Lab3: Reducing Crime

w203 Lab3

Harith Elrufaie and Gaurav Desai

Introduction

We have been tasked to help shape up a political campaign in North Carolina. We are equipped with "Crime Statistics" data of year 1987 for selected counties in North Carolina and our task is to decipher this data and understand various factors that could affect the crime rate and make statistics backed suggestions applicable to local government to improve the Crime rate in North Carolina.

Setup

First, we load the necessary libraries.

```
suppressMessages(library(dplyr))
suppressMessages(library(stargazer))
suppressMessages(library(corrplot))
suppressMessages(library(ggplot2))
```

Data Load

```
rawCrimeData = read.csv("crime_v2.csv")
dim(rawCrimeData)
```

```
## [1] 97 25
```

```
summary(rawCrimeData)
```

```
##      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
##  0.068376102: 1   Median :0.4234   Median : 9.100   Median :0.001485
##  0.140350997: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702
##  0.154451996: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877
##  0.203724995: 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
```

```r
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",..: 62 88 12 61 51 2 58 77 41 85 ...
##  $ 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** observations. Now lets see if there are any bad data that needs to be cleaned up.

Data Quality/Clean-up

Convert county to factor

Since county is not a measurement, it won't make sense to roll it up for aggregation or do any mathematical operation (like taking average) on it. Hence lets convert it into factor.

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

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

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

```
## [1] 6
```

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

```
rawCrimeData[duplicated(rawCrimeData[!is.na(rawCrimeData$county),]), c("county","crmrte")]
```

```
##    county    crmrte
## 89    193 0.0235277
```

```
#so lets delete the duplicate row
rawCrimeData <- rawCrimeData[!duplicated(rawCrimeData[!is.na(rawCrimeData$county),]),]
nrow(rawCrimeData) #after removal of duplicate we are left with 96 observations..
```

```
## [1] 96
```

Convert prbconv to number

Now lets convert prbconv from factor to number because it is a ratio of convictions to arrest so it is actual measurement and should be stored as number for aggregations and other mathematical operations.

```
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       6
```

Remove NAs

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

```
## [1] 6
```

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

```
## [1] 6
```

The data set contains 6 NA rows, lets remove them

```
crimeData <- rawCrimeData[!is.na(rawCrimeData$county),]
min(complete.cases(crimeData))
```

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

EDA

Now, 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 right variable combinations for our regression model.

Univariate Analysis

crmrte: crimes committed per person

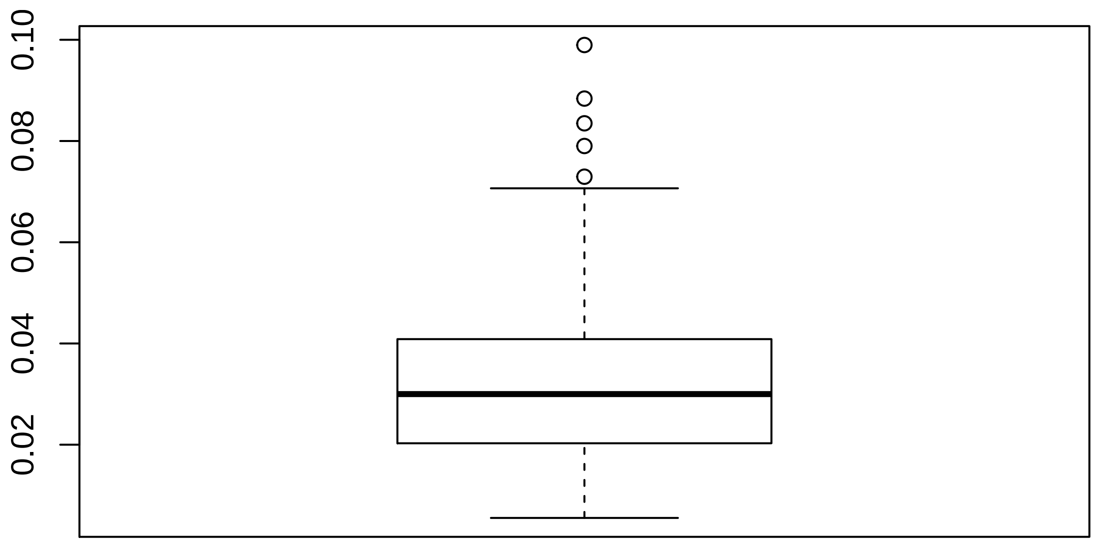
This is outcome variable for our regresison model where we will try and derive relation between various independent variables and crime rate. Looking at the quantiles of crmrte we can see large difference between 3rd quantile and max. So there are few outliers counties with very high crime rates than rest. Same is evident from histogram. To take care of outliers and fit the variable into normal distribution, lets take a log of crime rate. But we note that these are crimes rates per person and all the values are between 0 and 1. This range is not suitable for logarithms. So lets change the scale by creating new variable for crime rate per 1000 people (crmrtepk) and then lets take log(log_crmrtepk). The new variable is log_crmrtepk which shows nice normal distribution. Going forward whenever we talk about crime rate, we will use log_crmrtepk (log of crmrt per k)

Also we not the righ most outlier, county=119 has crime rate of 98 for every 1000 people, that is 1 crime per every 10 people which is very high. Population Density also is highest among all counties. More information is required to understand what is so different about this county so that appropriate remedial action can be suggested.

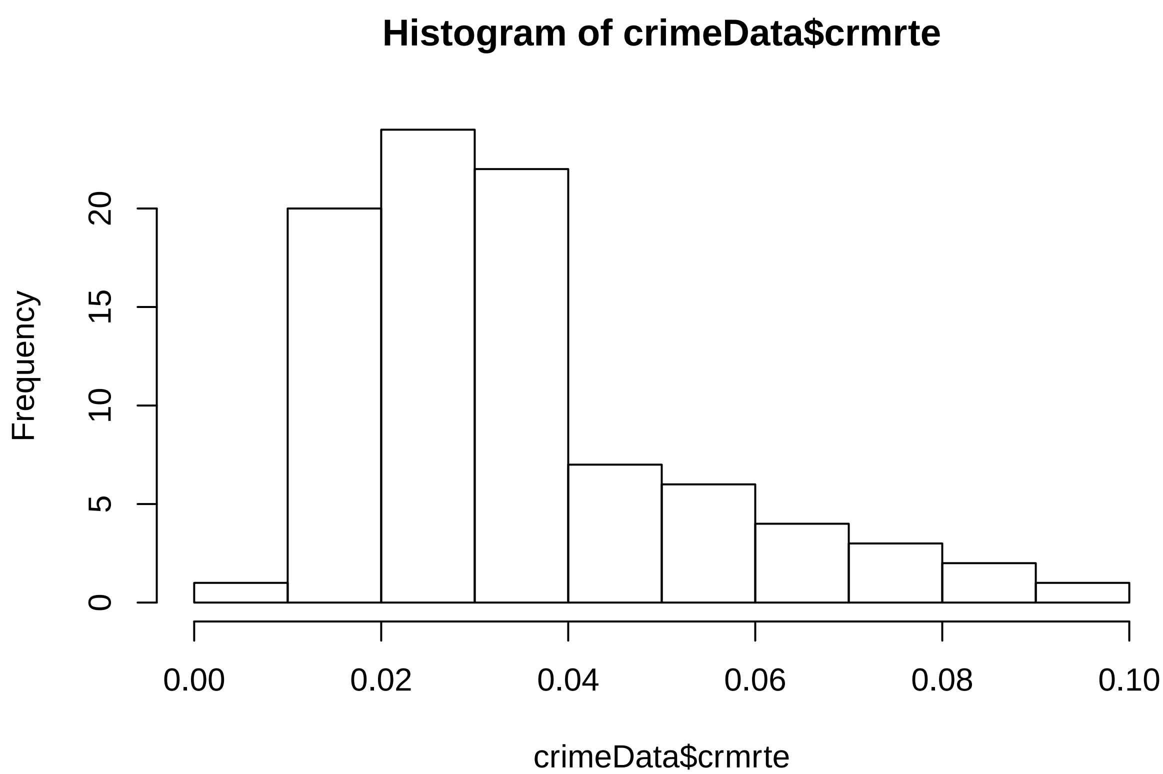
```
summary(crimeData$crmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966
```

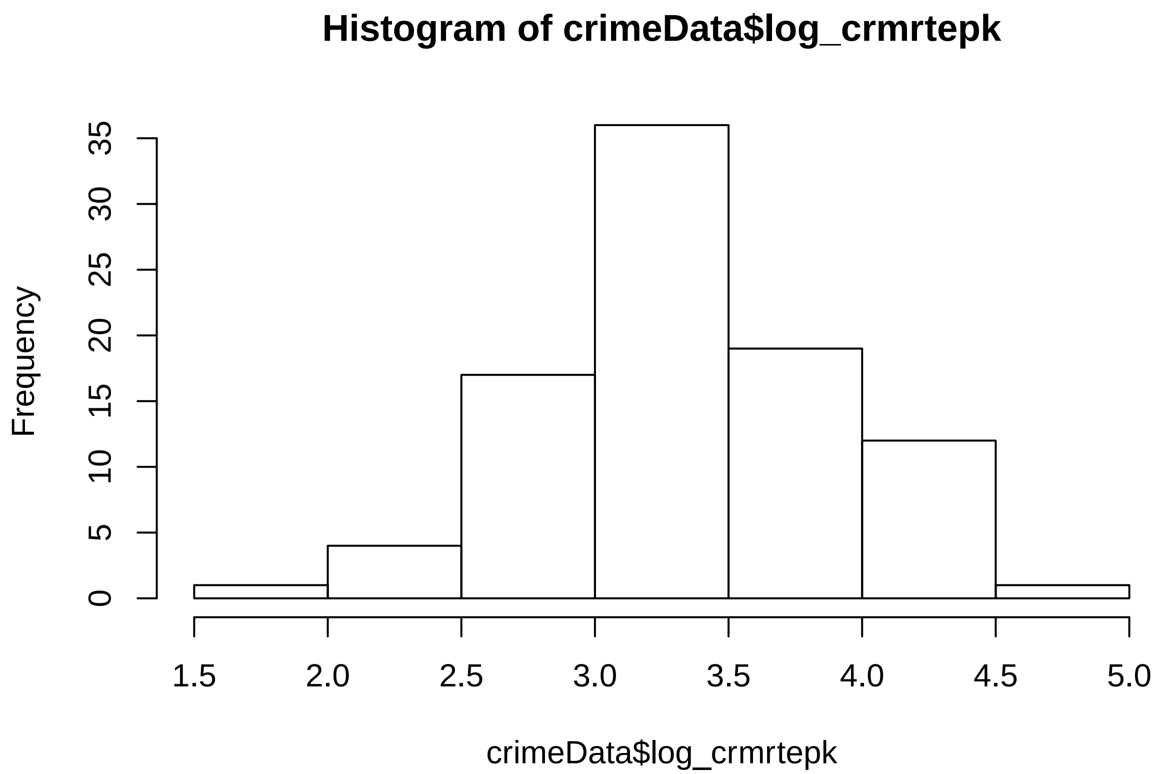
```
boxplot(crimeData$crmrte)
```




```
hist(crimeData$crmrte)
```




```
crimeData$crmrtepk <- crimeData$crmrte * 1000
crimeData$log_crmrtepk <- log(crimeData$crmrtepk)
hist(crimeData$log_crmrtepk)
```




```
crimeData[crimeData$crmrtepk>90,c("county","crmrtepk", "density")]
```

```
##    county crmrtepk  density
## 53    119  98.9659 8.827652
```

Also convert polpc from per capita to per 1000 people to keep the scale

Since we have converted crimerate from per capita to per K people, lets also convert other per capita variable polpc to same scale. While scaling we notice that for county 115 the police per 1000 people is highest at 9 while average is just 1.7. Noteably the second highest police per 1k is 4.5. Crime rate and density in this county is not high, but prbarr is highest at 1.09 and avgsen is highest at 20.7. Which means County 115 has highest police numbers which would logically translate into highest arrests. Though higher police numbers can not logcally explain highest average sentence in that county. We sould need more information about this county, may be there is a central jail for all of western counties of North Carolina which would explain highest police population and highest average sentences.

```
crimeData$polpk <- crimeData$polpc * 1000
summary(crimeData$polpk)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7459  1.2378  1.4897  1.7080  1.8856  9.0543
```

```
crimeData[crimeData$polpk>4,]
```

```
##    county year    crmrte    prbarr  prbconv  prbpris avgsen      polpc
## 25     55   87 0.0790163 0.224628 0.207831 0.304348  13.57 0.00400962
## 51    115   87 0.0055332 1.090910 1.500000 0.500000  20.70 0.00905433
## 90    195   87 0.0313973 0.201397 1.670520 0.470588  13.02 0.00445923
##      density     taxpc west central urban pctmin80     wcon     wtuc
## 25 0.5115089 119.76145    0       0     0  6.49622 309.5238 445.2762
```

```
## 51 0.3858093  28.19310    1       0     0  1.28365 204.2206 503.2351
## 90 1.7459893  53.66693    0       0     0 37.43110 315.1641 377.9356
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 25 189.7436 284.5933 221.3903 319.21 338.91 361.68 326.08 0.08437271
## 51 217.4908 342.4658 245.2061 448.42 442.20 340.39 386.12 0.10000000
## 90 246.0614 411.4330 296.8684 392.27 480.79 303.11 337.28 0.15612382
##       pctymle crmrtepk log_crmrtepk   polpk
## 25 0.07613807  79.0163     4.369654 4.00962
## 51 0.07253495   5.5332     1.710766 9.05433
## 90 0.07945071  31.3973     3.446722 4.45923
```

```
crimeData[crimeData$avgsen>15,]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 19     41   87 0.0257713 0.307246 0.45283 0.520833  17.41 0.00149399
## 51    115   87 0.0055332 1.090910 1.50000 0.500000  20.70 0.00905433
## 56    127   87 0.0291496 0.179616 1.35814 0.335616  15.99 0.00158289
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 19 0.7417582 41.76929    0       0     0 42.64210 256.4102 379.0005
## 51 0.3858093 28.19310    1       0     0  1.28365 204.2206 503.2351
## 56 1.3388889 32.02376    0       0     0 34.27990 290.9091 426.3901
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 19 238.5589 271.7391 232.5916 332.07 451.84 389.99 312.05 0.09872611
## 51 217.4908 342.4658 245.2061 448.42 442.20 340.39 386.12 0.10000000
## 56 257.6008 441.1413 305.7612 329.87 508.61 380.30 329.71 0.06305506
##       pctymle crmrtepk log_crmrtepk   polpk
## 19 0.06355526  25.7713     3.249261 1.49399
## 51 0.07253495   5.5332     1.710766 9.05433
## 56 0.07400288  29.1496     3.372441 1.58289
```

Check if there are any abnormal probabilities

```
#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")]
```

```
##    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. Infact there is one county=185 which has more than 2 convictions per arrest. Out of these 10 counties, one county (115) also has prbarr greater than 1 indicating more arrests than offences. We have talked about this county in detail while analysing polpc variable earlier.

Under normal curcomstances pribabilities should not cross 0 to 1 range, but in this case the probabilitis are mere proxies to actual police and judicioury data. One of the possible explanation to more convictions than arrest could be transfers of arrested people from base location to newar court locations within or outside of North Carolina. In absense of more details on these probabilities we keep the probabilities above 1 as it is and proceed further with our analysis

```
data.probabilities <- cbind(crimeData$prbarr,crimeData$prbconv,crimeData$prbpris,deparse.level = 2)
colnames(data.probabilities) <- c("prbarr", "prbconv", "prbpris")
summary(data.probabilities)
```

```
##      prbarr            prbconv            prbpris
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.1500
##  1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.:0.3642
##  Median :0.27146   Median :0.45170   Median :0.4222
##  Mean   :0.29524   Mean   :0.55086   Mean   :0.4106
##  3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:0.4576
##  Max.   :1.09091   Max.   :2.12121   Max.   :0.6000
```

Now lets look look in detail at outliers in these probabilities. Outlier in prbarr is county 115 which has been already discussed in earlier section for polpc. Lets look at outlier in prbconv which is county 185

```
crimeData[crimeData$prbconv>2,]
```

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
##      density    taxpc west central urban pctmin80     wcon    wtuc
## 84 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 84 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944
##       pctymle crmrtepk log_crmrtepk  polpk
## 84 0.07008217  10.8703     2.386034 1.2221
```

```
summary(crimeData)
```

```
##      county         year        crmrte             prbarr
##  1      : 1   Min.   :87   Min.   :0.005533   Min.   :0.09277
##  3      : 1   1st Qu.:87   1st Qu.:0.020604   1st Qu.:0.20495
##  5      : 1   Median :87   Median :0.030002   Median :0.27146
##  7      : 1   Mean   :87   Mean   :0.033510   Mean   :0.29524
##  9      : 1   3rd Qu.:87   3rd Qu.:0.040249   3rd Qu.:0.34487
##  11     : 1   Max.   :87   Max.   :0.098966   Max.   :1.09091
##  (Other):84
##     prbconv            prbpris            avgsen           polpc
##  Min.   :0.06838   Min.   :0.1500   Min.   : 5.380   Min.   :0.0007459
##  1st Qu.:0.34422   1st Qu.:0.3642   1st Qu.: 7.375   1st Qu.:0.0012378
##  Median :0.45170   Median :0.4222   Median : 9.110   Median :0.0014897
##  Mean   :0.55086   Mean   :0.4106   Mean   : 9.689   Mean   :0.0017080
##  3rd Qu.:0.58513   3rd Qu.:0.4576   3rd Qu.:11.465   3rd Qu.:0.0018856
##  Max.   :2.12121   Max.   :0.6000   Max.   :20.700   Max.   :0.0090543
##
##     density            taxpc             west           central
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.54718   1st Qu.: 30.73   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.97925   Median : 34.92   Median :0.0000   Median :0.0000
##  Mean   :1.43567   Mean   : 38.16   Mean   :0.2444   Mean   :0.3778
##  3rd Qu.:1.56926   3rd Qu.: 41.01   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##
```

```
##      urban           pctmin80          wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.:10.024   1st Qu.:250.8   1st Qu.:374.3
##  Median :0.00000   Median :24.852   Median :281.2   Median :404.8
##  Mean   :0.08889   Mean   :25.713   Mean   :285.4   Mean   :410.9
##  3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0   3rd Qu.:440.7
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##
##       wtrd            wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.7   1st Qu.:285.6   1st Qu.: 229.3   1st Qu.:288.6
##  Median :203.0   Median :317.1   Median : 253.1   Median :321.1
##  Mean   :210.9   Mean   :321.6   Mean   : 275.3   Mean   :336.0
##  3rd Qu.:224.3   3rd Qu.:342.6   3rd Qu.: 277.6   3rd Qu.:359.9
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##
##       wfed            wsta            wloc            mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:398.8   1st Qu.:329.3   1st Qu.:297.2   1st Qu.:0.08060
##  Median :448.9   Median :358.4   Median :307.6   Median :0.10095
##  Mean   :442.6   Mean   :357.7   Mean   :312.3   Mean   :0.12905
##  3rd Qu.:478.3   3rd Qu.:383.2   3rd Qu.:328.8   3rd Qu.:0.15206
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##
##     pctymle           crmrtepk       log_crmrtepk       polpk
##  Min.   :0.06216   Min.   : 5.533   Min.   :1.711   Min.   :0.7459
##  1st Qu.:0.07437   1st Qu.:20.604   1st Qu.:3.025   1st Qu.:1.2378
##  Median :0.07770   Median :30.002   Median :3.401   Median :1.4897
##  Mean   :0.08403   Mean   :33.510   Mean   :3.366   Mean   :1.7080
##  3rd Qu.:0.08352   3rd Qu.:40.249   3rd Qu.:3.695   3rd Qu.:1.8856
##  Max.   :0.24871   Max.   :98.966   Max.   :4.595   Max.   :9.0543
##
```

We observe an interesting combination of extremes for County 185. It has highest Arrest to Conviction ratio of 2.1. At the same time least average sentense of 5.4 days. It has highest % of minority as of 1980 at 64%. And very high weekly wage in service industry at 2177. It is difficult to conclude by such extremes without knowing more about that county. But a best guess would be there are more convictions for small petit crimes for which there are no arrest, may be just community service or warnings. Hence conviction ration is very high while average sentence is lowest.

avgsen : Average sentence (in days)

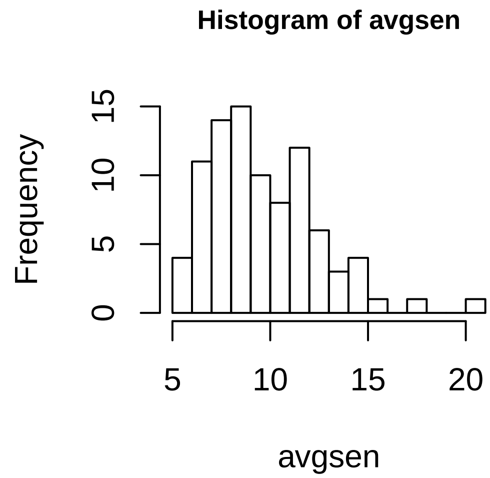
avgsen shows normal distribution with couple of outliers on right. Out of top 3 counties with average sentence, we have already analysed county 115 while analysing polpc. The other two counties 41 and 127 have very high % of minority (42% and 34% respectively). It is difficult to draw conclusion as to why higher average sentence in these areas without any spike in crime rate. Concerned authorities should imvestigate this further.

```
summary(crimeData$avgsen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.380   7.375   9.110   9.689  11.465  20.700
```

```
hist(crimeData$avgsen, breaks=20, main = "Histogram of avgsen"
     , cex.main=0.8, xlab="avgsen")
crimeData[crimeData$avgsen>15,c("county","avgsen","pctmin80", "crmrtepk")]
```

```
##    county avgsen pctmin80 crmrtepk
## 19     41  17.41 42.64210  25.7713
## 51    115  20.70  1.28365   5.5332
## 56    127  15.99 34.27990  29.1496
```



density: people per sq. mile

Density distribution is skewed with high concentration between .5 to 1.5 people per sq. mile. But there are ouliers at both end. Lets look at them.

```
summary(crimeData$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.54718 0.97925 1.43567 1.56926 8.82765
```

```
crimeData[crimeData$density<.3 | crimeData$density>7,]
```

```
##    county year    crmrte    prbarr   prbconv  prbpris avgsen      polpc
## 53    119   87 0.0989659 0.149094 0.347800 0.486183   7.13 0.00223135
## 79    173   87 0.0139937 0.530435 0.327869 0.150000   6.64 0.00316379
##         density    taxpc west central urban pctmin80     wcon     wtuc
## 53 8.8276519780 75.67243    0       1     1  28.5460 436.7666 548.3239
## 79 0.0000203422 37.72702    1       0     0  25.3914 231.6960 213.6752
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 53 354.6761 509.4655 354.3007 494.30 568.40 329.22 379.77 0.1686990
## 79 175.1604 267.0940 204.3792 193.01 334.44 414.68 304.32 0.4197531
##       pctymle crmrtepk log_crmrtepk   polpk
## 53 0.07916495  98.9659     4.594775 2.23135
## 79 0.07462687  13.9937     2.638607 3.16379
```

We have already talked about county 119 having highest density 8.8 people per square mile. Whereas county 173 has very low density of 0.00002 with highest mix of 0.42 i.e. it has highest % of face o face crimes. The population density is so low that mix could be at its peak even by chance. The population density is unrealistically low hence we replace it with mean of density from rest of the counties

```
crimeData[crimeData$density<.3,]$density <- mean(crimeData[crimeData$density>.3,]$density)
```

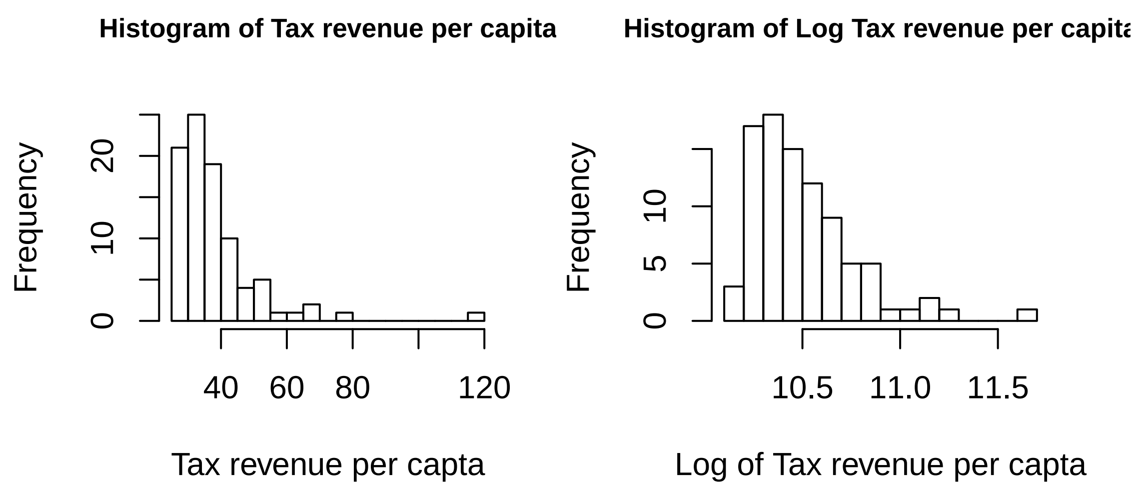
taxpc: tax revenue per capita

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. We will also scale this to per 1000 people to bring in line with crime rate. The linear regressions would benefit from this transformation. The one outlier with 119 taxpc does not show any other extreme value not does it show any super high wages. So this county looks to be wealthy county in general.

```
crimeData[crimeData$taxpc>100,]
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 25     55   87 0.0790163 0.224628 0.207831 0.304348  13.57 0.00400962
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 25 0.5115089 119.7615    0       0     0  6.49622 309.5238 445.2762
##         wtrd      wfir     wser   wmfg    wfed    wsta    wloc        mix
## 25 189.7436 284.5933 221.3903 319.21 338.91 361.68 326.08 0.08437271
##       pctymle crmrtepk log_crmrtepk   polpk
## 25 0.07613807  79.0163     4.369654 4.00962
```

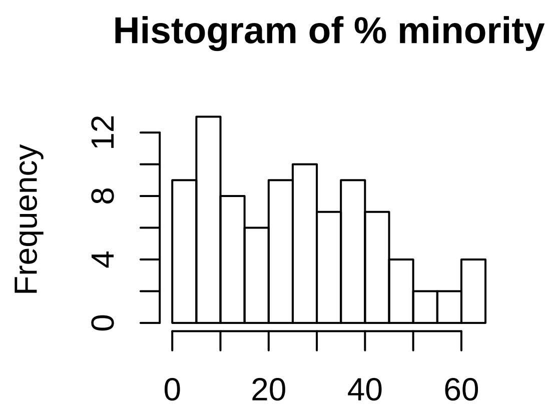
```
hist(crimeData$taxpc, breaks=20, main = "Histogram of Tax revenue per capita"
     , cex.main=0.8, xlab="Tax revenue per capta")
hist(log(crimeData$taxpc*1000), breaks=20, main = "Histogram of Log Tax revenue per capita"
     , cex.main=0.8, xlab="Log of Tax revenue per capta")
crimeData$taxpcpk <- log(crimeData$taxpc*1000)
```



pctmin80: perc. minority, 1980

Looking at the histogram of % of minority as of 1980, it is equally distributed. There are no suprises or any outliers that interests us.

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



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

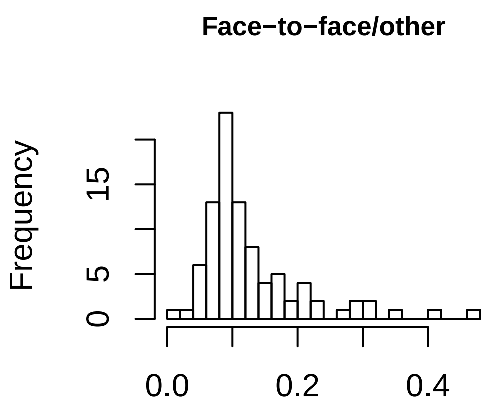
Looking at the histogram, the distribution appears to be slightly positively skewed with few outliers. But otherwise this is fairly normallly distributed. Looking at the top 2 counties for mix are located in the western region. Difficult to draw any conclusion based on this but something for authorities to look into.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01961 0.08060 0.10095 0.12905 0.15206 0.46512
```

```
crimeData[crimeData$mix>.4,c("county", "west", "central", "urban", "mix")]
```

```
##    county west central urban       mix
## 3       5    1       0     0 0.4651163
## 79    173    1       0     0 0.4197531
```



pctymle: percent young male

Looking at the histogram, the distribution appears to be positively skewed with a long tail and one distant outlier. 24% young male population might indicate a large manufacturing industry or some sort of labour intesive work setup in this county though manufacturing or any other wage does nto support this deduction. In absense of any other evidence we will keep this outlier without any modification.

```
summary(crimeData$pctymle)
```

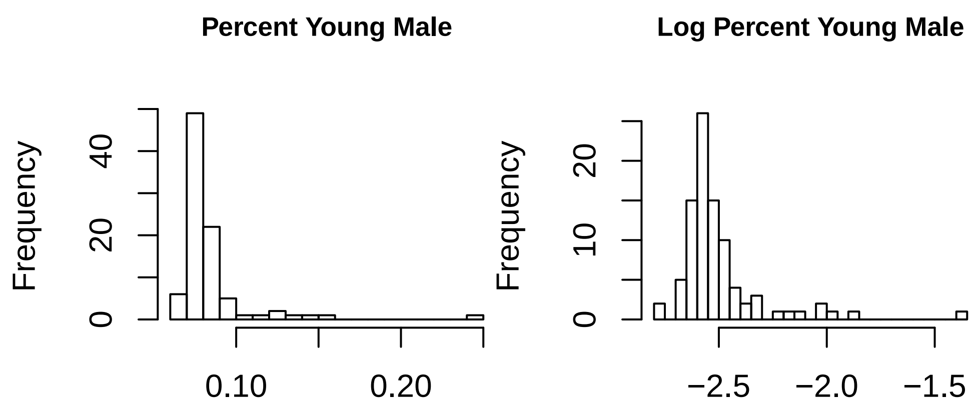
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## 0.06216 0.07437 0.07770 0.08403 0.08352 0.24871
```

```
crimeData[crimeData$pctymle>.2,]
```

```
##    county year    crmrte  prbarr  prbconv  prbpris avgsen      polpc
## 59    133   87 0.0551287 0.26696 0.271947 0.334951   8.99 0.00154457
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 59 1.650066 27.46926    0       0     0  26.3814 264.0406 318.9644
##         wtrd     wfir     wser   wmfg  wfed    wsta    wloc       mix
## 59 183.2609 265.1232 230.6581 258.25 326.1 329.43 301.64 0.1217632
##      pctymle crmrtepk log_crmrtepk    polpk  taxpcpk
## 59 0.2487116  55.1287      4.00967 1.54457 10.22082
```

```
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 = "")
```



wages

Now lets look at all wages together. We will also calculate average wage across all wage categories. Overall all wages look well distributed with some spikes in each othe wages.Total wage is almost perfectly normally distributed. The red line represents average for each of the category. Inrerestingly retal has least of the wages and fed has the higes wage.

```
crimeData$wtotal<-crimeData$wcon+crimeData$wtuc+crimeData$wtrd
                          +crimeData$wfir+crimeData$wser+crimeData$wmfg
```

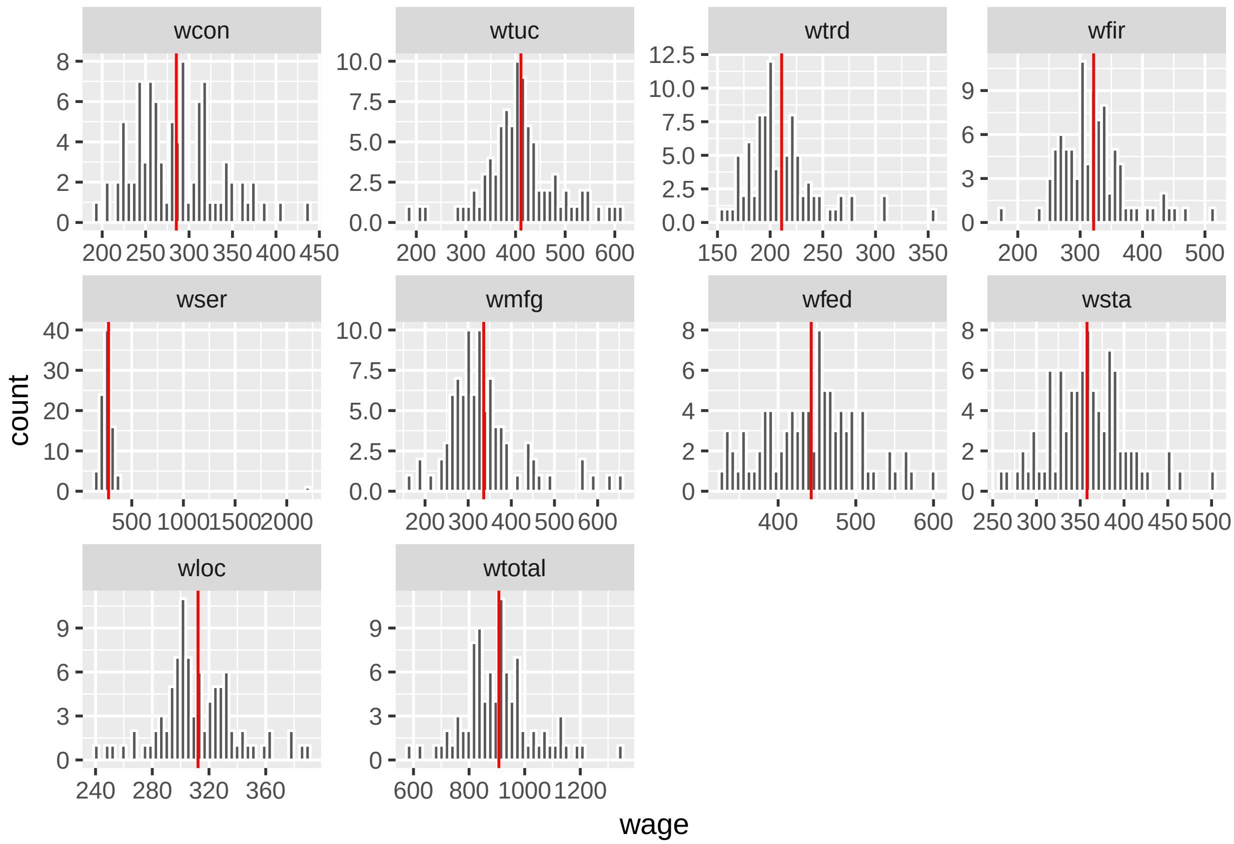
```
##  [1] 1061.8897  751.2527  753.2913  819.5865  795.3958  754.3157  872.9328
##  [8]  797.4570  816.3483 1123.6675 1061.0306  967.5013 1032.2563  918.5653
## [15]  865.1721  997.4808  858.9161  812.6532  836.4007  962.3981  964.8537
## [22]  992.3221 1023.8902  780.2143  825.1936  924.5911  950.5421  797.5878
## [29] 1473.2688  846.4644  784.2533 1374.4425 1002.2055  871.0111  802.7482
## [36] 1174.9006  896.7371  823.6719 1120.9221 1036.2078  835.9969  735.3245
## [43]  968.2335  808.1925  864.4711  950.8972  960.8112  885.9290  839.7436
## [50]  812.6404 1036.0919  838.7150 1358.0662  792.0945  897.5513 1076.7725
## [57] 1120.4807  759.0204  754.0313 1070.8275  614.7363  827.4683  715.4416
## [64]  524.9746  856.4821 1042.3844  853.0897  894.8974  805.8287  879.7144
## [71]  966.8900  969.3879  865.5895  812.7347  936.3476  930.9279  844.6708
```

```
## [78]  881.8954  664.4832 1202.2328  962.7986  829.2993 1222.7951 2689.2112
## [85]  762.3415  927.3813  853.7903  956.5847 1100.5714  883.8838
```

```
                          +crimeData$wfed+crimeData$wsta+crimeData$wloc
```

```
##  [1] 1081.58 1074.26  971.88 1039.45 1087.40  992.41 1084.32 1084.24
##  [9]  989.96  960.92 1212.79 1133.20 1254.88 1044.73 1033.14 1219.51
## [17] 1175.44  949.56 1153.88 1136.54 1054.11 1062.47 1134.99 1052.15
## [25] 1026.67 1135.23 1124.50 1056.35 1341.86 1221.07 1133.21 1285.17
## [33] 1208.12 1211.39 1061.58 1323.83 1136.48 1161.94 1129.29 1110.00
## [41] 1042.73 1164.18 1162.24 1121.22 1127.29 1074.02 1153.84 1145.88
## [49] 1043.47  957.85 1168.71 1046.57 1277.39  998.86 1109.88 1218.62
## [57] 1288.93 1063.40  957.17 1342.59  966.47 1086.11 1050.05 1023.06
## [65] 1008.41 1338.62 1065.49 1067.64 1119.57 1127.68 1138.68 1189.71
## [73] 1030.71 1047.21 1109.02 1087.69 1061.67 1118.36 1053.44 1000.08
## [81] 1222.93 1057.21 1414.02 1048.71 1014.93 1176.82 1037.62 1154.88
## [89] 1121.18 1084.22
```

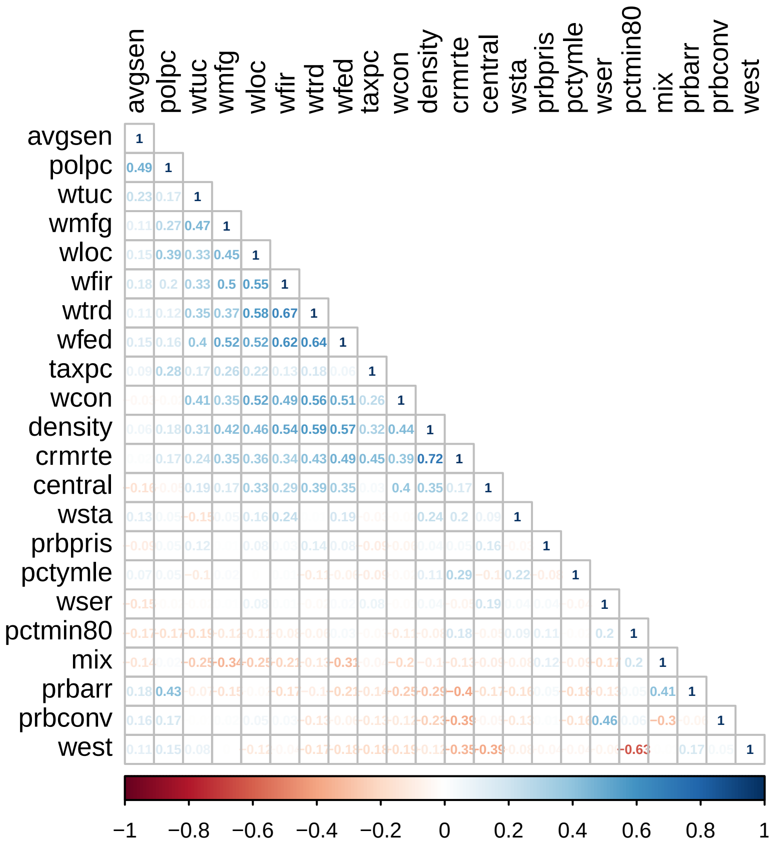
```
wages <- rbind(data.frame(wageType="wcon", wage=crimeData$wcon, meanWage=mean(crimeData$wcon)),
               data.frame(wageType="wtuc", wage=crimeData$wtuc, meanWage=mean(crimeData$wtuc)),
               data.frame(wageType="wtrd", wage=crimeData$wtrd, meanWage=mean(crimeData$wtrd)),
               data.frame(wageType="wfir", wage=crimeData$wfir, meanWage=mean(crimeData$wfir)),
               data.frame(wageType="wser", wage=crimeData$wser, meanWage=mean(crimeData$wser)),
               data.frame(wageType="wmfg", wage=crimeData$wmfg, meanWage=mean(crimeData$wmfg)),
               data.frame(wageType="wfed", wage=crimeData$wfed, meanWage=mean(crimeData$wfed)),
               data.frame(wageType="wsta", wage=crimeData$wsta, meanWage=mean(crimeData$wsta)),
               data.frame(wageType="wloc", wage=crimeData$wloc, meanWage=mean(crimeData$wloc)),
               data.frame(wageType="wtotal", wage=crimeData$wtotal, meanWage=mean(crimeData$wtotal)))

ggplot(wages, aes(x=wage)) + geom_histogram(bins=40, color="white") +
  facet_wrap(~wageType, scales="free") + geom_vline(aes(xintercept=meanWage), color="red")
```



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.

```
cex.before <- par("cex")
par(cex=.8)
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"))])
                   ,tl.pos = "lt", tl.col="black", order="AOE",  number.cex=.5, type="lower"
            , method="number", number.digits=2
                   )
```




```r
par(cex = cex.before)
```

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

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

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

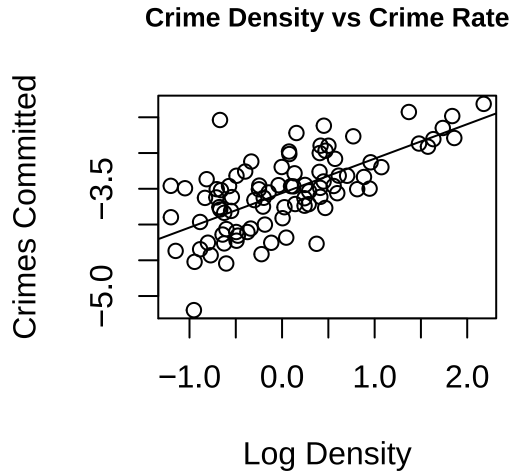
1. Density (density).
2. Tax revenue per capita (taxpc).
3. All wage variables.
4. Young Male (pctymle)

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)
```

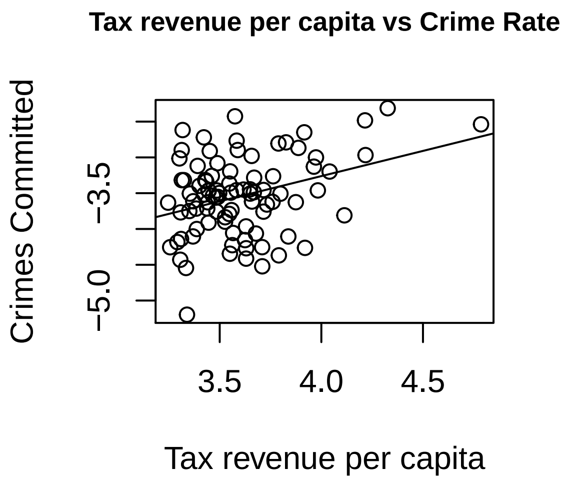
```
## [1] 0.7209483
```



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

```
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)
```

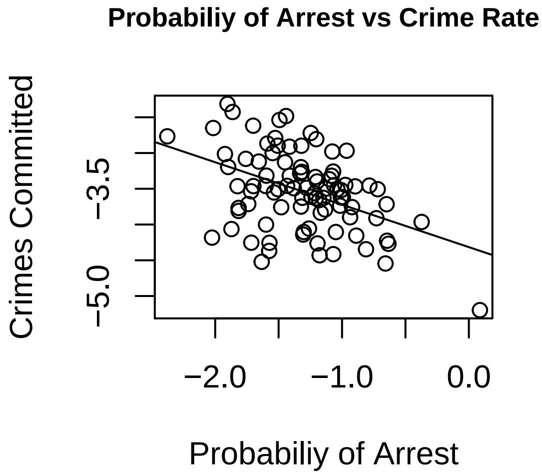
```
## [1] 0.4487151
```



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

```
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)
```

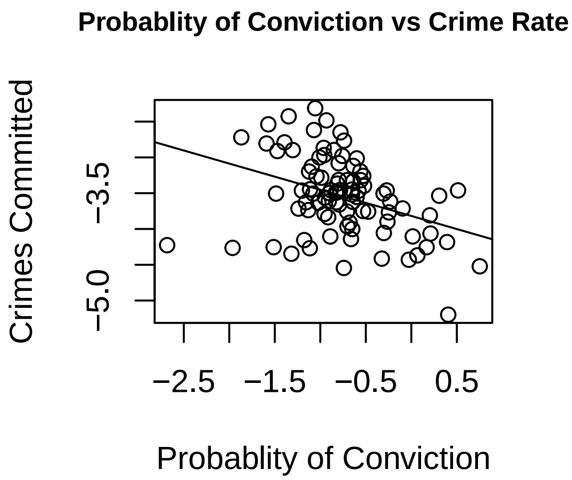
```
## [1] -0.395283
```



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

```r
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] -0.3859656
```



Proposed Models

Model 1: with only the explanatory variables

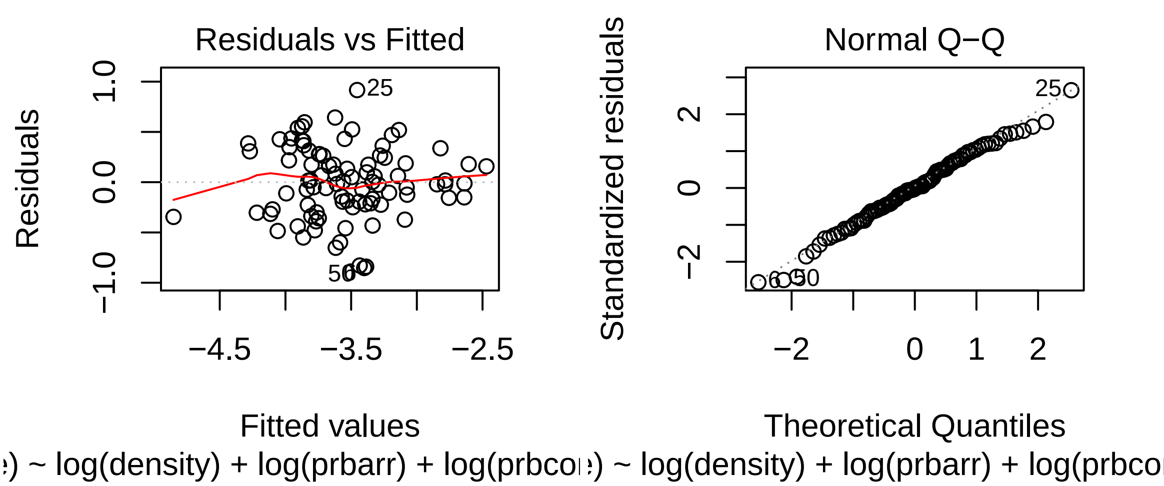
Using a combination of key positive (prbarr, prbconv) and negative attributes (density) to crime rate, we're recommending the following model:

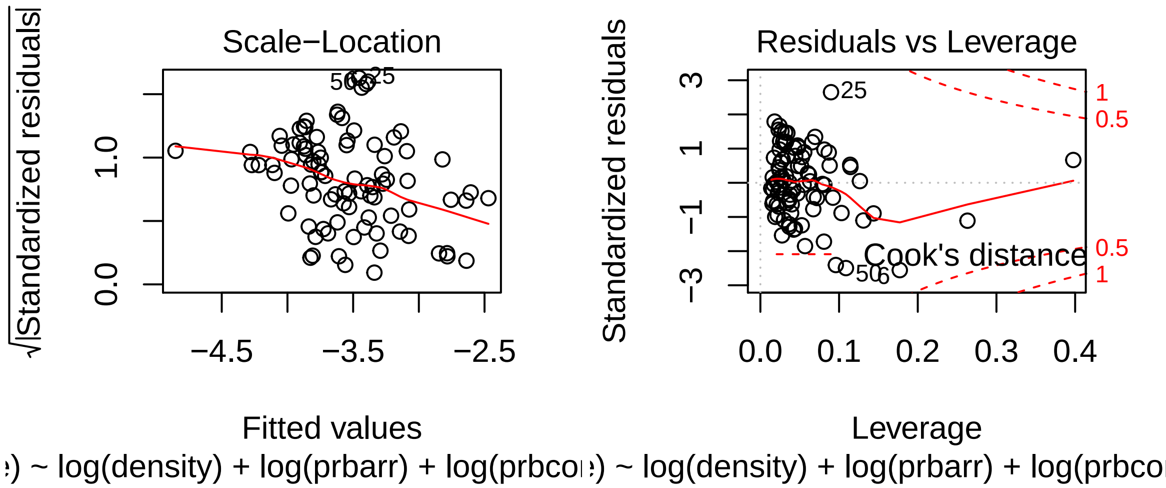
$$crimeDeterm = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(prbarr) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(pctymle)$$

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

```
##
## Call:
## lm(formula = log(crmrte) ~ log(density) + log(prbarr) + log(prbconv) +
##     log(pctymle), data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85369 -0.22285 -0.00454  0.26488  0.91643
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.07512    0.61271  -6.651 2.67e-09 ***
## log(density)  0.33799    0.05754   5.874 8.04e-08 ***
## log(prbarr)  -0.43186    0.11299  -3.822 0.000251 ***
## log(prbconv) -0.31078    0.07744  -4.013 0.000128 ***
## log(pctymle)  0.11105    0.21174   0.524 0.601332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3623 on 85 degrees of freedom
## Multiple R-squared:  0.5836, Adjusted R-squared:  0.564
## F-statistic: 29.79 on 4 and 85 DF,  p-value: 1.735e-15
```

```r
plot(model1)
```





Model 2: with key explanatory variables and only covariates

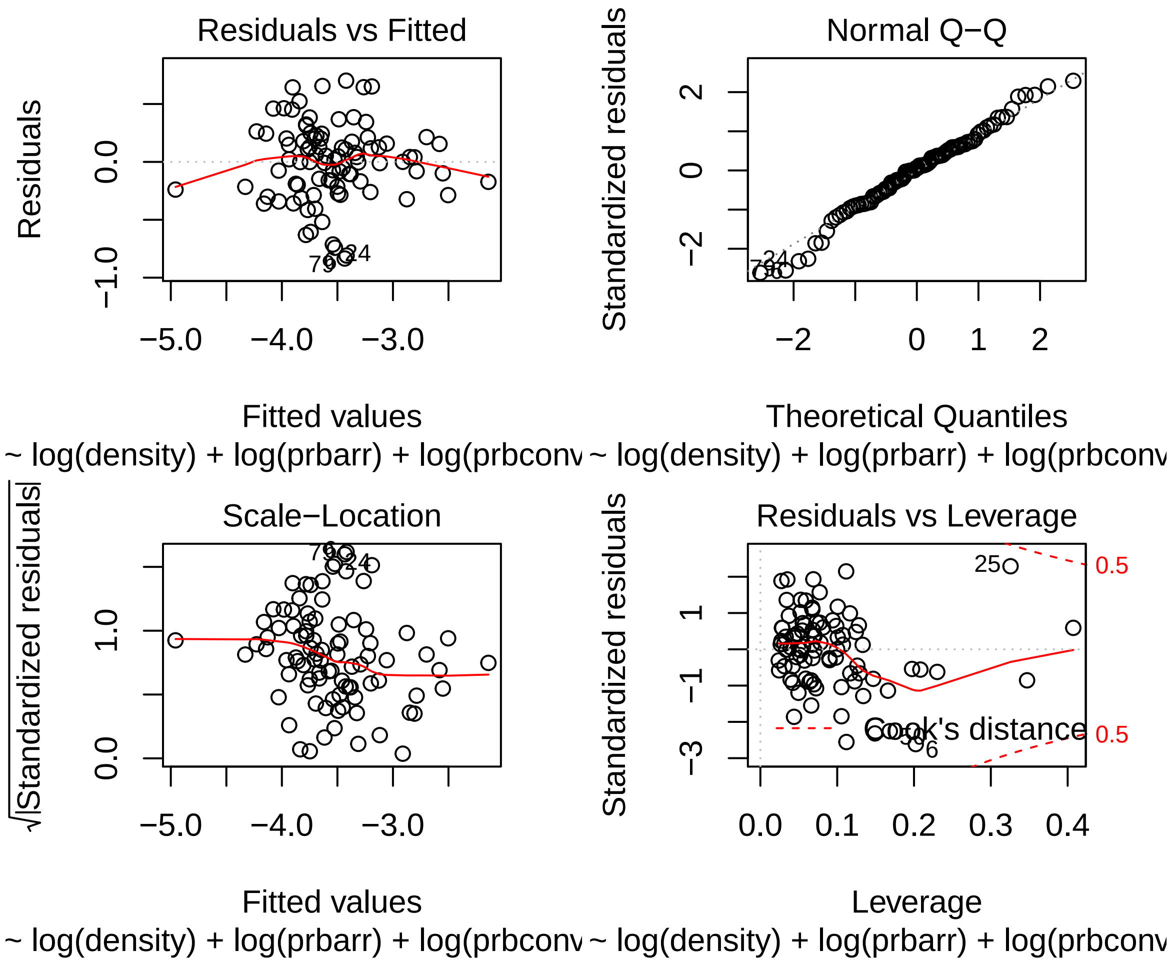
In this model, we'll include the variables (avgsen, mix, taxpc), as we think they will contribute to the accuracy of your results without introducing substantial bias.

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

```r
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
## -0.83479 -0.21034  0.02083  0.20191  0.70182
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.45768    0.79261  -5.624 2.51e-07 ***
## log(density)  0.33269    0.05585   5.957 6.15e-08 ***
## log(prbarr)  -0.49686    0.11926  -4.166 7.63e-05 ***
## log(prbconv) -0.20228    0.08137  -2.486   0.0150 *
## log(pctymle)  0.26433    0.20957   1.261   0.2108
## log(avgsen)  -0.02142    0.13466  -0.159   0.8740
## log(mix)      0.19789    0.08444   2.343   0.0215 *
## log(taxpc)    0.34625    0.14832   2.334   0.0220 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3467 on 82 degrees of freedom
## Multiple R-squared:  0.6323, Adjusted R-squared:  0.6009
## F-statistic: 20.14 on 7 and 82 DF,  p-value: 1.818e-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.

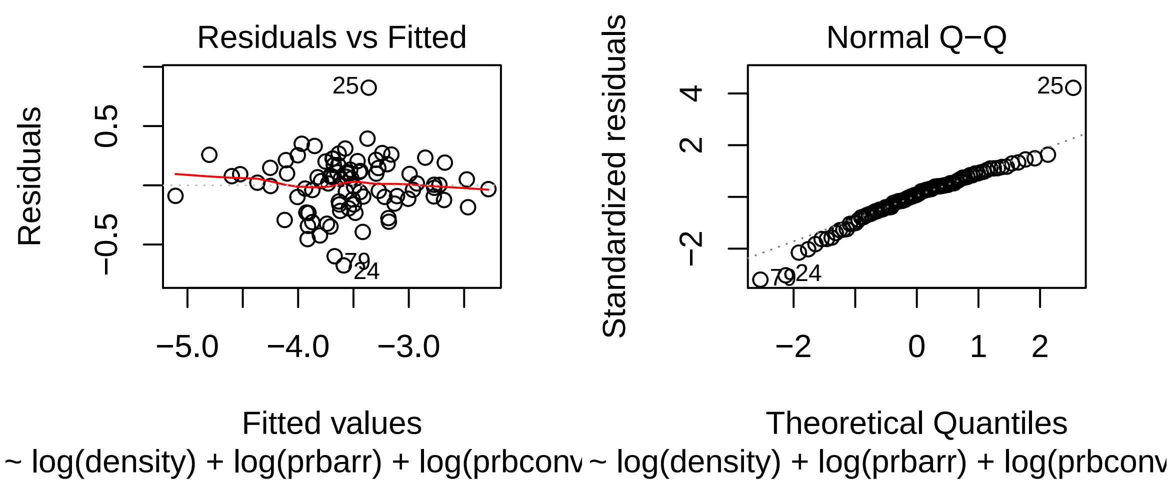
$$crimeDeterm = \beta_0 + \beta_1 \cdot log(density) + \beta_2 \cdot log(prbarr) + \beta_3 \cdot log(prbconv) + \beta_4 \cdot log(pctymle) + \beta_5 \cdot log(avgsen) + \beta_6 \cdot log(mix) + \beta_7 \cdot lo$$

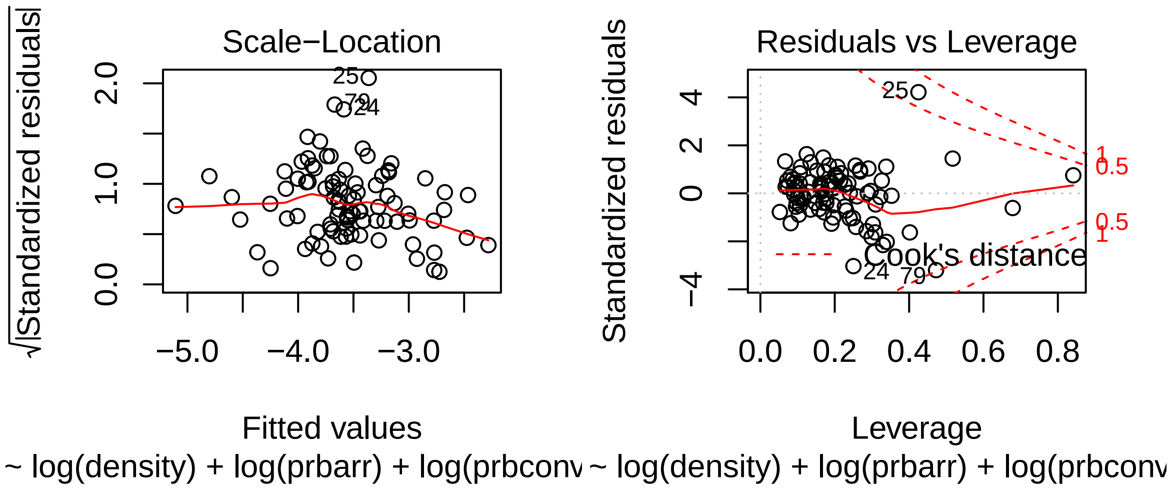
```r
model3 <- lm(log(crmrte) ~ log(density) + log(prbarr) + log(prbconv)
             + log(pctymle) + log(avgsen) + log(mix) + + log(taxpc)
             + prbpris + log(polpc)
             + log(pctmin80) + log(wcon) + log(wtrd) + wfir + log(wser) + log(wmfg)
             + log(wfed) + log(wsta) + 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
## -0.67661 -0.12308  0.01889  0.14228  0.82421
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.8632325  3.0655889  -2.239 0.028305 *
## log(density)   0.2135795  0.0609076   3.507 0.000791 ***
## log(prbarr)   -0.4952077  0.0940338  -5.266 1.42e-06 ***
## log(prbconv)  -0.2974700  0.0702459  -4.235 6.75e-05 ***
## log(pctymle)   0.2146154  0.1746706   1.229 0.223247
## log(avgsen)   -0.0946312  0.1179649  -0.802 0.425115
## log(mix)       0.0247883  0.0737707   0.336 0.737848
## log(taxpc)     0.1328496  0.1401862   0.948 0.346514
## prbpris        0.0164548  0.3609723   0.046 0.963769
## log(polpc)     0.2551010  0.1117918   2.282 0.025493 *
## log(pctmin80)  0.2241753  0.0349575   6.413 1.37e-08 ***
## log(wcon)      0.2257163  0.2322965   0.972 0.334512
## log(wtrd)      0.1445918  0.3245645   0.445 0.657318
## wfir          -0.0012859  0.0008211  -1.566 0.121774
## log(wser)     -0.2927780  0.1136340  -2.577 0.012061 *
## log(wmfg)      0.1771242  0.1585258   1.117 0.267624
## log(wfed)      0.7472945  0.3483480   2.145 0.035354 *
## log(wsta)     -0.2910889  0.2683262  -1.085 0.281666
## wloc          -0.0003693  0.0014835  -0.249 0.804122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2578 on 71 degrees of freedom
## Multiple R-squared:  0.8239, Adjusted R-squared:  0.7793
## F-statistic: 18.46 on 18 and 71 DF,  p-value: < 2.2e-16
```

```
plot(model3)
```





All 3 Regression models at a glance

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

```
##
## ==========================================================================================
##                                          Dependent variable:
##                       --------------------------------------------------------------------
##                                               log(crmrte)
##                              (1)                     (2)                       (3)
## ------------------------------------------------------------------------------------------
## log(density)              0.338***                0.333***                  0.214***
##                           (0.058)                 (0.056)                   (0.061)
##
## log(prbarr)              -0.432***               -0.497***                 -0.495***
##                           (0.113)                 (0.119)                   (0.094)
##
## log(prbconv)             -0.311***               -0.202**                  -0.297***
##                           (0.077)                 (0.081)                   (0.070)
##
## log(pctymle)               0.111                   0.264                     0.215
##                           (0.212)                 (0.210)                   (0.175)
##
## log(avgsen)                                       -0.021                    -0.095
##                                                   (0.135)                   (0.118)
##
## log(mix)                                          0.198**                    0.025
##                                                   (0.084)                   (0.074)
##
## log(taxpc)                                        0.346**                    0.133
##                                                   (0.148)                   (0.140)
##
## prbpris                                                                      0.016
##                                                                             (0.361)
##
## log(polpc)                                                                  0.255**
```

```
##                                                                      (0.112)
##
## log(pctmin80)                                                       0.224***
##                                                                      (0.035)
##
## log(wcon)                                                            0.226
##                                                                      (0.232)
##
## log(wtrd)                                                            0.145
##                                                                      (0.325)
##
## wfir                                                                -0.001
##                                                                      (0.001)
##
## log(wser)                                                           -0.293**
##                                                                      (0.114)
##
## log(wmfg)                                                            0.177
##                                                                      (0.159)
##
## log(wfed)                                                            0.747**
##                                                                      (0.348)
##
## log(wsta)                                                           -0.291
##                                                                      (0.268)
##
## wloc                                                                -0.0004
##                                                                      (0.001)
##
## Constant                  -4.075***             -4.458***            -6.863**
##                            (0.613)               (0.793)              (3.066)
##
## ------------------------------------------------------------------------------------------
## Observations                  90                    90                   90
## R2                          0.584                 0.632                0.824
## Adjusted R2                 0.564                 0.601                0.779
## Residual Std. Error    0.362 (df = 85)       0.347 (df = 82)      0.258 (df = 71)
## F Statistic         29.787*** (df = 4; 85) 20.142*** (df = 7; 82) 18.460*** (df = 18; 71)
## ==========================================================================================
## 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.

% population below poverty line

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.