W203 Lab 3 Draft

Mihir Sathe, Kathryn Papandrew, Anu Sankar, Rahul Vaswani

7/22/2018

1. Introduction

As part of our affiliation with the current political race in North Carolina, our team explored some of the factors that are associated with high crime rates. By examining several features and then conducting a model comparison of a few Ordinary Least Squares models, we are able to test our hypothesis that **whether or not a county is considered urban, the density, the percent of young males, the tax revenue per capita, the police per capita, and probability of arrests** in a specific county can be associated with a variety of crime rates.

Our motive behind our hypothesis is essentially the feasibility of proposing and then actually implementing these propositions. We can utilize established economic reforms, public recreation policy, hiring caps for public service, rehabilitation centers, and more stringent rules leading to arrests in order to influence these factors.

Thus our goal is to come up with an OLS model that will approximate crime rates and which we can ultimately use to generate policies that will help local governments.

2. Exploratory Data Analysis

Load in the data:

```
library(ggplot2, quietly = TRUE)
library(car, quietly = TRUE)
library(stargazer, quietly = TRUE)
library(sandwich, quietly = TRUE)
library(lmtest, quietly = TRUE)

crime = read.csv(
  "https://docs.google.com/spreadsheets/d/e/2PACX-1vRGHJsTAmsBXex_Ui7KjOD3K54-oDpjobMiXVfTHtNhdmEGA-lKofWz
  )
```

2.0 Preliminary Data Checks & Scrubbing

Upon loading the data, we noticed the following abnormalities:

- **Rows with fully null data**, rendering them ~~fully~~ useless in the analysis.
- **The variable `prbconv` needed to be munged** in order to remove odd characters that affected numerical calculations. (Note that this data was part of a row where the rest of the columns had null values)
- **Duplicate entry for county 193** which was removed.

Other than those items, the data we used in our exploratory data analysis was raw.

```
c <- suppressWarnings(transform(subset(crime, !is.na(county)),
                                prbconv = as.numeric(levels(prbconv))[prbconv]))
c = c[!duplicated(c$county), ]
```

2.1 Outcome Variable

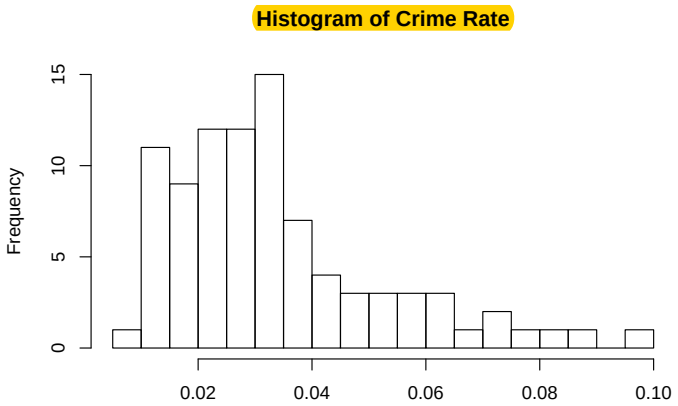
2.1.1 Crime Rate (crmrte)

The variable crmrte represents crimes commited by a person in the year 1987. ~~The 1980s in general were a very bad decade for crimes. There was a crack-cocaine epidemic, gun violence doubled, and homicide rates were at an all time high. Thus, this report will focus on developing a model we can use to draw policy to reduce crime rates for this upcoming political campaign.~~

According to the Uniform Crime Reporting Statitistics, crime rate can be defined as, "crimes are per capita (number of crimes per 100,000 persons)"(UCR).

Taking a look at the data:

```
hist(c$crmrte,
     breaks = 30,
     main = "Histogram of Crime Rate",
     xlab = NULL)
```



We see above that the distribution is postively skewed with the bulk observations being focused around 2%. We also notice a maximum value that is approximately 10%. We should take a further look to see what county this high crame rate belongs to:

```
c[c$crmrte == max(c$crmrte), c("county", "crmrte")]
```

```
##    county    crmrte
## 53    119 0.0989659
```

We use a lookup table to map that the county with the highest crime rate is 119 which is Mecklenburg, the county containing Charlotte which is North Carolina's largest city. ~~From our histogram above we can see that there are quite a few values that are greater than 4%, thus our goal is to try and come up with a model that will allow us to associate a few features to approximate a reduction in crime rate.~~

2.2 Key Predictor Variables

In this section, the variables selected for our core model will be explored & explained in both their univariate & relevant bivariate analysis.

Variables explored are: prbarr, urban, density, pctymle, taxpc, polpc

2.2.1 'Probability' of Arrest (prbarr)

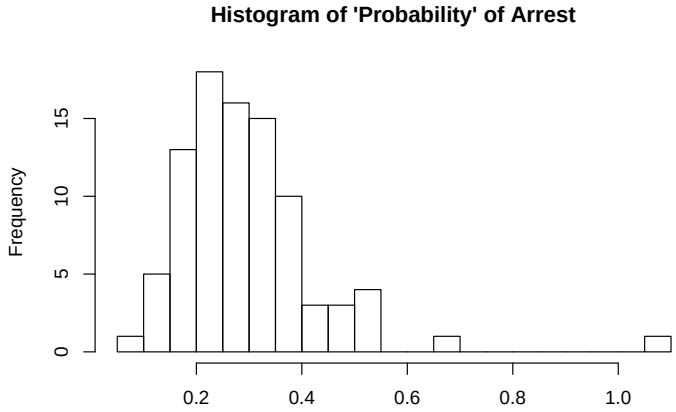
This variable prbarr describes the ratio of arrests to offenses.

It's important to note that the term "probability" is quoted because of its loose definition here where there can be over a 100% "chance" of being arrested as we see in Madison County (county 115).

2.2.1.1 Univariate Analysis

The data is right skewed with the previously-stated maximum value for Madison County of 1.09. Most counties are centered around 0.2-0.3 range where there are 0.2-0.3 arrests for each offense committed.

```
hist((c$prbarr), breaks = 20, main = "Histogram of 'Probability' of Arrest", xlab = NULL)
```




```
summary(c$prbarr)
```

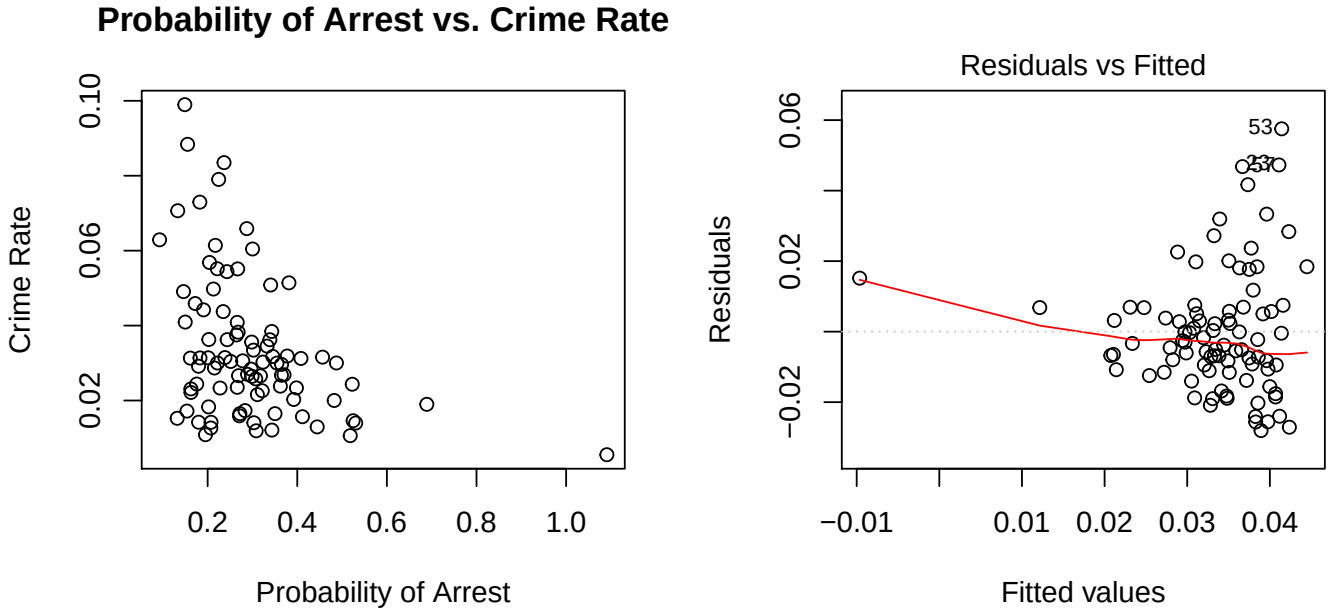
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20490 0.27150 0.29520 0.34490 1.09100
```

2.2.1.2 Bivariate Analysis

2.2.1.2.1 Probability of Arrest vs. Crime Rate

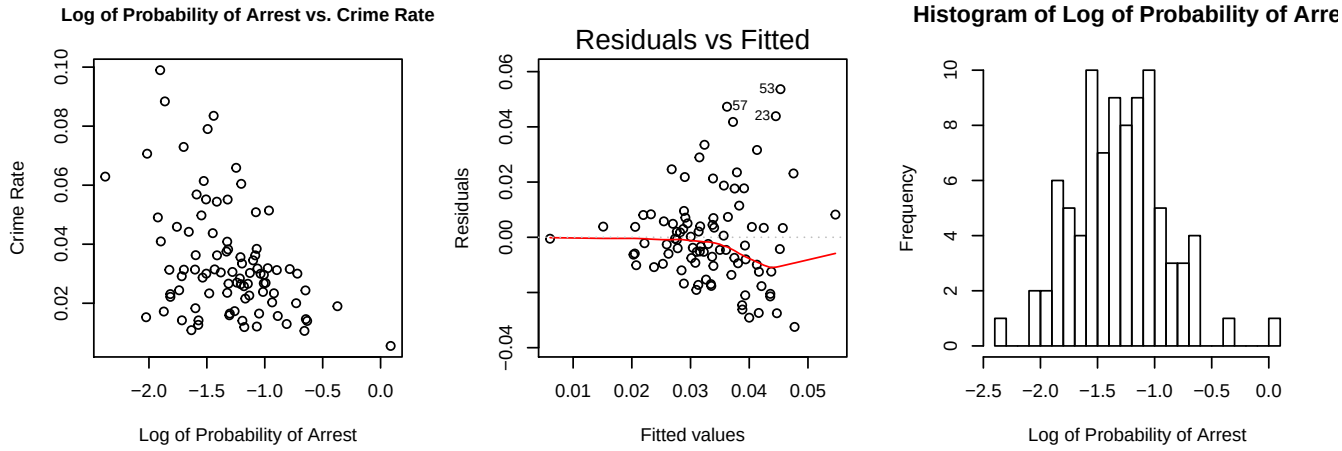
When analyzing the 'Probability' of Arrest data with respect to our outcome variable, we see that the relationship does not appear linear. This is accentuated with a simple linear model that shows the residuals vs. fit plot is also not well-distributed & linear.

```
par(mfrow = c(1,2))

plot(c$crmrte ~ c$prbarr, main="Probability of Arrest vs. Crime Rate",
     xlab = "Probability of Arrest", ylab = "Crime Rate")
plot(lm(c$crmrte ~ c$prbarr), which=1)
```



A log transformation was applied to see if a linear relationship exists between the crime rate & the log of the probability of arrest. Once the log was applied, a check of the distribution of values confirmed that the log transformation yielded a more normal distribution~~, as well.~~ These are visualized below.

```
par(mfrow = c(1,3))

plot(c$crmrte ~ log(c$prbarr), main="Log of Probability of Arrest vs. Crime Rate",
     cex.main = 1, xlab = "Log of Probability of Arrest", ylab = "Crime Rate")
plot(lm(c$crmrte ~ log(c$prbarr)), which=1)
hist(log(c$prbarr), breaks = 20, main="Histogram of Log of Probability of Arrest",
     xlab = "Log of Probability of Arrest")
```



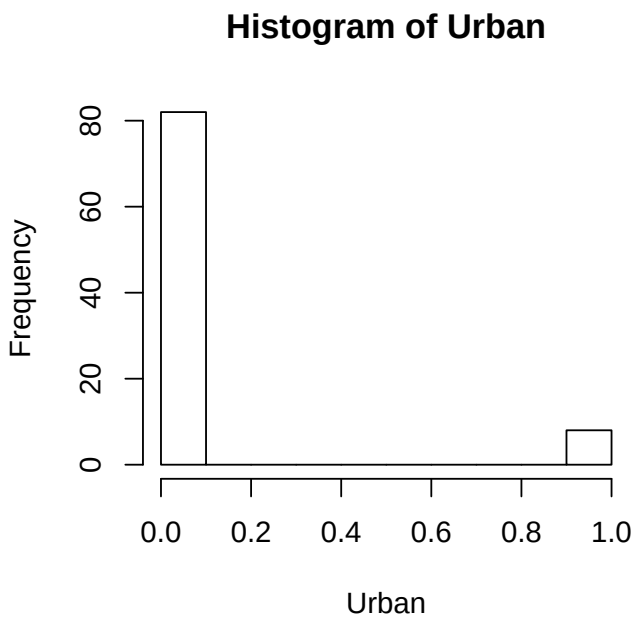
2.2.2 Urban (`urban`)

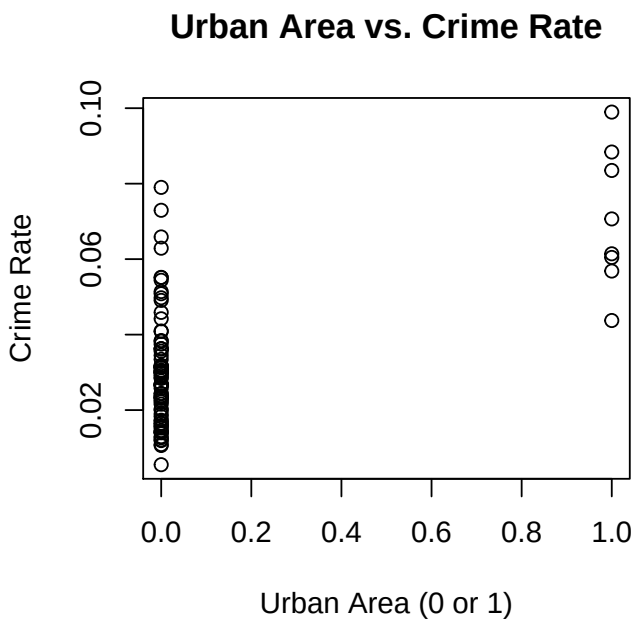
Urban is an indicator variable with the values `0` and `1`. This variable is helpful in a treatment effect capacity to eliminate gray area - is an area urban or not?

2.2.2.1 Univariate Analysis

A cursory check of the variable shows that the majority of counties do not contain urban areas. This makes sense as North Carolina only has a couple major metropolitan areas.

```
par(mfrow = c(1,2))

hist(c$urban, breaks = 10,
     main="Histogram of Urban",
     xlab="Urban")
plot(c$crmrte ~ c$urban, main = "Urban Area vs. Crime Rate",
     xlab = "Urban Area (0 or 1)",
     ylab = "Crime Rate" )
```





2.2.2.2 Bivariate Analysis

When compared against crime rate, we see (above) that the highest crime rate occurs in the counties having an urban area.

2.2.3 Density (density)

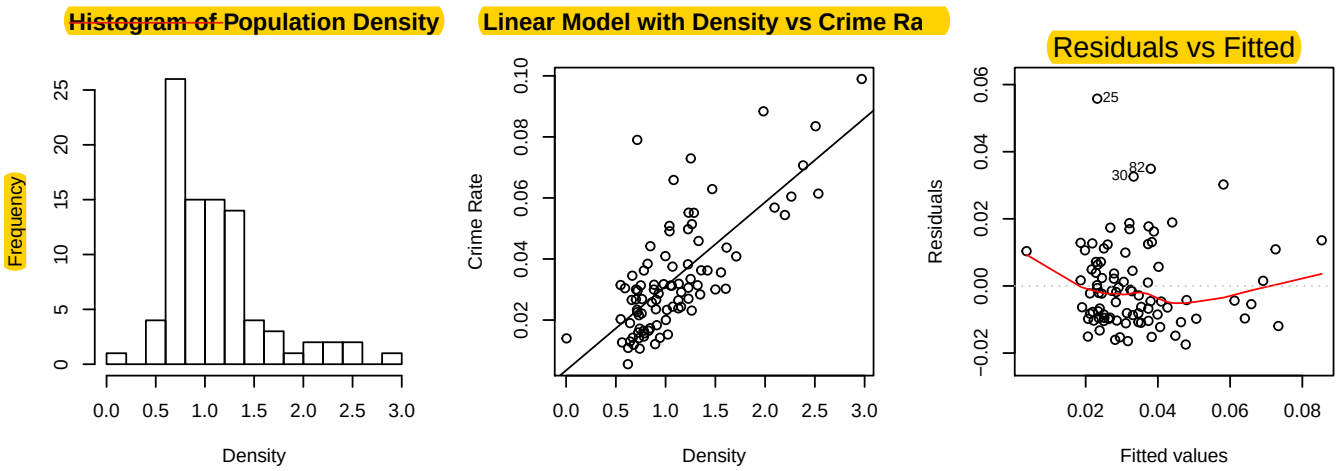
2.2.3.1 Univariate Analysis

The variable density is the number of people per square mile. It is skewed to the left which might be a result of higher population in the cities. We definitely want to explore how the crime rate changes with population density. It would be of interest to use the 'urban' variable along with 'density' in our model as crime rates tend to be higher in cities than in rural areas in general.

2.2.3.2 Bivariate Analysis

We see that density has a linear relationship with crime rate.

```
par(mfrow = c(1,3))

hist(sqrt(c$density),
     main = "Histogram of Population Density",
     xlab = "Density",
     breaks=20)

plot(sqrt(c$density), c$crmrte,
     main = "Linear Model with Density vs Crime Rate",
     xlab = "Density",
     ylab = "Crime Rate")

cr_d = lm(c$crmrte ~ sqrt(c$density), data = c)
abline(cr_d)
plot(cr_d, which=1)
```



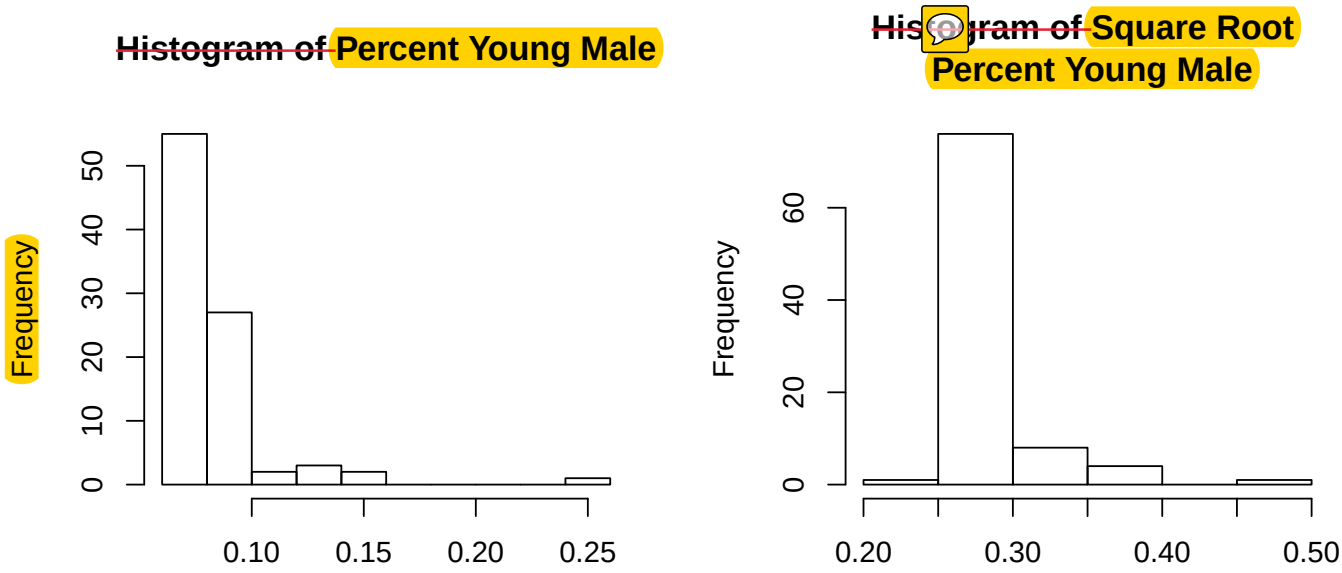
2.2.4 Percent Young Male (`pctymle`)

2.2.4.1 Univariate Analysis

We define a young male as a 15-24 year old male. We choose to examine this variable because we suspect that some crimes could be attributed to misguided youth or gangs. The United States Census suggests that one in three young males live at home, and of those, one in four do not have jobs. If our model approximates significant correlations between young males and crime rate, we will examine some initiatives that can help reduce this relationship.

We first begin by looking at the histogram:

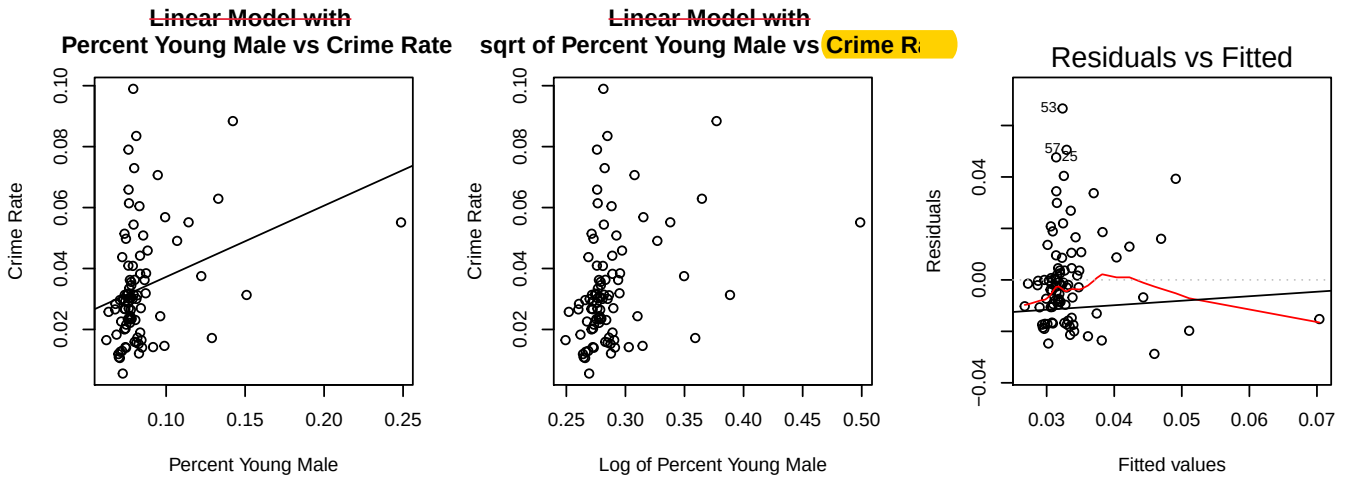
```
par(mfrow = c(1,2))

hist(c$pctymle,
     main = "Histogram of Percent Young Male",
     xlab = NULL)

hist(sqrt(c$pctymle),
     main = "Histogram of Square Root \n Percent Young Male",
     xlab = NULL)
```



The first thing we notice is that the distribution is positively skewed and will require a transformation. We see above that a log transformation slightly helps normalize the distribution. There does not generally seem to be a large concentration of young men in the majority of counties in North Carolina. The second thing (and perhaps more

importantly) we notice is that we have a large outlier. We will further examine this outlier in our bivariate analysis to test its influence on crime rate.

```
par(mfrow = c(1,3))

plot(c$pctymle, c$crmrte,
     main = "Linear Model with \n Percent Young Male vs Crime Rate",
     xlab = "Percent Young Male",
     ylab = "Crime Rate")


cr_pym = lm(c$crmrte ~ c$pctymle, data = c)
abline(cr_pym)

plot(sqrt(c$pctymle), c$crmrte,
     main = "Linear Model with \n sqrt of Percent Young Male vs Crime Rate",
     xlab = "Log of Percent Young Male",
     ylab = "Crime Rate")


cr_lpym = lm(c$crmrte ~ sqrt(c$pctymle), data = c)
plot(cr_lpym, which =1)
abline(cr_lpym)
```




```
cor(c$crmrte, log(c$pctymle))
```

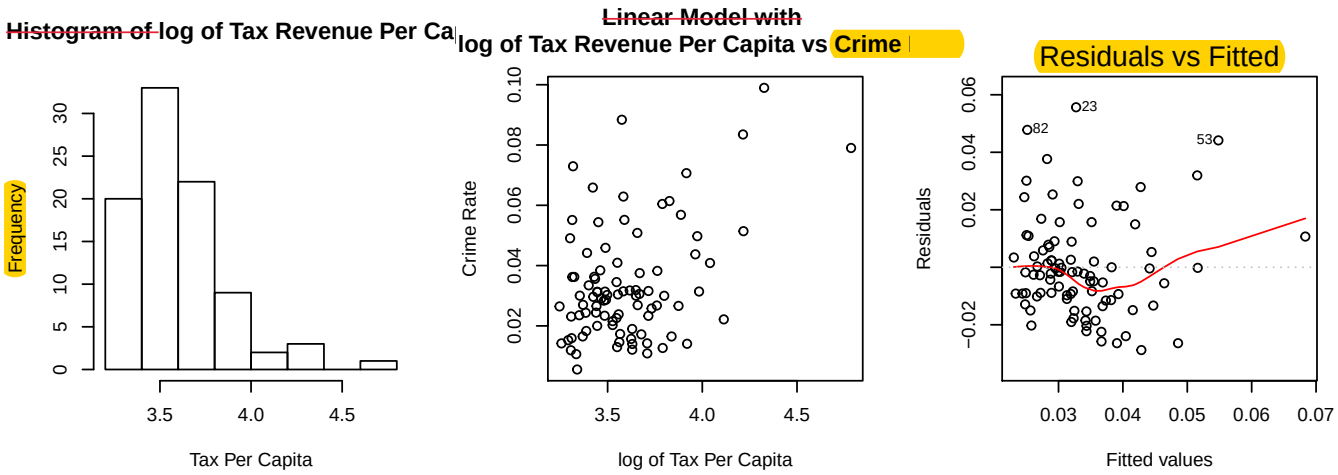
```
## [1] 0.3241625
```

2.2.5 Tax Revenue per Capita (taxpc)

The variable 'taxpc' is the tax revenue collected per person. A higher the tax revenue implies that the population is more productive and has a higher income. So, one notion can be that people who are gainfully employed will not indulge in crime. But on the contrary, they can be targets to crime. The positive correlation(0.45) between 'taxpc' and 'crmrte' variable points out the credence of the second hypothesis. As the 'taxpc' variable is skewed to the left, we tried a log transformation on it in our model.

```
par(mfrow = c(1,3))

hist(log(c$taxpc),
     main = "Histogram of log of Tax Revenue Per Capita",
     xlab = "Tax Per Capita"
    )
```

```
plot(log(c$taxpc), c$crmrte,
     main = "Linear Model with \n log of Tax Revenue Per Capita vs Crime Rate",
     xlab = "log of Tax Per Capita",
     ylab = "Crime Rate"
    )
cr_ltpc = lm(c$crmrte ~ log(c$taxpc), data = c)
plot(cr_ltpc, which =1)


abline(cr_ltpc)
```



2.2.6 Police per Capita (polpc)

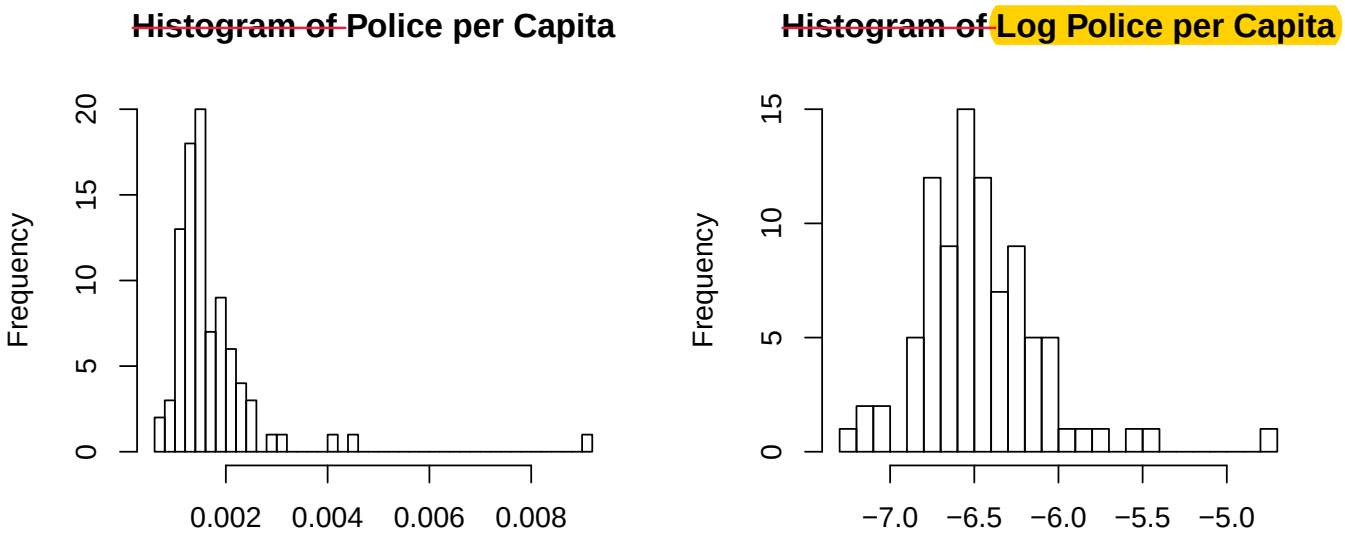
The variable polpc represents police per capita. Police per capita essentially measures the number of police officers per number of people in a specific county. Thus a low number would signfiy a low police presence and a high number would signify a large police presence.

Our rationale for choosing to focus on this feature is that there have already been successful policies in place in terms of hiring quotas for police officers (McCrary, 2006), thus we could also implement state policies in order to limit or increase hiring quotes for the police.

2.2.6.1 Univariate Analysis

We first look at the distribution of the variable:

```
par(mfrow = c(1,2))

hist(c$polpc,
    breaks = 30,
    main = "Histogram of Police per Capita",
    xlab = NULL)

hist(log(c$polpc),
    breaks = 30,
    main = "Histogram of Log Police per Capita",
    xlab = NULL)
```



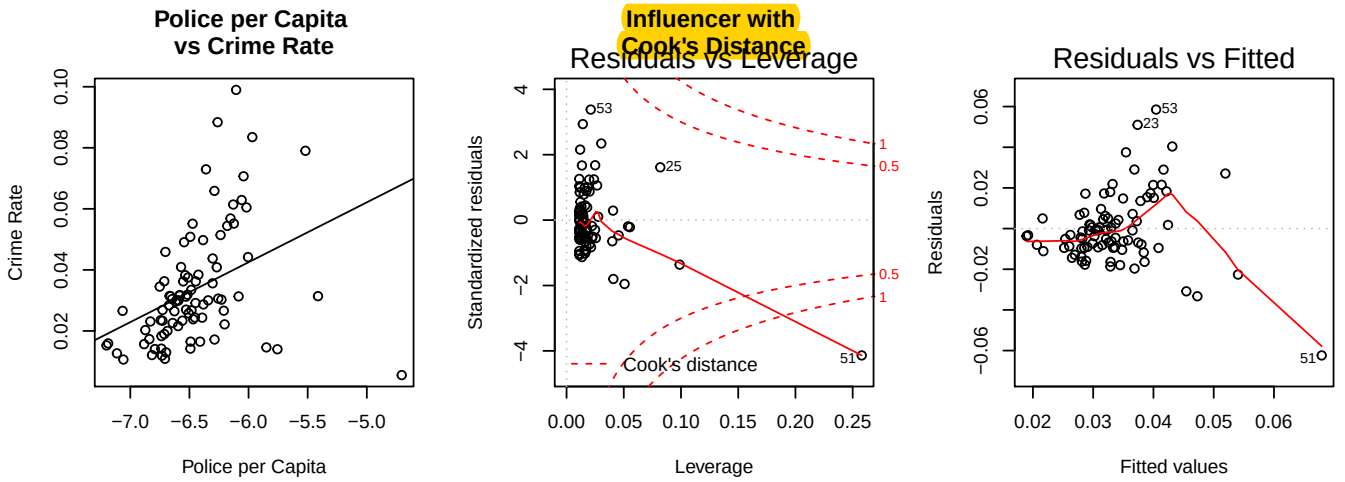
We see a very large positive skew with a big outlier. Generally, the number of police per capita seems quite low; hoever, we do see a few outliers around 0.004 and then a larger outlier past 0.008. This outlier is large enough that we should explore the influence in our bivariate analysis. Because of the skew, we would need to apply some sort of transformation to the variable.

We see above that the natural logarithm does indeed give us a more normal distribution than the untransformed variable, thus we will be using this transformation in our model. Therefore, the interpretation will be that crime rate will change as the *percentage* of police per capita changes.

2.2.6.2 Bivariate Analysis

We now examine the relationship between police per capita and crime rate, along with the Cooks distance to see how influential that outlier is to crime rate.

```
par(mfrow = c(1,3))
options(repr.plot.width = 15, repr.plot.height=2)

cr_ppc = lm(crmrte ~ log(polpc), data = c)
plot(log(c$polpc),
     c$crmrte, main = "Police per Capita \nvs Crime Rate",
     xlab = "Police per Capita",
     ylab="Crime Rate")
abline(cr_ppc)
plot(cr_ppc, which = 5, main = "Influencer with \nCook's Distance")
plot(cr_ppc, which =1)
```



We see that our suspicions about the influence of that outlier we saw in our univariate analysis is justified. The point has a Cook's distance greater than 1, which means that it will indeed influence our regression (although, with more variables, the effect may be less prominent). Despite the large influence, we maintain that we keep this variable in our analysis because there is no evidence that this is a faulty data point.

2.3 Key Covariate Variables

In this section, the variables selected for our covariate models will be explored & explained in both their univariate & relevant bivariate analysis.

Variables explored are: `avgsen`, `wfir`, `wser`, `wfed`, `wsta`, `wloc`, `wcon`, `wtuc`, `wtrd`, `wmfg`, `minorities`, and `prbconv`

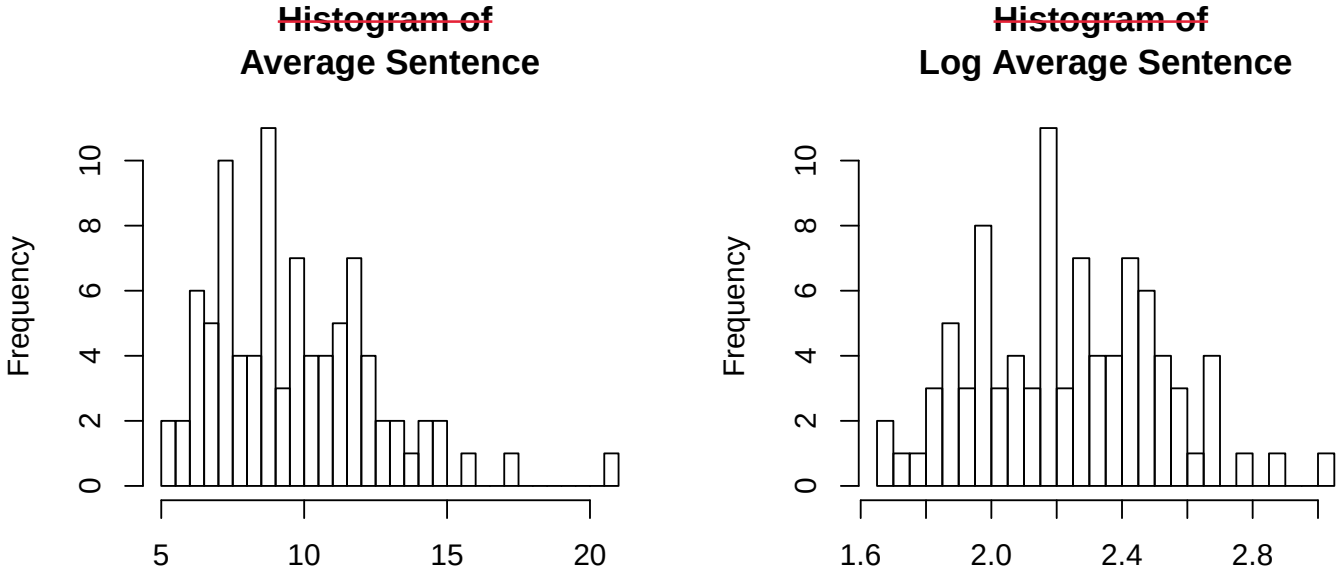
2.3.1 Average Sentence (`avgsen`)

Average sentence represents the number of days a person spends in jail. We notice that the max is 21 days, thus the jail time represented in our data is for misdemeanors. We know that the 1980s is famous for harsh sentencing laws (Eissen and Chettiar, 2016); however, we explore this variable to see the distribution and relationship with crime rate.

2.3.1.1 Univariate Analysis

We first begin by exploring the distribution and noting if we need to do any transformations:

```
par(mfrow = c(1,2))

hist(c$avgsen,
    breaks = 30,
    main = "Histogram of \nAverage Sentence",
    xlab = NULL)

hist(log(c$avgsen),
    breaks = 30,
    main = "Histogram of \nLog Average Sentence",
    xlab = NULL)
```

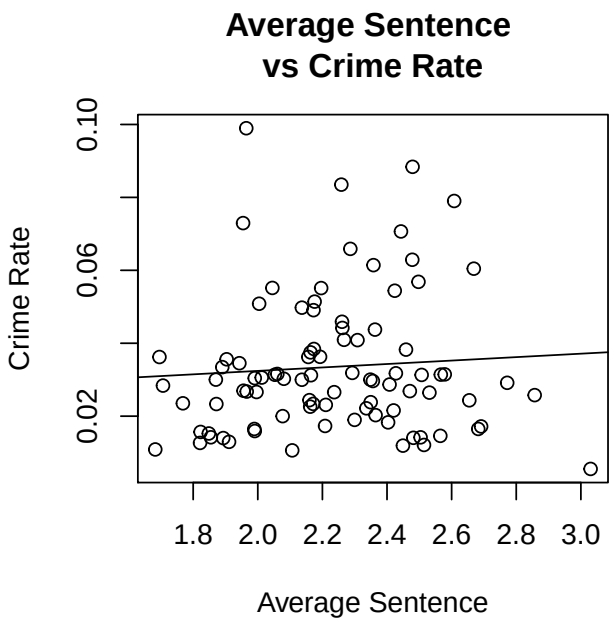


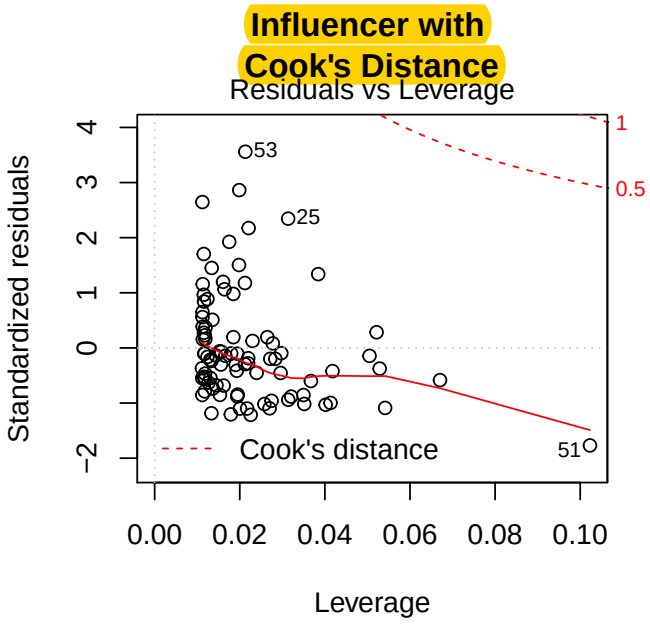
The original distribution is positively skewed; however, taking the log transformation does help normalize the distribution. We also notice an outlier (the maximum of 21 days) and will further examine its influence on our outcome.

2.3.1.1 Bivariate Analysis

We now examine the relationship between average sentence and crime rate, as well as the Cook's Distance to see if our outlier has a lot of influence:

```
par(mfrow = c(1,2))
cr_avs = lm(crmrte ~ log(avgsen), data = c)
plot(log(c$avgsen), c$crmrte,
     main = "Average Sentence \nvs Crime Rate",
     xlab = "Average Sentence", ylab="Crime Rate")
abline(cr_avs)
plot(cr_avs, which = 5, main = "Influencer with \nCook's Distance")
```





We see a (slightly) positive relationship between average sentence and crime rate, and also notice that the outlier does not influence our regression

2.3.2 'Probability' of Conviction (prbconv)

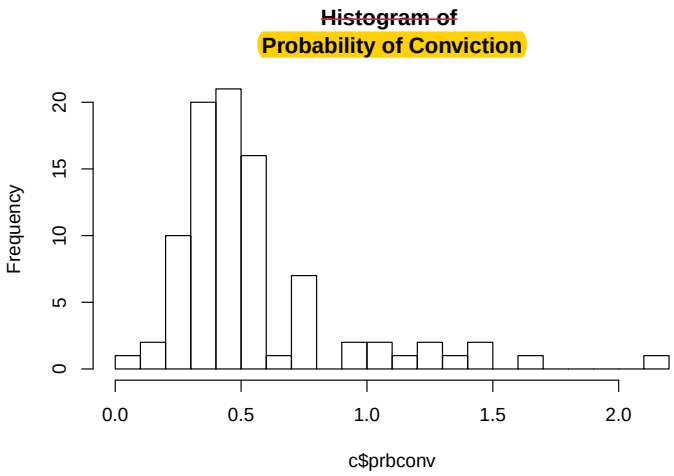
In probability of conviction, where each value is the ratio of convictions to arrests. This describes the second layer of the criminal justice system where defendents are convicted for crimes (assumed for which there is a corresponding arrest).

2.3.2.1 Univariate Analysis

The data is very right-skewed with mode in the range of 0.5 convictions to 1 arrest.

When looking at the range of values, the maximum is 2.12 indicating that there are multiple convictions for a single arrest. It's believed that this is due to the fact that a defendent can be convicted on multiple counts for which they were apprehended in one single arrest.

```
hist(c$prbconv, breaks=20,
     main = "Histogram of \nProbability of Conviction")
```




```
summary(c$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34420 0.45170 0.55090 0.58510 2.12100
```

2.3.2.2 Bivariate Analysis

2.3.2.2.1 Concerns of Multicollinearity with 'Probability' of Arrest & Prison Sentence

Within the dataset, there are three variables that provide insight into the criminal justice system. Those variables are 'Probability' of Arrest (Arrests:Offenses)1, 'Probability' of Conviction (Convictions:Arrests), & 'Probability' of Prison Sentence (Convictions resulting in Prison: Total Convictions)2

Because these seem so closely tied, a check of correlation as a cursory indicator of multicollinearity is safe to see if we could use Conviction or Prison Sentence in addition to Arrests.

1 See section 2.2.1 to read about 'Probability' of Arrest

2 See section 2.4.3 to read about 'Probability' of Prison Sentence

```
paste("Correlation of Probability of Arrest & Probability of Conviction: ",
      cor(c$prbarr, c$prbconv))
```

```
## [1] "Correlation of Probability of Arrest & Probability of Conviction:  -0.0557962059194347"
```

```
paste("Correlation of Probability of Arrest & Probability of Prison Sentence: ",
      cor(c$prbarr, c$prbpris))
```

```
## [1] "Correlation of Probability of Arrest & Probability of Prison Sentence:  0.045833244681947"
```

```
paste("Correlation of Probability of Prison Sentence & Probability of Conviction: ",
      cor(c$prbpris, c$prbconv))
```

```
## [1] "Correlation of Probability of Prison Sentence & Probability of Conviction:  0.0110226453110083"
```

Because the correlations between each criminal justic variable is very small, it is safe to say these are not multicollinear hence can be considered independent.

2.3.3 Wage

Since there are many wage variables, we have grouped weekly wage categories into blue collar & white collar:

Job Name	Category
Finance, Insurance, Real Estate (`wfir`)	White Collar
Service Industry (`wser`)	White Collar

Job Name	Category
Federal Employees (`wfed`)	White Collar
State Employees (`wsta`)	White Collar
Local Government Employees (`wloc`)	White Collar
Construction (`wcon`)	Blue Collar
Transportation, Utilities, Communication (`wtuc`)	Blue Collar
Wholesale, Retail Trade (`wtrd`)	Blue Collar
Manufacturing (`wmfg`)	Blue Collar

The rationale behind this is that we could suggest labor policies for a wider spectrum of occupations instead of at a more granular level.

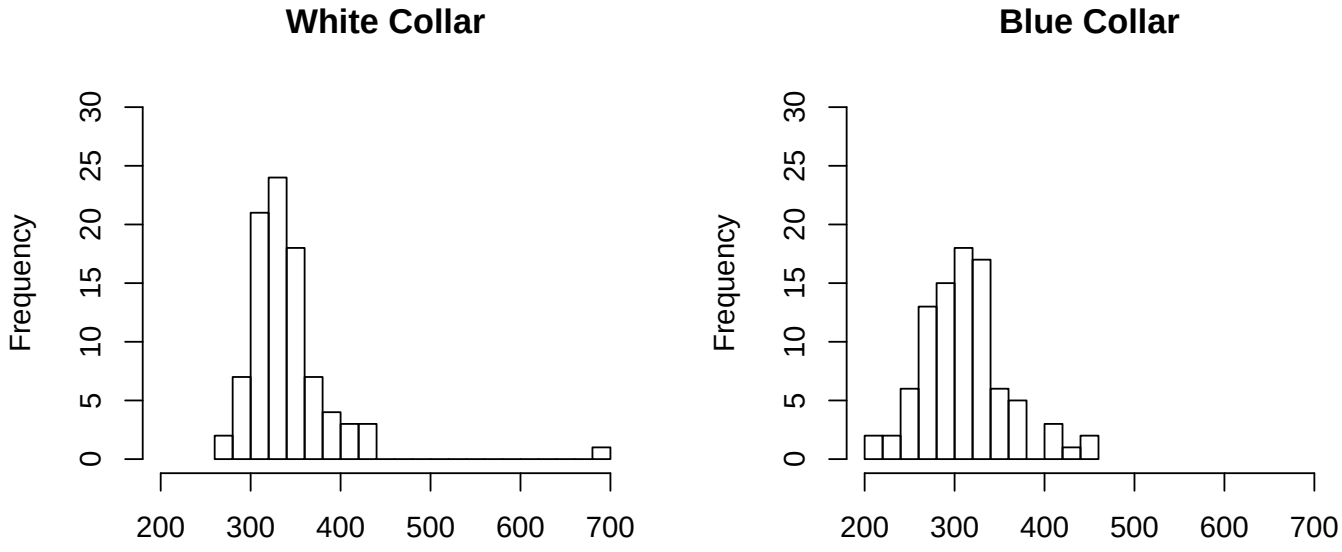
2.3.3.1 Univariate Analysis

We understand that our groupings can skew the wage averages for each category (White Collar and Blue Collar) if one of the wage categories is much higher or lower than the others, an example being if Federal Employees have a much higher average wage it will make the entire White Collar category skew higher than Blue Collar.

```
par(mfrow = c(1,2))
c['white_collar'] = (c$wfir+c$wser+c$wfed+c$wsta+c$wloc)/5
c['blue_collar'] = (c$wcon+c$wtuc+c$wtrd+c$wmfg)/4

wage_category = c('white_collar','blue_collar')
median_wage = c(sprintf("$%.2f", median(c$white_collar)),
                sprintf("$%.2f", median(c$blue_collar)))

median_table <- data.frame(wage_category, median_wage)
median_table
```

```
##   wage_category median_wage
## 1  white_collar     $333.43
## 2   blue_collar     $309.91
```

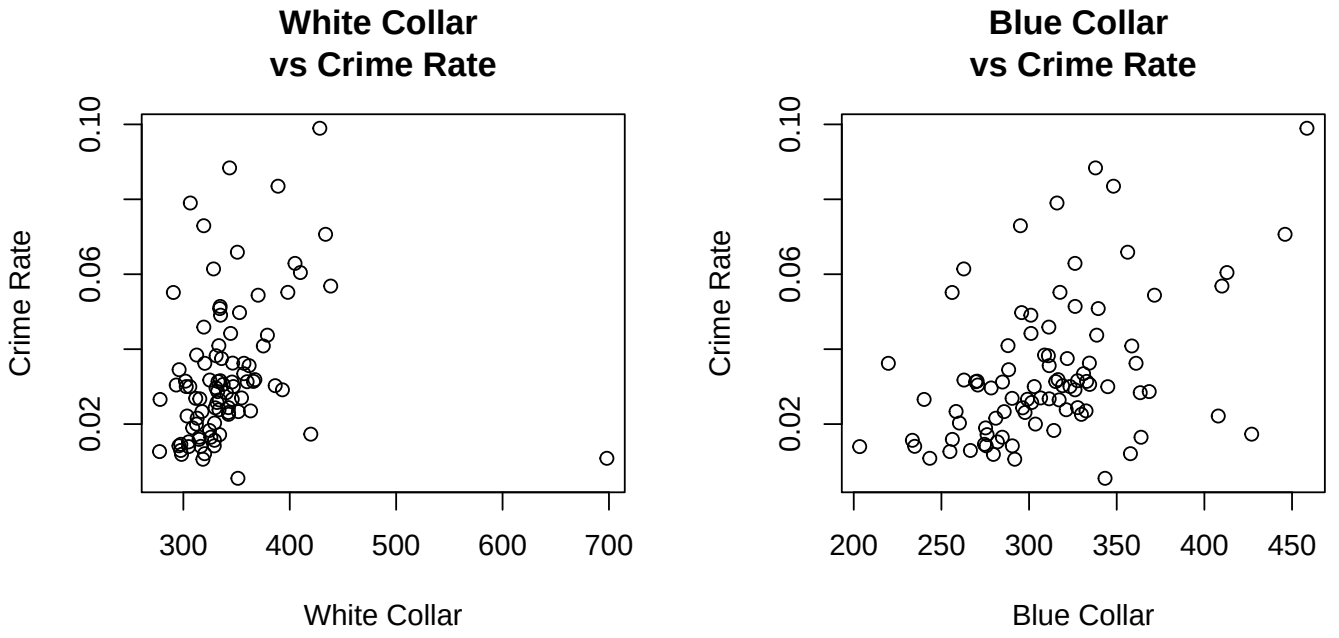
```
hist(c$white_collar, breaks=18,
    main="White Collar", xlab=NA,
    xlim=c(200,700),ylim=c(0,30))
hist(c$blue_collar, breaks=18,
    main="Blue Collar", xlab=NA,
    xlim=c(200,700),ylim=c(0,30))
```



2.3.3.2 Bivariate Analysis

Similar to the `taxpc` category, our initial thought was that a higher wage would indicate a more productive and wealthy county and be associated with lower crime rates. What we found, however, is that there is generally a positive correlation between wage and crime rate. We think that this positive correlation may be because counties with higher average incomes have income disparity.

```
par(mfrow = c(1,2))

plot(c$white_collar, c$crmrte,
     main="White Collar \nvs Crime Rate",
     xlab="White Collar",ylab="Crime Rate")
plot(c$blue_collar, c$crmrte,
     main="Blue Collar \nvs Crime Rate",
     xlab="Blue Collar",ylab="Crime Rate")
```

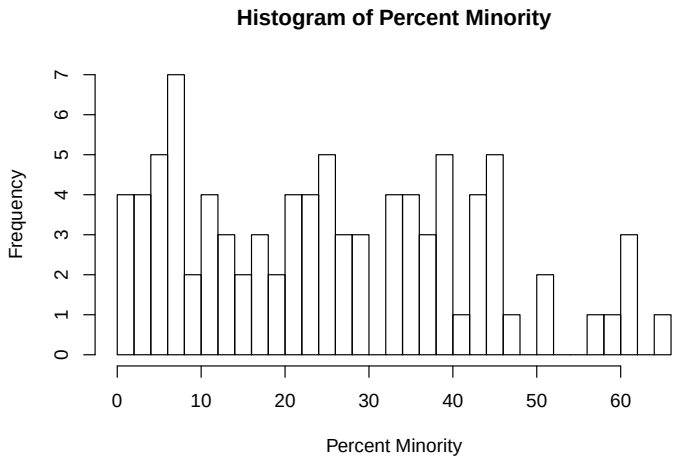


2.2.4 Percent Minority (`pctmin80`)

2.3.4.1 Univariate Analysis

Percent Minority is the percent of total population that did not report their ethnicity and race as something other than nonHispanic White alone in the 1980 census. We acknowledge that grouping all non 100% European white ethnicities into one category as a "minority" seems archaic when the USA is composed of 38% "minority" today, but we will still use this variable as it is what has been given from the 1980 census.

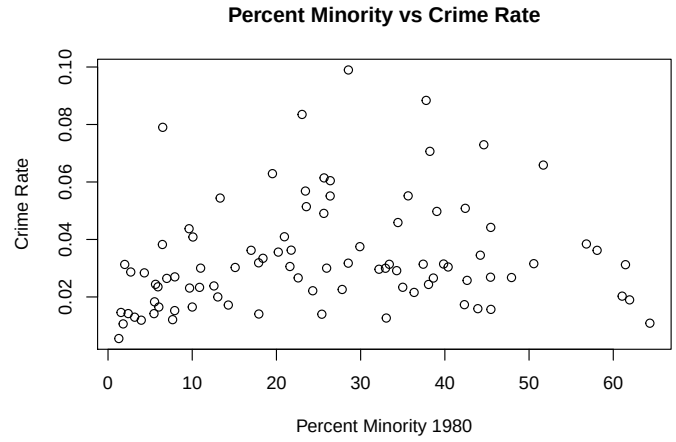
```
hist(c$pctmin80, breaks=30,
     main = "Histogram of Percent Minority",
     xlab = 'Percent Minority')
```



2.3.4.2 Bivariate Analysis

Our hypothesis is that a higher percent minority in a county will be associated with a higher average crime rate. We believe this because black and hispanic minorities, which make up a large portion of the "Percent Minority" category, have lower income averages and lower high school and college completion rates than white Americans. We think that this may cause counties with a higher percent of minorities to have higher crime rates. We found that the correlation with crime rate was .18. We think the impact may be diminished due to the clumping of all minorities into on category, including blacks, hispanics, and asians.

```
plot(c$pctmin80, c$crmrte,
    main='Percent Minority vs Crime Rate', xlab='Percent Minority 1980', ylab='Crime Rate')
```



2.4 Other Variables

This section contains variables in our dataset that will not be going into the models. Any graphical representation for the variables is presented in the Appendix.

2.4.1 County

The county variable is a sequence of odd numbers corresponding to the FIPS code. FIPS code values go from 1 to 193 for the state of North Carolina.

There are a total of 100 counties in North Carolina (and were such in the 1980s as well). In the data, there are 91 rows of data with complete data with the following counties missing from this dataset:

County Number	County Name
29	Camden County

County Number	County Name
31	Carteret County
43	Clay County
73	Gates County
75	Graham County
95	Hyde County
103	Jones County
121	Mitchell County
177	Tyrrell County

2.4.2 Year

All data within the provided set is from 1987 with the exception of the county's minority mix taken from the 1980 census. Having one year of data is nice as a control for exogenous factors not taken into account in our data (e.g. stock market crash in October 1987).

2.4.3 'Probability' of Prison Sentence (`prbpris`)

The final metric of criminal justice system is the ratio of convictions resulting in prison time to total convictions.

2.4.3.1 Univariate Analysis

Here the data is fairly normally-distributed with a maximum value of 0.6 and mean around 0.4.

All visuals & a further breakdown pertaining to this variable can be found in the Appendix - section X.

2.4.3.1 Bivariate Analysis

When run against the crime rate, this data has no clear pattern & seems to be little correlation. When checked, the correlation between the two is 0.047.

3. Modeling

Below we examine four models:

- Key Effects Model:

$$\begin{aligned} crmrte = \beta_0 + \beta_1 * urban + \beta_2 * \sqrt{density} + \beta_3 * \sqrt{pctymle} \\ + \beta_4 * \log(taxpc) + \beta_5 * \log(polpc) + \beta_6 * \log(prbarr) \end{aligned}$$

- Key Covariate Model:

$$\begin{aligned} crmrte = \beta_0 + \beta_1 * urban + \beta_2 * \sqrt{density} + \beta_3 * \sqrt{pctymle} \\ + \beta_4 * \log(taxpc) + \beta_5 * \log(polpc) + \beta_6 * \log(prbarr) + \beta_7 * \log(avgsen) \\ + \beta_8 * white\_collar + \beta_8 * blue\_collar \end{aligned}$$

- Additional Covariate Model:

$$\begin{aligned} crmrte = \beta_0 + \beta_1 * urban + \beta_2 * \sqrt{density} + \beta_3 * \sqrt{pctymle} \\ + \beta_4 * \log(taxpc) + \beta_5 * \log(polpc) + \beta_6 * \log(prbarr) + \beta_6 * \log(prbarr) + \beta_7 * \log(avgsen) \\ + \beta_8 * white\_collar + \beta_8 * blue\_collar + \beta_9 * pctmin80 + \beta_{10} * prbconv \end{aligned}$$

- Logged Key Effects Model:

$$\log(crmrte) = \beta_0 + \beta_1 * urban + \beta_2 * \sqrt{density} + \beta_3 * \sqrt{pctymle} + \beta_4 * \log(taxpc) + \beta_5 * \log(polpc) + \beta_6 * \log(prbarr)$$

We use the Akaike Information Criterion (AIC) as a judge for how well our model performs.

3.1 Key Effects Model

The key effects model explores our original hypothesis that the following variables will have a strong correlational relationship with our outcome when put in a mulitvariate ordinary least squares model.

3.1.1 Model Equation in R

```
(model1 = lm(crmrte ~ urban + sqrt(density) + log(pctymle)
             + log(taxpc) + log(polpc) + log(prbarr), data = c))
```

```
##
## Call:
## lm(formula = crmrte ~ urban + sqrt(density) + log(pctymle) +
##     log(taxpc) + log(polpc) + log(prbarr), data = c)
##
## Coefficients:
##   (Intercept)          urban  sqrt(density)   log(pctymle)     log(taxpc)
##      0.034067       0.004623       0.018694       0.015596       0.013725
##    log(polpc)    log(prbarr)
##      0.006333      -0.007012
```

3.1.2 Model Evaluation

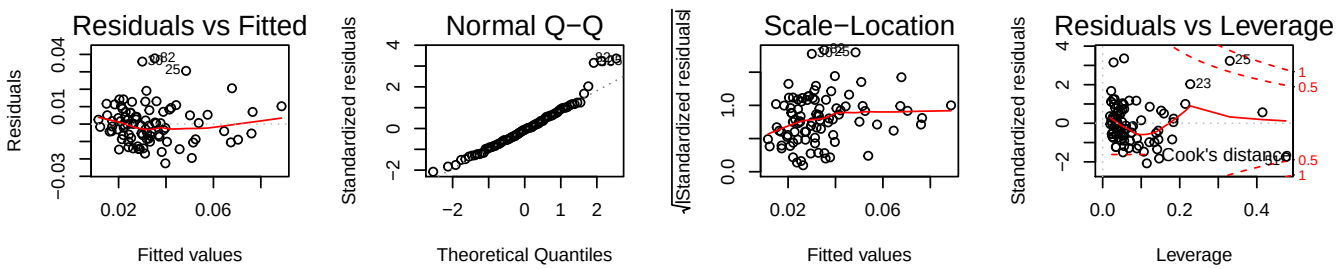
Looking at our model output, we see that the features that have the highest coefficients are:

- `density`: All things held equal, a square-root increase in density approximates to a 0.018 increase in crime rate.
- `pctymle`: All things held equal, a percent increase (in this case, percent) increase in pctmyle approximates to a 0.016 increase in crime rate.
- `taxpc`: All things held equal, a percent ~~increase (in this case, revenue)~~ increase in tax revenue per capita approximates to a 0.014 increase in crime rate.

This output is also important because it tells us that our original hypothesis for key indicator variables may be wrong. The p-values (see appendix for model summary) also validate that apart from density, percent of young males, and tax per capita, we cannot reject the null hypothesis for any other variables.

With that being said, we can still draft policies that will address the features with the high coefficients. This will be further discussed in the Model Comparisson and Key Takeaways section and the Conclusion.

```
par(mfrow = c(1,4))
plot(model1)
```



We look at the MLS assumptions:

1. Linear in parameters:
   - y is linear in the βs.
2. Random sampling:
   - We assume that the dataset given us has independently and identically distributed data.
   - We did not see any clustering or autocorrelation in our EDA.
   - Since this is panel data, each observation is independent of any previous observation.
3. No multicollinearity:

```
mult_col = vif(model1)
mult_col[mult_col >= 4]
```

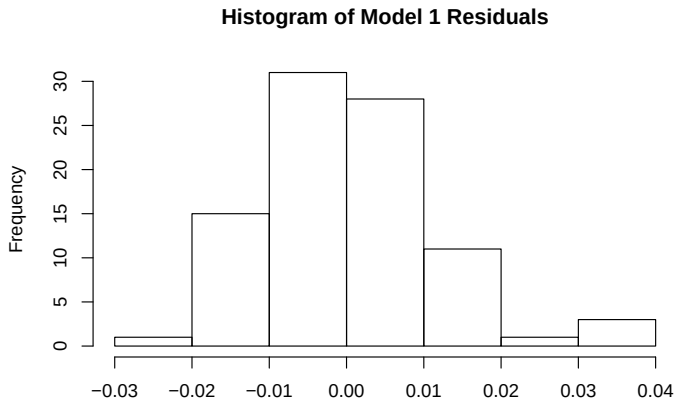
```
- The above VIF test has no values greater than 4 and our OLS model did not drop any variables.
```

4. Zero Conditional Mean/Exogeneity
   - Observe that E(U|X) = E(X|U)E(U)* (1/E(X))
   - We show below that E(U) = 0, thus E(U|X) = 0

   `mean(model1$residuals)`
   - This equals 2.89355866259261e-19 which is practically zero.
5. Homoskedacity:
   - Obsercing the Residuals vs Fitted graph above, we do see some outliers; however it is mainly banded around zero which is good.
   - Running a bptest gives us a p-value of 0.03057

   `bptest(model1)`
   - While low, it still isn't sufficient enough to say that this model fails in rejecting the null hypothesis of homoskedacity.
6. Normality of Residuals: If we look at the QQ plot we notice that the majority of our residuals are on the diagonal, with the tail ends slighly skewing outward. We also plot a histogram of the residuals below and see some normality, thus we can say that the normality of residuals is also satisfied.

```
hist(model1$residuals, main="Histogram of Model 1 Residuals", xlab= NULL)
```



We score the model with the AIC value:

```
AIC(model1)
```

```
## [1] -539.0493
```

3.2 Key Covariate Model

This model involves our key feature model including `avgsen`, `white_collar`, and `blue_collar` to our original model as key covariates.

3.2.1 Model Equation in R

```
(model2 = lm(crmrte ~ urban + sqrt(density) + log(pctymle) + log(taxpc)
             + log(polpc) + log(prbarr) + log(avgsen) + white_collar + blue_collar,
             data = c))
```

```
##
## Call:
## lm(formula = crmrte ~ urban + sqrt(density) + log(pctymle) +
##     log(taxpc) + log(polpc) + log(prbarr) + log(avgsen) + white_collar +
##     blue_collar, data = c)
##
## Coefficients:
##   (Intercept)          urban  sqrt(density)   log(pctymle)     log(taxpc)
##     9.233e-02      6.486e-03      1.978e-02      1.393e-02      1.257e-02
##    log(polpc)    log(prbarr)    log(avgsen)   white_collar    blue_collar
##     1.004e-02     -8.563e-03     -8.297e-03     -5.169e-05     -4.760e-06
```

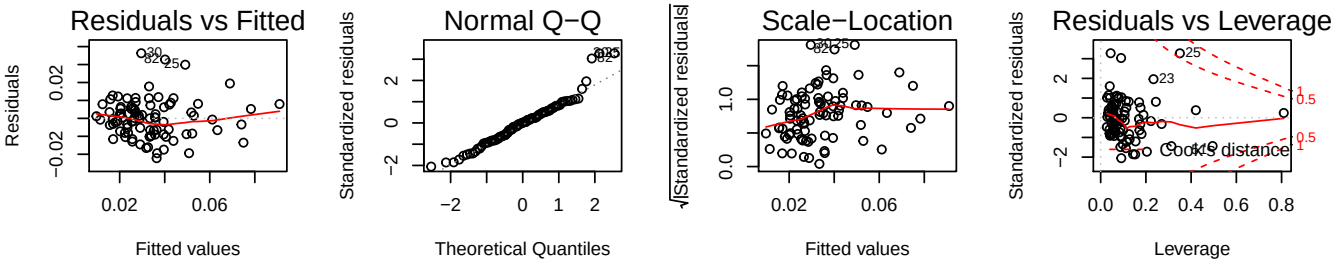
3.2.2 Model Evaluation

Looking at our model output, we see that the features that the stand-out features are:

- `density`: Similar interpretation to our key feature model; however, the approximation of density has a slightly higher coefficient.
- `pctymle`: Similar interpretation to our key feature model; however, the approximation of percent young male has a slightly lower coefficient.
- `taxpc`: Tax percentage has a very similar interpretation as our key feature model.
- `polpc`: Unlike the key feature model, the coefficient of police per capita increased due to our covariates. This can be interpreted as with all things held equal, a unit increase in police per capita (more officers per prson), results in a percentage * 0.01 unit increase in crime rates.

The interesting variable here is `polpc`. Police per Capita increased due to the added covariates, which means that a correlational relationship between police per capita, average sentence, and wages should be looked at further.

```
par(mfrow = c(1,4))
plot(model2)
```



We look at the MLS assumptions:

1. Linear in parameters:
   - Unchanged from Model 1
2. Random sampling:
   - Unchanged from Model 1
3. No multicollinearity:

```
mult_col = vif(model2)
mult_col[mult_col >= 4]
```

```
- The above VIF test has no values greater than 4 and our OLS model did not drop any variables.
```

4. Zero Conditional Mean/Exogeneity
   - Observe that E(U|X) = E(X|U)E(U)* (1/E(X))
   - We show below that E(U) = 0, thus E(U|X) = 0

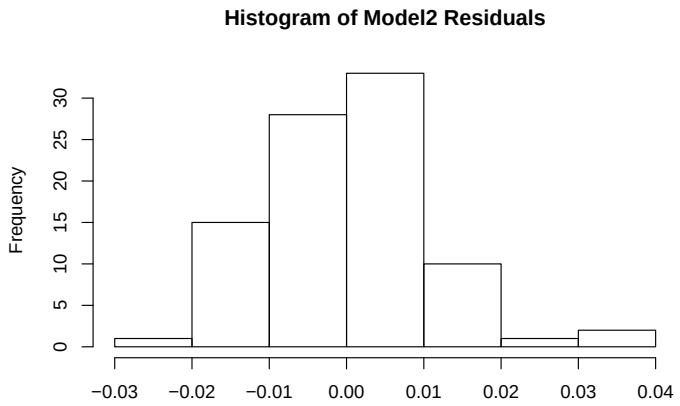
`mean(model2$residuals)`

- This equals 4.90234435439214e-19 which is practically zero.

5. Homoskedacity:
   - Obsercing the Rediduals vs Fitted graph above, we do see some outliers; however it is mainly banded around zero which is good.
   - Running a bptest gives us a p-value of 0.02802

   `bptest(model2)`
   - While lower, it still isn't sufficient enough to say that this model fails in rejecting the null hypothesis of homoskedacity.
6. Normality of Residuals: If we look at the QQ plot we notice that the majority of our residuals are on the diagonal, with the tail ends slighly skewing outward. We also plot a histogram of the residuals below and see some normality, thus we can say that the normality of residuals is also satisfied.

```
hist(model2$residuals, main="Histogram of Model2 Residuals", xlab= NULL)
```



We score the model with the AIC value:

```
AIC(model2)
```

```
## [1] -539.5856
```

The AIC of model2 is slightly lower which means that our model did improve slightly with our new added variables.

3.3 Additional Covariate Model

This model involves our key covariate model including `pctmin80` and `prbconv` to our original model as additional covariates.

3.3.1 Model Equation in R

```
(model3 = lm(crmrte ~ urban + sqrt(density) + log(pctymle) + log(taxpc)
             + log(polpc) + log(prbarr) + log(avgsen) + white_collar
             + blue_collar + pctmin80 + prbconv,
             data = c))
```

```
##
## Call:
## lm(formula = crmrte ~ urban + sqrt(density) + log(pctymle) +
##     log(taxpc) + log(polpc) + log(prbarr) + log(avgsen) + white_collar +
##     blue_collar + pctmin80 + prbconv, data = c)
##
## Coefficients:
##   (Intercept)          urban  sqrt(density)   log(pctymle)     log(taxpc)
##     9.248e-02      1.193e-03      1.714e-02      3.958e-03      5.985e-03
##    log(polpc)    log(prbarr)    log(avgsen)   white_collar    blue_collar
```

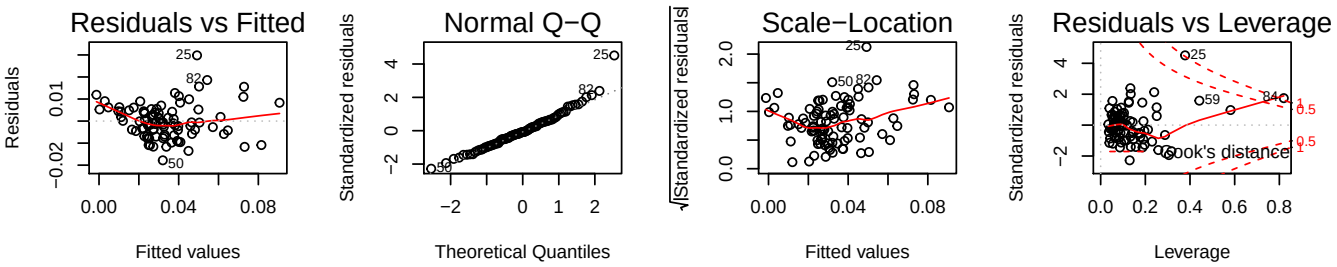
```
##     1.506e-02     -1.606e-02     -3.246e-03     -3.509e-05      1.657e-05
##      pctmin80        prbconv
##     3.854e-04     -1.616e-02
```

3.3.2 Model Evaluation

Looking at our model output, we see that the additional covariates resulted in a few changes to our key variables:

- `density`: The coefficient of density decreased slightly.
- `pctymle`: The coeffcient of percent young male decreased by a substantial amount. One reason for this is that we reduced the ommited variable bias in this variable by adding our new features. Perhaps percent young male was also taking into account the minority percentage or probability of conviction.
- `taxpc`: The coeffcient of tax reveneue per capita decreased by a substantial amount. As above, tax revenue per capita could also have had ommited variable bias reduced by our new variables.
- `polpc`: The coefficient of police per capita increased by a small margin.

```
par(mfrow = c(1,4))
plot(model3)
```



We look at the MLS assumptions:

1. Linear in parameters:
   - Unchanged from Model 1
2. Random sampling:
   - Unchanged from Model 1
3. No multicollinearity:

```
mult_col = vif(model2)
mult_col[mult_col >= 4]
```

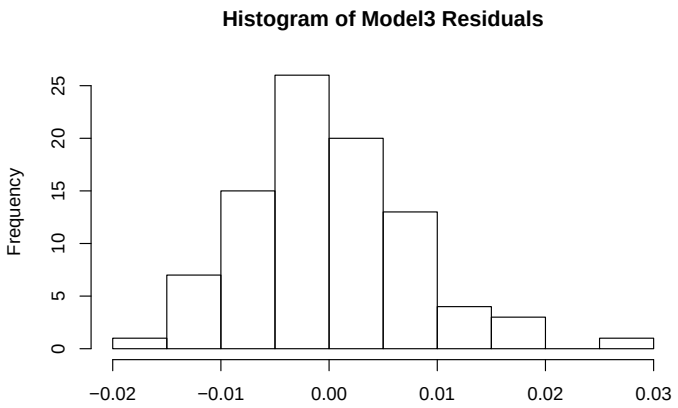
```
- The above VIF test has no values greater than 4 and our OLS model did not drop any variables.
```

4. Zero Conditional Mean/Exogeneity
   - Observe that E(U|X) = E(X|U)E(U)* (1/E(X))
   - We show below that E(U) = 0, thus E(U|X) = 0

   `mean(model3$residuals)`
   - This equals 6.33063013300672e-20 which is practically zero.
5. Homoskedacity:
   - Obsercing the Rediduals vs Fitted graph above, we do see some outliers; however it is mainly banded around zero which is good.
   - Running a bptest gives us a p-value of 0.0007179

   `bptest(model3)`
   - Model 3 is the lowest p-value which rejects the null hypothesis of homoskedacity and thus could suffer from heteroskedacity.
6. Normality of Residuals: If we look at the QQ plot we notice that the majority of our residuals are on the diagonal, with the tail ends slighly skewing outward. We also plot a histogram of the residuals below and see some normality, thus we can say that the normality of residuals is also satisfied.

```
hist(model3$residuals,
     main="Histogram of Model3 Residuals", xlab= NULL)
```



We score the model with the AIC value:

```
AIC(model3)
```

```
## [1] -592.3165
```

The AIC of this model is signficantly lower than our previous models, thus this model performs the best.

3.4 Log-Log Model

This model is a play on our key features model; however, we decided to log our dependent variable in order to get percent level changes on both sides and to make our equation more linear.

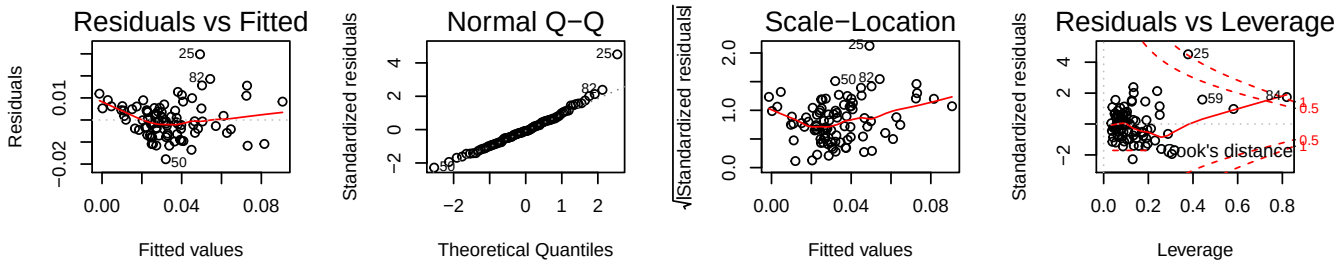
3.4.1 Model Equation in R

```
(model4 = lm(log(crmrte) ~ log(1+urban) + log(density) + log(pctymle)
             + log(taxpc) + log(polpc) + log(prbarr),
             data = c))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(1 + urban) + log(density) + log(pctymle) +
##     log(taxpc) + log(polpc) + log(prbarr), data = c)
##
## Coefficients:
##    (Intercept)  log(1 + urban)    log(density)    log(pctymle)
##        -2.7054          0.5424          0.1144          0.4347
##     log(taxpc)      log(polpc)     log(prbarr)
##         0.3328          0.2140         -0.3144
```

3.4.2 Model Evaluation

```
par(mfrow = c(1,4))
plot(model3)
```



We see from above that our residuals are not quite normal; however, they follow a more normal distribution that Model 3.1 (Key features). The regular vs fitted does show some curvature (and thus some bias); however, we feel that with a dataset that have enough observations (at least 30), we should expect some bias and that there is no way to control it all.

This model can be interpreted as a reflection of percent changes on the dependent variable. Thus, this model is measuring percentile changes on both sides of the equation.

```
AIC(model4)
```

```
## [1] 102.9209
```

The above AIC score gives us an indicator that while our logged OLS model may satisfy the requirements a bit more stringently than our key effects model, the overall model score is lower.

3.5 Model Comparisson and Key Takeaways

We compare our models below:

```
stargazer(model1, model2, model3, model4, header = FALSE, omit.stat = "f",
          star.cutoffs = c(0.05, 0.01, 0.001))
```

We see above that Model 3 has has more variables that are statistically significant features. We do see some interesting observations in our variables. Density does not seem to fluctuate too much when we add more features, thus perhaps some other ommited variable is influencing it's coefficient. Alternatively, percent of young males, tax revenue per capita, police per capita, and probabiliy of arrest all seemed to fluctuate drastically once our other variables were added in; however, significant fluctuations happend with the variables that we did not believe to be key covariates. Thus, in future studies, we should re-examine these variables to understand the effects that they have on our other variables.

While the log-log did give us better coefficient results, the results from our EDA did not immediately indicate that we needed to log crime rate, thus this exercise was an experiment in curiousity.

3.6 Ommited Variable Bias

Here we explore variables that are not in the given dataset; but that we believe would affect our key features:

3.6.1 Gun Sales

We believe that gun sales could be causing some ommited variable bias, specifically for police per capita and probability of arrest. We believe that gun sales would have a positive relationship with crime rate, police per capita, and probability of arrest. Thus, the coefficients for these two variables are higher than they should be if we were to take gun sales into account.

3.6.2 Unemployment Rates

We believe that unemployment rates could also be causing some ommited variable bias. The reason for this is that we believe that people might get desperate and resort to drastic means in order to improve their living situation. This variable could have a positive relationship with crime rate and a negative relationship with tax revenue per capita. Thus by adding this variable, we could see the coefficient of tax revenue per capita increase.

3.6.3 Years of Education

Years of education is important because we believe it affects percent of young males and percent minorites. We would expect that years of education would have a negative relationship with crime rate and a positive relationship with percent of young males and percent minorities. The rationale behind this is that usually parents push their children to do better in school than they did, and also minorities may see education as a valuable tool to improve

Table 3:

	Dependent variable:			
	crmrte			log(crmrte)
	(1)	(2)	(3)	(4)
urban	0.005	0.006	0.001	
	(0.007)	(0.007)	(0.005)	
sqrt(density)	0.019***	0.020***	0.017***	
	(0.004)	(0.004)	(0.003)	
log(1 + urban)				0.542*
				(0.264)
log(density)				0.114**
				(0.036)
log(pctymle)	0.016*	0.014*	0.004	0.435
	(0.007)	(0.007)	(0.005)	(0.239)
log(taxpc)	0.014*	0.013*	0.006	0.333
	(0.005)	(0.005)	(0.004)	(0.192)
log(polpc)	0.006	0.010*	0.015***	0.214
	(0.004)	(0.004)	(0.003)	(0.133)
log(prbarr)	−0.007*	−0.009*	−0.016***	−0.314*
	(0.004)	(0.004)	(0.003)	(0.122)
log(avgsen)		−0.008	−0.003	
		(0.005)	(0.004)	
white_collar		−0.0001	−0.00004	
		(0.00003)	(0.00002)	
blue_collar		−0.00000	0.00002	
		(0.00003)	(0.00003)	
pctmin80			0.0004***	
			(0.0001)	
prbconv			−0.016***	
			(0.003)	
Constant	0.034	0.092*	0.092**	−2.705*
	(0.036)	(0.044)	(0.033)	(1.241)
Observations	90	90	90	90
R^2	0.652	0.676	0.828	0.483
Adjusted R^2	0.627	0.640	0.803	0.446
Residual Std. Error	0.012 (df = 83)	0.011 (df = 80)	0.008 (df = 78)	0.408 (df = 83)
Note:				*p<0.05; **p<0.01; ***p<0.001

wealth. Thus, by adding in years of education, we would expect to see the coefficients for these two variables go down.

3.6.4 Written Arrests (Citations)

The data provided by the Federal Bureau of Investigation (FBI) Uniform Crime Reporting service only takes into account "arrests" and not "citations" - written arrests given out for low-level crimes that meet certain criteria and is at somewhat of the officer's discretion. This data would be helpful in seeing the full picture of crimes in each county.

3.6.5 Rehabilitation

When the right mentoring programs are made available to the criminals, it provides them an opportunity to correct their behavior. By providing targeted rehabilitation, misguided young males and minorities can avail better opportunities and therefore reduce crime rate. So, the availablity of rehabilitation centers can be a omitted bias.

3.6.6 US Economy

The Dow Jones Index has been an indicator of the US Economy and would be an interesting variable to compare to crime rate over a number of years. The significant increases and decreases of the stock market can be a major driver of wages and unemployment, and could therefore have a potential causal affect on crime rate

3.6.7 Gangs

The percent of young males statistic is interesting because it could point to a few underlying reasons for a concentration of men between the ages of 15 & 24. One theory is gangs. If gang density or gang membership count data were available, it would help not only with the young male statistic but also with seeing amount of organized crime in an area. We theorize that men in that age group are most prone to be active in gangs. This would help in differentiating between young men in gangs & young men in, say, a university.

3.6.8 Age Demographics

The age demographics would give us a more comprehensive look at what the percent young male category is giving us right now. This would allow us to explore how having more of certain age groups, male or female, would effect crime rate.

3.6.9 Crime Categories

The omitted variable of the categories of crime might influence the crime rate. Though the 'mix' variable gives some information on the offense mix which is not very clear to us and so we have not used it in our models.

4. Conclusion

We begin by reitterating the famous saying, "correlation does not imply causation." By no means would we argue that our models can be used for a causal analysis; however, we believe that the correlational relationship can be utilized for new policies.

If we were to take the results of our key effects model, we would argue that higher density, tax revenue per capita, and percent of young males approximate to higher crime. Therefore, we would first try to build policies around these variables.

Density is a tricky topic because controlling the flow of people to a particular city is very difficult to do; however, we can manipulate the flow of people to other cities. If we propose incentives to entrepreneurs across various cities, we could mitigate migration from various cities and encourage people to stay in their home town, thus potentially lowering densely populated areas.

Tax revenue per capita would also be an interesting problem to solve. We approximated that higher tax revenue is associated with higher crime, thus we could play around with tax brackets and lower the rates for people who are struggling to pay their taxes. This would be the most difficult policy to put forward because it would be met with opposition on both a local and national scale; however, it is worth exploring.

Higher percent of young males also was approximated to be associated with higher crime rates. We hypothesized that this could be due to gangs, and we explored the possibility of education in the ommited variable bias section. We believe that if we could implement more after school activities or recreation leagues, we could influence this correlation. This policy would be the easiest to push forward and we believe it will gain unanimous support.

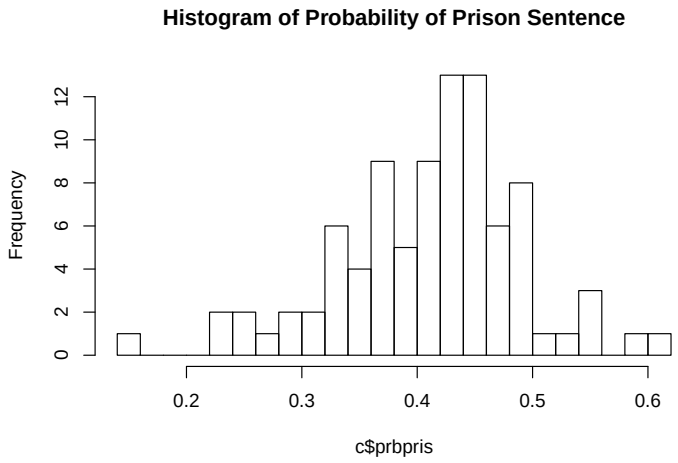
We are choosing to focus on three areas of various difficulty; however, we do indeed believe that having more data will allow us to make better approximations and thus better policies.

Appendix

Other Key Variable Analysis

A.1.1 Probability of Prison Sentence

```
hist(c$prbpris, breaks=20, main = "Histogram of Probability of Prison Sentence")
```



Correlation Table

```
round(cor(c[-(1:2)],c[-(1:2)]), 2)
```

```
##              crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## crmrte         1.00  -0.40   -0.39    0.05   0.02  0.17    0.73  0.45
## prbarr        -0.40   1.00   -0.06    0.05   0.18  0.43   -0.30 -0.14
## prbconv       -0.39  -0.06    1.00    0.01   0.16  0.17   -0.23 -0.13
## prbpris        0.05   0.05    0.01    1.00  -0.09  0.05    0.08 -0.09
## avgsen         0.02   0.18    0.16   -0.09   1.00  0.49    0.07  0.09
## polpc          0.17   0.43    0.17    0.05   0.49  1.00    0.16  0.28
## density        0.73  -0.30   -0.23    0.08   0.07  0.16    1.00  0.32
## taxpc          0.45  -0.14   -0.13   -0.09   0.09  0.28    0.32  1.00
## west          -0.35   0.17    0.05   -0.04   0.11  0.15   -0.14 -0.18
## central        0.17  -0.17   -0.05    0.16  -0.16 -0.05    0.36  0.03
## urban          0.62  -0.21   -0.20    0.05   0.14  0.16    0.82  0.35
## pctmin80       0.18   0.05    0.06    0.11  -0.17 -0.17   -0.07 -0.03
## wcon           0.39  -0.25   -0.12   -0.06  -0.03 -0.02    0.45  0.26
## wtuc           0.24  -0.07   -0.01    0.12   0.23  0.17    0.33  0.17
## wtrd           0.43  -0.10   -0.13    0.14   0.11  0.12    0.59  0.18
```

```
## wfir           0.34  -0.17    0.03    0.03   0.18  0.20    0.55  0.13
## wser          -0.05  -0.13    0.46    0.04  -0.15 -0.02    0.04  0.08
## wmfg           0.35  -0.15    0.02    0.01   0.11  0.27    0.44  0.26
## wfed           0.49  -0.21   -0.06    0.08   0.15  0.16    0.59  0.06
## wsta           0.20  -0.16   -0.13   -0.03   0.13  0.05    0.22 -0.03
## wloc           0.36  -0.02    0.05    0.08   0.15  0.39    0.46  0.22
## mix           -0.13   0.41   -0.30    0.12  -0.14  0.02   -0.14 -0.04
## pctymle        0.29  -0.18   -0.16   -0.08   0.07  0.05    0.12 -0.09
## white_collar   0.22  -0.22    0.35    0.06  -0.01  0.12    0.38  0.12
## blue_collar    0.44  -0.18   -0.05    0.07   0.16  0.21    0.57  0.29
##                west central urban pctmin80  wcon  wtuc  wtrd  wfir  wser
## crmrte        -0.35    0.17  0.62     0.18  0.39  0.24  0.43  0.34 -0.05
## prbarr         0.17   -0.17 -0.21     0.05 -0.25 -0.07 -0.10 -0.17 -0.13
## prbconv        0.05   -0.05 -0.20     0.06 -0.12 -0.01 -0.13  0.03  0.46
## prbpris       -0.04    0.16  0.05     0.11 -0.06  0.12  0.14  0.03  0.04
## avgsen         0.11   -0.16  0.14    -0.17 -0.03  0.23  0.11  0.18 -0.15
## polpc          0.15   -0.05  0.16    -0.17 -0.02  0.17  0.12  0.20 -0.02
## density       -0.14    0.36  0.82    -0.07  0.45  0.33  0.59  0.55  0.04
## taxpc         -0.18    0.03  0.35    -0.03  0.26  0.17  0.18  0.13  0.08
## west           1.00   -0.39 -0.09    -0.63 -0.19  0.08 -0.17 -0.04 -0.06
## central       -0.39    1.00  0.16    -0.05  0.40  0.19  0.39  0.29  0.19
## urban         -0.09    0.16  1.00     0.02  0.32  0.23  0.43  0.40  0.06
## pctmin80      -0.63   -0.05  0.02     1.00 -0.11 -0.19 -0.06 -0.08  0.20
## wcon          -0.19    0.40  0.32    -0.11  1.00  0.41  0.56  0.49 -0.01
## wtuc           0.08    0.19  0.23    -0.19  0.41  1.00  0.35  0.33 -0.02
## wtrd          -0.17    0.39  0.43    -0.06  0.56  0.35  1.00  0.67 -0.02
## wfir          -0.04    0.29  0.40    -0.08  0.49  0.33  0.67  1.00  0.01
## wser          -0.06    0.19  0.06     0.20 -0.01 -0.02 -0.02  0.01  1.00
## wmfg           0.00    0.17  0.40    -0.12  0.35  0.47  0.37  0.50  0.01
## wfed          -0.18    0.35  0.43     0.03  0.51  0.40  0.64  0.62  0.02
## wsta          -0.08    0.09  0.30     0.09 -0.02 -0.15  0.01  0.24  0.04
## wloc          -0.12    0.33  0.34    -0.11  0.52  0.33  0.58  0.55  0.08
## mix           -0.01   -0.09 -0.06     0.20 -0.20 -0.25 -0.13 -0.21 -0.17
## pctymle       -0.04   -0.10  0.09    -0.02 -0.02 -0.10 -0.11  0.01 -0.04
## white_collar -0.13    0.35  0.32     0.16  0.27  0.16  0.34  0.47  0.84
## blue_collar  -0.04    0.33  0.44    -0.17  0.69  0.80  0.64  0.61 -0.01
##                wmfg  wfed  wsta  wloc   mix pctymle white_collar
## crmrte         0.35  0.49  0.20  0.36 -0.13    0.29         0.22
## prbarr        -0.15 -0.21 -0.16 -0.02  0.41   -0.18        -0.22
## prbconv        0.02 -0.06 -0.13  0.05 -0.30   -0.16         0.35
## prbpris        0.01  0.08 -0.03  0.08  0.12   -0.08         0.06
## avgsen         0.11  0.15  0.13  0.15 -0.14    0.07        -0.01
## polpc          0.27  0.16  0.05  0.39  0.02    0.05         0.12
## density        0.44  0.59  0.22  0.46 -0.14    0.12         0.38
## taxpc          0.26  0.06 -0.03  0.22 -0.04   -0.09         0.12
## west           0.00 -0.18 -0.08 -0.12 -0.01   -0.04        -0.13
## central        0.17  0.35  0.09  0.33 -0.09   -0.10         0.35
## urban          0.40  0.43  0.30  0.34 -0.06    0.09         0.32
## pctmin80      -0.12  0.03  0.09 -0.11  0.20   -0.02         0.16
## wcon           0.35  0.51 -0.02  0.52 -0.20   -0.02         0.27
## wtuc           0.47  0.40 -0.15  0.33 -0.25   -0.10         0.16
## wtrd           0.37  0.64  0.01  0.58 -0.13   -0.11         0.34
## wfir           0.50  0.62  0.24  0.55 -0.21    0.01         0.47
## wser           0.01  0.02  0.04  0.08 -0.17   -0.04         0.84
## wmfg           1.00  0.52  0.05  0.45 -0.34    0.02         0.29
## wfed           0.52  1.00  0.19  0.52 -0.31   -0.06         0.47
```

```
## wsta          0.05  0.19  1.00  0.16 -0.08    0.22         0.31
## wloc          0.45  0.52  0.16  1.00 -0.25    0.00         0.44
## mix          -0.34 -0.31 -0.08 -0.25  1.00   -0.09        -0.30
## pctymle       0.02 -0.06  0.22  0.00 -0.09    1.00        -0.01
## white_collar  0.29  0.47  0.31  0.44 -0.30   -0.01         1.00
## blue_collar   0.82  0.65 -0.04  0.58 -0.34   -0.06         0.33
##              blue_collar
## crmrte              0.44
## prbarr             -0.18
## prbconv            -0.05
## prbpris             0.07
## avgsen              0.16
## polpc               0.21
## density             0.57
## taxpc               0.29
## west               -0.04
## central             0.33
## urban               0.44
## pctmin80           -0.17
## wcon                0.69
## wtuc                0.80
## wtrd                0.64
## wfir                0.61
## wser               -0.01
## wmfg                0.82
## wfed                0.65
## wsta               -0.04
## wloc                0.58
## mix                -0.34
## pctymle            -0.06
## white_collar        0.33
## blue_collar         1.00
```