W203 Lab 3: Reducing Crime Analysis

Chitra Agastya, Dean Wang, Katayoun Borojerdi

12/10/2018

Abstract

The intent of the following report is to explore the determinants of crime in order to provide insights to be used by a political campaign for policy recommendations using OLS regression analysis.

Our recommendation supported by the analysis setforth, to the politician for the political campaign we have been hired by, is to get tougher on crime through:

(a) more arrests and convictions and
(b) seeking immigration reform.

Introduction

As we begin to shed light on this important topic, we recognize crime is complicated, with an extensive range of factors, many of which may be interrelated. For this preliminary OLS regression analysis we will primarily use the 25 variables that we have been provided but recognize there may be other important omitted variables, that we currently do not have access to. Lastly our research will initially set out to explore various relationships between crime and other socioeconomic or demographic factors and lead to a model building process that will highlight the relationships which the political campaign should be focused on in developing effective policy to target crime and ultimately make communities in North Carolina safer.

The Data

We first load the dataset, 'crime_v2.csv', using the read.csv command into a dataset called 'crimedata' and inspect it for correctness, completeness and potential anomalies. To begin with, we can note that the 'crimedata' has 97 observations and 25 variables.

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
crimedata <- read.csv('crime_v2.csv')
cat("Total rows of data:", nrow(crimedata))
```

Total rows of data: 97

```
stargazer(crimedata, type="latex", header=FALSE, title="Summary of crimedata")
```

```
cat('\\newpage')
```

Table 1: Summary of crimedata

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
county	91	101.615	58.794	1.000	52.000	152.000	197.000
year	91	87.000	0.000	87.000	87.000	87.000	87.000
crmrte	91	0.033	0.019	0.006	0.021	0.040	0.099
prbarr	91	0.295	0.137	0.093	0.206	0.344	1.091
prbpris	91	0.411	0.080	0.150	0.365	0.457	0.600
avgsen	91	9.647	2.847	5.380	7.340	11.420	20.700
polpc	91	0.002	0.001	0.001	0.001	0.002	0.009
density	91	1.429	1.514	0.00002	0.547	1.568	8.828
taxpc	91	38.055	13.078	25.693	30.662	40.948	119.761
west	91	0.253	0.437	0.000	0.000	0.500	1.000
central	91	0.374	0.486	0.000	0.000	1.000	1.000
urban	91	0.088	0.285	0.000	0.000	0.000	1.000
pctmin80	91	25.495	17.017	1.284	9.845	38.142	64.348
wcon	91	285.358	47.487	193.643	250.782	314.795	436.767
wtuc	91	411.668	77.266	187.617	374.632	443.436	613.226
wtrd	91	211.553	34.216	154.209	190.864	225.126	354.676
wfir	91	322.098	53.890	170.940	286.527	345.354	509.466
wser	91	275.564	206.251	133.043	229.662	280.541	2,177.068
wmfg	91	335.589	87.841	157.410	288.875	359.580	646.850
wfed	91	442.901	59.678	326.100	400.240	478.030	597.950
wsta	91	357.522	43.103	258.330	329.325	382.590	499.590
wloc	91	312.681	28.235	239.170	297.265	329.250	388.090
mix	91	0.129	0.081	0.020	0.081	0.152	0.465
pctymle	91	0.084	0.023	0.062	0.074	0.083	0.249

Data Prepration

In searching for missing values we see that the values for rows 92-97 are all NA, as shown using the tail function below. Note, we know we caught all the missing values in the dataset because with the exception of *prbconv*, the other 24 variables all have NAs for the 6 rows so we know $24 * 6 = 144$ the number of missing values in the dataset. We decide to exclude these blank rows and use only rows 1 through 91 of crimedata as we continue to explore.

```
sum(is.na(crimedata))
```

```
## [1] 144
```

```
tail(crimedata[92:97,(1:8)])
```

```
##    county year crmrte prbarr prbconv prbpris avgsen polpc
## 92     NA   NA     NA     NA               NA     NA    NA
## 93     NA   NA     NA     NA               NA     NA    NA
## 94     NA   NA     NA     NA               NA     NA    NA
## 95     NA   NA     NA     NA               NA     NA    NA
## 96     NA   NA     NA     NA               NA     NA    NA
## 97     NA   NA     NA     NA        `      NA     NA    NA
```

Upon further inspection, we see that one of the rows (row # 89 of 'crimedata', associated with county ID 193) has been repeated. We eliminate this duplicated row. We notice that the values for columns *county* and *year* are numeric numeric instead of factor. We keep them this way and use a character representation of *county* in graphs. we also notice that that the values for the column probability of conviction (*prbconv*) are of type factor instead of numeric. We convert the column *prbconv* originally assigned as factor to numeric values.

```
data <- crimedata[1:91,]
data <- unique(data)
data$prbconv <- as.numeric(paste(data$prbconv))
```

```
cat("Number of unique rows",nrow(data), "\n")
```

```
## Number of unique rows 90
```

After our preliminary data preparation we now have 90 observations on 25 variables which we will be using for our Exploratory Data Analysis (EDA).

Anomalies in the Data

Before progressing further into the analysis we would like to list a number of anomalies that we recognized in the dataset and mention they will be further dealt with throughout our analysis. Note, based on the region information of the county from the dataset, we have strong reason to believe that the county ID used, is the FIPS code used by the US government to identify counties. With this assumption, we have further analysed the anomalies seen in the data. (source: https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina)

The anomalies we observe in our dataset are summarized below:

1. First off, there are 10 rows (~ 10% of our dataset) that have probability values greater than 1 and most of these values appear in the *prbconv* variable and in one case in *prbarr* value. We talk about plausible reasons for these in our EDA section. However despite these anomalies, we believe the rest of the data associated with these same rows will provide valuable information for our analysis and we have chosen to keep them rather than eliminate these rows. See below for rows with suspect probabilities.

```
data_invalid <- subset(data, prbconv > 1 | prbarr > 1 | prbpris > 1)
cat("Number of rows with invalid probability values: ", nrow(data_invalid), "\n")
```

```
## Number of rows with invalid probability values:  10
```

2. Percentage young male is at 25% for one data point, Onslow County (ID 133) in the North Carolina data provided. The city of Jacksonville, which is the commercial hub of Onslow county, is home to Marine Corps Base Camp Lejeune and Marine Corps Air Station New River. This explains why the percentage of young male is much higher in Onslow county compared to the rest of the counties in our data frame. (source: https://www.military.com/base-guide/jacksonville-north-carolina-military-bases)
3. Police per capita for Madison County (ID 115) is 90 which is very high compared to the median police per capita which is at 14 and mean police per capita which is at 17.
4. Weekly wage for services in Warren County (ID 185) is 2177 USD, which is significantly higher than the mean (275 USD) and median (253 USD) weekly wages for the service industry. After some research of Warren County we do not believe there is any reason why this data is correct and will plan to replace this data point with the mean value for weekly wage of services industry jobs when used in our analysis.
5. taxpc for Dare County (ID 55) surpasses that of every other county in the dataset. Dare county is known for tourism and generates a large tax revenue due to visitor spending. We also notice that the crime in the county is quite high at 0.079 and the police per capita is also high (40 cops for every 10000 people) even though the population density is only 0.51, given the high floating population. And yet probability of arrest is only at 0.2, well below the average value. It is plausible that crimes are being committed by visitors thereby less likely to result in an arrest. (source: https://outerbanksvoice.com/2016/08/18/tourism-related-spending-in-dare-county-tops-1-billion-again/)

As we go into our EDA section on Analysis of key variables, we will be transforming these anomalies where it makes sense.

Criteria for Analysis

Now that we have thoroughly prepared our data and noted obvious anomalies in the data, we can begin to discuss the data analysis. Our goal is to find the determinants of crime, so that we can make policy suggestions that a political candidate can run on. To do this, we need to understand the key relationships between each of the variables provided and crimes committed per person *crmrte*, which we will be using as our dependent variable. By building a robust multiple regression model we believe we can quantify the relationships between *crmrte* and the other independent variables policy recommendations can be based on.

Now that we have established crimes committed per person *crmrte* is the outcome variable, we also want to consider which variables intuitively make sense to use as independent variables, in a regression model. Also when considering these variables, we will be keeping in mind the political campaign, as they have to be factors that can be feasibly influenced by a politician in office. For example, variables such as wage, police presence, probability of arrest, tax revenues are all attractive variables because they are features a politician can influence, by passing legislation when in office. Other factors such as density, percent minority, percentage male or average prison sentence are harder to build policy recommendations around, as they are not easily controlled by politicians and policy recommendations are more complex.

Lastly it's important to address the fact that, there are a number of additional desired variables that would be helpful in building a more accurate crime predicting regression model. Variables such as additional demographics data from the counties, crime types, violent versus non violent crime, education level of residents, employment data, as well as others would have been helpful in providing more depth to this analysis.

Hypothesis for Determinants of Crime

Before exploring the data and investigating any relationships between the dependent variable, *crmte* and the other variables, we will generate our hypothesis of which key elements to explore. We will systemically explore these variables, discuss any problematic relationships between them and other variables in the dataset.

Based on the 24 variables we are considering the key variables we believe we need to test the effects of in terms of crime rates are as follows:

1. Probability Of Arrest/Conviction/Prison and Average Sentence:

We believe actions taken by law enforcement like arrests, conviction, prison sentence and severity of punishment to deter time ($prbarr, prbconv, prbpris, avgsen$) will have an impact on crime

2. Police Presence:

We believe the right police presence will make a considerable impact as a deterrence of crime and believe those areas that are more heavily populated have more opportunity to benefit for the optimal police presence.

3. Wages:

We believe many types of crime can be tied to financial factors and therefore believe these variables will have an impact of crime rates.

Note, we recognize that other factors like density, tax per capita, percentage young male, percentage minority may affect the key variables above either as covariates or confounding variables. We have eliminated some variables from our analysis and we will explain the reason for that in the EDA section of this report.

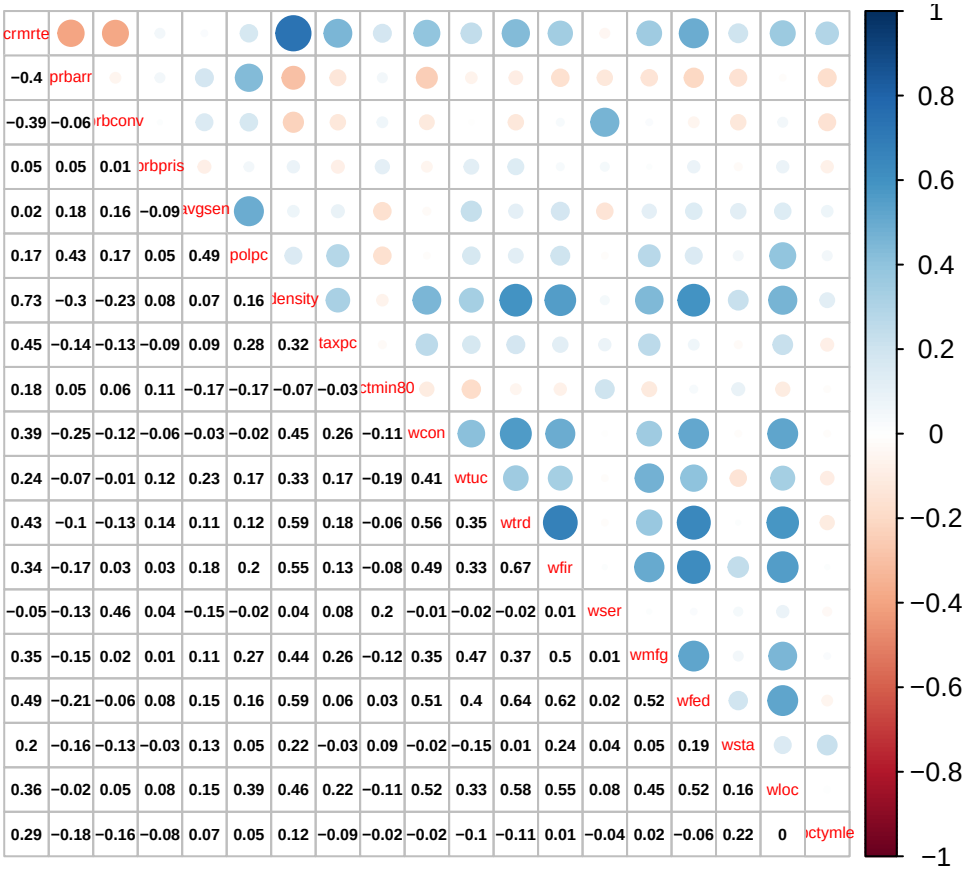
Exploratory Data Analysis

To initiate the EDA we take the correlation matrix of all variables of interest that we have mentioned in our hypothesis (including covariates and confounding variables) to show correlation with one another and especially with the predictor variable *crmrte*. Below we generate the correlation matrix for these variables of interest.

```
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
cor_matrix <- round(cor(data[c(3:10, 14:23, 25)]), 2)
corrplot.mixed(cor_matrix, lower.col = "black",
               number.cex = .5, tl.cex=0.5)
```



Quickly we can see that there is a strong positive correlation between *density* and *crmrte*, indicating that the more dense areas have high crime rates in our data. Additionally, there is a negative correlation between *prbarr* and *crmrte* and *prbconv* and *crmrte* but a positive correlation with *polpc*. This would indicate as probability of arrest goes up crimes rates are decreasing but that police presence is shown to be increasing. These correlations make intuitive sense but it is important to remind ourselves these can not be considered causation, rather artifacts of what we see in this sample data of counties in North Carolina provided.

Exploratory Analysis of Key Variables

```
# Function to summarize variable and show historgram
sumhist <- function(x, str, xlable) {
  print(summary(x))
  hist(x, main=paste("Histogram of",str), xlab=xlable, cex.main=0.8)
}

# Function for doing regression model
rmodel <- function (x, y,xlabel, ylabel)
{
  plot(x, y, xlab=xlabel, ylab=ylabel,
       main= paste(ylabel, " by ", xlabel), cex.main=0.8)
  abline(lm(y ~ x), col="blue")
  cat ("Correlation of", xlabel, "with", ylabel, ":", cor(x,y), "\n")
}
```

Crime Rate

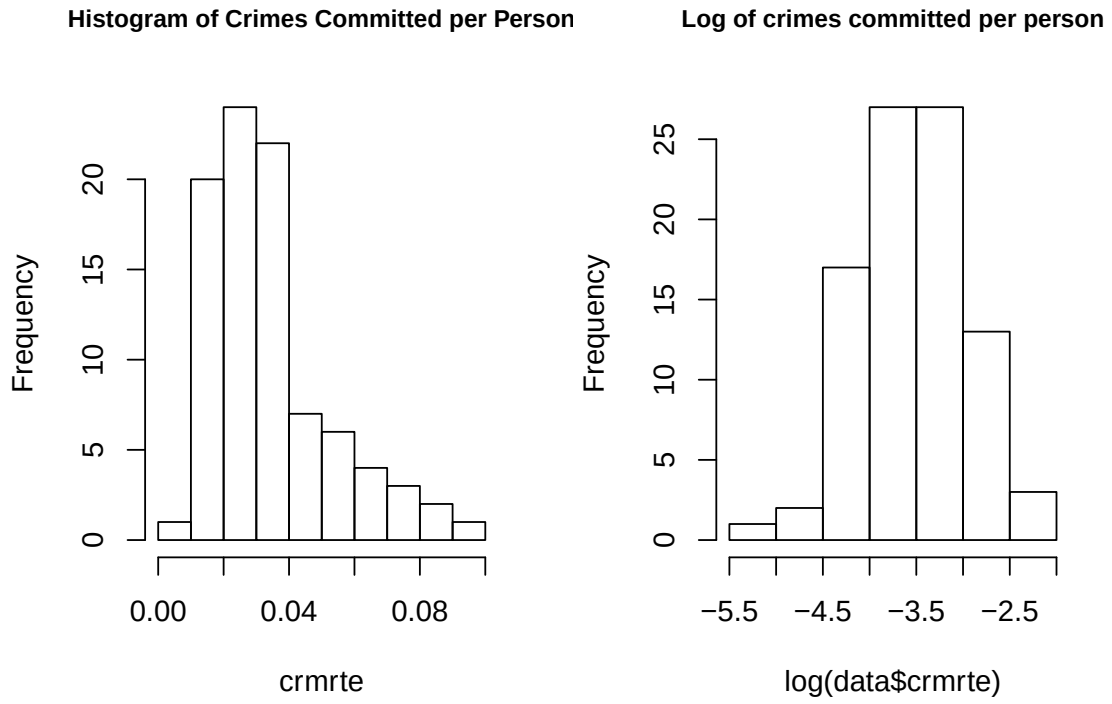
We begin with an exploratory analysis on the *crmrte* variable, which is described as "crimes committed per person" in a county. A benefit of measuring crimes per person rather than just absolute number of crimes in a county, is that a county will not have "more crime" simply by being larger. This normalization allows us to compare all counties

irrelevant of the total population. It is also important to note that all crimes committed provided in the dataset are lumped into one variable and we are assuming that various types of crimes that such as violent or more serious crimes are counted in the same category as nonviolent and less serious crimes. This nuance can be problematic as it does not allow us to do a more granular analysis to account for these differences in our model. It may be that the determinant of more violent or serious crimes are different in nature than those that are less serious or nonviolent.

```
par(mfrow=c(1,2))
sumhist(data$crmrte, "Crimes Committed per Person", "crmrte")
```

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

```
hist(log(data$crmrte), main= "Log of crimes committed per person", cex.main=0.8)
```



Looking at a summary of the numerical *crmrte* measure, we find that the median number of crimes per person is 0.03. The mean number of crimes is 0.03351, greater than the median number of crimes; this may indicate a right-skew. Indeed, looking at the histogram of *crmrte*, we find a long right tail, which means that this distribution is right-skewed.

We would expect this variable not to be less than 0, and the lowest value is 0.005533. There is also no evidence of a sudden increase in number of observations near 0. This means this distribution is not truncated significantly at 0. We take a log transformation to decrease the skew and maximize the spread and will help us infer the impact on crime rate better.

To summarize in practical terms, the range of crime committed data extends from 0 to approximately 1 crime per 10 people with the mean being about 3.3 crimes for every 100 number of people.

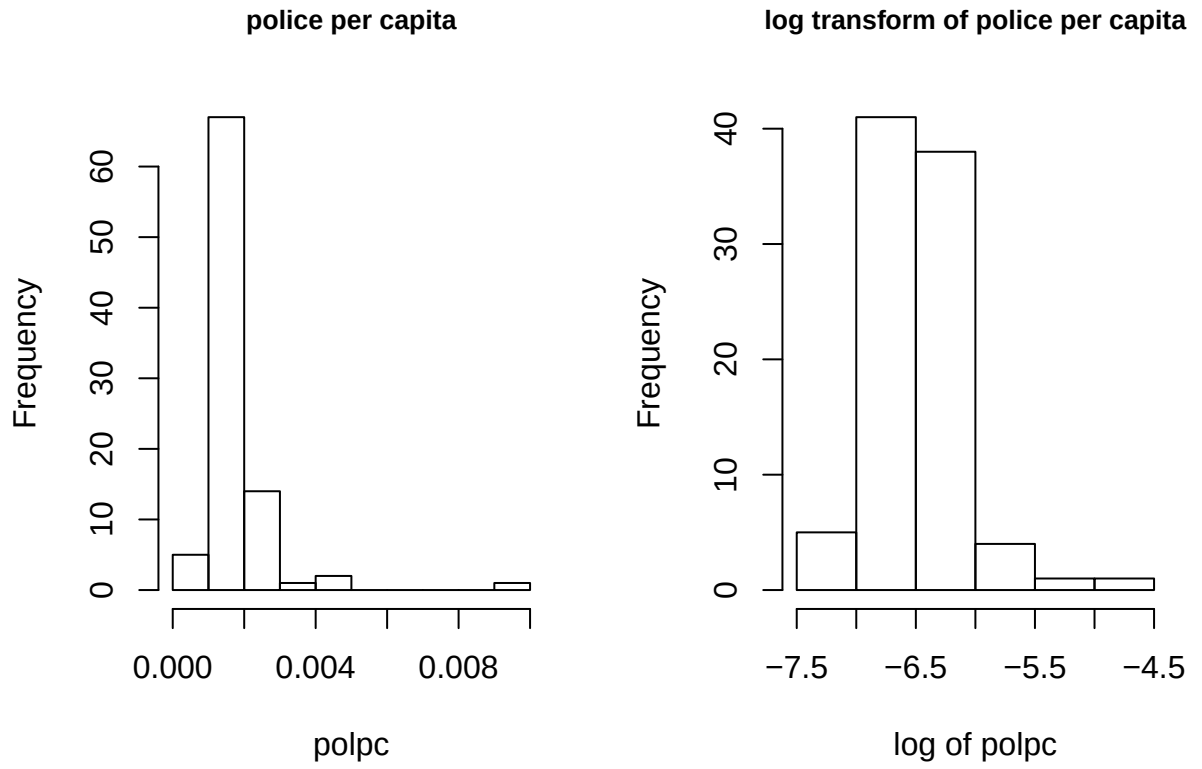
Police Presence

```
summary(data$polpc * 10000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.459  12.378  14.897  17.080  18.856  90.543
```

From the summary for the *polpc* we can see that the median police per capita for every 10000 people is 15 and the mean is at 17. The maximum value for this variable is at 90 which is significantly high.

```
par(mfrow=c(1,2))
hist((data$polpc), main="police per capita", xlab="polpc", cex.main=0.8)
hist(log(data$polpc),
     main="log transform of police per capita",
     xlab="log of polpc", cex.main=0.8)
```

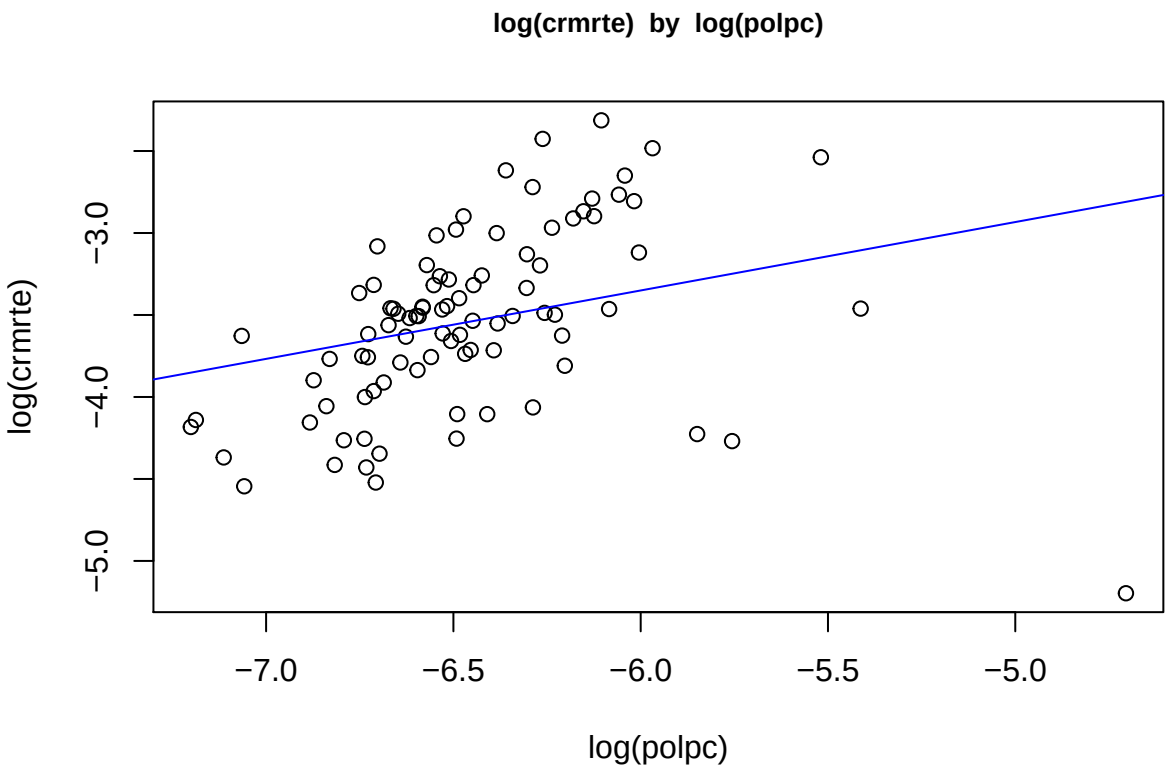



```
cat("Correlation of polpc with crmrte:", cor(data$polpc, data$crmrte), "\n")
```

```
## Correlation of polpc with crmrte: 0.1672816
```

The histogram of *polpc* shows a strong right skew. Applying a log transformation on this variable increases its correlation with our outcome variable *crmrte* thereby increasing the linear representation. This transformation makes sense because data distribution is heavily skewed to the right, due many counties having similar and lower police presence with a few counties, which for reason we will explore further into the analysis, have much higher police presence. So we use the log transform of this variable going forward.

```
rmodel(log(data$polpc), log(data$crmrte), "log(polpc)", "log(crmrte)")
```



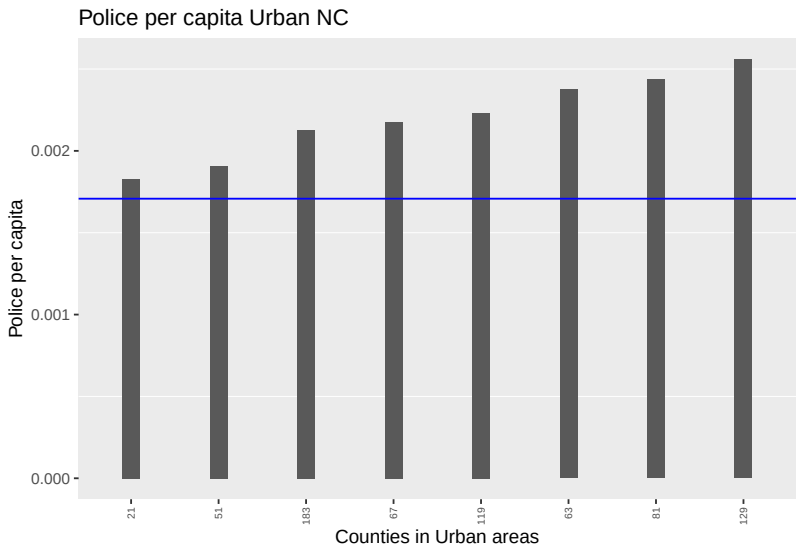

```
## Correlation of log(polpc) with log(crmrte) : 0.2845396
```

From the plot we can see that police per capita is positively correlated with crime rate indicating that there is more police when crime rate is high. This can be explained, as mentioned earlier, that more police is stationed in areas with higher crime rates with the goal of bringing the crime rates down.

Population density has a positive correlation of 0.35 with police per capita. It looks like population density might affect the decision on how many policemen are staffed in a county. We next look at the police per capita in urban counties.

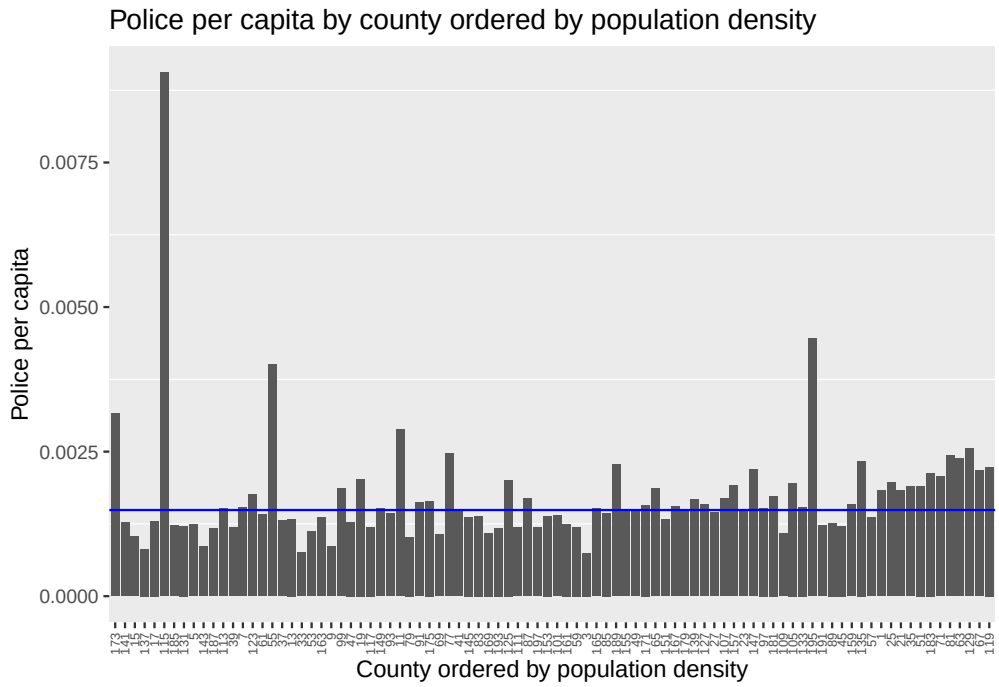
```
library(ggplot2)

urban <- subset(data, urban == 1)
ggplot(urban, aes(x=reorder(county, polpc), y=polpc)) +
  geom_bar(stat="identity", width=0.2) +
  geom_hline(yintercept = mean(data$polpc), color="blue") +
  #scale_x_discrete(breaks = data$county[c(T,F)]) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5, size=6),
        panel.grid.major = element_blank()) +
  labs(y="Police per capita", x = "Counties in Urban areas") + ggtitle("Police per capita Urban NC")
```



Counties in urban areas have higher population densities. The police per capita for all the urban counties is higher than the mean police per capita values. We next look at the county wise police per capita numbers. The graph below shows the police per capita numbers for counties ordered by population density.

```
# Police per capita based on population density
par(mfrow=c(2,1))
ggplot(data, aes(x=reorder(county, density), y=polpc)) +
  geom_bar(stat="identity") +
  geom_hline(yintercept = median(data$polpc), color="blue") +
  #scale_x_discrete(breaks = data$county[c(T,F)]) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5, size=6),
        panel.grid.major = element_blank()) +
  labs(y="Police per capita", x = "County ordered by population density") +
  ggtitle("Police per capita by county ordered by population density")
```




```
cat("Correlation between Tax Revenue per capita and Police per capita: ",
    cor(data$polpc, data$taxpc), "\n")
```

```
## Correlation between Tax Revenue per capita and Police per capita:  0.2805532
```

From the bar graph above, we can see that population totals don't fully reflect demands placed on law enforcement. Clearly, counties 115, 173 and 55 which have among of the higher police per capita are also among counties that have relatively lower population densities. While we see that many counties which have higher population densities have more police per capita than counties with lower population densities, this is not always true.

In addition to crime occurrences and population density, other factors like politically driven mandates, higher tax revenues, budgeting constraints etc. could have potential impact on the police staffing numbers. We do see a high correlation of 0.6 between tax revenue per capita and police per capita indicating counties with higher tax revenues appoint more police personnel. However, the data frame provided has limited to no insight into other factors that could potentially influence police per capita numbers and a follow up study to determine these factors would be recommended to help understand crime rates better.

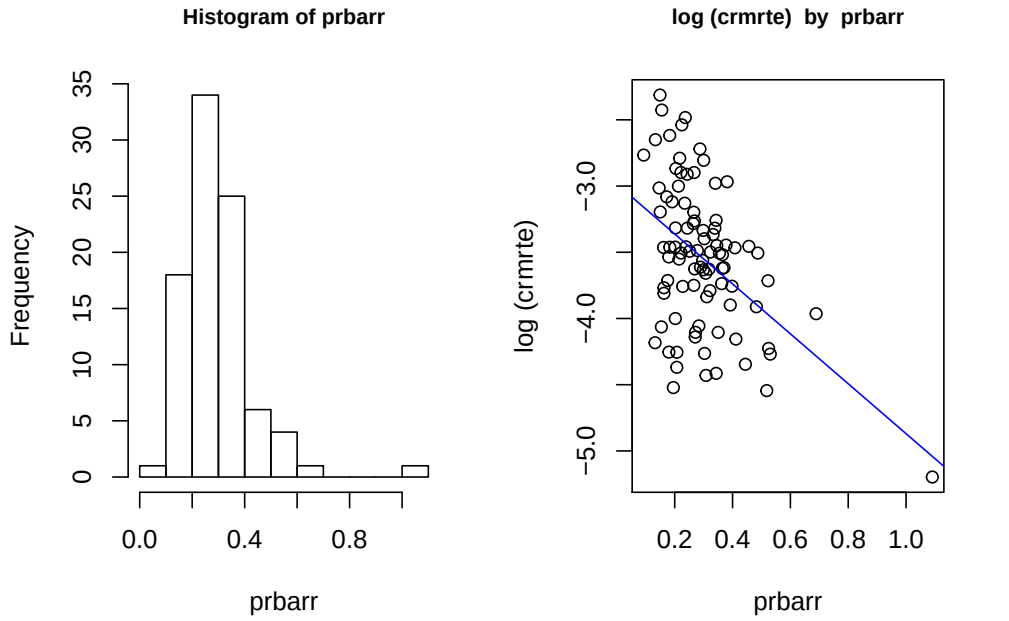
Probability of Arrest/Conviction/Prison and Average Sentence

In the data set provided we have four deterrents of crime. Of these correlation of probability of prison sentence with crime rate is 0.0479954 and correlation of average Sentence with crime rate is 0.01979653. Since these two variables are not really correlated much with our outcome variable we will not be including them in our regression models of interest. (i.e. model1 and model2)

Probability of arrest (i.e. ratio of arrests to offenses registered) and probability of conviction are variables of interest that are natural deterrents to crime occurrences. Higher probability values for *prbarr* and *prbconv* indicate that more perpetrators are being held accountable for crimes that have been recorded. In a way, these two variables act as a success indicators to evaluate how effective law enforcement is in controlling crime. Our intuition is that being tougher on crime will help bring the crime rate down.

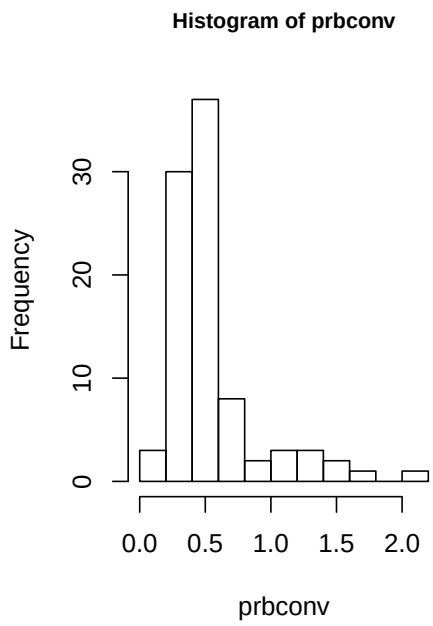
We next perform EDA on variables *prbarr* and *prbconv* to get a better understanding of their impact on crime rate in North Carolina.

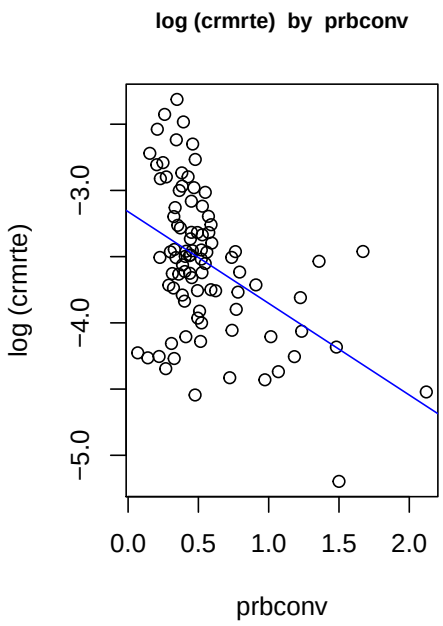
```
par(mfrow=c(1,2))
hist((data$prbarr), main="Histogram of prbarr", xlab="prbarr", cex.main=0.8)
rmodel((data$prbarr), log(data$crmrte), "prbarr", "log (crmrte)")
```




```
## Correlation of prbarr with log (crmrte) : -0.4727669
```

```
hist((data$prbconv), main="Histogram of prbconv", xlab="prbconv", cex.main=0.8)
rmodel((data$prbconv), log(data$crmrte), "prbconv", "log (crmrte)")
```






```
## Correlation of prbconv with log (crmrte) : -0.4468136
```

The histogram for probability of arrest shows a positive right skew. It has a negative correlation with crime rate. This indicates lower crime rates in areas with high probability of arrest.

The histogram of probability of conviction shows values ranging from 0 to 2. In fact 10 of the records have values greater than 1. We believe it is plausible that the conviction trials in some cases may have carried forward to subsequent years and are tallied in the year they are settled rather than the year the crime was registered thereby causing these artificially higher numbers. Another plausible reasoning could be that a perpetrator has been arrested after committing multiple recorded crimes and is being charged for all of those crimes within the same arrest. Probability of conviction has a negative correlation with crime rate.

```
cat("Cor between probabilities of arrest and conviction: ",
    cor(data$prbarr, data$prbconv), "\n")
```

```
## Cor between probabilities of arrest and conviction:  -0.05579621
```

We further notice that probabilities of arrest and conviction have very less correlation with each other. So we like to understand our outcome variable crmrte as a function of both of these predictor variables.

Wage

There are 9 different wage type categories specified in the dataset. We will look at a summary of these different wage categories in order to assess their values and consider how they may impact crime rate.

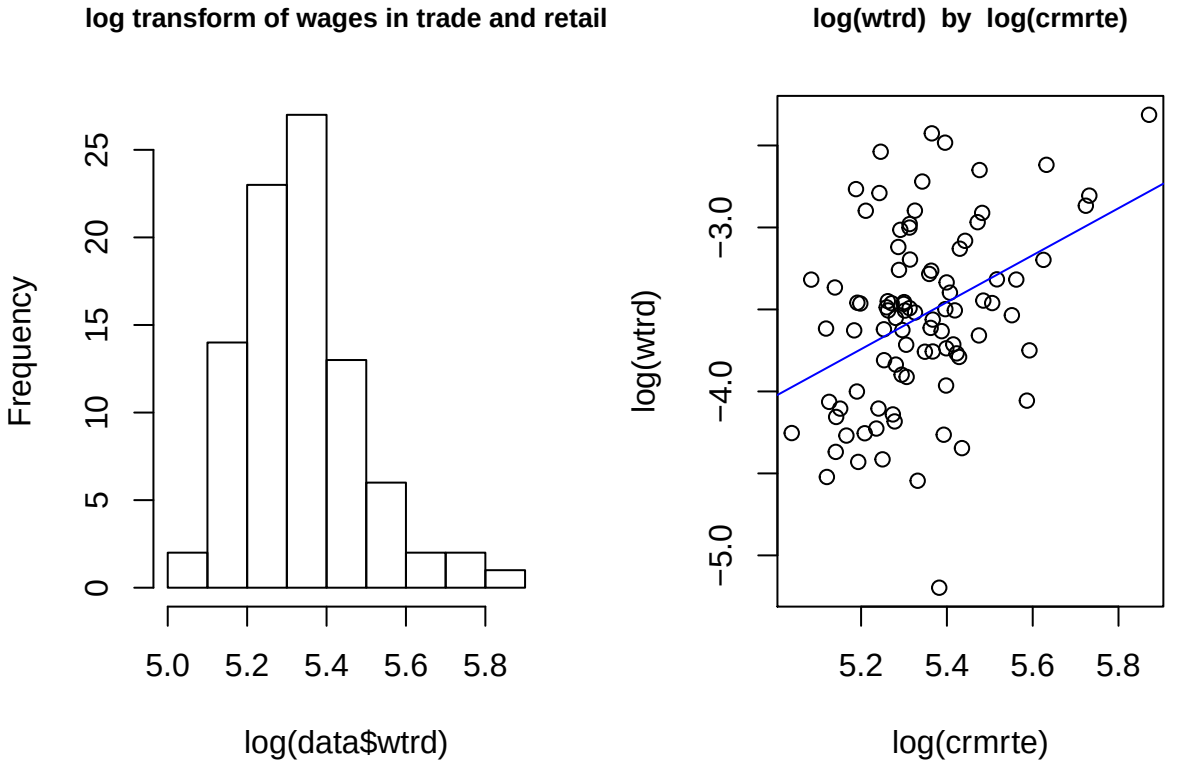
```
stargazer(crimedata[,15:23], header=FALSE, title = "Summary of all wages")
```

From the table above we can see that *wtrd*, wages for "wholesale and retail trade" has the lowest mean value. The standard deviation for this variable is also among the smallest out of the wage variables, indicating that workers in this category are consistently paid the lowest wages out of all the different categories. We predict that this group of lowest-paid workers is most at risk of committing crime. People in this group may have the least to lose and most to gain by committing crimes. Another reason why this group may be particularly susceptible to crime is that they themselves along with people around them may have lower education and skills. Our background knowledge related to crime leads us to believe that educated people tend to commit less crime, so less educated people would commit more crime in addition to having low wages. We will use this *wtrd* variable as our proxy for wage because based on our research we believe low wages could have a direct impact on crime rate. We choose this variable in particular since it has the lowest mean. Next, we consider the histogram and scatter plot of *wtrd* and *crmrte*.

Table 2: Summary of all wages

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
wcon	91	285.358	47.487	193.643	250.782	314.795	436.767
wtuc	91	411.668	77.266	187.617	374.632	443.436	613.226
wtrd	91	211.553	34.216	154.209	190.864	225.126	354.676
wfir	91	322.098	53.890	170.940	286.527	345.354	509.466
wser	91	275.564	206.251	133.043	229.662	280.541	2,177.068
wmfg	91	335.589	87.841	157.410	288.875	359.580	646.850
wfed	91	442.901	59.678	326.100	400.240	478.030	597.950
wsta	91	357.522	43.103	258.330	329.325	382.590	499.590
wloc	91	312.681	28.235	239.170	297.265	329.250	388.090

```
par(mfrow = c(1,2))
hist(log(data$wtrd), main = "log transform of wages in trade and retail",
     cex.main=0.8)
rmodel(log(data$wtrd), log(data$crmrte), "log(crmrte)", "log(wtrd)")
```




```
## Correlation of log(crmrte) with log(wtrd) : 0.3896589
```

```
cor(log(data$wtrd), log(data$crmrte))
```

```
## [1] 0.3896589
```

We take the log transform of the *wtrd* wage variable, which is a standard practice in the labor economics literature to get the wage elasticity. Another effect of taking the log transformation is that it helps to de-skew the data, which may make our residuals more normal. From the scatter plot, we can see that the log-transformed *wtrd* variable has a positive correlation (0.39) with the log-transformed *crmrte* (crime rate) variable. One plausible explanation for this

could be that people earning less wages may be taking to crime. Another plausible reasoning could be that these people are paid in cash and are potential crime targets. We are not trying to infer causality at this point. However, it is important to note our outcome variable seems to be a function of wage making this a key variable of interest. Also with a campaign policy targeted towards low wages, this group might be of interest to a political campaign.

Going forward we will not be considering the other wage variables in our first two modeling as they all seem to be more or less comparable in values and might not be and ideal fit for a political campaign.

Covariates

We will next look at variables in the data set which are not easy to influence through a political campaign (e.g. density, percentage young male etc.) and at the same time may have correlation with the predictor variables and/or with the outcome variables.

Density

The *density* variable, which is represented by people per square mile, is a relevant variable to consider for this analysis, as it has a strong correlation with the outcome variable. Also intuitively this makes sense because it is long known that many urban areas are generally associated with higher crime rates and are in need of more police support, as there is a higher density of people. Additionally, we find it is collinear with variables provided such as urban, western and central NC and will therefore not pursue those, in order to eliminate introducing multicollinearity to the model. Also because of their categorical binary nature (0 and 1s), we felt that using the density variable was sufficient in capturing the impact of urban vs rural crime characteristics in a stepwise, continuous data perspective.

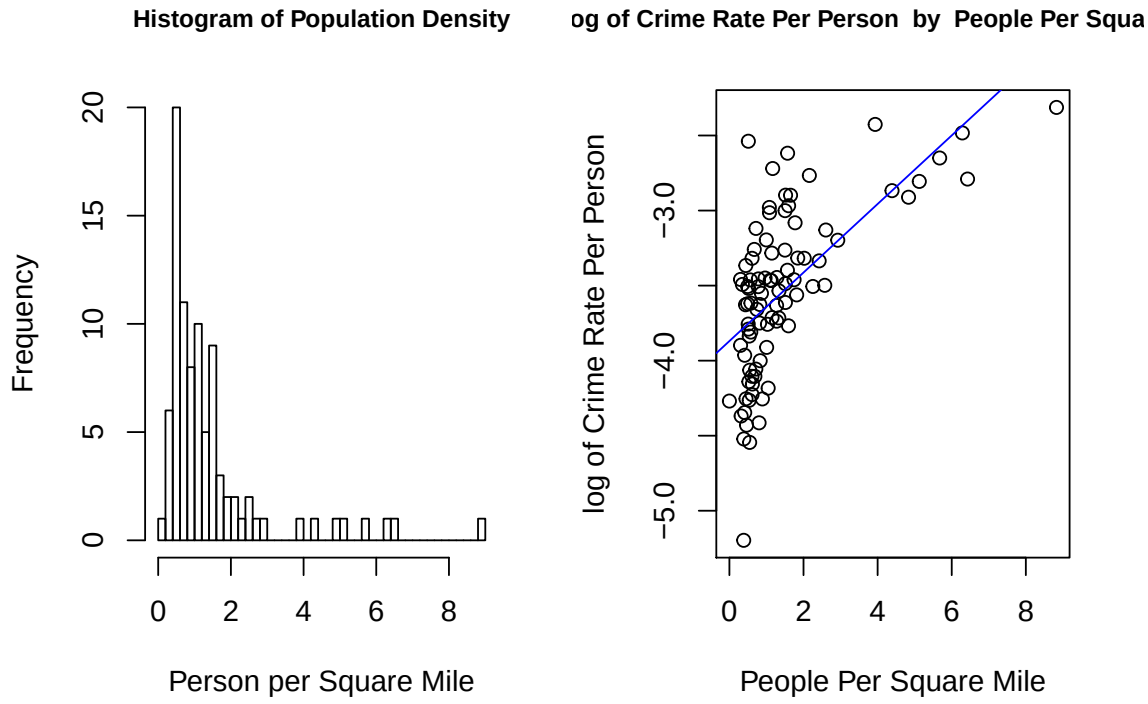
```
summary(data$density)
```

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

Taking a quick look, using the summary function, the density variable has a minimum value of 0.00002 and maximum value of almost 9, with a mean of approximately 1.5 indicating we expect the data distribution to have a right skew. It is important to note that we believe the 0.00002 data point is going to be problematic because it is essentially zero and suspect of being a potential error. Next creating a histogram of the density variable you can see that in fact the data does have a major right skew indicating that the vast majority of the North Carolina county data are rural areas as we suspected.

Now, let's plot the *density* variable against *crmrte* to investigate further.

```
par(mfrow=c(1,2))
hist(data$density, main = "Histogram of Population Density",
     xlab = "Person per Square Mile", breaks = 50, cex.main =0.8)
rmodel (data$density, log(data$crmrte), "People Per Square Mile",
        "log of Crime Rate Per Person")
```




```
## Correlation of People Per Square Mile with log of Crime Rate Per Person : 0.6330234
```

The correlation between the outcome variable, crimes committed per person *crmrte* to people per square mile, *density* has a strong positive correlation of 0.72, which can also clearly be seen in the plot. Another observation we can make from the plot, is that the majority of the data is clustered in the 0 to 3 *density* pat of the x axis of the graph with much fewer data point in the what Would be more considered urban areas and therefore we need to test for influence of those data points on the slope of this line.

Density of a county is correlated with many other factors. A denser county seems to experience more crime, as well as have a higher police presence per capita. The increased police presence may be a reaction to increased crime in a denser area. We control for these relationships by incorporating the density variable as a covariate into our regression.

Percentage Young Male

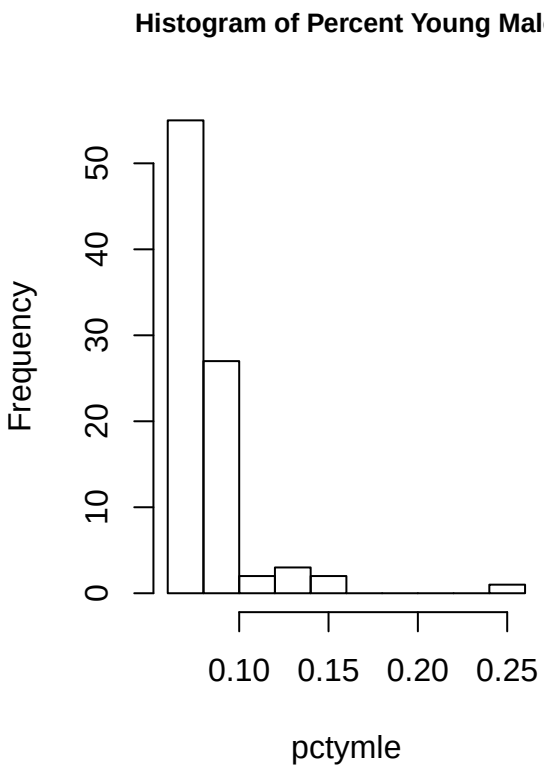
Another confounding relationship is that the percentage of young males in a county can also be associated with our explanatory and outcome variables. *pctymle* has a positive correlation of 0.3 with crime rate indicating that when percentage of young male is high, the crimes committed per person is high. But it also has a much lower negative correlation with probabilities of arrest and conviction, which makes it an interesting confounding variable.

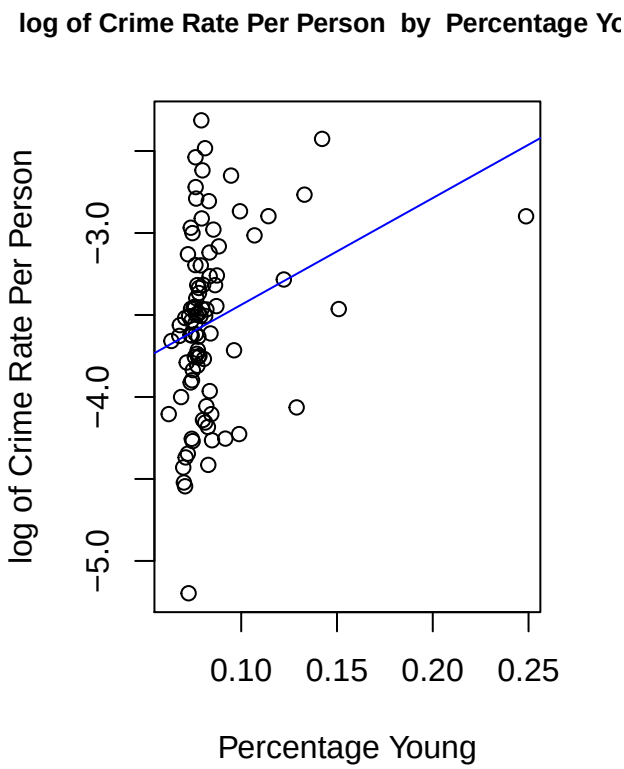
Let's look at histogram of this variable and its relationship with crime rate.

```
par(mfrow = c(1,2))
sumhist((data$pctymle), "Percent Young Male", "pctymle")
```

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

```
rmodel (data$pctymle, log(data$crmrte), "Percentage Young",
        "log of Crime Rate Per Person")
```






```
## Correlation of Percentage Young with log of Crime Rate Per Person : 0.2781547
```

It is theorized that young men have a greater tendency to commit crime.Thus, having more young males in an area can result in an increased police presence to combat crime, but also an increased crime rate. We control for this relationship by incorporating the percentage of young males in an area into our regression. Also *pctymle* has a reasonably good correlation with our outcome variable but not a very high correlation with all other predictor variables making it a viable candidate for selection.

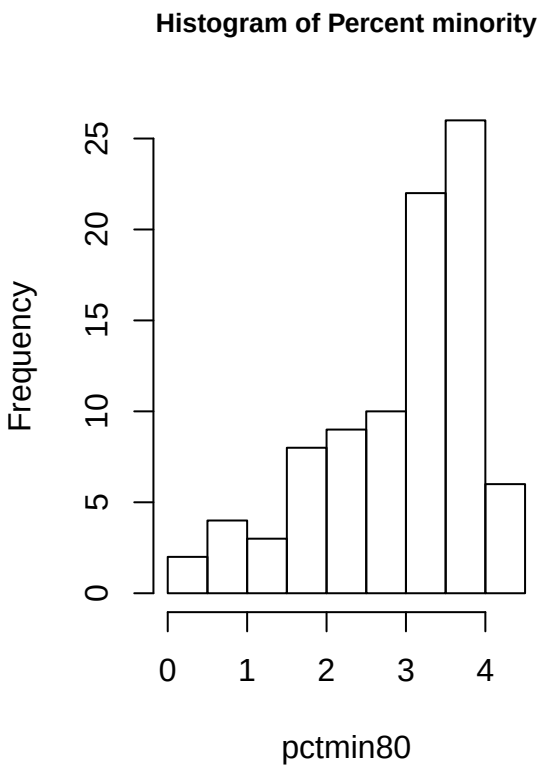
Percentage Minority

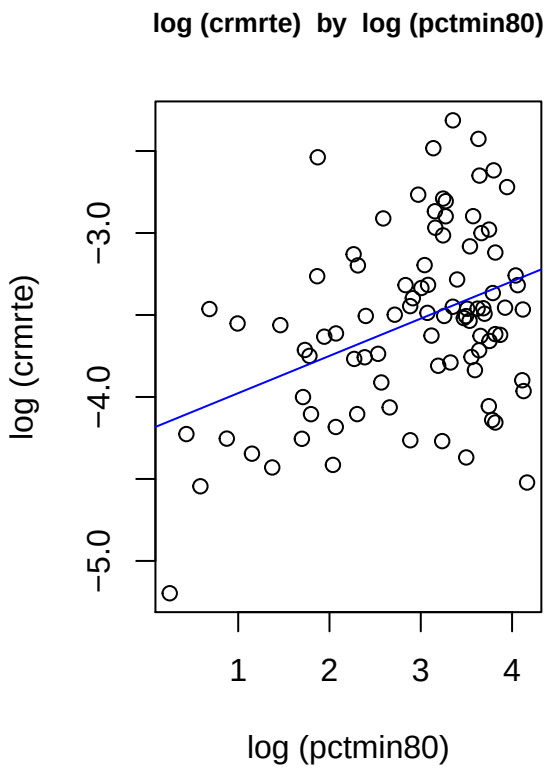
Let's look at histogram of this variable and its relationship with crime rate.

```
par(mfrow = c(1,2))
sumhist(log(data$pctmin80), "Percent minority", "pctmin80")
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2497  2.3050  3.2127  2.9134  3.6424  4.1643
```

```
rmodel (log(data$pctmin80), log(data$crmrte),
        "log (pctmin80)", "log (crmrte)")
```






```
## Correlation of log (pctmin80) with log (crmrte) : 0.397291
```

We have shown the histogram for a log transformation of *pctmin*80. We are performing a log transform because that increases the correlation with the outcome variable $log(crmrte)$ to 0.4 and gives us a better linear representation of the outcome variable. The scatter plot shows that as percentage minority increases more crimes are observed.

```
cor(log(data$pctmin80), log(data$polpc))
```

```
## [1] -0.1618985
```

This variable has an interesting relationship with police presence. The correlation between the log of both of these is -0.16. Police presence is low where percentage minority is high. However, this variable has a high positive correlation with crime rate indicating that higher the percentage minority higher the crime. There is a counter balancing act going on between these three variables making *pctmin*80 an interesting confounding variable in our regression model. Also while *pctmin*80 has a correlation with the outcome variable, it barely has any correlation with our predictor variables, making it a good candidate to add to our regression model.

Other Variables

Tax Revenue Per Capita

Tax revenue per capita seems to be an interesting variable at first. It has a high correlation of 0.45 with our outcome variable, which is desired for OLS regression. But it also has a correlation of 0.28 with *polpc*, a correlation of 0.32 with *density* and a correlation of 0.18 with *wtrd*. Since we have already picked *density* as one of our covariates (density has high correlation with *polpc*, *wtrd* but also very high correlation with our outcome variable too), it will not be parsimonious to add *taxpc* to our regression model.

Offense mix

Offense mix is defined as the ratio of face to face offense to other offenses. This has a positive correlation of 0.41 with probability of arrest, which makes sense. More arrests are made in cases where the perpetrator can be identified

owing to a face to face offense. This variable is an indicator of the type of crimes committed in a county and not a determinant of crime. So we will exclude this from our analysis.

West, Central and Urban

There are three region variables in our dataset: $west, central, urban$. We believe that density can be used as a proxy in place of these three as it gives us the impact of both urban and rural areas. And this way we will not need to add additional dummy variables and potential interaction terms. So we will be excluding these three variables from our analysis.

Multiple OLS Regression Modeling

Model Specifications

We construct the following linear regression models that predict the log-transformed crime rate:

1. model_1 with only the key explanatory variables mentioned in our hypothesis, to include:

- Probability of arrest and conviction
- Police per capita
- Weekly wages, Wholesale, Retail Trade

The crime rate, police per capita and wage variables are log-transformed as discussed in the EDA section of the report.

2. model_2 where we have added the following covariates to model_1:

- density
- percentage young male and
- percentage minority All of the covariates above are correlated with both crime rate and one or more of our key explanatory variables. We retain the log transforms from model 1 in model 2

3. model_3 with all variables from the dataset excluding year (which does not vary). Police per capita and probability of arrest are still log-transformed.

```
library(car)
```

```
## Loading required package: carData
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)

# generate our models
model_1 <- lm(log(crmrte) ~ log(polpc) + prbarr + prbconv
              + log(wtrd), data = data)

model_2 <- lm(log(crmrte) ~ log(polpc) + prbarr + prbconv
              + log(wtrd) + density + pctymle + log(pctmin80),
              data = data)

model_3 <- lm(log(crmrte) ~ log(polpc) + prbarr + prbconv
              + log(wtrd) + density + pctymle + log(pctmin80)
              + prbpris + avgsen + taxpc + west + central
              + urban + wcon + wtuc + wfir + wser + wmfg + wfed
              + wsta + wloc + mix, data=data)
```

Omitted Variables, Confounding Variables and Bias

Omitted Variables

As mentioned earlier in this report, there are a number of additional desired variables that would be helpful in building a more accurate crime predicting regression model. The dataset we have is very limited in the amount and type of information it provides to fully understand the determinants of crime. Additional demographics data from the counties e.g. age, gender of residents, education level of residents, employment data, unemployment status, household income etc. would help do a more thorough analysis.

Furthermore, additional data related to law enforcement like types of crime committed, violent versus non violent crime, severity of punishment, level of surveillance would be more meaningful in understanding crime deterrents better and thereby understanding how they can impact crime rate. Bucketing all crimes under the same umbrella might not give us the best representation as different types of crime might have different motivations (e.g. rape vs petty theft vs murder) and hence different determinants.

Lastly, we were only provided data for 97 counties out of the total 100 counties which were in existence in 1987. We do not have information as to why these 97 were provided but have to consider whether this sample data would be different if data for all 100 counties was provided.

Confounding Variables

We have covered the confounding variables in our dataset in the EDA section. Another confounding variable could be police policy. Policies made by the police could influence both the severity and likelihood of punishment as well as the crime rate. However, they are not contained in the dataset we have. There may be other unobserved characteristics of a county associated with our explanatory variables that would cause them to seem correlated even when the relationships are not causal

In a perfect world to reduce the impact of confounding variables, we would be able to run an experiment to randomly change police per capita, blue collar wages, and the probability of arrest to see these impacts on the crime rate. However, it may not be practically possible to conduct such an experiment.

Another way to reduce the impact of confounding variables would be to observe the change in crime rates in these counties over time. By holding characteristics of the counties themselves more or less constant over time, those confounding characteristics' impact can be reduced.

Bias

We believe many of the variables mentioned above would play an important role determining crime. Omitting them bias our estimates of slope co-efficient for the variables in our model and possibly over/under state the effects of the predictor variables in our model depending on the direction of the bias. Following are some of the omitted variables bias we look into:

1. Level of education: Level of education is correlated with crime rate and wages. As level of education increases we expect crime rate to decrease and wages to go up. So we can see that our omitted variable bias is negative and has understated our slope co-efficient for wages. This bias moves our slope co-efficient in a downward direction towards zero.
2. Unemployment rate: As unemployment rate increases, expect crime rate to increase and wages to go down. Our omitted variable bias will be negative. So we can see that our omitted variable bias is negative and has understated our slope coefficient for wages. This bias moves our slope co-efficient in a downward direction towards zero.
3. Household income: Household income is correlated with crime rate and wages. As household income increases we expect crime rate to decrease and wages to go up. So we can see that our omitted variable bias is negative and has understated our slope coefficient for wages. This bias moves our slope co-efficient in a downward direction towards zero.
4. Level of surveillance: Higher surveillance will lead to higher probability of arrests and lower crimes. Our omitted variable bias will be negative. This bias moves our slope coefficient in a negative direction away from zero.

5. Average age of a county: We expect a higher average age of a county would mean a decreased crime rate and also increased wages. Because the correlation directions are different, our omitted variable bias will be negative. The bias moves our slope coefficient in a downward direction towards zero.

Regression Table

```
# calculate robust standard errors
se.model_1 = sqrt(diag(vcovHC(model_1)))
se.model_2 = sqrt(diag(vcovHC(model_2)))
se.model_3 = sqrt(diag(vcovHC(model_3)))

# display regression table with robust errors
suppressWarnings(stargazer(model_1, model_2, model_3,
          type="latex", header=FALSE, single.row=TRUE,
          omit.stat="f",
          title = "Linear Models Predicting Crime Rate",
          omit.table.layout = "n",
          se=list(se.model_1, se.model_2, se.model_3),
          star.cutoffs=c(0.05, 0.01, 0.001)))
```

Table 3: Linear Models Predicting Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
log(polpc)	0.530*** (0.113)	0.490*** (0.130)	0.473** (0.182)
prbarr	−2.227*** (0.354)	−1.713*** (0.250)	−1.471*** (0.235)
prbconv	−0.690*** (0.107)	−0.605*** (0.096)	−0.569*** (0.166)
log(wtrd)	0.789** (0.303)	0.415 (0.252)	0.259 (0.391)
density		0.075** (0.027)	0.106* (0.045)
pctymle		1.826 (1.238)	2.186 (1.582)
log(pctmin80)		0.211*** (0.030)	0.165* (0.068)
prbpris			0.154 (0.567)
avgsen			−0.013 (0.018)
taxpc			0.001 (0.008)
west			−0.148 (0.173)
central			−0.161 (0.085)
urban			−0.131 (0.200)
wcon			0.0005 (0.001)
wtuc			0.0002 (0.001)
wfir			−0.001 (0.001)
wser			−0.0001 (0.002)
wmfg			−0.0002 (0.001)
wfed			0.002 (0.001)
wsta			−0.001 (0.001)
wloc			−0.0001 (0.002)
mix			−0.548 (0.622)
Constant	−3.297 (2.105)	−2.631 (1.860)	−1.983 (1.924)
Observations	90	90	90
R^2	0.650	0.817	0.863
Adjusted R^2	0.633	0.801	0.818
Residual Std. Error	0.332 (df = 85)	0.245 (df = 82)	0.234 (df = 67)

Across the different regression models and using heteroskedasticity-robust standard errors, we find that the effects for

probability of arrest, log police per capita and probability of conviction are significant at at least the 5% level across having only the explanatory variables, several confounding variables, and all the remaining applicable variables of the dataset in the regression. They maintain their direction. The coefficients lose some of their magnitude after adding in more covariates, suggesting that some of their original effect was due to positive omitted variables bias.

The coefficient for the log transform of police per capita is positive and ranges from 0.473 to 0.530. For every increase of police per capita, the crime rate increases. The magnitude of the increase is a 0.47% to 0.53% increase in crimes per 1% increase in number of police officers per capita. The significance of the police per capita coefficient fits with our hypothesis that the different numbers of police per capita can change the crime rate. The direction of this effect may seem counter intuitive. A policy might predict that increasing the number of police would reduce crime. However, we see the opposite. One possible explanation is that there is a third variable both crime rate and police per capita are correlated with that causes them to both be positively correlated with each other. Another possible explanation is that increasing the number of police increases the police department's capacity to detect crime, causing a seeming increase in crime when the only thing increasing is the observation of crime.

The coefficient for the probability of arrest, *prbarr*, is negative and ranges from -2.227 to -1.471. As the probability of arrest increases, the crime rate decreases. Depending on the model, as the probability of arrest increases by by 1% the number of crimes per 100 people decreases by 0.022% to 0.014%. An interpretation of this effect is that when police policy is "tougher", criminals are more discouraged from committing crimes. This fits with our policy intuitions that changing the stance on crime can impact the crime rate.

The coefficient for the probability of conviction, *prbconv*, is negative and ranges from -0.569 to -0.690. As the probability of conviction increases, the crime rate decreases. Depending on the model, as the probability of conviction increases by by 1% out of 100% the crime rate decreases by 0.006% to 0.007%. Like the probability of arrest coefficient above, an interpretation of this effect is that when police policy is "tougher", criminals are more discouraged from committing crimes. This also fits with our policy intuitions that changing the stance on crime can impact the crime rate.

The coefficient for the log-transformed average weekly retail and wholesale wages in an area, *wtrd*, starts significant in our first model and loses significance as we add more variables. The direction is always positive. The effect also decreases in size as more variables are added. The coefficient ranges from 0.259 to 0.789 depending on the model, which means for every 1% increase of the average weekly wage, the number of crimes committed per 100 people increases by 0.26% to 0.79%. The effect seen in the service wage coefficient may not be significant, which may not support our alternative hypothesis that changing wages impacts the crime rate.

Practical Significance

We now examine the practical significance for each of the explanatory variables.

As we have seen in the coefficient for the police per capita, across our models there is a 0.47% to 0.53% increase in crimes per 1% increase in number of police officers per capita, keeping all other variables constant. This is roughly in the same magnitude (an increase of police per capita by a percentage is associated with an increase of half that percentage in crime rate), which implies that this relationship is highly practically significant.

For probability of arrest and probability of conviction, the coefficient size is much smaller. Maintaining ceteris paribus, as the probability of arrest increases by by 1% the number of crimes per 100 people decreases by 0.022% to 0.014%, and as the probability of conviction increases by by 1% out of 100% the crime rate decreases by 0.006% to 0.007%. This is an effect size on the order of 10 to 100 times smaller than the one seen for police per capita.

Finally, we have seen that for every 1% increase of the average weekly wage, the number of crimes committed per 100 people increases by 0.26% to 0.79%, keeping all other factors constant. This effect size is roughly comparable to that for the police per capita, so we can say this is also practically significant.

In terms of fit, using the adjusted R squared values, we can see that model_1 is able to explain 63% of the variation in our dataset while model_2 and model_3 can explain 80% and 81% of our dataset respectively.

Block Tests

We perform several block tests across several model coefficients to test whether coefficients for related variables are jointly significant.

Stance on Crime

One of our hypotheses is that the stance on crime can affect crime rate. While we have established that there are robust statistically significant effects for individual "probability" variables above (such as probability of arrest and probability of conviction), we should also perform a test to establish that there is a general effect of stance towards crime on the crime rate. The fact that we have found significance above is a strong indicator that we will find a statistically significant result in our joint significance test.

Running a joint significance test on model 2, we find that the coefficients for *prbarr* and *prbconv* are jointly highly significant at the 1% level. We reject the null hypothesis that both of these coefficients are 0.

```
linearHypothesis(model_2, c("prbarr = 0", "prbconv = 0"),
                 vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ log(polpc) + prbarr + prbconv + log(wtrd) + density +
##     pctymle + log(pctmin80)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     84
## 2     82  2 31.089 8.945e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running a joint significance test on model 3, we find that the coefficients for *prbarr*, *prbconv*, *prbpris*, and *avgsen* are jointly highly significant at the 1% level. We reject the null hypothesis that both of these coefficients are 0.

```
linearHypothesis(model_3,
                 c("prbarr = 0", "prbconv = 0",
                   "prbpris = 0", "avgsen = 0"),
                 vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## prbpris = 0
## avgsen = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ log(polpc) + prbarr + prbconv + log(wtrd) + density +
##     pctymle + log(pctmin80) + prbpris + avgsen + taxpc + west +
##     central + urban + wcon + wtuc + wfir + wser + wmfg + wfed +
##     wsta + wloc + mix
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     71
## 2     67  4 11.877 2.376e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The conclusion from the block tests on the "stance on crime" variables is that since both these tests are significant across two models, there is a significant effect on crime rates from the stance on crime.

Wages

We conduct a joint significance test on all provided wage variables. While the coefficient for $log(wtrd)$ is significant for model 1, it loses significance in model 2. In model 3, it is not significant, along with all other wage variables.

Performing the joint significance test, we find a p-value of 0.28, so we fail to reject the null hypothesis that there is no effect of wage on crime rate.

```
linearHypothesis(model_3,
                 c("wcon = 0", "wtuc = 0", "log(wtrd) = 0",
                   "wfir = 0", "wser = 0", "wmfg = 0",
                   "wfed = 0", "wsta = 0", "wloc = 0"),
                 vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## wcon = 0
## wtuc = 0
## log(wtrd) = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wfed = 0
## wsta = 0
## wloc = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ log(polpc) + prbarr + prbconv + log(wtrd) + density +
##     pctymle + log(pctmin80) + prbpris + avgsen + taxpc + west +
##     central + urban + wcon + wtuc + wfir + wser + wmfg + wfed +
##     wsta + wloc + mix
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     76
## 2     67  9 1.2369 0.2881
```

Information Loss

Looking at the AIC (Akaike Information Criterion) for all three models, we find that model 2 (key explanatory variables with covariates) has the lowest AIC at 11.72. This means that adjusted for parsimony, this model is the best fit out of the three we have constructed.

```
paste("Model 1 AIC: ", AIC(model_1))
```

```
## [1] "Model 1 AIC:  64.0165507406399"
```

```
paste("Model 2 AIC: ", AIC(model_2))
```

```
## [1] "Model 2 AIC:  11.7256237782597"
```

```
paste("Model 3 AIC: ", AIC(model_3))
```

```
## [1] "Model 3 AIC:  15.4713476390451"
```

The Six Classical Linear Model (CLM) Assumptions

Having discussed each of the model specifications we are using for our analysis, we will take this opportunity to deep dive into one (model_2), checking the 6 CLM assumptions as a sample of the validation that we have conducted in this report, for each of the models.

CLM 1: A Linear Model

Model_2 is built such that *crmrte* is a linear function to the outcome variables that we have selected to include in this model. We have not constrained the error term. Additionally the scatter plots between the dependent variable and each of the independent variables show that there are linear relationships between these. For this reason we can validate this assumption and move to the second CLM assumption.

CLM 2: Random Sampling

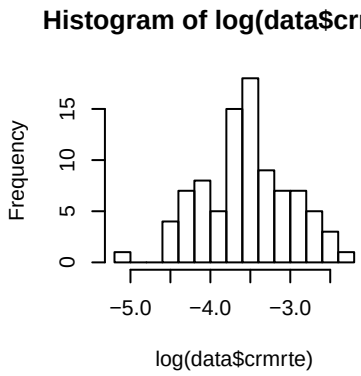
Along with the data, we have been provided minimal information regarding how the data was captured. For several of the variables we are analyzing we can look at the distribution along with the sources mentioned, to determine whether random sampling condition is met.

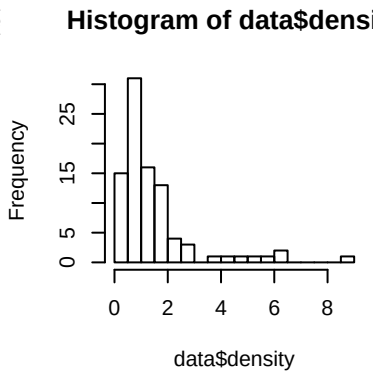
Model_2 includes, the dependent variable, *crmrte*, which we are not given information about how the data was collected. Looking at the distribution of the log transformed version of this variable we can see that it is generally normally distributed. This along with applying the Central Limit Theorem we can validate random sampling for the outcome variable.

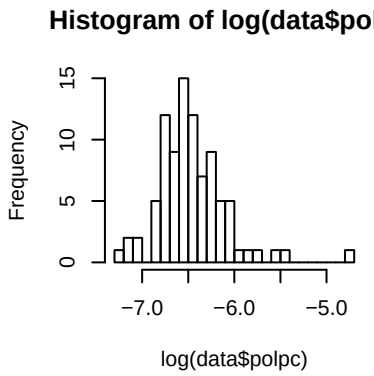
Next, in regards to the explanatory variables, which we have 7 of, by looking at the distribution that are provided here we can see that with the exception of *polpc* which is log transformed, and closer to normal distribution, all have positive skews. Considering the *pctymle* and *pctmin*80 are census data we can be confident that these are also random samples. For the Density variable we don't have information regarding how the data may have been collected and the fact that most counties are leaning towards less density makes it unclear whether this data is a random sample. Lastly *prbarr* is from the Uniform Crime Reports data and is not a random sample. Further *prbconv* is coming from the North Carolina Department of Correction and not likely random sample.

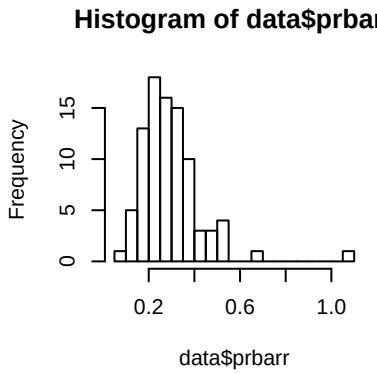
Overall we don't believe that this data is a completely passes the test of random sampling but also does not exhibit large biases as well, which makes it acceptable for use in our analysis.

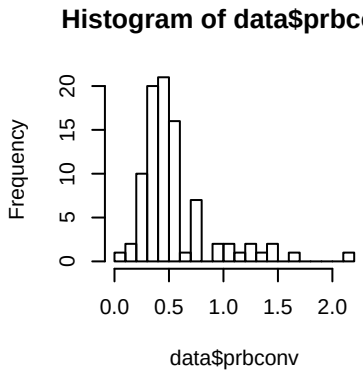
```
par(mfrow = c(2,3))
hist(log(data$crmrte), breaks = 20)
hist(data$density, breaks = 20)
hist(log(data$polpc), breaks = 20)
hist(data$prbarr, breaks = 20)
hist(data$prbconv, breaks = 20)
hist(data$pctymle, breaks = 20)
```

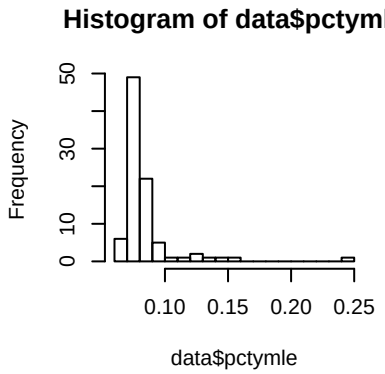




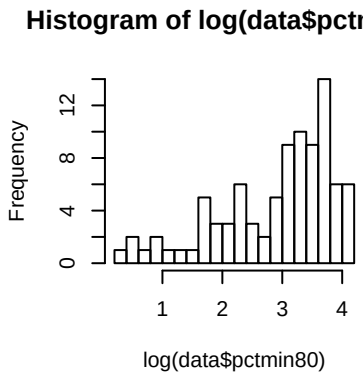









```r
hist(log(data$pctmin80), breaks = 20)
```



CLM 3: Multicollinearity

To quickly test for the multicollinearity condition, we can check the correlation of all the explanatory variables and their Variance Inflation Factors (VIF).

In the correlation matrix we can see that there is no perfect correlation between any two of the independent variables in the model. Additionally, we can see that the VIFs are all much less than 10 meaning they are low. Based on this we can confidently say model_2 passes this assumption of no perfect collinearity.

```
# Create correlation matrix for model 2 explanatory variables
X = data.matrix(subset(data, select=c("density", "polpc",
                                      "prbarr",  "prbconv",
                                      "pctymle", "pctmin80", "wtrd")))
(Cor = round(cor(X),2))
```

```
##          density polpc prbarr prbconv pctymle pctmin80  wtrd
## density     1.00  0.16  -0.30   -0.23    0.12    -0.07  0.59
## polpc       0.16  1.00   0.43    0.17    0.05    -0.17  0.12
## prbarr     -0.30  0.43   1.00   -0.06   -0.18     0.05 -0.10
## prbconv    -0.23  0.17  -0.06    1.00   -0.16     0.06 -0.13
## pctymle     0.12  0.05  -0.18   -0.16    1.00    -0.02 -0.11
## pctmin80   -0.07 -0.17   0.05    0.06   -0.02     1.00 -0.06
## wtrd        0.59  0.12  -0.10   -0.13   -0.11    -0.06  1.00
```

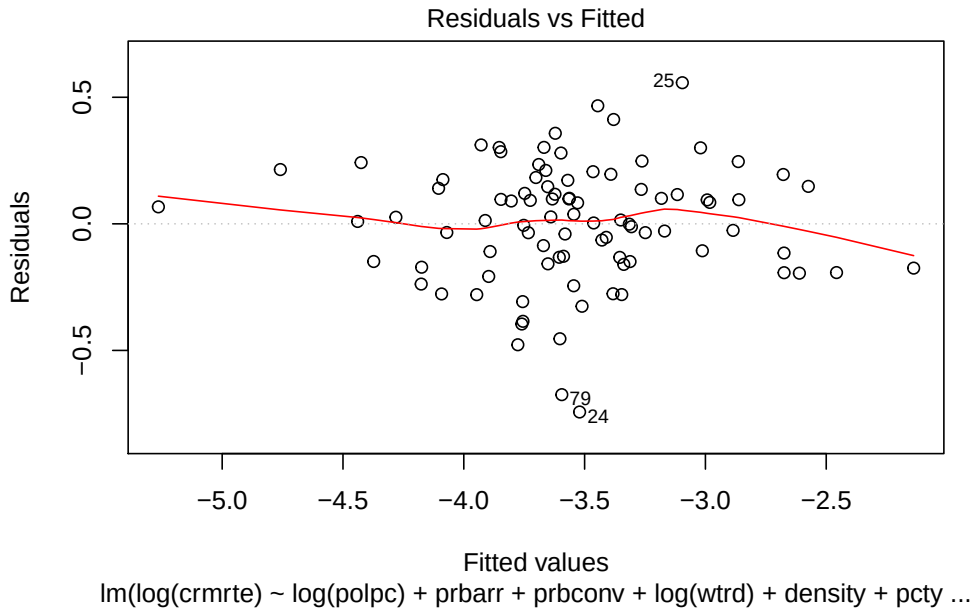
```
vif(model_2)
```

```
##    log(polpc)        prbarr       prbconv     log(wtrd)       density
##      1.387396      1.379862      1.134915      1.559221      2.022335
##       pctymle log(pctmin80)
##      1.166353      1.049313
```

CLM 4: Zero Conditional Mean

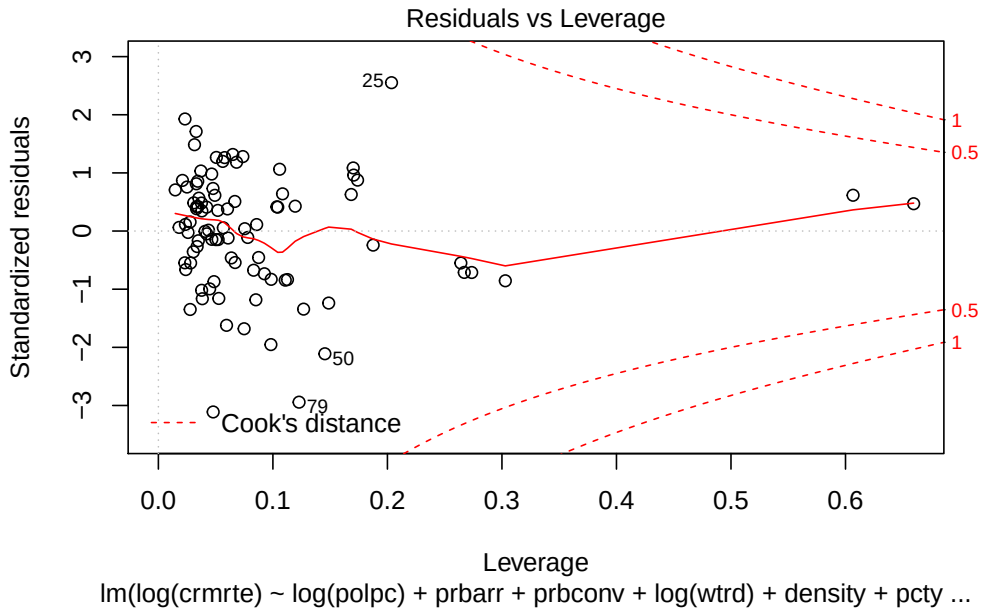
To check model_2 for zero conditional mean we will first look at the the Residual vs. Fitted plot followed by the Residuals vs. Leverage plot.

```
plot(model_2, which=1)
```



```
cat('\n')
```

```
plot(model_2, which=5)
```

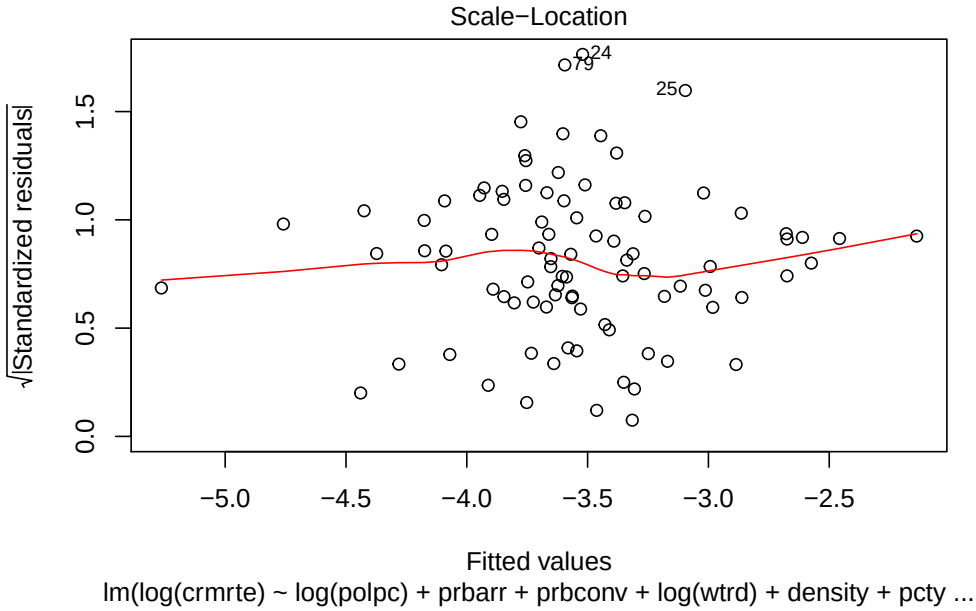


The Residual vs. Fitted Plot shows a fairly flat red spline which indicates the residuals are fairly balanced. One thing to note is that the red spline is going up int he beginning and down at the end but as we can see there is only one data point at those extremes which is likely causing them. The residual vs. leverage plot also indicates toward meeting zero conditional mean as the line is hovering around zero throughout. Last thing to note from this plot is taking into account where the cook's distance lines are that there are no data points that have major influence on the model.

CLM 5: Homoscedasticity

To check for homoscedasticity, first we look back at the thickness of the data band from the residual vs. fitted plot we looked at for the last assumption and the scale-location plot. Upon visual inspection we see that the band starts out much thinner, thickens up and then thins out again, indicating that we do not meet the Homoscedasticity assumptions.

```
plot(model_2, which=c(3))
```




```
# Breusch-Pagan-Test
bptest(model_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 12.286, df = 7, p-value = 0.09153
```

```
# Score-test for non-constant error variance
ncvTest(model_2)
```

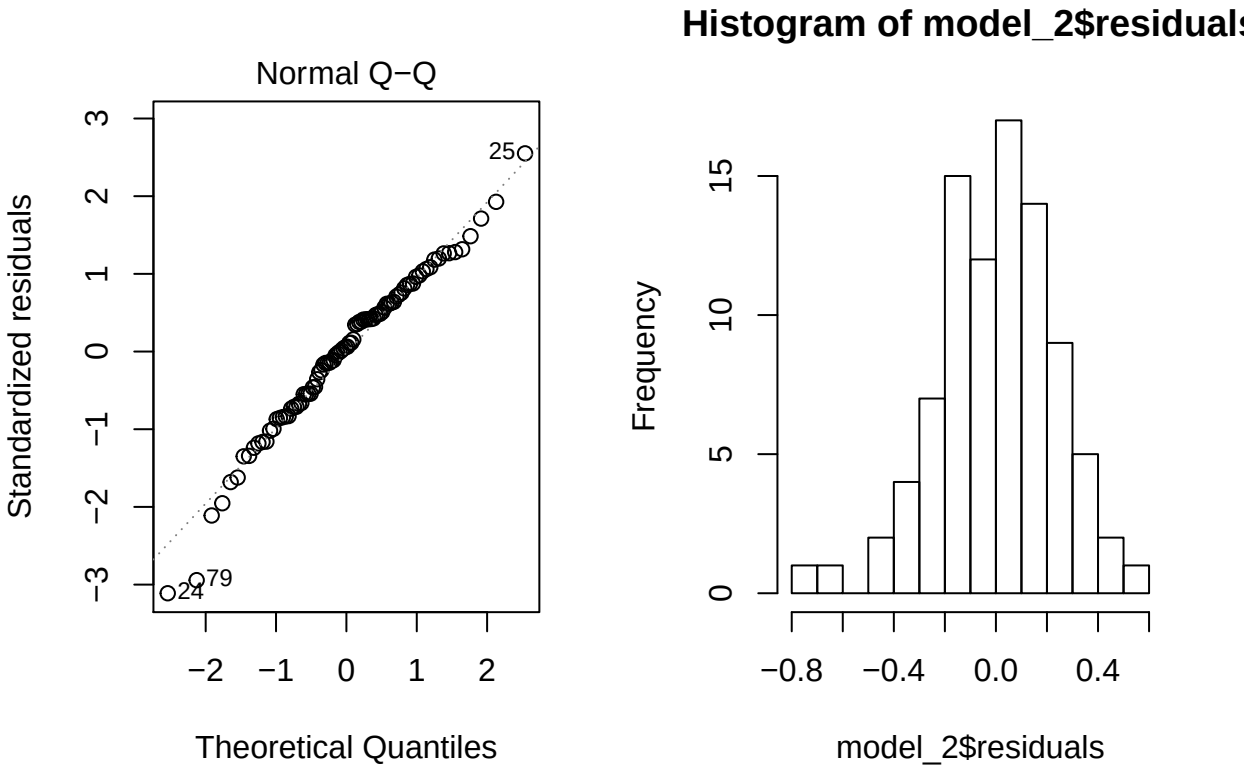
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.008117053, Df = 1, p = 0.92821
```

just to confirm we run the bptest and ncvTest and see that we based on the p-values we fail to reject that null hypothesis that the model is homoscedastic. Since we have mixed evidence of homoscedasticity (visual vs pvalues of bptest etc.), we are going to address the problem of potential heteroscedasticity by using standard robust errors for our model.

CLM 6: Normality of Residuals

By visually observing the Normal Q-Q plot of of the residuals, as well as the distribution of the residuals for model_2 we can assess whether this assumption is valid.

```
par(mfrow=c(1,2))
plot(model_2, which=2)
hist(model_2$residuals, breaks=10)
```



The Normal Q-Q plot is quite linear with only some deviations near each end. The model's distribution of the residuals is also generally normal with the exception of a few outliers one the left side of the graph. Overall we can say that model_2 passes the assumption for normality of residuals.

In summary, after testing out each of the 6 CLM assumptions with the exception of homoscedasticity, which we mentioned we will address by using robust standard errors, model_2 passes all other assumptions.

CLM Assumptions Summary for Model 1 and Model 3

For our other models we used the same methodology to test for the CLM assumptions. However, model_1 did not pass as many assumptions as model_2. for model_1 the zero conditional mean was less valid, as was normality of residuals. Additionally for model_3 these same assumptions of zero conditional mean and the normality of residuals were less robust than the ones in model_1. All these models were heteroscedastic and robust standard errors would be recommended.

Conclusion

In this report we have sought to operationalize the determinants of crime and supported our decisions by providing thorough and step by step breakdown of our analysis in developing our models. More specifically we have three different specifications of OLS models to understand the determinants of crime. Model_1 is a parsimonious model, designed to include only the key variables from our hypothesis, which were based on our intuitive knowledge on factors that could potentially impact crime. To eliminate endogeneity from model_1, we have conceptualized model_2 to include covariates that have correlation with our outcome variable and predictor variables. Even while creating model_2, we have been careful in selecting only covariates that we have justification for and can proxy other covariates without trading off parsimony for fit. Model_3 includes all variables from the dataframe (with the exception of countyID and year) and retains all the transformations from model_1 and model_2. This model serves as a datapoint to check the robustness of model_1 and model_2.

Based on the results of our models, we see that our hypothesis holds with respect to both crime deterrents ($prbarr, prbconv$) and law enforcement ($polpc$), each having statistical significance in all three models. Additionally, our covariate selections for *density* and *pctmin*80 have also proven highly statistically significant as important

determinants of crime, as indicated by model_2. We looked at wage from different aspects and while it shows some significance in model_1, we did not find it to be a significant factor in determining crime in models 2 and 3. Based on our findings, *pctymle* does not seem to be a determinant of crime.

To summarize we have found that model_2 is the best representation of our dataset, taking into account AIC scores, model fit, statistical significance and practical significance.

Policy Recommendation

Based on selecting model_2 as our best representation of our results and findings, the following are the policy recommendations we are making to the political campaign:

1. Get tough on crime! By increasing the number of arrest and conviction, we should see a decrease in crime rates. This will send crimanals the message that law enforcement is tough on crime and make communities safer through detterance.
2. Immigration reform. By reforming immigration laws to better vet immigrants to the state and ultimately the counties in North Carolina can help aid in reducing crime as our findings indicate.
3. While our findings indicate that an increase in police presence will increase crime, this seems to be counterintuitive. We have reason to believe this relationship could be a false positive in terms of assessing safety, which is what reducing crime is trying to achieve. We believe, based on research and knowledge of the criminal justice system that decreasing police presence may falsely lower crime by catching less criminals overall. Our recommendation in terms of this finding is that more research and analysis should take place to validate this relationship before any policy recommendations are made.