W203 Lab 3: Reducing Crime

Chi Iong Ansjory, Tsung-Chin Han, Marcelo Queiroz

7/31/2018

Introduction

The motivation of this analysis is to understand the determinants of crime and to generate policy suggestions in order to reduce crime. Imagine that we have been hired to provide research for a political campaign, our data source is primarily the dataset of crime statistics for a selection of counties in North Carolina.

The Initial EDA

Set up the working directory by putting data file and Rmd file in the same directory.

Load all necessary libraries for the R functions.

```
library(car)
library(lmtest)
library(sandwich)
library(stargazer)
```

Load the cross-section data set into R and inspect it.

```
Data <- read.csv("crime_v2.csv", header=TRUE, sep=",")
summary(Data)
```

```
##      county           year        crmrte             prbarr       
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277  
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568  
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095  
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492  
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438  
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091  
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6        
##         prbconv      prbpris           avgsen           polpc          
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746  
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231  
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485  
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702  
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877  
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054  
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6         
##     density            taxpc             west           central      
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000  
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736  
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000  
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000  
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6       
##      urban            pctmin80           wcon            wtuc      
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6  
```

```
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd            wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed            wsta            wloc            mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

The data set consists of 97 observations and 25 variables. From the summary, there are 6 of the observations with data consistently missing across variables. *prbconv* is a factor variable, and some of the variables that are supposed to be probabilities are actually greater than 1. In order to fix these problems, following cleansing of data are performed:

- Convert *prbconv* from factor to numeric.
- Eliminate 6 observations missing data based *county*.
- Eliminate 10 observations with probability values greater than 1 from *prbarr*, *prbconv*, *prbpris*.
- Eliminate 1 observation by reassigning the indices to country number.

```
Data$prbconv = as.numeric(paste(Data$prbconv))
subcases = !is.na(Data$county) & !Data$prbarr>1 & !Data$prbconv>1 & !Data$prbpris>1
crime_data = Data[subcases, ]
crime_data[duplicated(crime_data$county),]
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle
## 89 0.07819394
```

```r
crime_data <- crime_data[1:80,]
row.names(crime_data) <- crime_data$county
```

Now, the new data frame has 80 observations, which can be assessed to improve our policy suggestions for counties of North Carolina. The available descriptions of variables are:

variable	label
year	1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	avg. sentence, days
polpc	police per capita
density	people per sq. mile
taxpc	tax revenue per capita
west	=1 if in western N.C.
central	=1 if in central N.C.
urban	=1 if in SMSA
pctmin80	perc. minority, 1980
wcon	weekly wage, construction
wtuc	wkly wge, trns, util, commun
wtrd	wkly wge, whlesle, retail trade
wfir	wkly wge, fin, ins, real est
wser	wkly wge, service industry
wmfg	wkly wge, manufacturing
wfed	wkly wge, fed employees
wsta	wkly wge, state employees
wloc	wkly wge, local gov emps
mix	offense mix: face-to-face/other
pctymle	percent young male

As counties of North Carolina are interested in policy suggestions that could address the crime problem, the dependent variable will be *crmrte*, or crimes committed per person.

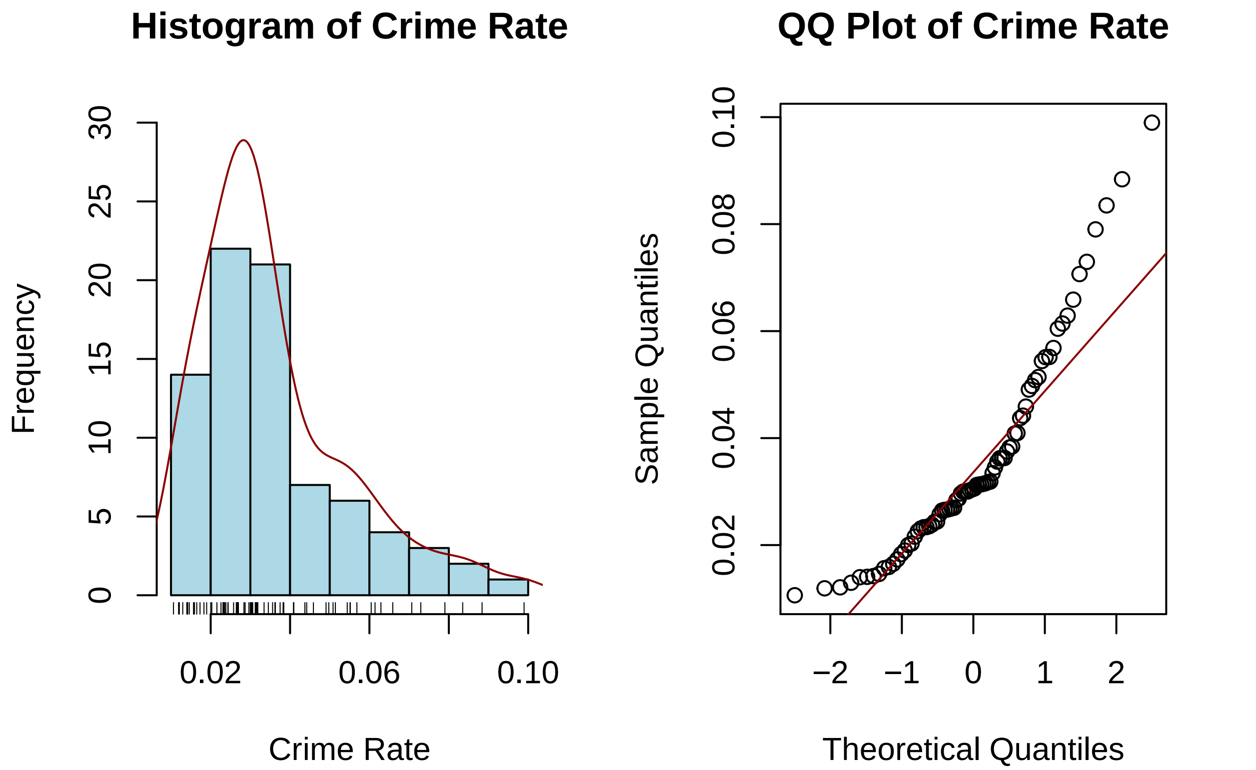
Additionally, as analyzing 25 variables would be inefficient, we decided to divide our analysis into 3 groups based on natures of variables. We will have a group of variables for models that explains how convictions and police enforcement relates to crime rates, another group for models that explains how econo-geographic data influences crime rates, and last group for models that covers variations in wages and industry differences.

This division may be useful to figure out variables that may be used for building model specifications later, more robust and contemplating all kinds of variables. Also this was chosen in order to make the campaign decision making process easier since policies usually have well defined areas of impact, such as housing, employment, police forces, and so on.

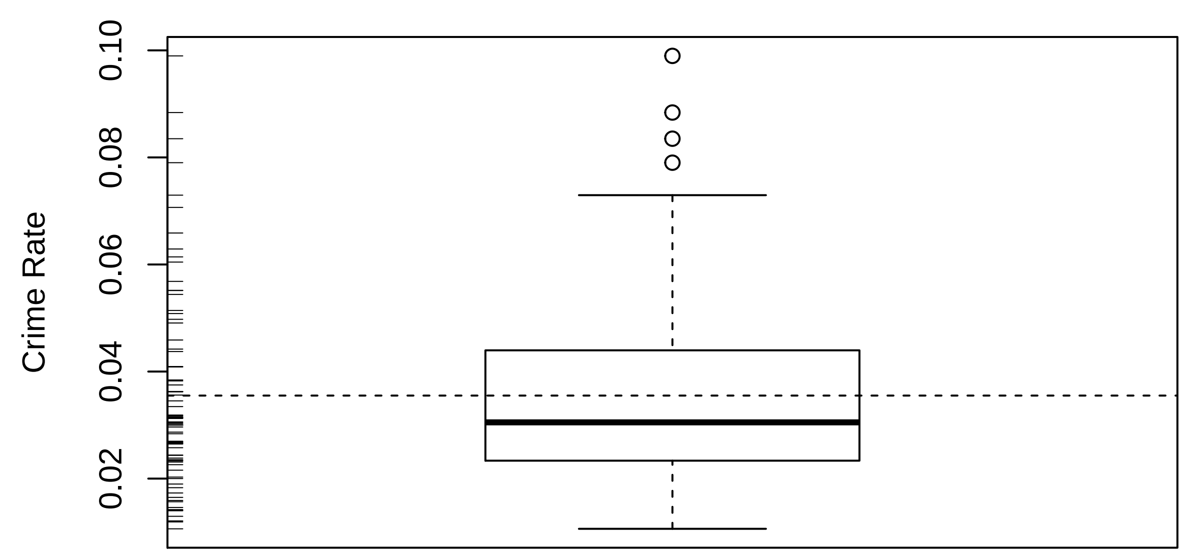
First of all, our goal is to understand the determinants of crime, crimes committed per person *crmrte* is more direct as to what we want to measure. Therefore, our dependent variable will be *crmte* (%). Let's first look at the un-transformed data.

```r
# to better understand the skewness distribution and it's spread graphically
par(mfrow=c(1,2))
hist(crime_data$crmrte, xlab="Crime Rate",
     col="light blue",
     main="Histogram of Crime Rate", ylim=c(0,30))
```

```
lines(density(crime_data$crmrte, na.rm=T),
      col="dark red")
rug(jitter(crime_data$crmrte))
qqnorm(crime_data$crmrte, main="QQ Plot of Crime Rate")
qqline(crime_data$crmrte, col="dark red")
```



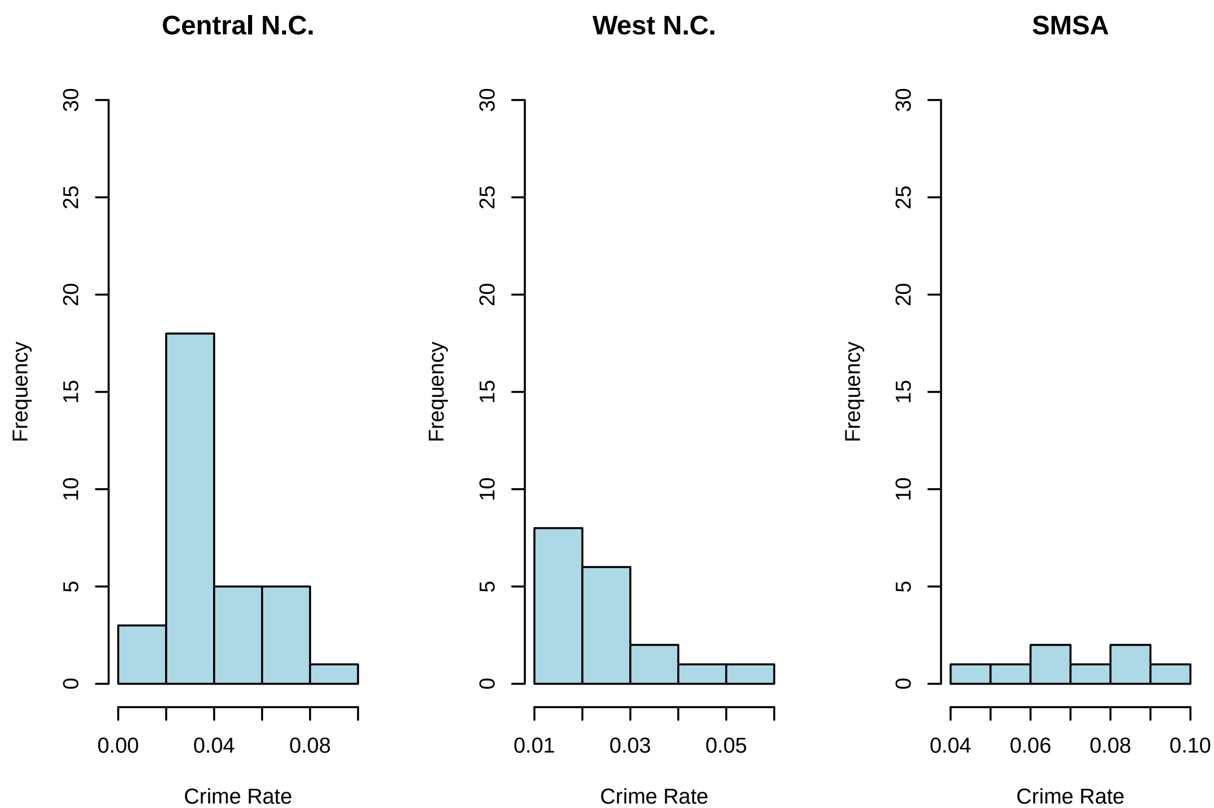

```
# boxplot
par(mfrow=c(1,1))
boxplot(crime_data$crmrte, ylab="Crime Rate")
rug(jitter(crime_data$crmrte), side=2)
abline(h=mean(crime_data$crmrte, na.rm=T), lty=2)
```



The crime rate has right skew with the mean at 0.033, and median at 0.030. The distribution is not normally distibuted. The box plot also shows more possible outliers have distorted the value of the mean as a statistic of centrality. Also, the variable *crmrte* has a distribution of the observed values concentrated on low values, thus with a positive skew.

One other observation is central N.C. tends to have higher frequency of crime rates than west N.C. and SMSA.

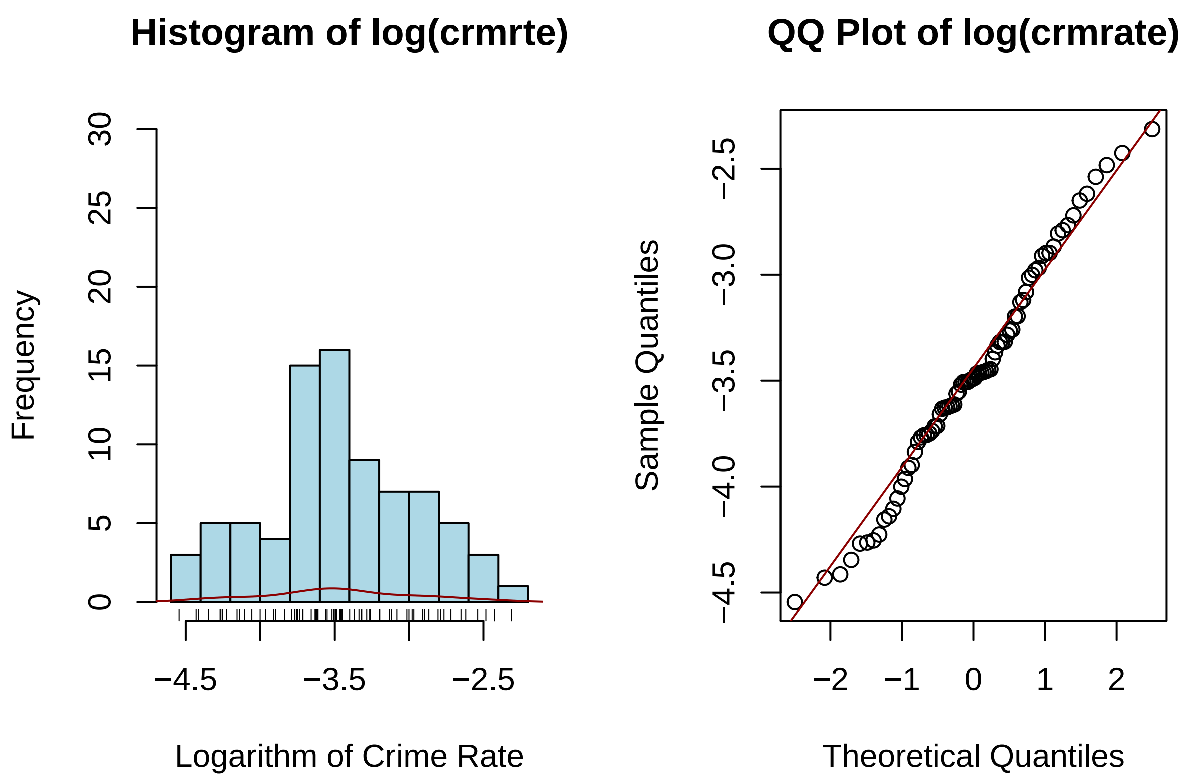
```r
par(mfrow=c(1,3))

# Histogram of Crime Rate in Central N.C.
hist(crime_data[crime_data$central == 1, ]$crmrte, col="light blue",
     main="Central N.C.", xlab="Crime Rate", ylim=c(0,30))

# Histogram of Crime Rate in West N.C.
hist(crime_data[crime_data$west == 1, ]$crmrte, col="light blue",
     main="West N.C.", xlab="Crime Rate", ylim=c(0,30))

# Histogram of Crime Rate in SMSA
hist(crime_data[crime_data$urban == 1, ]$crmrte, col="light blue",
     main="SMSA", xlab="Crime Rate", ylim=c(0,30))
```



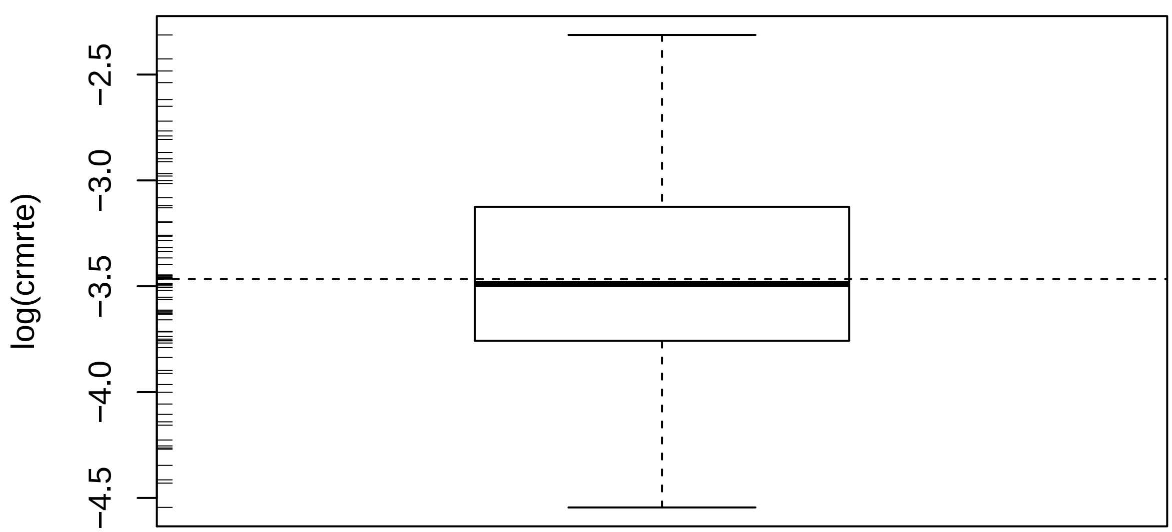
Now, let's see what happens if we apply log transformation on the dependent variable *crmrte*.

```r
# to better understand the skewness distribution and it's spread graphically
par(mfrow=c(1,2))
hist(log(crime_data$crmrte), xlab="Logarithm of Crime Rate",
     col="light blue",
     main="Histogram of log(crmrte)", ylim=c(0,30))
lines(density(log(crime_data$crmrte), na.rm=T),
```

```
       col="dark red")
rug(jitter(log(crime_data$crmrte)))
qqnorm(log(crime_data$crmrte), main="QQ Plot of log(crmrate)")
qqline(log(crime_data$crmrte), col="dark red")
```




```
# boxplot
par(mfrow=c(1,1))
boxplot(log(crime_data$crmrte), ylab="log(crmrte)")
rug(jitter(log(crime_data$crmrte)), side=2)
abline(h=mean(log(crime_data$crmrte), na.rm=T), lty=2)
```



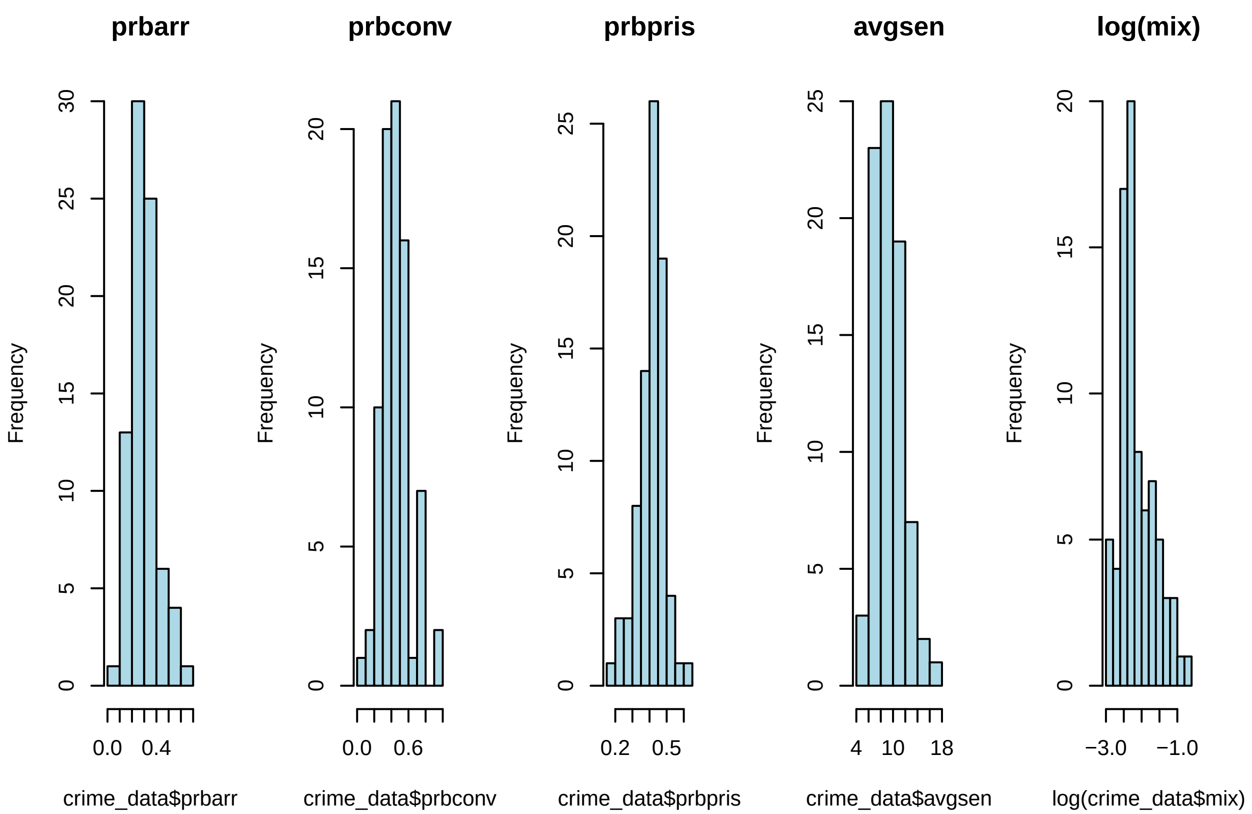
Clearly, if we apply log transformation on crime rate, our distribution becomes normally distibuted with mean and median to be very close, almost no skew and symmetric. This log transformed crime rate could be more ideal when it comes to modelling for OLS.

Next, we break the independent variables into 3 groups to examine the relationship against crime rate.

First group is crime-related variables: $prbarr, prbconv, prbpris, avgsen, mix$. This group could explain how convictions and police enforcement relate to crime rates. Inspecting histograms of each variable and turns out mix needs to be log transformed.

variable	label
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	avg. sentence, days
mix	offense mix: face-to-face/other

```
par(mfrow=c(1,5))
hist(crime_data$prbarr, col="light blue", main="prbarr") # close to normal
hist(crime_data$prbconv, col="light blue", main="prbconv") # close to normal
hist(crime_data$prbpris, col="light blue", main="prbpris") # close to normal
hist(crime_data$avgsen, col="light blue", main="avgsen") # close to normal
hist(log(crime_data$mix), col="light blue", main="log(mix)") # close to normal
```



First scatterplot matrix is crime rate with variables related to the nature of crime: probabilities of arrest, conviction and prison sentence, average sentence days, and log transformation of offense mix.

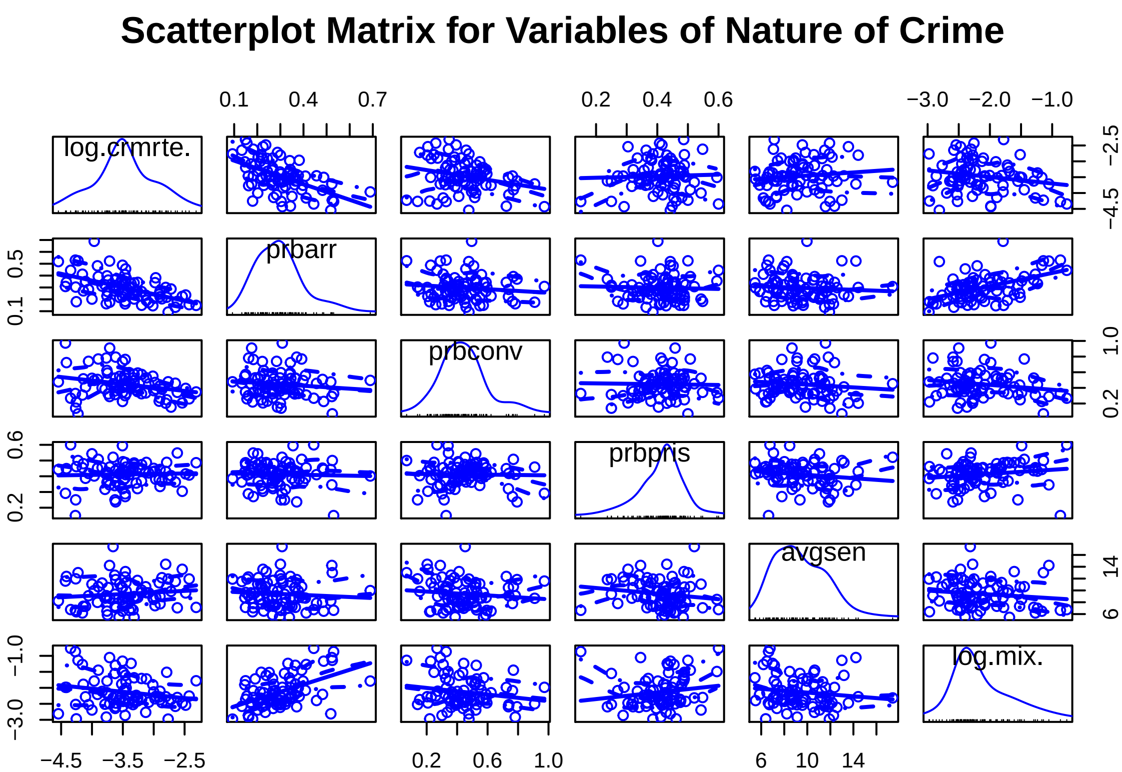
Here are some features noticed from the matrix:

- There are noticable negative relationship between crime rate and probability of arrest, crime rate and probability of conviction.

- There is strong positive relationship between probability of arrest and offense mix.
- Probability of prison sentence and average sentence days do not seem to have a strong relationship with any other variables in this group.

Additionally, it is interesting to the point that probabilty of arrest *prbarr* and probability of convicition *prbconv* are not highly correlated as we could expect from common sense. This indicates that keeping the two variables in an analysis will weaken our model due to multicolinearity, but furthter investigation will be necessary.

```
scatterplotMatrix(~ log(crmrte) + prbarr + prbconv + prbpris + avgsen + log(mix),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Nature of Crime")
```




```
cor(log(crime_data$crmrte), crime_data$prbarr, use="complete.obs")
```

```
## [1] -0.531073
```

```
cor(log(crime_data$crmrte), crime_data$prbconv, use="complete.obs")
```

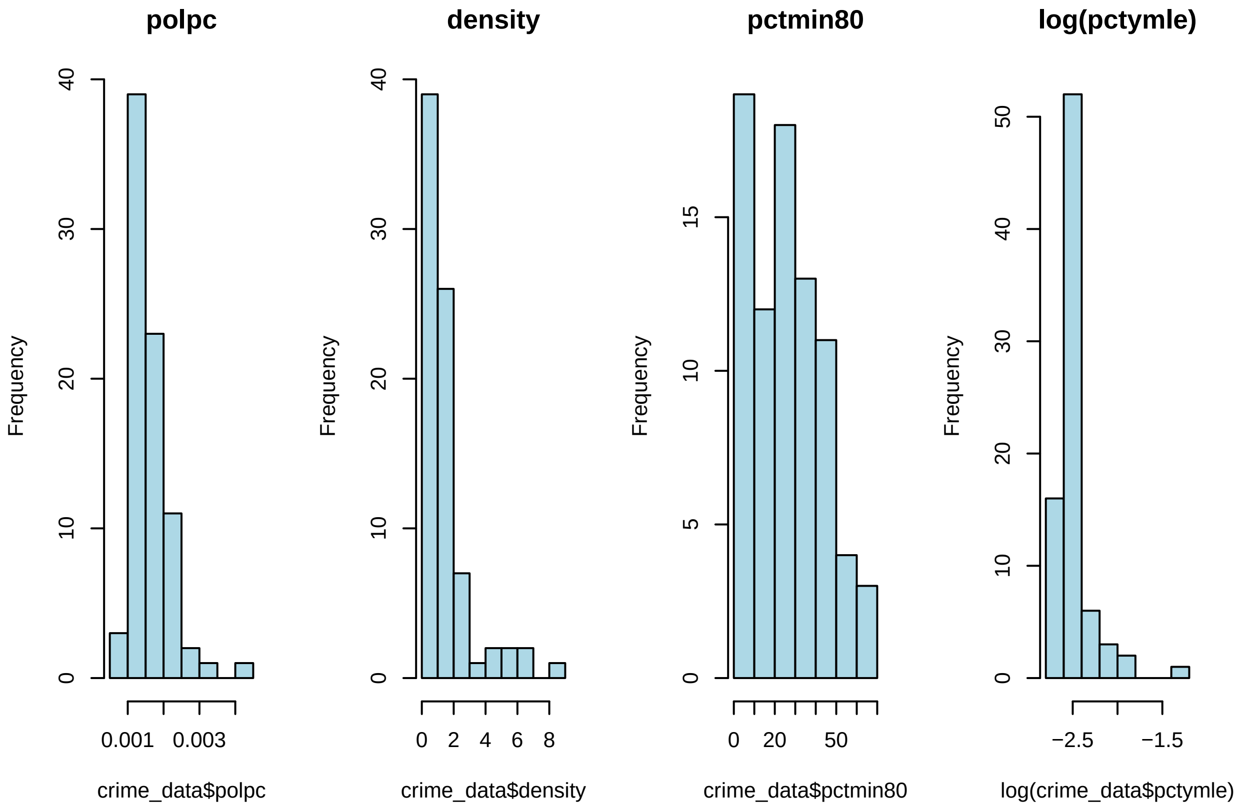
```
## [1] -0.2609103
```

Second group is population-related variables: $polpc, density, pctmin80, pctymle$. This group could explain how econo-geographic data influences crime rate. Inspecting histograms of each variable and turns out *pctymle* needs to be log transformed.

variable	label
crmrte	crimes committed per person
polpc	police per capita
density	people per sq. mile
pctmin80	perc. minority, 1980

variable	label
pctymle	percent young male

```
par(mfrow=c(1,4))
hist(crime_data$polpc, col="light blue", main="polpc") # close to normal
hist(crime_data$density, col="light blue", main="density") # right skew
hist(crime_data$pctmin80, col="light blue", main="pctmin80") # close to normal
hist(log(crime_data$pctymle), col="light blue", main="log(pctymle)") # right skew
```



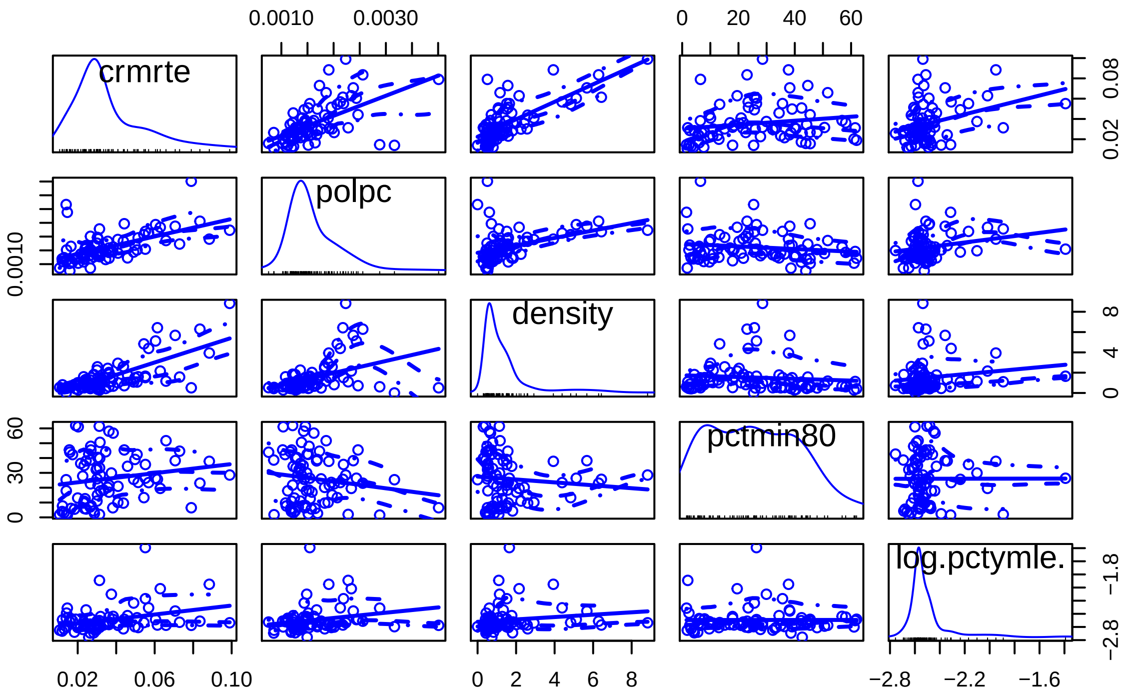
Second scatterplot matrix is crime rate with variables related to population: police per capita, people per square mile, % minority, and log transformation of % young male.

Here are some features noticed from the matrix:

- There are noticable positive relationship between crime rate and police per capita, crime rate and people per sq. mi., % young male and crime rate.
- Positive relationship between crime rate and police per capita seems to be an anomaly since crime rate is supposed to go down if there is more police per capita. Therefore, *polpc* could be a top-coded variable with data not reflected with appropriate variable name. This could also be explained by local governments increasing police presence in higher crime rate areas as an attempt to reduce crimies. If this is true, however, we can see that increasing police only can't reduce crime rate.
- % minority does not seem to have a strong relationship with any other variables in this group.

```
scatterplotMatrix(~ crmrte + polpc + density + pctmin80 + log(pctymle),
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Population")
```

Scatterplot Matrix for Variables of Population




```
cor(log(crime_data$crmrte), crime_data$density, use="complete.obs")
```

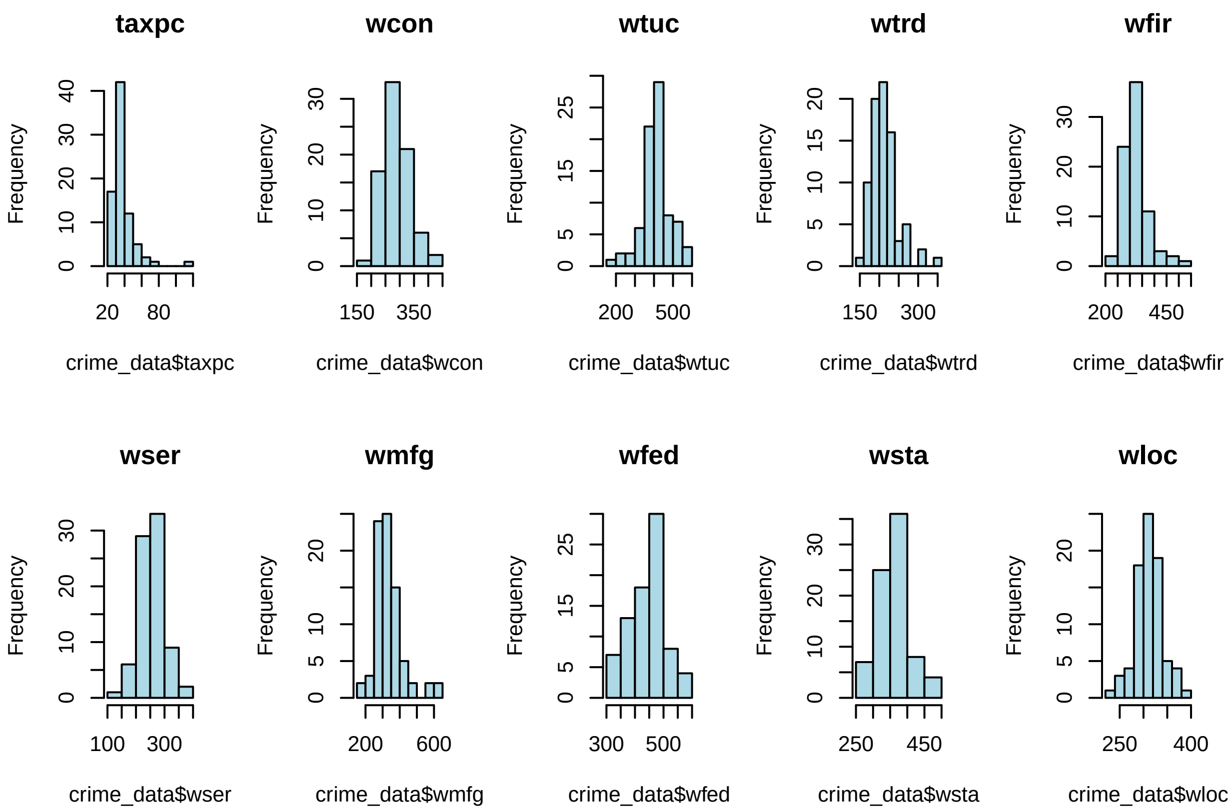
```
## [1] 0.6440777
```

Third group is economy-related variables: $taxpc, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wtoc$. This group could cover variations in wages and industry differences. Inspecting histograms of each variable.

variable	label
taxpc	tax revenue per capita
wcon	weekly wage, construction
wtuc	wkly wge, trns, util, commun
wtrd	wkly wge, whlesle, retail trade
wfir	wkly wge, fin, ins, real est
wser	wkly wge, service industry
wmfg	wkly wge, manufacturing
wfed	wkly wge, fed employees
wsta	wkly wge, state employees
wloc	wkly wge, local gov emps

```
par(mfrow=c(2,5))
hist(crime_data$taxpc, col="light blue", main="taxpc") # right skew
hist(crime_data$wcon, col="light blue", main="wcon") # close to normal
hist(crime_data$wtuc, col="light blue", main="wtuc") # close to normal
hist(crime_data$wtrd, col="light blue", main="wtrd") # close to normal
hist(crime_data$wfir, col="light blue", main="wfir") # close to normal

hist(crime_data$wser, col="light blue", main="wser") # close to normal
hist(crime_data$wmfg, col="light blue", main="wmfg") # close to normal
hist(crime_data$wfed, col="light blue", main="wfed") # close to normal
```

```r
hist(crime_data$wsta, col="light blue", main="wsta") # close to normal
hist(crime_data$wloc, col="light blue", main="wloc") # close to normal
```



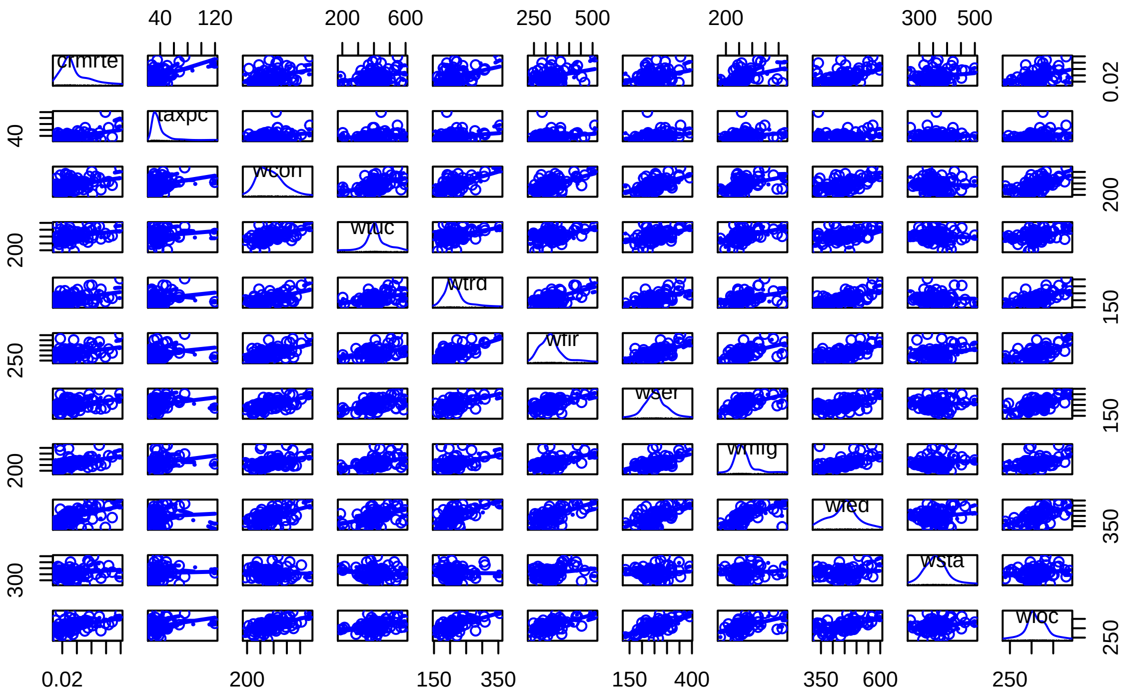
Third scatterplot matrix is crime rate with variables related to wages: tax revenue per capita, weekly wages of 6 different industries, and wages of federal, state, and local employees.

Here are some features noticed from the matrix:

- There are strong relationship between crime rate and all variables in this group.

```r
scatterplotMatrix(~ crmrte + taxpc + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc,
                  data = crime_data,
                  main = "Scatterplot Matrix for Variables of Wages" )
```

Scatterplot Matrix for Variables of Wages




```
cor(log(crime_data$crmrte), crime_data$wcon, use="complete.obs")
```

```
## [1] 0.3439327
```

```
cor(log(crime_data$crmrte), crime_data$wtrd, use="complete.obs")
```

```
## [1] 0.3700086
```

```
cor(log(crime_data$crmrte), crime_data$wfed, use="complete.obs")
```

```
## [1] 0.5307735
```

The Model Building Process

The purpose of this analysis is to identify independent variables relevant to the concerns of the political campaign in order to reduce crime rate.

Those variables found correlated to crime rate from EDA as follow:

- *prbarr*, *prbconv*, *taxpc*: these variables could potentially be applicable and implementable for policy suggestions.
- *density*, *pctymle*, *wcon*, *wtuc*, *wtrd*, $wfir$, *wser*, $wmfg$, $wfed$, *wsta*, *wloc*: these variables could not be directly applicable for policy suggestions.

The covariates that help us further identify a causal effect are *prbarr* and *prbconv*, *density* and *pctymle* based on output from scatterplots. On the other hand, the problematic covariates due to multicollinearity are *taxpc* and $w*$ (all wages variables) seen from the scatterplot above since they will absorb some of causal effect we want to measure.

We will consider building 3 model specifications:

1. Model with only the explanatory variables of key interest and no other covariates.

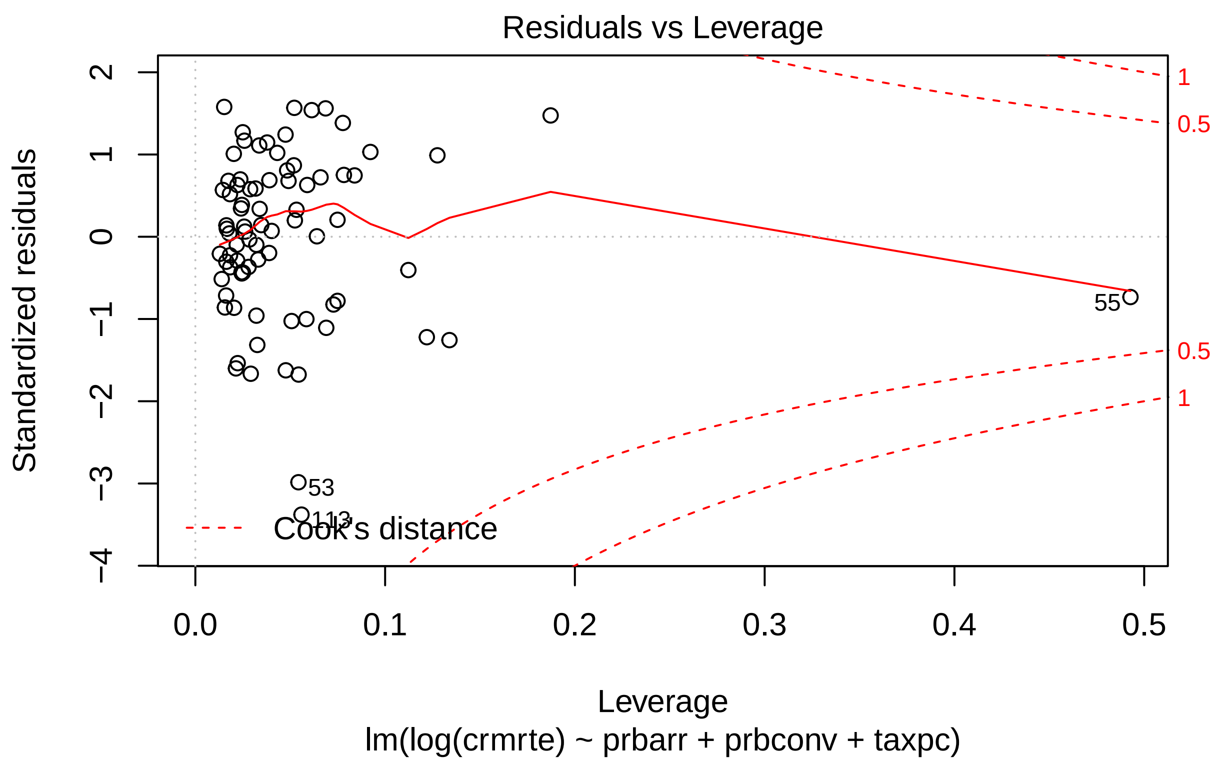
$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + \beta_3 taxpc + u$$

Picking variables which are only applicable for policy suggestions as the key interest with no other covariates from each variable. As discussed earlier, we decided to keep probabilities of arrest and conviction in our model, since they are not highly correlated as common sense could infer.

```
(model1 = lm(log(crmrte) ~ prbarr + prbconv + taxpc, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + prbconv + taxpc, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      prbconv        taxpc
##   -2.768861    -2.480395    -0.723735     0.009519
```

```
plot(model1, which = 5)
```




```
summary(model1)$r.square
```

```
## [1] 0.4432557
```

```
summary(model1)$adj.r.squared
```

```
## [1] 0.4212789
```

```
AIC(model1)
```

```
## [1] 80.72369
```

```
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.7688614  0.3417654 -8.1016 7.033e-12 ***
## prbarr      -2.4803950  0.4890282 -5.0721 2.704e-06 ***
## prbconv     -0.7237350  0.3511749 -2.0609   0.04273 *
## taxpc        0.0095188  0.0037368  2.5473   0.01288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model suggests that points 53, 55, and 113 have high Cook's distance, so further investigations were made to define why they are outliers comparing to the expectations of the variables of all other counties.

```r
avg_county <- colMeans(crime_data)
outliers_compare <- data.frame(t(avg_county))
outliers_compare$county <- 999
outliers_compare <- rbind(crime_data[c("53","55","113"),], outliers_compare)
head(outliers_compare)
```

```
##     county year     crmrte    prbarr  prbconv   prbpris  avgsen
## 53      53   87 0.01406550 0.3031910 0.140351 0.2500000 11.9600
## 55      55   87 0.07901630 0.2246280 0.207831 0.3043480 13.5700
## 113    113   87 0.01420710 0.1798780 0.220339 0.4615380  6.3900
## 1      999   87 0.03551236 0.2971094 0.446486 0.4119534  9.4055
##           polpc   density     taxpc  west central urban pctmin80     wcon
## 53  0.001122250 0.5351562  50.38139 0.000     0.0   0.0 17.90960 266.4504
## 55  0.004009620 0.5115089 119.76145 0.000     0.0   0.0  6.49622 309.5238
## 113 0.001516000 0.4487427  40.80142 1.000     0.0   0.0  2.39865 244.7552
## 1   0.001615639 1.5170512  38.16108 0.225     0.4   0.1 26.02239 287.9047
##         wtuc     wtrd     wfir     wser     wmfg     wfed     wsta
## 53  202.4292 219.7802 305.9441 223.8502 250.4200 371.7900 383.7200
## 55  445.2762 189.7436 284.5933 221.3903 319.2100 338.9100 361.6800
## 113 412.0879 154.2090 256.4102 265.1301 291.1000 337.0900 374.1100
## 1   410.0088 212.4555 322.0438 254.6922 336.1615 444.9141 359.8099
##         wloc        mix    pctymle
## 53  296.6400 0.08045977 0.08476309
## 55  326.0800 0.08437271 0.07613807
## 113 246.6500 0.05128205 0.09171820
## 1   311.6226 0.13611190 0.08462820
```

We see that no major deviations are found. When analyzing county 55, we see that this is an area with considerable high crime rates, but with all variables not very different from the others. Based on population density and the wages, we can infer this is a rural county, which economy is based on construction and transportation industries. The highlight here is the *taxpc* variable: it is more than 3 times the average. This discrepancy probably is generating our leverage, but as we don't have enough evidence that this is a erroneous error, we will keep that point in the data set.

The counties represented by 53 and 113 are the opposite. They have crime rates as low as 40% of the state average. Small population density and no assumptions can be made based only in the wages per industry. As we did with 55, we found no strong evidence that thi data is wrong, thus keeping the data in our data set. It is also interesting to note that the 3 outliers have completely different values for $pctmin80$, yet still have similar values for most variables. This endorses our decision to not use that variable.

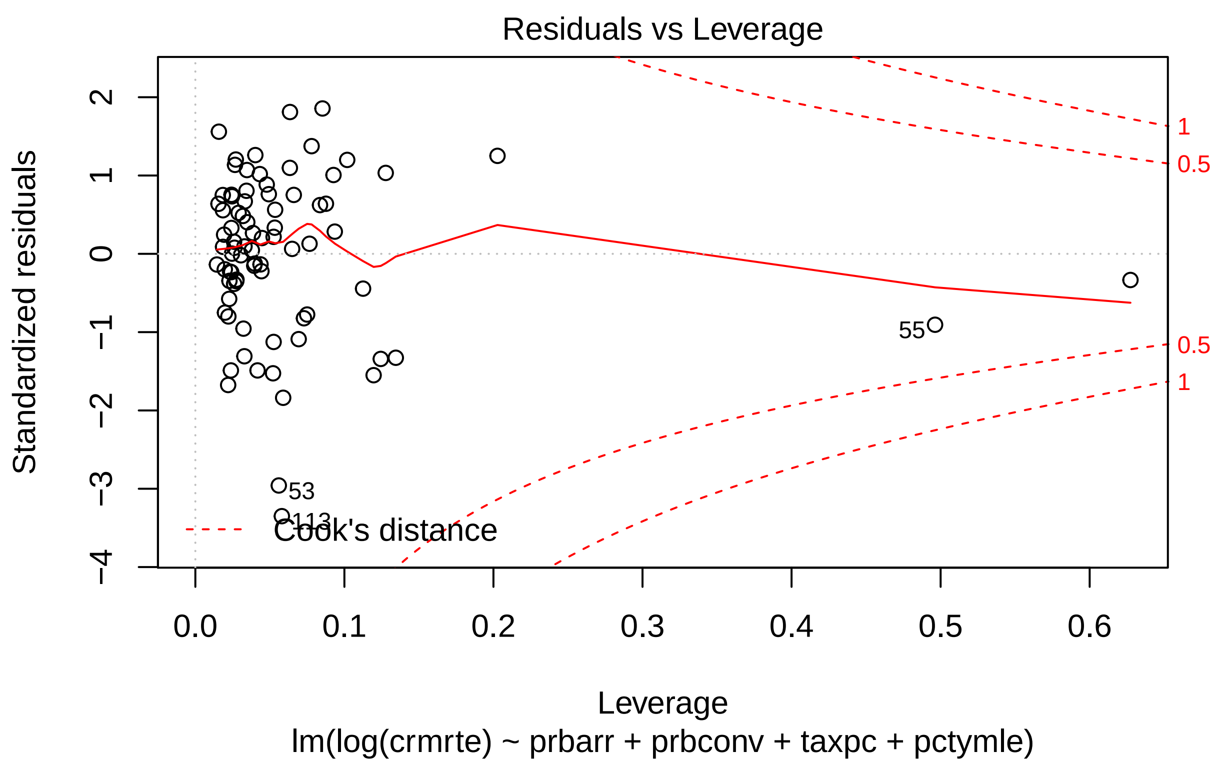
2. Model that includes key explanatory variables and only covariates that we believe increase the accuracy of your results.

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 taxpc + \beta_3 pctymle + u$$

```
(model2 = lm(log(crmrte) ~ prbarr + prbconv + taxpc + pctymle, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + prbconv + taxpc + pctymle,
##     data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      prbconv        taxpc      pctymle
##    -3.27553     -2.26146     -0.55978      0.01109      3.64335
```

```
plot(model2, which = 5)
```




```
summary(model2)$r.square
```

```
## [1] 0.4684787
```

```
summary(model2)$adj.r.squared
```

```
## [1] 0.4401309
```

```
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -3.2755260  0.4622750 -7.0857 6.414e-10 ***
```

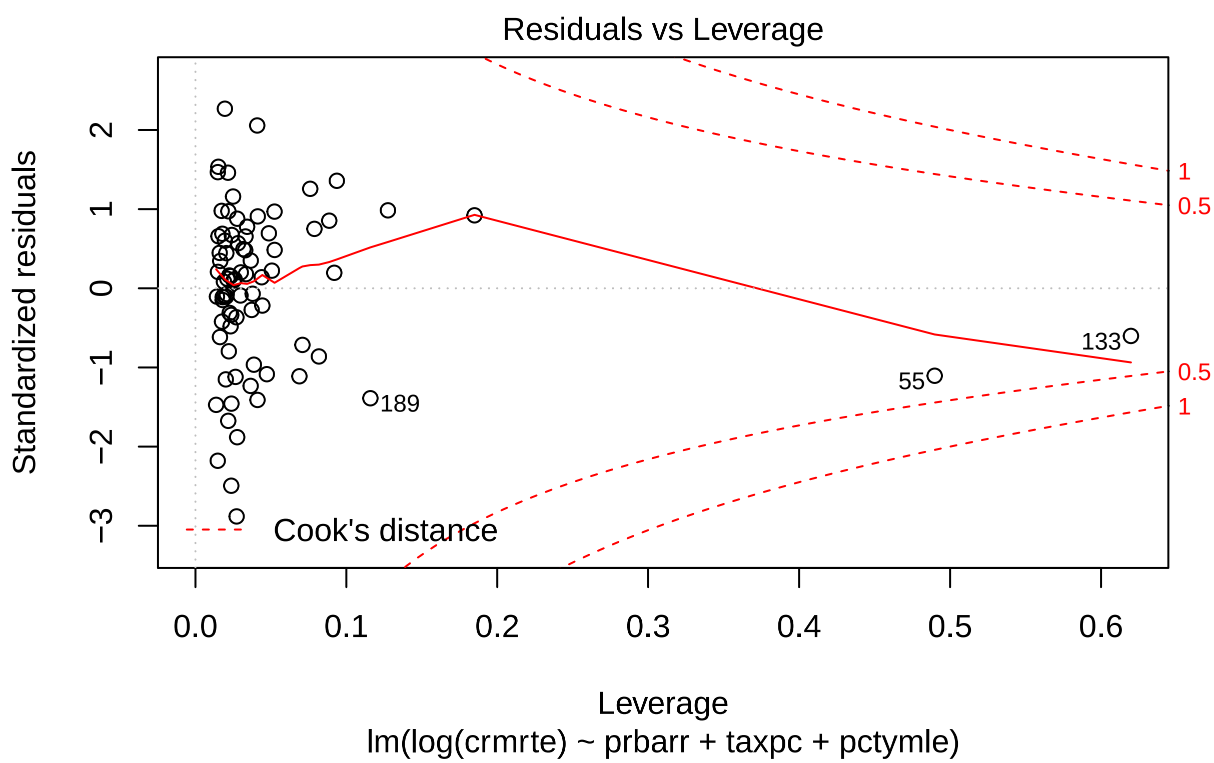
```
## prbarr      -2.2614640  0.4883784 -4.6306 1.501e-05 ***
## prbconv     -0.5597843  0.3829771 -1.4617   0.14801
## taxpc        0.0110934  0.0042646  2.6013   0.01118 *
## pctymle      3.6433462  1.6436939  2.2166   0.02968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note by the coefficients that the *prbconv* variable is the only one without statistical significance in this test, so the model was rebuilt to remove that variable. As stated earlier, this variable is not correlatd to *prbarr*, however it can have relationship with the other (even omitted ones that were lumped under the error term and are not available for analysis).

```r
(model2 = lm(log(crmrte) ~ prbarr + taxpc + pctymle, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + taxpc + pctymle, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr        taxpc      pctymle
##    -3.78901     -2.06521      0.01379      4.85466
```

```r
plot(model2, which = 5)
```




```r
summary(model2)$r.square
```

```
## [1] 0.4398461
```

```r
summary(model2)$adj.r.squared
```

```
## [1] 0.4177347
```

```r
AIC(model2)
```

```
## [1] 81.21213
```

```r
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.7890097  0.2994521 -12.6531 < 2.2e-16 ***
## prbarr      -2.0652054  0.4027572  -5.1277 2.173e-06 ***
## taxpc        0.0137853  0.0043696   3.1548  0.002301 **
## pctymle      4.8546614  1.8997658   2.5554  0.012604 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted R^2 increases by 0.84% by adding one additional variable, and AIC decreases by 0.61% to indicate improvements on parsimony. However, there is not a significant improvement as the solid red line still getting very close to the danger zone of Cook's distance.

Additionally, there is a new outlier, county 133, which we can investigate.

```r
outliers_compare <- rbind(crime_data[c("133"),], outliers_compare)
outliers_compare[c("133","1"),]
```

```
##     county year     crmrte    prbarr  prbconv   prbpris avgsen       polpc
## 133    133   87 0.05512870 0.2669600 0.271947 0.3349510 8.9900 0.001544570
## 1      999   87 0.03551236 0.2971094 0.446486 0.4119534 9.4055 0.001615639
##      density    taxpc  west central urban pctmin80     wcon     wtuc
## 133 1.650066 27.46926 0.000     0.0   0.0 26.38140 264.0406 318.9644
## 1   1.517051 38.16108 0.225     0.4   0.1 26.02239 287.9047 410.0088
##         wtrd     wfir     wser     wmfg     wfed     wsta     wloc
## 133 183.2609 265.1232 230.6581 258.2500 326.1000 329.4300 301.6400
## 1   212.4555 322.0438 254.6922 336.1615 444.9141 359.8099 311.6226
##           mix   pctymle
## 133 0.1217632 0.2487116
## 1   0.1361119 0.0846282
```

As expected, the *pctymle* variable is substantially higher than the state average. However, there are no evidences that there is an error in our data. So the observation will still be used in our data set.

3. Model that includes the previous covariates, and most, if not all, other covariates.

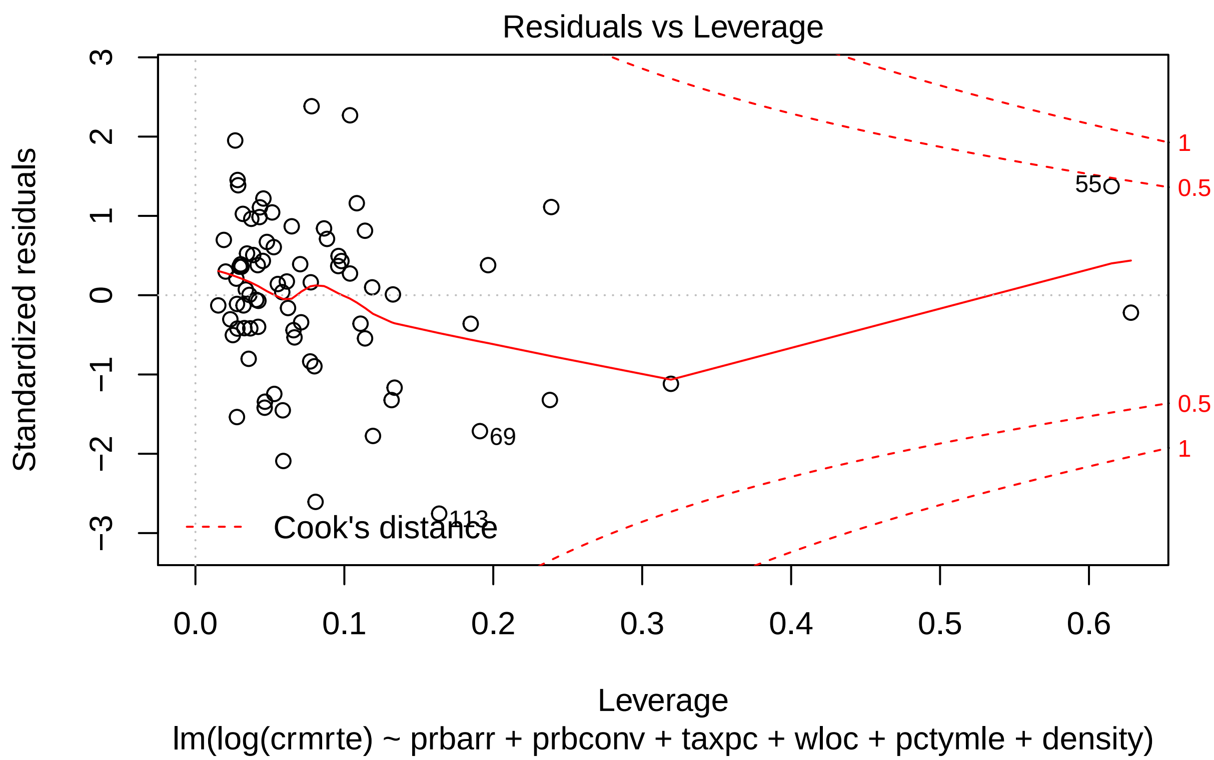
$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + \beta_3 taxpc + \beta_4 wloc + \beta_5 pctymle + \beta_6 density + u$$

```r
(model3 = lm(log(crmrte) ~ prbarr + prbconv + taxpc + wloc + pctymle + density,
             data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + prbconv + taxpc + wloc +
##     pctymle + density, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      prbconv        taxpc         wloc
```

```
##    -4.154101    -1.487671    -0.345699     0.007008     0.001733
##      pctymle      density
##     3.561869     0.115910
```

```r
plot(model3, which = 5)
```




```r
summary(model3)$r.square
```

```
## [1] 0.5937206
```

```r
summary(model3)$adj.r.squared
```

```
## [1] 0.5603277
```

```r
AIC(model3)
```

```
## [1] 61.5185
```

Adjusted R^2 increases by 33.0% by adding 3 additional variables, and AIC decreases by 23.8% to indicate further improvements on parsimony. Moreover, there is a significant improvement since the solid red line moves away from the danger zone of Cook's distance.

The Regression Table

Now consolidating all statistical findings from these 3 models to a regression table.

```r
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))
stargazer(model1, model2, model3, type = "latex",
          title = "Linear Models Predicting Crime Rate",
          omit.stat = "f",
```

```
         se = list(se.model1, se.model2, se.model3),
         star.cutoffs = c(0.05, 0.01, 0.001))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Aug 06, 2018 - 16:06:07

Table 5: Linear Models Predicting Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
prbarr	−2.480***	−2.065***	−1.488**
	(0.489)	(0.403)	(0.544)
prbconv	−0.724*		−0.346
	(0.351)		(0.371)
taxpc	0.010*	0.014**	0.007
	(0.004)	(0.004)	(0.006)
wloc			0.002
			(0.002)
pctymle		4.855*	3.562**
		(1.900)	(1.300)
density			0.116**
			(0.036)
Constant	−2.769***	−3.789***	−4.154***
	(0.342)	(0.299)	(0.911)
Observations	80	80	80
R^2	0.443	0.440	0.594
Adjusted R^2	0.421	0.418	0.560
Residual Std. Error	0.386 (df = 76)	0.387 (df = 76)	0.337 (df = 73)
Note:			*p<0.05; **p<0.01; ***p<0.001

According to Table 5[1], for Model 1, increasing the probability of arrest will reduce crime rate with minimal effect from tax revenue per capita. For Model 2, on top of Model 1, decreasing % of young male will reduce crime rate. For Model 3, on top of Model 2, increasing both probabilities of arrest and conviction, decreasing people per sq. mi. will reduce crime rate.

The Model Assumptions and Statistical Inference Discussion

Model 2 is being picked as our most important model specification as all 3 independent variables (*prbarr*, *taxpc*, *pctymle*) are statistically significant. A detailed assessment of all 6 classical linear model assumptions will be performed.

[1] Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

1. Linear population model

We can assume that this model has linear coefficients only because we have not constrained our error term. The assumption that the error term will incorporate non-linearities is true, and so the model is linear.

2. Random sampling

While background data was not provided for this analysis, we notice that our sample has 80 different counties. A quick search on North Carolina website shows 100 counties in that state, with the youngest one created in 1911 and none incorporated by other since then. Under this fact, the assumption that the numbers are official from each county. We can assume that we have analyzed data referent to 80% of the population, being enough to reduce the non-random sampling effect to minimum.

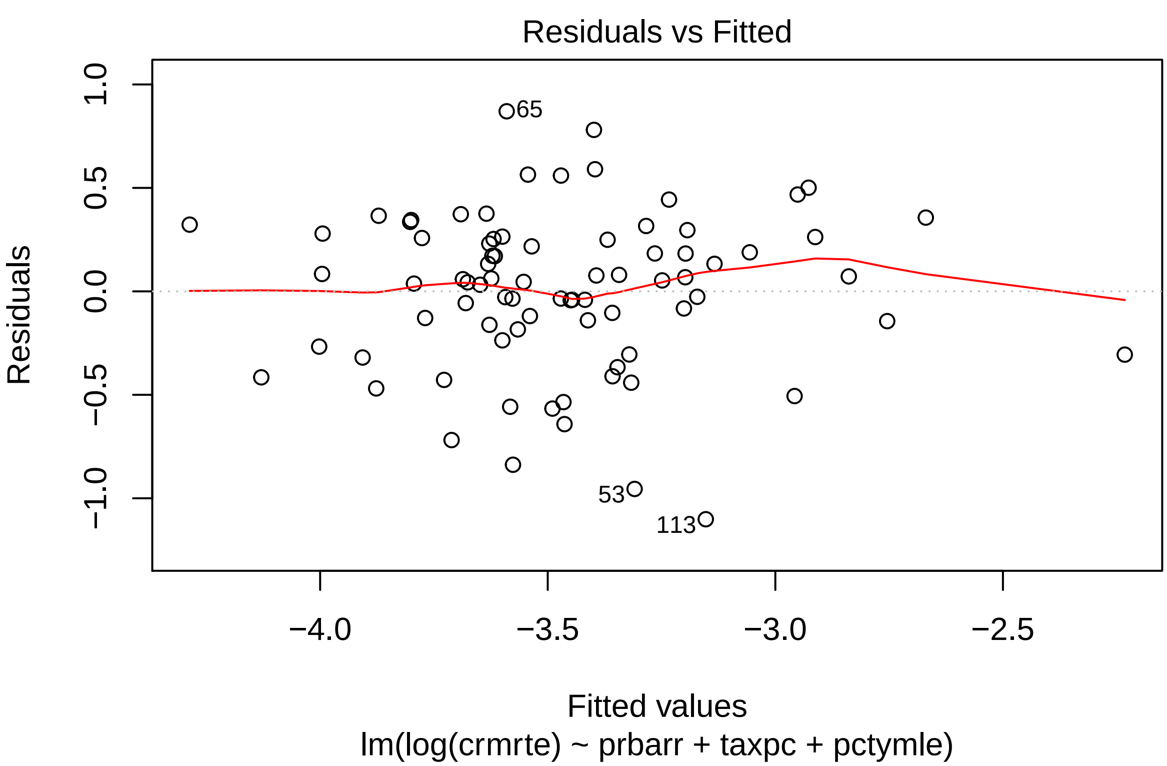
3. No perfect multicollinearity

As R didn't warn for any perfect collinearity, this assumption is met. Additionally we visually checked for that using `scatterplotMatrix` and the correlation index.

4. Zero-conditional mean

We start looking at the diagnostic plot:

```r
plot(model2, which=1)
```

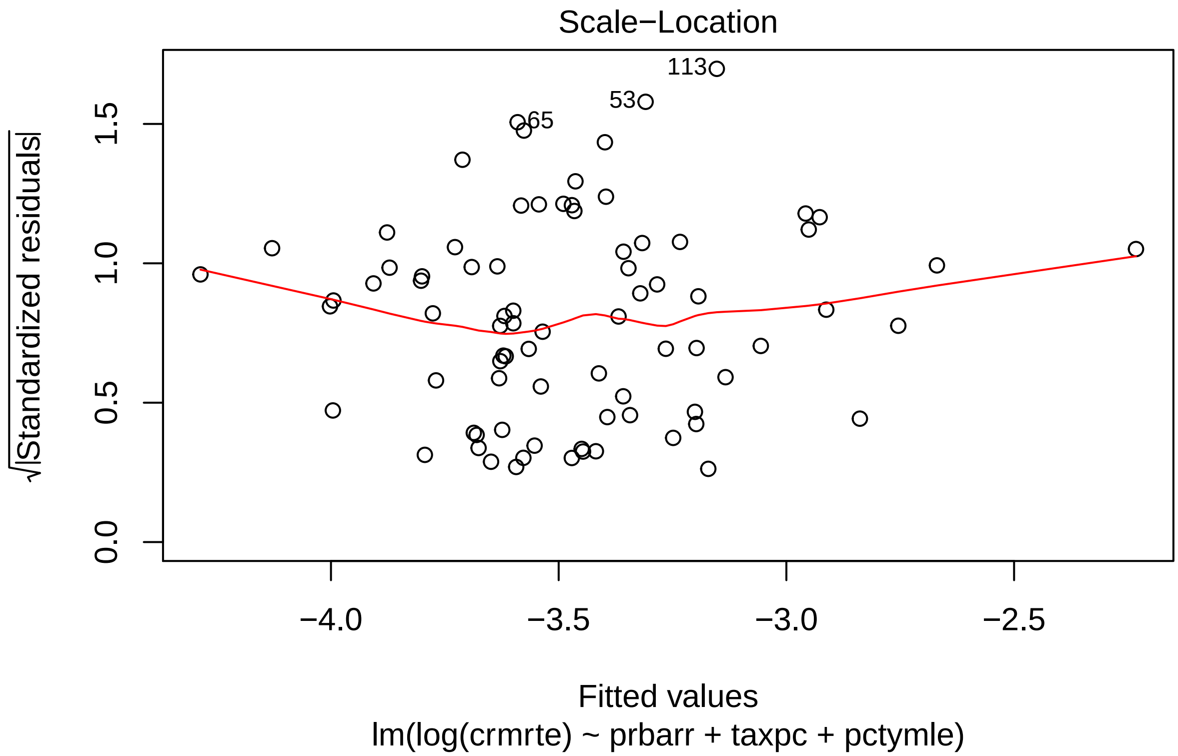


As evidenced in residuals versus fitted values plot, there is no clear deviation from zero conditional mean indicated by the red line. Therefore, we can consider the zero-conditional assumption met.

5. Homoskedasticity

The residuals versus fitted values plot doesn't seem to indicate heterskedasticity, because the band seems to have even thickness. The scale location plot gives us another way to access this assumption:

```r
plot(model2, which=3)
```




```r
bptest(model2)
```

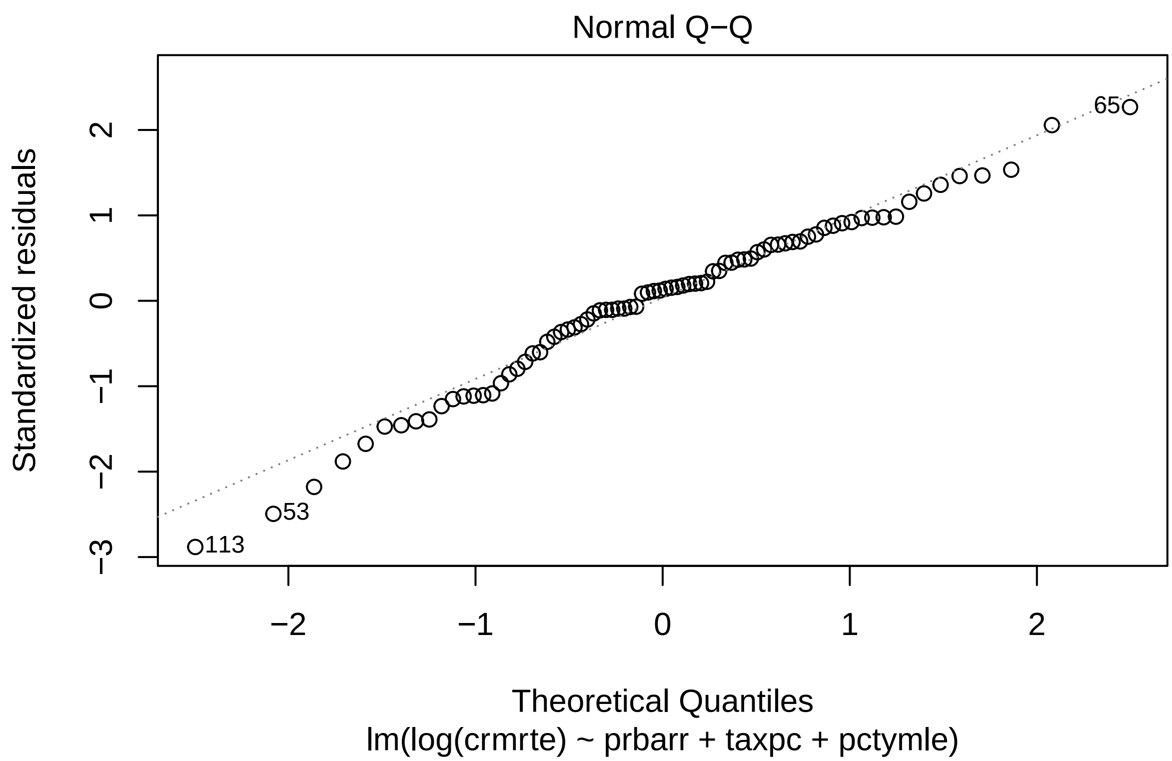
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 0.45592, df = 3, p-value = 0.9285
```

The fairly flat red line also suggests homoskedasticity. Despite this evidence, we will proceed with robust standard errors, because that is good conservative practice. Also, through a Breusch-Pagan test, the null hypothesis is the model has homoskedasticity. p-value indicates we can't reject the null hypothesis, meaning heteroskedasticity is not present.

6. Normality of errors

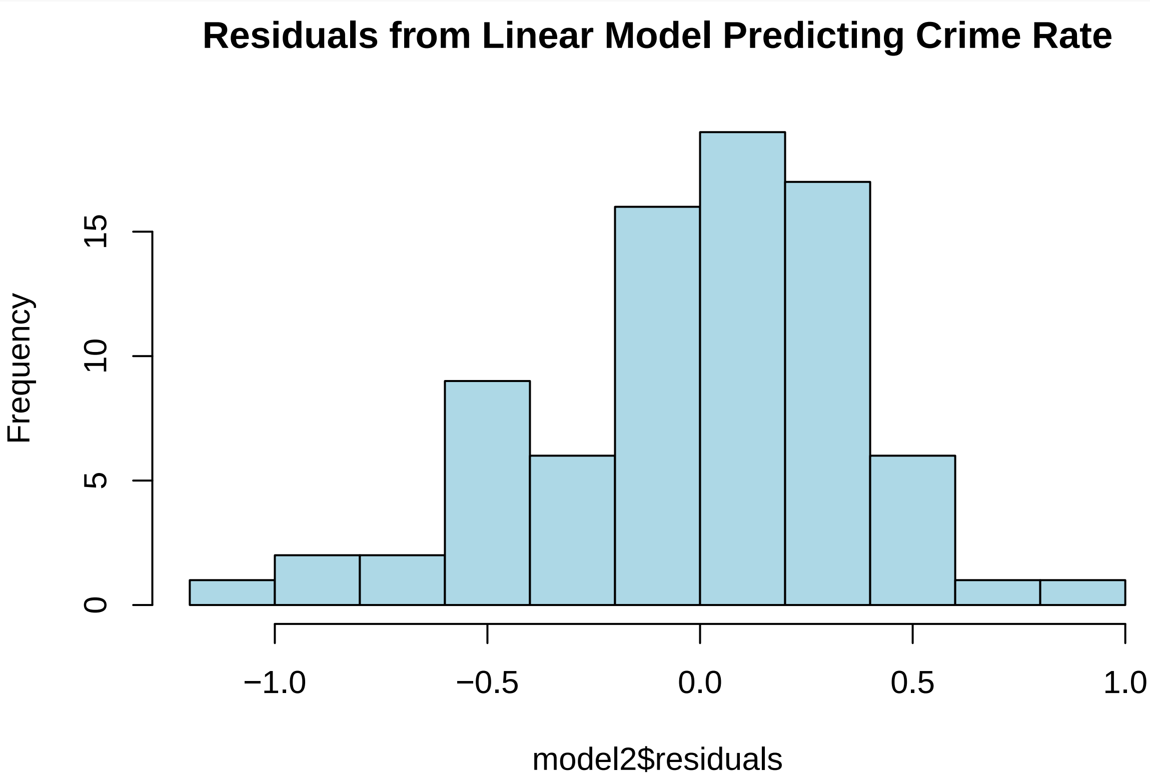
To check normality of errors, we can look at the qqplot that is part of R's standard diagnostics:

```r
plot(model2, which=2)
```



We can also visually look at the residuals directly:

```r
hist(model2$residuals, breaks=10, col="light blue",
     main="Residuals from Linear Model Predicting Crime Rate")
```



We have a sample size > 30, so the CLR tells us that our estimators will have a normal sampling distribution. We might also consider the formal Shapiro-Wilk test of normality. The null hypothesis is the residuals are normally distributed. p-value indicates it can't be rejected, meaning residuals are with normal distribution.

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97907, p-value = 0.2136
```

Next, inference for linear regression and standard errors via statistical tests will be inspected through model coefficients completed with standard errors that are valid given our diagnostics. We noticed that *prbarr*, *taxpc*, and *pctymle* are all statistically significant.

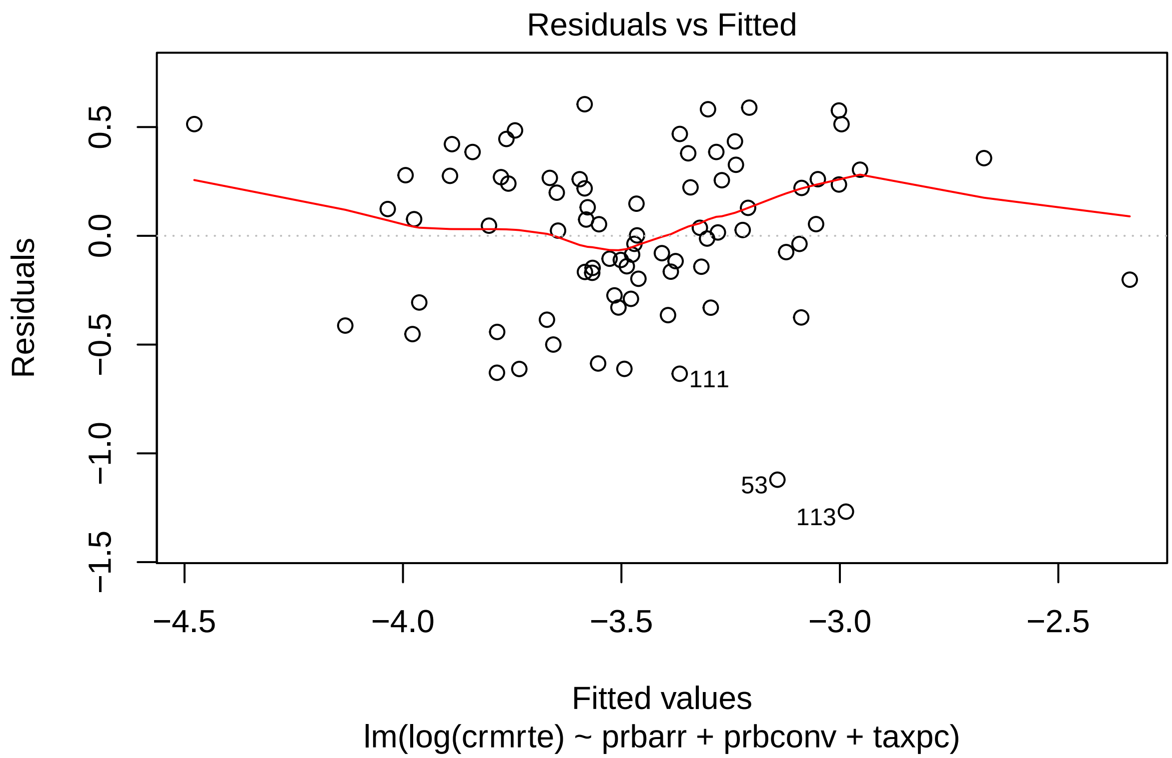
```
coeftest(model2, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.7890097  0.2994521 -12.6531 < 2.2e-16 ***
## prbarr      -2.0652054  0.4027572  -5.1277 2.173e-06 ***
## taxpc        0.0137853  0.0043696   3.1548  0.002301 **
## pctymle      4.8546614  1.8997658   2.5554  0.012604 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

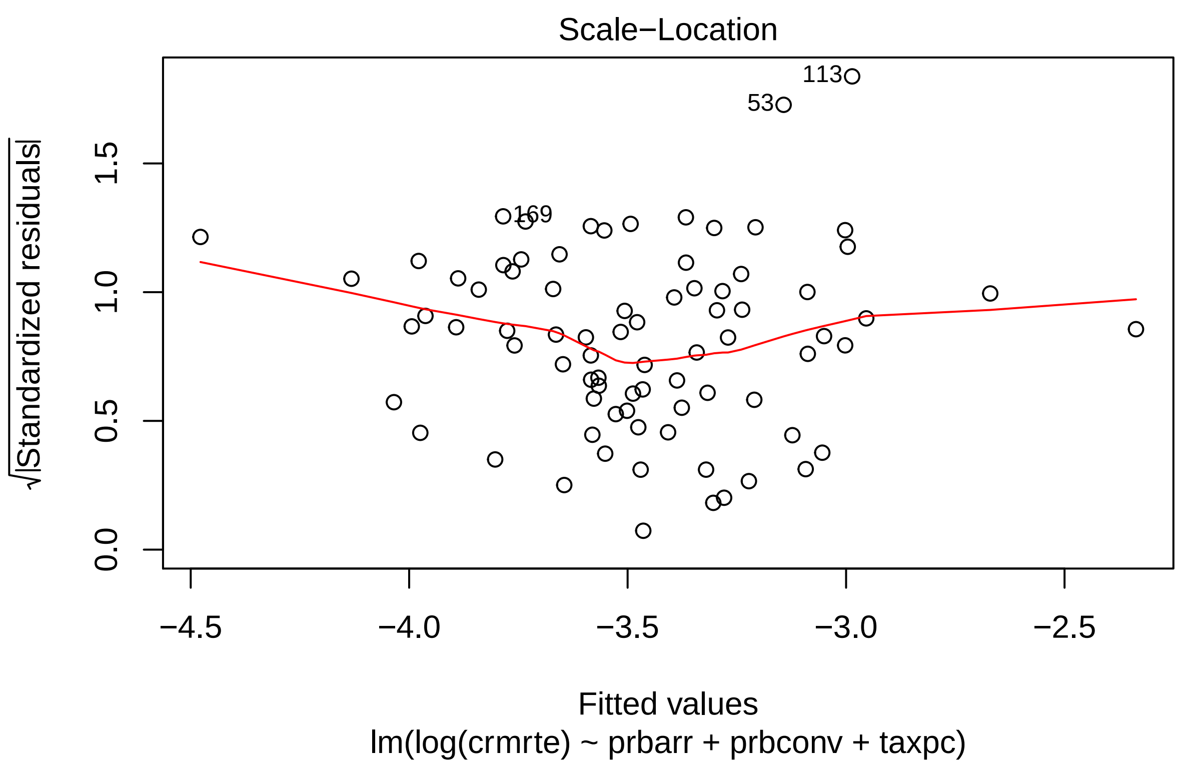
In general, Model 2 doesn't seem to violate any of the 6 linear model assumptions.

However, Model 1 demostrates violation of zero-conditional mean, homoskedasticity, and normality of errors:

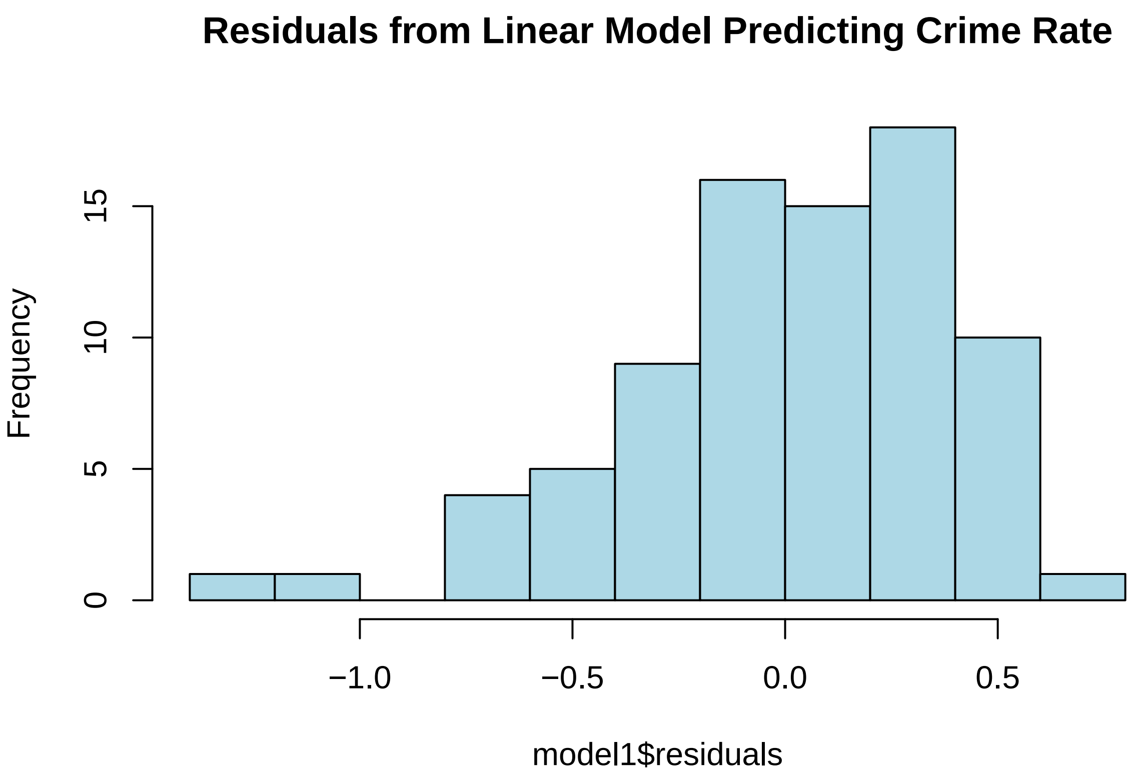
```
plot(model1, which=1) # red line is not flat enough
```




```r
plot(model1, which=3) # red line is parabolic
```

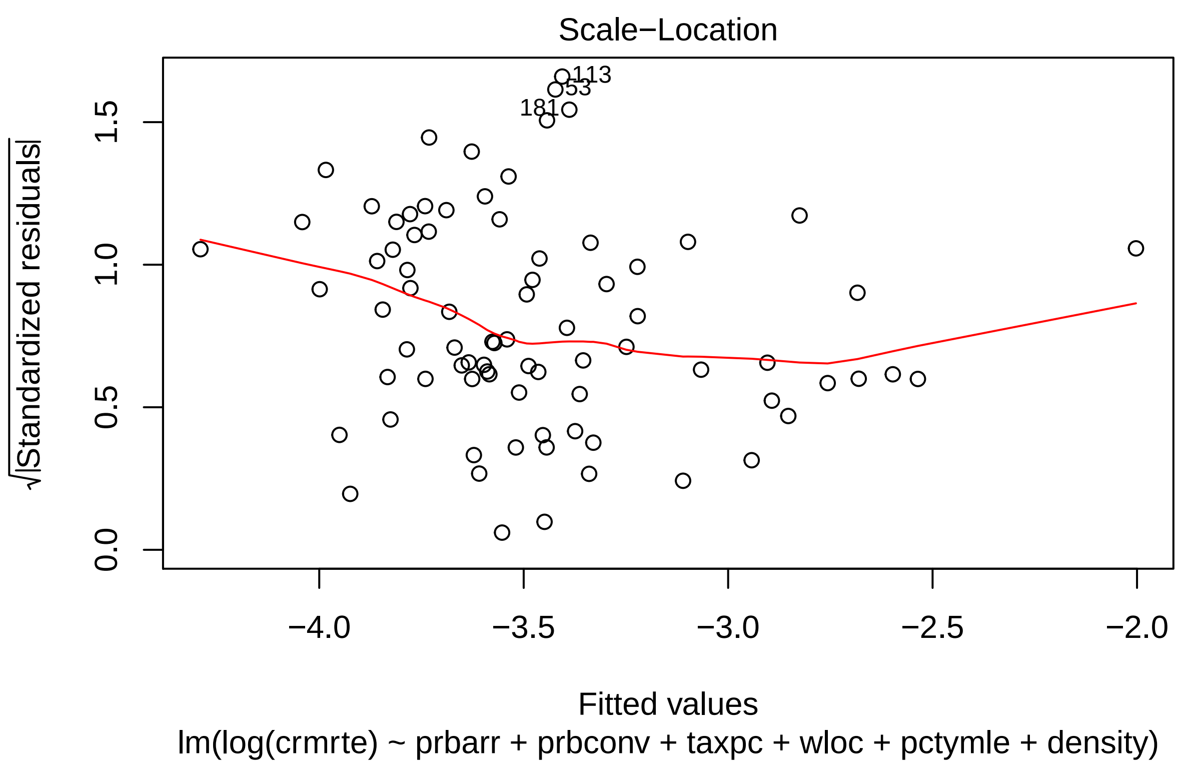



```r
hist(model1$residuals, breaks=10, col="light blue",
     main="Residuals from Linear Model Predicting Crime Rate") # right skew
```

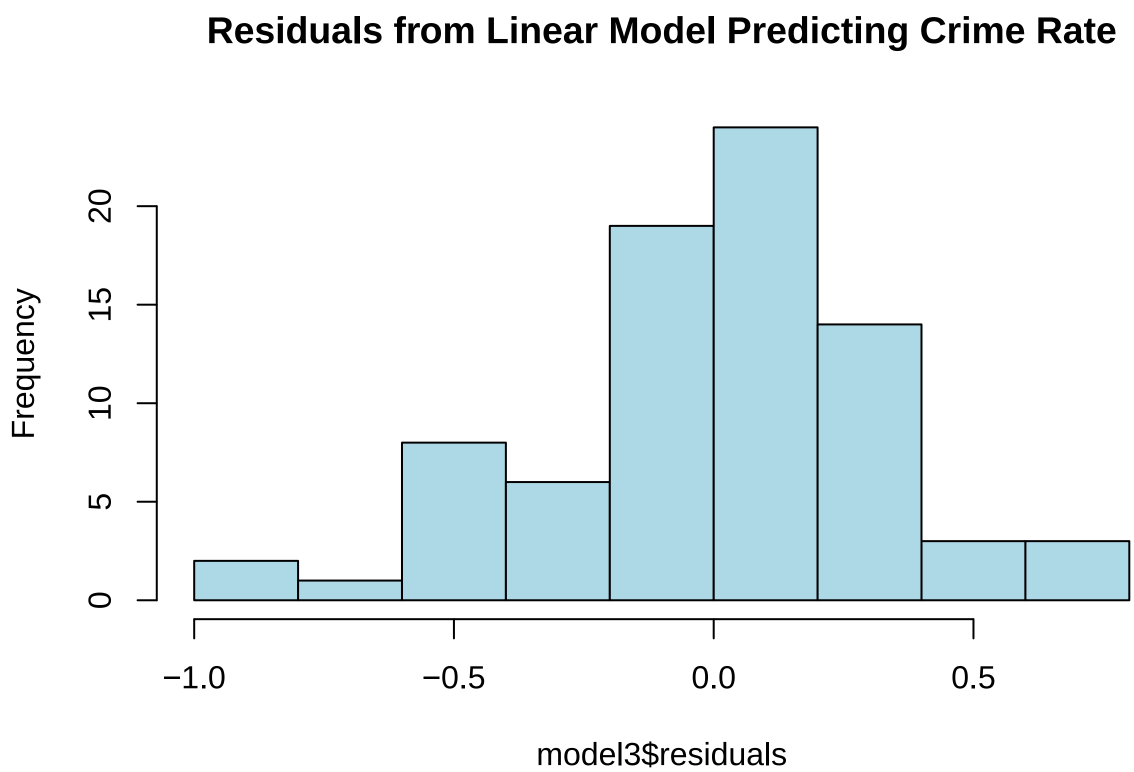


Model 3 demostrates violation of homoskedasticity and normality of errors:

```r
plot(model3, which=3) # red line is parabolic
```




```r
hist(model3$residuals, breaks=10, col="light blue",
     main="Residuals from Linear Model Predicting Crime Rate") # left skew
```



To test whether the difference in fit is significant, we use the wald test, which generalizes the usual F-test of overall significance, but allows for a heteroskedasticity-robust covariance matrix. p-value indicates that the difference in fit is statistically significant.

```r
waldtest(model3, model2, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: log(crmrte) ~ prbarr + prbconv + taxpc + wloc + pctymle + density
## Model 2: log(crmrte) ~ prbarr + taxpc + pctymle
##   Res.Df Df      F    Pr(>F)
## 1     73
## 2     76 -3 6.9932 0.0003364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we could test the additional 3 variables in Model 3 and see if they are jointly significant. In fact, they are and there is probably a great deal of multicollinearity.

```r
linearHypothesis(model3, c("prbconv = 0", "wloc = 0", "density = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbconv = 0
## wloc = 0
## density = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + taxpc + wloc + pctymle + density
##
## Note: Coefficient covariance matrix supplied.
```

```
##
##   Res.Df Df      F    Pr(>F)
## 1     76
## 2     73  3 6.9932 0.0003364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we could test if coefficients of *prbarr* and *prbconv* are the same. In turns out that this hypothesis is statistically significant.

```r
linearHypothesis(model3, "prbarr = prbconv", vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr - prbconv = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + taxpc + wloc + pctymle + density
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df     F  Pr(>F)
## 1     74
## 2     73  1 6.026 0.01648 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Omitted Variables Discussion

The omitted variables discussion will be based on Model 1 with *taxpc* dropped since its effect is minimal with following 5 variables omitted one at a time.

1. Omitted *taxpc*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 taxpc + u$$

$$taxpc = \alpha_0 + \alpha_1 prbarr + u$$

```r
(omit1_pri = lm(log(crmrte) ~ prbarr + taxpc, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + taxpc, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr        taxpc
##    -3.26307     -2.30354      0.01262
```

```r
(omit1_sec = lm(taxpc ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = taxpc ~ prbarr, data = crime_data)
##
```

```
## Coefficients:
## (Intercept)       prbarr
##       42.09       -13.21
```

Since $\beta_2 = 0.01279$ and $\alpha_1 = -12.89$, then $OMVB = \beta_2\alpha_1 = -0.1649$. Since $\beta_1 = -2.2938 < 0$, the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

2. Omitted *prbconv*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + u$$

$$prbconv = \alpha_0 + \alpha_1 prbarr + u$$

```r
(omit2_pri = lm(log(crmrte) ~ prbarr + prbconv, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + prbconv, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      prbconv
##     -2.2458      -2.6519      -0.9676
```

```r
(omit2_sec = lm(prbconv ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = prbconv ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##      0.5022      -0.1877
```

Since $\beta_2 = -0.9807$ and $\alpha_1 = -0.1921$, then $OMVB = \beta_2\alpha_1 = 0.1884$. Since $\beta_1 = -2.647 < 0$, the OLS coefficient on *prbarr* will be scaled toward zero (less negative) losing statistical significance.

3. Omitted *pctymle*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 pctymle + u$$

$$pctymle = \alpha_0 + \alpha_1 prbarr + u$$

```r
(omit3_pri = lm(log(crmrte) ~ prbarr + pctymle, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + pctymle, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      pctymle
##      -3.106       -2.295        3.811
```

```r
(omit3_sec = lm(pctymle ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = pctymle ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##     0.09828     -0.04594
```

Since $\beta_2 = 3.870$ and $\alpha_1 = -0.04568$, then $OMVB = \beta_2\alpha_1 = -0.1768$. Since $\beta_1 = -3.119 < 0$, the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

4. Omitted *density*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 density + u$$

$$density = \alpha_0 + \alpha_1 prbarr + u$$

```r
(omit4_pri = lm(log(crmrte) ~ prbarr + density, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + density, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr      density
##     -3.2608      -1.5302       0.1646
```

```r
(omit4_sec = lm(density ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = density ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##       3.214       -5.711
```

Since $\beta_2 = 0.1657$ and $\alpha_1 = -5.682$, then $OMVB = \beta_2\alpha_1 = -0.9415$. Since $\beta_1 = -1.5169 < 0$, the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

5. Omitted *mix*

$$crmrte = \beta_0 + \beta_1 prbarr + \beta_2 mix + u$$

$$mix = \alpha_0 + \alpha_1 prbarr + u$$

```r
(omit5_pri = lm(log(crmrte) ~ prbarr + mix, data = crime_data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ prbarr + mix, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr          mix
##    -2.73194     -2.47444      0.01041
```

```r
(omit5_sec = lm(mix ~ prbarr, data = crime_data))
```

```
##
## Call:
## lm(formula = mix ~ prbarr, data = crime_data)
##
## Coefficients:
## (Intercept)       prbarr
##     0.01929      0.39319
```

Since $\beta_2 = 0.02237$ and $\alpha_1 = 0.3936$, then $OMVB = \beta_2\alpha_1 = 0.0088$. Since $\beta_1 = -2.4674 < 0$, the OLS coefficient on *prbarr* will be scaled toward zero (less negative) losing statistical significance.

6. Other omitted variables

While our dataset has 25 variables, we noticed that more socioeconomic and infrastructure variables could improve our model. Examples of extra variables that could be added, and the theories we could test with them are:

- Education degree of population (better skilled residents may commit less crime).
- Average number of years of residents (transient population may commit more crimes).
- Umemployment numbers (people that are not working tends to recur to crime).
- Weather (harsh weather may reduce incentives for crime).
- Some way to measure the cultural acceptance to small crimes (crime rate scales from the minor misdemeanors, as New York crime reduction in the 90's suggests).

Conclusion

Based on the analysis and comparison on several models, the determinants of crime are essentially probability of arrest, tax revenue per capita, and % young male. In order to anticipate reduction of crime, the actionable policy suggestions would be as follow for local government:

- Increase the probability of arrest when offense occurs. This doesn't necessarily mean increasing the number of police officers on the street as seen in our analysis. This change could be addressed with programs that incentivizes crime reporting practices and population confidence in the law enforcement. Our best model suggests that an improvement of 1 percentual point in arresting people that committed crimes may improve crime rate of 2%.
- Decrease the tax revenue per capita by reducing local tax rate. Less tax means more money in counties' economy, so it may be a way to improve earnings by the population and decrease criminality. The effect, however may not be really big, is that reducing 1 percentual point in the tax revenue per capita may reduce crime rate by approximately 0.13%. In other words, this variable may have a high statistical significance but not a practical one.
- Decrease the percentage of young male population in communities. While this can turn into a highly unethical advice, we can try to address this matter with making other areas attractive to young male population using government fostered jobs, for example.