Lab 3: Reducing Crime

Vinicio De Sola, Sam Tosaria, Joanna Yu (W203 Tuesday 4pm Fall 2018)

12/11/2018

Introduction

Public policy is focused on maximizing the well-being of the population and reducing the harmful effects of crime on the society while ensuring equitable treatment of all residents within the justice system. Crime poses sustantial economic and intangible costs to the society. Based on The Bureau of Justice statistics and FBI data, the U.S. crime rate reached a broad peak between the 1970s through early 1990s and has declined significantly after that. The current crime rates are about the same as those in the 1960s. This is partly the result of a series of policies targeted at reducing crime rate. A good understanding of factors associated with crime is important for effective policymaking and resource allocation.

In this study, we will research the crime data for a selection of counties in North Carolina from the year 1987 to understand the determinants of crime and generate policy suggestions for the local or state governments.

Exploratory Data Analysis

Data Description

The data set includes 25 variables that describe the various statistics of each county. The variables have been grouped by the following major categories:

Table 1: Data Description

No.	Category	Fields
1	Crime Rate	crmrte
2	Crime Punishment	prbarr, prbconv, prbpris, avgsen
3	Population	density, pctmin80, pctymle
4	Economic	taxpc
5	Geographic	county, west, central, urban
6	Income	wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc
7	Crime Type	mix
8	Law Enforcement	polpc
9	Time Period	year

Data Cleaning

In the 97 observations, 6 consists of missing values in all fields. These rows have been removed. The "probconv" variable should be numeric but is expressed as characters. That has been converted to numbers. Also, one data point was repeated in the dataset, so we remove it by using the unique function.

```
dfCrime <- read.csv('crime_v2.csv', stringsAsFactors = FALSE)
dfCrime <- na.omit(dfCrime)
dfCrime$prbconv <- as.numeric(dfCrime$prbconv)
```

```
dfCrime <- dfCrime[!duplicated(dfCrime),]
dimCrime <- data.frame(Dimensions = dim(dfCrime), row.names = c("Data Points","Variables"))
kable(dimCrime, booktabs=T, digits = 3,
        caption = 'Final number of data points and Variables') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 2: Final number of data points and Variables

	Dimensions
Data Points	90
Variables	25

Therefore, our final dataset contains 90 observations and 25 variables. In North Carolina, there is a total of 100 counties [1], thus our dataset covers 90% of all North Carolina. We must assume that the counties not included were selected at random, so our dataset is built from a random sample. If not, we know that there will be some biases in our results.

Let's finish this section by providing a summary of the data: we will tabulate the minimum, maximum, median, and mean of the numeric variables. However, not all numeric variables are equal: some are ratios, some are binary variables, and finally some are continuous variables. Nevertheless, let's tabulate them to produce a summary of the data.

```
v_mean <- lapply(dfCrime, mean, na.rm=TRUE)
v_mean <- round(as.numeric(v_mean), 4)
v_min <- lapply(dfCrime, min, na.rm=TRUE)
v_min <- round(as.numeric(v_min), 4)
v_max <- lapply(dfCrime, max, na.rm=TRUE)
v_max <- round(as.numeric(v_max), 4)
v_med <- lapply(dfCrime, median, na.rm=TRUE)
v_med <- round(as.numeric(v_med), 4)
v_tab <- cbind(Min=v_min, Mean=v_mean, Median=v_med , Max=v_max)
rownames(v_tab) <- colnames(dfCrime)
kable(v_tab, booktabs=T, digits = 4,
        caption = 'Summary dataset') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

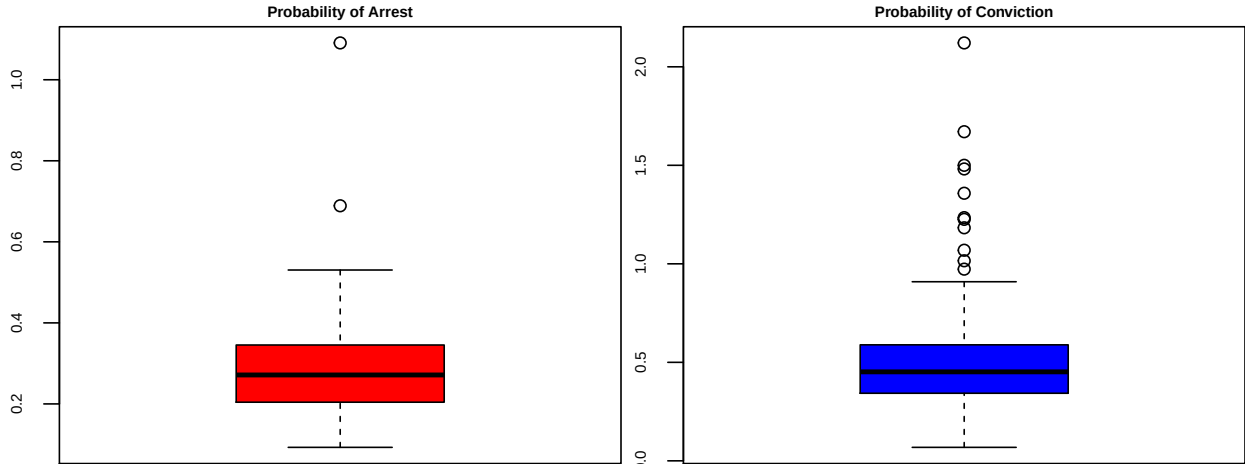
From this table we spot some strange values, especially the probabilities. The variable *prbarr* or probability of arrests and *prbconv* or probability of convictions have values well above 1, which is against the definition of probability. However, this may be caused by multiple arrests and multiple convictions for the same felony. Or this may be due to some data error or incorrect definition of the variable as a 'probability'. Additionally, there could be issues at registration of the values or some problems at how the proxy variable was used for defining a "probability". Let's take a look at some boxplots to try to make sense of it.

```
par(mfrow=c(1,2), mar=c(0,2,1,0))
boxplot(dfCrime$prbarr,
        col = "red", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Probability of Arrest")
boxplot(dfCrime$prbconv,
        col = 'blue', cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Probability of Conviction")
```

[1] https://www.ncpedia.org/geography/counties

Table 3: Summary dataset

	Min	Mean	Median	Max
county	1.0000	100.6000	103.0000	197.0000
year	87.0000	87.0000	87.0000	87.0000
crmrte	0.0055	0.0335	0.0300	0.0990
prbarr	0.0928	0.2952	0.2715	1.0909
prbconv	0.0684	0.5509	0.4517	2.1212
prbpris	0.1500	0.4106	0.4222	0.6000
avgsen	5.3800	9.6889	9.1100	20.7000
polpc	0.0007	0.0017	0.0015	0.0091
density	0.0000	1.4357	0.9792	8.8277
taxpc	25.6929	38.1610	34.9161	119.7615
west	0.0000	0.2444	0.0000	1.0000
central	0.0000	0.3778	0.0000	1.0000
urban	0.0000	0.0889	0.0000	1.0000
pctmin80	1.2837	25.7129	24.8516	64.3482
wcon	193.6432	285.3532	281.1624	436.7666
wtuc	187.6173	410.9065	404.7800	613.2261
wtrd	154.2090	210.9214	202.9879	354.6761
wfir	170.9402	321.6213	317.1257	509.4655
wser	133.0431	275.3379	253.1188	2177.0681
wmfg	157.4100	336.0327	321.0500	646.8500
wfed	326.1000	442.6189	448.8550	597.9500
wsta	258.3300	357.7402	358.4000	499.5900
wloc	239.1700	312.2801	307.6500	388.0900
mix	0.0196	0.1290	0.1009	0.4651
pctymle	0.0622	0.0840	0.0777	0.2487



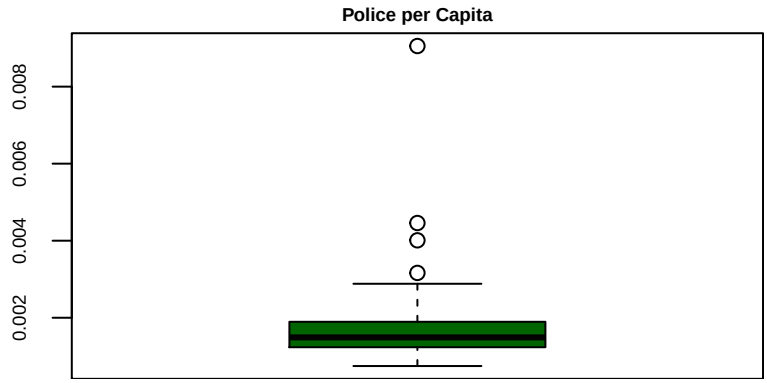
From the first boxplot, only one observation is higher than 1, and is quite a bit higher than the second one. This is county 115. If this point is not an error on record, this means that in this county, there were more arrests than crimes in a given year. Police may be arresting multiple suspects for a single crime. Or, this may be a mistake due to a data entry error.

On the other hand, for the probability of convictions, the issue doesn't seem to be trivial. There are a lot more convictions than arrests in several counties, which could mean that convictions from previous years are

being processed on this year, while arrests are below average. As reasoned earlier, this may also be due to multiple convictions for the same felony in a given year. Additionally, there may be a waiting time between being arrested, arreigned, and convicted. Many counties may have a significant number of people in jail or in bail waiting for trial, which means higher backlog in the judicial system. Therefore, none of these points are considered real outliers and will be included in our models, with the caveat that the term probability should be changed to ratio.

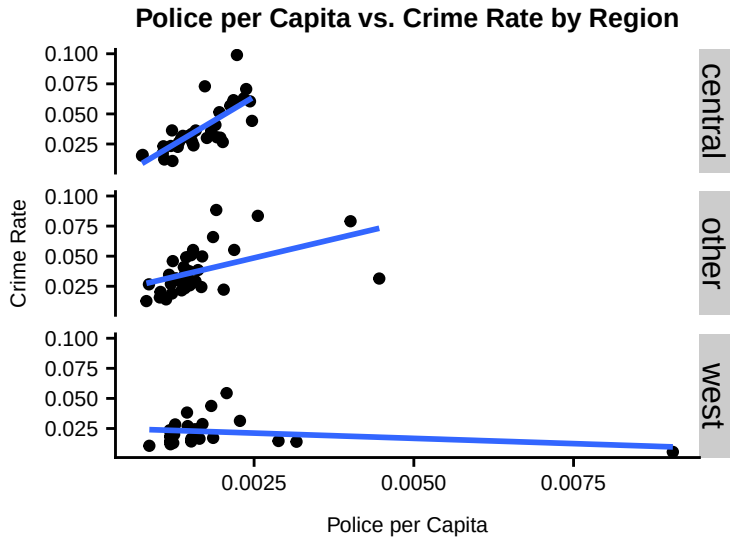
Now let's move our attention to *polpc* or Police per Capita. Let's also start by doing a boxplot to see the pattern of the possible outlier. Also, let's check this variable per region, to see if there is a pattern in the region that the possible outlier occurs when plotted against our dependent variable.

```
par(mfrow=c(1,1), mar=c(0,2,1,0))
boxplot(dfCrime$polpc,
        col = "darkgreen", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Police per Capita")
```




```
dfCrime$region <- ifelse(dfCrime$west == 1,"west",
                         ifelse(dfCrime$central == 1, "central", "other"))
```

```
ggplot(dfCrime, aes(polpc, crmrte))+
    geom_point() + facet_grid(region~.)+
    geom_smooth(method = 'lm', se = FALSE) + xlab("Police per Capita") +
    ylab("Crime Rate") + ggtitle("Police per Capita vs. Crime Rate by Region")+
    theme(plot.title = element_text(size = 10),
            axis.title = element_text(size = 8),
            axis.text = element_text(size = 8))
```



Once again, the max of this variable is alone at the top, and it comes from the same county as the outlier on probability of arrests, 115. To understand if this observation is a product of an error in the data process generation or a useful value, let's do some research on the amount of police per capita in the US. According to governing.com [2], a media that covers politics and government issues, for 2010, the highest police per capita county in the US was Washington D.C. with 0.00656, 31% less than the reported value in our dataset. From this research, and knowing that if we standarize this point, we get 7.41, we can make the decision that this point is an outlier that is not represenative of the rest of the population and can be removed.

Another possible explanation for this outlier may be due to the geographical or economic nature of that county. Since it is an urban county, it may have a higher concentration of commericial or non-residential infrastructure and thus a high police per capita figure.

To convince us that this county is not representative of the State, let's check the correlation between polpc and prbarr, the two variables where this data point have outliers. We will check the correlation between these two variable with and without this observation.

```
corout <- data.frame('Text' = c("Correlation With Outlier",
                                "Correlation Without Outlier"),
                     'Cor' = c(cor(dfCrime$polpc, dfCrime$prbarr),
                               cor(dfCrime[-51,]$polpc, dfCrime[-51,]$prbarr)))
colnames(corout)<- c("","Correlations")
kable(corout, booktabs=T,
        caption = 'County 115 Analysis of Correlations') %>%
        row_spec(0,bold=TRUE) %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 4: County 115 Analysis of Correlations

	Correlations
Correlation With Outlier	0.4259648
Correlation Without Outlier	-0.1261030

[2] http://www.governing.com/gov-data/safety-justice/law-enforcement-police-department-employee-totals-for-cities.html

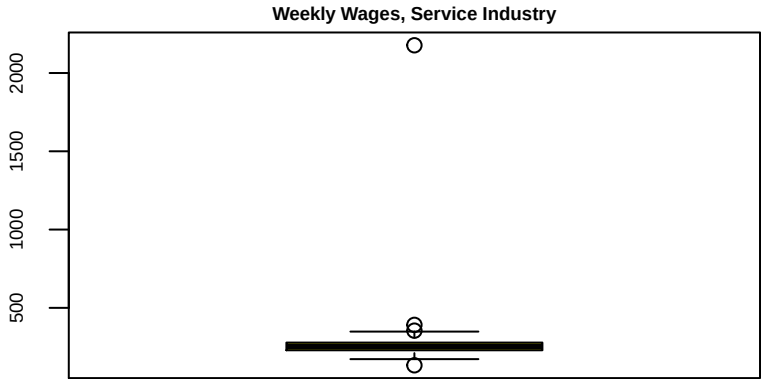
The sign even changes when we remove this point. To avoid removing a complete observation, because we don't have many on this dataset, we will replace the value of this data point with its mean. To have consistency, if the same county has problems with one outlier (polpc) and with other (prbarr), we will also substitute the value of pbrarr with its mean for that variable.

```
dfCrime$polpc <- ifelse(dfCrime$polpc == max(dfCrime$polpc),
                        mean(dfCrime$polpc[!dfCrime$polpc == max(dfCrime$polpc)]),
                        dfCrime$polpc)
```

```
dfCrime$prbarr <- ifelse(dfCrime$prbarr == max(dfCrime$prbarr),
                         mean(dfCrime$prbarr[!dfCrime$prbarr==max(dfCrime$prbarr)]),
                         dfCrime$prbarr)
```

Now let's move to wages. From the summary table, the only wage's variable that seem to have a clear outlier is *wser* or weekly wages in the service industry. Let's take a look at the boxplot.

```
par(mfrow=c(1,1), mar=c(0,2,1,0))
boxplot(dfCrime$wser,
        col = "yellow", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Weekly Wages, Service Industry")
```



The data on wages per week for numerous industries without the weights of those sectors does not help in assessing the economic situation of the county. We only have one value that is extremely higher than the rest. If we standarize the value, we get 9.17 standard deviations from the mean.

From the original paper [3] where the data comes from, we know that the wage data were provided by the North Carolina Employment Security Commission. Looking at this value and other wage ranges, we believe this value was an error and is not representative of the rest of the counties or the state. We will substitute its value with a mean of the sample.

```
dfCrime$wser <- ifelse(dfCrime$wser == max(dfCrime$wser),
                       mean(dfCrime$wser[!dfCrime$wser==max(dfCrime$wser)]),
                       dfCrime$wser)
```

As a final note, density seems to be in a different type of units that the one presented in the description. The average density in North Carolina in 1987 was [4] 131.63 people per square mile, according to our calculation.

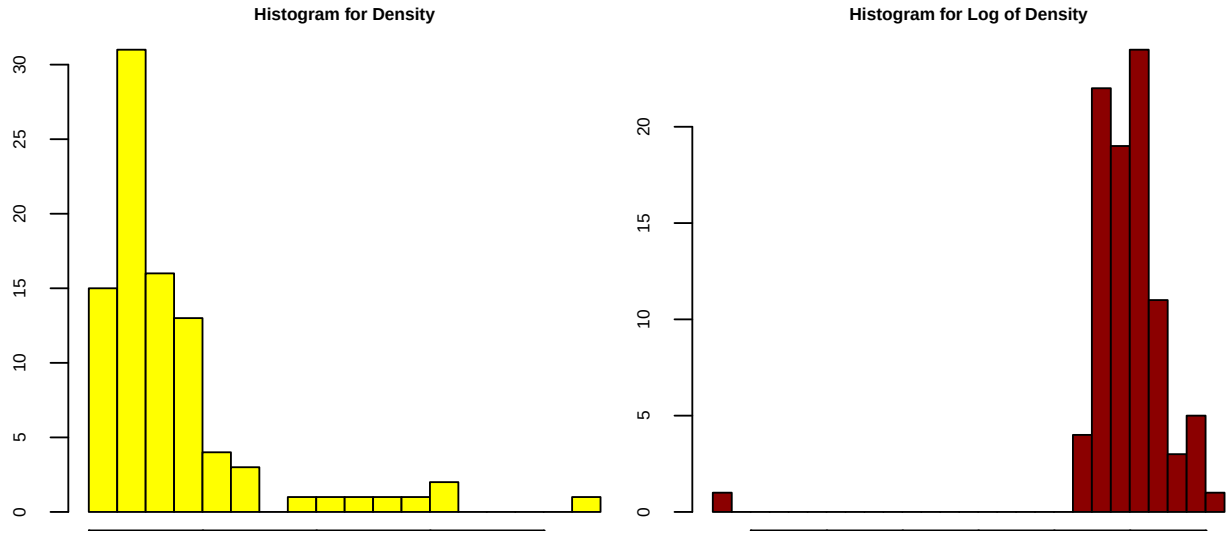
[3] https://www.amherst.edu/media/view/121570/original/CornwellTrumbullCrime%2BElasticities.pdf

[4] https://www.statista.com/statistics/206270/resident-population-in-north-carolina/, https://www.indexmundi.com/facts/united-states/quick-facts/north-carolina/land-area#map

This is far from the range that the variable density is presented to us. Using a quick transformation, we can see that what we have is hundreds of people per square mile. To avoid a cumbersome interpretation of regression coefficients, we will multiply the variable by 100.

Also, let's plot an histogram of density and of its logarithm. We can do this transformation because it's highly unlikely that a county will have a density of 0. Because density is a ratio, it's bounded by 0, which means that it will likely be positively skewed. So the log transformation makes sense.

```
# dfCrime$density <- 100*dfCrime$density
par(mfrow=c(1,2), mar=c(0,2,1,0))
hist(dfCrime$density, breaks = 20,
        col = "yellow", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Histogram for Density", xlab = 'Density')
hist(log(dfCrime$density), breaks = 20,
        col = 'darkred', cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Histogram for Log of Density", xlab = "Log of Density")
```



While it's clear that we should use a logarithmic transformation for this variable, in both cases we have possible outliers. In the normal plane, there is a county with a density of 8, which means a urban city or a highly populated metro area. For policy purposes, we need to live with this observation. However, on the transformation, we spot a point that is quite small. From what we gather, it correspond to a density of 0.00002 people per square mile, which using our estimate for the are of North Carolina, it's about 1 * 100 persons living in the county. This may be an error, and if is not, from a policy perspective the remedies for this remote area which is quite different from the rest of the state may need a targeted policy separate from that of the rest of the state. Additionally, public policy should focus on maximizing the benefits for the largest number of people possible, thus we will ignore this point for the purpose of the analysis and substitute it by the mean.

```
dfCrime$density <- ifelse(dfCrime$density == min(dfCrime$density),
                          mean(dfCrime$density[!dfCrime$density
                                               ==min(dfCrime$density)]),
                          dfCrime$density)
```

Let's now tabulate the summary of the variables that we change to present the removal of the outliers.

```
vars <- c("prbarr","polpc","wser","density")
v_mean <- lapply(dfCrime[vars], mean, na.rm=TRUE)
v_mean <- round(as.numeric(v_mean), 4)
v_min <- lapply(dfCrime[vars], min, na.rm=TRUE)
v_min <- round(as.numeric(v_min), 4)
v_max <- lapply(dfCrime[vars], max, na.rm=TRUE)
v_max <- round(as.numeric(v_max), 4)
v_med <- lapply(dfCrime[vars], median, na.rm=TRUE)
v_med <- round(as.numeric(v_med), 4)
v_tab <- cbind(Min=v_min, Mean=v_mean, Median=v_med , Max=v_max)
rownames(v_tab) <- vars
kable(v_tab, booktabs=T, digits = 4,
        caption = 'Summary dataset of changed Variables') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 5: Summary dataset of changed Variables

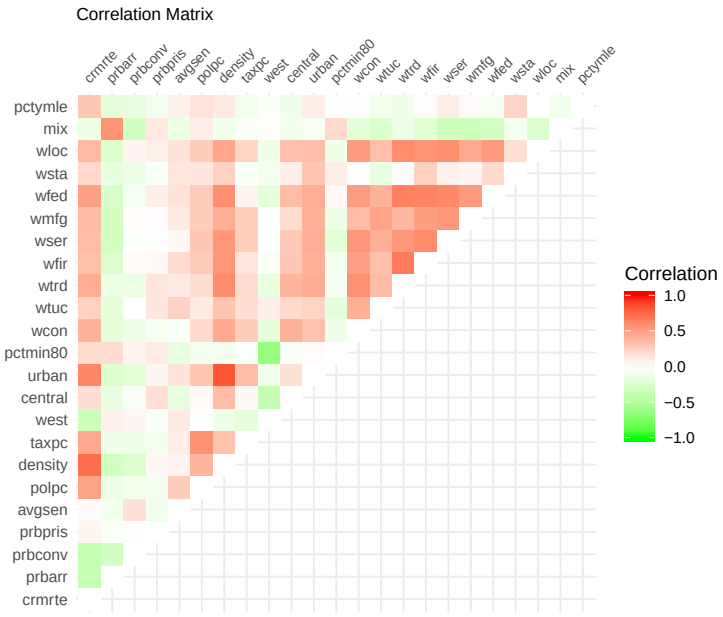
	Min	Mean	Median	Max
prbarr	0.0928	0.2863	0.2715	0.6890
polpc	0.0007	0.0016	0.0015	0.0045
wser	133.0431	253.9701	253.1188	391.3081
density	0.3006	1.4518	1.0008	8.8277

Correlations

For a final EDA, let's build a heatmap of the correlations between the all the variables to check effects between them, and to see if there are some negative effects like multicolinearity.

```
cormat <- round(cor(dfCrime[,colnames(dfCrime)[3:25]]),3)
cormat[lower.tri(cormat)] <- NA
melted_cormat <- melt(cormat, na.rm = TRUE)
melted_cormat[melted_cormat$value==1,]$value <- 0

ggplot(data=melted_cormat, aes(Var1,Var2,fill=value))+
    geom_tile()+
    scale_fill_gradient2(low = "green", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name = "Correlation")+
    theme_minimal()+
    scale_x_discrete(position = "top")+
    theme(axis.text.x = element_text(angle=45, vjust=1, size=8,hjust=0),
          axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          plot.title = element_text(size=10))+
    coord_fixed()+
    ggtitle("Correlation Matrix")
```



From this heatmap we can have the next observations for our models

- The dependent variable *crmrte* (crime rate) has positive correlation with almost all the wages, density, percent of minorities and young males. It has a negative correlation with probability of conviction and arrests, and being on the west location. The variables mix, average sentence, and probability of prison sentence have low correlations.
- Wages are all highly correlated with each other. This will become a problem in the multicolinear regression. We will have some issues with multicolinearity.

Summary of variables

Before doing the Model Building Process, let's do a summary of the variables in the dataset, how they're related to the dependent variable, crime rate, and quickness of changes in policy that could be implemented. For this last part, we divide the data in two: long term effects and short term effects.

```
labels <- c("crimes committed per person", "probability of arrest",
  "probability of conviction", "probability of prison sentence",
  "avg. sentence, days", "police per capita", "100s of people per sq. mile",
  "tax revenue per capita", "=1 if in western N.C.", "=1 if in central N.C.",
  "=1 if in SMSA", "perc. minority, 1980", "weekly wage, construction",
  "wkly wge, trns, util, commun", "wkly wge, whlesle, retail trade",
  "wkly wge, fin, ins, real est", "wkly wge, service industry",
  "wkly wge, manufacturing", "wkly wge, fed employees",
  "wkly wge, state employees", "wkly wge, local gov emps",
  "offense mix: face-to-face/other", "percent young male")
control <- c("NA", "Long Term", "Long Term", "Short Term", "Short Term",
  "Long Term", "Long Term", "Long Term","No", "No", "No", "Long Term",
  "Long Term", "Long Term", "Long Term", "Long Term", "Long Term", "Long Term",
  "Long Term", "Short Term", "Long Term", "No", "Long Term")
cormat1 <- round(cor(dfCrime[,colnames(dfCrime)[3:25]])[1,],3)
desc <- data.frame(colnames(dfCrime)[3:25], labels, cormat1, control,
                   row.names = NULL)
colnames(desc) <- c("Explanatory Variables", "Explanation",
```

```
                    "Correlation w/ Crime Rate","Potential Policy Impact Timeframe")
kable(desc, booktabs = TRUE, align = c("llcc"),
      caption = 'Overview of the variables and their impact') %>%
  kable_styling(latex_options = c("scale_down",'hold_position'),
                full_width = FALSE, position = "center") %>%
  row_spec(0, bold = TRUE) %>%
  column_spec(1, width = "7em") %>%
  column_spec(2, width = "12em") %>%
  column_spec(3, width = "8em") %>%
  column_spec(4, width = "10em")
```

Table 6: Overview of the variables and their impact

Explanatory Variables	**Explanation**	**Correlation w/ Crime Rate**	**Potential Policy Impact Timeframe**
crmrte	crimes committed per person	1.000	NA
prbarr	probability of arrest	-0.378	Long Term
prbconv	probability of conviction	-0.386	Long Term
prbpris	probability of prison sentence	0.048	Short Term
avgsen	avg. sentence, days	0.020	Short Term
polpc	police per capita	0.477	Long Term
density	100s of people per sq. mile	0.721	Long Term
taxpc	tax revenue per capita	0.449	Long Term
west	=1 if in western N.C.	-0.346	No
central	=1 if in central N.C.	0.166	No
urban	=1 if in SMSA	0.615	No
pctmin80	perc. minority, 1980	0.182	Long Term
wcon	weekly wage, construction	0.393	Long Term
wtuc	wkly wge, trns, util, commun	0.236	Long Term
wtrd	wkly wge, whlesle, retail trade	0.427	Long Term
wfir	wkly wge, fin, ins, real est	0.336	Long Term
wser	wkly wge, service industry	0.345	Long Term
wmfg	wkly wge, manufacturing	0.353	Long Term
wfed	wkly wge, fed employees	0.490	Long Term
wsta	wkly wge, state employees	0.200	Short Term
wloc	wkly wge, local gov emps	0.360	Long Term
mix	offense mix: face-to-face/other	-0.132	No
pctymle	percent young male	0.290	Long Term

The Model Building Process

Overview

Now that we have cleaned our dataset, let's build our models with our objectives in mind: build policies designed to impact the crime rate in the State that the party could campaign on, while being possible to implement. We will build a model based on the main drivers: the probability of being arrested, convicted, and sentenced to prison. We will create a variable called from Arrest to Prison that will assume that each process is independent from the other (as the judicial system should be).

Our first model will use this variable, and we will use two more explanatory variables to decrease the bias and add more explanatory power to our model: density and percent of young male, two strongly correlated variables.

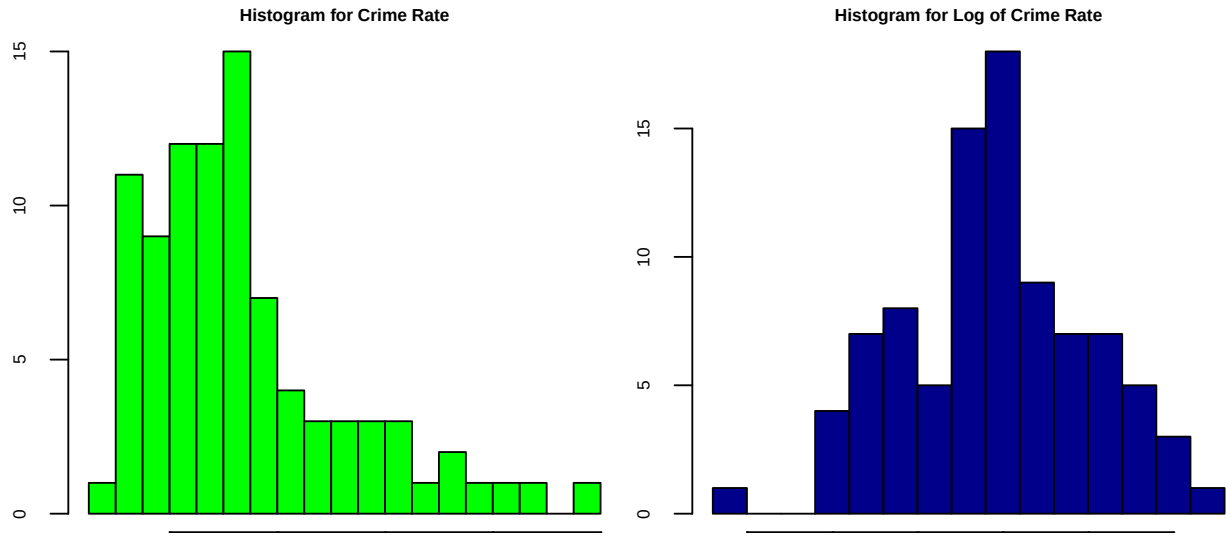
Our second model will build from the first, including more geographical variables like region, urban areas, minority and police per capita.

Finally, we will include a model with all the variables. From this we will decide if all variables are statistically significant or not, and if our previous model is robust.

Outcome Variable: *crmrte*

But first, let's do an EDA on the dependent variable of our study: crime rate. Because crime rate is a ratio, it's bounded by 0, which will make a skewed distribution. Also, because the crime rate may be related to the probabilities in a multiplicative model, a logarithmic transformation makes sense to avoid the skewness. Let's plot both graphs to see this effect.

```
par(mfrow=c(1,2), mar=c(0,2,1,0))
hist(dfCrime$crmrte, breaks = 20,
        col = "green", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Histogram for Crime Rate", xlab = 'Crime Rate')
hist(log(dfCrime$crmrte), breaks = 20,
        col = 'darkblue', cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Histogram for Log of Crime Rate", xlab = "Log of Crime Rate")
```



From the graphs, it's pretty clear that a *log* transformation is quite useful for our OLS regression, having a close to normal variable. The only issue with this transformation is when a county has a crime rate of 0,

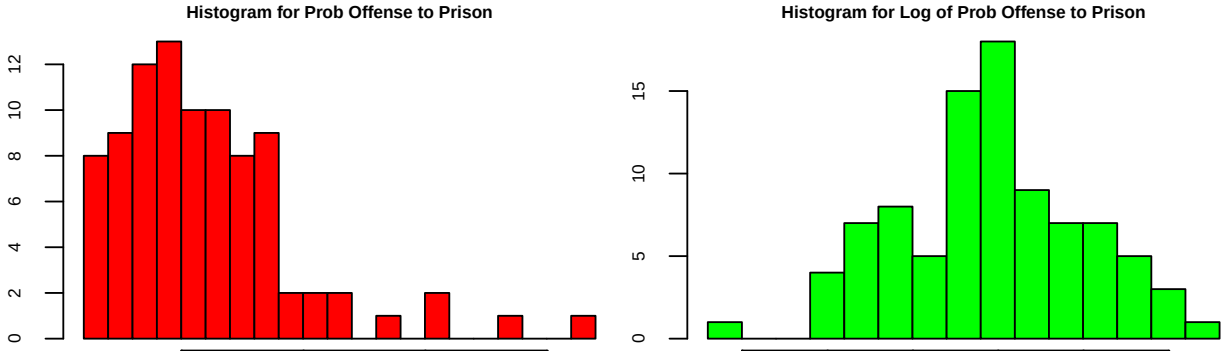
which is theoretically possible. Also, other positive inclusion of doing this transformation is that now we can interpret the coefficients as elasticities: for any unit change in our covariates, the change in the dependent variable can be interpreted as a per $100\beta_i\%$ change. This will help explain the practical significance of our coefficients.

Model #1

In this section, we will investigate the effect of punishment on crime rate. Specifically, punishment is broken into two major categories:

1) Certainty of punishment, proxied by various ratio 1) arrest per offense, 2) conviction per arrest, and 3) prison per conviction. We created a new variable that combines all of them (by multiplying) the 3 certainty of punishment ratios to obtain the ratio for prison per offense to get a better sense of the extent in which criminals expect to be in prisoned after committing a crime. We expect the probability of prison to be strong deterrants for crime. Because this is a ratio, we will use a log transformation for this new variable. This variable will be called *prbPrisonToCrime*

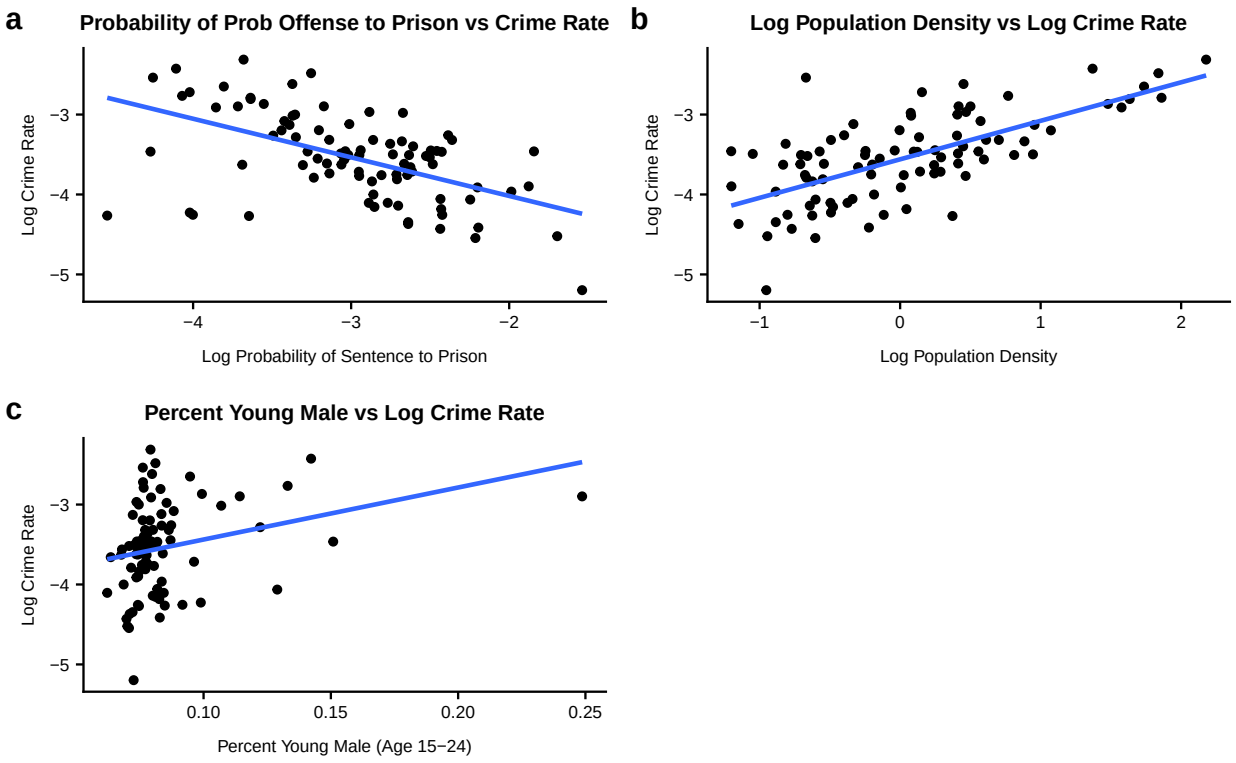
```
par(mfrow=c(1,2), mar=c(0,2,1,0))
dfCrime$prbPrisonToCrime <-  dfCrime$prbarr * dfCrime$prbconv * dfCrime$prbpris
hist(dfCrime$prbPrisonToCrime, breaks = 20,
        col = "red", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Histogram for Prob Offense to Prison",
        xlab = 'Prob Offense to Prison')
hist(log(dfCrime$crmrte), breaks = 20,
        col = 'green', cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Histogram for Log of Prob Offense to Prison",
        xlab = "Log of Prob Offense to Prison")
```



2) In addition, based on the heatmap and our understanding of crime, we expect variables "density" and "percent of young male" to also show strong correlation with crime rate. As confirmed by the following scatterplots, the correlations seem strong.

```
# 1 Scatterplot
sc1 <- ggplot(dfCrime, aes(log(prbPrisonToCrime), log(crmrte)))+
    geom_point()+
    geom_smooth(method = 'lm', se = FALSE) +
    xlab("Log Probability of Sentence to Prison") +
    ylab("Log Crime Rate") +
    ggtitle("Probability of Prob Offense to Prison vs Crime Rate")+
```

```
    theme(plot.title = element_text(size = 10),
            axis.title = element_text(size = 8),
            axis.text = element_text(size = 8))
# 2 Scatterplot
sc2 <- ggplot(dfCrime, aes(log(density), log(crmrte)))+
    geom_point()+
    geom_smooth(method = 'lm', se = FALSE) +
    xlab("Log Population Density") +
    ylab("Log Crime Rate") +
    ggtitle("Log Population Density vs Log Crime Rate")+
    theme(plot.title = element_text(size = 10),
            axis.title = element_text(size = 8),
            axis.text = element_text(size = 8))
# 3 Scatterplot
sc3 <- ggplot(dfCrime, aes(pctymle, log(crmrte)))+
    geom_point()+
    geom_smooth(method = 'lm', se = FALSE) +
    xlab("Percent Young Male (Age 15-24)") +
    ylab("Log Crime Rate") +
    ggtitle("Percent Young Male vs Log Crime Rate")+
    theme(plot.title = element_text(size = 10),
            axis.title = element_text(size = 8),
            axis.text = element_text(size = 8))
plot_grid(sc1, sc2, sc3, labels = "auto")
```



Now we can create our first model. For checking that our model is balancing number of variables and information given, we calculate the AIC of the overall model, and the AIC of each individual OLS. We want to select the model with the lower AIC.

```
# Overall Model
model_1 <- lm(log(crmrte) ~ log(prbPrisonToCrime)
              +log(density)+pctymle, data = dfCrime)
model_1$AIC <-  AIC(model_1)
# OLS of Prob
model_prob <- lm(log(crmrte) ~ log(prbPrisonToCrime), data = dfCrime)
model_prob$AIC <- AIC(model_prob)
# OLS of density
model_density <- lm(log(crmrte) ~ log(density), data = dfCrime)
model_density$AIC <- AIC(model_density)
# OLS of Male
model_ymale <- lm(log(crmrte) ~ pctymle, data = dfCrime)
model_ymale$AIC <- AIC(model_ymale)
# Combined OLS Prison and Density
model_prob_density <- lm(log(crmrte) ~ log(prbPrisonToCrime)
                         + log(density), data = dfCrime)
model_prob_density$AIC <- AIC(model_prob_density)
# Combined OLS Prison and Male
model_prob_ymale <- lm(log(crmrte) ~ log(prbPrisonToCrime)
                         + pctymle, data = dfCrime)
model_prob_ymale$AIC <- AIC(model_prob_ymale)
# Combined OLS Density and Male
model_density_ymale <- lm(log(crmrte) ~ log(density)
                         + pctymle, data = dfCrime)
model_density_ymale$AIC <- AIC(model_density_ymale)
# Table of results
stargazer(model_1, model_prob, model_density, model_ymale,
          model_prob_density, model_prob_ymale, model_density_ymale,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "adj.rsq", "n", "aic"), header = FALSE)
```

The regression table shows the results of 7 models: 3 involving single variables (#1.2-1.4), 3 involving joint variables (#1.5-1.7), and 1 with all 3 variables of interest (#1.1). The results for model #1.1 (all 3 variables) and #1.5 (probability of prison and density) are very similar. The resulting adjusted R^2 and AIC show a good balance between fit and parsimony. However, we should drop the "young male" variable since it is not statistically significant and it has no effect on the adjusted R^2 while increasing AIC slightly. Therefore, our initial model is #1.5 and will have the transformed probability and density as its main variables. The adjusted R^2 is 0.514, which means we are explaining near 51% of the dependent variance with our model.

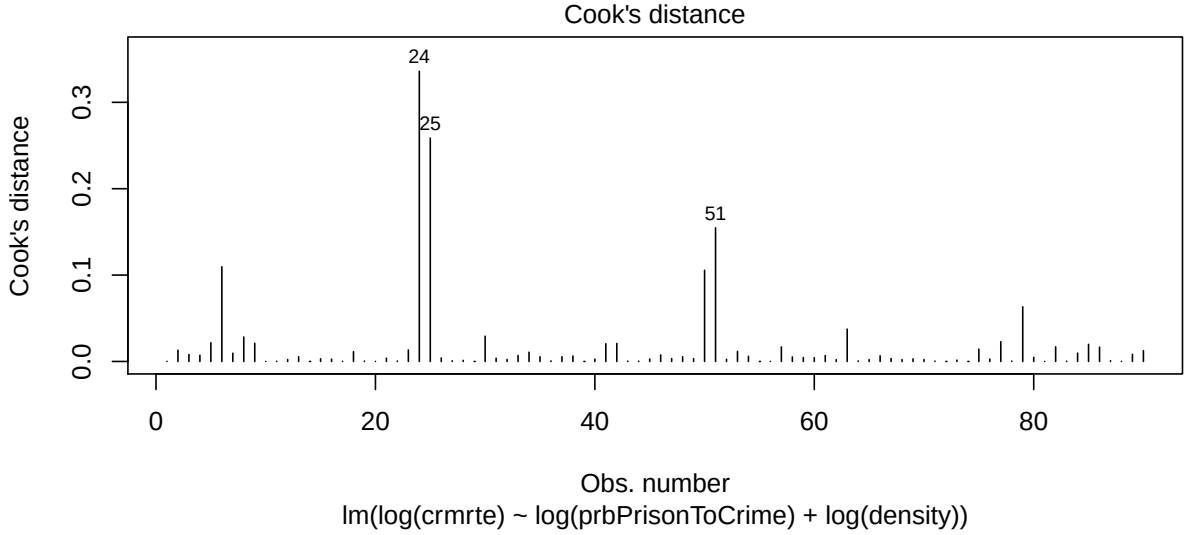
Now let's do a quick analysis of the model's outliers.

```
plot(model_prob_density, which = 4)
```

Table 7: Linear Models Predicting Crime Rate

	Dependent variable:						
	log(crmrte)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
log(prbPrisonToCrime)	-0.255^{***}	-0.483^{***}			-0.276^{***}	-0.450^{***}	
	(0.077)	(0.082)			(0.074)	(0.087)	
log(density)	0.386^{***}		0.482^{***}		0.389^{***}		0.460^{***}
	(0.059)		(0.057)		(0.059)		(0.057)
pctymle	1.701			6.509^{***}		2.421	3.536^{*}
	(1.851)			(2.396)		(2.253)	(1.862)
Constant	-4.460^{***}	-4.984^{***}	-3.560^{***}	-4.089^{***}	-4.381^{***}	-5.089^{***}	-3.856^{***}
	(0.239)	(0.249)	(0.043)	(0.209)	(0.223)	(0.267)	(0.162)
Observations	90	90	90	90	90	90	90
R^2	0.530	0.284	0.449	0.077	0.525	0.294	0.471
Adjusted R^2	0.514	0.276	0.443	0.067	0.514	0.277	0.459
Akaike Inf. Crit.	88.430	122.297	98.729	145.134	87.309	123.110	97.073

Note: *p<0.1; **p<0.05; ***p<0.01



There is only a couple of big influence points in observation (Observation 24, 25, and 51) that we need to keep an eye.

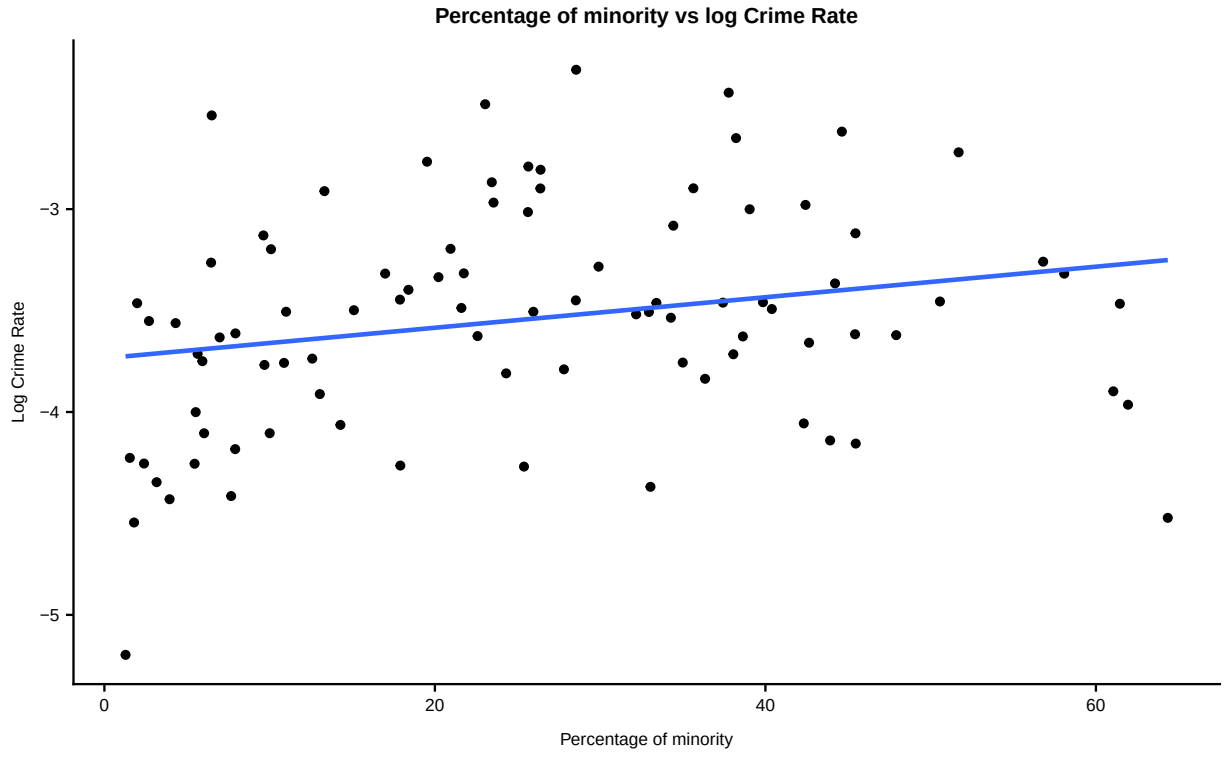
Model #2

Now we will reconstruct a comprehensive model while being mindful of the correlations between the independent variables, and the omitted variable biases. As explained before, Model 1 will not include the variable *pctymle*. However, as we move to more inclusive models, we will continue to include it because of the high correlation that this variable has with the dependent variable.

The variables that we are going to include are

- *pctmin80*: Minority percentage. This will include some socioeconomics variables in our study (We present the scatterplot below)
- *central, west*: Regions dummy variables. We don't need to encode the *Other* variable to avoid having three variables that are perfectly correlated. We don't include *urban* because the low number of data points that we have. For this model, we won't include any interactions between the models. The main reason is that for policy, we don't expect changes in the slope of crime rare if we change the region. What we expect is a fundamental change or jump in the crime rate if geography is significant in the model.

```
# 1 Scatterplot
sc1 <- ggplot(dfCrime, aes(pctmin80, log(crmrte)))+
    geom_point()+
    geom_smooth(method = 'lm', se = FALSE) +
    xlab("Percentage of minority") +
    ylab("Log Crime Rate") +
    ggtitle("Percentage of minority vs log Crime Rate")+
    theme(plot.title = element_text(size = 10),
            axis.title = element_text(size = 8),
            axis.text = element_text(size = 8))
sc1
```



We'll use a similar process as before, using several models to try to strike a good balance between fit and parsimony.

```
# Model 2
model_2 <- lm(log(crmrte) ~ log(prbPrisonToCrime) + pctmin80 +
                log(density) + central + west, data = dfCrime)
model_2$AIC <- AIC(model_2)
# OLS of Minority
model_mino <- lm(log(crmrte) ~ pctmin80, data = dfCrime)
model_mino$AIC <- AIC(model_mino)
# OLS of Region
model_region <- lm(log(crmrte) ~ central + west, data = dfCrime)
model_region$AIC <- AIC(model_region)
# Combined OLS Region and Minority
model_mino_region <- lm(log(crmrte) ~ central + west
                            + pctmin80, data = dfCrime)
model_mino_region$AIC <- AIC(model_mino_region)
# Combined OLS Model 1 and Minority
model_1_mino <- lm(log(crmrte) ~ log(prbPrisonToCrime)+ log(density)+
                            pctmin80, data = dfCrime)
model_1_mino$AIC <- AIC(model_1_mino)
# Combined OLS Model 1 and Region
model_1_region <- lm(log(crmrte) ~ log(prbPrisonToCrime)+ pctymle + log(density)+
                            + central+west, data = dfCrime)
model_1_region$AIC <- AIC(model_1_region)
# Table of results
stargazer(model_2, model_mino, model_region, model_mino_region,
          model_1_mino, model_1_region ,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "adj.rsq", "n", "aic"), header = FALSE,
          table.placement = "H")
```

Table 8: Linear Models Predicting Crime Rate

	Dependent variable:					
	log(crmrte)					
	(1)	(2)	(3)	(4)	(5)	(6)
log(prbPrisonToCrime)	−0.308***				−0.348***	−0.256***
	(0.060)				(0.060)	(0.064)
pctmin80	0.008***	0.008**		−0.001	0.014***	
	(0.003)	(0.003)		(0.005)	(0.002)	
pctymle						0.454
						(1.540)
log(density)	0.434***				0.419***	0.426***
	(0.050)				(0.047)	(0.053)
central	−0.168**		0.031	0.013		−0.256***
	(0.084)		(0.119)	(0.132)		(0.082)
west	−0.322***		−0.513***	−0.555***		−0.562***
	(0.116)		(0.134)	(0.193)		(0.084)
Constant	−4.541***	−3.735***	−3.428***	−3.375***	−4.949***	−4.128***
	(0.240)	(0.103)	(0.084)	(0.191)	(0.198)	(0.211)
Observations	90	90	90	90	90	90
R^2	0.721	0.054	0.172	0.173	0.694	0.694
Adjusted R^2	0.705	0.044	0.153	0.144	0.684	0.676
Akaike Inf. Crit.	45.362	147.361	137.355	139.254	49.695	53.681
Note:						*p<0.1; **p<0.05; ***p<0.01

The regression table shows the results of 6 models. Models #2.2-2.3 explore regressions using a single category variable - minority and geographic variables "central" and "west". Model #2.4 includes both categories. Model #2.5 and #2.6 explore the scenario of adding each of the two new category variables to our base model in the prior section. Finally, model #2.1 includes both new category variables and the two variables from our base model.

Model #2.1 is clearly more robust with a lowest AIC (45.362) and the highest adjusted R^2, 0.705, which means that in this model we are able to explain 70.5% of the dependent variable variance.

When we recalculated the probability of prison sentence through the regression anatomy approach, we see there is a slight bias towards zero for the probability of conviction for a given offensevariable (prbPrisonToCrime) due to the covariance between some of the independent variables such as percentage of minorities, and density.

```
cov_prbPrisonToCrime_xi <- lm(log(prbPrisonToCrime) ~ pctmin80 +
                          log(density) , data = dfCrime)
true_prbPrisonToCrime <- lm(log(crmrte) ~ resid(cov_prbPrisonToCrime_xi),
                          data = dfCrime)
coef <- data.frame("Multi Step Process Coefficients Prob"
                   =coef(true_prbPrisonToCrime))
```

```
kable(coef, booktabs=T, digits = 3,
        caption = 'Coefficients of OLS using multistep process for Prob') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 9: Coefficients of OLS using multistep process for Prob

	Multi.Step.Process.Coefficients.Prob
(Intercept)	-3.542
resid(cov_prbPrisonToCrime_xi)	-0.348

Moreover, the downward bias in percentage of minorities is more substantial, as shown below, the true coefficient is markedly higher than the multiple regression illustrates. For every 1.4% increase in the percent of minority population in a given county, the reported crime rate increases by 1%. This is an important policy implication to keep in mind.

```
cov_pctmin80_xi <- lm(pctmin80 ~ log(prbPrisonToCrime) +
                        log(density) , data = dfCrime)
true_pctmin80 <- lm(log(crmrte) ~ resid(cov_pctmin80_xi), data = dfCrime)
coef <- data.frame("Multi Step Process Coefficients for Minority"
                   =coef(true_pctmin80))
kable(coef, booktabs=T, digits = 3,
        caption = 'Coefficients of OLS using multistep process for Minority') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 10: Coefficients of OLS using multistep process for Minority

	Multi.Step.Process.Coefficients.for.Minority
(Intercept)	-3.542
resid(cov_pctmin80_xi)	0.014

Finally, here are the results of the multiple regression coefficients alongside their VIF. The percentage of minority variable and the probability of conviction for a given offense variables are underestimated as explained above. The 'west' variable has a relation with the the density variable. Though, none can be considered problematic for analysis purposes.

```
model_2 <- lm(log(crmrte) ~ log(prbPrisonToCrime) + pctmin80 +
                  log(density) + central + west, data = dfCrime)
model_summary <- data.frame("Model 2 Coefficients" = coef(model_2),
                            "VIF" = c(0,vif(model_2)))
kable(model_summary, booktabs=T, digits = 3,
        caption = 'Model 2 Summary') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

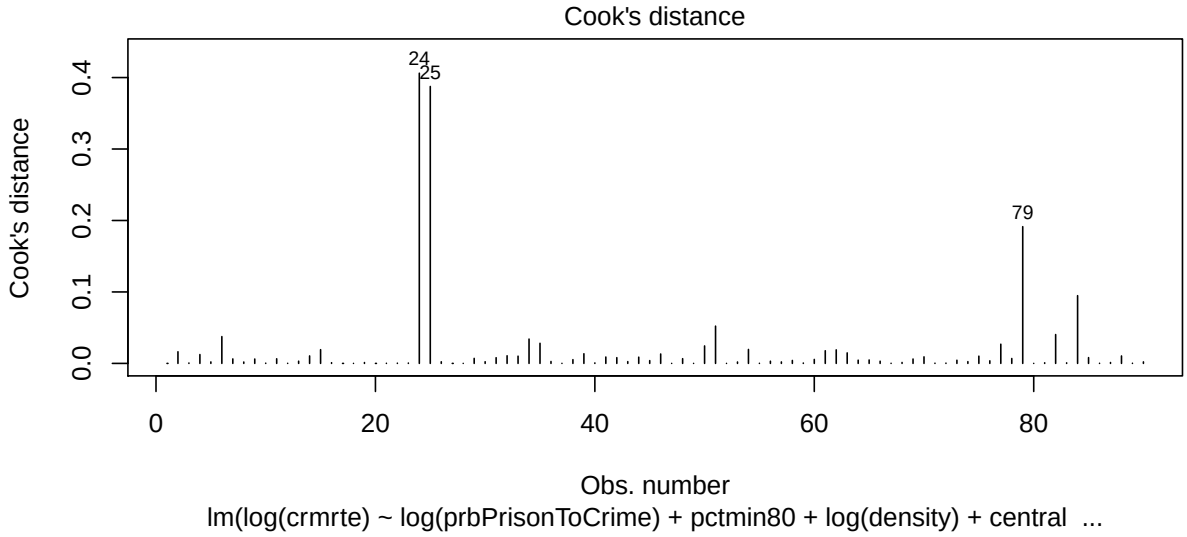
Though the police variable has a high correlation with the crime rate, the causality relationship between the two variables is questionable. Intuitively, an increase in presence of policy should not agitate an increase in crime reporting, in normal circumstances. And police coverage could be driven by crime rate so the causality relationship could be of a different direction.

A deeper study in causality of the police factor can only be attempted if we collect confounding data by increasing police in certain counties, while keeping other factors constant. Finally, let's plot the outliers of *Model 2* to spot any further issues.

Table 11: Model 2 Summary

	Model.2.Coefficients	VIF
(Intercept)	-4.541	0.000
log(prbPrisonToCrime)	-0.308	1.340
pctmin80	0.008	2.262
log(density)	0.434	1.478
central	-0.168	1.666
west	-0.322	2.517

```
plot(model_2, which = 4)
```



In this case, there is no clear outlier that can leverage our model, besides observation 25. The residuals seem to be uncorrelated with each other, as there is no clear pattern in the first graph. Also, if we want to assume normal errors, it is not far from the truth.

As a final analysis, in model 2.6 from table 8, the coefficient test of individual coefficients illustrates that the *pctmin80* variable is not statistically significant, however as we showed earlier by partialling out the coefficient of *pctmin80* the coefficient is understated in the model due to the variability and multicollinearity with the density variable. The partialled coefficient of *pctmin80* would be statistically significant. Using robust estimation we found the folowing final table of results:

```
stargazer(coeftest( model_2, vcov = vcovHC),
          title = "Model 2 with Robust Estimation", header = FALSE,
          table.placement = "H")
```

Table 12: Model 2 with Robust Estimation

	Dependent variable:
log(prbPrisonToCrime)	-0.308^{***}
	(0.103)
pctmin80	0.008^{*}
	(0.004)
log(density)	0.434^{***}
	(0.067)
central	-0.168^{*}
	(0.095)
west	-0.322^{*}
	(0.183)
Constant	-4.541^{***}
	(0.303)
Note:	*p<0.1; **p<0.05; ***p<0.01

The Wald test shows that the coefficients are jointly significant.

```
kable(waldtest(model_2, vcov = vcovHC), booktabs=T, digits = 3,
        caption = 'Joint Test Summary') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 13: Joint Test Summary

Res.Df	Df	F	Pr(>F)
84	NA	NA	NA
89	-5	71.27	0

Alternative Per Person Model

To illustrate the importance of the understated *pctmin80* variable we consider another alternative model with an additional transformation. We transform the observed variable crime rate by multiplying it with the probability of prison sentence (prbarr x prbconv x prbpris x crmrte) to obtain a new variable depicting convictions resulting in prison sentence per person in a given county. This variable is represented by *convpp*. Next we regress this new variable on the percentage of minorities, density, and the dummy region variable.

Additionally, taking a per person perspective on the data, rather than a summary statistic or probability based independent terms, which mask the per person impact or potency of the variables associated with crime.

```
df_alt <- dfCrime
df_alt$convpp <- df_alt$crmrte * df_alt$prbarr * df_alt$prbconv * df_alt$prbpris * 1
alt_model <- lm(convpp ~ pctmin80 + log(density) + west + central , data = df_alt)
stargazer(coef(alt_model),
          title = "Alternative Model", header = FALSE,
          table.placement = "H")
```

Table 14: Alternative Model

(Intercept)	pctmin80	log(density)	west	central
0.001	0.00003	0.0003	-0.0001	-0.0001

The residual term variability in this model is much more narrower and thus the practical and statistical significance of the *pctmin80* variable is clearly visible.

Model #3

Finally, let's include all variables in our model. This will help us analyze the robustness of our *Model 2* and which variables are not statistically significant in the end. The table will be in a tiny font so it can fit in the page.

```
# Model 2
model_3 <- lm(log(crmrte) ~ log(prbPrisonToCrime) + pctmin80 + urban+
                  pctymle + log(density) + central + west+avgsen+polpc+taxpc+
                  wcon+wtuc+wtrd+wfir+wser+wmfg+wfed+wsta+
                  wloc+mix, data = dfCrime)
model_3$AIC <- AIC(model_3)
# Table of results
stargazer(model_3,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "adj.rsq", "n", "aic"), header = FALSE,
          table.placement = "H", font.size = 'tiny')
```

Table 15: Linear Models Predicting Crime Rate

	Dependent variable:
	log(crmrte)
log(prbPrisonToCrime)	-0.272^{***}
	(0.066)
pctmin80	0.008^{***}
	(0.003)
urban	-0.069
	(0.166)
pctymle	3.052^{*}
	(1.535)
log(density)	0.338^{***}
	(0.081)
central	-0.162^{*}
	(0.085)
west	-0.217^{*}
	(0.123)
avgsen	-0.025^{*}
	(0.013)
polpc	79.705
	(78.007)
taxpc	0.006^{*}
	(0.004)
wcon	0.001
	(0.001)
wtuc	0.001
	(0.0005)
wtrd	0.001
	(0.002)
wfir	-0.001
	(0.001)
wser	-0.002^{**}
	(0.001)
wmfg	0.00001
	(0.0005)
wfed	0.003^{***}
	(0.001)
wsta	-0.001
	(0.001)
wloc	-0.0003
	(0.002)
mix	-0.355
	(0.444)
Constant	-5.348^{***}
	(0.619)
Observations	90
R^2	0.806
Adjusted R^2	0.750
Akaike Inf. Crit.	42.800
Note:	*p<0.1; **p<0.05; ***p<0.01

From this table, we can observe how robust our model is. Although there is a couple of new variables that seem to be statistically significant (wages and average sentences), in terms of policy, they don't seem to add anything to our selected model. The change in AIC is almost insignificant, and the adjusted R^2 increase only by 3 % of variance explained. For policy, *Model 2* seem to be a clear winner for modeling the crime rate in the State of North Carolina.

CLM Assumption Analysis

Let's now shift our attention to the Classic Linear Model (CLM) Assumptions of our models. We're going to focus our attention on Model #2, although we will highlight any major deviation from the other two models when they occur.

MLR #1: Linearity in Parameters

We made log transformations of the dependent variable (*crmrte*), and made log transformations of our transformed variables of probabilities. Although our transformed variable is a multiplication of three independent variables, being inside a log transformation means the model is linear in terms of the parameters. The three models *don't violate MLR #1.*

MLR #2: Random Sampling

Let's now focus on the dataset and the random sampling assumption. As we mentioned during our EDA, the observations in the data include 90% of all counties in North Carolina, thus the data seems to be representative of the population. However, since it's unknown how the remaining 10% was removed from the dataset, we don't know if it was selected randomly. Let's look at the variances of several variables to try to understand the randomness of the sampling. Let's start by the outcome variable.

```
vars <- "crmrte"
v_mean <- lapply(dfCrime[vars], mean, na.rm=TRUE)
v_mean <- round(as.numeric(v_mean), 4)
v_min <- lapply(dfCrime[vars], min, na.rm=TRUE)
v_min <- round(as.numeric(v_min), 4)
v_max <- lapply(dfCrime[vars], max, na.rm=TRUE)
v_max <- round(as.numeric(v_max), 4)
v_sd <- lapply(dfCrime[vars], sd, na.rm=TRUE)
v_sd <- round(as.numeric(v_sd), 4)
v_tab <- cbind(Min=v_min, Mean=v_mean, St.Dev=v_sd , Max=v_max)
rownames(v_tab) <- vars
kable(v_tab, booktabs=T, digits = 4,
        caption = 'Summary dataset of crmrate') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Table 16: Summary dataset of crmrate

	Min	Mean	St.Dev	Max
crmrte	0.0055	0.0335	0.0189	0.099

The range seems big enough and the standard deviation is not small compared to the mean. Also, we know that the log transformation is normally distributed, so we can assume there wasn't any foul play in the selection.

Finally, let's run a Durbin-Watson test on our three models to check for autocorrelation on the residuals:

```
dw1 <- dwtest(model_prob_density, alternative = "two.sided")$p.value
dw2 <- dwtest(model_2, alternative = "two.sided")$p.value
dw3 <- dwtest(model_3, alternative = "two.sided")$p.value
DW <- data.frame(Model1 = dw1, Model2 = dw2, Model3 = dw3)
rownames(DW) <- "P-Values"
kable(DW, booktabs=T, digits = 4,
        caption = 'p-values of Durbin-Watson test') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

As we can observe, Model #1 and #2 don't have issues of autocorrelation. Model 3 have serious problem of autocorrelation of the residuals, which means a *violation of MLR #2.*

Table 17: p-values of Durbin-Watson test

	Model1	Model2	Model3
P-Values	0.6663	0.067	0.026

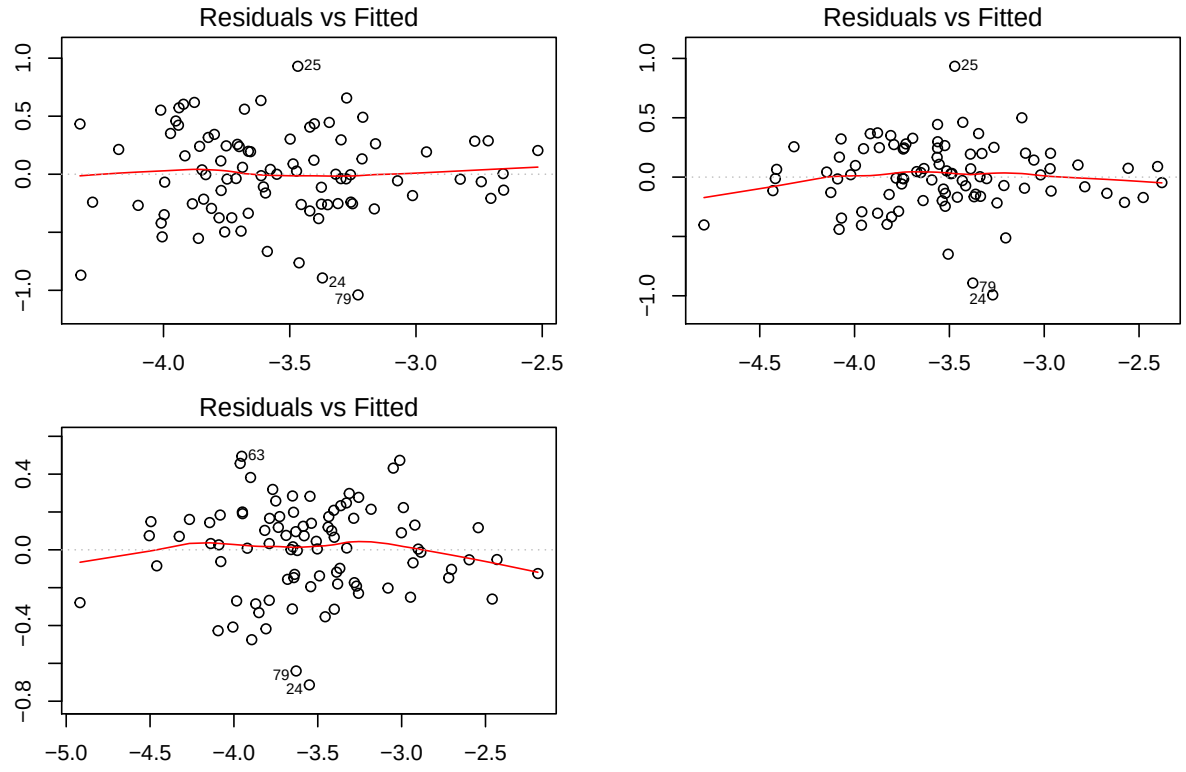
MLR #3: No Perfect Collinearity

As we observed in our heatmap on the EDA, many of the covariates are correlated with each other. However, besides the indicator variables that are automatically removed from the final model by R, there are no perfectly correlated variables. Thus, although we need to use robust estimation of the errors to account for the inflation of multicollinearity, the assumption is not entirely violated. Finally, in our Model 2 summary, no VIF was greater than 10, thus the collinearity problem is not a big deal.

MLR #4: Zero Conditional Mean

Let's check the diagnostics plots to determine if there is a violation of this assumption.

```
par(mfrow=c(2,2), mar=c(2,3,2,2))
plot(model_prob_density, which = 1)
plot(model_2, which = 1)
plot(model_3, which = 1)
```

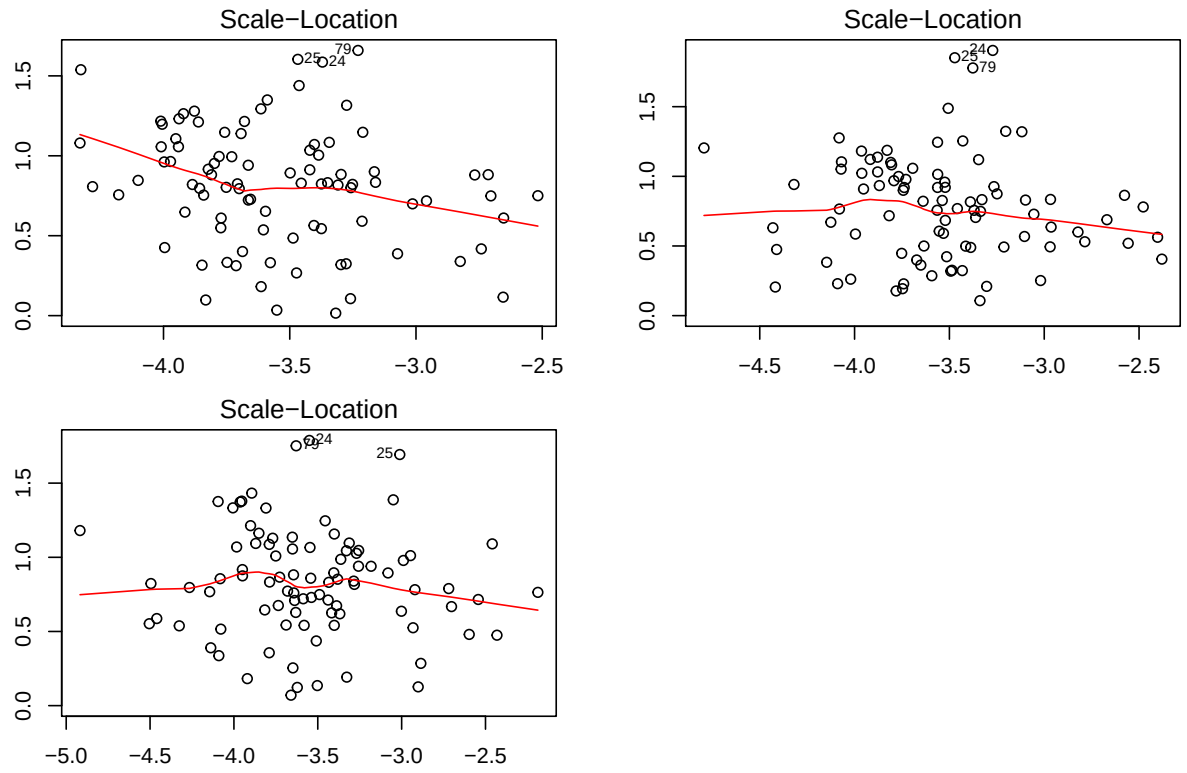


On the plots, we don't observe a huge deviation from zero conditional mean, as the red lines tend to be flat on the three plots. However, the first and third models seem to have a more pronounced deviation from "flatness", especially at the ends, which could mean some problems regarding zero conditional mean, but nothing major. Any deviation from zero conditional mean could be explained by omitted variables.

MLR #5: Homoskedasticity

For homoskedasticity, let's first turn our attention to the Scale-Location plot.

```
par(mfrow=c(2,2), mar=c(2,3,2,2))
plot(model_prob_density, which = 3)
plot(model_2, which = 3)
plot(model_3, which = 3)
```



Here the picture seems to be very clear. All models suffer from heteroskedasticity, as the mean lines are not flat, presenting clear patterns in each case. This again leads us to use robust estimation for our standard errors for our final regression tables. To be sure, let's also run a Breusch-Pagan test on the three models.

```
bp1 <- bptest(model_prob_density)$p.value
bp2 <- bptest(model_2)$p.value
bp3 <- bptest(model_3)$p.value
BP <- data.frame(Model1 = bp1, Model2 = bp2, Model3 = bp3)
rownames(BP) <- "P-Values"
kable(BP, booktabs=T, digits = 4,
        caption = 'p-values of Breusch-Pagan test') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

In all cases, we have enough statistical evidence to reject the null hypothesis of homoskedasticity. *All models violate MLR #5, thus we need to use robust estimators for our standard errors*

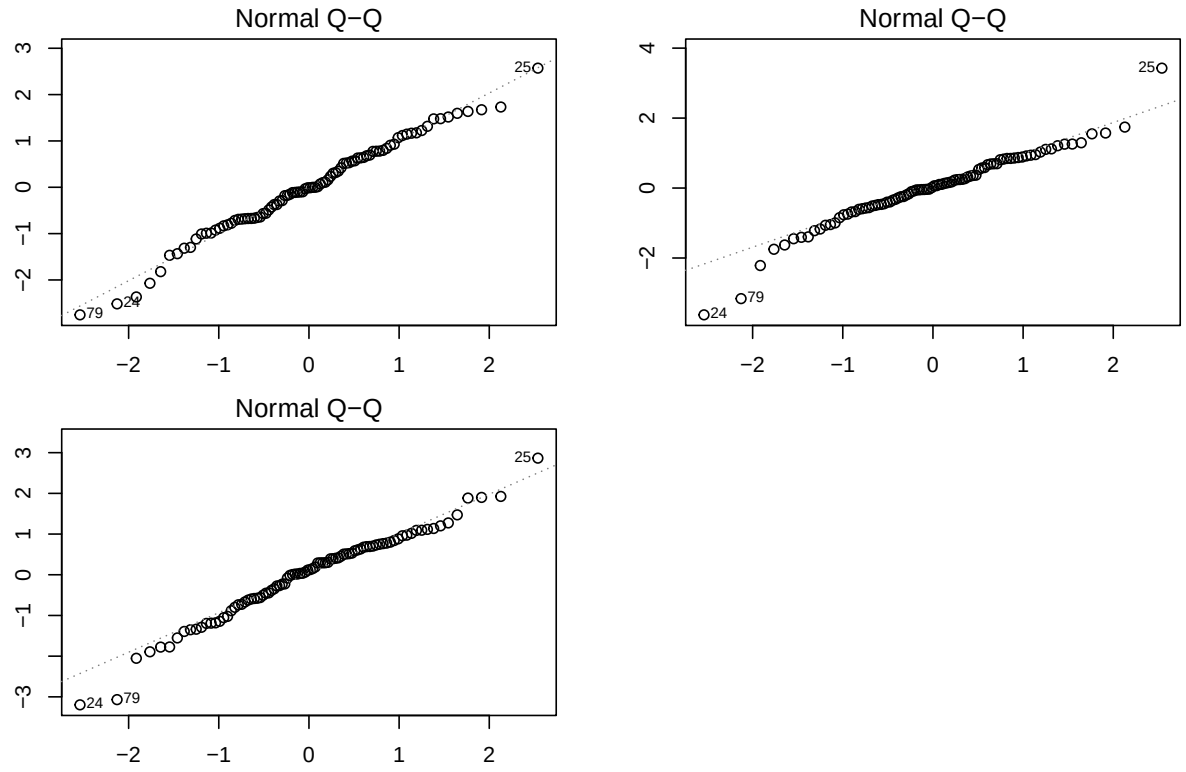
MRL #6: Normality of Errors

Finally, let's check the QQ plots of the residuals to check if errors are normally distributed.

Table 18: p-values of Breusch-Pagan test

	Model1	Model2	Model3
P-Values	4e-04	0.0036	0.0112

```
par(mfrow=c(2,2), mar=c(2,3,2,2))
plot(model_prob_density, which = 2)
plot(model_2, which = 2)
plot(model_3, which = 2)
```



Here we don't have any major departure from normality other than some observations at the tails. To be sure, let's run a Shapiro-Wilk test on the residuals.

```
sw1 <- shapiro.test(model_prob_density$residuals)$p.value
sw2 <- shapiro.test(model_2$residuals)$p.value
sw3 <- shapiro.test(model_3$residuals)$p.value
SW <- data.frame(Model1 = sw1, Model2 = sw2, Model3 = sw3)
rownames(SW) <- "P-Values"
kable(SW, booktabs=T, digits = 4,
        caption = 'p-values of Shapiro-Wilk test') %>%
        kable_styling(latex_options='hold_position',position = 'center')
```

Only model 2 violates assumption MLR #6

Table 19: p-values of Shapiro-Wilk test

	Model1	Model2	Model3
P-Values	0.7178	0.0115	0.2561

Omitted Variables and Possible Biases

In this section we will discuss 5 omitted variables and their effect on our model. Current we have

$log(crmrte) = \beta_0 + \beta_1 * prbPrisonToCrime + \beta_2 * log(density) + \beta_3 * pctmin80 + \beta_4 * central + \beta_5 * west$

The following table shows our assessment of the expected correlation between each omitted variable and the dependent variable as well as the explanatory variables.

Type and Severity of Crime

More detailed and categorical data on the dependent variable such as the types and severity of crime would help us shed more light on the crime rate. For example, the nature of face-to-face crime such as assault is very different from property damage or graffitting in a neighborhood. Though they each count as crime, the probability of prison sentence for a minor property crime with no witness may not result in a prison sentence. As many minor crimes are often unreported, there should be a moderately positive correlation between *crmrte* and severity of crime. When a crime is more severe, there will be more investigation, which leads to higher rate of convinction. Therefore we expect a strong positive correlation between severity of crime and *prbPrisonToCrime*, causing a positive bias on β_1, toward zero. We also see a correlation with *density* since more populated areas will likely have different types of crime so β_2 has positive bias, away from zero. The dataset provides a variable *mix* that can proxy the severity of crime to a certain extent, but *mix* only indicates whether the crime is face-to-face without further detail.

Poverty Level

Poverty level can have a strong influence on crime. Although many wage variables are present in the dataset and can serve as a proxy, there is no conlusive way to consolidate the wage variables to get a sense of the overall economic status of each county or the population. As discussed earlier, without knowing the weights of each industry, the industry-specific wages do not help assess the economic situation of the county. Therefore, unfortunately, our models could not make good use of the wage variables. We expect poverty level to have a strong positive correlation with *crmrte*. In particular, property thefts tend to occur more when criminals feel a sense of financial hardship and even food insecurity. We also expect a moderate positive correlation with *density* and a mild positive correction with geographic variables since poverty is unevenly distributed across space and concentrate primarily in inner-city America and rural places. Therefore, β_2 has positive bias, away from zero.

Education

We expect a moderately negative correlation between education and *crmrte* because higher level of education often lead to higher income, making the opportunity cost of an educated individual being in jail higher. Education can also strengthen a person's moral standards so in that respect, it can help reduce crime rate. Due to the disparities in educational opportunties, communities of color and minorities often achieve lower level of education than whites. Therefore, the correlation between education level and *pctmi80* is expected to be moderately negative, which means that β_3 is positively biased, away from zero. The dataset does not provide any information to serve as proxy on level of education so additional data collection would be needed.

Unemployment Rate

Higher levels of unemployment rate is often associated with elevated levels of crime. While there are public assistance programs such as food stamp and unemployment benefits to alleviate the situation, we expect a moderately positive correlation between unemployment rate and *crmrte*. In particular, property theft often increase as unemployment rate increases since criminals commit this kind of crime as an income substitution. The dataset does not provide employment information, however, the wage variables could serve as a proxy if the unemployment rate could not be collected. Higher wages typically indicate economic bloom, which could mean lower unemployment rate. Likewise, when unemployment rate is high, the county may be experiencing economic issues, which could impact the amount of resources for police and negatively affect *prbPrisonToCrime*. This would cause β_1 to have small negative bias, away from zero.

Illegal Drug Use

Illegal drug use has a strong correlation with crime rate. In fact, the possession of illegal drug is already a crime. And illegal drug usage can impair a person's mental health and lead to addiction. While intoxicated, the person can commit crimes of varied severity, ranging from shoplifting to rape. Counties with high level of illegal drug use tend to have higher probability of prison so the correlation with *prbPrisonToCrime* is positive. This is because people who use illegal drugs are more likely to be arrested as a suspect of crime, leading to higher probability of arrest, which causes *prbPrisonToCrime* to be higher. Therefore, we expect β_1 to have positive bias, toward zero. The dataset does not provide any information on level of illegal drug use so additional data collection would be needed.

Bias in the Measurement Variable

Most importantly, any bias in the observed variable in this study, crime rate, would have the most profound impact on this study. The crime rate only considers the reported cases of crime. The number of crimes that may go unreported, such as property theft in a rural area with low police density, long response time and lack of witnesses, may deter reporting of such crimes. Our analysis reveals a positive relation in conviction rates and percentage of minorities, thus a higher incarceration rate may deter reporting of such crimes by certain segments of the population if they believe the law does not treat them fairly.

Conclusion

For a campaign to be successful, it has to address the concerns of the constituency. Since crime rate is a universal concern, effective policies can help the campaign tremendously. Based on the data given, we arrived at a model that allows us to see how different factors affect the crime rate.

The probability of prison, which is proxy by the ratio of conviction to offense, is a strong deterrent of crime. The coefficient of the log-log regression between crime rate and probability of prison suggests that for each 1% increase in probability of prison per offense, crime rate decreases by 0.31%. Therefore, the campaign should include policies to increase the probability of prison per offense, which ultimately means a higher success rate at solving crimes. The campaign could suggest a new incentive system that rewards detectives and their teams based on the success and speed of the investigation and the difficulty of the cases.

Punishment and efficiency in pursuing a resolution of a reported crime can be strong deterrents for future crimes. However, the thorny relation between race and reported crimes encompasses a multitude of social issues and biases. Policy makers need to take cohesive and well-rounded approach to tackling crimes swiftly through an efficient and equitable justice system. At the same time, in the denser counties, they should introduce policies to promote social programs which will gather richer data on the socio-economic issues faced by densely populated and high minority population counties. The policies should also extend to unconscious bias training programs for the law enforcement agencies.

In geographical terms the 'central' and the 'west' counties face distinctive issues. The 'west' counties tend to be low density rural areas, while the 'central' counties are more densly populated with higher minority populations. In the 'west' counties crime reduction policies such as rural crime watch associations may help reduce incidences of crime. As mentioned earlier caution must be exercised in interpreting the low crime rates in rural counties.

Public policy is focused on maximizing the well-being of the population and reducing the harmful effects of crime on the society while ensuring equitable treatment of all residents within the justice system. Crime poses sustantial economic and intangible costs to the society.

At the second stage of the analysis we can conduct tests to assess the robustness of our model. Additionally we can check for endogeneity by plotting the explanatory variables against the error term to check for bias and inconsistency in the model. We can check the variance of the explanatory variables and assess their consistency. We will also check if the main assumptions of linear model are met.

The model construction tells us that the multicollinearity in multivariate models gives rise to substantial tradeoffs as the size of coefficients can vary widely and modify causal relations in the model. Some variables may need to be dropped even though they increase the R-squared of the model.