Lab3: Reducing Crime

w203 Lab3

Harith Elrufaie and Gaurav Desai

Introduction

The goal of this report is to analyze and study dataset of crime statistics for a selection of counties in North Carolina to help a political campaign to understand the determinants of crime. The outcome of this study is policy suggestions that are applicable to local government.

Exploratory Data Analysis

First, we'll conduct an Exploratory Data Analysis of the given dataset. This process will help us gain a solid understanding of our variables, which will eventually be essential to choose the right combinations for our ultimate model.

Setup

First, we load the necessary libraries.

```
suppressMessages(library(stargazer))
suppressMessages(library(corrplot))
suppressMessages(library(dplyr))
suppressMessages(library(car))
```

```
## Warning: package 'car' was built under R version 3.4.4
```

```
suppressMessages(library(dplyr))
```

Then we load the dataset, which is in the same directory as this RMD.

```
# Load the data
rawCrimeData = read.csv("crime_v2.csv")
str(rawCrimeData)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
```

```
## $ wcon    : num  281 255 227 375 292 ...
## $ wtuc    : num  409 376 372 398 377 ...
## $ wtrd    : num  221 196 229 191 207 ...
## $ wfir    : num  453 259 306 281 289 ...
## $ wser    : num  274 192 210 257 215 ...
## $ wmfg    : num  335 300 238 282 291 ...
## $ wfed    : num  478 410 359 412 377 ...
## $ wsta    : num  292 363 332 328 367 ...
## $ wloc    : num  312 301 281 299 343 ...
## $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
## $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

The dataset contains **25** variables and **97** records. Now, we'll need to do few rounds of data cleanup before analyzing any further.

Data Quality/Clean-up

Convert county to factor

Since county is not a measurement, we should convert it into factor so that it can be easily used for further analysis.

```
rawCrimeData$county <- as.factor(rawCrimeData$county)
length(levels(rawCrimeData$county))
```

```
## [1] 90
```

Interestingly we have 91 rows but only 90 levels. Eyeballing the data shows there are two identical rows for county 193, same can be verified using duplicated function.

```
rawCrimeData[duplicated(rawCrimeData),"county"]
```

```
## [1] 193  <NA> <NA> <NA> <NA>
## 90 Levels: 1 3 5 7 9 11 13 15 17 19 21 23 25 27 33 35 37 39 41 45 ... 197
```

```
#so lets delete the duplicate row
#before..
nrow(rawCrimeData)
```

```
## [1] 97
```

```
rawCrimeData <- rawCrimeData[!duplicated(rawCrimeData),]
#after..
nrow(rawCrimeData)
```

```
## [1] 92
```

Convert prbconv to number

Now lets convert prbconv from factor to number because it is a probability value.

```
rawCrimeData$prbconv <- as.numeric(levels(rawCrimeData$prbconv))[rawCrimeData$prbconv]
```

```
## Warning: NAs introduced by coercion
```

```
summary(rawCrimeData$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.06838 0.34422 0.45170 0.55086 0.58513 2.12121       2
```

Remove NAs

```r
#let us find how many records we have first..
sum(is.na(rawCrimeData$crmrte))
```

```
## [1] 2
```

```r
sum(is.na(rawCrimeData$county))
```

```
## [1] 2
```

```r
#The data set containsrows with NA county and NA crime rate.
rawCrimeData <- rawCrimeData[!is.na(rawCrimeData$crmrte),]
crimeData <- rawCrimeData[!is.na(rawCrimeData$county),]
min(complete.cases(crimeData))
```

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

Ceil probabilities to 1 to fit them in valid probability range

```r
#Now lets see if any of the probability is crossing 0 to 1 range
filter(crimeData, prbarr< 0 | prbarr>1 |
         prbconv < 0 | prbconv > 1 |
         prbpris < 0 | prbpris > 1) [,c("county", "prbarr", "prbconv", "prbpris")]
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
##    county   prbarr prbconv  prbpris
## 1       3 0.132029 1.48148 0.450000
## 2      19 0.162860 1.22561 0.333333
## 3      99 0.153846 1.23438 0.556962
## 4     115 1.090910 1.50000 0.500000
## 5     127 0.179616 1.35814 0.335616
## 6     137 0.207143 1.06897 0.322581
## 7     149 0.271967 1.01538 0.227273
## 8     185 0.195266 2.12121 0.442857
## 9     195 0.201397 1.67052 0.470588
## 10    197 0.207595 1.18293 0.360825
```

We have 10 counties where prbconv is greater than 1, which means there are more convictions than arrests. Out of these 10 counties, one county (115) also has prbarr greater than 1 indicating more arrests than offences.

We have two ways to clean this data, either we remove these 10 counties or we cap the max probabilities at 1. For this analysis we take second approach of capping the probabilities at 1.

```r
crimeData$prbconv <- ifelse(crimeData$prbconv>1,1,crimeData$prbconv)
crimeData$prbarr <- ifelse(crimeData$prbarr>1,1,crimeData$prbarr)
summary(crimeData$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.09277 0.20495 0.27146 0.29423 0.34487 1.00000
```

```r
summary(crimeData$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.06838 0.34422 0.45170 0.50799 0.58513 1.00000
```

```r
summary(crimeData$prbpris)
```

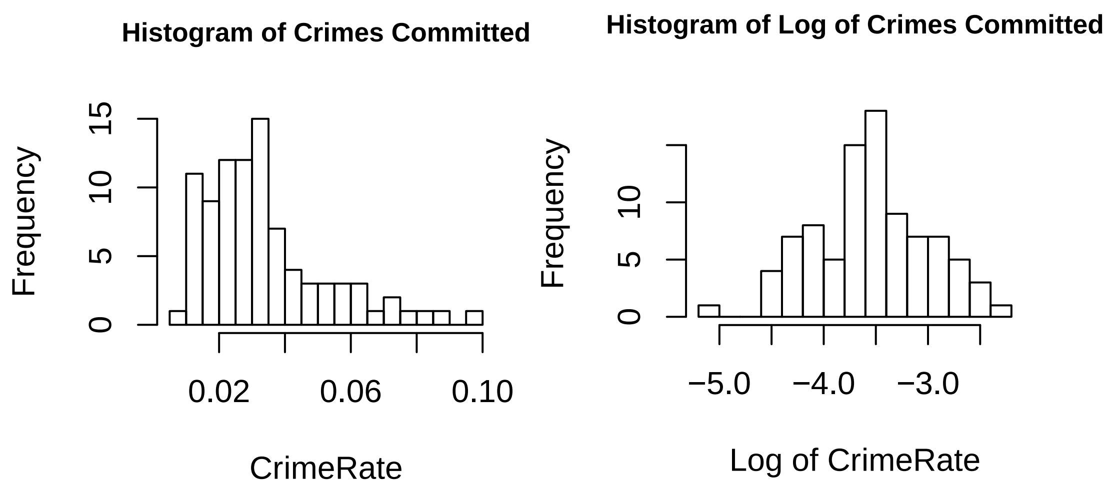
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1500  0.3642  0.4222  0.4106  0.4576  0.6000
```

Univariate Variable Analysis of Key Variables

crmrte: crimes committed per person

Looking at the histogram of crime per person, the distribution appears to be positively skewed. Applying `log()` on crime shows the histogram to appear normally distributed. The linear regressions would benefit from this transformation.

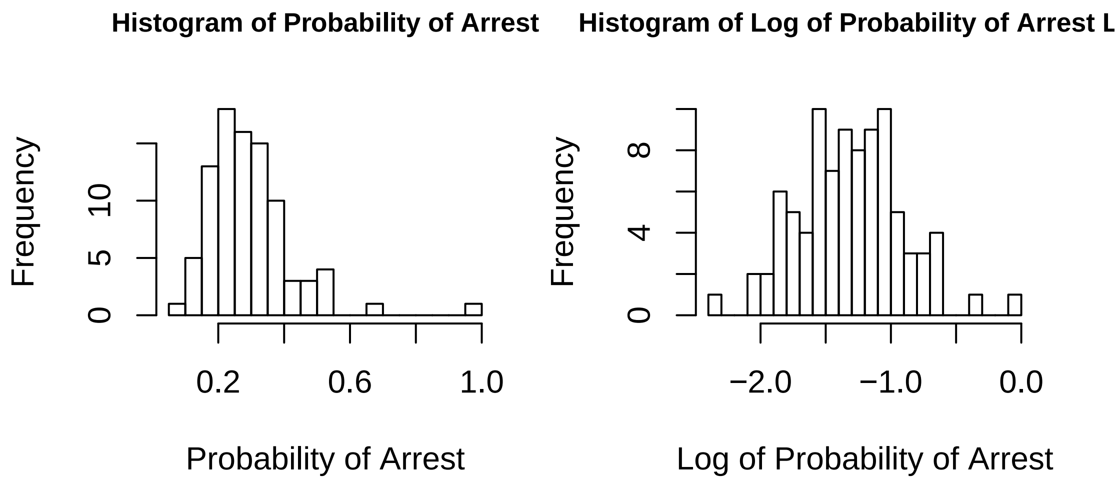
```r
hist(crimeData$crmrte, breaks=20, main = "Histogram of Crimes Committed"
     , cex.main=0.8, xlab="CrimeRate")
hist(log(crimeData$crmrte), breaks=20, main = "Histogram of Log of Crimes Committed"
     , cex.main=0.8, xlab="Log of CrimeRate")
```



prbarr: probability of arrest

Looking at the histogram of arrest per person, the distribution appears to be positively skewed. Applying `log()` shows the histogram to appear normally distributed. The linear regressions would benefit from this transformation.

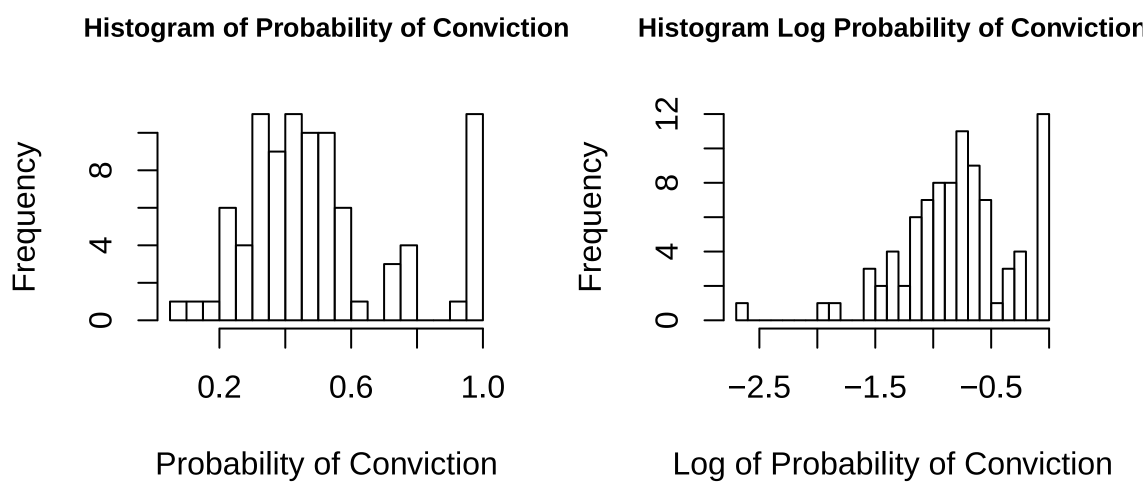
```r
hist(crimeData$prbarr, breaks=20, main = "Histogram of Probability of Arrest"
     , cex.main=0.8, xlab="Probability of Arrest")
hist(log(crimeData$prbarr), breaks=20, main = "Histogram of Log of Probability of Arrest Log"
     , cex.main=0.8, xlab="Log of Probability of Arrest")
```



prbconv: probability of conviction

Looking at the histogram of probability of conviction, the distribution appears to be positively skewed. Applying `log()` shows the histogram to appear normally distributed. The linear regressions would benefit from this transformation.

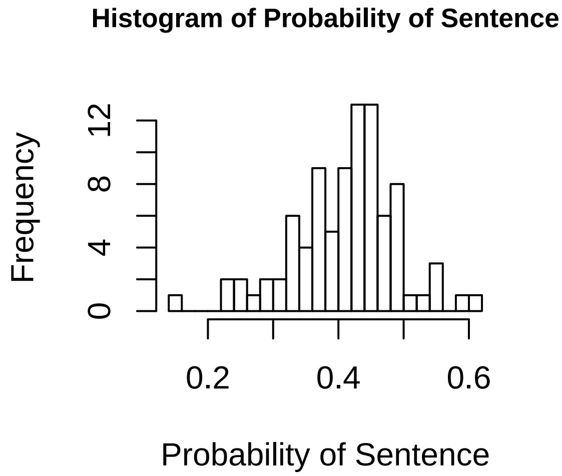
```
hist(crimeData$prbconv, breaks=20, main = "Histogram of Probability of Conviction"
     , cex.main=0.8, xlab="Probability of Conviction")
hist(log(crimeData$prbconv), breaks=20, main = "Histogram Log Probability of Conviction"
     , cex.main=0.8, xlab="Log of Probability of Conviction")
```



prbpris: probability of prison sentence

Looking at the histogram of probability of sentence, the distribution appears to be relatively normal.

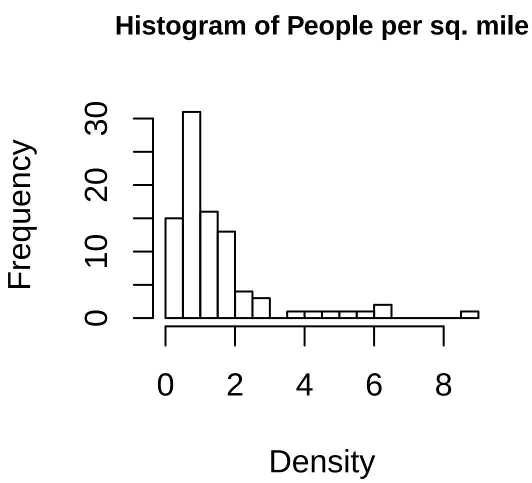
```
hist(crimeData$prbpris, breaks=20, main = "Histogram of Probability of Sentence"
     , cex.main=0.8, xlab="Probability of Sentence")
```

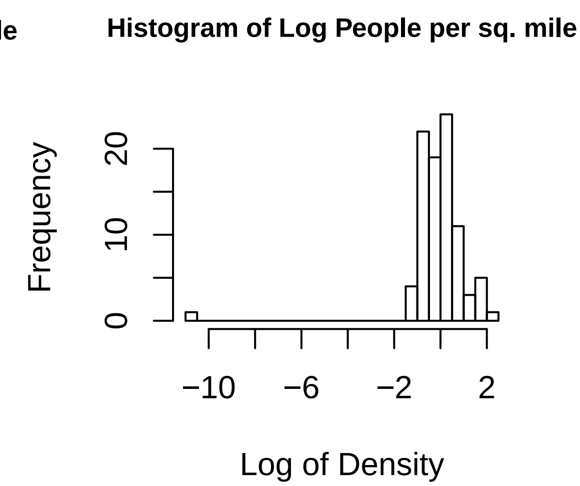


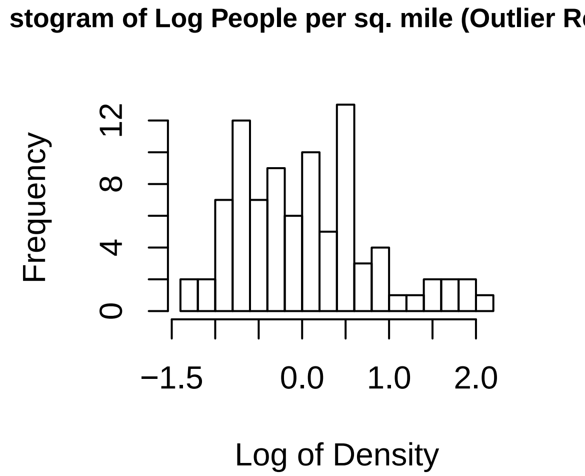
density: people per sq. mile

Looking at the histogram of People per sq mile, the distribution appears to be positively skewed. Applying `log()` shows the histogram to appear *more* normally distributed, with one outlier, which we should remove.

```
hist(crimeData$density, breaks=20, main = "Histogram of People per sq. mile"
     , cex.main=0.8, xlab="Density")
hist(log(crimeData$density), breaks=20, main = "Histogram of Log People per sq. mile"
     , cex.main=0.8, xlab="Log of Density")
#summary(crimeData$density)
crimeData <- filter(crimeData, density > 0.02)
hist(log(crimeData$density), breaks=20, main = "Histogram of Log People per sq. mile (Outlier Removed)"
     , cex.main=0.8, xlab="Log of Density")
```



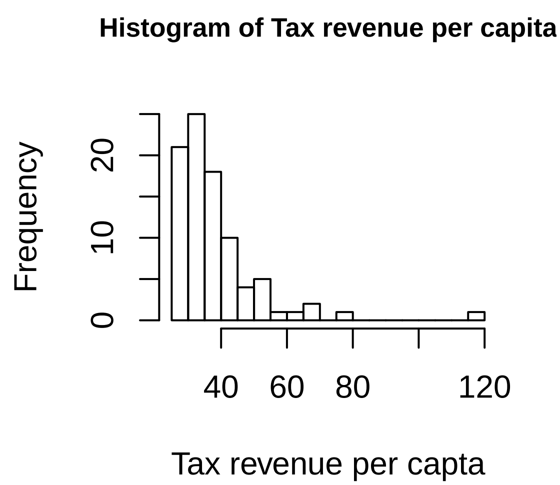


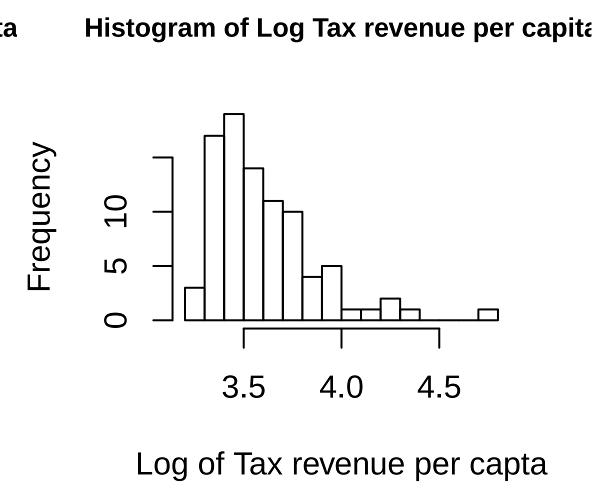


taxpc: tax revenue per capita

Looking at the histogram of probability of sentence, the distribution appears to be positively skewed. Applying `log()` shows the histogram to appear slightly positively skewed. The linear regressions would benefit from this transformation.

```
hist(crimeData$taxpc, breaks=20, main = "Histogram of Tax revenue per capita"
     , cex.main=0.8, xlab="Tax revenue per capta")
hist(log(crimeData$taxpc), breaks=20, main = "Histogram of Log Tax revenue per capita"
     , cex.main=0.8, xlab="Log of Tax revenue per capta")
```

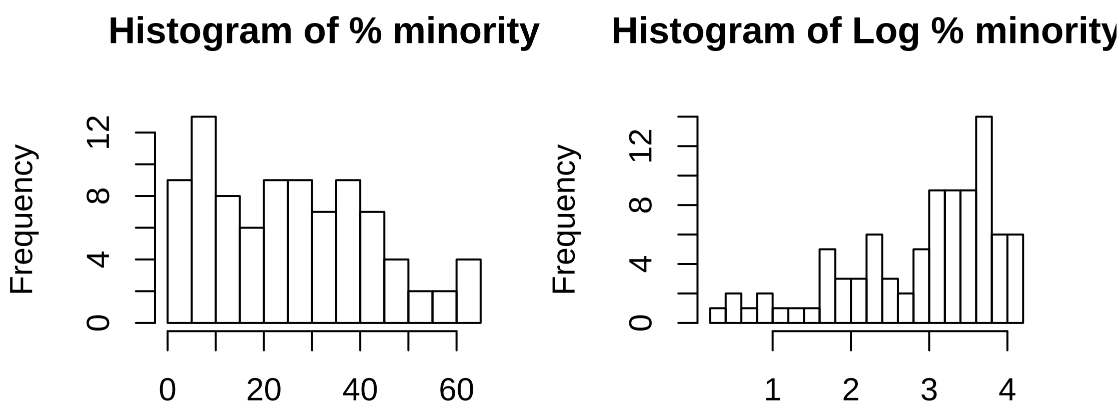




pctmin80: perc. minority, 1980

Looking at the histogram of probability of sentence, the distribution appears to be slightly positively skewed. Applying `log()` shows the histogram to appear negatively skewed. The linear regressions would benefit from this transformation.

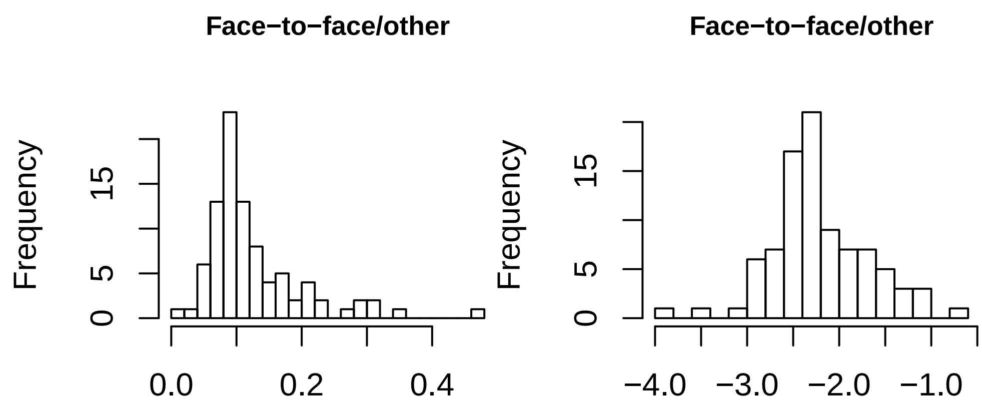
```
hist(crimeData$pctmin80, breaks=20, main = "Histogram of % minority", xlab = "")
hist(log(crimeData$pctmin80), breaks=20, main = "Histogram of Log % minority", xlab = "")
```



mix: offense mix: face-to-face/other

Looking at the histogram, the distribution appears to be slightly positively skewed. Applying `log()` shows the histogram to appear normally distributed. The linear regressions would benefit from this transformation.

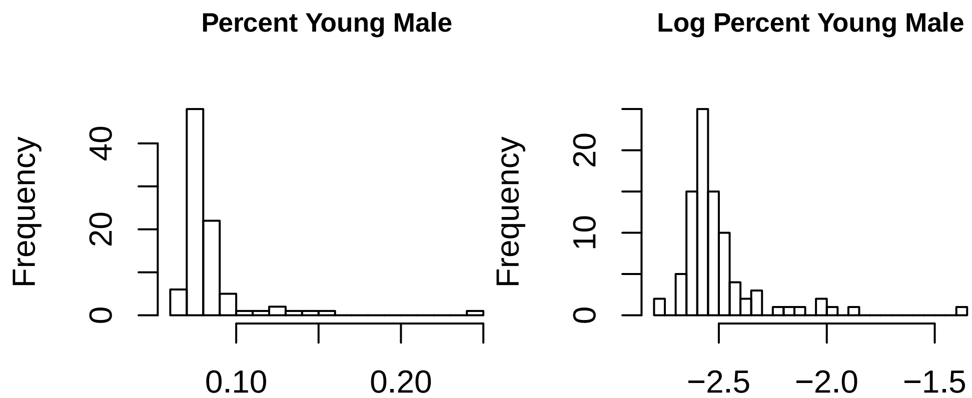
```
hist(crimeData$mix, breaks=20, main = "Face-to-face/other"
     , cex.main=.8, xlab = "")
hist(log(crimeData$mix), breaks=20, main = "Face-to-face/other"
     , cex.main=.8, xlab = "")
```

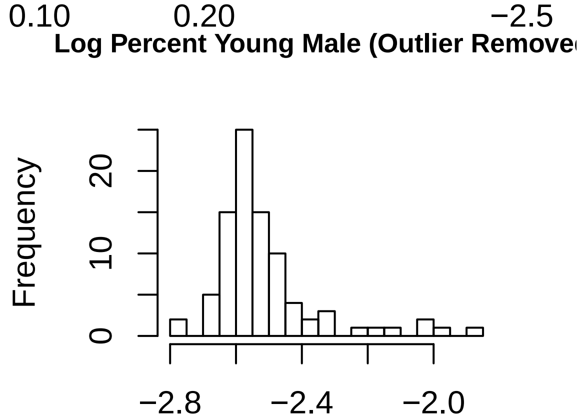


pctymle: percent young male

Looking at the histogram, the distribution appears to be positively skewed. Applying `log()` shows the histogram to appear positively skewed with one outlier that we should remove.

```
hist(crimeData$pctymle, breaks=20, main = "Percent Young Male"
     , cex.main=.8, xlab = "")
hist(log(crimeData$pctymle), breaks=20, main = "Log Percent Young Male"
     , cex.main=.8, xlab = "")
crimeData <- filter(crimeData, pctymle < .20)
hist(log(crimeData$pctymle), breaks=20, main = "Log Percent Young Male (Outlier Removed)"
     , cex.main=.8, xlab = "")
```

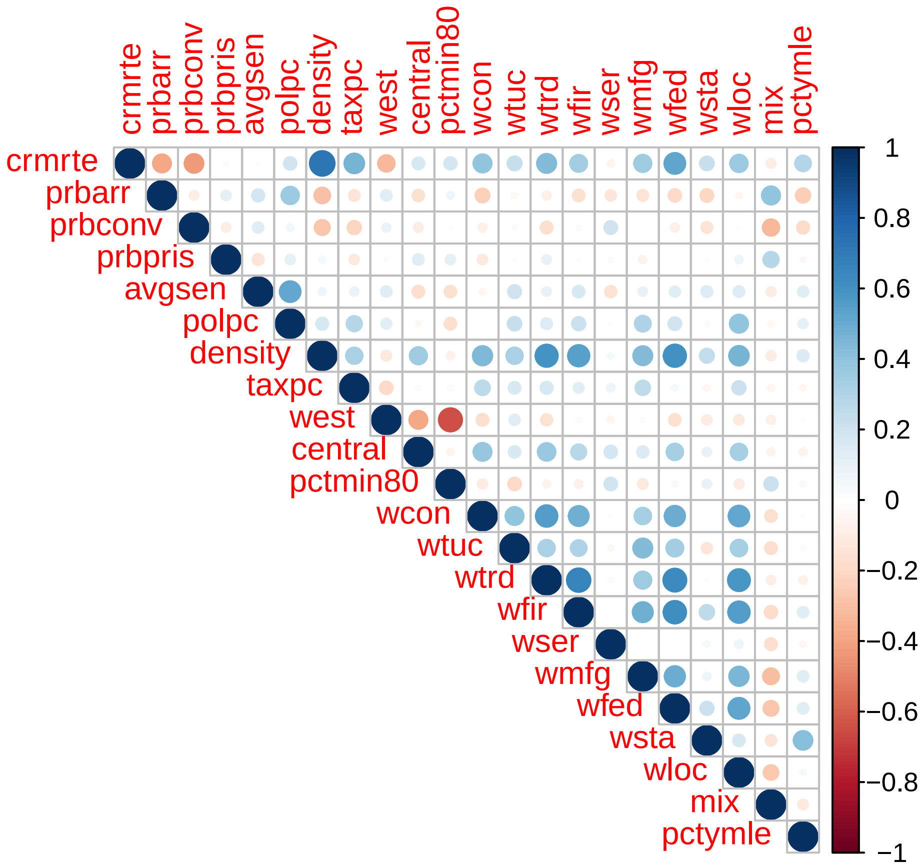




Analysis of Key Relationships

It is very imperative to realize the relationship between crime rate and all the data available to us. We'll use `corrplot` to make the exploration of key relationships clearer.

```
corrplot(cor(crimeData[ , (names(crimeData) %in%
                                    c("crmrte", "prbarr", "prbconv", "prbpris", "avgsen"
                                      , "polpc", "density", "taxpc", "west", "central"
                                      , "uraban", "pctmin80", "wcon", "wtuc", "wtrd"
                                      , "wfir", "wser", "wmfg", "wfed", "wsta", "wloc"
                                      , "mix", "pctymle"))])
         , method="circle", type="upper")
```



The above plot indicates the following *positive* relationships with crime rate:

1. Density (density).
2. Tax revenue per capita (taxpc).
3. All wage variables.
4. Young Male (pctymle)

The above plot also indicates the following *negative* relationships with crime rate:

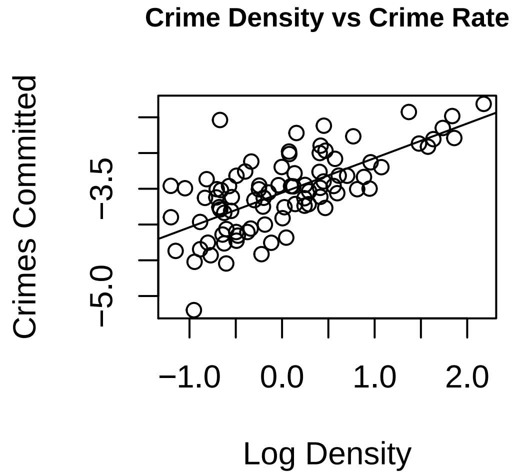
1. Probability of Arrest (prbarr)
2. Probability of Conviction (prbconv)
3. West region of NC (west)

Crimes Committed per person (crmrte) & People per sq. (density)

As you can see from the correlation plot below, there is a positive linear relationship between crime rate and density.

```r
plot(log(crimeData$density), log(crimeData$crmrte),
    main="Crime Density vs Crime Rate",
    xlab="Log Density",
    ylab="Crimes Committed", cex.main=0.8)
abline(lm(log(crimeData$crmrte) ~ log(crimeData$density)))
cor(crimeData$crmrte, crimeData$density)
```

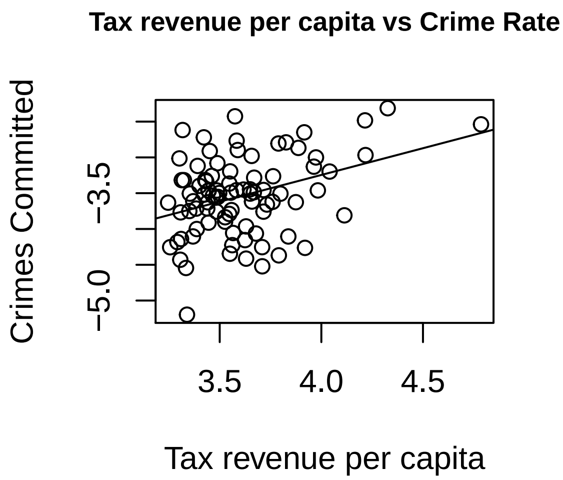
```
## [1] 0.7291395
```



Crimes Committed per person (crmrte) & Tax revenue per capita (taxpc)

```r
plot(log(crimeData$taxpc), log(crimeData$crmrte),
    main="Tax revenue per capita vs Crime Rate",
    xlab="Tax revenue per capita",
    ylab="Crimes Committed", cex.main=0.8)
abline(lm(log(crimeData$crmrte) ~ log(crimeData$taxpc)))
cor(crimeData$crmrte, crimeData$taxpc)
```

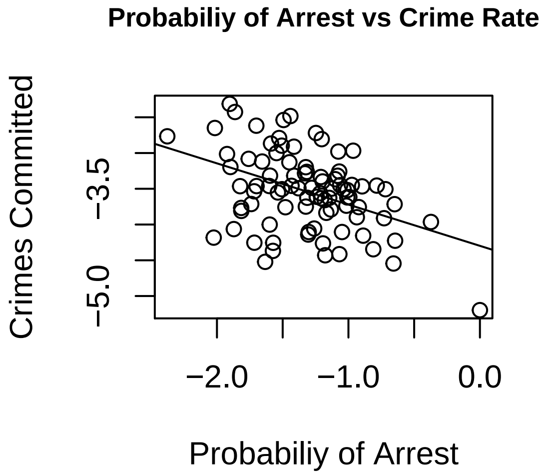
```
## [1] 0.4668575
```



Crimes Committed per person (crmrte) & Probabiliy of Arrest (prbarr)

```r
plot(log(crimeData$prbarr), log(crimeData$crmrte),
    main="Probabiliy of Arrest vs Crime Rate",
    xlab="Probabiliy of Arrest",
    ylab="Crimes Committed", cex.main=0.8)
abline(lm(log(crimeData$crmrte) ~ log(crimeData$prbarr)))
cor(crimeData$crmrte, crimeData$prbarr)
```

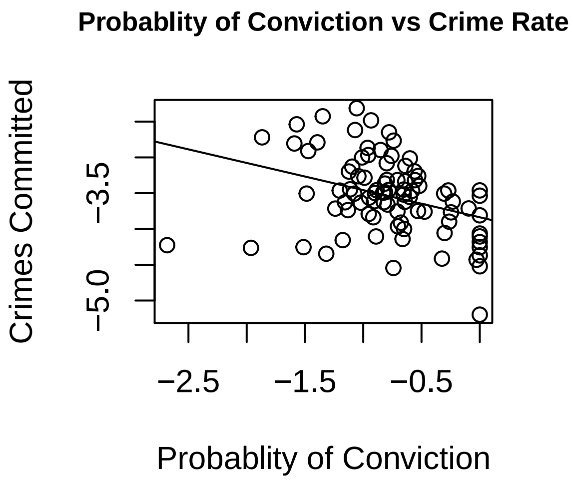
```
## [1] -0.3898229
```



Crimes Committed per person (crmrte) & Tax revenue per capita (prbconv)

```
plot(log(crimeData$prbconv), log(crimeData$crmrte),
    main="Probablity of Conviction vs Crime Rate",
    xlab="Probablity of Conviction",
    ylab="Crimes Committed", cex.main=0.8)
abline(lm(log(crimeData$crmrte) ~ log(crimeData$prbconv)))
cor(crimeData$crmrte, crimeData$prbconv)
```

```
## [1] -0.4216284
```



Proposed Models

Model 1: with only the explanatory variables

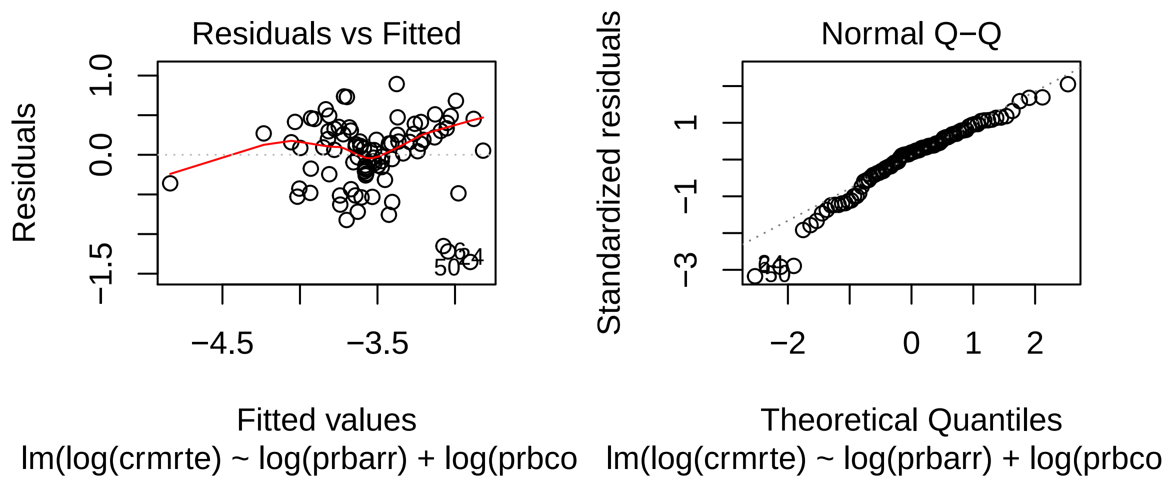
Using a combination of key positive and negative attributes to crime rate, we're recommending the following model:

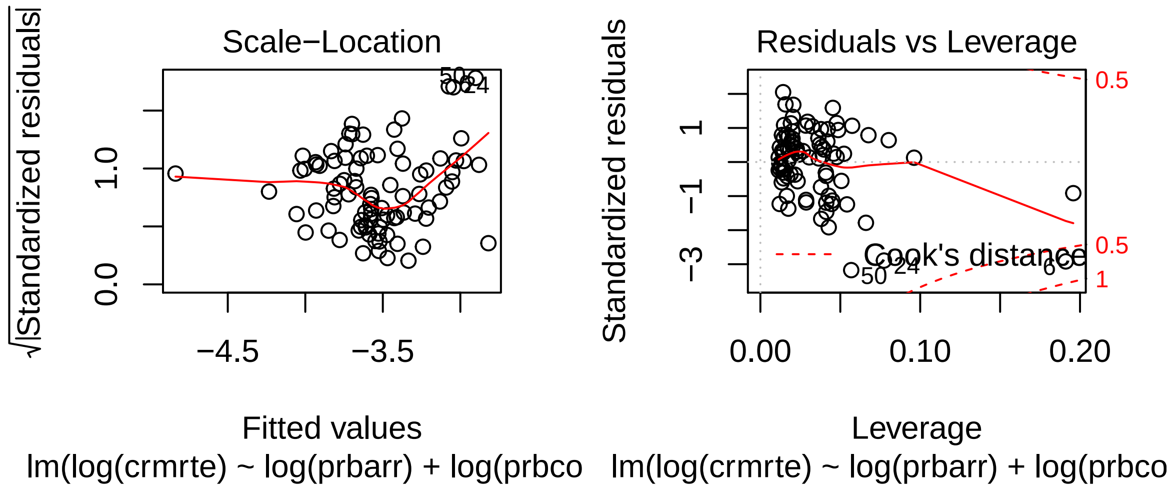
$$crimeDeterm = \beta_0 + \beta.log(prbarr) + \beta_2 \cdot log(prbconv) + \beta_3 \cdot log(pctymle)$$

```
model1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv), data=crimeData)
summary(model1)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv), data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35254 -0.21604  0.08342  0.29790  0.89327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.83710    0.19413 -24.917  < 2e-16 ***
## log(prbarr)  -0.69673    0.12067  -5.774 1.24e-07 ***
## log(prbconv) -0.48949    0.09681  -5.056 2.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4387 on 85 degrees of freedom
## Multiple R-squared:  0.3672, Adjusted R-squared:  0.3523
## F-statistic: 24.66 on 2 and 85 DF,  p-value: 3.581e-09
```

```
plot(model1)
```





Model 2: with key explanatory variables and only covariates

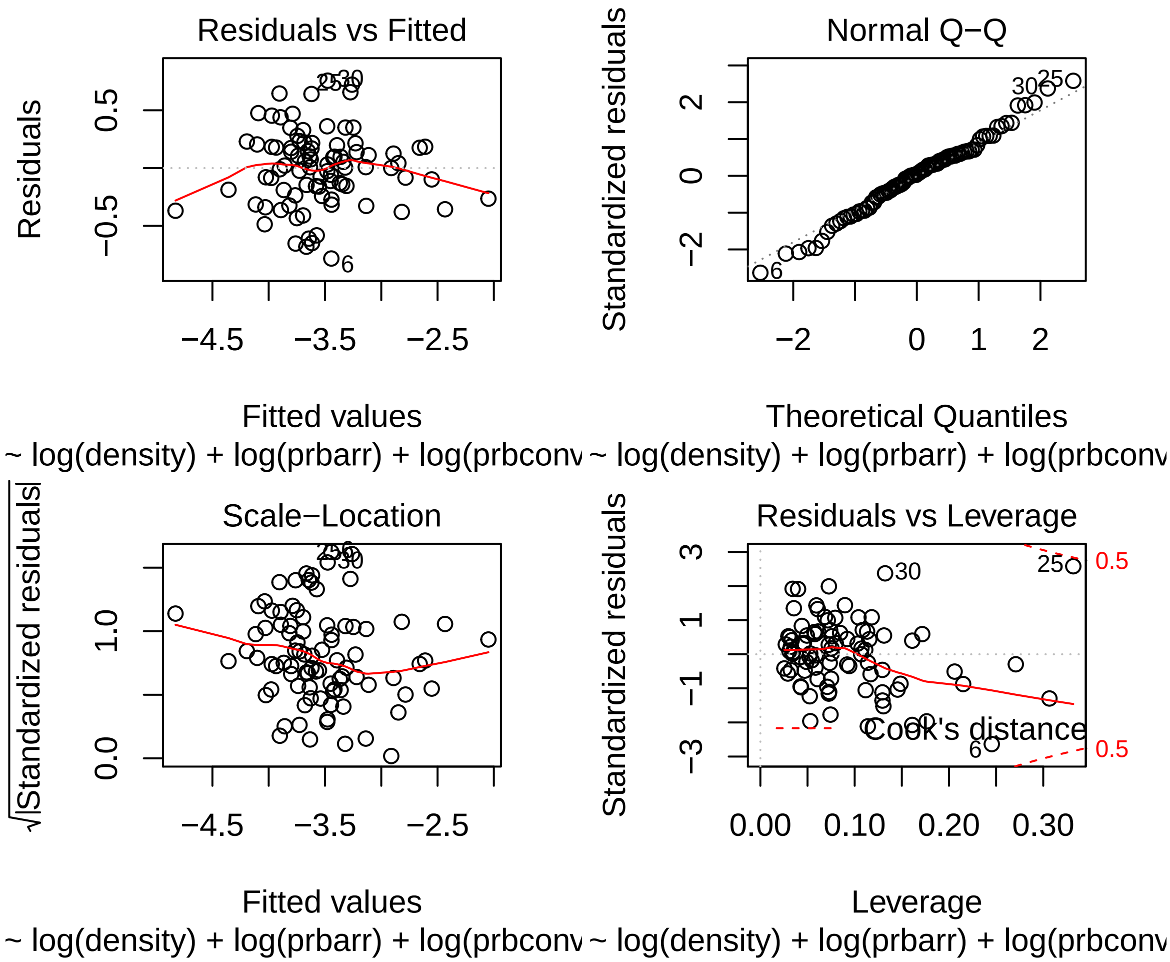
In this model, we'll include the variables (avgsen, mix, taxpc), as we think they will contribute to the accuracy of your results without introducing substantial bias.

$$crimeDeterm = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(prbarr) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(pctymle) + \beta_5 \cdot log(avgsen) + \beta_6 \cot log(mix) + +$$

```
model2 <- lm(log(crmrte) ~ log(density) + log(prbarr) + log(prbconv)
             + log(pctymle) + log(avgsen) + log(mix) + log(taxpc), data=crimeData)
summary(model2)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(density) + log(prbarr) + log(prbconv) +
##     log(pctymle) + log(avgsen) + log(mix) + log(taxpc), data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78273 -0.20167  0.01464  0.18740  0.75630
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.26094    0.92921  -4.586 1.65e-05 ***
## log(density)  0.36485    0.05542   6.583 4.43e-09 ***
## log(prbarr)  -0.45912    0.12231  -3.754 0.000329 ***
## log(prbconv) -0.15926    0.09176  -1.736 0.086476 .
## log(pctymle)  0.24205    0.26358   0.918 0.361199
## log(avgsen)  -0.06984    0.13423  -0.520 0.604280
## log(mix)      0.26106    0.08431   3.097 0.002699 **
## log(taxpc)    0.36793    0.14809   2.485 0.015058 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.342 on 80 degrees of freedom
## Multiple R-squared:  0.6381, Adjusted R-squared:  0.6064
## F-statistic: 20.15 on 7 and 80 DF,  p-value: 2.525e-15
```

```r
plot(model2)
```



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

In this model, we'll include all the data available to us to demonstrate the robustness of results to model specification.

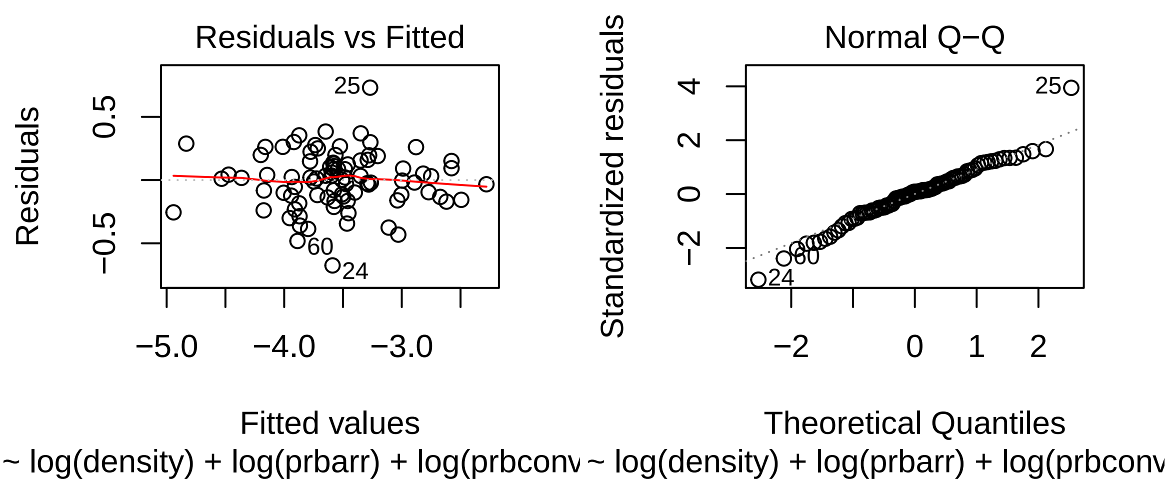
$$crimeDeterm = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(prbarr) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(pctymle) + \beta_5 \cdot log(avgsen) + \beta_6 \cdot log(mix) + \beta_7 \cdot lo$$

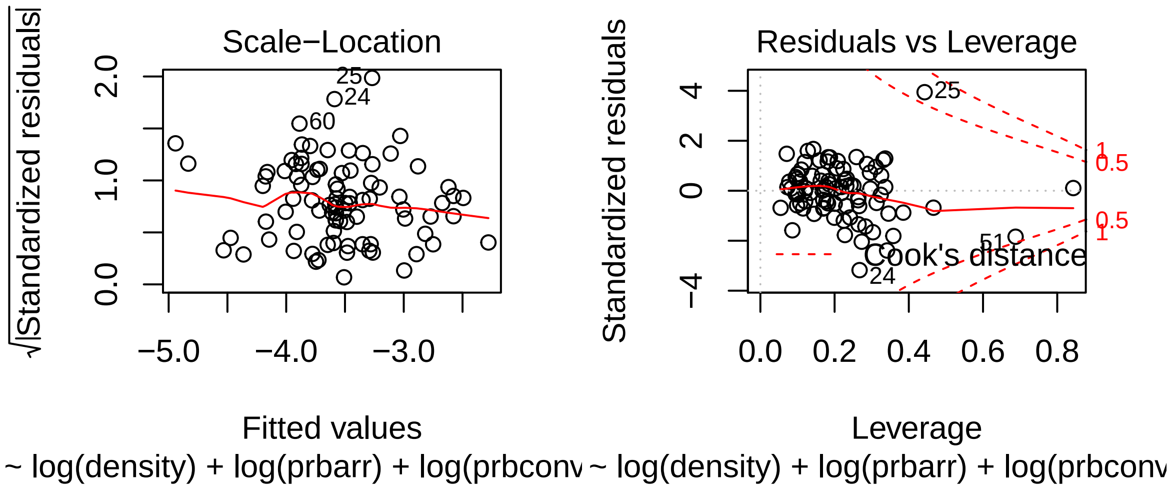
```r
model3 <- lm(log(crmrte) ~ log(density) + log(prbarr) + log(prbconv)
             + log(pctymle) + log(avgsen) + log(mix) + + log(taxpc)
             + prbpris + log(polpc)
             + log(pctmin80) + log(wcon) + log(wtrd) + wfir + log(wser) + log(wmfg)
             + log(wfed) + log(wsta) + wloc, data=crimeData)
summary(model3)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(density) + log(prbarr) + log(prbconv) +
##     log(pctymle) + log(avgsen) + log(mix) + +log(taxpc) + prbpris +
##     log(polpc) + log(pctmin80) + log(wcon) + log(wtrd) + wfir +
```

```
##     log(wser) + log(wmfg) + log(wfed) + log(wsta) + wloc, data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67516 -0.13090  0.01483  0.13978  0.73100
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.2627357  3.2597246  -1.921 0.058832 .
## log(density)   0.2659222  0.0616361   4.314 5.23e-05 ***
## log(prbarr)   -0.4760816  0.0931306  -5.112 2.72e-06 ***
## log(prbconv)  -0.2603837  0.0768946  -3.386 0.001173 **
## log(pctymle)   0.0202421  0.2131946   0.095 0.924632
## log(avgsen)   -0.2287424  0.1193767  -1.916 0.059491 .
## log(mix)       0.0725884  0.0709180   1.024 0.309621
## log(taxpc)     0.1270078  0.1355286   0.937 0.351961
## prbpris       -0.5293938  0.3875821  -1.366 0.176410
## log(polpc)     0.3672425  0.1167592   3.145 0.002448 **
## log(pctmin80)  0.2213139  0.0337383   6.560 8.27e-09 ***
## log(wcon)      0.2520582  0.2232093   1.129 0.262705
## log(wtrd)      0.1779501  0.3172517   0.561 0.576675
## wfir          -0.0014873  0.0007962  -1.868 0.066032 .
## log(wser)     -0.3825086  0.1079785  -3.542 0.000716 ***
## log(wmfg)      0.1057123  0.1559167   0.678 0.500036
## log(wfed)      0.6240522  0.3609127   1.729 0.088264 .
## log(wsta)      0.0219402  0.2712762   0.081 0.935773
## wloc          -0.0006701  0.0014339  -0.467 0.641715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2482 on 69 degrees of freedom
## Multiple R-squared:  0.8356, Adjusted R-squared:  0.7927
## F-statistic: 19.48 on 18 and 69 DF,  p-value: < 2.2e-16
```

```r
plot(model3)
```





All 3 Regression models at a glance

```r
stargazer(model1, model2, model3, type = "text", title="Comparison of 3 Regression models", float=FALSE)
```

```
##
## ==========================================================================================
##                                         Dependent variable:
##                         ------------------------------------------------------------------
##                                            log(crmrte)
##                               (1)                   (2)                     (3)
## ------------------------------------------------------------------------------------------
## log(density)                                     0.365***                0.266***
##                                                  (0.055)                 (0.062)
##
## log(prbarr)               -0.697***             -0.459***               -0.476***
##                            (0.121)               (0.122)                 (0.093)
##
## log(prbconv)              -0.489***             -0.159*                 -0.260***
##                            (0.097)               (0.092)                 (0.077)
##
## log(pctymle)                                     0.242                   0.020
##                                                  (0.264)                 (0.213)
##
## log(avgsen)                                      -0.070                  -0.229*
##                                                  (0.134)                 (0.119)
##
## log(mix)                                         0.261***                0.073
##                                                  (0.084)                 (0.071)
##
## log(taxpc)                                       0.368**                 0.127
##                                                  (0.148)                 (0.136)
##
## prbpris                                                                  -0.529
##                                                                          (0.388)
##
## log(polpc)                                                               0.367***
```

```
##                                                                       (0.117)
##
## log(pctmin80)                                                        0.221***
##                                                                       (0.034)
##
## log(wcon)                                                              0.252
##                                                                       (0.223)
##
## log(wtrd)                                                              0.178
##                                                                       (0.317)
##
## wfir                                                                 -0.001*
##                                                                       (0.001)
##
## log(wser)                                                           -0.383***
##                                                                       (0.108)
##
## log(wmfg)                                                              0.106
##                                                                       (0.156)
##
## log(wfed)                                                             0.624*
##                                                                       (0.361)
##
## log(wsta)                                                              0.022
##                                                                       (0.271)
##
## wloc                                                                  -0.001
##                                                                       (0.001)
##
## Constant                  -4.837***             -4.261***             -6.263*
##                            (0.194)               (0.929)              (3.260)
##
## ---------------------------------------------------------------------------------------
## Observations                 88                    88                    88
## R2                         0.367                 0.638                 0.836
## Adjusted R2                0.352                 0.606                 0.793
## Residual Std. Error    0.439 (df = 85)       0.342 (df = 80)       0.248 (df = 69)
## F Statistic         24.661*** (df = 2; 85) 20.148*** (df = 7; 80) 19.484*** (df = 18; 69)
## =======================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Omitted Variables

We believe that following omitted variables may contribute towards crime rate regression results.

1. Literacy: Higher the literacy, crime rate should go down. In general terms as literacy increases, it is easier for people to find jobs, which deters them from conducting crimes.
2. Poverty: If per capita income is not distributed equally then there is high chance of crimes in that area. Tax per capita tries to proxy this variable but it does not capture the high to low distribution of income. If per capita income has huge variance from mean then crime rate should go up. Different wages provided in the data may act as proxy as they cover most of the wage range except may be farming and other self-employed people.

3. Corruption: Higher the corruption, more the crime rate in the area. More corruption generally disrupts employment and effectively pushes people into criminal activity.
4. Historic criminal rate of the area: If previous generation had high criminal rate in a particular area then new generation would grow in that area and continue following same foot steps. So we should also measure this continuity effect. It is much easier for new people to turn to criminals where there are already plenty of established criminals than areas where crime is low.

Conclusion

Our Regression Model (Model 1) indicates that as population density increases and the young male percentage increases, the crime rate grows. So policymakers need to pay attention to more urbanized or highly dense regions with a high male ratio. Also, steps should be taken to improve gender by diversifying the community, for instance bringing more women and men of different age groups, which potentially can bring down crime rate.

More important aspect is the effect of strong arrest and conviction ratio on the crime rate. Having strong and capable police has a noticeable deterrent effect on crime rate. Therefore, policymakers should concentrate on strengthening the police and judiciary system and deter people from committing crimes by setting strong examples of arrests and convictions.