Lab3: Reducing Crime

w203 Lab3

Harith Elrufaie and Gaurav Desai

Contents

Introduction 2
Setup . . . 2
Data Load . . . 2
Data Quality/Clean-up . . . 2

Exploratory Data Analysis 3
Univariate Analysis . . . 3
Analysis of Key Relationships . . . 11

Proposed Models 13
Model 1: with only the explanatory variables . . . 13
Model 2: with key explanatory variables and only covariates . . . 16
Model 3: includes the previous covariates, and most, if not all, other covariates . . . 20
All 3 Regression models at a glance . . . 22
Omitted Variables . . . 22

Conclusion 23

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
```

```
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.

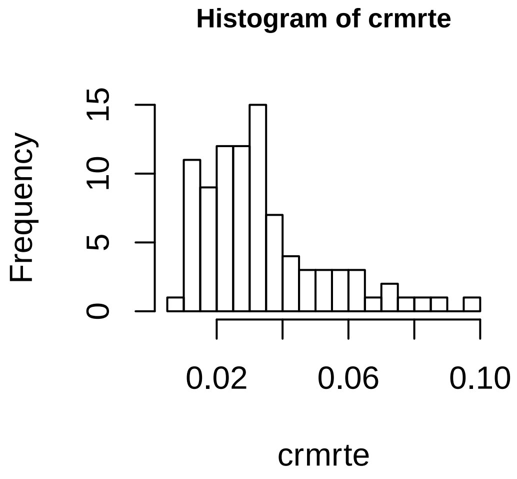
Univariate Analysis

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 crmrte"
     , cex.main=0.8, xlab="crmrte")
```

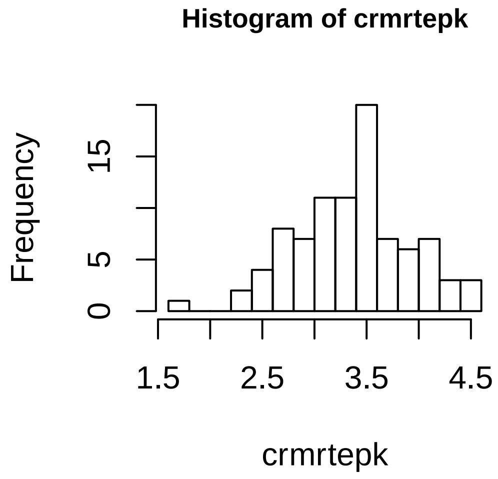


To take care of outliers and fit the variable into normal distribution, we can easily take a log of crime rate. However, we observed that the values of crimes rates per person are between 0 and 1. This range is not suitable for logarithms. Instead, we decided to scale by creating new variable for crime 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 crmrtepk"
     , cex.main=0.8, xlab="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

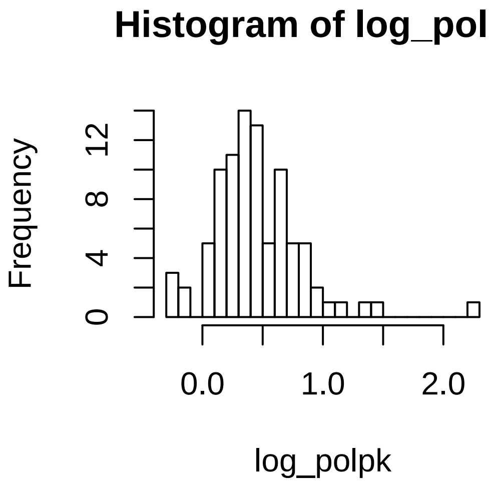
Since we have converted crimerate from per capita to per K people, lets also convert other per capita variable polpc to same scale. While scaling we notice that for county 115 the police per 1000 people is highest at 9 while average is just 1.7. Notably the second highest police per 1k is 4.5. Crime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

```r
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

```r
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)

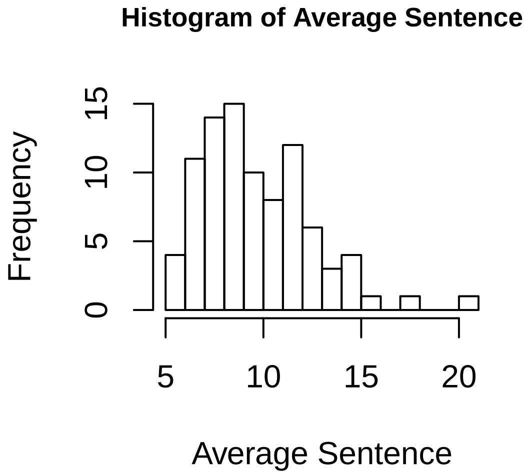
avgsen shows normal distribution with couple of outliers on right. Out of top 3 counties with average sentence, we have already analysed county 115 while analyzing polpc. The other two counties 41 and 127 have very high % of minority (42% and 34% respectively). It is difficult to draw conclusion as to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

```r
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$avgsen, breaks=20, main = "Histogram of Average Sentence"
     , cex.main=0.8, xlab="Average Sentence")
crimeData[crimeData$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

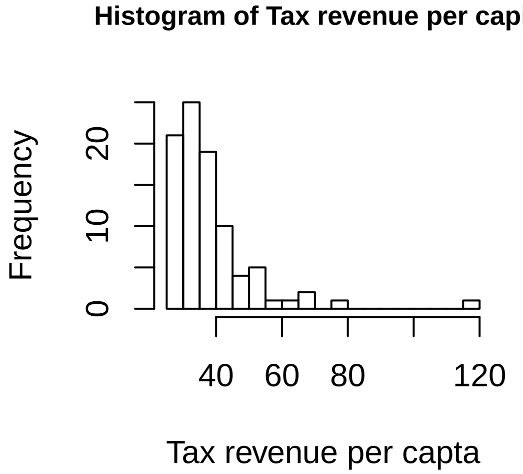
```
crimeData[crimeData$density<.3,]$density <- mean(crimeData[crimeData$density>.3,]$density)
```

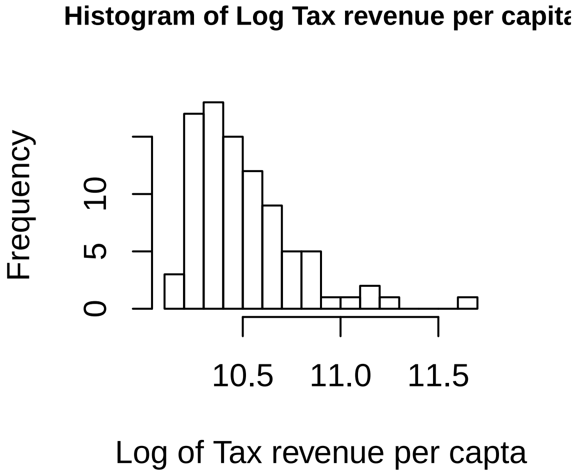
taxpc: tax revenue per capita

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

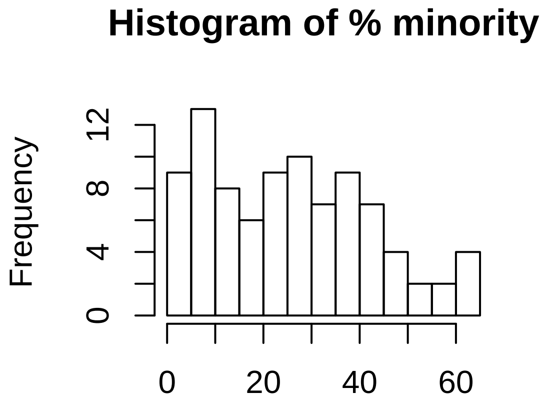




pctmin80: perc. minority, 1980

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

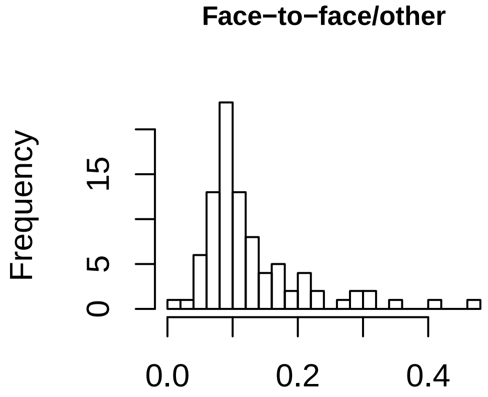
Looking at the histogram, the distribution appears to be slightly positively skewed with few outliers. But otherwise this is fairly normally distributed. Looking at the top 2 counties for mix are located in the western region. Difficult to draw any conclusion based on this but something for authorities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

```r
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
```

```r
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.

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

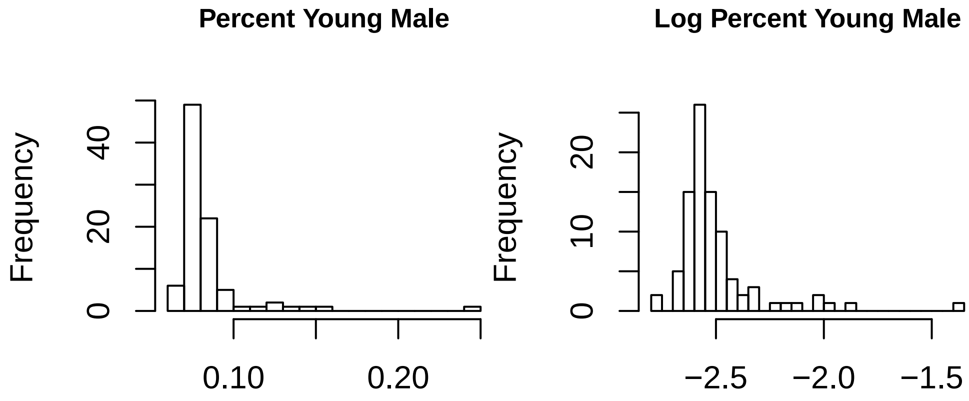
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06216 0.07437 0.07770 0.08403 0.08352 0.24871
```

```r
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
```

```r
hist(crimeData$pctymle, breaks=20, main = "Percent Young Male"
     , cex.main=.8, xlab = "")
```

```r
hist(log(crimeData$pctymle), breaks=20, main = "Log 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..

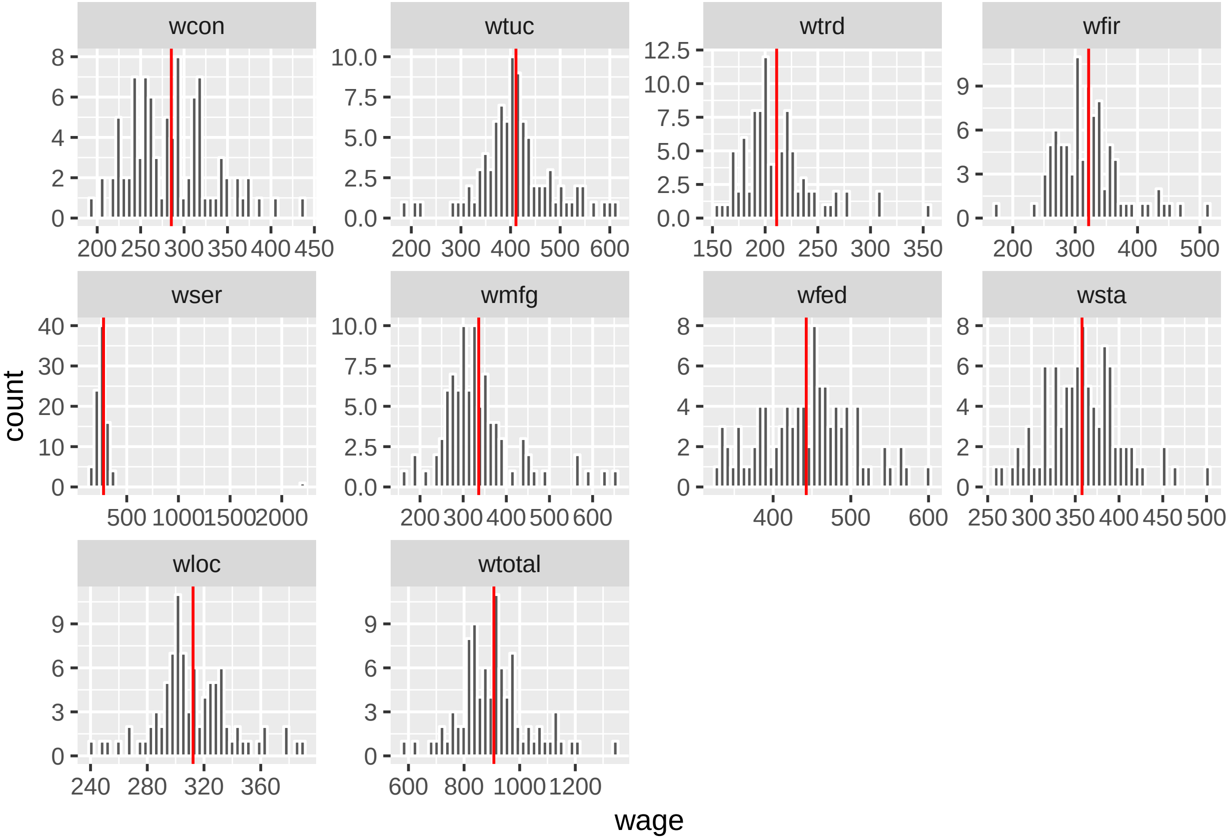
```r
crimeData$wtotal<-crimeData$wcon+crimeData$wtuc+crimeData$wtrd
                          +crimeData$wfir+crimeData$wser+crimeData$wmfg
```

```
##  [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
```

```r
                          +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
```

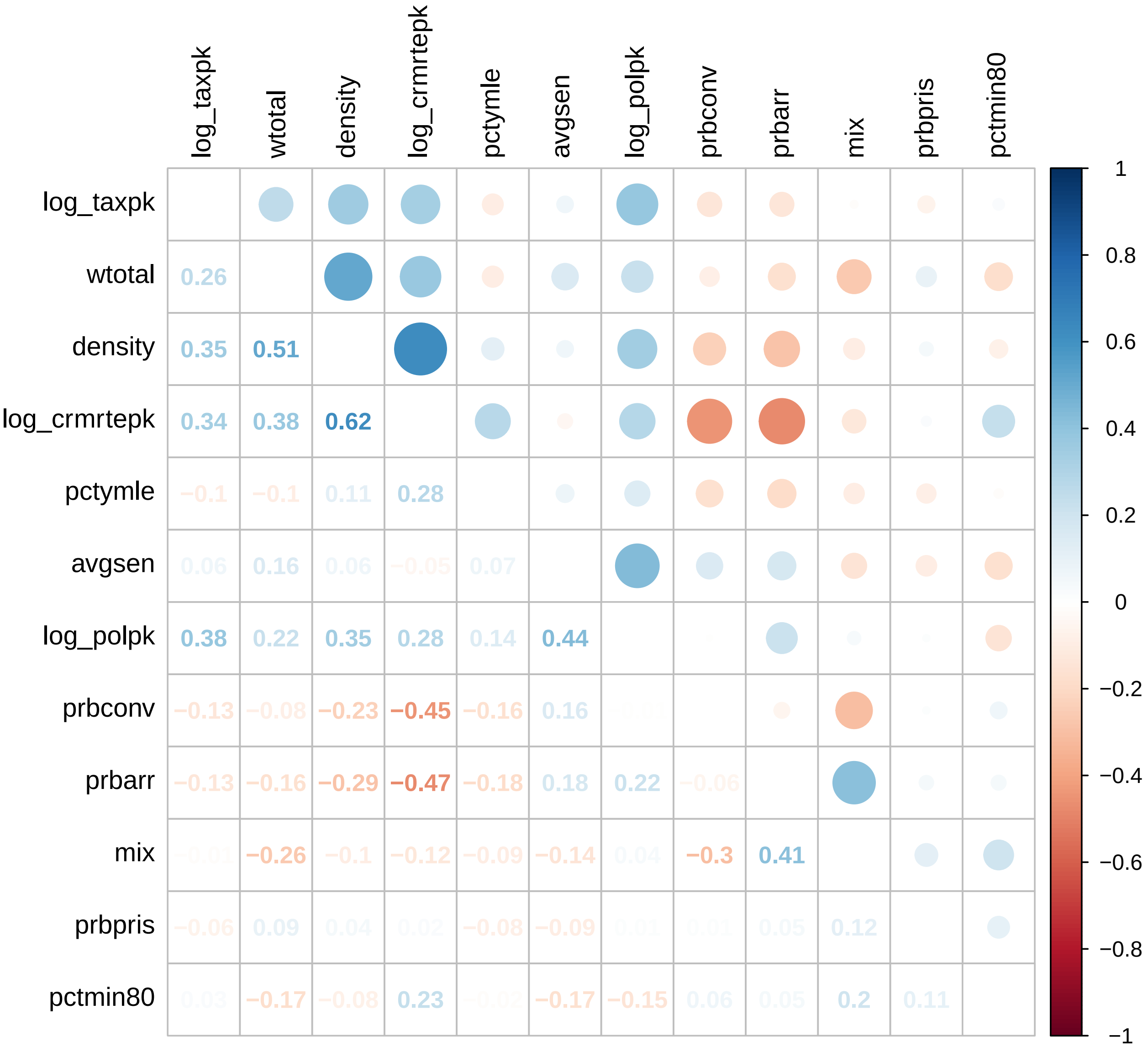
```
wages <- rbind(data.frame(wageType="wcon", wage=crimeData$wcon, meanWage=mean(crimeData$wcon)),
               data.frame(wageType="wtuc", wage=crimeData$wtuc, meanWage=mean(crimeData$wtuc)),
               data.frame(wageType="wtrd", wage=crimeData$wtrd, meanWage=mean(crimeData$wtrd)),
               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.

```
corrplot.mixed(cor(crimeData[ , (names(crimeData) %in%
                                  c("log_crmrtepk", "prbarr", "prbconv", "prbpris", "avgsen"
                                    , "log_polpk", "density", "log_taxpk", "pctmin80", "wtotal"
                                    , "mix", "pctymle"))]),
               tl.pos = "lt", tl.col="black", order="hclust", number.cex=.9, number.digits=2)
```



We can see *strong positive* correlation (>.25) between

1. crime rate (log_crmrtepk) and population density (density), total wages (wtotal) and taxes (log_taxpk). Which is logical in the sense that as population density increases, 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 and taxes, so we can take only one of these variables in our model to avoid multi co-linearity.

2. crime rate (log_crmrtepk) and % of male population (pctymle). Historically it has been observed that men are involved in crime more than woman. So this relation makes logical sense. As male population increases there is higher number of criminal activity and hence more crime rate in that area.

On the opposite side, we can see *strong negative* correlation (< -.25) 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.

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 percent of young male (pctymle) 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.

Given that, we're recommending the following model:

$$model1 = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot prbarr + \beta_3 \cdot prbconv + \beta_4 \cdot log_polpk$$

```
#crimeData <- crimeData[-c(51),  ]

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

```
##
## Call:
## lm(formula = log_crmrtepk ~ density + prbarr + prbconv + pctymle,
##     data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91562 -0.18500 -0.00608  0.22506  0.88656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.63178    0.20837  17.429  < 2e-16 ***
## density      0.15514    0.02629   5.902 7.13e-08 ***
## prbarr      -1.39431    0.28594  -4.876 4.98e-06 ***
## prbconv     -0.53909    0.10930  -4.932 3.99e-06 ***
## pctymle      2.59006    1.62086   1.598    0.114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3471 on 85 degrees of freedom
## Multiple R-squared:  0.6178, Adjusted R-squared:  0.5999
## F-statistic: 34.36 on 4 and 85 DF,  p-value: < 2.2e-16
```

```
cov1 <- vcovHC(model1, type = "HC")
robust.se1 <- sqrt(diag(cov1))
```

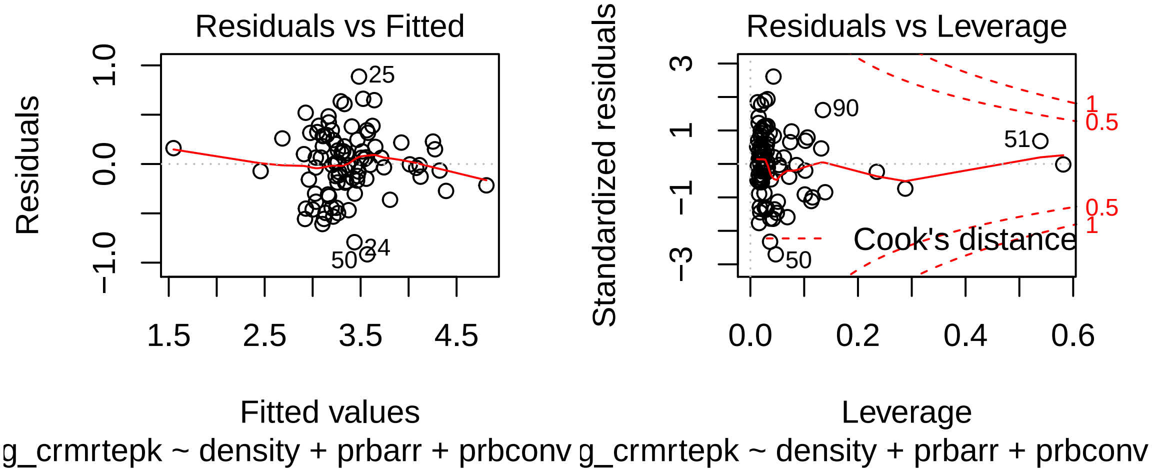
TODO: mix shows weak negative correlation with wage. See detailed correlation matrix for mix vs all wages and see if we can find something.

Model1 CLM Assumptions Analysis

CLM.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 the outlier (#51 to be precise) did not cross the Cook's distance, which is a good sign that the outlier influence is limited.

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

CLM.3 - Linear in Parameters

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

CLM.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.

CLM.4 - Multicollinearity

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

```
vif(model1)
```

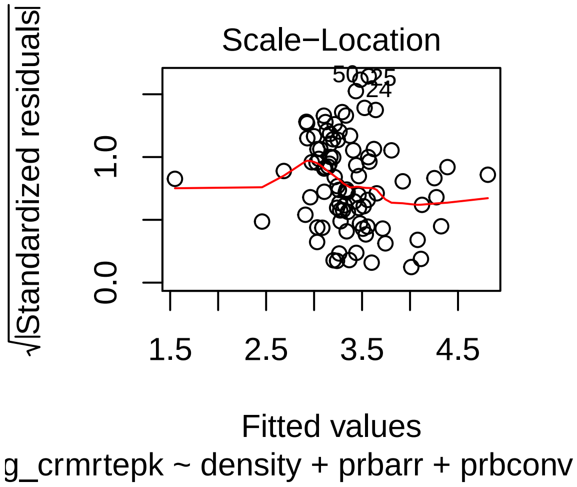
```
## density   prbarr  prbconv  pctymle
## 1.169543 1.144704 1.107103 1.067067
```

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

CLM.5 - Homoskedasticity

We will generate the Scale-Location plot to asses if the model meets the assumption.

```
plot(model1, which=3)
```



It is not clear whether we have met, or violated the assumption, so we'll use robust standard error to address any possible heteroskedasticity.

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

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3.631776   0.206876 17.5553 < 2.2e-16 ***
## density      0.155144   0.024023  6.4581 6.292e-09 ***
## prbarr      -1.394310   0.311464 -4.4766 2.340e-05 ***
## prbconv     -0.539095   0.130291 -4.1376 8.220e-05 ***
## pctymle      2.590063   1.021138  2.5364   0.01303 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

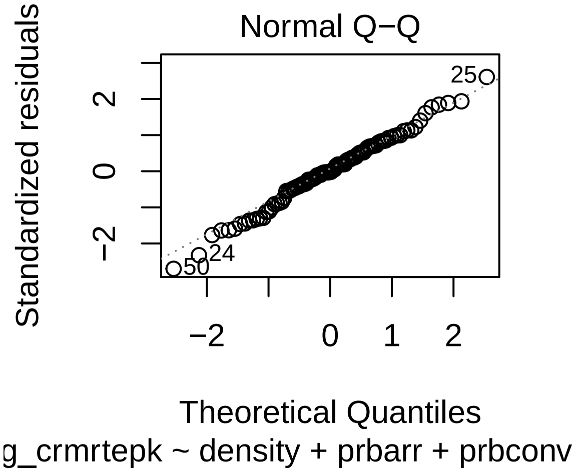
```
vcovHC(model1)
```

```
##              (Intercept)       density       prbarr     prbconv
## (Intercept)  0.042797825 -0.0032764687 -0.045257064 -0.01927407
## density     -0.003276469  0.0005771055  0.003303294  0.00172675
## prbarr      -0.045257064  0.0033032937  0.097009576  0.01118936
## prbconv     -0.019274073  0.0017267502  0.011189362  0.01697577
## pctymle     -0.131295080  0.0019687867  0.057584244  0.03644985
##                  pctymle
## (Intercept) -0.131295080
## density      0.001968787
## prbarr       0.057584244
## prbconv      0.036449849
## pctymle      1.042722789
```

CLM.6 - Normality of Residuals

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

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



We can notice few outliers at each end, but we're still considering this condition is met.

Model 2: with key explanatory variables and only covariates

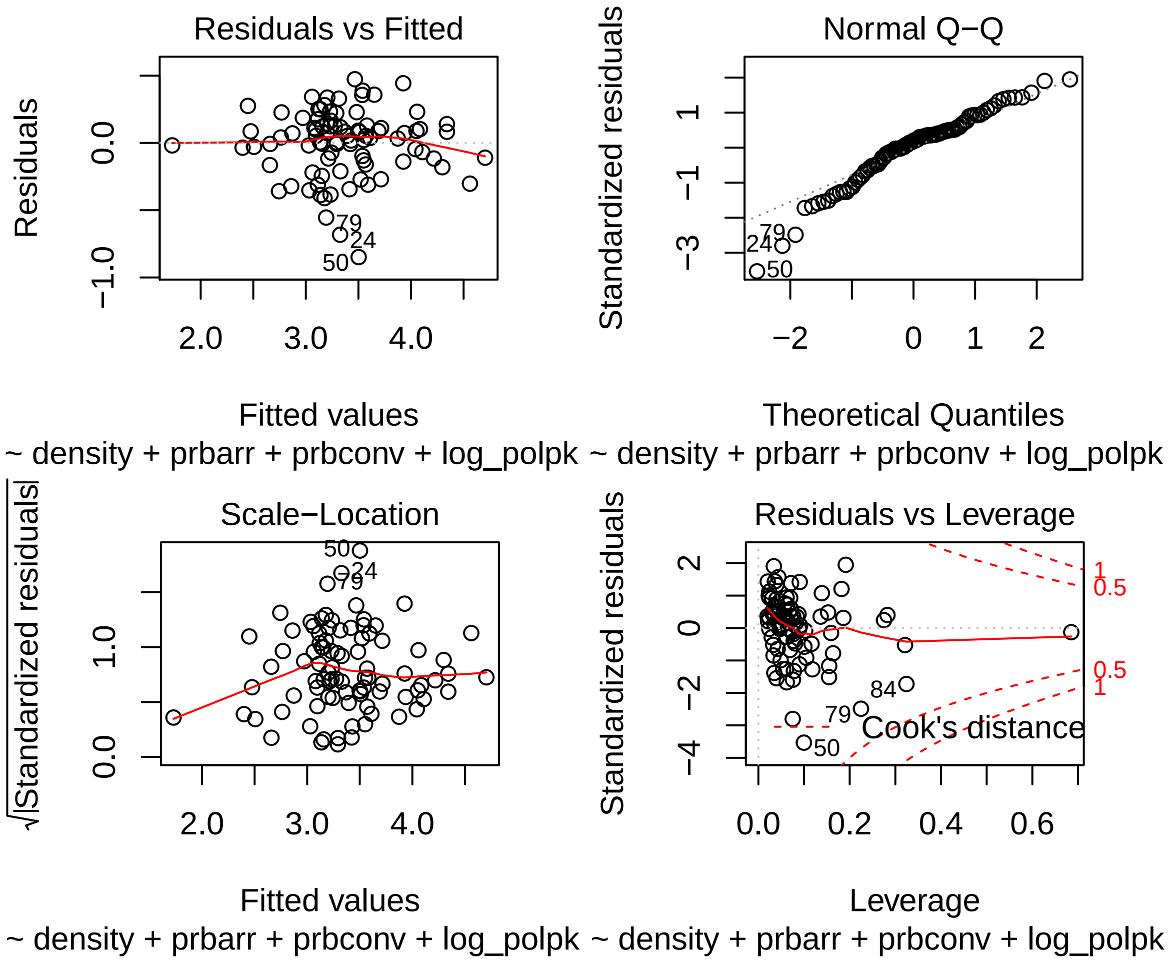
In this model, we'll include the variables (avgsen, mix, pctmin80), as we think they will contribute to the accuracy of your results without introducing substantial bias. These variables show good degree of correlation with crime rate, as well as, linear relationship as observed in the EDA.

$$crimeDeterm = \beta_0+\beta_1\cdot density+\beta_2\cdot prbarr+\beta_3\cdot prbconv+\beta_4\cdot log_polpk+\beta_5\cdot avgsen+\beta_6 \cot mix++\beta_7\cdot pctmin80$$

```
model2 <- lm(log_crmrtepk ~ density + prbarr + prbconv + log_polpk
             + avgsen + mix + pctmin80, data=crimeData)
summary(model2)
```

```
##
## Call:
## lm(formula = log_crmrtepk ~ density + prbarr + prbconv + log_polpk +
##     avgsen + mix + pctmin80, data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84848 -0.12600  0.04642  0.12824  0.47374
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.777008   0.140478  26.887  < 2e-16 ***
## density      0.099825   0.021550   4.632 1.34e-05 ***
## prbarr      -1.792313   0.240353  -7.457 8.23e-11 ***
## prbconv     -0.728017   0.084852  -8.580 4.93e-13 ***
## log_polpk    0.521299   0.090062   5.788 1.26e-07 ***
## avgsen      -0.004998   0.011027  -0.453   0.6515
## mix         -1.001898   0.395775  -2.531   0.0133 *
## pctmin80     0.012360   0.001657   7.457 8.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2531 on 82 degrees of freedom
## Multiple R-squared:  0.804,  Adjusted R-squared:  0.7873
## F-statistic: 48.05 on 7 and 82 DF,  p-value: < 2.2e-16
```

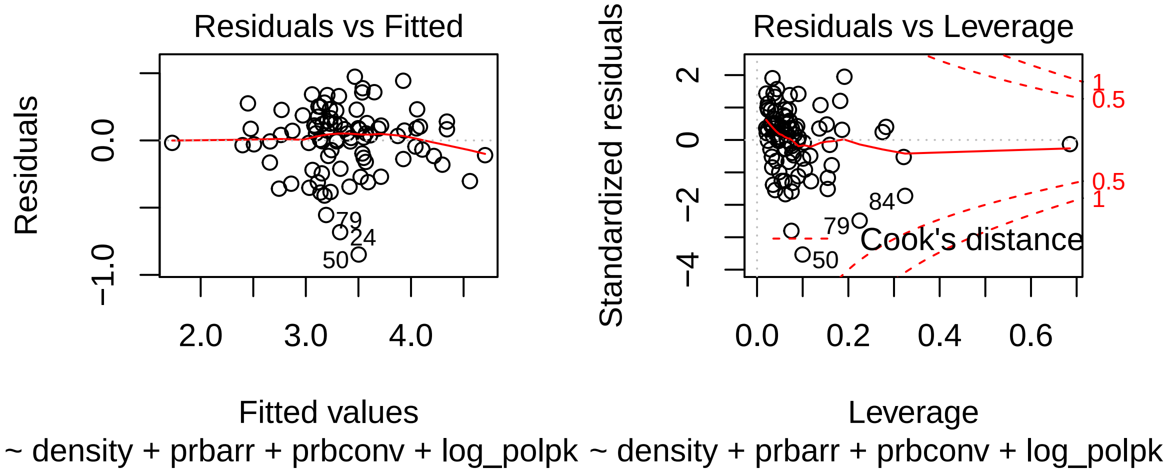
```
plot(model2)
cov2 <- vcovHC(model2, type = "HC")
robust.se2 <- sqrt(diag(cov2))
```



Model2 CLM Assumptions Analysis

CLM.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.

CLM.3 - Linear in Parameters

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

CLM.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.

CLM.4 - Multicollinearity

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

```
vif(model2)
```

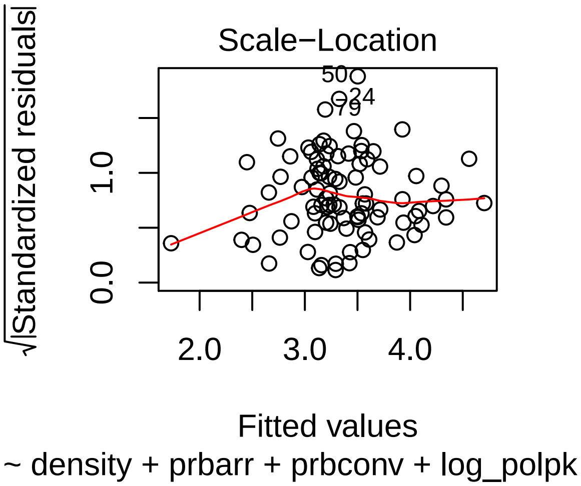
```
##   density    prbarr   prbconv log_polpk    avgsen       mix  pctmin80 
##  1.478657  1.521263  1.254915  1.578312  1.356970  1.454871  1.100927
```

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

CLM.5 - Homoskedasticity

We will generate the Scale-Location plot to asses if the model meets the assumption.

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



It is not clear whether we have met, or violated the assumption, so we'll use robust standard error to address any possible heteroskedasticity.

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

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3.7770082  0.1916772 19.7050 < 2.2e-16 ***
## density      0.0998253  0.0242604  4.1147 9.173e-05 ***
## prbarr      -1.7923132  0.1836667 -9.7585 2.243e-15 ***
## prbconv     -0.7280166  0.1234874 -5.8955 8.002e-08 ***
## log_polpk    0.5212994  0.1168590  4.4609 2.573e-05 ***
## avgsen      -0.0049982  0.0139337 -0.3587   0.72073
## mix         -1.0018979  0.4998199 -2.0045   0.04831 *
## pctmin80     0.0123597  0.0015950  7.7492 2.189e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

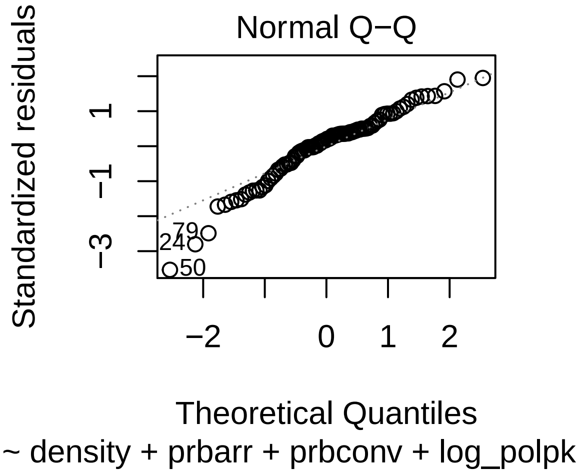
```
vcovHC(model2)
```

```
##               (Intercept)       density        prbarr       prbconv
## (Intercept)  0.0367401384 -3.257539e-03 -1.341637e-02 -1.158992e-02
## density     -0.0032575391  5.885666e-04  2.586156e-03  1.705387e-03
## prbarr      -0.0134163732  2.586156e-03  3.373345e-02  5.141196e-03
## prbconv     -0.0115899162  1.705387e-03  5.141196e-03  1.524914e-02
## log_polpk    0.0095989938 -1.662945e-03 -1.052299e-02 -2.765677e-03
## avgsen      -0.0013997553  4.549240e-05 -1.547131e-04 -4.813974e-04
## mix         -0.0432194988  2.629025e-03  3.101812e-03  2.106347e-02
## pctmin80    -0.0001652245  1.517918e-05  5.884565e-05  6.932898e-05
##                 log_polpk        avgsen           mix      pctmin80
## (Intercept)  9.598994e-03 -1.399755e-03 -4.321950e-02 -1.652245e-04
## density     -1.662945e-03  4.549240e-05  2.629025e-03  1.517918e-05
## prbarr      -1.052299e-02 -1.547131e-04  3.101812e-03  5.884565e-05
## prbconv     -2.765677e-03 -4.813974e-04  2.106347e-02  6.932898e-05
## log_polpk    1.365604e-02 -8.075368e-04  1.151516e-02 -5.972991e-05
## avgsen      -8.075368e-04  1.941481e-04 -8.390171e-04  5.525685e-06
## mix          1.151516e-02 -8.390171e-04  2.498199e-01 -9.019801e-05
## pctmin80    -5.972991e-05  5.525685e-06 -9.019801e-05  2.543949e-06
```

CLM.6 - Normality of Residuals

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

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



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

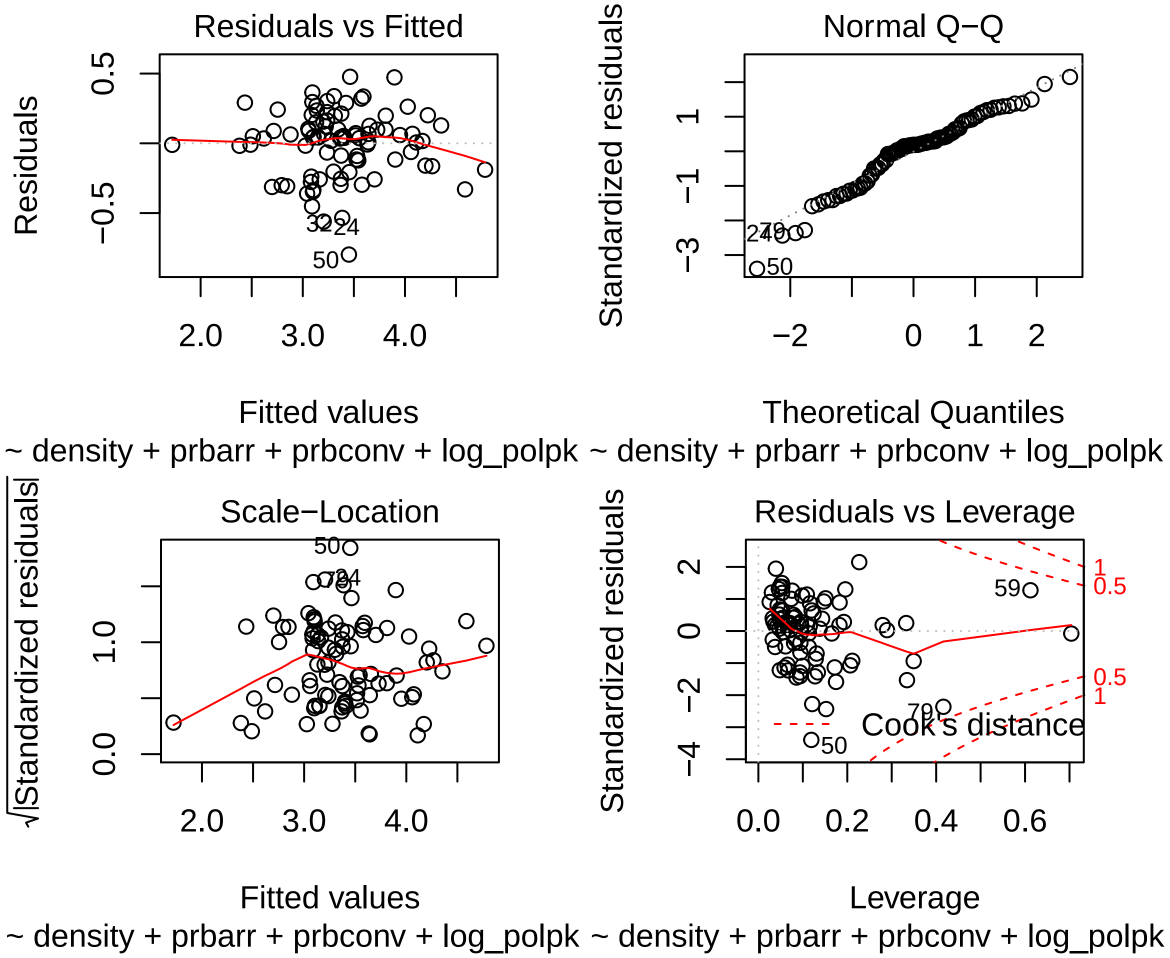
In this model, we'll include all the data available to us to demonstrate the robustness of results to model specification.

$$crimeDeterm = \beta_0+\beta_1\cdot density+\beta_2\cdot prbarr+\beta_3\cdot prbconv+\beta_4\cdot pctymle+\beta_5\cdot avgsen+\beta_6\cdot mix+\beta_7\cdot pctmin80+\beta_8\cdot prbpris+\beta_9\cdot log_p$$

```
model3 <- lm(log_crmrtepk ~ density + prbarr + prbconv + log_polpk
             + avgsen + mix + pctmin80
             + pctymle + prbpris  + wtotal, data=crimeData)
summary(model3)
```

```
##
## Call:
## lm(formula = log_crmrtepk ~ density + prbarr + prbconv + log_polpk +
##     avgsen + mix + pctmin80 + pctymle + prbpris + wtotal, data = crimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79756 -0.15063  0.04218  0.14503  0.47688
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.1638659  0.3310823   9.556 8.01e-15 ***
## density      0.0800850  0.0236374   3.388   0.0011 **
## prbarr      -1.7531408  0.2429020  -7.217 2.86e-10 ***
## prbconv     -0.6952287  0.0865489  -8.033 7.53e-12 ***
## log_polpk    0.4997279  0.0905507   5.519 4.20e-07 ***
## avgsen      -0.0078086  0.0110223  -0.708   0.4808
## mix         -0.7735857  0.4063833  -1.904   0.0606 .
## pctmin80     0.0125679  0.0016516   7.609 5.01e-11 ***
## pctymle      1.5141320  1.2355666   1.225   0.2240
## prbpris     -0.0294291  0.3380535  -0.087   0.9308
## wtotal       0.0005502  0.0002640   2.084   0.0404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2501 on 79 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7922
## F-statistic: 34.93 on 10 and 79 DF,  p-value: < 2.2e-16
```

```
plot(model3)
cov3 <- vcovHC(model3, type = "HC")
robust.se3 <- sqrt(diag(cov3))
```



Model3 CLM Assumptions Analysis

CLM.1 - Linear in Parameters

CLM.2 - Random Sampling

CLM.3 - Multicollinearity

CLM.4 - Zero Conditional Mean

CLM.5 - Homoskedasticity

CLM.6 - Normality of Residuals

All 3 Regression models at a glance

```r
#stargazer(model1, model2, model3, se = list(robust.se1,robust.se2, robust.se3),
#          dep.var.labels = "Log of Crime Rate per 1000 People",
#          order = c("prbarr", "prbconv", "density"), single.row = T, align = T,no.space=T,
#          title="Comparison of 3 Regression models", float=FALSE, header = FALSE)

stargazer(model1, model2, model3, se = list(robust.se1,robust.se2, robust.se3)
          ,dep.var.labels = "Log of Crime Rate per 1000 People"
          ,order = c("prbarr", "prbconv", "density"), no.space=TRUE, single.row = TRUE
          ,title="Comparison of 3 Regression models", float=FALSE, header = FALSE)
```

	Dependent variable:		
	Log of Crime Rate per 1000 People		
	(1)	(2)	(3)
prbarr	−1.394*** (0.249)	−1.792*** (0.162)	−1.753*** (0.171)
prbconv	−0.539*** (0.116)	−0.728*** (0.099)	−0.695*** (0.099)
density	0.155*** (0.021)	0.100*** (0.021)	0.080*** (0.022)
pctymle	2.590*** (0.927)		1.514 (1.073)
prbpris			−0.029 (0.374)
wtotal			0.001* (0.0003)
log_polpk		0.521*** (0.099)	0.500*** (0.102)
avgsen		−0.005 (0.012)	−0.008 (0.011)
mix		−1.002** (0.420)	−0.774* (0.399)
pctmin80		0.012*** (0.001)	0.013*** (0.001)
Constant	3.632*** (0.179)	3.777*** (0.172)	3.164*** (0.415)
Observations	90	90	90
R^2	0.618	0.804	0.816
Adjusted R^2	0.600	0.787	0.792
Residual Std. Error	0.347 (df = 85)	0.253 (df = 82)	0.250 (df = 79)
F Statistic	34.356*** (df = 4; 85)	48.051*** (df = 7; 82)	34.933*** (df = 10; 79)
Note:			*p<0.1; **p<0.05; ***p<0.01

TODO: change stargazer type to latex

Note that density coefficient drastically changes in model 3 because variables like wtotal and log_polpk distribute are synonymous to increase in density and they absorb some of the causality of density.

TODO: Talk about Q-Q plot and fitment and slight distortion due to outliers.

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.

TODO: Also talk about small conclusions drawn from EDA which did not reach to regression model (outliers etc)