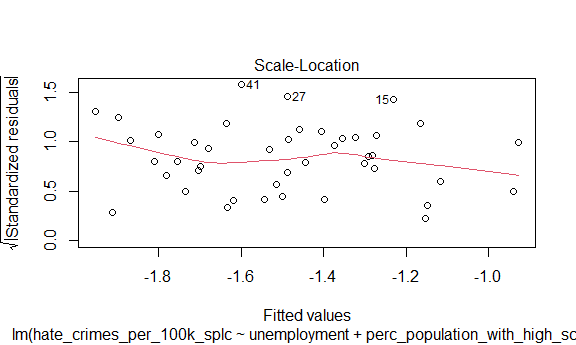
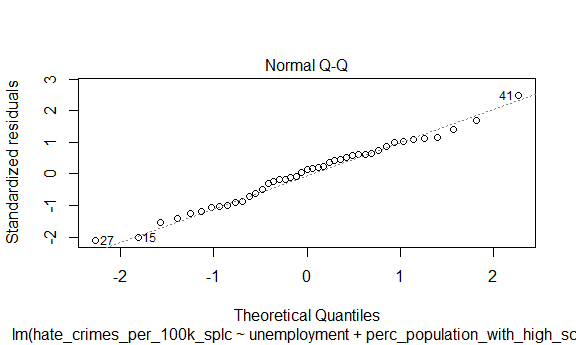
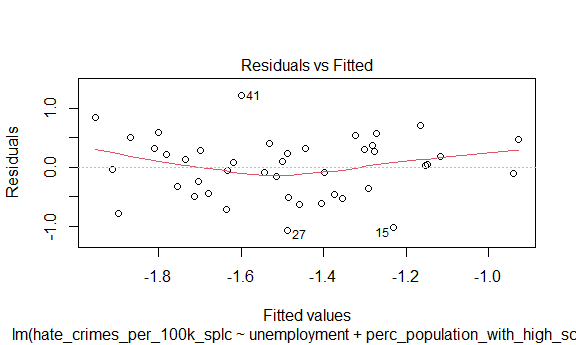
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| term | df | sumsq | meansq | statistic | p.value |
| perc\_population\_with\_high\_school\_degree | 1 | 1.2012907 | 1.2012907 | 4.0336007 | 0.0515635 |
| urbanization | 1 | 0.3131419 | 0.3131419 | 1.0514437 | 0.3114931 |
| perc\_population\_with\_high\_school\_degree:urbanization | 1 | 0.2719223 | 0.2719223 | 0.9130397 | 0.3451949 |
| Residuals | 39 | 11.6150159 | 0.2978209 | NA | NA |

Further test shows no significant interaction

### Model Diagnosis:

Use diagnostic plots to check model assumptions:

plot(final\_rec)

 1. Residuals vs. Fitted Plot: The points show a random patter and are evenly spread above and below the line of 0. The red line is approximately horizontal and is bouncing around the line of 0. This graph shows that this model fit the assumption of homoscedasiticity.

1. Normal Q-Q Plot: All points align in an approximately straight line with no significant departure. This graph indicates that the residuals are normal.
2. Scale-Location Plot: Similar with the residual vs. fitted plot, points in the graph are randomly and equally spread. The red line is approximately horizontal indicating that the variance of the model is equal.
3. Residuals vs. Leverage Plot: Although there is a bounce in the line, the overall line is approximately horizontal around the 0. Points are not randomly spread and are lump together between 0.05 and 0.10. There is no point beyond the Cook’s distance so we could assume that there is no significant influential point.

Overall, these four graphs show that this fitted model is good for represent the data.

Check for outlier:

stu\_res <- rstandard(final\_rec)  
outliers\_y <- stu\_res[abs(stu\_res)>2.5]  
outliers\_y

## named numeric(0)

Result: there is no outlier in Y.

Check for multicollinearity:

vif(final\_rec) %>%  
 knitr::kable()

|  |  |
| --- | --- |
|  | x |
| unemploymentlow | 1.335118 |
| perc\_population\_with\_high\_school\_degree | 1.970627 |
| gini\_index | 1.806199 |

Result: None of the parameter has variance inflation factor (VIF) value greater than 5 indicating that there is no multicollinearity in this model.