W203_Lab_3

April 15, 2019

1 W203 Lab 3: Reducing Crime in North Carolina

Spyros Garyfallos, Ross MacLean, Paul Petit

1.1 Introduction

The topic of crime is expected to play a central role in the forthcoming North Caroline election. As such, our consultancy firm has been hired by one of the political campaigns to identify its determinants in an effort to inform policy. Using statistical analysis, this report seeks to outline the key drivers of crime, culminating in a series of policy recommendations to address its root causes and reduce crime rates over time.

Our consultancy firm is broadly interested in the factors that influence crime rates that are within the remit of control for policymakers. Specifically, our research will assess whether there is evidence to support making changes to the legislation related to the certainty and severity of criminal punishment. We'll approach analyzing this evidence from two perspectives, a simpler perspective focused on just the determinants of crime that policy can influence and a more holistic perspective that considers a broader range of factors.

1.2 Exploratory Data Analysis

The crime dataset that will be used to inform these policy decisions contains county-level information on a range of metrics pertaining to crime such as policing levels, probability of arrest/conviction, average sentence and crime rate. It also contains demographic and economic data that can be used to explore the factors that are associated with varying crime rates.

Research indicates that there are 100 counties in North Carolina currently, yet only 90 counties are present in the dataset. It is not possible to identify which counties are missing which introduces some uncertainty to the extent our policy recommendations are truly applicable North Carolina in its entirety, especially so if some of the counties with larger populations are absent from the sample. For the purposes of this assignment, we'll assume that these 90 counties were obtained using random sampling so inferences can be made about the population.

1.2.1 Data Cleansing and Manipulation

```
In [37]: # Install libraries
         if(!require(stargazer))
             suppressMessages(install.packages('stargazer'))
         if(!require(corrplot))
             suppressMessages(install.packages('corrplot'))
```

```
 if(!require(plyr))
     suppressMessages(install.packages('plyr'))
 if(!require(sandwich))
     suppressMessages(install.packages('sandwich'))
 if(!require(lmtest))
     suppressMessages(install.packages('lmtest'))
 if(!require(car))
     suppressMessages(install.packages('car'))
```

In [2]:
```
# Load libraries
suppressMessages(library('stargazer'))
suppressMessages(library('corrplot'))
suppressMessages(library('plyr'))
suppressMessages(library('sandwich'))
suppressMessages(library('lmtest'))
suppressMessages(library('car'))
```

In [3]:
```
# Read crime data
crime.data <- read.csv(file='crime_v2.csv', header=TRUE, sep=',')

# Remove the trailing blank rows
crime.data <- crime.data[complete.cases(crime.data), ]

# Convert to numeric
crime.data$prbconv <- as.numeric(as.character(crime.data$prbconv))

# Check there still 91 rows of complete data
paste('Number of rows of data: ', nrow(crime.data))

# Identify if there are duplicate counties
crime.data.county.count <- count(crime.data, c('county'))

stargazer(crime.data.county.count[crime.data.county.count$freq > 1, ],
          type = 'text')

# Remove duplicate row (select unique)
crime.data <- unique(crime.data[, , ] )
paste('Remove duplicate county')

paste('Rows:', nrow(crime.data))
paste('Columns:', ncol(crime.data))

# Counties not in west/central region
paste('Number of counties neither west nor central: ',
      length(crime.data[crime.data$west == 0 & crime.data$central == 0,
                        'county']))

# Create the east region
```

```
        crime.data$east <- ifelse(crime.data$west == 0 & crime.data$central == 0, 1, 0)

        cat('\n')
```

'Number of rows of data: 91'

```
==========================================================
Statistic N  Mean   St. Dev. Min Pctl(25) Pctl(75) Max
----------------------------------------------------------
county    1 193.000          193   193      193    193
freq      1  2.000            2     2        2      2
----------------------------------------------------------
```

'Remove duplicate county'
'Rows: 90'
'Columns: 25'
'Number of counties neither west nor central: 35'

The crime dataset is rich source of information, but it needs to be cleaned before it can be used effectively. The original dataset contained 6 largely blank trailing rows that were removed. The variable representing the probability of conviction also needed to be converted to a numeric data type for later computations. The primary key in the dataset is county and so it's necessary to check for uniqueness within this field. A quick count of distinct values revealed there was a duplicate entry that needed to be removed for county '193'. Following the removal of the duplicate entry, the final dataset consisted of observations for 90 counties across 25 variables.

The table below contains a summary of various descriptive statistics for each variable and provides an indication of those that require further investigation, guiding the subsequent section on outliers.

```
In [4]: # Relative SD statistic
        rsd <-  function(x)
        {
            return(sd(x)/abs(mean(x)))
        }
        min.distance <-  function(x)
        {
            minmax <- max(abs(min(x)-mean(x)), abs(max(x)-mean(x)))
            return(minmax/sd(x))
        }

        # Descriptive statistics
        crime.descriptive.statistics <- function(data, excluded.colums, column.names)
            {
            vars <- colnames(data)
```

```
    mins <- format(round(sapply(data, min), 3), nsmall = 3)
    means <- format(round(sapply(data, mean), 3), nsmall = 3)
    maxs <- format(round(sapply(data, max), 3), nsmall = 3)
    sds <- format(round(sapply(data, sd), 3), nsmall = 3)
    rsds <- format(round(sapply(data, rsd), 3), nsmall = 3)
    min.distances <- format(round(sapply(data, min.distance), 3), nsmall = 3)

    # Bind to df and format
    result.data.frame <- data.frame(cbind(vars,
                                          mins,
                                          means,
                                          maxs,
                                          sds,
                                          rsds,
                                          min.distances))
    result.data.frame <- subset(
        result.data.frame,
        !(vars %in% excluded.colums))

    names(result.data.frame) <- column.names
    row.names(result.data.frame) <- NULL

    return (result.data.frame)
}

crime.descriptive.statistics(crime.data,
                             c('county',
                               'year',
                               'west',
                               'central',
                               'east',
                               'urban'),
                             c('',
                               'Min',
                               'Mean',
                               'Max',
                               'SD',
                               'Relative SD',
                               'Max SDs from Mean'))

cat('\n')
```

	Min	Mean	Max	SD	Relative SD	Max SDs from Mean
crmrte	0.006	0.034	0.099	0.019	0.564	3.466
prbarr	0.093	0.295	1.091	0.138	0.466	5.779
prbconv	0.068	0.551	2.121	0.354	0.643	4.434
prbpris	0.150	0.411	0.600	0.081	0.196	3.231
avgsen	5.380	9.689	20.700	2.834	0.293	3.885
polpc	0.001	0.002	0.009	0.001	0.580	7.414
density	0.000	1.436	8.828	1.522	1.060	4.858
taxpc	25.693	38.161	119.761	13.112	0.344	6.223
pctmin80	1.284	25.713	64.348	16.985	0.661	2.275
wcon	193.643	285.353	436.767	47.753	0.167	3.171
wtuc	187.617	410.907	613.226	77.355	0.188	2.887
wtrd	154.209	210.921	354.676	33.870	0.161	4.244
wfir	170.940	321.621	509.466	53.999	0.168	3.479
wser	133.043	275.338	2177.068	207.396	0.753	9.170
wmfg	157.410	336.033	646.850	88.231	0.263	3.523
wfed	326.100	442.619	597.950	59.951	0.135	2.591
wsta	258.330	357.740	499.590	43.294	0.121	3.276
wloc	239.170	312.280	388.090	28.132	0.090	2.695
mix	0.020	0.129	0.465	0.082	0.634	4.110
pctymle	0.062	0.084	0.249	0.023	0.279	7.023

The final column on the right is a measure of how many SDs are the maximum or minimum values from the mean, indicating just how extreme the min/max values are for each variable. We will revisit this table after our data clean up, when we've dealt with the outliers, either by transforming the scale of the variables or by imputing the outlier values.

1.3 Outliers

1.3.1 County 71

It appears that county 71 is both in west and central regions. This is the only county with this characteristic, so we'll assume it's an error. Since we don't know which region is actually correct, we will remove both categories, and we intentionally don't classify this county as east.

```
In [5]: # Removing the west and central from 71
        crime.data[crime.data$county == 71, ]$west = 0
        crime.data[crime.data$county == 71, ]$central = 0
```

1.3.2 County 115

Further analysis was conducted to understand whether a small number of counties were contributing to these outliers. Closer inspection revealed that county 115 was an extreme outlier in terms the probability of arrest, probability of conviction, number of police per capita, average sentence and percent minority. In the next analysis we show the variables for which this county is either a minimum or a maximum of the entire dataset. We also notice that, compared to the variable mean, the county 115 values are extreme.

```
In [6]: removed.columns <- c('county', 'year', 'west',
                             'central', 'east','urban',
                             'color3c', 'prbarr.log', 'prbconv.log')

        column.names <- c('', 'Min', 'Mean',
                          'Max', 'SD', 'Relative SD')

        keep.rows <- c('crmrte','prbarr','avgsen',
                       'polpc','pctmin80','pctmin80')

        overall.statistics <- crime.descriptive.statistics(crime.data,
                                    removed.columns,
                                    column.names)

        outlier <- crime.data[crime.data$county == 115,]
        outlier <- t(outlier)
        outlier <- format(round(outlier, 3), nsmall = 3)
        outlier <- outlier[!rownames(outlier) %in% removed.columns, ]
        df <- cbind(overall.statistics, outlier)
        df <- df[ -c(1, 6) ]
        df[rownames(df) %in% keep.rows, ]

        cat('\n')
```

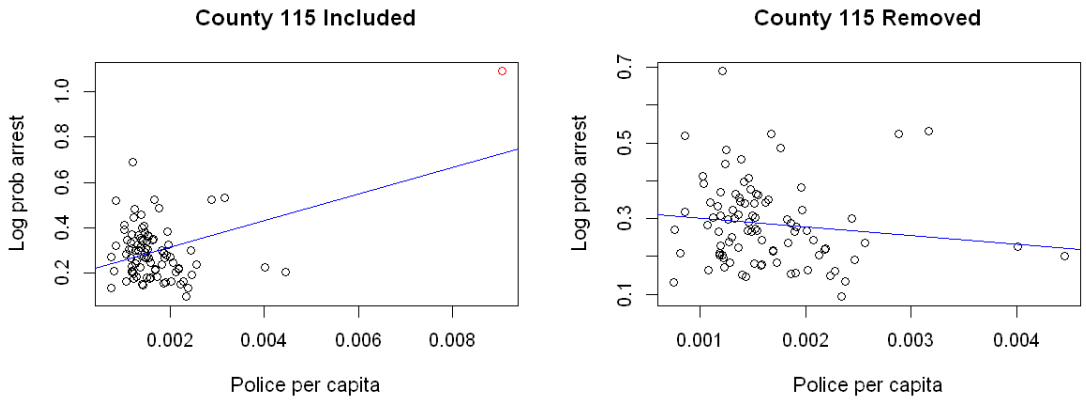
	Min	Mean	Max	SD	NA	outlier
crmrte	0.006	0.034	0.099	0.019	3.466	0.006
prbarr	0.093	0.295	1.091	0.138	5.779	1.091
avgsen	5.380	9.689	20.700	2.834	3.885	20.700
polpc	0.001	0.002	0.009	0.001	7.414	0.009
pctmin80	1.284	25.713	64.348	16.985	2.275	1.284

This is interesting case and presents an opportunity to demonstrate the extent to which a single observation can influence the nature of association between variables across the entire dataset, and ultimately on the resulting regression model. The scatter plots below show the relationship between the number of police per capita and the probability of arrest with the extreme values of county 115 included, and excluded.

```
In [7]: # Setup the plots

        plot.settings <- function(ratio, width, rows, columns)
            {
            options(repr.plot.width=width, repr.plot.height=width*ratio)
            par(mfrow = c(rows, columns))
        }

        plot.settings(0.4, 10, 1, 2)
```

```r
# Color county 115
crime.data$color3c <- ifelse(crime.data$county == 115, 'red', 'black')
crime.data.without115 = crime.data[crime.data$county != 115,]

plot.outlier.comparison <- function(x, y, main, col, xlab, ylab)
    {
    plot(x,
         y,
         main = main,
         col = col,
         cex.main = 1.2,
         cex.lab = 1.1,
         cex.axis = 1,
         xlab = xlab,
         ylab = ylab)

    abline(lm(y ~ x), col = 'blue')
}

# Police per capita plots
plot.outlier.comparison(
    crime.data$polpc,
    crime.data$prbarr,
    'County 115 Included',
    crime.data$color3c,
    'Police per capita',
    'Log prob arrest')

# County 115 removed
plot.outlier.comparison(
    crime.data.without115$polpc,
    crime.data.without115$prbarr,
    'County 115 Removed',
    crime.data.without115$color3c,
    'Police per capita',
    'Log prob arrest')

paste('Correlation of probability of arrest and police per capita with outlier:',
      round(cor(crime.data$prbarr,
                crime.data$polpc), 2))
paste('Correlation of probability of arrest and police per capita with outlier removed
      round(cor(crime.data.without115$prbarr,
                log(crime.data.without115$polpc),
                use='pairwise.complete.obs'), 2))

cat('\n')
```

'Correlation of probability of arrest and police per capita with outlier: 0.43'
'Correlation of probability of arrest and police per capita with outlier removed: -0.13'



It's clear that this observation dictates the nature of the relationship between the number of police per capita and the probability of arrest. With county 115 included, the correlation between *polpc* and *prbarr* is positive and fairly strong. Removal of this observation dramatically changes the nature of the relationship to a negative correlation. We therefore decide to permanently remove county 115 from our data.

```
In [8]: crime.data <- crime.data.without115
```

1.3.3 Logarithmic Scaling

It is important to note here that data for the number of arrests is obtained from FBI's Uniform Crime Reports while the number of convictions is obtained from the North Carolina Department of Correction. This is concerning and introduces some uncertainty in terms of interpreting the probability of conviction variable, as the numerator and denominator stem from different data sources. Issues of different definitions and methods of collection come into effect when data is obtained from different sources, and this comes into sharp focus when the probability of conviction variable is scrutinized. While this limits the usefulness of *prbconv* the outliers observed are nonetheless deemed valid because they simply highlight an incongruous artifact of the data, an artifact that is prevalent across all counties - different underlying data sources for arrests and convictions.

This leads us to think that all probabilities and wages variables should be used in the logarithmic scale. This decision is justified based on the nature of these variable categories. Specifically, we've already seen that the probability variables are rates, in reality, between two different data sources or timeframes. For example, we saw that the probability of conviction is the fraction

$$\frac{convictions}{arrests}$$

.

When we study the effect of a variable in a complex system, we have the notion of **ceteris paribus** which means "other (relevant) factors being equal" which plays an important role in causal analysis. On the other hand, the rate variables can be controlled in two different ways, the numerator, the number of convictions in this case, and the denominator, the number of arrests. In other words, one could conclude that an increase of the conviction probability means an increase of the convictions, but this is not necessarily the case, it could in practice mean a decrease of the arrests.

Also, another challenge is the relative comparison of these rates between counties of different clusters (counties with different arresting behaviors). When working on the logarithmic scale of these rate variables, there's the advantage of removing the denominator effect - assuming ceteris paribus - and comparing percentage increases in the numerator, something that can better reflect our variable interpretation.

$$\Delta(log(prbconv)) = log(prbconv_2) - log(prbconv_1)$$

$$\Delta(log(prbconv)) = log(convictions_2/arrests) - log(convictions_1/arrests)$$

$$\Delta(log(prbconv)) = log(convictions_2) - log(arrests) - log(convictions_1) + log(arrests)$$

$$\Delta(log(prbconv)) = log(convictions_2) - log(convictions_1)$$

$$\Delta(log(prbconv)) = log(convictions_2/convictions_1)$$

$$\Delta(log(prbconv)) \simeq \Delta(convictions) * 100\%$$

The same idea applies to the wage-based variables. The wage is meaningful when measured relatively to the local cost of life. Because the wage variables are not normalized in this way, we cannot any more interpret the absolute values but only percentages of changes.

Next, we transform some of the variables to the logarithmic scale but for simplicity we keep using their original name when referring to them.

1.3.4 Extreme Outliers Threshold

Extreme outliers have the potential to influence the nature of associations between variables and so it is necessary to mitigate against the overbearing effects that these values may exert on our regression model. **An extreme outlier is defined as any value that is greater than 3fs from the nearest fourth (Devore, 2015)**.

1.3.5 Variables prbarr and prbconv

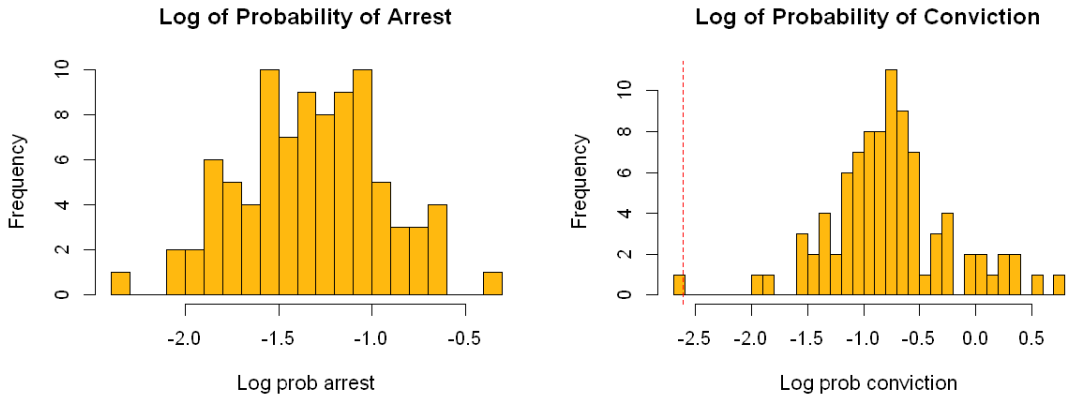
Variable *prbarr* is the probability of arrest proxied by the ratio of arrests to offenses. Based on the above analysis, we will transform them in the logarithmic scale and plot the histograms to check for outliers.

In [9]:
```r
# Arrest outlier threshold
hist.with.threshold <- function(x,
                                main,
                                xlab,
                                col,
                                print.left.text = FALSE,
                                print.right.text = FALSE)
    {
    stats <- boxplot.stats(x)
    lower <- stats$stats[2]
    upper <- stats$stats[4]
    fs <- upper - lower

    hist(x,
         breaks = 25,
         main =  main,
         col = col, cex.main = 1.2,
         cex.lab = 1.1, cex.axis = 1,
         xlab = xlab, ylab = 'Frequency')

    abline(v = upper + 3 * fs,
           col = 'red',
           lty = 2)

    abline(v = lower - 3 * fs,
           col = 'red',
           lty = 2)

    if(print.right.text)
        text(x = (upper + 3 * fs) + 0.2,
             y = 15,
             cex = 0.6,
             labels = c('Extreme threshold'))

    if(print.left.text)
        text(x = (lower - 3 * fs) + 0.2,
             y = 15,
             cex = 0.6,
             labels = c('Extreme threshold'))

    return ((sum(x > upper + 3 * fs | x < lower - 3 * fs)) )
}

impute.outliers <- function(x)
    {
    stats <- boxplot.stats(x)
    lower <- stats$stats[2]
    upper <- stats$stats[4]
```

```
        fs <- upper - lower

        x <- ifelse(x > upper + 3 * fs, mean(x),
        ifelse(x < lower - 3 * fs, mean(x), x))

        return (x)
    }
```

```
In [10]: plot.settings(0.4, 10, 1, 2)

        crime.data$prbarr.log <- log(crime.data$prbarr)
        crime.data$prbconv.log <- log(crime.data$prbconv)

        outliers <- hist.with.threshold(crime.data$prbarr.log,
                            'Log of Probability of Arrest',
                            'Log prob arrest',
                            'darkgoldenrod1')

        paste('Number logarithmic probability of arrest outliers :',
            outliers)

        outliers <- hist.with.threshold(crime.data$prbconv.log,
                            'Log of Probability of Conviction',
                            'Log prob conviction',
                            'darkgoldenrod1')

        paste('Number logarithmic probability of conviction outliers :',
            outliers)
```

'Number logarithmic probability of arrest outliers : 0'
'Number logarithmic probability of conviction outliers : 1'



We want to examine what is the effect of this variables transformation to the correlation with the rate of crime:

```
In [11]: paste('Correlation of crmrte with prbarr:',
               round(cor(crime.data$crmrte,
                         crime.data$prbarr), 2))
         paste('Correlation of crmrte with log(prbarr):',
               round(cor(crime.data$crmrte,
                         crime.data$prbarr.log,
                         use='pairwise.complete.obs'), 2))

         paste('Correlation of crmrte with prbconv:',
               round(cor(crime.data$crmrte,
                         crime.data$prbconv), 2))
         paste('Correlation of crmrte with log(prbconv):',
               round(cor(crime.data$crmrte,
                         crime.data$prbconv.log,
                         use='pairwise.complete.obs'), 2))
         cat('\n')
```

'Correlation of crmrte with prbarr: -0.38'
'Correlation of crmrte with log(prbarr): -0.39'
'Correlation of crmrte with prbconv: -0.36'
'Correlation of crmrte with log(prbconv): -0.34'

We notice that for *prbarr* the correlation slightly increased, and for *prbconv* decreased. In both cases, the effect was negligible given the fact that we were able to keep all of our data in the dataset and not impute any outlier values. Also, we note here the correction in the skewness, something that will become important further on in our analysis.

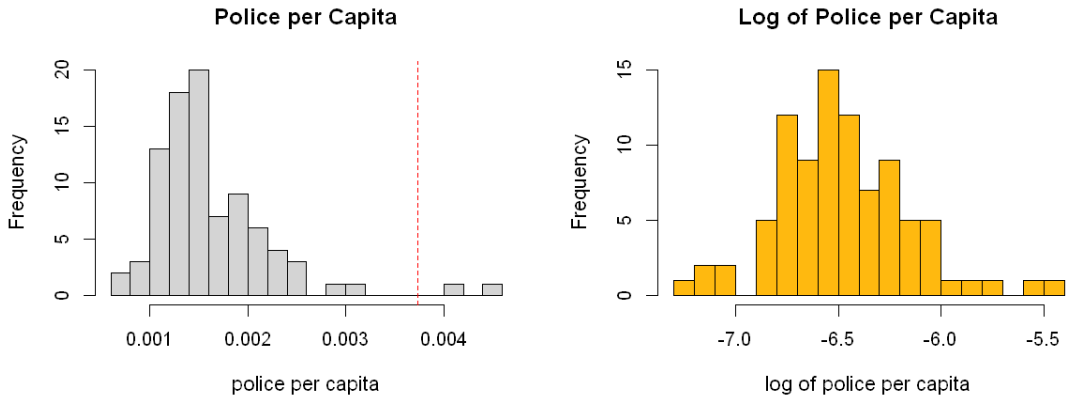
1.3.6 Variable polpc

Next we analyse the variable *polpc*.

```
In [12]: plot.settings(0.4, 10, 1, 2)

         outliers <- hist.with.threshold(crime.data$polpc,
                                         'Police per Capita',
                                         'police per capita',
                                         'grey83')


         paste('Number of police per capita outliers :',
             outliers)

         crime.data$polpc.log <- log(crime.data$polpc)

         outliers <- hist.with.threshold(crime.data$polpc.log,
                                         'Log of Police per Capita',
```

```
                        'log of police per capita',
                        'darkgoldenrod1')

paste('Number of police per capita outliers imputed:',
    outliers)
```

'Number of police per capita outliers : 2'
'Number of police per capita outliers imputed: 0'



We can see that there are two extreme outliers which we won't have to impute by transforming *polpc* with a log. We also notice the skewness improvement.

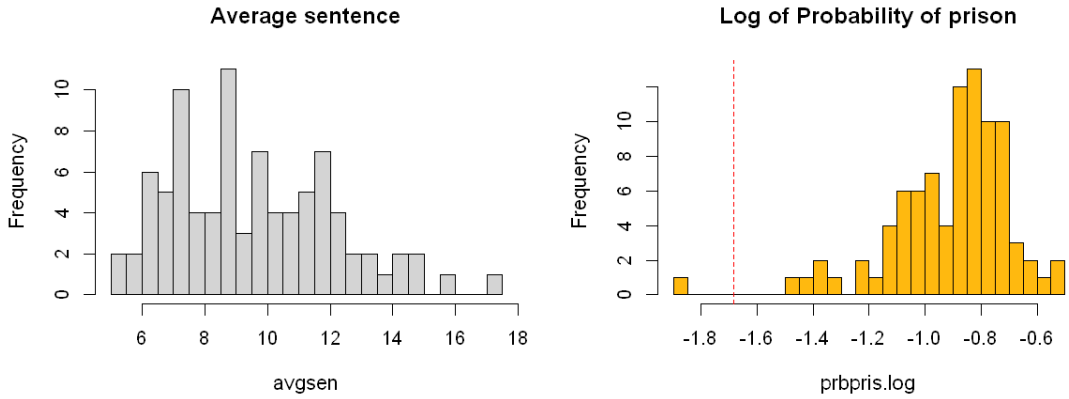
1.3.7 Variables avgsen and prbpris

Additional variables under consideration for inclusion in our linear regression models were also checked for the presence of extreme outliers such as *avgsen* and *prbpris* - none were identified. Nevertheless, we decided to transform the probability of prison to the logarithmic scale, because it makes more sense to interpret this variable as a percentage of effect. By doing so, an extreme outlier is revealed, which we impute using the same method as before.

```
In [13]: plot.settings(0.4, 10, 1, 2)

         crime.data$prbpris.log <- log(crime.data$prbpris)

         # Tax revenue plots
         par(mfrow = c(1, 2))

         outliers <- hist.with.threshold(crime.data$avgsen,
                             'Average sentence',
                             'avgsen',
                             'grey83')

         paste('Number of average sentence outliers:',
```

```
    outliers)

outliers <- hist.with.threshold(crime.data$prbpris.log,
                    'Log of Probability of prison',
                    'prbpris.log',
                    'darkgoldenrod1')

paste('Number of log of probability of prison outliers:',
    outliers)

crime.data$prbpris.log.imp <- impute.outliers(crime.data$prbpris.log)
```

'Number of average sentence outliers: 0'
'Number of log of probability of prison outliers: 1'



1.3.8 Variables pctmin80 and taxpc

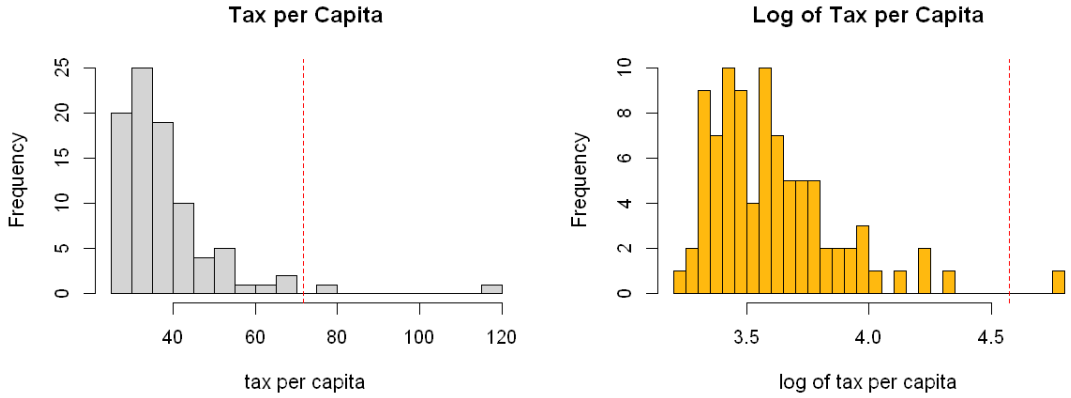
The percentage of population that are a minority is under consideration for inclusion in our regression models but no extreme outliers were identified when checked.

Likewise, we examined *taxpc* and identified extreme outliers. As we already briefly mentioned, we will treat all wage-based variables in the logarithmic scale. The two initial data points exceeding the 3fs threshold in the logarithmic scale become a single extreme outlier, which subsequently is replaced with the mean value for log tax revenue which had a notable effect on the correlation between *taxpc* and *crmrte*.

```
In [14]: plot.settings(0.4, 10, 1, 2)

crime.data$taxpc.log <- log(crime.data$taxpc)

outliers <- hist.with.threshold(crime.data$taxpc,
                    'Tax per Capita',
                    'tax per capita',
```

```
                    'grey83')

paste('Number of probability of tax per capita outliers :',
    outliers)


outliers <- hist.with.threshold(crime.data$taxpc.log,
                    'Log of Tax per Capita',
                    'log of tax per capita',
                    'darkgoldenrod1')

paste('Number of probability of log of tax per capita outliers :',
    outliers)

crime.data$taxpc.log.imp <- impute.outliers(crime.data$taxpc.log)

paste('Correlation with taxpc:',
      round(cor(crime.data$crmrte,
                crime.data$taxpc), 2))
paste('Correlation with log(taxpc):',
      round(cor(crime.data$crmrte,
                crime.data$taxpc.log,
                use='pairwise.complete.obs'), 2))
paste('Correlation with imputed log(taxpc):',
      round(cor(crime.data$crmrte,
                crime.data$taxpc.log.imp,
                use='pairwise.complete.obs'), 2))
```

'Number of probability of tax per capita outliers : 2'
'Number of probability of log of tax per capita outliers : 1'
'Correlation with taxpc: 0.44'
'Correlation with log(taxpc): 0.4'
'Correlation with imputed log(taxpc): 0.32'



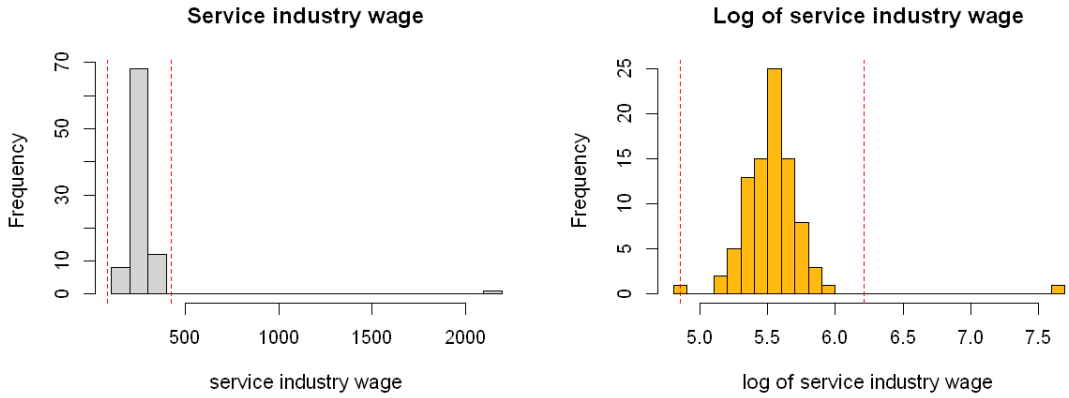
1.3.9 Variable wser

The final variable under consideration for the removal of outliers is average service industry wage. The mean average wage in this sector is 275, and the maximum average wage is 2,177, which is over 9 standard deviations from the mean. The service industry isn't typically a high paying sector, and as this value varies so considerably from the mean, it is feasible that this data point is an error.

```
In [15]: plot.settings(0.4, 10, 1, 2)

         paste('Number of standard deviations max value is from the mean:',
               round((max(crime.data$wser) -
                      mean(crime.data$wser)) /
                      sd(crime.data$wser), 2))

         crime.data$wser.log <- log(crime.data$wser)

         outliers <- hist.with.threshold(crime.data$wser,
                             'Service industry wage',
                             'service industry wage',
                             'grey83')

         paste('Number of probability of service industry wage outliers :',
             outliers)


         outliers <- hist.with.threshold(crime.data$wser.log,
                             'Log of service industry wage',
                             'log of service industry wage',
                             'darkgoldenrod1')

         paste('Number of probability of log service industry wage outliers :',
             outliers)

         crime.data$wser.log.imp <- impute.outliers(crime.data$wser.log)

         paste('Correlation with wser:',
               round(cor(crime.data$crmrte,
                         crime.data$wser), 2))
         paste('Correlation with log(wser):',
               round(cor(crime.data$crmrte,
                         crime.data$wser.log,
                         use='pairwise.complete.obs'), 2))
         paste('Correlation with imputed log(wser):',
               round(cor(crime.data$crmrte,
```

```
                    crime.data$wser.log.imp,
                    use='pairwise.complete.obs'), 2))
```

'Number of standard deviations max value is from the mean: 9.12'
'Number of probability of service industry wage outliers : 1'
'Number of probability of log service industry wage outliers : 1'
'Correlation with wser: -0.06'
'Correlation with log(wser): 0.09'
'Correlation with imputed log(wser): 0.32'

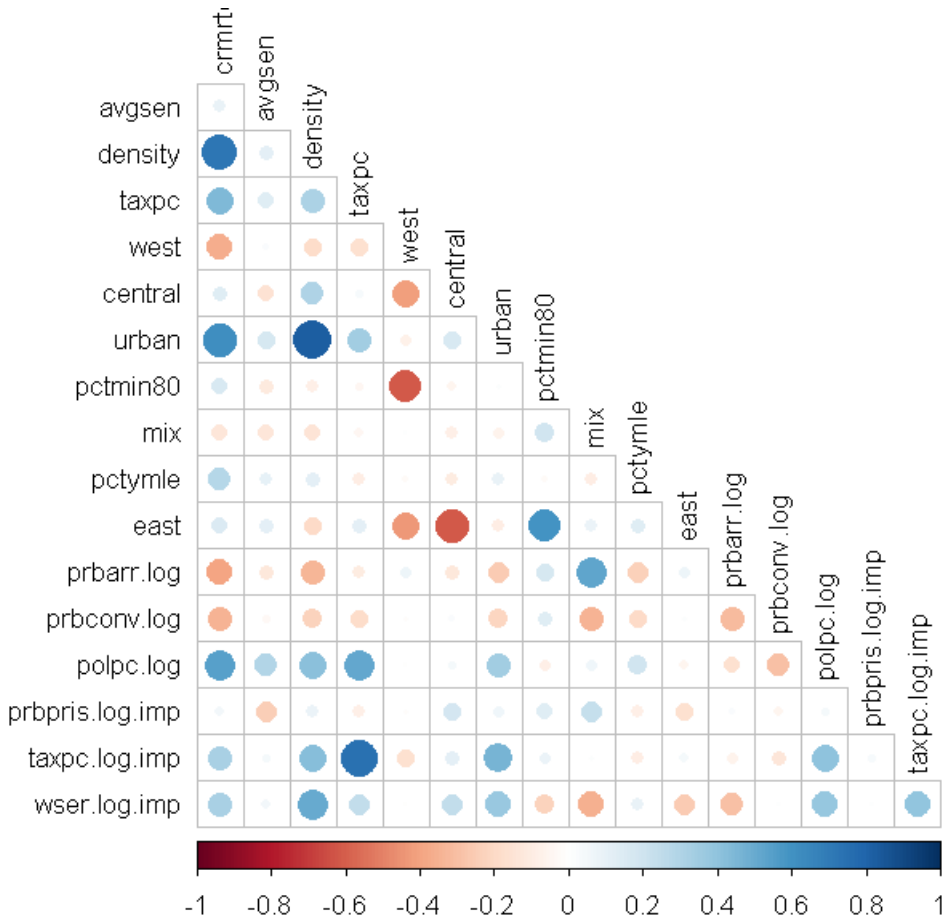


For reasons of interpretation as analyzed in the introduction, we will be using the logarithmic scale of the service industry wage. There, we still see the same extreme outlier, so we replace it with the mean. The log scale and the replacement of this observation changed the direction of the relationship between average service industry wage and crime rate. With the outlier included, the correlation between average service industry wage and crime rate was negative, -0.06. Following the removal of the outlier, the nature of the correlation changed to a positive 0.32.

1.4 Correlation Matrix

```
In [16]: # Resize the plot
         plot.settings(1, 5, 1, 1)

         # Remove redundant variables
         C <- crime.data[ , !names(crime.data) %in% c('year',
                                                      'county',
                                                      'prbarr',
                                                      'prbconv',
                                                      'polpc',
                                                      'prbpris',
                                                      'prbpris.log',
                                                      'taxpc.log',
                                                      'wcon',
                                                      'wtuc',
```

```
                                        'wtrd',
                                        'wfir',
                                        'wmfg',
                                        'wfed',
                                        'wsta',
                                        'wloc',
                                        'wcon',
                                        'wser',
                                        'wser.log')]

# Create correlation matrix
C <- cor(C[sapply(C, is.numeric)])

# Plot matrix
corrplot(C, type = 'lower',
         method = 'circle',
         tl.col = 'black',
         tl.cex = 0.87,
         diag = FALSE)
```



```
In [42]: crmrte.sorted.correlations <- sort(C[, "crmrte"], decreasing = TRUE)
         data.frame(crmrte.sorted.correlations)

         cat('\n')
```

	crmrte.sorted.correlations
crmrte	1.00000000
density	0.72783412
urban	0.61792107
polpc.log	0.54674727
taxpc	0.44291442
wser.log.imp	0.32273903
taxpc.log.imp	0.32021599
pctymle	0.28606600
pctmin80	0.16135085
east	0.15413561
central	0.13056154
avgsen	0.09477333
prbpris.log.imp	0.05736180
mix	-0.13983223
prbconv.log	-0.34094461
west	-0.36117669
prbarr.log	-0.39498529

The correlation matrix above is a useful tool that will help guide aspects of the model building process.

At first glance, we observe a few intuitive relationships: density appears positively associated with crime while probability of conviction is negatively associated with it. Higher wages are associated with density. Counties in the west region have very low percentage of minorities while the other regions have a higher percentage of minorities, indicating that the population demographic of these regions differs considerably. It can also be seen that regions west and east have a weak negative correlation with the average wages across sectors. In contrast, the central region is positively correlated with average wages, which is likely attributable to the fact that it is a densely populated, urban area.

The subsequent section will focus on the model building process and will explore the casual relationships of these variables in greater detail.

1.5 Regression Model Building

1.5.1 Preface

Now that we've reviewed our data, we'll fulfill the first part of our commission for this campaign by building three OLS regression models to describe the relationship between crime and the variables we've observed that drive it. The following values will guide our decisions on how to build and evaluate our models:

1. **Fitness**: our models should aim to describe a maximum of variance in crime
2. **Parsimony**: our models should aim to introduce a minimum of complexity in their description of crime
3. **Usefulness to the campaign**: our models should describe a causal, actionable relationship between crime and the factors that contribute to its occurrence, especially those that policymakers can influence

Naturally, we'll trade off value across these dimensions, but the optimal model should attempt to optimize across all of them.

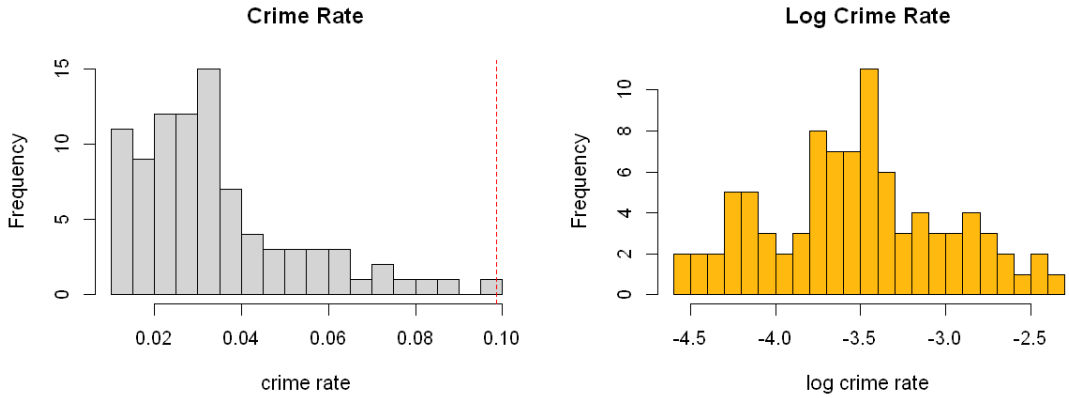
As a preview, our base regression model will prioritize usefulness and parsimony by focusing on the impact of the probability of arrest, probability of prison sentence, and average sentence length on crime rates in our data. Our second model will represent our best effort to prioritize usefulness, fitness, and parsimony by examining the impact of the key explanatory variables on crime rate. And our third model will prioritize usefulness and fitness by examining the impact of all non-problematic explanatory variables on crime rate.

1.5.2 Regressand: Crime Rate

Before we make a case for which explanatory variables to include in our model, we'll take a closer look at our key outcome variable of interest, *crmrte*.

```
In [18]: plot.settings(0.4, 10, 1, 2)

         # Summary stats for crmrte
         summary(crime.data$crmrte)

         crime.data$crmrte.log <- log(crime.data$crmrte)

         # Tax revenue plots
         par(mfrow = c(1, 2))

         outliers <- hist.with.threshold(crime.data$crmrte,
                              'Crime Rate',
                              'crime rate',
                              'grey83')

         outliers <- hist.with.threshold(crime.data$crmrte.log,
                              'Log Crime Rate',
                              'log crime rate',
                              'darkgoldenrod1')
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01062 0.02157 0.03002 0.03382 0.04086 0.09897
```



crmrte is a ratio of crimes per person and is distributed between 0.01 and 0.1. The right skewness of the distribution indicates that most counties' crime rates are less than average, which matches intuition, as we'd expect there to be a minority of urban counties where crime rates are high and a majority of rural or suburban counties where crime rates are more modest.

Given *crmrte* is a ratio that we expect to be slightly above 0 and very likely not exactly 0, which would imply 0 recorded crimes in a county in a year, we'll transform this variable by taking its log in the regression models that follow. This should improve the fitness of our models given the right skewness in the distribution of *crmrte* since transforming a variable with a log reduces the impact of outliers. The upshot for interpreting regression coefficients is that a unit change in an explanatory variable—for instance, a unit change in *avgsen*—will result in a 100 * β% change in *crmrte*.

1.5.3 Key Explanatory Variables of Interest

Justifying the Inclusion of Key Explanatory Variables We're most interested in 3 key explanatory variables of crime rate and will use these in our base regression model: *prbarr*, *avgsen*, and *polpc*. We restrict our initial model to these variables because our party believes there's an opportunity to influence them with policies concerning law and order.

prbarr, measured as the ratio of arrests to offenses, can be influenced with changes in policing strategies and programs designed to improve the ratio such that more arrests are made on appropriate offenses, and ultimately, fewer offenses are committed, which would further improve the ratio.

prbpris, measured as the ratio of convictions that result in a prison sentence to total convictions, and *avgsen*, days of the average prison sentence, can be influenced with policy that changes the severity of punishment for crime we've historically mishandled.

Justifying the Exclusion of Other Explanatory Variables It's worth noting that we didn't include in our base model other factors in the domain of law of order that policy might also influence, specifically *prbconv* and *polpc*.

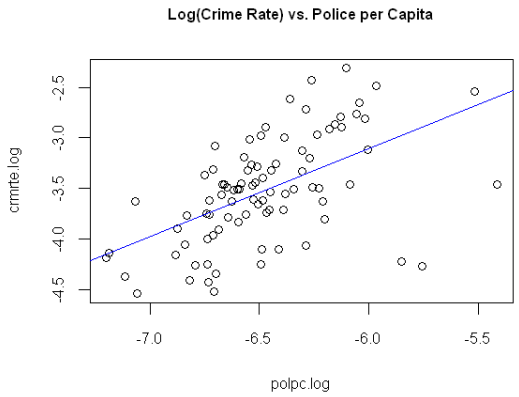
prbconv, measured as the ratio of convictions to arrests, doesn't seem ethically appropriate to call influence-able because we're assuming fairness of judicial decisions in accordance with laws regarding conviction. While policy may change the law related to severity of punishment, it

should not have an influence on the fairness of decisions made in the judicial process. While we recognize that judicial fairness is not guaranteed practically for all people and every judge, we'll assume for this analysis that defendants are convicted if they are found guilty of their crime in a fair trial.

polpc, measured as the police per capita in a county, is an interesting variable that deserves special attention in this analysis. The positive correlation between it and *crmrte* represents a possibly counterintuitive association: shouldn't the presence of police deter crime?

We described *polpc* briefly in our EDA to identify outliers, but we'll take a closer look at its positive relationship with *crmrte* below.

```
In [19]: # Resize the plot
         plot.settings(0.4, 10, 1, 2)

         # Plot crmrte.log vs. polpc.log
         plot(crime.data$polpc.log,
              crime.data$crmrte.log,
              main = "Log(Crime Rate) vs. Police per Capita",
              cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
              xlab = 'polpc.log', ylab = 'crmrte.log')

         abline(fit <- lm(crime.data$crmrte.log ~ crime.data$polpc.log), col = 'blue')
```



While this relationship may seem surprising, we should also acknowledge that police deployment decisions are intentional and not made agnostic to crime. Following this line of reasoning, crime rate would be expected to rise where police are deployed because there are more police in the area to catch and report crime. Furthermore, an increase in the presence of police may signal to residents that they're living in a dangerous location. This could drive wealthier residents to locations with less crime and fewer police which, in turn, drives crime rate up because the location perceived to be dangerous has a smaller population and those remaining are the poorer residents, likely more prone to criminal activity due to their poverty and increasingly disenfranchised, and potentially aggressive, with the police. Under this theory of change, socio-economic segregation would accelerate in a location as *crmrte* or *polpc* rose. The upshot is that crime rate and policing work together to drive each other up or down.

And yet there is still the intuition that effective policing should reduce criminal activity.

This makes the decision to include *polpc* in our models tricky. While we'd like to observe the causal effect on crime of incrementing the police force within a county, our cross-sectional data is limited in its usefulness to this end. We'll address this limitation further in our omitted variables discussion but given *polpc* in our dataset is likely dependent on *crmrte*, we've decided to exclude it from our base model because it's not a factor we could confidently recommend influencing.

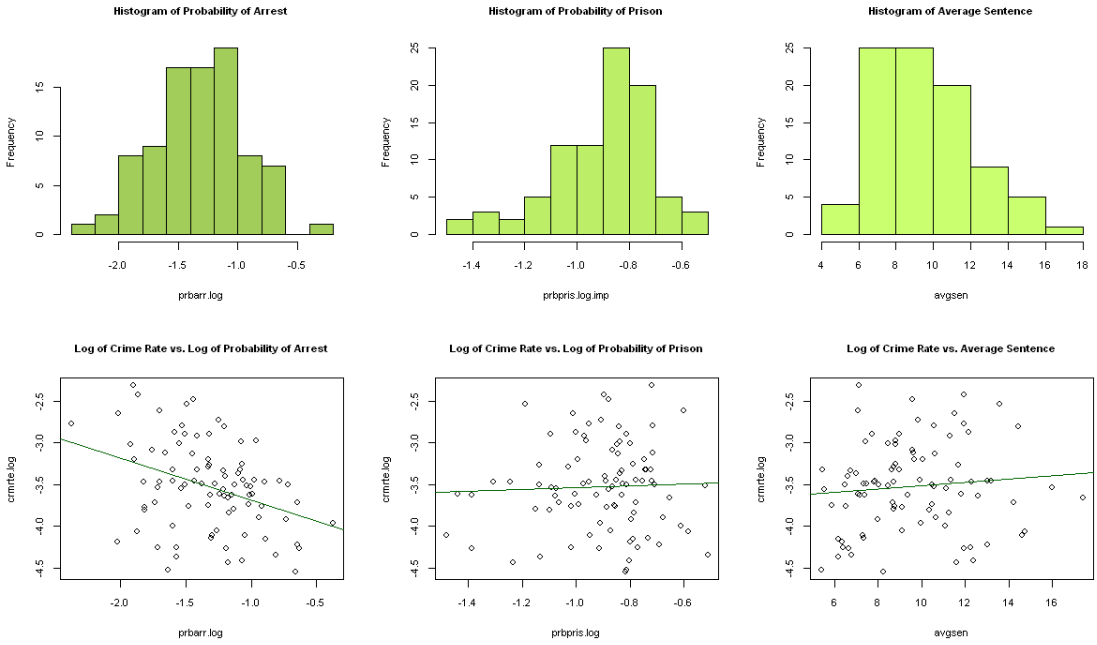
As we've stated, our party's interest in influencing factors in the domain of law and order has constrained the building of our base model. As such, we'll exclude economic and fiscal factors at this stage. We'll also exclude county regional and demographic factors until we build our more robust second model, thereby preserving parsimony.

Finally, we won't include *mix* in this model as it is descriptive of *crmrte*, not plausibly seen as a driver of it.

Exploratory Analysis of Explanatory Variables We previously conducted a thorough analysis of each of our explanatory variables—*prbarr*, *prbpris*, and *avgsen*—in the Outliers section, providing justification for transforming *prbarr* and *prbpris* with a natural logarithm. The distributions of those variables and their association with *crmrte* is re-summarized below.

```
In [20]: plot.settings(0.6, 10, 2, 3)

         # Histogram of prbarr.log
         hist(crime.data$prbarr.log,
             main = "Histogram of Probability of Arrest",
             xlab = "prbarr.log", ylab = "Frequency",
             cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
             col = "darkolivegreen3")

         # Histogram of prbpris.log.imp
         hist(crime.data$prbpris.log.imp,
             main = "Histogram of Probability of Prison",
             xlab = "prbpris.log.imp", ylab = "Frequency",
             cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
             col = "darkolivegreen2")

         # Histogram of avgsen
         hist(crime.data$avgsen,
             main = "Histogram of Average Sentence",
             xlab = "avgsen", ylab = "Frequency",
             cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
             col = "darkolivegreen1")

         # Plot of prbarr.log
         plot(crime.data$prbarr.log,
              crime.data$crmrte.log,
              main = "Log of Crime Rate vs. Log of Probability of Arrest",
              cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
              xlab = 'prbarr.log', ylab = 'crmrte.log')
         abline(lm(crime.data$crmrte.log ~ crime.data$prbarr.log), col = 'darkgreen')
```

```r
# Plot of prbpris.log.imp
plot(crime.data$prbpris.log.imp,
     crime.data$crmrte.log,
     main = "Log of Crime Rate vs. Log of Probability of Prison",
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
     xlab = 'prbpris.log', ylab = 'crmrte.log')
abline(lm(crime.data$crmrte.log ~ crime.data$prbpris.log.imp), col = 'darkgreen')

# Plot of avgsen
plot(crime.data$avgsen,
     crime.data$crmrte.log,
     main = "Log of Crime Rate vs. Average Sentence",
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
     xlab = 'avgsen', ylab = 'crmrte.log')
abline(lm(crime.data$crmrte.log ~ crime.data$avgsen), col = 'darkgreen')
```



The log transformed *prbarr* is approximately normally distributed and displays an obvious negative relationship with the log transformed *crmrte*. The log transformed *prbpris*, despite conversion, is still slightly skewed to the left and displays a very weak positive relationship with the log transformed *crmrte*. *avgsen* is skewed to the right and displays a weak positive relationship with *crmrte*.

We'll examine how these variables' relationships with *crmrte* change when we consider them in a multivariate regression model below, and we'll examine how their skewness influences our homoskedasticity assumption in our model evaluation.

1.5.4 Base Regression Model

Our base regression model contains our three key explanatory variables, *prbarr*, *prbpris*, and *avgsen*. Thus, our model 1 regression equation is:

$$crmrte.log = \beta_0 + \beta_1 prbarr.log + \beta_2 prbpris.log.imp + \beta_3 avgsen + u$$

```
In [21]: # Base regression model
         model1 <- lm(crime.data$crmrte.log ~
                      crime.data$prbarr.log +
                      crime.data$prbpris.log.imp +
                      crime.data$avgsen,
                      data = crime.data,  na.action = 'na.exclude')

         stargazer(model1, type = 'text')

         cat('\n')
```

```
================================================
                        Dependent variable:
                    ----------------------------
                               crmrte.log
------------------------------------------------
prbarr.log                      -0.490***
                                 (0.141)

prbpris.log.imp                  0.182
                                (0.277)

avgsen                           0.014
                                (0.021)

Constant                        -4.142***
                                 (0.329)

------------------------------------------------
Observations                       89
R2                                0.137
Adjusted R2                       0.107
Residual Std. Error         0.494 (df = 85)
F Statistic             4.506*** (df = 3; 85)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

The coefficients of our model show us that (all statements ceteris paribus):

1. For a 1% increase in probability of arrest, there's a 0.49% decrease in crime rate
2. For a 1% increase in the probability of inprisonmnet, there's a 0.18% increase in crime rate
3. For a 1 day increase in the average sentence, there's a 1.4% increase in crime rate

The direction of each of the marginal effects of our outcome variables matches the direction of their respective correlations. However, the sign on the estimators for *prbpris* and *avgsen* are counterintuitive—common sense would suggest an increase in *prbpris* or *avgsen* would result in a decrease in *crmrte*. Regardless, no result seems practically significant given the magnitude of each coefficient is small.

This result lends credence to the hypothesis that omitted variables are biasing our results and that we'll need to include them to draw meaningful conclusions about the impact of our influence-able variables on *crmrte*. Additionally, the fitness of our model is poor, explaining a modest 11% of the variation in *crmrte*, further indicating that our model would benefit from additional explanatory variables.

By controlling for regional, demographic, and fiscal variables in our second model, we hope to limit the bias in these coefficients.

1.5.5 Second Regression Model

Our second regression model will add the following variables to our base model: *prbconv*, *density*, *west*, *I(density* west), pctmin80, prbconv, polpc, wavg* (average wage across sectors), and *taxpc*.

Thus, our model 2 regression equation is:

$$crmrte.log = \beta_0 + \beta_1 prbarr.log + \beta_2 prbpris.log.imp +$$
$$\beta_3 avgsen + \beta_4 prbconv.log + \beta_5 I(density * west) + \beta_6 taxpc.log.imp +$$
$$\beta_7 pctmin80 + \beta_8 polpc.log + \beta_9 wavg.log + u$$

We've chosen these variables primarily because of their comparatively strong correlation with *crmrte*.

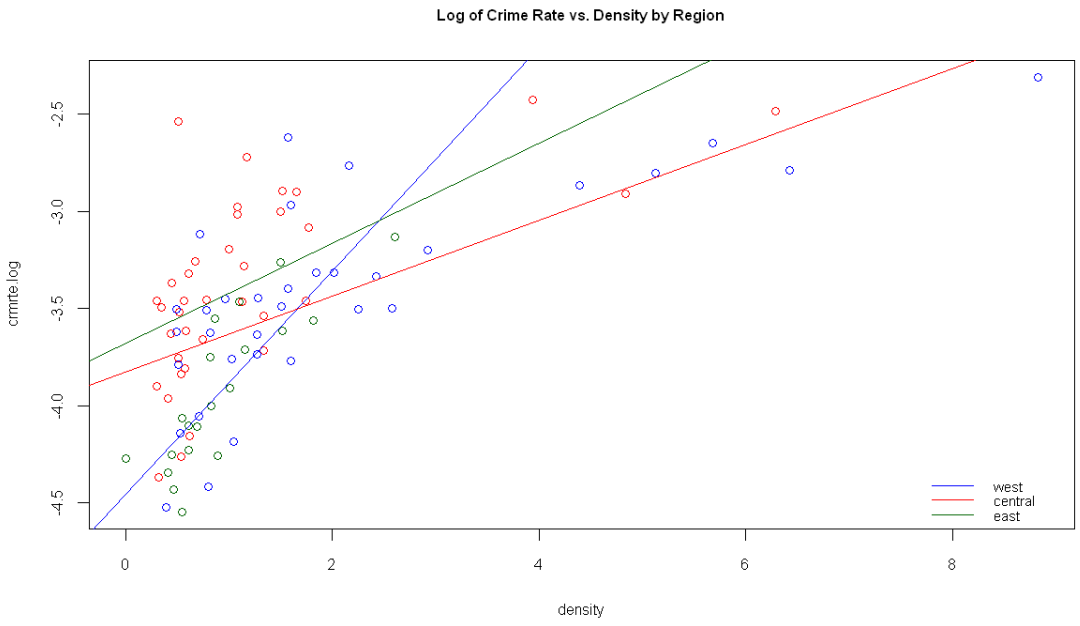
As we noted in the justification section for our base model, we don't think *prbconv* is ethically influence-able, and for that reason, we didn't include it in our base model. However, because of its strong correlation with *crmrte* (-0.34), we've included it here to improve model fitness and to minimize the bias in our influence-able estimators. As stated in our EDA, and like *prbarr*, we took the log of *prbconv* to ease interpretability and reduce the impact of outliers.

For our demographic variables, density is the variable with the strongest correlation with *crmrte* (0.73). For that reason and because of its value as a measure of urbanicity, we'll include it. We're also including *pctmin80* at this stage because of its correlation with *crmrte* (0.18).

Upon closer inspection, as pictured below, we noted significant discrepancies between *crmrte* and *density* by region, so we decided to include an interaction term of the two to get a sense for the variation in *crmrte* by the variable most correlated with it on the regional dimension that seemed most unique.

```
In [44]: # Resize the plot
         plot.settings(0.6, 10, 1, 1)

         # Add color column to crime.data
         crime.data$region.col <- ifelse(crime.data$west == 1,
                                         'darkgreen',
```

```
                                    ifelse(crime.data$central == 1,
                                           'blue', 'red'))

#filter crime.data by region
crime.data.west <- crime.data[crime.data$west == 1, ]
crime.data.central <- crime.data[crime.data$central == 1, ]
crime.data.east <- crime.data[crime.data$east == 1, ]

plot(crime.data$density,
     crime.data$crmrte.log,
     main = "Log of Crime Rate vs. Density by Region",
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
     xlab = 'density', ylab = 'crmrte.log',
     col = crime.data$region.col)

abline(lm(crime.data.west$crmrte.log ~ crime.data.west$density),
       col = 'blue')
abline(lm(crime.data.central$crmrte.log ~ crime.data.central$density),
       col = 'red')
abline(lm(crime.data.east$crmrte.log ~ crime.data.east$density),
       col = 'darkgreen')

legend('bottomright',
       c('west', 'central', 'east'),
       lty=1, col=c('blue', 'red', 'darkgreen'),
       bty='n', cex=0.75)
```



Finally, we've included two fiscal/economic variables, *taxpc*, and *wavg*. The correlation of the former with *crmrte* (0.449) suggests it should improve fitness in our model. In an effort to improve model parsimony, we decided to take an average of the 9 wage category fields and create *wavg* to observe the influence a county's average wage across sectors has on *crmrte*. Again, per our EDA above, we transform both variables with a logarithm.

```
In [23]: # Take the average of the wages
         crime.data$wavg.log <- log((crime.data$wcon + crime.data$wtuc + crime.data$wtrd +
                                crime.data$wfir + exp(1)^crime.data$wser.log.imp + crime.data$wmf
                                crime.data$wfed + crime.data$wsta + crime.data$wloc) / 9)

In [24]: # Second regression model
         model2 <- lm(crime.data$crmrte.log ~
                      crime.data$prbarr.log +
                      crime.data$prbpris.log.imp +
                      crime.data$avgsen +
                      crime.data$prbconv.log +
                      crime.data$density * crime.data$west +
                      crime.data$taxpc.log.imp +
                      crime.data$pctmin80 +
                      crime.data$polpc.log +
                      crime.data$wavg.log,
                      data = crime.data,  na.action = 'na.exclude')

         stargazer(model2, type = 'text')

         cat('\n')


================================================
                        Dependent variable:
                    ---------------------------
                              crmrte.log
------------------------------------------------
prbarr.log                     -0.378***
                                (0.092)

prbpris.log.imp                 -0.131
                                (0.150)

avgsen                          -0.015
                                (0.012)

prbconv.log                    -0.280***
                                (0.063)

density                        0.102***
                                (0.027)
```

```
west                            -0.639***
                                 (0.144)

taxpc.log.imp                   -0.410***
                                 (0.140)

pctmin80                        0.007***
                                 (0.002)

polpc.log                       0.593***
                                 (0.111)

wavg.log                         0.518
                                (0.373)

west                            0.401***
                                 (0.110)

Constant                         -2.161
                                 (2.550)

-----------------------------------------------
Observations                       89
R2                                0.785
Adjusted R2                       0.754
Residual Std. Error         0.259 (df = 77)
F Statistic            25.550*** (df = 11; 77)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

The coefficients of our model show us that (all statements ceteris peribus):

1. (From -0.49%) For a 1% increase in probability of arrest, there's a 0.378% decrease in crime rate
2. (From +0.18%) For a 1% increase in probability of prison, there's a 0.13% decrease in crime rate
3. (From +1.4%) For a 1 day increase in average sentence, there's a 1.5% decrease in crime rate
4. For a 1% increase in probability of conviction, there's a 0.28% decrease in crime rate
5. For central and easter counties, for a unit increase in density (100 people per square mile), there's a 10.2% increase in crime rate. For western counties, for a unit increase in density, there's a 50.3% increase in crime rate.
6. For counties in the west, there's a 63.9% decrease in crime rate compared with central or eastern counties
7. For a 1% increase in taxpc, there's a 0.41% decrease in crime rate
8. For a 1 percentage point increase in percent minority, there's a 0.7% increase in crime rate

9. For a 1% increase in police per capita, there's a 0.593% increase in crime rate
10. For a 1% increase in average wage, there's a 0.518% increase in crime rate
11. For the interaction between density and west, there's a 40.1% increase in crime rate for western counties per unit increase in density on top of the 10.2% increase per unit increase in density

All signs on our coefficients match the direction of correlation except for a few notable switches: *prbpris*, *avgsen*, and *taxpc* for which the direction of correlation flipped, indicating that they had absorbed much of the positive relationship between *crmrte* and other variabes we added to this model, like *density*, in our base model. Between this model and our base model, the sign on the coefficients for *prbpris* and *avgsen* switched as well to match our intuition, which is encouraging. There was little change in *avgsen* between models.

Our interaction term shows what the graph justifying its inclusion displays: while counties in the west start with 64% discount on *crmrte* due to their lower density on average, a unit increase in density in a western county will drive a 40% greater increase in *crmrte* compared with central and eastern counties ceteris peribus.

Despite apparent statistical significance on many of our estimators, their practical significance is questionable. Only *density* has a > 1.5% impact on *crmrte* in either direction. *avgsen* at -1.5% seems next most practically significant, however, per our EDA, its range is between just 4 and 18 days, so there's limitation to the available unit-increase in this variable. No other variable's estimator seems clearly practically significant.

The set of variables with positive estimators in relation to *crmrte* are *density*, *pctmin80*, *polpc*, and *wavg*. As we mentioned in our examination of *polpc* earlier, we need to be careful interpreting this variable's estimator as *polpc* is just as likely the effect of *crmrte* as the cause of it. Our estimator on *wavg* is also fishy: intuitively, it would seem increasing wages, other factors held constant, would reduce *crmrte*. This leads us to believe that there are still likely omitted variables biasing these estimators up and away from 0 and biasing our negative coefficients as well.

The fitness of this model improved dramatically from our base model, now explaining 75% of the variation in *crmrte* validating that the variables we decided to include based on their correlation with *crmrte* improve the bias in our estimators.

```
In [25]: coeftest(model2, vcov = vcovHC, level = 0.05)

         cat('\n')
```

```
t test of coefficients:

                                Estimate Std. Error t value  Pr(>|t|)
(Intercept)                   -2.1610257  3.5966335 -0.6008 0.5497066
crime.data$prbarr.log         -0.3779736  0.1008773 -3.7469 0.0003441 ***
crime.data$prbpris.log.imp    -0.1308290  0.1687783 -0.7752 0.4406233
crime.data$avgsen             -0.0154319  0.0121058 -1.2748 0.2062292
crime.data$prbconv.log        -0.2804046  0.1095413 -2.5598 0.0124316 *
crime.data$density             0.1018902  0.0291295  3.4978 0.0007825 ***
crime.data$west               -0.6385657  0.1672550 -3.8179 0.0002705 ***
crime.data$taxpc.log.imp      -0.4102477  0.2017395 -2.0336 0.0454422 *
crime.data$pctmin80            0.0073580  0.0026782  2.7473 0.0074793 **
```

```
crime.data$polpc.log                 0.5934514  0.1635743  3.6280 0.0005114 ***
crime.data$wavg.log                  0.5177905  0.5169842  1.0016 0.3196935
crime.data$density:crime.data$west   0.4008536  0.1434597  2.7942 0.0065631 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With heteroskedacticity-robust standard errors, we have 8 significant variables for which we can reject the null hypothesis: *prbarr*, *prbconv*, *density*, *west*, *taxpc*, *pctmin*, *polpc*, and *I(density x west)*.

1.5.6 Second Regression Model (Restricted)

The intial specification for model 2 yielded three insignificant coefficients for the explantory variables: *avgsen*, *prbpris*, and *wavg*. In the interest of parsimony and to address a matter relevant to our research question on puruing a policy change in the space of law enforcement, a hypothesis test shall be conducted to ascertain whether *avgsen* and *prbpris* have an effect on crime rate once the remaining explantory variables in model 2 have been controlled for.

$$\text{Null hypothesis:} \quad H_0 : \beta_2 = 0, \beta_3 = 0$$
$$\text{Alternate hypothesis:} \quad H_1 : H_0 \text{ is not true.}$$

In [26]:
```
# Restricted model 2
model2.restricted <- lm(crime.data$crmrte.log ~
                crime.data$prbarr.log +
  #              crime.data$prbpris.log.imp +
  #              crime.data$avgsen +
                crime.data$prbconv.log +
                crime.data$density * crime.data$west +
                crime.data$taxpc.log.imp +
                crime.data$pctmin80 +
                crime.data$polpc.log +
                crime.data$wavg.log,
                 data = crime.data,  na.action = 'na.exclude')

# Check restricted model coefficients for significance
coeftest(model2.restricted, vcov = vcovHC)

# Test for joint significance
model2.rest.ftest <- waldtest(model2, model2.restricted, vcov = vcovHC)

# Create table of df, F statistic and p-value.
stargazer(model2.rest.ftest, type = 'text',
          omit.summary.stat = c('min', 'max', 'sd', 'p25', 'p75'))
```

```
        model2 <- model2.restricted

        cat('\n')
```

```
t test of coefficients:

                                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)                       -2.2015860  3.5694590 -0.6168 0.5391505
crime.data$prbarr.log             -0.3728721  0.0958304 -3.8910 0.0002072 ***
crime.data$prbconv.log            -0.2794686  0.1062610 -2.6300 0.0102597 *
crime.data$density                 0.1015771  0.0299615  3.3902 0.0010925 **
crime.data$west                   -0.6529515  0.1684014 -3.8774 0.0002171 ***
crime.data$taxpc.log.imp          -0.3932713  0.1976600 -1.9896 0.0500916 .
crime.data$pctmin80                0.0071934  0.0027271  2.6378 0.0100473 *
crime.data$polpc.log               0.5570449  0.1632619  3.4120 0.0010194 **
crime.data$wavg.log                0.4706950  0.5154238  0.9132 0.3639058
crime.data$density:crime.data$west 0.4103253  0.1503030  2.7300 0.0078065 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
==================
Statistic N  Mean
------------------
Res.Df    2 78.000
Df        1 -2.000
F         1 1.022
Pr(> F)   1 0.365
------------------
```

The resulting F-statistic represents the relative increase in the sum of square residuals (SSR) when moving from the unrestricted model to the restricted model. A F-statistic of 1.022 corresponds to a p-value of 0.365 which means we cannot reject the null hypothesis at the 5% significance level and so the variables *avgsen* and *prbpris* are jointly insignificant. That is to say, once the other explanatory variables have been controlled for, *prbpris* and *avgsen* have no effect on crime rate. In the interest of parsimony the jointly insignificant variables shall be removed from model 2, and we shall later refer to this test result in support of certain judicial policy recommendations.

1.5.7 Third Regression Model

Our third regression model will use almost all variables in our data set. The regression formula is:

$$crmrte.log = \beta_0 + \beta_1 prbarr.log + \beta_2 prbpris.log.imp +$$

$$\beta_3 avgsen + \beta_4 prbconv.log + \beta_5 polpc.log + \beta_6 density +$$
$$\beta_7 taxpc.log.imp + \beta_8 pctmin80 + \beta_9 west + \beta_{10} wcon.log +$$
$$\beta_{11} wtuc.log + \beta_{12} wtrd.log + \beta_{13} wfir.log +$$
$$\beta_{14} wser.log + \beta_{15} wmfg.log + \beta_{16} wfed.log +$$
$$\beta_{17} wsta.log + \beta_{18} wloc.log + \beta_{19} pctymle + u$$

We left out *urban* given its strong correlation with *density* and *mix* given it's descriptive of *crmrte*, not something that can be feasibly thought of as a contributing cause of it.

We broke out each wage category, and as before, transform them with a logarithm.

```
In [27]: # Wage variables all transformed with a log
         crime.data$wcon.log <- log(crime.data$wcon)
         crime.data$wtuc.log <- log(crime.data$wtuc)
         crime.data$wtrd.log <- log(crime.data$wtrd)
         crime.data$wfir.log <- log(crime.data$wfir)
         crime.data$wmfg.log <- log(crime.data$wmfg)
         crime.data$wfed.log <- log(crime.data$wfed)
         crime.data$wsta.log <- log(crime.data$wsta)
         crime.data$wloc.log <- log(crime.data$wloc)

In [28]: # Third regression model
         model3 <- lm(crime.data$crmrte.log ~
                      crime.data$prbarr.log +
                      crime.data$prbpris.log.imp +
                      crime.data$avgsen +
                      crime.data$prbconv.log +
                      crime.data$polpc.log +
                      crime.data$density +
                      crime.data$taxpc.log.imp +
                      crime.data$pctmin80 +
                      crime.data$west +
                      crime.data$wcon.log +
                      crime.data$wtuc.log +
                      crime.data$wtrd.log +
                      crime.data$wfir.log +
                      crime.data$wser.log.imp +
                      crime.data$wmfg.log +
                      crime.data$wfed.log +
                      crime.data$wsta.log +
                      crime.data$wloc.log +
                      crime.data$pctymle,
                      data = crime.data,  na.action = 'na.exclude')

         stargazer(model3, type = 'text')

         cat('\n')
```

	Dependent variable:
	crmrte.log
prbarr.log	-0.435***
	(0.103)
prbpris.log.imp	-0.190
	(0.161)
avgsen	-0.017
	(0.013)
prbconv.log	-0.308***
	(0.071)
polpc.log	0.480***
	(0.120)
density	0.096***
	(0.030)
taxpc.log.imp	-0.290*
	(0.163)
pctmin80	0.009***
	(0.003)
west	-0.148
	(0.100)
wcon.log	0.159
	(0.255)
wtuc.log	0.136
	(0.185)
wtrd.log	0.217
	(0.340)
wfir.log	-0.314
	(0.278)
wser.log.imp	-0.397
	(0.261)

```
wmfg.log                      0.206
                             (0.170)

wfed.log                     0.764**
                             (0.344)

wsta.log                     -0.423
                             (0.293)

wloc.log                      0.402
                             (0.500)

pctymle                       2.283
                             (1.452)

Constant                     -5.247
                             (3.345)

-----------------------------------------------
Observations                    89
R2                             0.793
Adjusted R2                    0.737
Residual Std. Error       0.268 (df = 69)
F Statistic           13.954*** (df = 19; 69)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

The coefficients of our model show us that (all statements ceteris peribus):

1. (Model 2: -0.378%) For a 1% increase in probability of arrest, there's a 0.435% decrease in crime rate
2. (Model 2: -0.131%) For a 1% increase in probability of imprisonment, there's a 0.190% decrease in crime rate
3. (Model 2: -1.5%) For a 1 day increase in the avgerage sentence, there's a 1.7% decrease in crime rate
4. (Model 2: -0.280%) For a 1% increase in probability of conviction, there's a 0.308% decrease in crime rate
5. (Model 2: +0.593%) For a 1% increase in police per capita, there's a 0.480% increase in crime rate
6. (Model 2: +10.2%) For a unit increase in density (100 people per square mile), there's an 9.6% increase in crime rate
7. (Model 2: -0.410%) For a 1% increase in taxpc, there's a 0.29% decrease in crime rate
8. (Model 2: 0.7%) For a 1 percentage point increase in percent minority, there's a 0.9% increase in crime rate
9. Counties in the west have a 14.8% lower *crmrte* than central or eastern counties
10. For a 1% increase in wcon, there's a 0.159% increase in crime rate

11. For a 1% increase in wtuc, there's a 0.136% increase in crime rate
12. For a 1% increase in wtrd, there's a 0.217% increase in crime rate
13. For a 1% increase in wfir, there's a 0.314% decrease in crime rate
14. For a 1% increase in wser, there's a 0.397% decrease in crime rate
15. For a 1% increase in wmfg, there's a 0.206% increase in crime rate
16. For a 1% increase in wfed, there's a 0.764% increase in crime rate
17. For a 1% increase in wsta, there's a 0.423% decrease in crime rate
18. For a 1% increase in wloc, there's a 0.402% increase in crime rate
19. For a 1 percentage point increase in percent young male, there's a 2.28% increase in crime rate

Compared with model 2, all changes in the coefficients in model 3 were modest, indicating that the reduction in bias we achieved by adding the variables we did was also modest. Likewise the story around the practical significance of our estimators didn't change noticeably between models 2 and 3.

Splitting out wages as we did in model 3 gives us more resolution into the impacts wages in differing industries have on *crmrte*, but given the practical insignificance of the estimators on these variables and their irrelevance to our research question, there isn't much here that begs for commentary.

The fitness of this model is good, but it's actually slightly worse than model 2, which contains 10 fewer variables, suggesting that model 2 is preferable to this one in usefulness, fitness, and parsimony.

```
In [29]: coeftest(model3, vcov = vcovHC, level = 0.05)

         cat('\n')
```

```
t test of coefficients:

                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -5.2471553  4.7769788 -1.0984 0.275837
crime.data$prbarr.log      -0.4346992  0.1288704 -3.3732 0.001222 **
crime.data$prbpris.log.imp -0.1902283  0.2041618 -0.9318 0.354713
crime.data$avgsen          -0.0165409  0.0152345 -1.0858 0.281366
crime.data$prbconv.log     -0.3083435  0.1270702 -2.4266 0.017859 *
crime.data$polpc.log        0.4803092  0.2426573  1.9794 0.051765 .
crime.data$density          0.0964037  0.0364682  2.6435 0.010148 *
crime.data$taxpc.log.imp   -0.2902644  0.2288864 -1.2682 0.209003
crime.data$pctmin80         0.0087743  0.0031667  2.7708 0.007180 **
crime.data$west            -0.1482138  0.1224098 -1.2108 0.230103
crime.data$wcon.log         0.1594351  0.2156411  0.7394 0.462200
crime.data$wtuc.log         0.1359078  0.2777515  0.4893 0.626171
crime.data$wtrd.log         0.2168253  0.3167925  0.6844 0.495990
crime.data$wfir.log        -0.3141947  0.3289754 -0.9551 0.342876
crime.data$wser.log.imp    -0.3974178  0.3031286 -1.3111 0.194186
crime.data$wmfg.log         0.2063289  0.1689996  1.2209 0.226286
crime.data$wfed.log         0.7636950  0.5258573  1.4523 0.150955
```

```
crime.data$wsta.log        -0.4230582  0.3608478 -1.1724 0.245068
crime.data$wloc.log         0.4016109  0.6281659  0.6393 0.524719
crime.data$pctymle          2.2828246  2.1665816  1.0537 0.295718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.5.8 Comparison of the three models

```
In [30]: stargazer(model1, model2, model3,
                   type = "text", omit.stat = "f",
                   star.cutoffs = c(0.05, 0.01, 0.001),
                   no.space = TRUE, align = TRUE)

         paste('Model 1 AIC: ', format(round(AIC(model1),3), nsmall = 3))
         paste('Model 2 AIC: ', format(round(AIC(model2),3), nsmall = 3))
         paste('Model 3 AIC: ', format(round(AIC(model3),3), nsmall = 3))

         cat('\n')
```

```
=================================================================
                              Dependent variable:
                    ---------------------------------------------
                                      crmrte.log
                          (1)             (2)             (3)
-----------------------------------------------------------------
prbarr.log             -0.490***       -0.373***       -0.435***
                        (0.141)         (0.092)         (0.103)
prbpris.log.imp         0.182                           -0.190
                        (0.277)                         (0.161)
avgsen                  0.014                           -0.017
                        (0.021)                         (0.013)
prbconv.log                             -0.279***       -0.308***
                                         (0.063)         (0.071)
density                                 0.102***        0.096**
                                         (0.027)         (0.030)
west                                    -0.653***       -0.148
                                         (0.143)         (0.100)
wcon.log                                                 0.159
                                                        (0.255)
wtuc.log                                                 0.136
                                                        (0.185)
wtrd.log                                                 0.217
```

```
                                                                (0.340)
wfir.log                                                        -0.314
                                                                (0.278)
wser.log.imp                                                    -0.397
                                                                (0.261)
wmfg.log                                                         0.206
                                                                (0.170)
wfed.log                                                        0.764*
                                                                (0.344)
wsta.log                                                        -0.423
                                                                (0.293)
wloc.log                                                         0.402
                                                                (0.500)
pctymle                                                          2.283
                                                                (1.452)
taxpc.log.imp                          -0.393**                 -0.290
                                        (0.140)                 (0.163)
pctmin80                                0.007**                 0.009**
                                        (0.002)                 (0.003)
polpc.log                              0.557***                 0.480***
                                        (0.107)                 (0.120)
wavg.log                                0.471
                                        (0.372)
west                                   0.410***
                                        (0.110)
Constant                  -4.142***     -2.202                   -5.247
                           (0.329)      (2.536)                  (3.345)
-------------------------------------------------------------------------
Observations                 89            89                       89
R2                          0.137         0.779                    0.793
Adjusted R2                 0.107         0.754                    0.737
Residual Std. Error 0.494 (df = 85) 0.259 (df = 79) 0.268 (df = 69)
=========================================================================
Note:                                      *p<0.05; **p<0.01; ***p<0.001
```

'Model 1 AIC: 132.902'
'Model 2 AIC: 23.597'
'Model 3 AIC: 37.648'

We can see that the coefficient on *prbarr* is relatively stable across models. *prbpris* and *avgsen* don't improve in significance in model 3 from model 1, further validating their exlusion from model 2.

prbconv, *density*, *pctmin80*, and *polpc* maintain their significance across models 2 and 3. Furthermore, none of their estimators change substantially (*polpc* changes most from 0.58 to 0.48) across model 2 and 3,

On the contrary, we can see that the population density has a significant effect to the crime rate. As a reminder, in model two, or a unit increase in density (100 people per square mile), there's an 11% increase in crime rate. Given the fact that it is very difficult to effectively control the population density by policy, we will see in the next section analysis how this variable contains a lot of bias of the omitted variables. Examining the model three metrics we can see that model two performs well relatively to the full model, especially if we include in our analysis the variables parsimony.

Finally, it's worth noting that our model 2 outperforms in fitness the full model 3, both in Adjusted R2 (75% vs 73%) and in Akaike's Information Criterion results (23.5 vs 37.6). The main reason for this is the combination of the model 2 parsimony and the combined indicator variable for the west and density variables.

1.6 CLM Assumptions

In this next section we shall assess the Classical Linear Model (CLM) assumptions. Model 2 contains a broad range of explanatory variables under policy maker control and strikes an optimal balance between goodness of fit and parsimony. It is considered the preferred model of choice and as such will be assessed more thoroughly against CLM assumptions to establish unbiasedness and to enable inferences to be made about the population (normality). In addition, interesting features from the other models shall be referenced so as to provide points of comparison.

1.6.1 1. Linear in Parameters

All three models are assumed to be linear in their parameters, enabling us to model the linear relationship between the selected explanatory variables and the log crime rate.

1.6.2 2. Random Sampling

In the Exploratory Data Analysis (EDA) section it was noted that the crime dataset contains observations for 90 counties (out of a possible 100 counties in North Carolina). For the purposes of this assignment, we are assuming that the data has been randomly sampled and that each observation is identical and independently distributed (i.i.d.). It is quite possible that some large, densely populated counties have been omitted from the dataset, but this point is mute providing the counties have been randomly sampled.

1.6.3 3. No Perfect Collinearity

The correlation matrix generated in the EDA section revealed no perfect multicollinearity. There's also no need to explicitly check for perfect collinearity as R will automatically notify us in the unlikely event this assumption is violated. An example of how this assumption could feasibly be violated was if all three the binary variables representing region ('west', 'central' or 'east') were included in the same model - they are mutually exclusive and so perfect multicollinearity would exist in this case.

While the presence of high (imperfect) multicollinearity in two or more explanatory variables does not violate CLM.3, it increases the standard error for each variable, resulting in an unstable coefficient and slope estimates. It is therefore worthwhile to assess the extent to which imperfect multicollinearity is present with the regression model. We notice that the variance inflation factor test characterizes the west binary (dummy) variable as the single variable with collinearity. This is not an issue because this variable is a binary variable used as an indicator in our model.

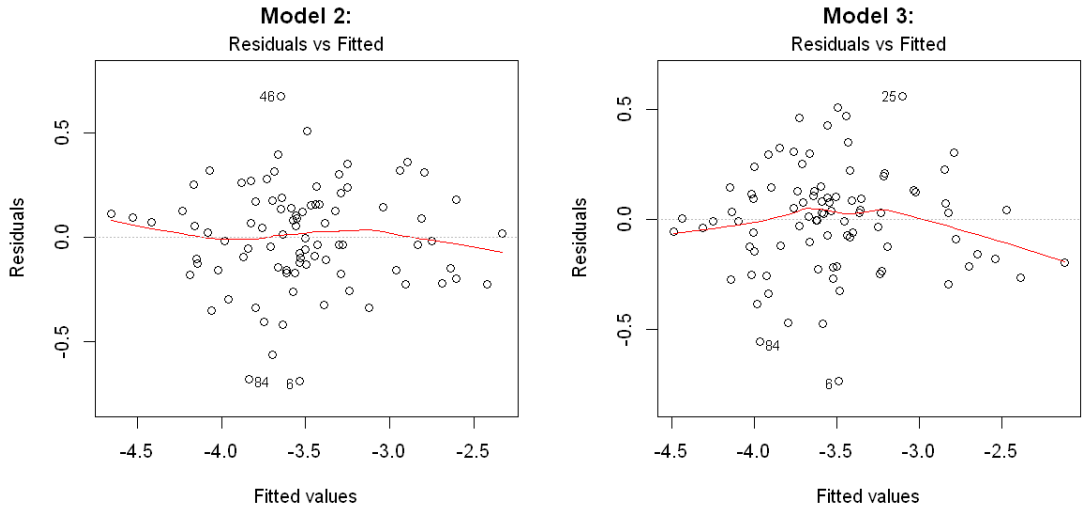
```
In [47]: # Variance inflation factor
         variance.inflation.factor.test<-vif(model2) > 4
         data.frame(variance.inflation.factor.test)

         cat('\n')
```

	variance.inflation.factor.test
crime.data$prbarr.log	FALSE
crime.data$prbconv.log	FALSE
crime.data$density	FALSE
crime.data$west	TRUE
crime.data$taxpc.log.imp	FALSE
crime.data$pctmin80	FALSE
crime.data$polpc.log	FALSE
crime.data$wavg.log	FALSE
crime.data$density:crime.data$west	FALSE

1.6.4 4. Zero Conditional Mean

The zero conditional mean assumption is the strongest we have encountered yet, and if met allows us to establish that the OLS estimators are unbiased, providing that CLM 1-3 have also be met. Plotting the residuals against the fitted values results in a flat line across zero, indicating that the zero conditional mean assumption has been met, i.e. $E(u|x_i) = 0$. Interestingly, model 3 appears to violate CLM.4, gauging by the moderate parabola. A further indication that the OLS estimators for model 2 are more optimal.

```
In [32]: plot.settings(0.5, 10, 1, 2)

         # Zero conditional mean
         plot(model2, which = 1, main='Model 2:')
         plot(model3, which = 1, main='Model 3:')
```



1.6.5 5. Homoskedasticity

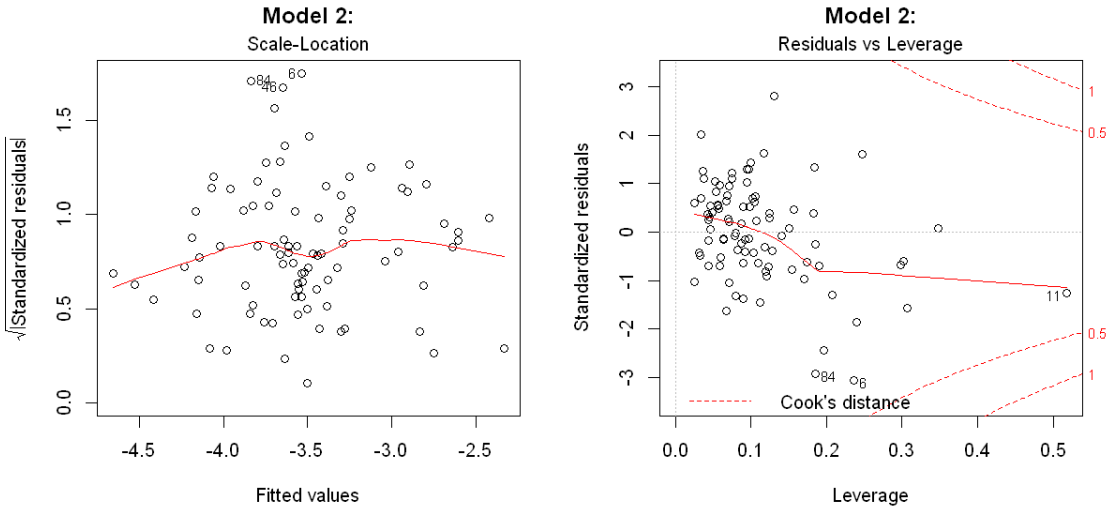
Another strong assumption is homoskedasticity, which if met, will allow us to establish that the model's errors exhibit constant variance for all values of x. The scale-location plots indicate that the assumption of homoskedasticity of error has been met by virtue of the flat line.

The Breusch-Pagan test however allows us to formally test for heteroskedasticity in our linear regression model. The null hypothesis is that there's homoskedasticity of error, and so a p-value < 0.05 means we can reject the null hypothesis and assume heteroskedasticity. A p-value of 0.1303 is greater than desired significance level, and so the null hypothesis cannot be rejected. The assumption of homoskedasticity has therefore been met allowing us to establish that the OLS estimators for model 2 are BLUE: Best Linear Unbiased Estimators.

The Cook's distance plot is also displayed to ascertain whether any data points are exerting excessive influence over OLS model 2. No data points have a Cook's distance greater than 0.5 indicating that this is a non-issue.

```
In [33]: plot.settings(0.5, 10, 1, 2)

         # Homoskedasticity
         plot(model2, which = 3, main = 'Model 2:')

         # Cook's Distance
         plot(model2, which = 5, main = 'Model 2:')

         # Breusch-Pagan test
         bptest(model2)
```

```
studentized Breusch-Pagan test

data:  model2
BP = 13.781, df = 9, p-value = 0.1303
```



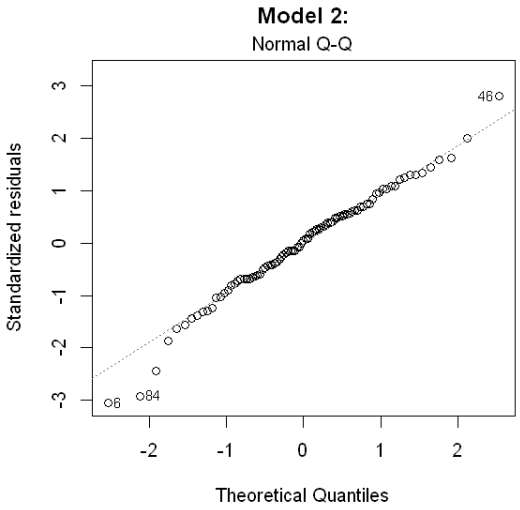
1.6.6 6. Normality

The final CLM assumption asserts that the distribution of error should be normally distributed. The following plots indicate that the residuals of model 2 are normally distributed. The Shapiro-Wilk test also allows us to formally test for normality. The null hypothesis is that population is normally distributed, and so a p-value < 0.05 means we can reject the null hypothesis. The resulting p-value is 0.5056 and so we cannot reject the null hypothesis, confirming that the population distribution of residuals for model 2 is normally distributed. The Shapiro-Wilk normality test was also conducted for models 1 and 3, both of which had p-values > 0.05, confirming that the population distribution of residuals is normal for all three models.

As model 2 meets all of the CLM assumptions, we have established it is BLUE and assert that the population distribution of OLS coefficients is normally distributed, and that it is appropriate to conduct inferential tests on their values.

```
In [34]: plot.settings(0.5, 10, 1, 2)

         # QQ plot
         plot(model2, which = 2, main = 'Model 2:')

         shapiro.test(model2$residuals)
```

```
Shapiro-Wilk normality test

data:  model2$residuals
W = 0.98674, p-value = 0.5056
```




```
In [35]: # Normality test - Models 1 and 3
         shapiro.test(model1$residuals)
         shapiro.test(model3$residuals)

         cat('\n')
```

```
Shapiro-Wilk normality test

data:  model1$residuals
W = 0.98246, p-value = 0.2733
```

```
Shapiro-Wilk normality test

data:  model3$residuals
W = 0.98731, p-value = 0.5442
```

1.7 Omitted variables

In this section we will examine variables that have been omitted from this research, but likely have a non-negligible effect on crime rate.

We know that the omitted variable bias on an OLS estimator β_1 is defined as:

$$Bias(\tilde{\beta_1}) = \hat{\beta_2}\tilde{\delta_1}$$

where $\tilde{\delta_1}$ is the slope from the simple regression of the omitted variable on the independent variables and $\hat{\beta_2}$ is the OLS slope estimator (if we could have it) from the multiple regression of the independent variables and the omitted variable.

Because we don't have any data of the omitted variables to do this analysis, we will roughly estimate the sign on the bias, based on our domain knowledge. Furthermore, depending on the OLS estimator, we'll say if the bias is moving the estimator away from zero (increasing its absolute value) or vice versa.

For reference, the independent variables and their estimated coefficients for model 2 are:

```
In [36]: Coefficients <- summary(model2)$coefficients[, "Estimate"]
         data.frame(Coefficients)
```

	Coefficients
(Intercept)	-2.20158599
crime.data$prbarr.log	-0.37287213
crime.data$prbconv.log	-0.27946862
crime.data$density	0.10157705
crime.data$west	-0.65295146
crime.data$taxpc.log.imp	-0.39327131
crime.data$pctmin80	0.00719337
crime.data$polpc.log	0.55704487
crime.data$wavg.log	0.47069500
crime.data$density:crime.data$west	0.41032528

1.7.1 Progressiveness of Judicial System Index

The legal system reflects common values that a society honors at any given time. Nevertheless, it lags behind the evolution of those values, which can act as a safety measure, but can prevent natural evolution of values. This impacts the definition of crime, when legislation delay resists the societal ethical principles it serves.

Indeed, in many historical cases, crime was dramatically reduced when the definition of what is illegal was revised; for example, alcohol and cannabis prohibition. The primary step of every policymaker trying to reduce the crime should always be to examine if the current laws are in line with the current society. In fact, this could be the first crime reduction method, to render some criminal laws obsolete and remove them. [6]

Another perspective of this is the Progressiveness of the punishment. Perhaps, the generalization of each crime and punishment pair is driven by the misconception of policy makers of what is actually important to each one of us. This is why education needs to be an integral part of the modern imprisonment punishment. In fact, modern scholars strongly believe in the holist impact of a conviction punishment and the actual rehabilitation impact of it. [2]

We estimate that the an index that indicates how progressive is the judicial system, will have a negative effect on crime rate $\hat{\beta_2}$<0.

Also, this omitted variable will have a negative correlation with the probability of arrest and conviction variables $\tilde{\delta_1} > 0$ resulting in negative bias. We cannot estimate a direction for the average sentence.

$$\hat{\beta_2} < 0$$

Dep. Var.	$\tilde{\delta_1}$	Bias sign	Bias effect
prbarr	> 0	negative	away from 0
prbconv	> 0	negative	away from 0

1.7.2 Financial Indices

Another major omitted variables category of this study is the one related with the poverty index and how this is addressed by the state. These omitted variables are:

1. The Unemployment Rate
2. The Poverty Index
3. The State Social Services Budget pre Capita

One could claim that people in many cases are forced to commit crimes for survival reasons. Unemployment and poverty are non-sustainable conditions that eventually will come to the surface if not properly handled by the state. Historically, the state approach against these groups has been the criminalization of them, driven by multiple politico-economic theories. It is clear in the modern days that this is wrong and it is not a solution but rather a repression of the problem. Indeed, modern political and legal acts like the "Homeless Bill of Rights" [5] or the "Poverty Task Force" [4] are a clear indication of the mentality shift against traditionally unprotected and sensitive lower layers of each social stratification. [3]

We estimate that the unemployment rate and the poverty index have a positive correlation with the crime rate $\hat{\beta_2}>0$. Also, these variables will have a positive correlation with the probability of arrest and tax per capita variables $\tilde{\delta_1} > 0$ resulting in positive bias and a negative correlation with the average wage $\tilde{\delta_1} < 0$ resulting in negative bias.

$$\hat{\beta_2} > 0$$

Dep. Var.	$\tilde{\delta_1}$	Bias sign	Bias effect
prbarr	> 0	positive	closer to 0
taxpc	> 0	positive	closer to 0
wavg	<0	negative	closer to 0

Regarding the State Social Services Budget per Capita variable, we estimate that $\hat{\beta_2}<0$ and $\tilde{\delta_1} > 0$ for tax per capita , resulting a negative bias for this independent variable.

$$\hat{\beta_2} < 0$$

Dep. Var.	$\tilde{\delta_1}$	Bias sign	Bias effect
taxpc	> 0	negative	away from 0

1.7.3 Education Index

Generally speaking, there is an inverse relationship between the education level and the possibility of committing a crime. These is a clear causal effect between education and the strengthening of peoples self-esteem. Self-esteem refers to a person's beliefs about their own worth and value. It also has to do with the feelings people experience that follow from their sense of worthiness. Self-esteem is important because it heavily influences people's choices and decisions. It has been shown that people with low self-importance and self-esteem feelings do not think that a possible punishment of lack of freedom, does not act as a significant deterrence of committing crimes.

We estimate that the education level will have a negative correlation with the crime rate $\hat{\beta}_2$<0. Also, we expect a positive correlation with the average wage variable.

$$\hat{\beta}_2 < 0$$

Dep. Var.	$\tilde{\delta}_1$	Bias sign	Bias effect
wavg	> 0	negative	closer to 0

1.7.4 Population Density and Crime Confounders

Income inequality In cities, money, services, wealth and opportunities are centralized. Many rural inhabitants come to the city to seek their fortune and alter their social position. Businesses, which provide jobs and exchange capital, are more concentrated in urban areas. Whether the source is trade or tourism, it is also through the ports or banking systems, commonly located in cities, that foreign money flows into a country. All these benefits come with a price though. There is a greater awareness of the income inequality between the rich and poor due to modern media.

We expect:

$$\hat{\beta}_2 > 0$$

Dep. Var.	$\tilde{\delta}_1$	Bias sign	Bias effect
prbarr	> 0	positive	closer to 0
prbconv	> 0	positive	closer to 0

Ghettoization Also, these areas have less social cohesion, and therefore less social control, something that acts as a social control factor of crime in suburban and rural areas. Finally, forced migration because of people rejection in more traditional and cohesive environment is another factor of increased crime.

We expect:

$$\hat{\beta}_2 > 0$$

Dep. Var.	$\tilde{\delta}_1$	Bias sign	Bias effect
prbarr	> 0	positive	closer to 0
prbconv	> 0	positive	closer to 0

Dep. Var.	$\tilde{\delta_1}$	Bias sign	Bias effect
density	> 0	positive	away from 0
pctmin80	>0	positive	away from 0
wavg	< 0	negative	closer to 0

It is obvious that the population density actually entails a large number of omitted variables, and the justification of any immediate causal effect between crime and population density would be wrong [6]

For wavg and pctmin80, the omitted variables move these estimators closer and away from zero respectively, we can tell that the effect of these variables will in practice become less and more important respectively.

For the rest variables, because we have opposite bias effects it's is difficult to estimate what would be the overall effect on the final model, especially given the fact that we have no data to estimate the size of the bias effects.

1.8 Conclusion and Recommendations

Controlling for counties' demographic and regional factors crucial for describing impact of influence-able factors A one-size-fits-all, state-wide policy in the domain of law enforcement or otherwise won't have the same impact on crime rate in every county. By including demographic, regional, and fiscal factors, we saw how bias in our base model changed. Notwithstanding policies directed toward increasing the probability of arrest, which was consistent across models, we'd recommend our party localizes or regionalizes policy decisions where possible.

Changing severity of punishment doesn't significantly impact crime rate Interestingly, our key second model could not show significance (even joint significance) of the effect of changes on probability of imprisonment or average sentence on crime rate. For that reason, we feel confident recommending that our party pursue policy to lower probability of imprisonment and reduce prison sentence length for non-violent crime. Opportunity to pursue alternatives to imprisonment, such as community service or mandatory surveillance, should reduce the burden on North Carolina citizens of expensive prison systems and not affect crime rates.

Focus on policy to increase the probability of arrest Our second model shows that increasing the probability of arrest, that is arrests per offense, should have a non-zero impact on crime rate. The practical significance of the impact is questionable, so to see a practically significant change in crime rate, there would need to be a substantial push to increase probability of arrest. We recommend investing in programs to improve police efficacy such that a greater proportion of offenders are arrested. For example, we may use IoT and machine learning to more accurately forecast where and when crime is most likely to occur. We may also consider how to signal a "tough-on-crime" stance such that public awareness of the likelihood of arrest is more deeply entrenched. Additionally, it's important that police don't arrest innocent offenders as probability of conviction can't significantly decrease if we hope to see the impact of additional arrests.

Develop creative policy to stymie the destructive interference of density and poverty Crime rate is most strongly associated with density and it also has a significant association with the percentage of the population in the racial or ethnic minority, often the underserved and systemically

oppressed groups in the state. We recommend pursuing a policy like affordable housing in school districts such that wealthier areas don't export their impoverished to dense neighborhoods where the poor are concentrated, which will drive up crime rate.

Modernize the laws In the omitted variables section we saw the impact of having an enhanced perspective can potentially have. What is clear from that analysis is that a more holistic approach is needed when it comes to understanding what is the causal driver of crimes. Out of all different aspects of this complex problem, we recognize the opportunity a crime reduction by revisioning the core definition of crime, the current laws. Having a modern and progressive judicial system sets the cornerstone of a more healthy society.

1.9 References:

1. : https://en.wikipedia.org/wiki/Legal_history_of_cannabis_in_the_United_States
2. : https://www.unodc.org/unodc/en/justice-and-prison-reform/prison-reform-and-alternatives-to-imprisonment.html
3. : https://en.wikipedia.org/wiki/Homeless_Bill_of_Rights
4. : https://www.cdss.ca.gov/inforesources/CDSS-Programs/Poverty-Task-Force
5. : https://journalistsresource.org/studies/government/criminal-justice/unemployment-property-crime-burglary/