Lab3: Reducing Crime

w203 Lab3

Harith Elrufaie and Gaurav Desai

Contents

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. 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))
suppressMessages(library(sandwich))
suppressMessages(library(car))
```

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

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

```
suppressMessages(library(lmtest))
```

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

Data Load

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

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

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, therefore we'll 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 analyzed 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$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
```

Exploratory Data Analysis

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.

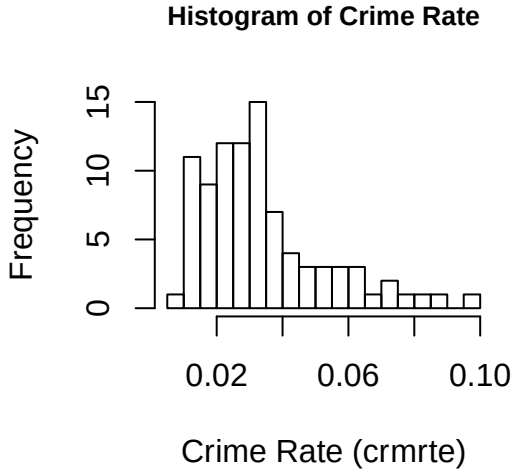
Univariate Analysis

crmrte: crimes committed per person

This is outcome variable for our regression model where we will try and derive relationships between various independent variables and crime rate.

Looking at the quarantines 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. This is also evident from histogram.

```
hist(crimeData$crmrte, breaks=20, main = "Histogram of Crime Rate"
     , cex.main=0.8, xlab="Crime Rate (crmrte)")
```

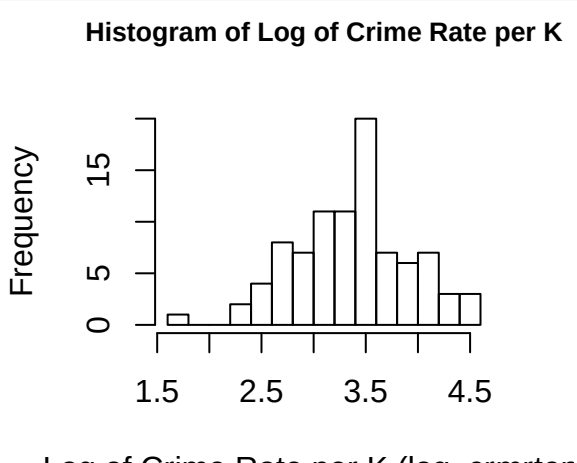


To take care of outliers and fit the variable into normal distribution, we can easily take a log of crime rate. However, we observed that the values of crimes rates per person are between 0 and 1. This range is not suitable for logarithms. Instead, we decided to scale by creating new variable for crime rate per 1000 people (crmrtepk) and then lets take `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)

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

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

```
crimeData$crmrtepk <- crimeData$crmrte * 1000
crimeData$log_crmrtepk <- log(crimeData$crmrtepk)
hist(crimeData$log_crmrtepk, breaks=20, main = "Histogram of Log of Crime Rate per K"
     , cex.main=0.8, xlab="Log of Crime Rate per K (log_crmrtepk)")
```




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

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

Also we noticed the right 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.

Convert polpc from per capita to per 1000 people to keep the scale

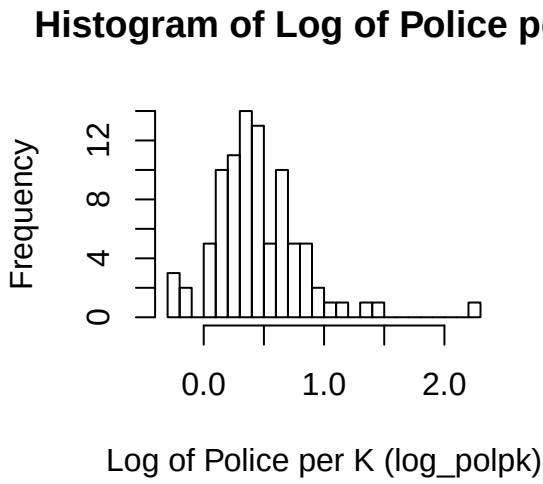
Since we have converted crimerate from per capita to per K people, lets also convert other per capita variable polpc to same scale. While scaling we notice that for county 115 the police per 1000 people is highest at 9 while average is just 1.7. Notably the second highest police per 1k is 4.5. Crime 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 logically explain highest average sentence in that county. We 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$log_polpk <- log(crimeData$polpc * 1000)
hist(crimeData$log_polpk, breaks=20
     , main = "Histogram of Log of Police per K"
     , xlab="Log of Police per K (log_polpk)")
summary(crimeData$polpk)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
crimeData[crimeData$polpc>.009,c("county", "polpc", "log_polpk", "avgsen")]
```

```
##    county      polpc log_polpk avgsen
## 51    115 0.00905433  2.203243   20.7
```



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")]
```

```
## 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. In fact 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 analyzing polpc variable earlier.

Under normal circumstances probabilities should not cross 0 to 1 range, but in this case the probabilities are mere proxies to actual police and judiciary data. One of the possible explanation to more convictions than arrest could be transfers of arrested people from outside counties where they were arrested to court locations within county. In absence 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,c("county","prbconv","avgsen","pctmin80","wser")]
```

```
##    county prbconv avgsen pctmin80     wser
## 84    185 2.12121   5.38  64.3482 2177.068
```

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 sentence 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 petite 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)

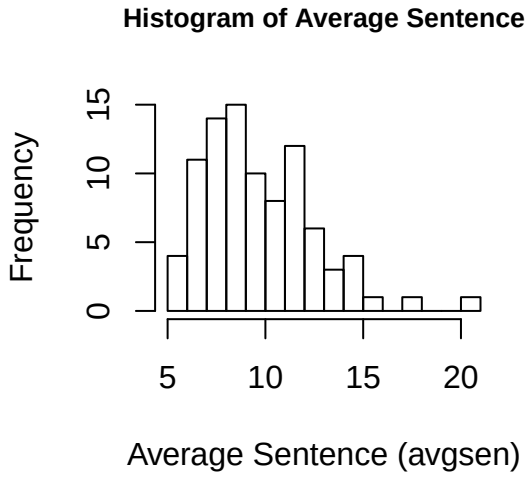
avgsen shows normal distribution with couple of outliers on right. Out of top 3 counties with average sentence, we have already analysed county 115 while analyzing polpc. The other two counties 41 and 127 have very high % of minority (42% and 34% respectively). It is difficult to draw conclusion as to why higher average sentence in these areas without any spike in crime rate. Concerned authorities should investigate 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$avgsen, breaks=20, main = "Histogram of Average Sentence"
     , cex.main=0.8, xlab="Average Sentence (avgsen)")
crimeData[crimeData$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 outliers 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,c("county", "density", "mix")]
```

```
##    county      density       mix
## 53    119 8.8276519780 0.1686990
## 79    173 0.0000203422 0.4197531
```

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

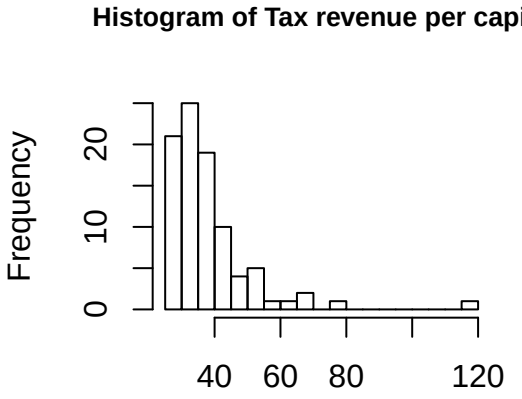
```
density.mean <- mean(crimeData[crimeData$density>.3,]$density)
crimeData[crimeData$density<.3,]$density <- density.mean
```

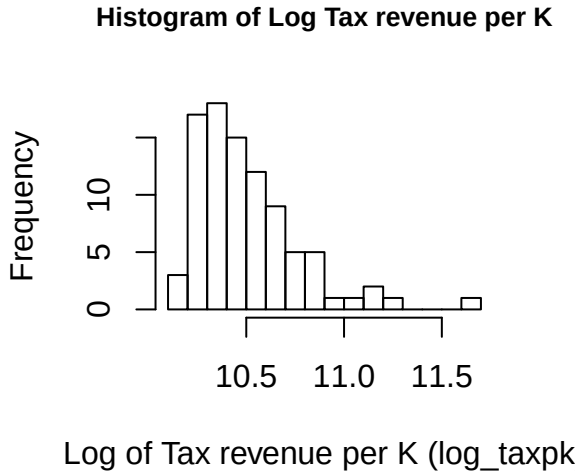
taxpc: tax revenue per capita

Looking at the histogram of tax revenue per capita, 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 nor does it show any super high wages to imply high taxes. So this county looks to be wealthy county in general with population paying high taxes from income outside wages.

```
hist(crimeData$taxpc, breaks=20, main = "Histogram of Tax revenue per capita"
     , cex.main=0.8, xlab="Tax revenue per capta (taxpc)")
crimeData$log_taxpk <- log(crimeData$taxpc*1000)
hist(crimeData$log_taxpk, breaks=20, main = "Histogram of Log Tax revenue per K"
     , cex.main=0.8, xlab="Log of Tax revenue per K (log_taxpk)")
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 log_polpk log_taxpk
## 25 0.07613807  79.0163     4.369654  1.388696  11.69326
```

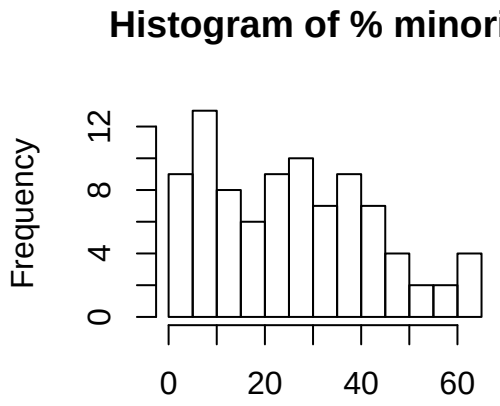




pctmin80: perc. minority, 1980

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

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



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

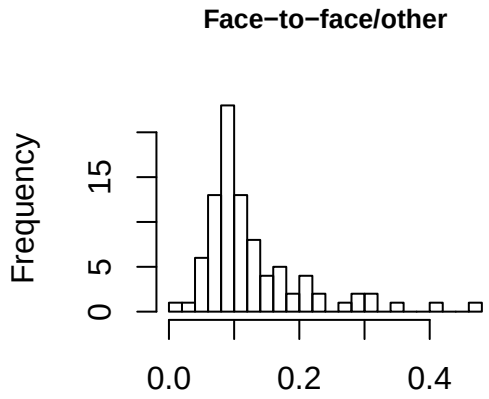
Looking at the histogram, the distribution appears to be slightly positively skewed with few outliers. But otherwise this is fairly normally distributed. Looking at the top 2 counties for mix are located in the western region. Difficult to draw any conclusion based on this but something for authorities 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 labor intensive work setup in this county though manufacturing or any other wage does not support this deduction. In absence of any other evidence we will keep this outlier without any modification.

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

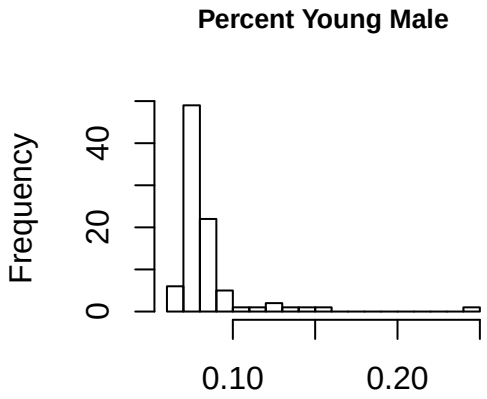
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## 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 log_polpk log_taxpk
## 59 0.2487116  55.1287      4.00967 0.4347456  10.22082
```

```
hist(crimeData$pctymle, breaks=20, main = "Percent Young Male"
     , 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.Total wage is almost perfectly normally distributed. The red line represents average for each of the category. Interestingly retail has least of the wages and fed has the highest wage.

Since we don't find significant difference in any of the wages, going forward we will use wtotal as proxy for various wages to see effect of wage on crime..

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

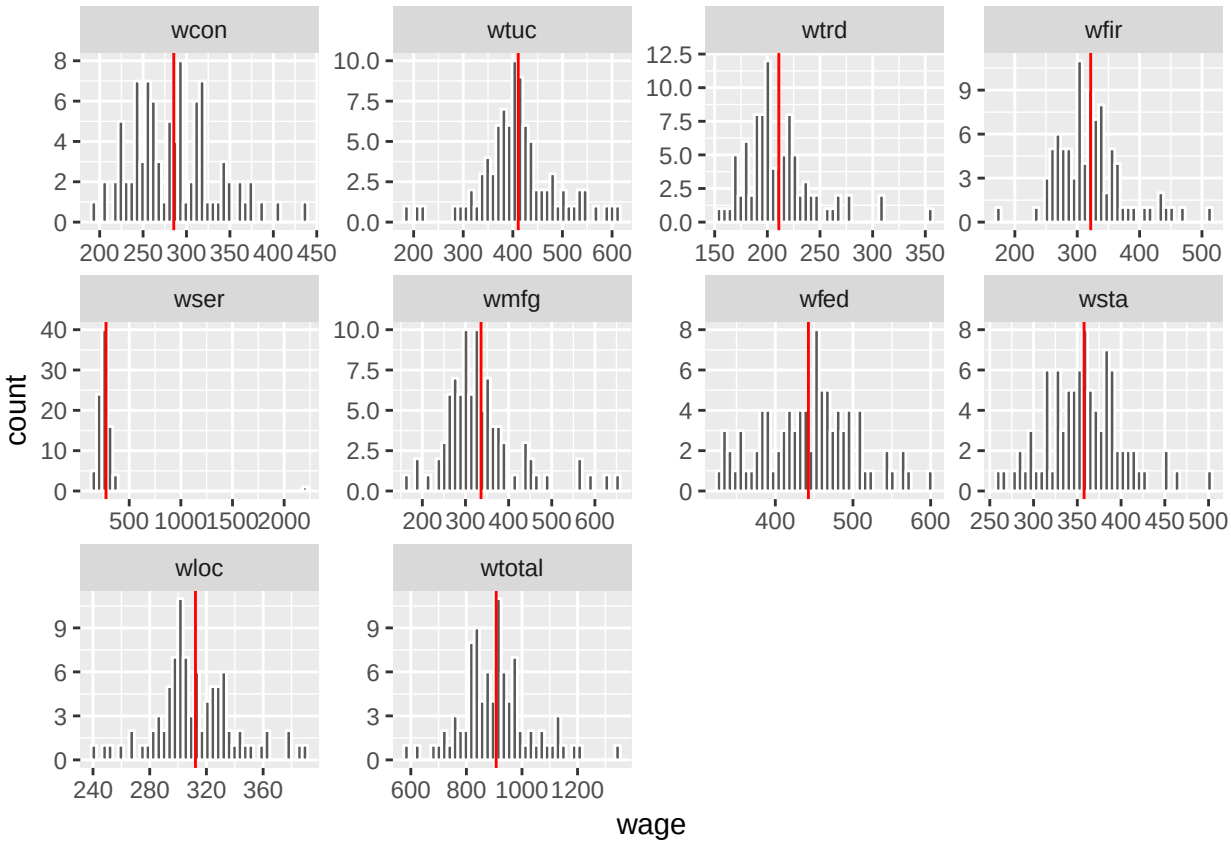
```
##  [1] 1061.8897  751.2527  753.2913  819.5865  795.3958  754.3157  872.9328
##  [8]  797.4570  816.3483 1123.6675 1061.0306  967.5013 1032.2563  918.5653
## [15]  865.1721  997.4808  858.9161  812.6532  836.4007  962.3981  964.8537
## [22]  992.3221 1023.8902  780.2143  825.1936  924.5911  950.5421  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")
```



Geographic Indicators

Lets look at the indicator flags for west and central region and indicator for urban counties. There are 22 counties marked under west region and 34 as central. Rest of the counties are neither in west or central region, so we assume they are in east region of North Carolina. Since west and central are not mutually exclusive we can use them for our regression model as is. Similarly 8 counties are marked as urban, so we assume rest of the counties as non urban counties.

```
sum(crimeData$west)
```

```
## [1] 22
```

```
sum(crimeData$central)
```

```
## [1] 34
```

```
sum(crimeData$urban)
```

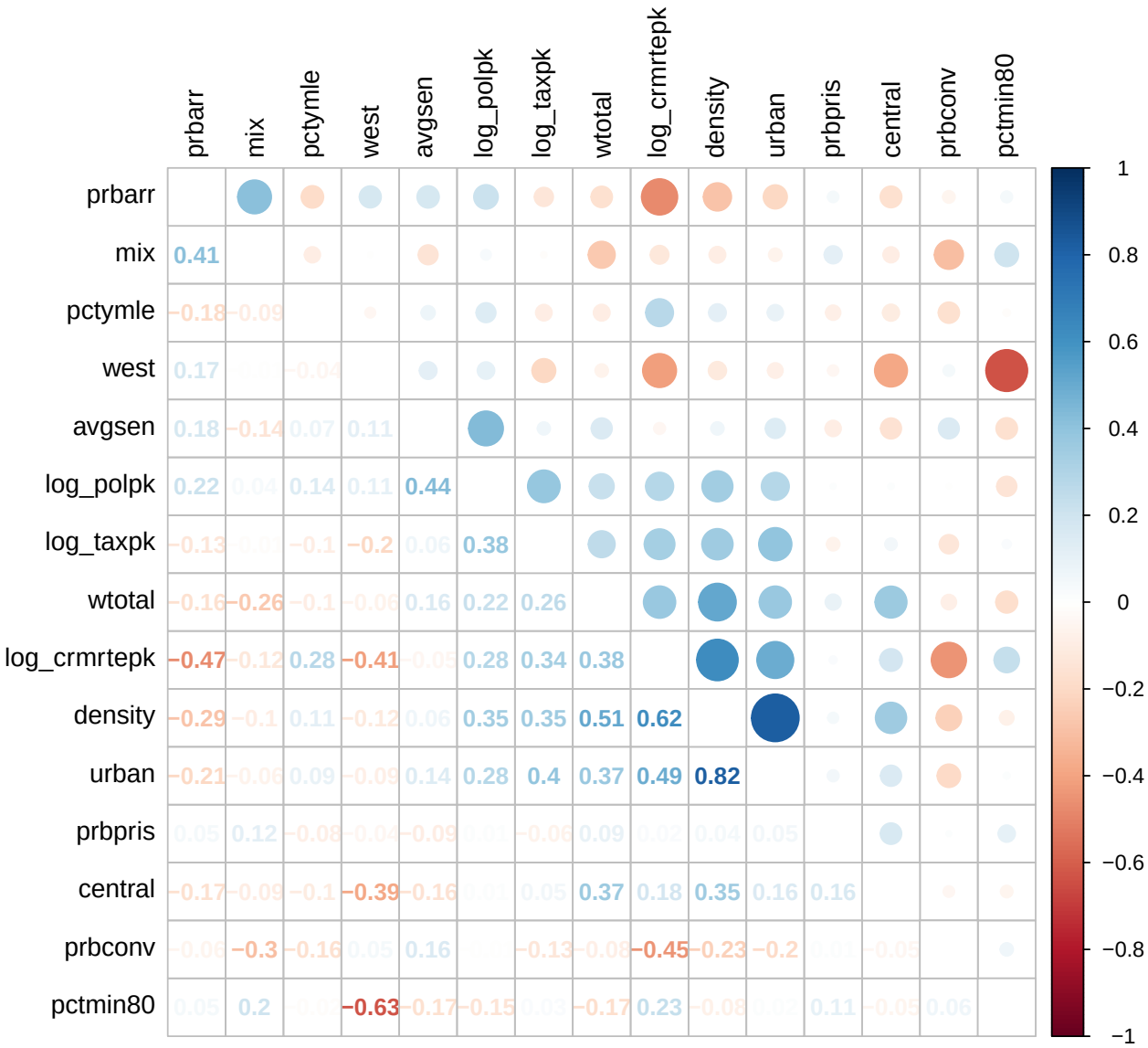
```
## [1] 8
```

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.mixed(cor(crimeData[ , (names(crimeData) %in%
                                 c("log_crmrtepk", "prbarr", "prbconv", "prbpris"
                                   , "avgsen", "log_polpk", "density", "log_taxpk"
```

```
                                , "pctmin80", "wtotal", "mix", "pctymle", "west"
                                , "central", "urban"))]),
                tl.pos = "lt", tl.col="black", order="hclust", number.cex=.9, number.digits=2)
```



We can see *strong positive* correlation (>.30) between crime rate (log_crmrtepk) and population density (density), total wages (wtotal), taxes (log_taxpk) and whether the counties is urban (urban). Which is logical in the sense that as population density increases because of urbanization, then wages and taxes would go up and so would the crimes rate in that area will increase. Note that Density itself is correlated with total wages, taxes and urbanization, so we can take only one of these variables in our model to avoid multicollinearity

On the opposite side, we can see *strong negative* correlation (< -.30) between crime rate (log_crmrtepk) and probability of arrest (prbarr) and probability of conviction (prbconv). And the two probabilities are not correlated with each other. We also see strong negative correlation with western counties. It indicates lower crimes in western counties.

Apart from these strong correlations, we also have *weak positive* correlation between crime rate and % of minority (pctmin80). The relation is not so strong and hence we need not include in our primary model.

Apart from effect on crime rate, there are some other interesting relations that can be seen here. For instance, the number of police per capita (log_polpk) increases as taxes (log_taxpk) and population density increases. And, as police force strengthens the Average sentence (avgsen) goes up. Maybe the additional police force catches serious criminals who get longer duration sentences?

There is another interesting trio of relationships. As mix of face to face crimes go up the probability of arrest goes up but probability of conviction goes down. Logical explanation of this situation would be since there are more face to face crimes, it is easier to identify the person involved and hence more and may be faster arrests, but these extra arrest do not translate to convictions and hence they drag down the conviction rate.

Proposed Models

Model 1: with only the explanatory variables

As observed during our EDA, probability of arrest (prbarr), probability of conviction (prbconv), density (density) and whether the county is in western region (west) show largest effect on crime rate (log_crmrtepk). Therefore, it is logical to include those variables in our model.

Although it is tempting to include log_polpk, we decided not to include it. We found it illogical to say crime rate increases as police per capita increases, whereas the reality is other way round, that is, police per capita increases as crime rate increases.

We have also considered (log_taxpk, pctymle) for this model, but we concluded that none of these is statistically significant.

Given all of the above, we're recommending the following model:

$$crimeRateDeterm = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot prbarr + \beta_3 \cdot prbconv + \beta_4 \cdot west$$

```
model1 <- lm(log_crmrtepk ~ density + prbarr + prbconv + west, data=crimeData)
summary(model1)$coefficients[,"Estimate"]
```

```
## (Intercept)     density      prbarr     prbconv        west
##   3.9240159   0.1492464  -1.2913195  -0.5491441  -0.3719293
```

```
summary(model1)$cov.unscaled
```

```
##              (Intercept)      density      prbarr      prbconv
## (Intercept)  0.147156558 -0.017401836 -0.23671077 -0.071572760
## density     -0.017401836  0.005752265  0.01834468  0.006101118
## prbarr      -0.236710770  0.018344681  0.67235134  0.034763111
## prbconv     -0.071572760  0.006101118  0.03476311  0.096615952
## west        -0.006009097  0.001120028 -0.03098886 -0.003150793
##                     west
## (Intercept) -0.006009097
## density      0.001120028
## prbarr      -0.030988856
## prbconv     -0.003150793
## west         0.062459026
```

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

```
## [1] 0.6743701
```

The model fits fir 67.4% of the population. This is fairly good fit. The crime rate is positively proportional to density while inversely proportional to rest of the explanatory variables. We also note that none of the explanatory variables is highly correlated with any other explanatory variable.

Now lets look at the coefficients for their practical significance. Every unit increase in density results in approximately 1% increase in crime rate. If we increase the probability of arrest by by 1% then crime rate would decrease by approximately 12% (1/100 of coefficient). Similarly for 1% increase in conviction ration will decrease crime rate by 5% (1/100 of coefficient) If the county is in western region then crime rate is 37% lesser than average crime rate.

Model 2: with key explanatory variables and only covariates

In this model, we'll include the variables (avgsen, mix, pctymle), as we think they will contribute to the accuracy of your results without introducing substantial bias. These variables show good degree of correlation with crime rate, as well as, linear relationship as observed in the EDA. Also these variables do not show high correlation with any of the model 1 explanatory variables so there is less chance of multicolinearity.

We also considered other variables such as taxpc, but because of the non-linear relationship with crime rate and the clear violation to the Linearity assumption, we decided not to include it.

pctmin80 was also considered for this model. Although it is statistically significant, we realized that a negative correlation between western counties and pctmin80.

$$crimeRateDeterm = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot prbarr + \beta_3 \cdot prbconv + \beta_4 \cdot west + \beta_5 \cdot avgsen + \beta_6 \cdot mix + \beta_7 \cdot pctymle$$

```
model2 <- lm(log_crmrtepk ~ density + prbarr + prbconv + west
             + avgsen + mix + pctymle, data=crimeData)
summary(model2)$coefficients[,"Estimate"]
```

```
## (Intercept)     density      prbarr     prbconv        west      avgsen
##  3.66701157  0.14347370 -1.17198298 -0.56853146 -0.38325557  0.01042442
##         mix     pctymle
## -0.41061564  2.32770444
```

```
summary(model2)$cov.unscaled
```

```
##              (Intercept)       density      prbarr       prbconv
## (Intercept)  0.457958359 -0.0146646049 -0.17238766 -1.032870e-01
## density     -0.014664605  0.0059478435  0.01990287  7.296741e-03
## prbarr      -0.172387660  0.0199028712  0.88250640  2.144027e-02
## prbconv     -0.103286965  0.0072967410  0.02144027  1.137620e-01
## west        -0.005238902  0.0014628977 -0.03266300 -3.630963e-05
## avgsen      -0.009731412 -0.0005207095 -0.01099706 -2.231323e-03
## mix         -0.363689545  0.0043003440 -0.58470638  1.458806e-01
## pctymle     -2.085843554  0.0031717819  0.72739663  3.153684e-01
##                     west        avgsen          mix      pctymle
## (Intercept) -5.238902e-03 -0.0097314117 -0.363689545 -2.085843554
## density      1.462898e-03 -0.0005207095  0.004300344  0.003171782
## prbarr      -3.266300e-02 -0.0109970602 -0.584706379  0.727396632
## prbconv     -3.630963e-05 -0.0022313233  0.145880649  0.315368449
## west         6.319433e-02 -0.0007489022  0.024732249  0.016606200
## avgsen      -7.489022e-04  0.0016271110  0.011060559 -0.024348085
## mix          2.473225e-02  0.0110605586  2.353139153  0.390847525
```

```
## pctymle      1.660620e-02 -0.0243480853  0.390847525 22.304595946
```

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

```
## [1] 0.6818515
```

We see slight increase in model fit. We are able to fit 68% of our population by adding 3 new covariates into model 1. But at the same time we see that some of the new covariates added in model 2 have high covariance with existing variables.

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.

$$\begin{aligned} crimeRateDeterm = \beta_0 &+ \beta_1 \cdot density + \beta_2 \cdot prbarr + \beta_3 \cdot prbconv + \beta_4 \cdot west \\ &+ \beta_5 \cdot avgsen + \beta_6 \cdot mix + \beta_7 \cdot pctmyle \beta_8 \cdot pctmin80 + \beta_9 \cdot log\_taxpk \\ &+ \beta_{10} \cdot urban + \beta_{11} \cdot central + \beta_{12} \cdot prbpris + \beta_{13} \cdot log\_polpk + \beta_{14} \cdot wtotal \end{aligned}$$

```
model3 <- lm(log_crmrtepk ~ density + prbarr + prbconv + west
             + avgsen + mix + pctymle + pctmin80 + log_taxpk + urban + central
             +  + prbpris  + log_polpk + wtotal, data=crimeData)
summary(model3)$coefficients[,"Estimate"]
```

```
##   (Intercept)       density        prbarr       prbconv          west
##  5.0875877549  0.1149128456 -1.6880903522 -0.6686695406 -0.3115946591
##        avgsen           mix       pctymle      pctmin80     log_taxpk
## -0.0142865861 -0.7096831764  0.6884650253  0.0074495568 -0.1633175220
##         urban       central       prbpris     log_polpk        wtotal
## -0.1293995909 -0.1837552281  0.0878940174  0.5505428418  0.0006051242
```

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

```
## [1] 0.8104533
```

The adjusted R square has jumped to 81% indicating the all inclusive model3 is able to predict 81% of the population. But we expect a lot of multicollinearity and overlap between various variables making it difficult to identify true effect of any one variable on crime rate.

All 3 Regression models at a glance

```
cov1 <- vcovHC(model1)
robust.se1 <- sqrt(diag(cov1))
cov2 <- vcovHC(model2)
robust.se2 <- sqrt(diag(cov2))
cov3 <- vcovHC(model3)
robust.se3 <- sqrt(diag(cov3))
robust.se <- list(robust.se1, robust.se2, robust.se3)

stargazer(model1, model2, model3
          ,dep.var.labels = "Log of Crime Rate per 1000 People"
          ,covariate.labels = c("Probability of Arrest", "Probability of Conviction"
```

```
                              ,"Population Density", "Is Western County"
                              , "Average Sentence", "Face to Face Crime %"
                              , "% of Male", "% of Minority"
                              , "Log of tax per K", "Is Urban County", "Is Central County"
                              , "Probability of Prison", "Log of Police per K"
                              , "Total Wage")
          ,order = c("prbarr", "prbconv", "density", "west"), single.row = TRUE
          ,title="Comparison of 3 Regression models", float=FALSE, header = FALSE, report="vc*sp"
          ,star.cutoffs = c(.05, .01, .001), se = robust.se)
```

	Dependent variable:		
	Log of Crime Rate per 1000 People		
	(1)	(2)	(3)
Probability of Arrest	-1.291^{**} (0.417)	-1.172^{***} (0.312)	-1.688^{***} (0.287)
	p = 0.002	p = 0.0002	p = 0.000
Probability of Conviction	-0.549^{***} (0.135)	-0.569^{***} (0.142)	-0.669^{***} (0.106)
	p = 0.0001	p = 0.0001	p = 0.000
Population Density	0.149^{***} (0.026)	0.143^{***} (0.023)	0.115^{*} (0.047)
	p = 0.000	p = 0.000	p = 0.014
Is Western County	-0.372^{***} (0.074)	-0.383^{***} (0.075)	-0.312^{**} (0.119)
	p = 0.00000	p = 0.00000	p = 0.009
Average Sentence		0.010 (0.014)	−0.014 (0.013)
		p = 0.449	p = 0.279
Face to Face Crime		p = 0.393	p = 0.160
		p = 0.009	p = 0.682
			p = 0.003
Log of tax per K			−0.163 (0.207)
			p = 0.431
Is Urban County			−0.129 (0.209)
			p = 0.535
Is Central County			-0.184^{*} (0.076)
			p = 0.017
Probability of Prison			0.088 (0.468)
			p = 0.852
Log of Police per K			0.551^{***} (0.135)
			p = 0.00005
Total Wage			0.001 (0.0003)
			p = 0.079
Constant	3.924^{***} (0.206)	3.667^{***} (0.233)	5.088^{*} (2.084)
	p = 0.000	p = 0.000	p = 0.015
Observations	90	90	90
R^2	0.689	0.707	0.840
Adjusted R^2	0.674	0.682	0.810
Residual Std. Error	0.313 (df = 85)	0.310 (df = 82)	0.239 (df = 75)
F Statistic	47.079^{***} (df = 4; 85)	28.249^{***} (df = 7; 82)	28.182^{***} (df = 14; 75)
Note:			*p<0.05; **p<0.01; ***p<0.001

The above summary demonstrates the following:

1. `prbarr` and `prbconv` have the best estimates in all three models.
2. `density` coefficient drastically changes in model 3 because variables like `wtotal` and `log_polpk` distribution are synonymous to increase in density and they absorb some of the causality of density.

3. The addition of `pctmin80` in model 3 reduces the effect of Is Western County as western counties have very low minority population hence `pctmin80` absorbs some of the causal effect of west variable in model 3.
4. Adding additional variables to model 2 could have increased the R squared but we'll introduce violations to CLM assumptions.
5. Although Model 3 has the highest R square, CLM assumptions such as linearity and multicollinearity are violated.

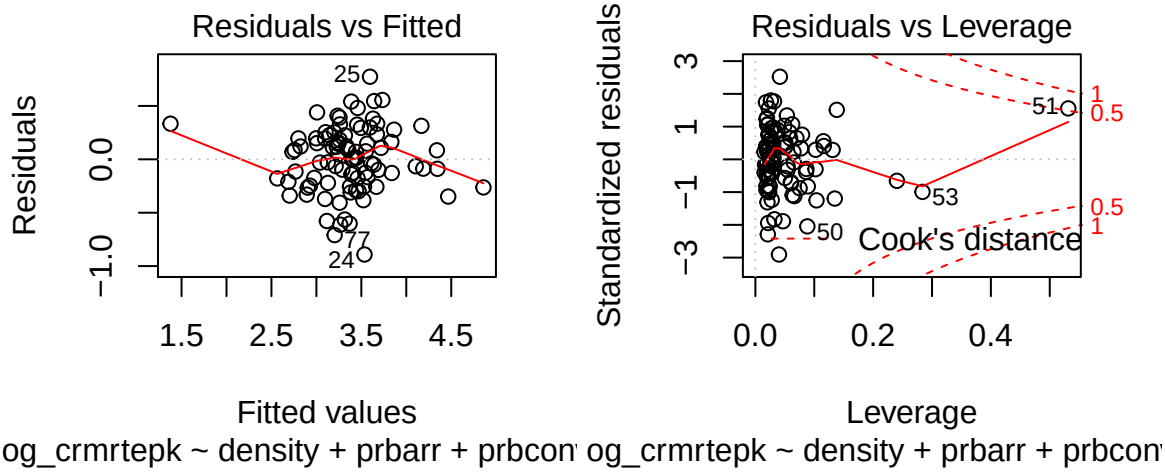
CLM Assumptions Analysis

Model1 CLM Assumptions Analysis

Assumption.1 - Zero Conditional Mean

We'll now plot model1 in order to assess if the model has zero conditional mean.

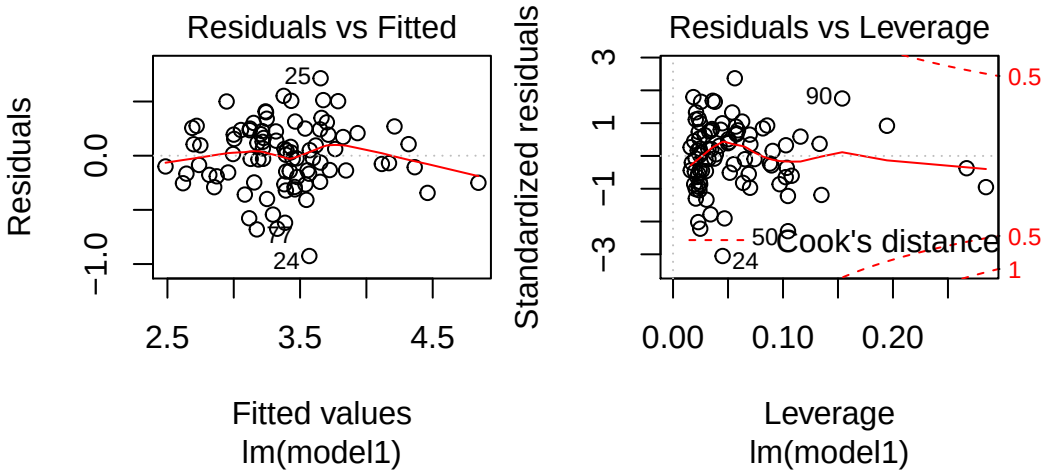
```
plot(model1, which=c(1, 5))
```



Looking at the above plots we observe the following:

1. The residual vs fitted indicates that while the red spline line remains close to 0, there is a slight dip and rise and both ends, which may be due to some outlier observations.
2. The residual vs leverage indicates that there are outlier (#51 and #53 to be precise). Observation #51 in particular, is just under Cook's distance of 1, which is concerning. We analyzed what is special about it and realized county 115 has highest values for three variables - police per capita, density and average sentence. There is something special about this county to show highest values in 3 separate explanatory variables. Our recommendation would be to remove this observation to reduce the effect of this outlier on our regression model. Since cook's distance does not cross critical value of 1 the assumption is not violated.

```
model1.no.outliers <- lm(model1, data = crimeData[-51, ])
plot(model1.no.outliers, which=c(1, 5))
```



Given #1 and #2, we're confident to say this assumption is met.

Assumption.2 - Linear in Parameters

Looking at the above residual vs leverage plot, we can say the assumption is met.

Assumption.3 - Random Sampling

It is not clear to us how the dataset was collected, but we only know it is from 90 counties. Given that North Carolina has 100 counties, it makes us believe this is good enough sampling to consider this assumption as met.

Assumption.4 - Multicollinearity

To test this assumption, we'll run the `vif` command

```
vif(model1)
```

```
##  density   prbarr  prbconv     west
## 1.173244 1.134185 1.078752 1.038208
```

Given the small values (less than 5) for all the variables, we'll consider this assumption is met.

Assumption.5 - Homoskedasticity

Lets run Breusch-Pagan test to verify Homosckedasticity

```
bptest(model1)
```

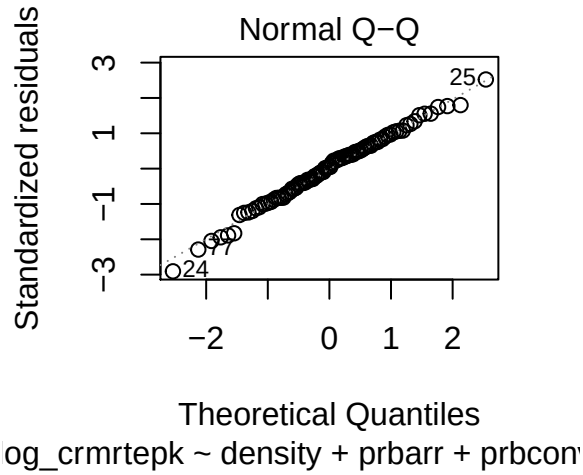
```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 5.6491, df = 4, p-value = 0.2269
```

Because of large p value (> 0.05), our null hypothesis that model is Homosckedastic can not be rejected. So the assumption is met.

Assumption.6 - Normality of Residuals

We will look at the QQ-plot to assess the normality of residuals.

```
plot(model1, which=2)
```

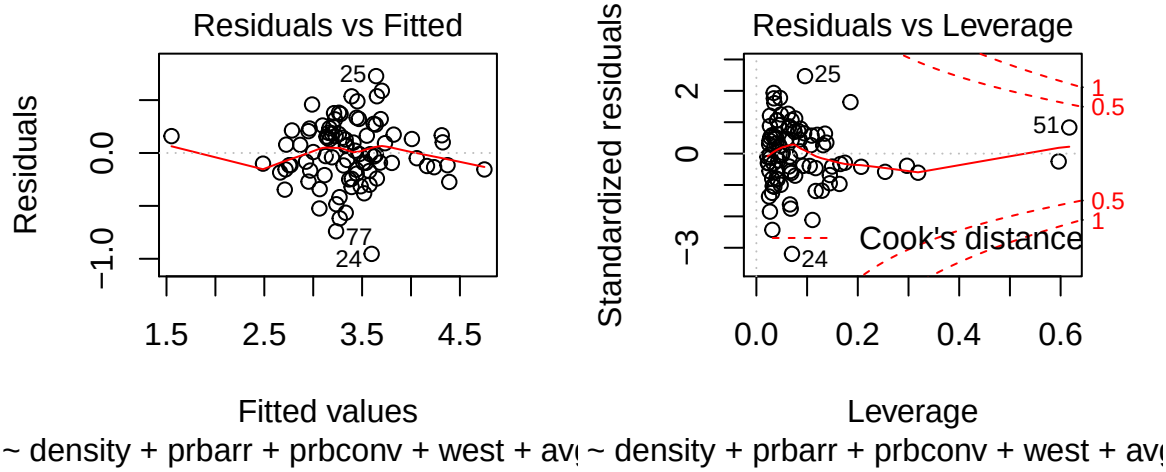


We see a beautiful solid line with few outliers at each end, but we're still considering this condition is met.

Model2 CLM Assumptions Analysis

Assumption.1 - Zero Conditional Mean

```
plot(model2, which=c(1, 5))
```



In the residual vs fitted plot we can observe a nice red spline line remains close to 0, there is a slight dip in one side, but it is not significant. The assumption is met.

Assumption.2 - Linear in Parameters

Looking at the above residual vs leverage plot, we can say the assumption is met.

Assumption.3 - Random Sampling

It is not clear to us how the dataset was collected, but we only know it is from 90 counties. Given that North Carolina has 100 counties, it makes us believe this is good enough sampling to consider this assumption as met.

Assumption.4 - Multicollinearity

To test this assumption, we'll run the `vif` command

```
vif(model2)
```

```
##  density   prbarr  prbconv     west   avgsen      mix  pctymle
## 1.213135 1.488694 1.270194 1.050430 1.163292 1.400100 1.091551
```

Given the small values for all the variables, we'll consider this assumption is met.

Assumption.5 - Homoskedasticity

Lets run Breusch-Pagan test to verify Homosckedasticity

```
bptest(model2)
```

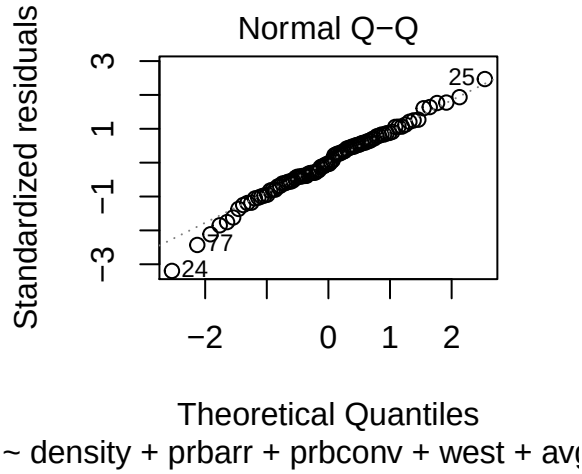
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 11.913, df = 7, p-value = 0.1035
```

Because of large p value (> 0.05), our null hypothesis that model is Homosckedastic can not be rejected. So the assumption is met.

Assumption.6 - Normality of Residuals

We will look at the QQ-plot to assess the normality of residuals.

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

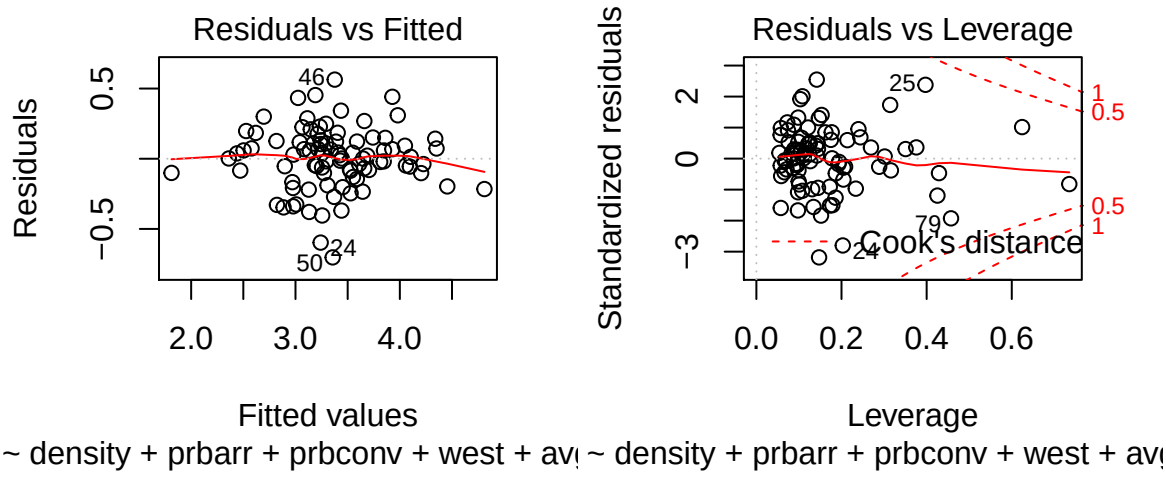


We can see a clear normality proof line.

Model3 CLM Assumptions Analysis

Assumption.1 - Zero Conditional Mean

```
plot(model3, which=c(1, 5))
```



In the residual vs fitted plot we can observe a nice red spline line remains close to 0, there is a slight dip in one side, but it is not significant. The assumption is met.

Assumption.2 - Linear in Parameters

Looking at the above residual vs leverage plot, we can say the assumption is met.

Assumption.3 - Random Sampling

Similar to model1 and model2, we are not clear to us how the dataset was collected, but we only know it is from 90 counties. Given that North Carolina has 100 counties, it makes us believe this is good enough sampling to consider this assumption as met.

Assumption.4 - Multicollinearity

To test this assumption, we'll run the `vif` command

```
vif(model3)
```

```
##   density    prbarr   prbconv      west    avgsen       mix   pctymle
##  4.787375  1.702598  1.391606  2.999454  1.580992  1.577732  1.318111
##  pctmin80 log_taxpk     urban   central   prbpris log_polpk    wtotal
##  2.558952  1.719155  3.720864  2.004572  1.106791  2.059592  1.706811
```

density coefficient is 4.8 which is very close to being termed as high value. This indicates there is multicollinearity between density and another variable added in model 3. This was evident from our earlier correlation matrix analysis as well. Density shows high correlation with 'urban' and 'wtotal' variables. This is logical as urbanization accelerates, population density increases and wages grow.

Assumption.5 - Homoskedasticity

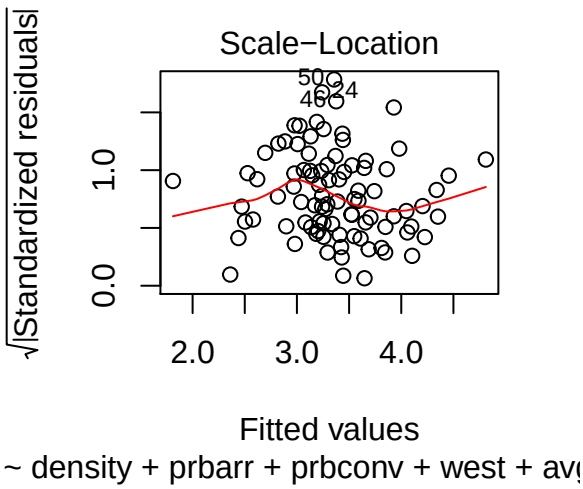
Lets run Breusch-Pagan test to verify Homosckedasticity

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  model3
## BP = 32.537, df = 14, p-value = 0.003358
```

```
plot(model3, which = 3)
```

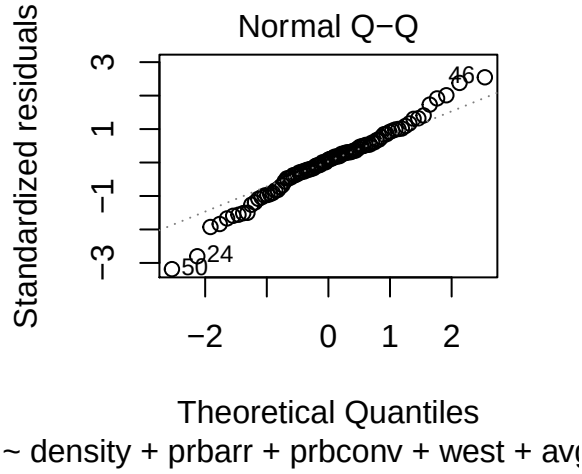


The p-values of < 0.05 indicates we can reject the null hypothesis that model is Homosckedastic i.e. model is Heteroskedestic. Same can be concluded by looking at Scale-Location plot where we can see residual variance range varies across fitted-values axes.

Assumption.6 - Normality of Residuals

We will look at the QQ-plot to assess the normality of residuals.

```
plot(model3, which=2)
```



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. Literacy can be measured by number of years of education per capita. This variable should have negative correlation with crime rate and positive correlation with tax per capita and wages.

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. It is difficult to measure corruption by observing any statistical figure. Only way to measure corruption is by conducting surveys and gathering public feedback. Corruption should have negative correlation with crime rate and tax per capita.
4. Historic criminal rate of the area: If previous generation had high criminal rate in a particular area then a next generation person being raised in that area has higher scope and encouragement to follow the same foot steps. So we should also measure this continuity effect. Of course there will always be exceptions and outliers in this measurement but crime rate is not something that spikes up or down rather it grows with time or comes down with time. So if we get 5 year, 10 year etc. time period average crime rates for that county we can better estimate future crime rate of a county and advise correctly to policy makers.

Conclusion

Our Regression Model (Model 1) indicates that as population density increases the crime rate goes up. The model also tells us that western counties have significantly lower (37%) crime rate than rest of the North Carolina. So policymakers need to pay attention to more urbanized or highly dense regions specially outside western region.

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. Between the two factors, 1 % increase in probability of arrest has higher impact on crime rate reduction vs 1 % increase in probability of conviction (12% vs 5%). So if policymakers have to choose one out of two due to budgetary or any other constraints then they should first look at strengthening the police force to increase arrest.

Apart from our analysis of regression model, we have following suggestions from our EDA that may help designing political campaign policies:

1. There is a strong correlation between crime rate and the probability of arrest and conviction. The campaign should focus on counties with the lowest number of arrests and convictions.
2. County 119 has 1 crime per every 10 people.The county also has highest population density. A detailed analysis is required to understand what are the causes of such a high crime rate. Possibly one or more of the omitted variables has a role in this.
3. We observe some drastic parameters for county 185. It has highest arrest to conviction ratio but least average sentence while having highest minority population. The county also shows very high weekly wage in service industry. This raises an alarm. It indicates lot of people are getting convicted but for smaller duration. The fact that county has highest population of minorities, policy makers need to be vigilant and ensure that minority is not harassed or abused. We need to check if there are lot of illegal immigrants working in service industry in this county. Also we need to verify if convictions are valid or law has been abused. If convictions are valid but for petite crimes then citizens should be educated about such crimes so that we can decrease the crimes and the arrest.