Lab 4: Reducing Crime

Kim Vignola, Kiersten Henderson, Aaron Yuen

August 17, 2017

Section 1: Introduction

The authors, Kim, Kiersten and Aaron, were hired to provide research for a political campaign in North Carolina to understand the determinants of crime (both correlational and causal) using exploratory data analysis and OLS regression. The end goal is to leverage the data to provide policy suggestions that are applicable to local government to reduce crime based on the available data we have.

The provided cross-sectional dataset consists of statistics for a selection of counties for a given year. Data for 90 counties and 25 variables for each county were provided. As part of this report, we will also make recommendations on what additional data would help better inform campaign recommendations, as there could be other variables that are currently not captured in the provided dataset.

For this analysis, the following assumptions were made:

- The 90 counties provided were randomly sampled among the 100 counties in North Carolina.

Other assumptions will be called out in the remaining sections as appropriate.

Section 2: Exploratory Analysis

Data Load and Library Imports

Reading the data and loading the right libraries:

```
library(car)
library(corrplot)
library(lmtest)
library(sandwich)
library(stargazer)

data = read.csv("crime_v2.csv")
```

Univariate Variable Analysis

There are 90 data points and 25 variables

```
nrow(data)
```

```
## [1] 90
```

```
colnames(data)
```

```
##  [1] "X"        "county"   "year"     "crime"    "probarr"  "probconv"
##  [7] "probsen"  "avgsen"   "police"   "density"  "tax"      "west"
## [13] "central"  "urban"    "pctmin"   "wagecon"  "wagetuc"  "wagetrd"
## [19] "wagefir"  "wageser"  "wagemfg"  "wagefed"  "wagesta"  "wageloc"
## [25] "mix"      "ymale"
```

There doesn't seem to be any NAs in the dataset.

```
apply(!is.na(data[,]), MARGIN = 2, mean)
```

```
##         X    county      year     crime   probarr probconv   probsen    avgsen
##         1         1         1         1         1         1         1         1
##    police   density       tax      west   central     urban    pctmin   wagecon
##         1         1         1         1         1         1         1         1
##   wagetuc   wagetrd   wagefir   wageser   wagemfg   wagefed   wagesta   wageloc
##         1         1         1         1         1         1         1         1
##       mix     ymale
##         1         1
```

The following summarizes the different variables types based on the variable desciptions and basic understanding of the data:

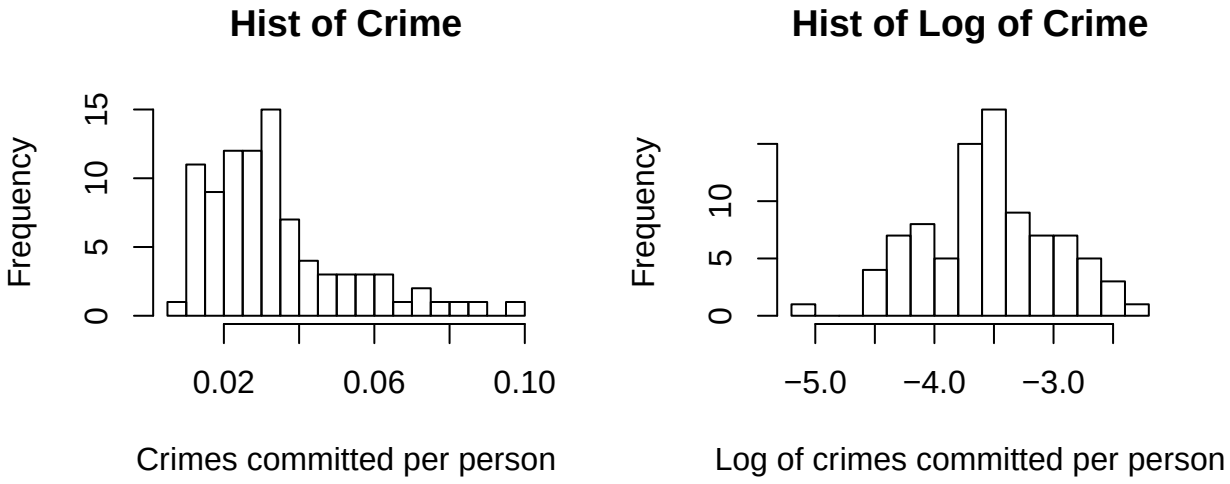
1. Rates, averages, and probabilities - crime, probarr, probconv, probsen, avgsen, police, density, pctmin, mix, ymale
2. $ variables - tax, wagecon, wagetuc, wagetrd, wagefir, wageser, wagemfg, wagefed, wagesta, wageloc
3. Indicator variables - west, central, urban. No base categories are in the dataset (e.g. non-west/central, rural), which is expected, given that adding the base category will lead to perfect collinearity.
4. Other miscellaneous variables - X, county, year

The remaining EDA focuses the analysis on only the key variables of interest.

Crimes Committed per Person (crime)

Crime is the main dependent variable of interest. Looking at the histogram of crime, the distribution tends to be right skewed. Taking the log of crime tends to make the histgram appear more normal. As a result, for our modeling, we will proceed with using log of crime as the dependent variable.

```
par(mfrow=c(1,2))
hist(data$crime, breaks=20,
     main="Hist of Crime",
     xlab="Crimes committed per person", cex=0.7)
hist(log(data$crime), breaks=20,
     main="Hist of Log of Crime",
     xlab="Log of crimes committed per person", cex=0.7)
```



Probability of Conviction (probconv)

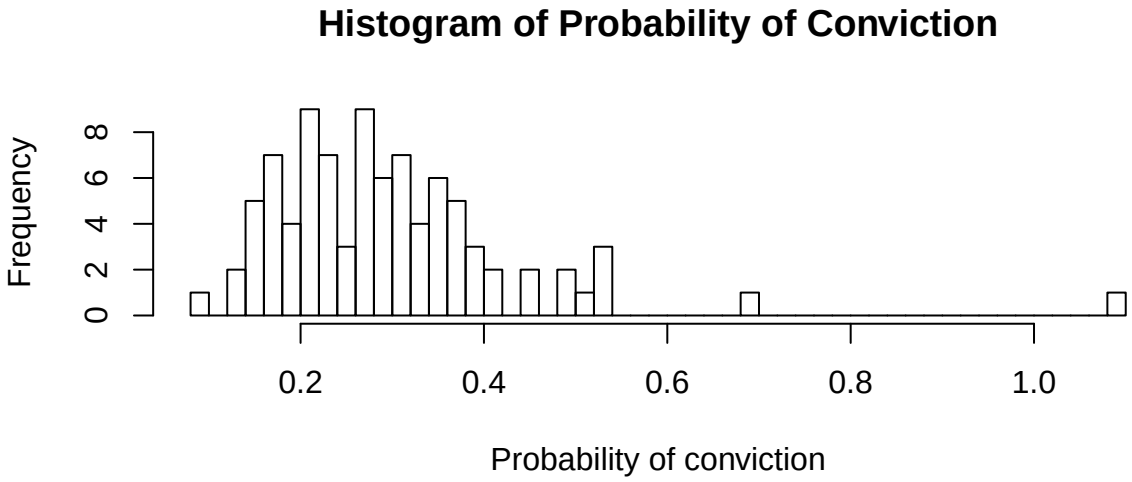
Probconv seems to have values greater than 1.0, which is unexpected given that probability values are supposed to be between 0.0 to 1.0. The variable description does not seem to indicate how the variable is calculated. After discussing this in the lectures and office hours, we will assume that probconv values above 1.0 is okay, and that the higher the value, the higher the probabilty of conviction. Essentially, we will treat this variable as an odds variable.

The histogram of probconv does not seem to look very normal. That said, taking the log of a probabilty does not seem to make sense from an interpretability perspective. For example, an "increase of 10% in probability" is not intuitively interpretable. As a result, for probconv no log transformation will be applied. To improve interpretability, in the transformed dataset we will multiply this variable by 100.

```
summary(data$probconv)
```

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

```
hist(data$probconv, breaks=50,
     main="Histogram of Probability of Conviction",
     xlab="Probability of conviction")
```



Probability of Prison Sentence (probsen)

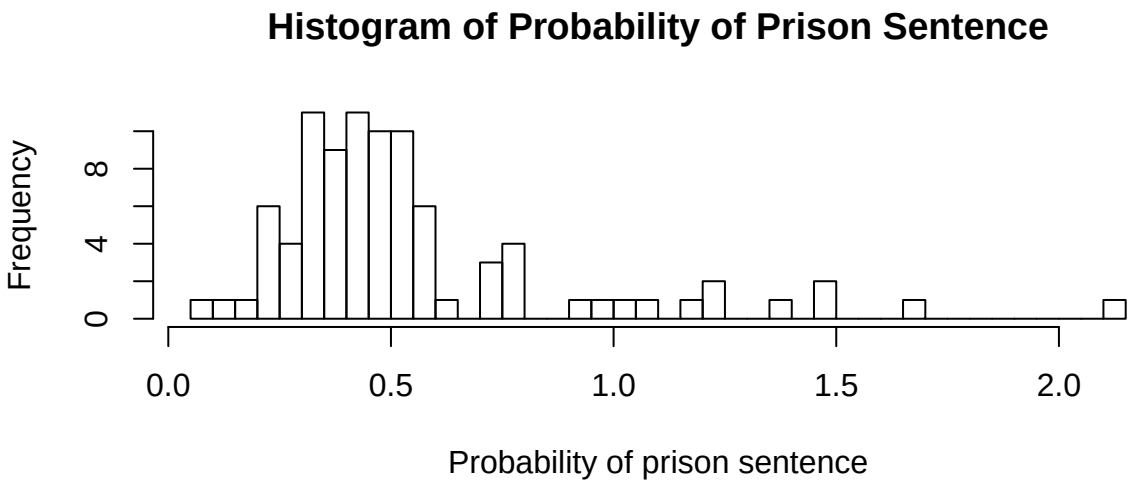
Similarly, probsen seems to have values greater than 1.0, which is unexpected given that probability values are supposed to be between 0.0 to 1.0. After discussing this in the lectures and office hours, we will assume that probsen values above 1.0 is okay, and that the higher the value, the higher the probabilty of prison sentence. Essentially, we will treat this variable as an odds variable.

Also, similar to probconv, for probsen no log transformation will be applied. To improve interpretability, in the transformed dataset we will multiply this variable by 100.

```
summary(data$probsen)
```

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

```
hist(data$probsen, breaks=50,
     main="Histogram of Probability of Prison Sentence",
     xlab="Probability of prison sentence")
```



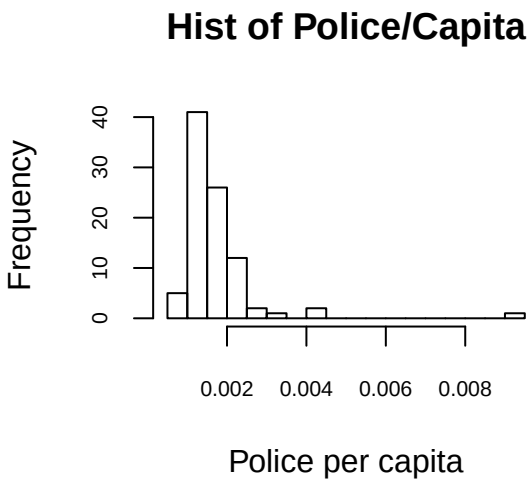
Police per Capita (police)

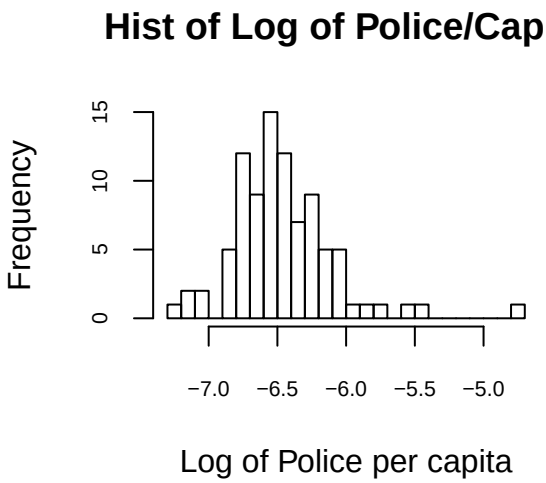
Looking at the histogram of police, the distribution tends to be right skewed. Taking the log of police tends to make the histgram appear more normal. As a result, for our modeling, we will proceed with using log of police. As well, taking the log of police allows us to look at the elasticity between police and crime.

```
summary(data$police)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012380 0.0014900 0.0017080 0.0018860 0.0090540
```

```
par(mfrow=c(1,2))
hist(data$police, breaks=20,
     main="Hist of Police/Capita",
     xlab="Police per capita", cex.axis = 0.7)
hist(log(data$police), breaks=20,
     main="Hist of Log of Police/Capita",
     xlab="Log of Police per capita", cex.axis = 0.7)
```





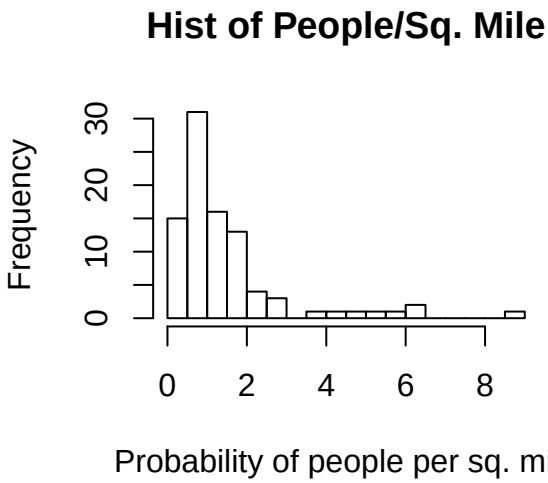
People per Sq. Mile (density)

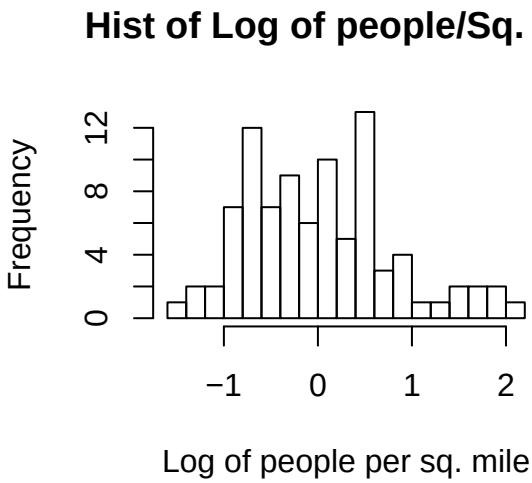
Looking at the histogram of density, the distribution tends to be right skewed. Taking the log of density tends to make the histgram appear more normal. As a result, for our modeling, we will proceed with using log of density.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2034  0.5472  0.9792  1.4380  1.5690  8.8280
```

```
par(mfrow=c(1,2))
hist(data$density, breaks=20,
     main="Hist of People/Sq. Mile",
     xlab="Probability of people per sq. mile")
hist(log(data$density), breaks=20,
     main="Hist of Log of people/Sq. Mile",
     xlab="Log of people per sq. mile")
```





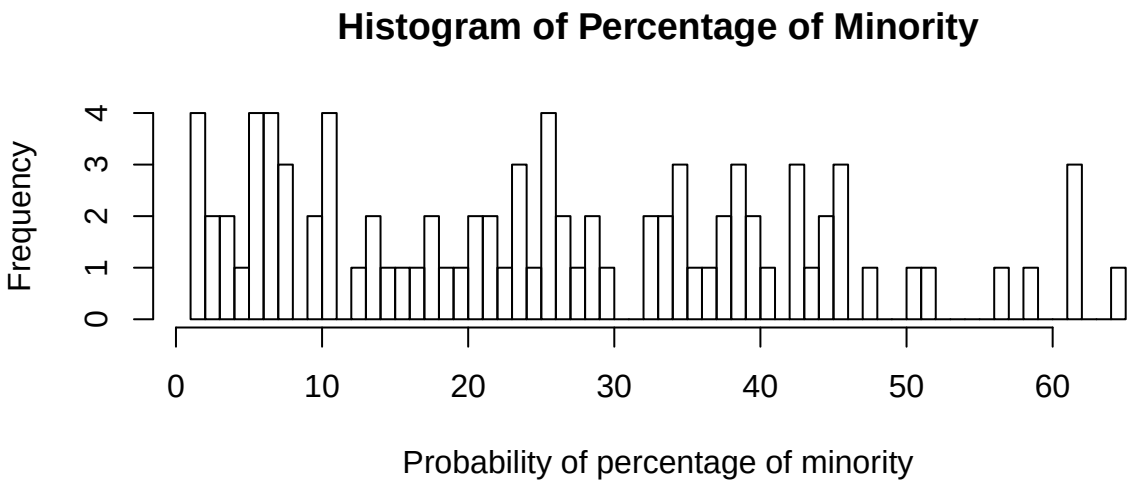
Percentage of Minority, 1980 (pctmin)

The data and histogram for pctmin looks as expected. However, the data is between 0 - 100 whereas other percentage variables is between 0.0 and 1.0.

```
summary(data$pctmin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.284  10.020  24.850  25.710  38.180  64.350
```

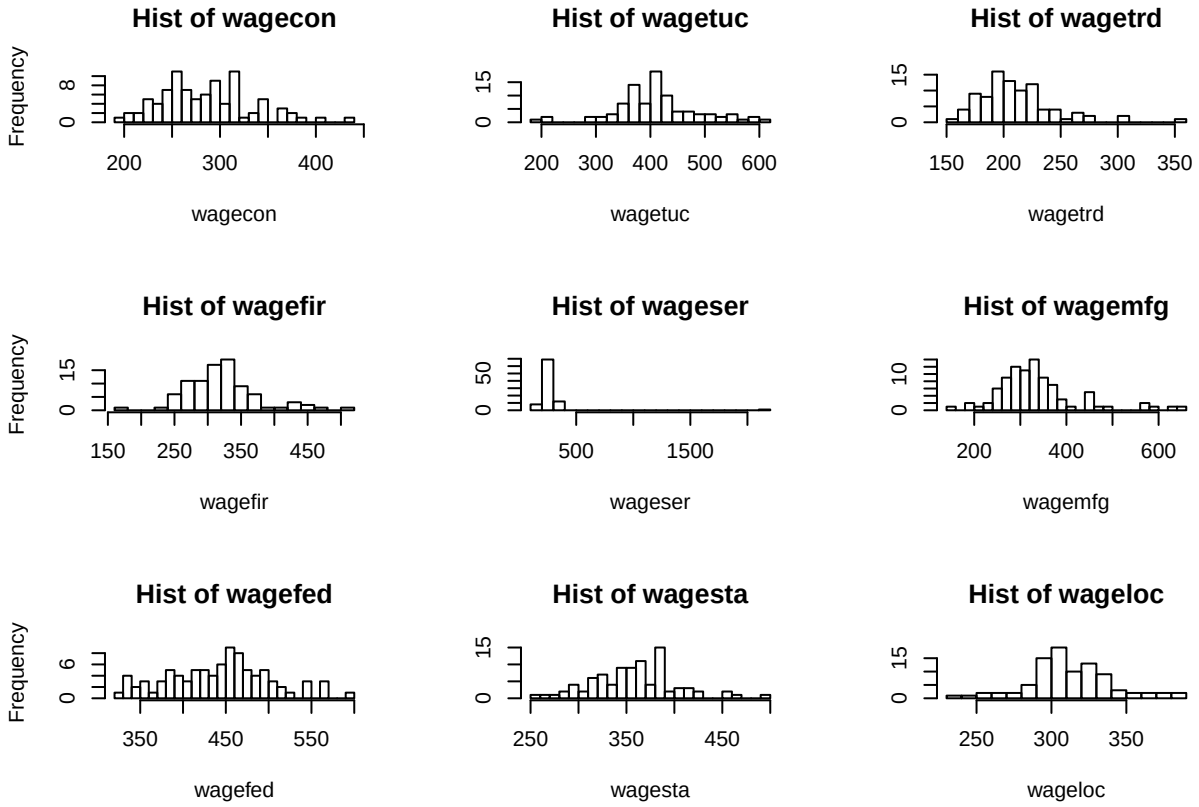
```
hist(data$pctmin, breaks=50,
     main="Histogram of Percentage of Minority",
     xlab="Probability of percentage of minority")
```



Wage variables (wage*)

Looking at the histogram of the wage variables, there is no obvious case for supporting any transforms on the variables, since for most cases the distribution does not look very skewed.

```
par(mfrow=c(3,3))
hist(data$wagecon, breaks=20, main="Hist of wagecon", xlab="wagecon", ylab="Frequency")
hist(data$wagetuc, breaks=20, main="Hist of wagetuc", xlab="wagetuc", ylab="")
hist(data$wagetrd, breaks=20, main="Hist of wagetrd", xlab="wagetrd", ylab="")
hist(data$wagefir, breaks=20, main="Hist of wagefir", xlab="wagefir", ylab="Frequency")
hist(data$wageser, breaks=20, main="Hist of wageser", xlab="wageser", ylab="")
hist(data$wagemfg, breaks=20, main="Hist of wagemfg", xlab="wagemfg", ylab="")
hist(data$wagefed, breaks=20, main="Hist of wagefed", xlab="wagefed", ylab="Frequency")
hist(data$wagesta, breaks=20, main="Hist of wagesta", xlab="wagesta", ylab="")
hist(data$wageloc, breaks=20, main="Hist of wageloc", xlab="wageloc", ylab="")
```



Focusing on wageser, there seems to be one data point that looks to be an extreme outlier. Looking further, it seems to be coming from data point 84. This data point has an extremely high value for probsen and wageser. For our model, we will remove this datapoint, as we have found that this data point tends to have high Cook's distance if we include it in our models.

```
data[data$X == 84,c("probsen", "wageser")]
```

```
##    probsen  wageser
## 84 2.12121 2177.068
```

Urban Indicator Variable

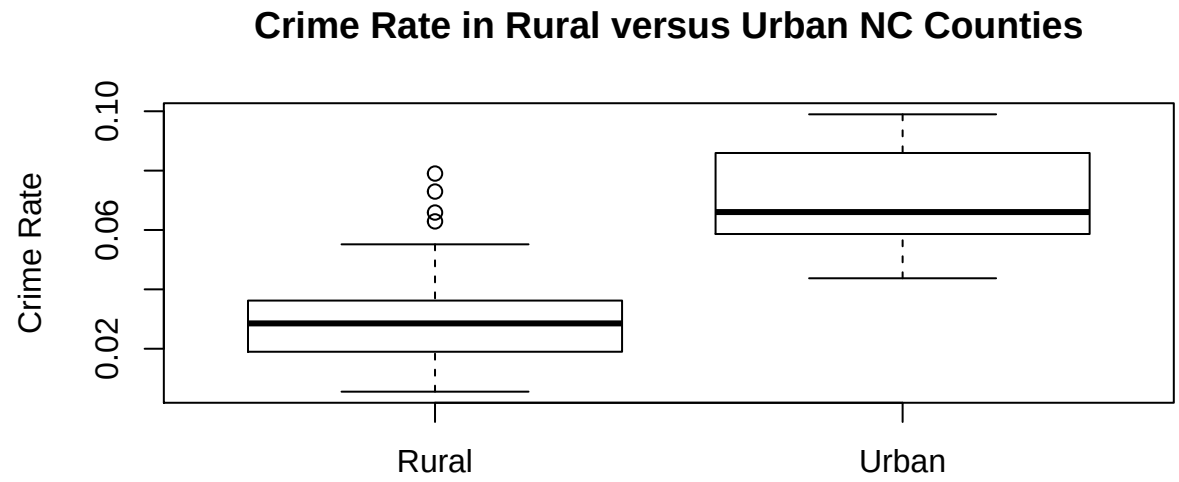
There are only 8 of the 90 counties in NC that are included in our dataset are urban.

```
sum(data$urban == 1)
```

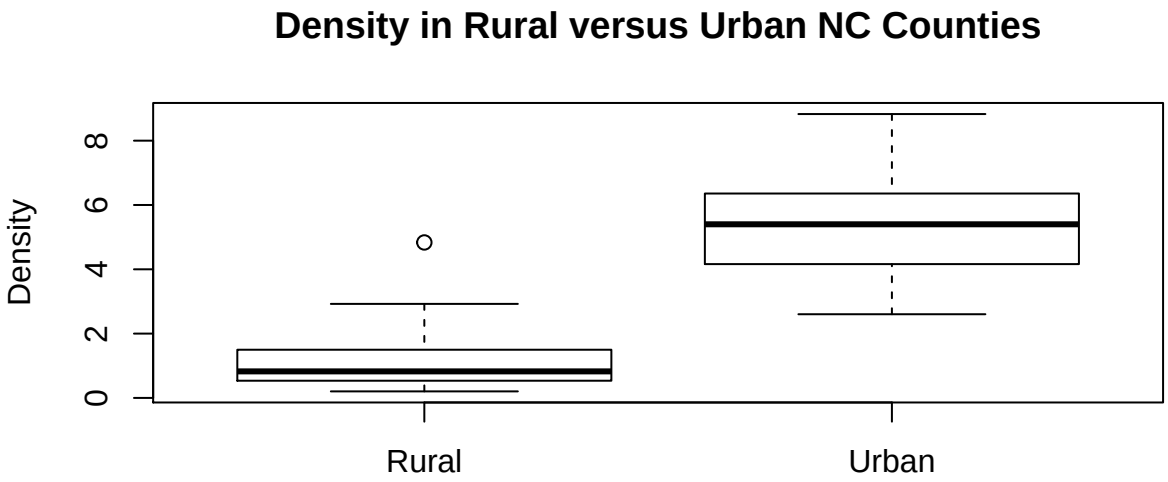
```
## [1] 8
```

When we look at crime rate in urban vs rural counties, it appears about 3 times higher in urban counties than rural ones. However, note the urban sample is only 8 observations of 90. As a result, given the low number of urban samples in the dataset as well as the fact that we feel the log of density variable already captures at a more granular level how urban or rural a county is, we decided not to leverage this indicator variable in our models.

```
boxplot(data$crime ~ data$urban, ylab = "Crime Rate",
        main = "Crime Rate in Rural versus Urban NC Counties", names=c("Rural", "Urban"))
```




```
boxplot(data$density ~ data$urban, ylab = "Density",
        main = "Density in Rural versus Urban NC Counties", names=c("Rural", "Urban"))
```



Transformed and Filtered Dataset

Based on the univariate analysis, the following transformations were proposed, as well as removal of one data point as described in the previous section. As well, the percentage male and probabilities are normalized to 0 - 100 for better model interpretability.

```
data$log_crime = log(data$crime)
data$log_police = log(data$police)
data$log_density = log(data$density)
data$ymale = data$ymale * 100
data$probsen = data$probsen * 100
data$probconv = data$probconv * 100
data$probarr = data$probarr * 100
data = data[data$X != 84,]

sort(colnames(data))
```

```
##  [1] "avgsen"      "central"     "county"      "crime"       "density"
##  [6] "log_crime"   "log_density" "log_police"  "mix"         "pctmin"
## [11] "police"      "probarr"     "probconv"    "probsen"     "tax"
## [16] "urban"       "wagecon"     "wagefed"     "wagefir"     "wageloc"
## [21] "wagemfg"     "wageser"     "wagesta"     "wagetrd"     "wagetuc"
## [26] "west"        "X"           "year"        "ymale"
```

Bi-variate analysis

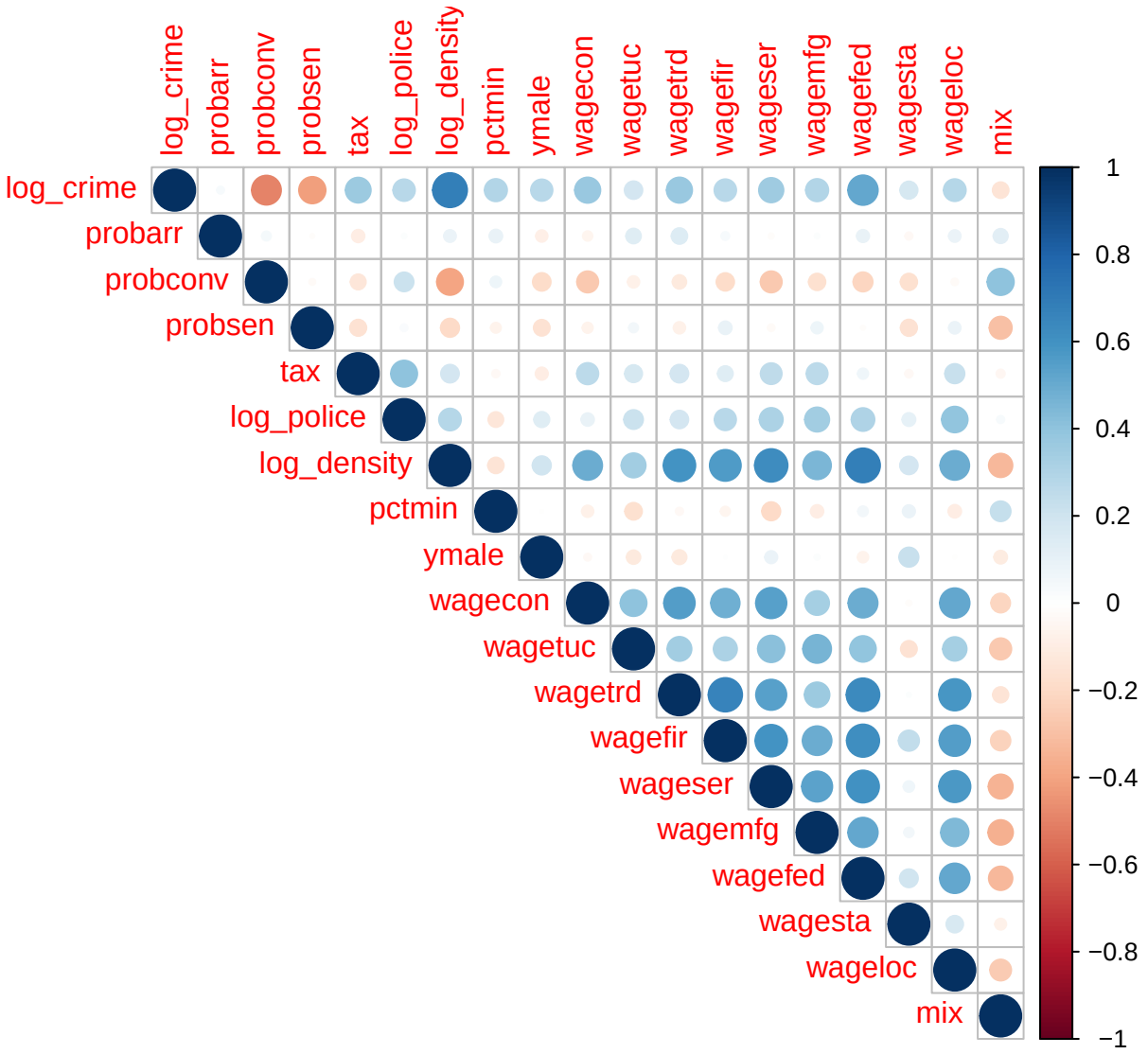
Looking at the correlation plot of the non-indicator variables, it looks like the following variables correlate with log of crime (first row in the plot) highly and positively:

1. Log of density
2. Tax
3. Wage variables

It looks like the following variables correlate with log of crime (second-last column in the plot) highly and negatively:

1. Probability of conviction
2. Probability of sentencing

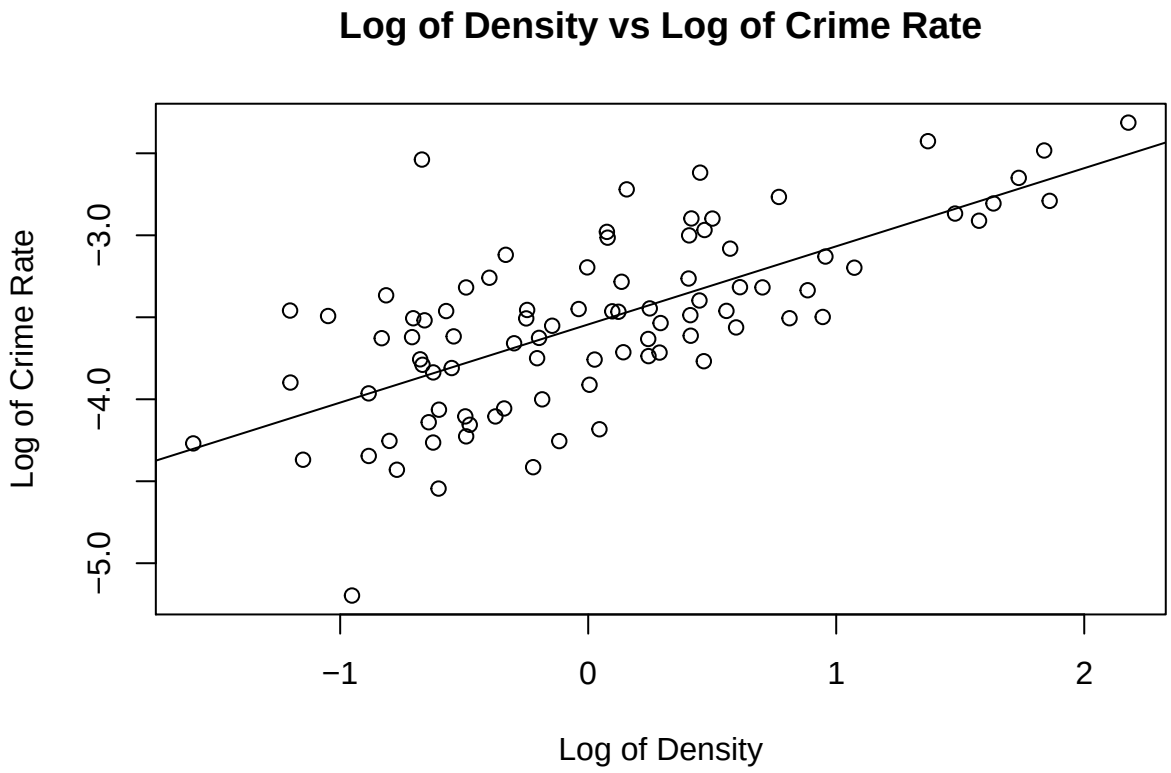
```
corrplot(cor(data[ ,
                        c("log_crime", "probarr", "probconv", "probsen", "tax", "log_police", "log_density",
                          "pctmin", "ymale", "wagecon", "wagetuc", "wagetrd", "wagefir", "wageser", "wagemfg"
                          "wagefed", "wagesta", "wageloc", "mix")]), method="circle", type="upper")
```



Log of Density vs Log of Crime Rate

As seen from the correlation plot, as well as the x-y plot below, there is a clear positive linear relationship between log of density and log of crime rate. Seeing a positive relationship makes sense, since the crime rate tends to increase in more populated areas.

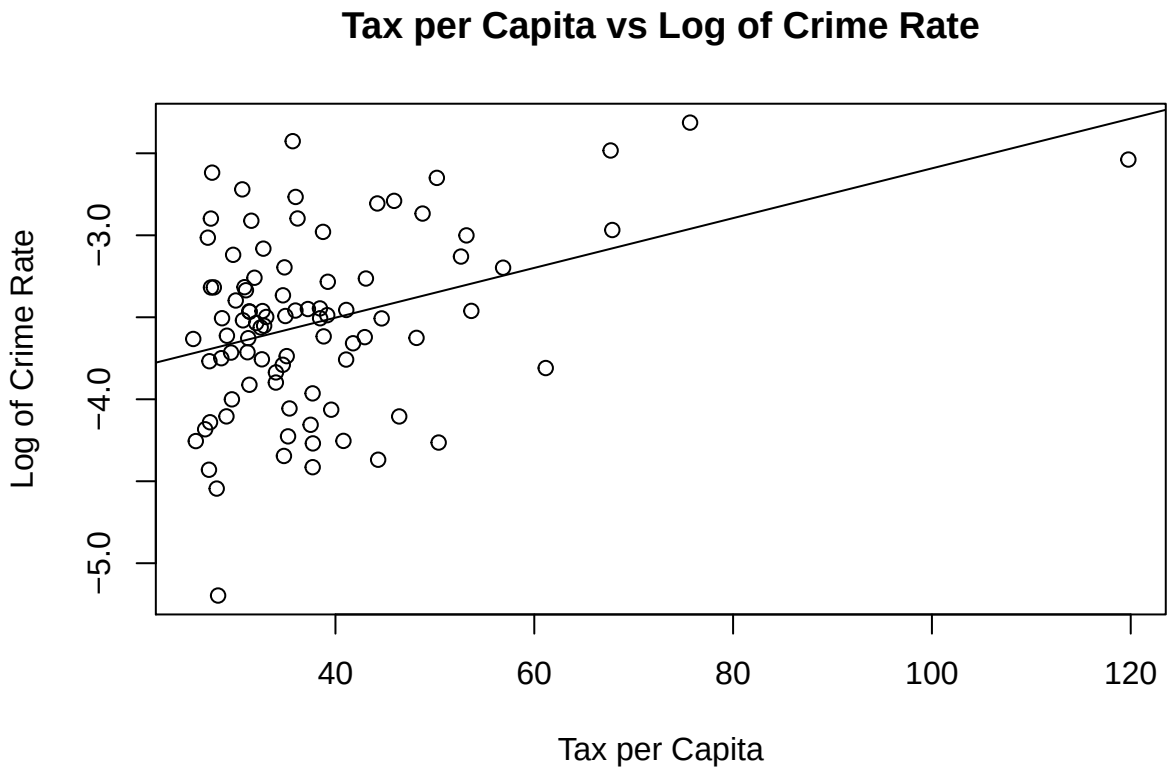
```
plot(data$log_density, data$log_crime,
     main="Log of Density vs Log of Crime Rate", ylab="Log of Crime Rate", xlab="Log of Density")
abline(lm(data$log_crime ~ data$log_density))
```



Tax vs Log of Crime Rate

Similar, as seen from the correlation plot, as well as the x-y plot below, there is a positive linear relationship between tax per capita and log of crime rate. Seeing a positive relationship also makes sense here since crime rate could increase in areas where there are more tax revenues collected.

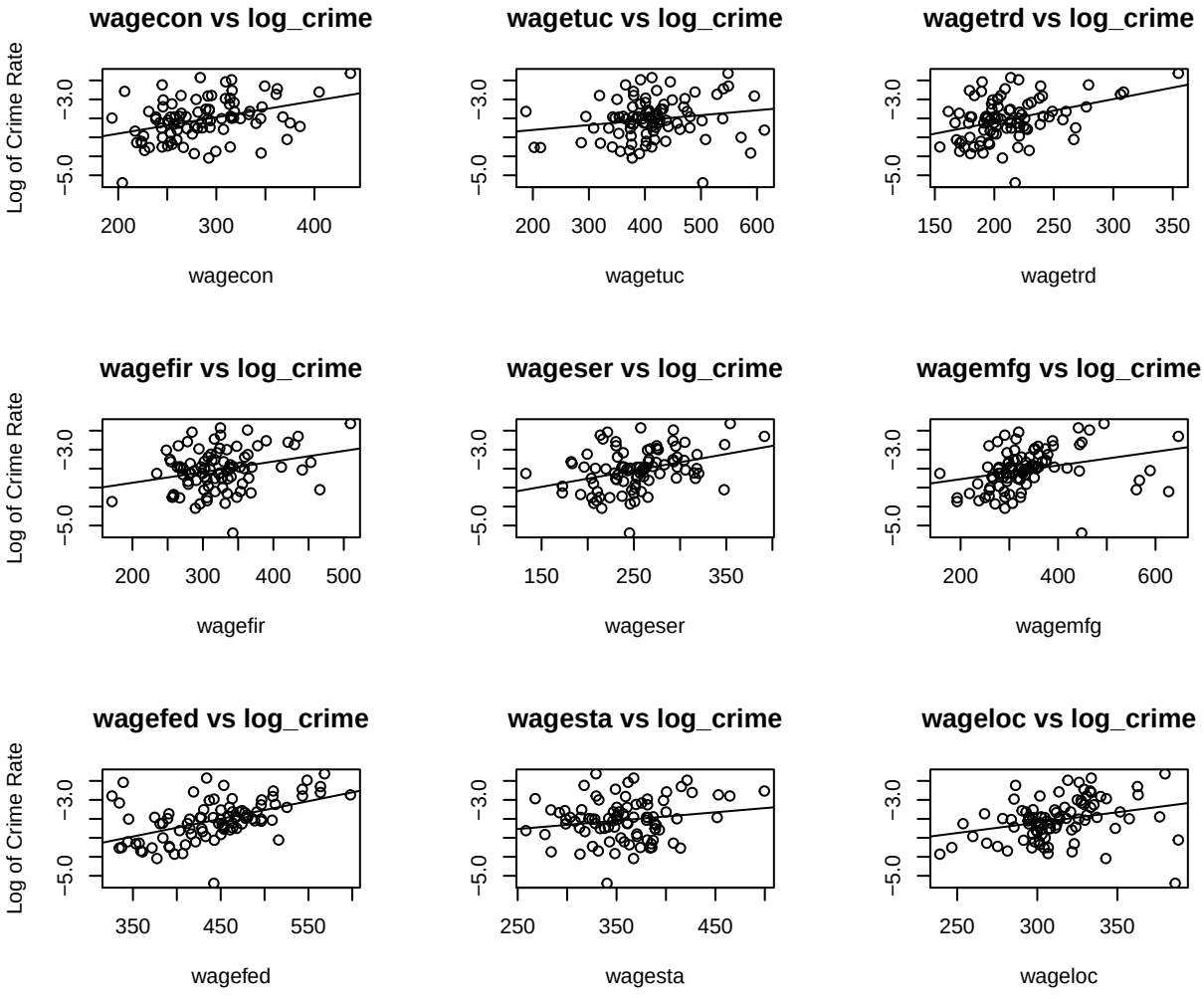
```
plot(data$tax, data$log_crime,
     main="Tax per Capita vs Log of Crime Rate", ylab="Log of Crime Rate", xlab="Tax per Capita")
abline(lm(data$log_crime ~ data$tax))
```



Wages vs Log of Crime Rate

From the correlation plot, as well as the individual x-y plots below of wages against log of crime rate, each of these seem to show some positive linear relationship. Seeing positive relationships here could be explained by the fact that crime rate could increase in areas where there are more tax revenues collected (and hence are counties with more wealth).

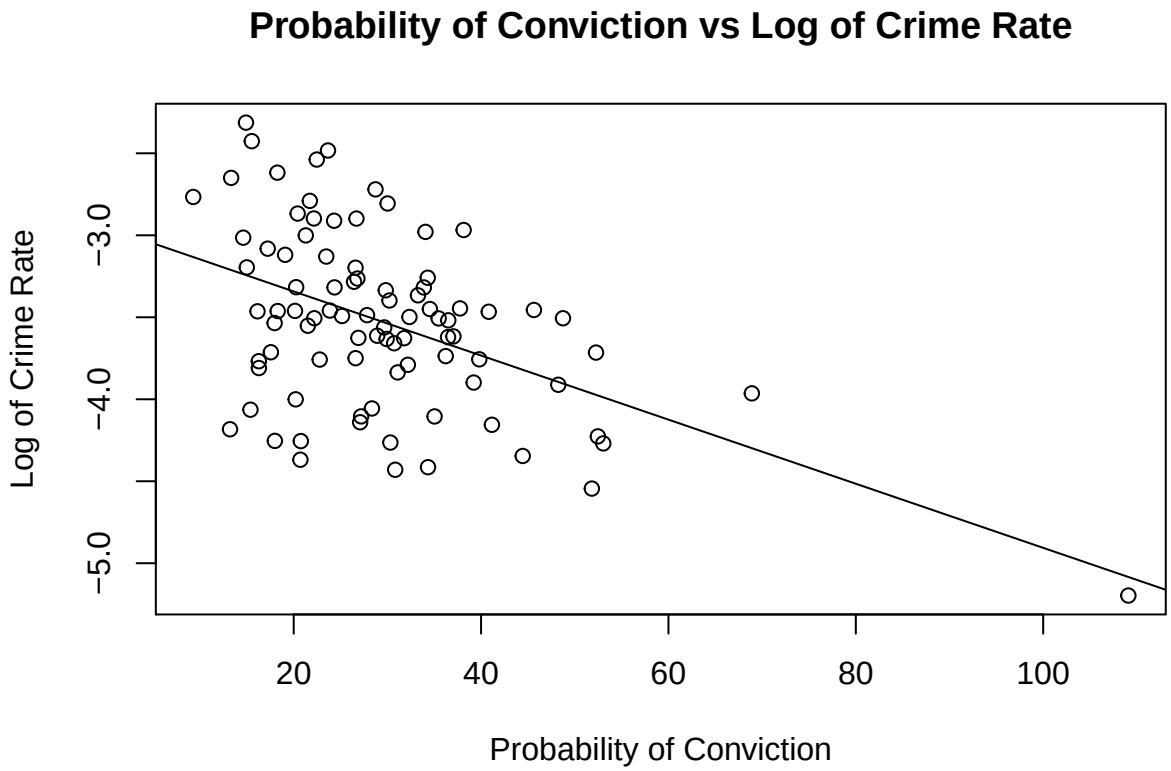
```
par(mfrow=c(3,3))
plot(data$wagecon, data$log_crime, main="wagecon vs log_crime", xlab="wagecon", ylab="Log of Crime Rate")
abline(lm(data$log_crime ~ data$wagecon))
plot(data$wagetuc, data$log_crime, main="wagetuc vs log_crime", xlab="wagetuc", ylab="")
abline(lm(data$log_crime ~ data$wagetuc))
plot(data$wagetrd, data$log_crime, main="wagetrd vs log_crime", xlab="wagetrd", ylab="")
abline(lm(data$log_crime ~ data$wagetrd))
plot(data$wagefir, data$log_crime, main="wagefir vs log_crime", xlab="wagefir", ylab="Log of Crime Rate")
abline(lm(data$log_crime ~ data$wagefir))
plot(data$wageser, data$log_crime, main="wageser vs log_crime", xlab="wageser", ylab="")
abline(lm(data$log_crime ~ data$wageser))
plot(data$wagemfg, data$log_crime, main="wagemfg vs log_crime", xlab="wagemfg", ylab="")
abline(lm(data$log_crime ~ data$wagemfg))
plot(data$wagefed, data$log_crime, main="wagefed vs log_crime", xlab="wagefed", ylab="Log of Crime Rate")
abline(lm(data$log_crime ~ data$wagefed))
plot(data$wagesta, data$log_crime, main="wagesta vs log_crime", xlab="wagesta", ylab="")
abline(lm(data$log_crime ~ data$wagesta))
plot(data$wageloc, data$log_crime, main="wageloc vs log_crime", xlab="wageloc", ylab="")
abline(lm(data$log_crime ~ data$wageloc))
```



Probability of Conviction vs Log of Crime Rate

From the correlation plot, as well as the individual x-y plots below for probability of conviction vs log of crime rate, there tends to be a somewhat clear negatively linear relationship. This makes sense since if the probability of convictions goes down, then there are more (potential) criminals out on the streets.

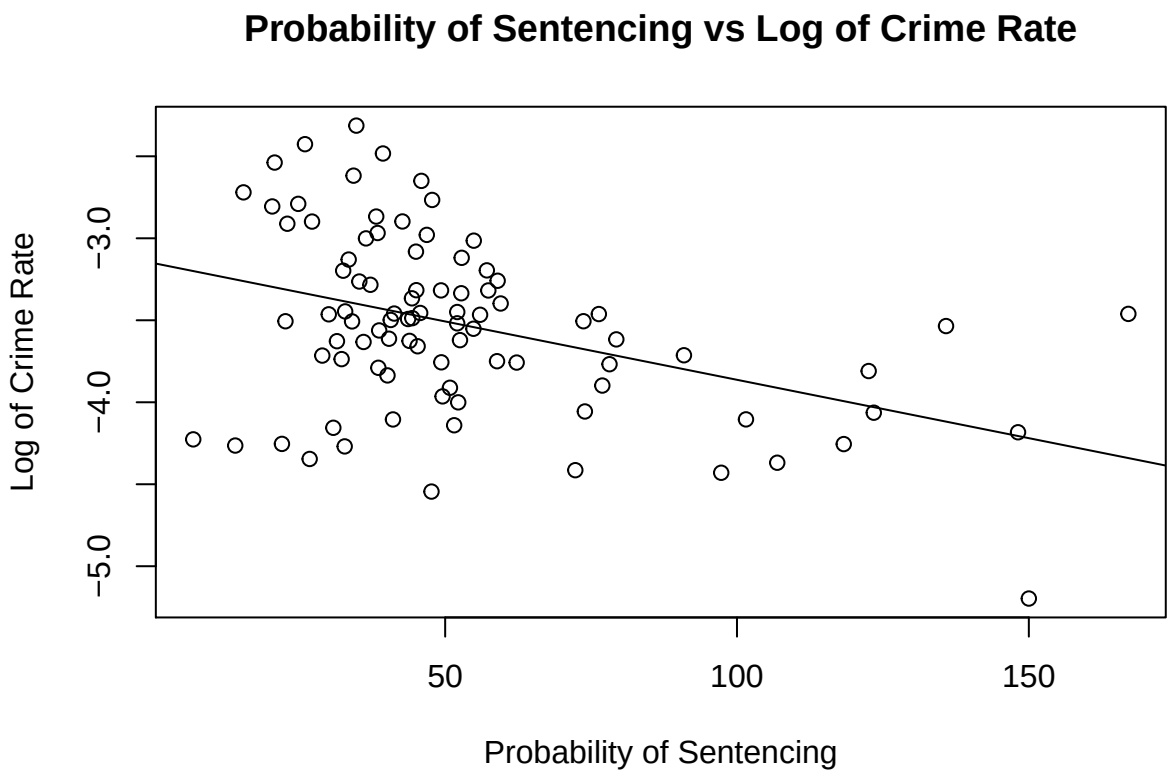
```
plot(data$probconv, data$log_crime,
     main="Probability of Conviction vs Log of Crime Rate", ylab="Log of Crime Rate", xlab="Probability of Co
abline(lm(data$log_crime ~ data$probconv))
```



Probability of Sentencing vs Log of Crime Rate

Similarly, from the correlation plot, as well as the individual x-y plots below for probability of sentencing vs log of crime rate, there tends to be a negatively linear relationship. This also makes sense since if the probability of sentencing goes down, then there are more (potential) criminals out on the streets.

```
plot(data$probsen, data$log_crime,
     main="Probability of Sentencing vs Log of Crime Rate", ylab="Log of Crime Rate", xlab="Probability of Se
abline(lm(data$log_crime ~ data$probsen))
```



Section 3 - Model Specification and Assumptions

In our exploratory analysis, we identified key independent variables that were positively and negatively correlated with log of crime rate. To create our first and simplest model that contains variables of key interest, we included the subset of these variables that we hyothesized might be the most important determinants of crime.

It is "common knowledge" that areas with higher denisty have more crime, therefore we included log_density in our first model. In addition, we hypothesized that a high probability of arrest, sentencing and coviction would be deterrents of crime. While researching determinants of crime, we found that historically, one of the strongest determinants of crime has been the percent of young males in population, perhaps because they are the most likely perpetrators of crime. We therefore included this variable in our first model as well. In addition, we thought that higher taxes might increase the effectiveness of police and criminal justice capacities.

$$log(CrimeRate) = \beta_0 + \beta_1 log(density) + \beta_2 conviction + \beta_3 sentencing + \beta_4 arrests + \beta_5 tax + \beta_6 youngmale + u$$

We estimated the above equation by least ordinary squares regression to produce the linear model summarized in Table 1.

```
final_model1 = lm(log_crime ~ log_density + probsen + probconv + probarr + ymale + tax, data = data)
se.final_model1 = sqrt(diag(vcovHC(final_model1)))
```

In order to determine if our OLS coefficients will be unbiased estimates of the population parameters and in order to be able to perform statistical inference, we examined the six classic linear model assumptions for model 1.

CLM.1 - Linear in Parameters

We chose our model specification so that the dependent variable is a linear function of the explanatory variables. Therefore, the CLM.1 assumption is met for our first model and all the other models we created.

CLM.2 - Random Sampling

There were originally only 90 counties in the dataset. During our data cleaning, we removed 1 county because we judged that it contained an error in the input for the weekly wage of service employees variable. During our research into North Carolina, we discovered that there have been 100 counties in North Carolina since 1911. Therefore, our dataset does *not* contain every county in North Carolina. However, we didn't identify any indications of non-random sampling during our analysis. We therefore assume that the counties are a random sample of the 100 counties in North Carolina. Thus, the MLR.2 assumption is met for our first model and all the other models we created.

CLM.3 - No Perfect Multicollinearity

From our EDA, it is apparent that none of our variables has constant values across the dataset. In addition, inspection of the correlation plot above indicates that there are no perfectly correlated variable pairs. In addition, analysis of the variance inflation factor for each variable does not provide evidence of multicollinearity.

```
vif(final_model1)
```

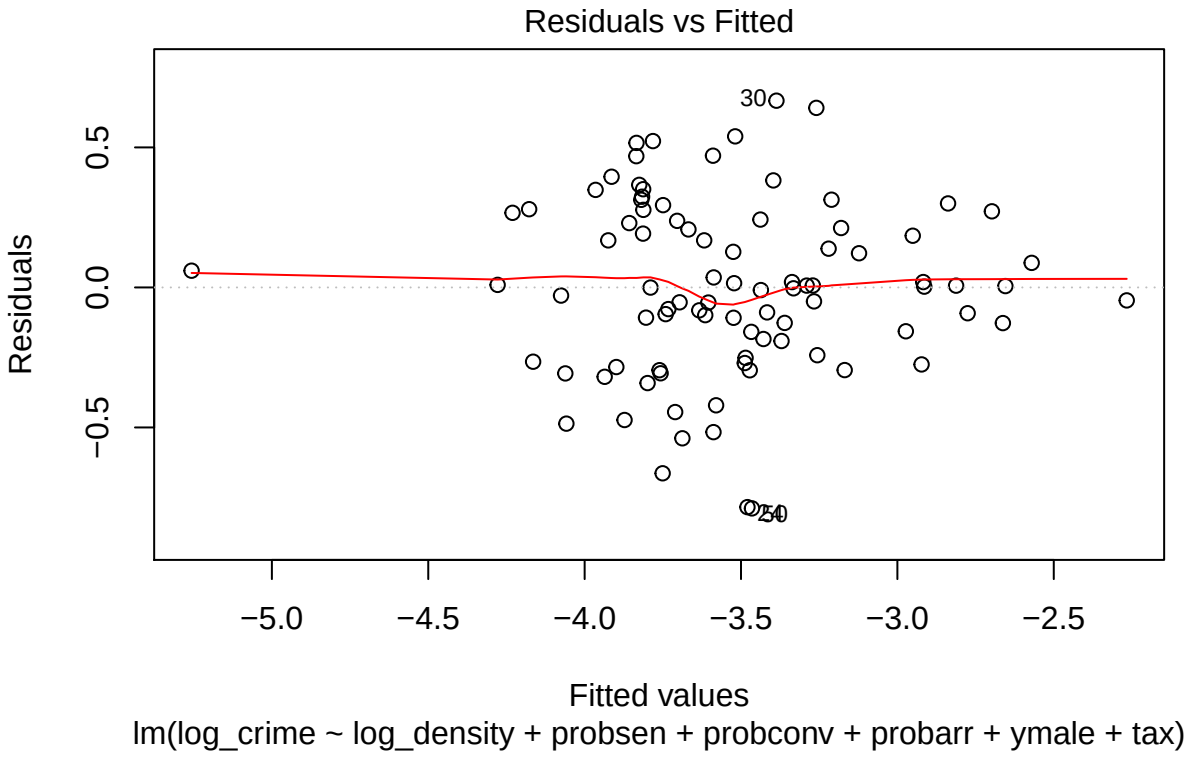
```
## log_density     probsen    probconv     probarr       ymale         tax
##    1.317732    1.098417    1.246982    1.039479    1.116629    1.105165
```

Thus, the MLR.3 assumption is met for our first model and all the other models we created.

CLM.4 - Zero Conditional Mean

By examining the residuals verus fitted values plot for our first model, we conclude that the assumption of zero conditional mean is met. The red spline curve does not deviate much from zero.

```
plot(final_model1, which = 1)
```

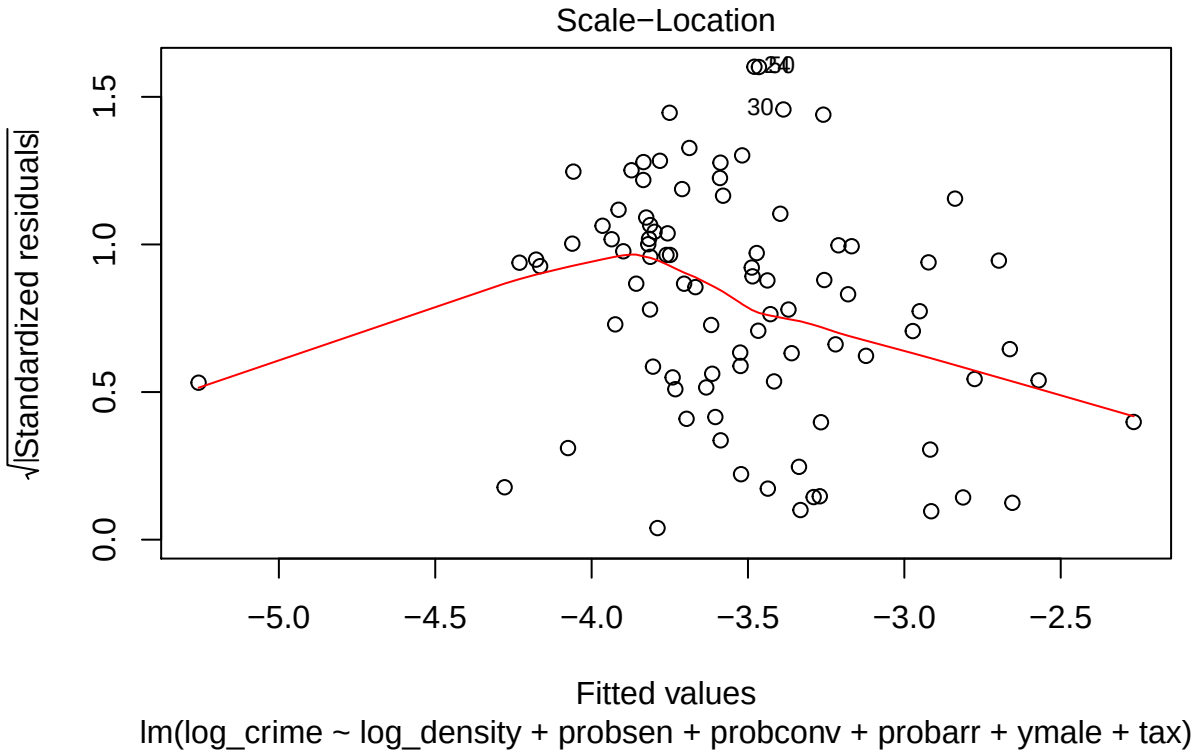

Residuals vs Fitted
30
54
Residuals
0.5
0.0
−0.5
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
Fitted values
lm(log_crime ~ log_density + probsen + probconv + probarr + ymale + tax)

CLM.5 - Homoskedasticity

When we examined the residuals versus fitted values plot, it was apparent that the variance of errors to the right of the plot is smaller than the variance of errors in the middle and left of the plot. This suggested heteroskedasticity, so we examined the scale-location plot. The spline curve on the scale-location plot is curved rather than flat, indicating heteroskedasticity.

Despite this clear violation of CLM.5, we are able to proceed with our OLS model by using heteroskedasticity-robust standard errors.

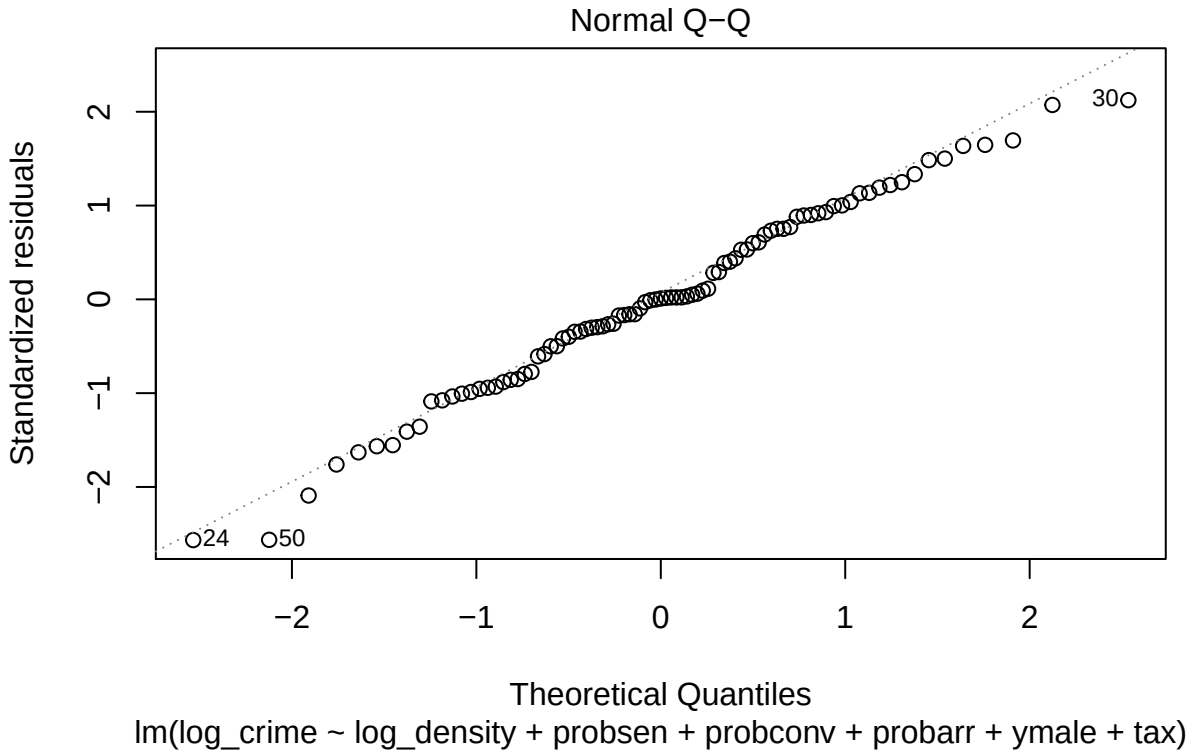
```
plot(final_model1, which = 3)
```



CLM.6 - Normality of Residuals

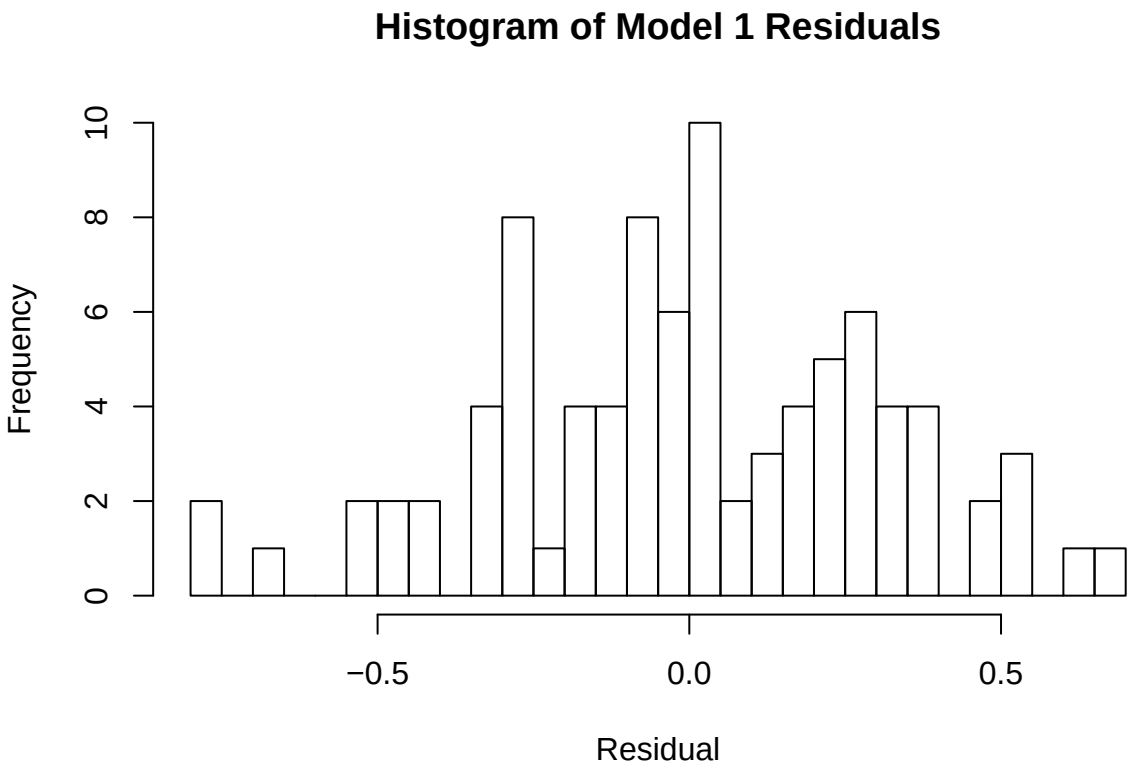
Analysis of the qqplot of residuals for model 1 suggested that the residuals for model 1 were approximately normally distributed.

```
plot(final_model1, which = 2)
```



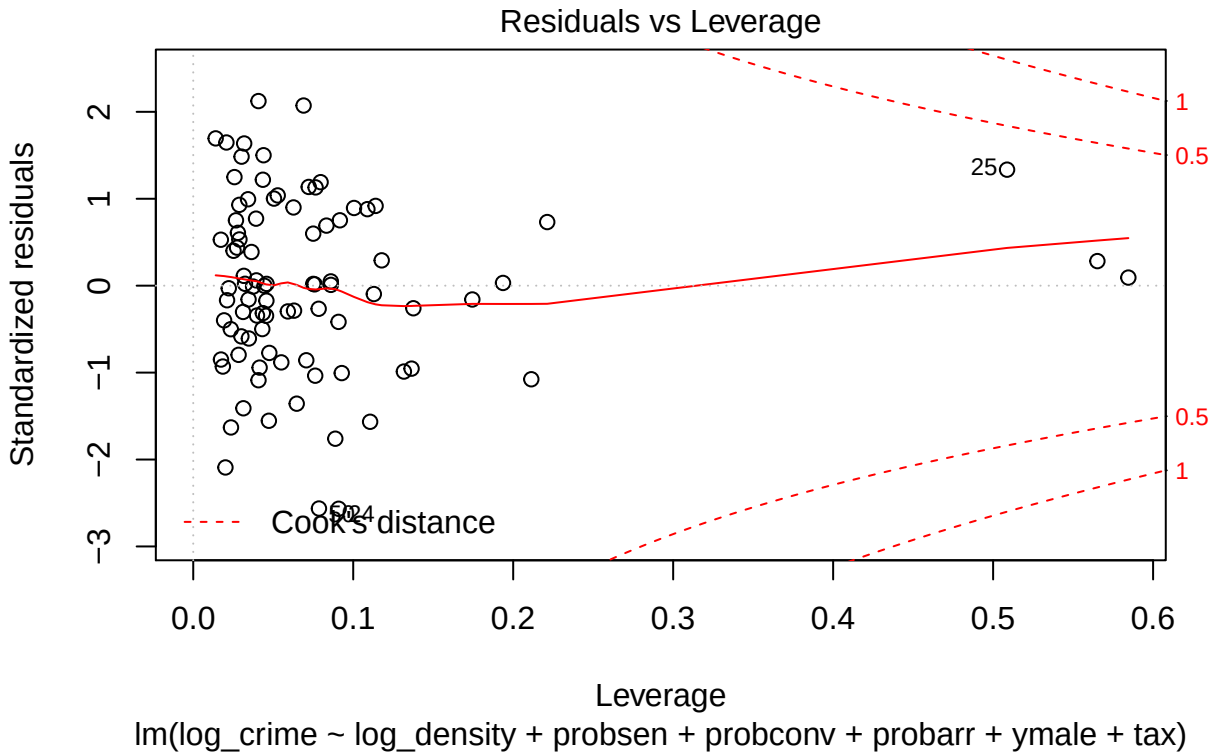
We further examined the distribution of the residuals in model 1 in a histogram. The residuals are somewhat normal, and in addition our sample size of 89 allows us to rely on asymptotics and the Central Limit Theorem. We can thus perform hypothesis testing of our OLS coefficients.

```
hist(final_model1$residuals, breaks = 50,
     main="Histogram of Model 1 Residuals", xlab="Residual")
```



Lastly, while there are several data points in our model with high leverage, their residuals are not large. Thus, none of the data points has a disconcertingly large Cook's distance and none cause undue influence on model fit.

```
plot(final_model1, which=5)
```



Conclusions from Examining CLM Assumptions and Interpretation of Model 1 Metrics

As seen in the regression table (Table 1), the F statistic for the omnibus test of model 1 is 28.2 with a statistically significant p value of $< 2.2e\text{-}16$. We therefore reject the null hypothesis that none of the independent variables help to describe log(crime rate).

The p values for the t tests for the coefficients on log_density, probconv, probsen, were all < 0.001 and ymale and tax were less than 0.05. For each coefficient, we therefore reject the null hypothesis that the coefficient is equal to zero. Because we will give a detailed summary of the coefficients for our favored model 2 (below), we won't discuss the practical significance of all the coefficients in model 1.

The adjusted R2 for model 1 is quite high for a social science study (0.65). However, the Akaike Information Criterion (AIC) is 58.8 - fairly large compared to other models.

```
AIC(final_model1)
```

```
## [1] 58.80868
```

Model 2: More Refined and More Robust than Model 1

After creating our first model, we carefully considered how robust the coefficients were for the variables we had included. We considered whether any of the variables we had not included in model 1 could be considered "omitted variables" - meaning, were they both correlated with log(crime) and with at least one other independent variable?

To address these issues, we created more than twenty other models (not shown). The results of these analyses suggested that we should include percent minority and log(police per capita) in our second model, and (to our surprise) exluce percent young male because the coefficient for this variable was not robust across multiple models. The same was true for the tax variable.

$$log(CrimeRate) = \beta_0 + \beta_1 log(density) + \beta_2 conviction + \beta_3 sentencing + \beta_4 minority + \beta_5 log(police) + u$$

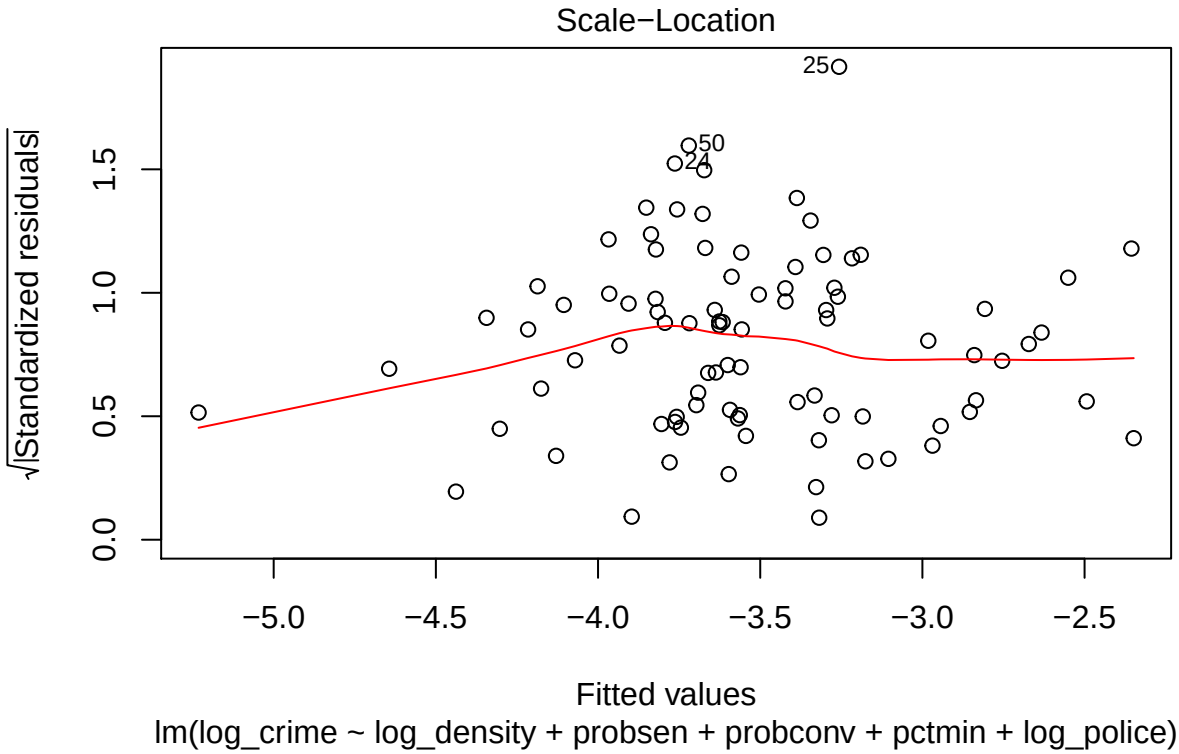
We estimated the above equation by least ordinary squares regression to produce the linear model summarized in Table 1.

```
final_model2 = lm(log_crime ~ log_density + probsen + probconv  + pctmin + log_police, data = data)
se.final_model2 = sqrt(diag(vcovHC(final_model2)))
```

Conclusions from Examining CLM Assumptions and Interpretation of Model 2 Coefficients

Similar to our first model, we tested the CLM assumptions for model 2 and found that we only had to address CLM.4 (homoskedasticity) by using heteroskedasticity robust standard errors.

```
plot(final_model2, which = 3)
```



Because our model only violates CLM.4 (homoskedasticity), we can conclude that our model coefficients will be unbiased estimators of the population parameters. In addition, we have little reason to suspect endogeneity and we feel comfortable interpreting particular coefficients in the model as causal.

As seen in the regression table (Table 1), the F statistic for the omnibus test of model 2 is 88.4 with a statistically significant p value of $<$ 2.2e-16. We therefore reject the null hypothesis that none of the independent variables help to describe log(crime rate).

The p values for the t tests for the coefficients on log_density, probsen, probconv, and pctmin are all less than 0.001. The p value for the t test for the coefficient of log_police is less than 0.01. Therefore, for every coefficient, we reject the null hypothesis that the coefficient is equal to zero.

The coefficient for log_density is 0.29. This suggests that for a one percentage point increase in density, there is a 0.29 percent point increase in crime rate, holding all other factors fixed. Thus, changes in density do not have a practically significant effect on crime rate.

The coefficient for probsen is -0.006. This suggests that for a one percent increase in the probability of sentencing, there is a 0.6 percentage point decrease in crime rate, holding all other factors fixed. Thus changes in probability of sentencing do have practical significance on crime rate, but not a very large effect.

The coefficient for probconv is -0.017. This suggests that for a one percent increase in the probability of conviction, there is a 1.7 percentage point decrease in crime rate, holding all other factors fixed. Thus changes in probability of conviction do have practical significance on crime rate, but not a very large effect.

The coefficient for pctmin is 0.013. This suggests that for a one percent increase in percent minority, there is a 1.3 percentage point increase in crime rate, holding all other factors fixed. Thus a change in percent of people who are minorities does have a practical significance on crime rate, but not a very large effect.

The coefficient for log_police is 0.45. This suggests that for a one percentage point increase in police per capita, there is a 0.45 percentage point increase in crime rate, holding all other factors fixed. Thus changes in the number of police per capita do have practical significance on crime rate, but not a particularly large effect.

The adjusted R2 for this model (0.83) is better than for model 1 (0.65). In addition the AIC for model 2 is -7.7, indicating that the model predicts a very large percent of the variability in crime rate while being highly parsimonious compared to model 1 whose AIC is 58.8.

```
AIC(final_model2)
```

```
## [1] -7.741259
```

Model 3: The "Kitchen Sink" Model

To generate model 3, we added most of the other covariates from the dataset:

$$log(CrimeRate) = \beta_0 + \beta_1 log(density) + \beta_2 conviction + \beta_3 sentencing + \beta_4 arrests + \beta_5 minority$$

$$+\beta_6 log(police) + \beta_7 yougmale + \beta_8 urban + \beta_9 tax + \beta_{10} wagecon + \beta_{11} wagetuc + \beta_{12} wagetrd$$

$$+\beta_{13} wagefir + \beta_{14} wagemfg + \beta_{15} wagefed + \beta_{16} wagesta + \beta_{17} wageloc + \beta_{18} wageser + u$$

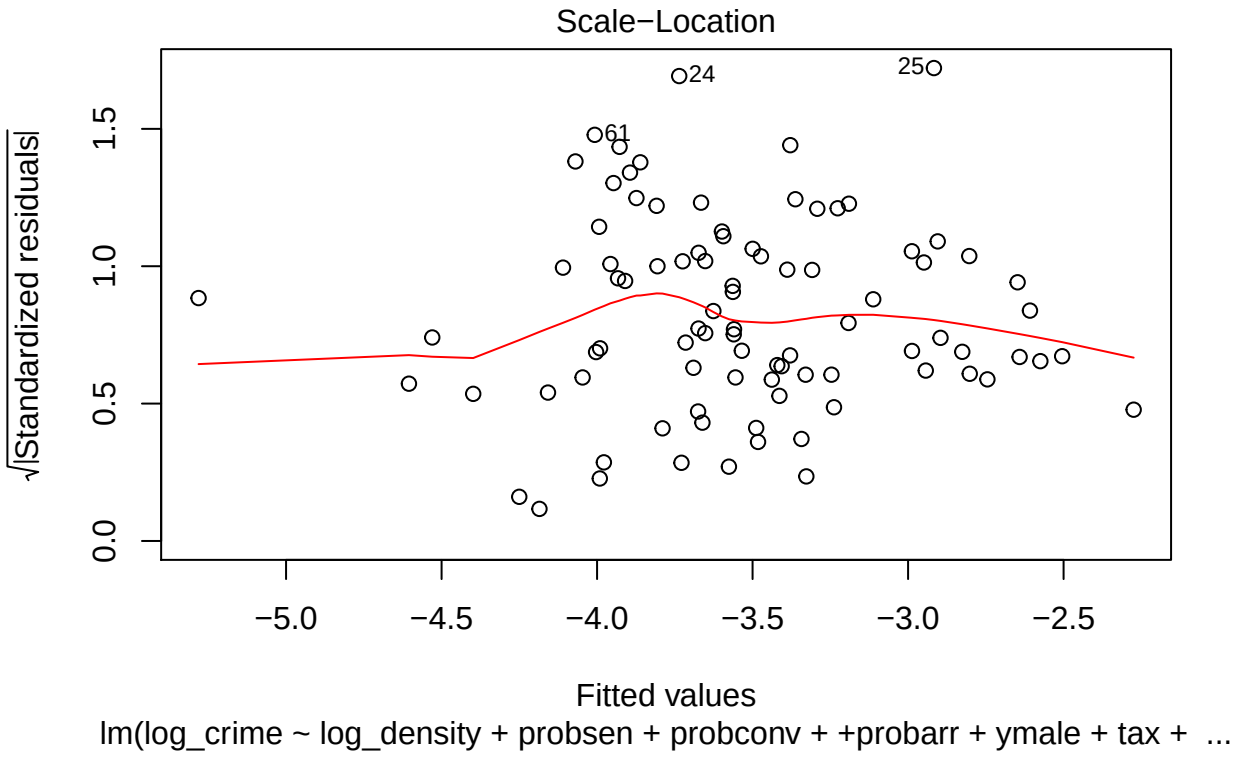
We estimated the above equation by least ordinary squares regression to produce the linear model summarized in Table 1.

```
final_model3 = lm(log_crime ~ log_density + probsen + probconv + +probarr + ymale + tax + pctmin +
                  log_police + urban + wagecon + wagetuc + wagetrd + wagefir + wagemfg + wagefed +
                  wagesta + wageloc + wageser, data = data)
se.final_model3 = sqrt(diag(vcovHC(final_model3)))
```

Conclusions from Examining CLM Assumptions and Interpretation of Model 3

We found that Model 3, like our other models, only violated CLM.4 - we identified heteroskedasticity.

```
plot(final_model3, which = 3)
```



To address this violation of CLM assumptions, we made use of heteroskedasticity-robust standard errors. Analysis of the variance inflation factor for the variables in model 3 did not indicate any serious multicollinearity. However, we note that the VIF for log(density) was greater than 4 in this model. We thus speculate that there is multicolinearity between the variables log(density), tax, and potentially log(police) in this model.

```
AIC(final_model3)
```

```
## [1] -9.591703
```

As seen in the regression table, the adjusted R2 for model 3 is 0.86. In addition, the AIC is -9.6. These metrics are indicators of a parsimonious model that can account for a huge amount of the variation in log(crime rate). However, we feel that this model is not as parsimonious as model 2 and includes variables that could unnecessarily increase variance of the coefficients resulting in potentially biased estimates. Thus, the primary purpose of this model is to support the idea that model 2 is a robust model. The coefficients of the variables from model 2 (log_density, probsen, probconv, pctmin & log_police) remain fairly constant compared to model 3 and these coefficients are still statistically significant in model 3 (see regression table below).

Another reason that we do not favor model 3 is that the coefficient for percent young male is statistically significant in model 3. Of more than 20 models we tested, this coefficient is only statistically significant in model 1 and model 3 (as discussed above while justifying the variables to include in model 2). This supports both the idea that the contribution of percent young male to log(crime rate) is not robust and also the idea that model 3 is not a robust model. In addition, model 3 contains many variables whose coefficients are not statistically significant from zero (see regression table). Thus we favor model 2, because it is the most robust model we have identified and it also accounts for a large amount of variation in log(crime rate).

```
stargazer(final_model1, final_model2, final_model3,
        se=list(se.final_model1, se.final_model2, se.final_model3),
        star.cutoffs=c(0.05, 0.01, 0.001), title = "Table 1: Linear Models to Predict log(Crime Rate)",
        column.labels = c("Model 1 - OK", "Model 2 - Great", "Model 3 - Kitchen Sink"), type="text")
```

```
##
## Table 1: Linear Models to Predict log(Crime Rate)
## ================================================================================================
##                                              Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                    log_crime
##                          Model 1 - OK         Model 2 - Great     Model 3 - Kitchen Sink
##                              (1)                   (2)                    (3)
## ------------------------------------------------------------------------------------------------
## log_density                0.317***              0.292***              0.319***
##                            (0.053)               (0.065)               (0.066)
##
## probsen                   -0.005***             -0.006***             -0.005***
##                            (0.001)               (0.001)               (0.001)
##
## probconv                  -0.011***             -0.017***             -0.016***
##                            (0.003)               (0.003)               (0.003)
##
## probarr                    0.002                                       -0.002
##                            (0.005)                                     (0.003)
##
## ymale                      0.025*                                      0.021*
##                            (0.010)                                     (0.009)
##
## tax                        0.009*                                       0.005
##                            (0.004)                                     (0.006)
##
## pctmin                                           0.013***              0.013***
##                                                  (0.002)               (0.002)
##
## log_police                                       0.452***              0.370**
##                                                  (0.135)               (0.121)
##
## urban                                                                  -0.105
##                                                                        (0.150)
##
## wagecon                                                                0.0004
##                                                                        (0.001)
##
## wagetuc                                                                0.0001
##                                                                       (0.0005)
##
## wagetrd                                                                0.001
##                                                                        (0.001)
##
## wagefir                                                                -0.001
##                                                                        (0.001)
##
## wagemfg                                                               -0.00003
##                                                                       (0.0004)
##
## wagefed                                                                0.001
##                                                                        (0.001)
##
```

```
## wagesta                                                                  -0.001
##                                                                          (0.001)
##
## wageloc                                                                   0.001
##                                                                          (0.002)
##
## wageser                                                                  -0.002
##                                                                          (0.001)
##
## Constant                  -3.601***              -0.155                  -0.968
##                            (0.281)               (1.004)                 (1.143)
##
## ------------------------------------------------------------------------------------------
## Observations                 89                     89                      89
## R2                          0.674                  0.842                   0.884
## Adjusted R2                 0.650                  0.832                   0.855
## Residual Std. Error    0.321 (df = 82)        0.222 (df = 83)         0.206 (df = 70)
## F Statistic         28.208*** (df = 6; 82) 88.447*** (df = 5; 83) 29.762*** (df = 18; 70)
## ==========================================================================================
## Note:                                                  *p<0.05; **p<0.01; ***p<0.001
```

Causality

The proposed model (model 2) indicates that harsher sentencing laws are the primary deterrent to crime, whereas police per capita does not improve the crime rate. While percentage minority does appear to have an impact on crime, the coefficient is small and the below analysis of omitted variables indicates that even this coefficient is likely overstated.

A number of metrics that are likely present in the true population model for predicting crime are not available for this analysis, and this impacts the overall predictability of our model beyond the current dataset. Police per person is a key variable exhibiting the wrong sign. In theory, police should reduce crime rather than increase it (as is currently predicted in our model). The below analysis of omitted variables underscores the complexity of assessing the impact of a police force among other factors on crime.

1. Police Profiling may increase the likelihood of minority groups and young males being arrested for crimes they do not commit. The coefficients of the pctmin and ymale variables are likely overstated without accounting for profiling.

2. Recidivism (the likelihood for people to become repeat offenders) is likely a very strong predictor of crime in specific regions. It would make sense that areas with higher crime overall, have higher incidences of recidivism, and therefore more police per capita. If recidivism and police per capita are positively correlated, then the police variable is absorbing some of the error term associated with the omitted variable, recidivism. Including this metric would likely help change the sign of the police variable; this would lead to a more intuitive model since police should be associated with reducing (not increasing) crime.

3. Police Motivation may also have an impact on probability of arrest. For example, if crime is very prevalent in one area while convictions or sentences in that same area are low, police may feel that their efforts are not having an impact. This could have either a positive or negative impact on the probarr variable, though a negative impact is more likely as police are less inspired to pursue arrests. In this case, fewer arrests would be reported for each crime and this factor would be unaccounted for in the model.

4. Security: Home security services such as ADT or Ring may reduce the prevalence of crime. Greater security around businesses (e.g. density of security guards) would have the same effect. Omitting these variables likely over-attributes crime reduction to stricter laws and more police than is accurate in the true population. Therefore probsen, probconv and police per capita are likely overstated in terms of reducing crime. (Note that this would make the police metric even more positive than it currently is). Additionally, with a security metric, the mix variable related to property crimes may have more meaning in the data.

5. The unemployment rate and education levels could help explain incidences of crime, and together these metrics may be better measures of socioeconomic status than total wages or tax per person. At the very least, one of the four metrics (or interactions between them) may allow socioeconomic status to have a significant role in the model.

6. Proximity to other high crime regions could be a notable factor to examine as well. For example, a region with low density and high crime may be contiguous to another region with high density/high crime, helping to further explain the

relationship between crime and density.

Conclusions

This data indicates that the main way we can deter crime is to employ stricter criminal justice practices, increasing the probability of convicting and sentencing those who have been arrested. In terms of practical significance, however, the coefficients for probability of conviction and probability of sentencing are so small that it would take an immense amount of effort — and severity — on the part of the criminal justice system to make a tangible difference in the crime rate.

Incorporating key omitted variables into the data set — and analyzing shifts in these variables over time — would help provide a holistic view of the human situations that lead to crime, and what can be done to deter it. For example, if additional data indicates that poverty, unemployment or a lack of an education are associated with overall crime and recidivism, we could focus efforts on better educating the nation's poor to give them positive alternatives to crime. Another recommendation could be to invest in rehabilitation programs or those that help folks get jobs following a prison sentence.

It is likely that taking a more humane approach than stricter sentencing laws will lead to a better vision for our society, and better outcomes for individuals. We recommend broadening this analysis so that all critical factors are taken into consideration.