W203_Lab3_Kramer_Liu

Beau Kramer, Rory Liu

April 1, 2018

Introduction

Our team of data scientists was hired to conduct research to help a political campaign understand the determinants of crime, and provide relevant policy suggestions for local governments.

Given this mandate, we look into the crime data compiled by Cornwell and Trumball, recording crime statistics along with demographics, policing, and wage variables in North Carolina in 1987. We will use this data to identify causal determinants of crime, focusing on the factors that are relevant and applicable to local governments. We will consider the crime rate our dependent variable. We will use this because it measures the crime per capita so is generalizable across counties and captures the crime level in a fairly direct way.

To begin, we first load the data and requisite packages into R.

```
options(warn = -1)
f = read.csv("crime_v2.csv")
library(car)
library(ggplot2)
library(corrplot)
library(stargazer)
library(gtable)
library(grid)
library(gridExtra)
library(sandwich)
library(dplyr)
```

Data Selection and Processing

We note that we have 97 observations and 25 variables.

We proceed to run the summary command on the full dataset, to get a snapshot view of all variables.

```
# Summarize dataset
summary(f)
```

```
##      county           year        crmrte             prbarr
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6
##         prbconv      prbpris           avgsen             polpc
##              : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6
##     density             taxpc             west           central
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6
##      urban            pctmin80           wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd            wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed            wsta            wloc            mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

Before further analysis we make the following notes:

- *prbconv* (probability of conviction) variable is a factor. It should instead be numeric like the other probability variables.
- outliers may exist for *wser* (wage in service industry), *pctymle* (percent male), and *taxpc* (tax revenue per capita)
- *pctmin80* is a percent but is in percentage points.
- 6 records have NA values. Upon further inspection these are completely blank rows in the CSV file.
- We have some entries in the data with prbarr and prbconv values greater than 1. Given these variables represent probabilities, theoratically we should not have values > 1. We will take a deeper look at these

rows to understand whether they are erroneous.

- The *density* values seem extremely low, with a mean of 1.43 people per square mile, and a max of 8.83 per square mile. We will dive into this when we look at this variable in more detail.

Next, we make the following actions to clean the data:

1) convert *prbconv* from a factor to a numeric
2) convert *pctmin80* from percentage points to percents
3) verify that NA rows are NA for all variables and drop them
4) remove duplicate records

```
options(warn = -1)
## Data cleaning for numeric data stored in strings
f$prbconv <- as.numeric(as.character(f$prbconv))
## Scale pctmin80 to be between 0 and 1
f$pctmin80 <- f$pctmin80/100
## Verify that the NA rows are NAs across all columns and drop
## them f[is.na(f$county),]
f <- f[!is.na(f$county), ]
# remove duplicate record
f = distinct(f)
## Examine probability entries > 1 f[f$prbarr>1,]
## f[f$prbconv>1,]
```

Upon investigation, we found that there is 1 entry with prbarr (probability of arrest) greater than 1, and 10 entries with prbconv (probability of conviction) greater than 1. Apart from having theoretically impossible probability values, these entries do not give any other indications to suggest they are erroneous (i.e. there are not whole numbers that look like recording errors). Therefore, we did not exclude them completely from the dataset. More analysis on these variables can be found in the univariate analysis later.

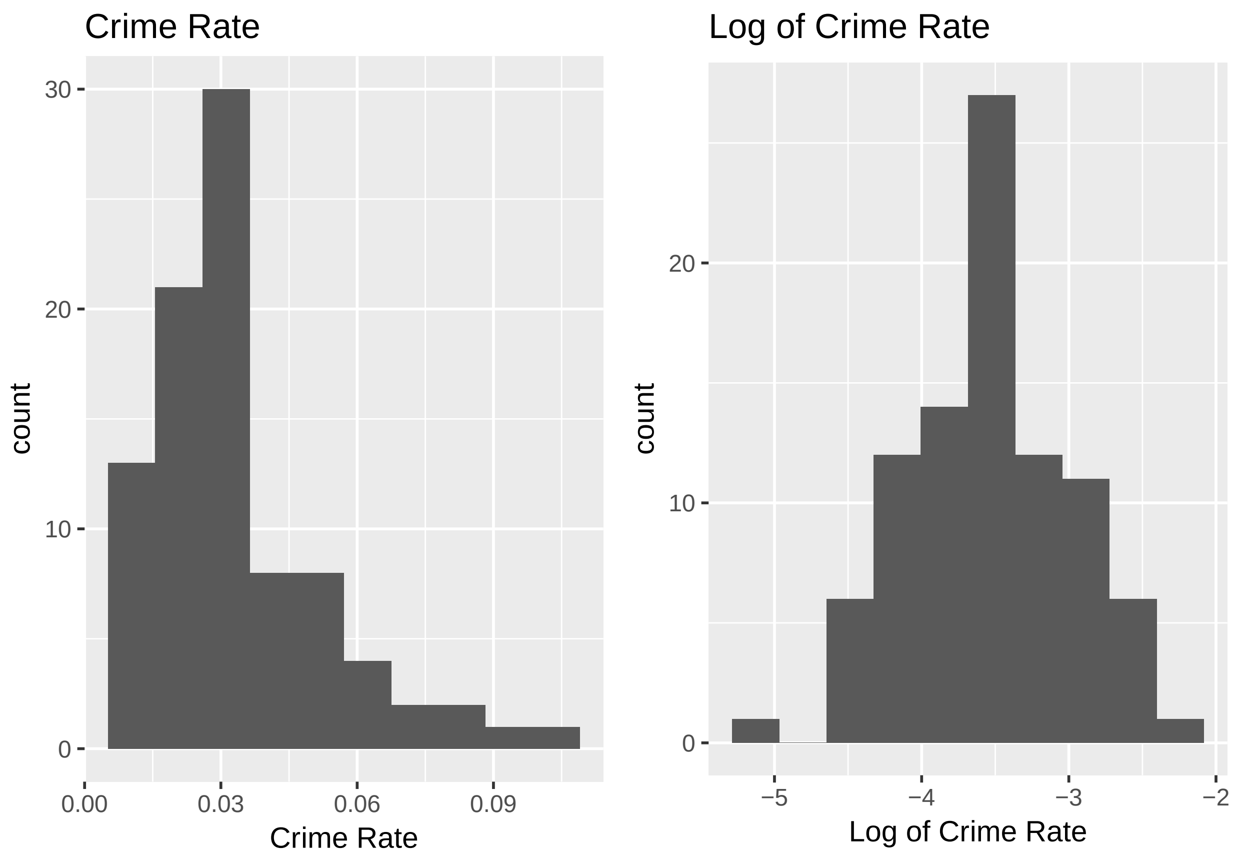
Intial Exploratory Data Analysis

We first examine the outcome variable *crmrte* (crime rate).

```
summary(f$crmrte)
```

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

```
p0 <- ggplot(data = f, aes(x = f$crmrte)) + geom_histogram(bins = 10) +
    xlab("Crime Rate") + ggtitle("Crime Rate")
p1 <- ggplot(data = f, aes(x = log(f$crmrte))) + geom_histogram(bins = 10) +
    xlab("Log of Crime Rate") + ggtitle("Log of Crime Rate")
grid.arrange(p0, p1, nrow = 1)
```



We note a range of 0.5% to 9.9% and a right skew in the data. This distribution is reasonable, with most counties experiencing a relatively low crime rate, and a handful of counties that are more crime-ridden. A quick cross check validates that in 2016, crime rate in North Carolina is 3109.7 per 100000 population(https://www.statista.com/statistics/301549/us-crimes-committed-state/), i.e. 0.031, which is slightly small than (but very close to) the average crime rate we see in this dataset from 1987.

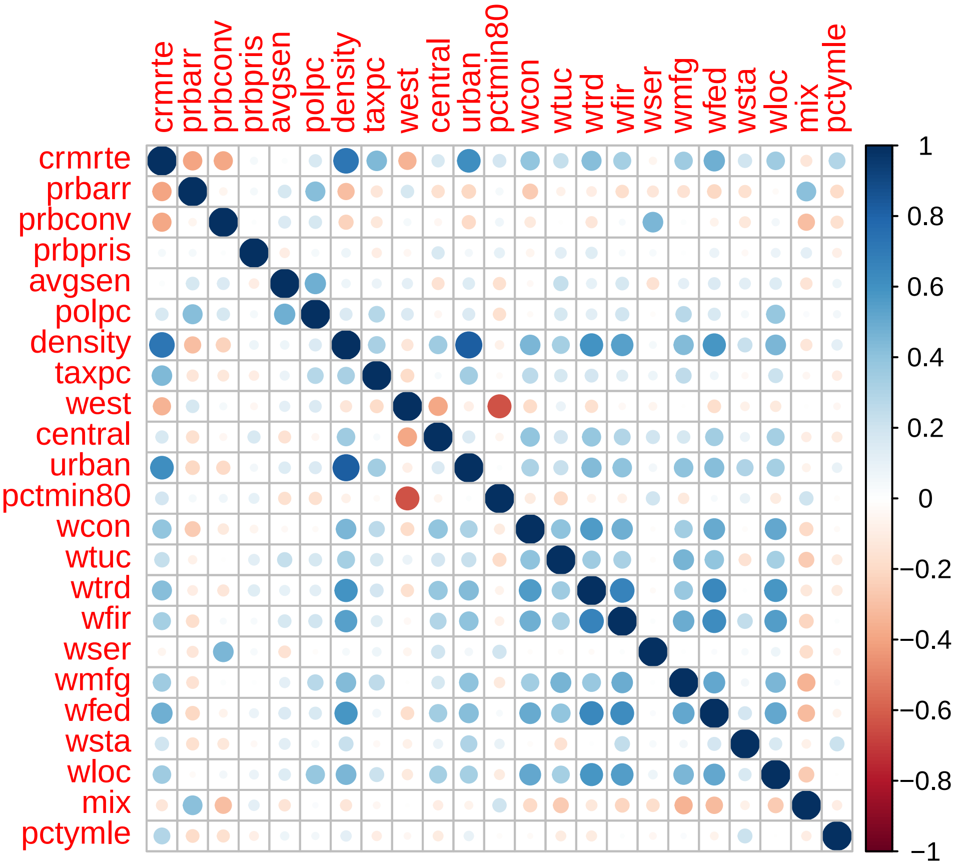
To correct for the right skew, we took the log of the crime rate variable here, and see that the distibution of log(crmrte) approaches normal. We will use log(crmrte) as our dependent variable for linear models, because it has a more normal distribution, and because we will be able to interpret the results in percent terms, instead of percentage point terms.

Next, we look at the cross correlations of the variables in the plot below. We conduct this analysis for 2 main purposes:

1) We want to check whether the correlations in the dataset are in accordance with our expectations and whether there are any anomalies. This should build our confidence in the dataset, and strengthen our trust in the conclusions we draw from the data.

2) We hope to understand which variables are strongly correlated with our outcome variable (*crmrte*), so that we can have a reasonable starting point in our univariate and regression analysis.

```
library(corrplot)
M = cor(f[3:25])
corrplot(M)
```



We first look into the existing relationships among the independent variables in our data. From the chart above, we observe strong correlations among the wage variables (*wcon, wtuc, wtrd, wfir, wmfg, wfed, wsta, wloc*). This is expected because when wage levels of a county are high, wage levels of other sectors in the same county tend to be high as well. The only wage variable that is not strongly correlated with others is *wser*. This is likely caused by the outlier we observed in the begining of our summary analsyis. Apart from the wage variables, we also observe a strong positive relationship between *urban* and *density*, and a strong negative relationship between *west* and *pctmin80*. Given the demographic distribution of North Carolina, these correlations are also expected. Lastly, we observe that *density* is highly correlated with many variables in the dataset. We need to take note of this in our regression analysis so we control for omitted variable bias.

Next, we examine the bivariate relationships between our outcome variable (*crmrte*) and our independent variables. Here we see that *density*, *urban*, and *taxpc* have strong positive correlations with crime rate, while the certainty of punishment proxies (*prbarr* and *prbconv*) are negatively correlated with crime rate.

Modeling process

Initial Model

Following these observations we start investigating the key independent variables for our base model. The key factors we selected for our base model are: punshiment, density, and income. We chose these because we believe that they have a direct causal relationship with crime. Punishment serves as a deterrant for offenses. We believe that high punishment will likely lead to lower crime rates. We believe that higher density contributes to greater opportunity for crime in general, especially for certain types of crimes (e.g. face to face robery). By providing so many opportunities, density leads to higher crime rates. In regards to income, we think of two potential causal factors: 1. poverty (i.e. extreme low levels of income) may induce criminal behavior to satisfy basic needs; 2. wealth may attract crime due to high potential reward. Next, we proceed to investigate our variables in pursuit of the best proxy (or proxies) for these key factors in our data.

Punishment

We first examine the various measures provided on punishment: *prbarr* (probability of arrest), *prbconv* (probability of conviction), *prbpris* (probability of prison sentence), and *avgsen* (average prison sentence).

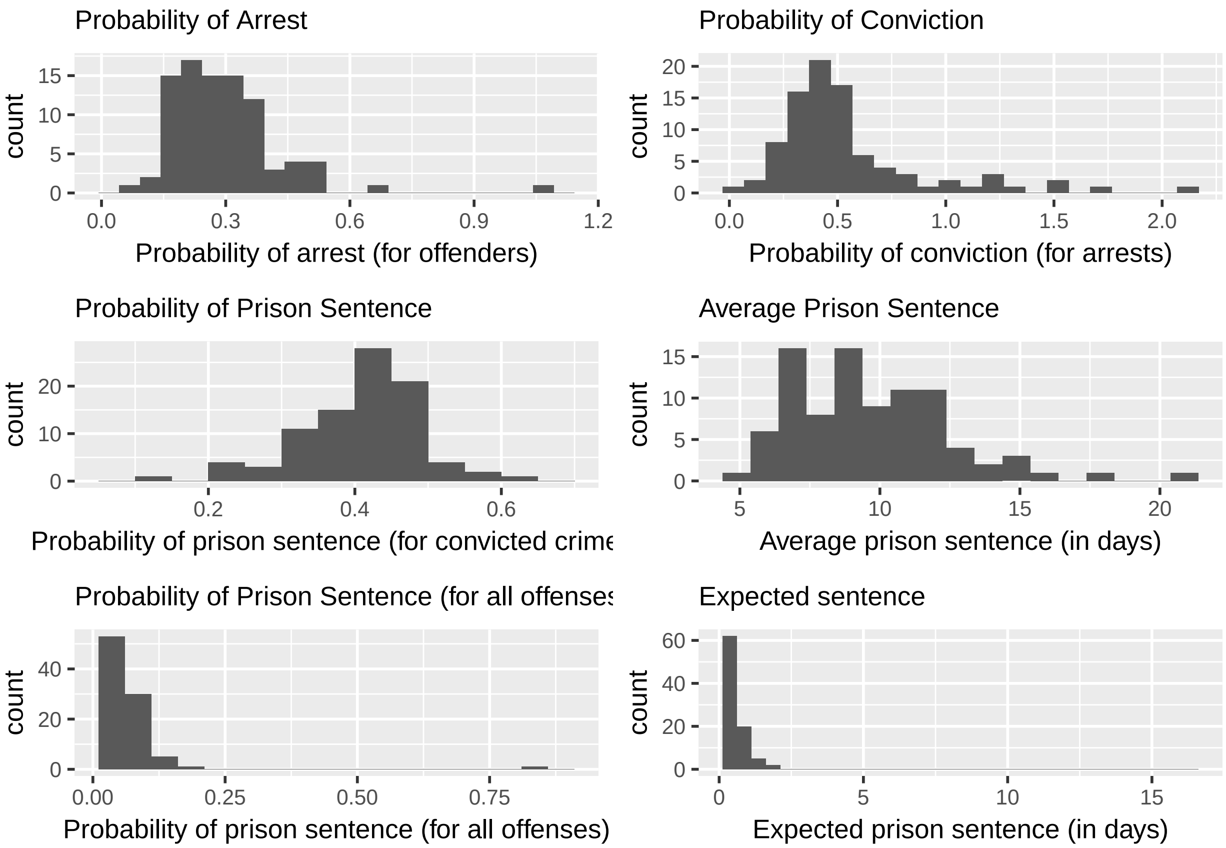
```
# Probablity of Arrest (of offenses)
p2 <- ggplot(data = f, aes(x = f$prbarr)) + geom_histogram(breaks = seq(min(f$prbarr) -
    0.1, max(f$prbarr) + 0.1, 0.05)) + xlab("Probability of arrest (for offenders)") +
    ggtitle("Probability of Arrest") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

# Probability of Conviction (of arrests)
p3 <- ggplot(data = f, aes(x = f$prbconv)) + geom_histogram(breaks = seq(min(f$prbconv) -
    0.1, max(f$prbconv) + 0.1, 0.1)) + xlab("Probability of conviction (for arrests)") +
    ggtitle("Probability of Conviction") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

# Probability of Prison Sentence (of convicted crimes)
p4 <- ggplot(data = f, aes(x = f$prbpris)) + geom_histogram(breaks = seq(min(f$prbpris) -
    0.1, max(f$prbpris) + 0.1, 0.05)) + xlab("Probability of prison sentence (for convicted crimes)") +
    ggtitle("Probability of Prison Sentence") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

# Average Length of Prison Sentence
p5 <- ggplot(data = f, aes(x = f$avgsen)) + geom_histogram(breaks = seq(min(f$avgsen) -
    1, max(f$avgsen) + 1, 1)) + xlab("Average prison sentence (in days)") +
    ggtitle("Average Prison Sentence") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

# Probability of Prison Sentence (of all offenses)
f$prbpunish <- f$prbarr * f$prbconv * f$prbpris
summary(f$prbpunish)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01064 0.03452 0.05267 0.06668 0.07301 0.81818
```

```
p6 <- ggplot(data = f, aes(x = f$prbpunish)) + geom_histogram(breaks = seq(min(f$prbpunish),
    max(f$prbpunish) + 0.1, 0.05)) + xlab("Probability of prison sentence (for all offenses)") +
    ggtitle("Probability of Prison Sentence (for all offenses)") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))

# Expected Prison Sentence (of all offenses)
f$expectsent <- f$prbpunish * f$avgsen
summary(f$expectsent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1169  0.3141  0.4577  0.7313  0.6638 16.9364
```

```
p7 <- ggplot(data = f, aes(x = f$expectsent)) + geom_histogram(breaks = seq(min(f$expectsent),
    max(f$expectsent), 0.5)) + xlab("Expected prison sentence (in days)") +
    ggtitle("Expected sentence") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
grid.arrange(p2, p3, p4, p5, p6, p7, nrow = 3)
```



Of the four provided punishment variables all exhibit some skew, but generally follow expected distributions. We note that 1 count of probability of arrest and 10 counts of probability of conviction are greater than 100%. Our first instinct was that these were erroneous datapoints, however the other data for these counties seemed correct. We believe this to be a quirk of the measurement system used. Perhaps convictions and arrests do not occur in the same year. There could be more convictions than arrests in a given year as convictions for crimes commited in prior years could not occur until the next year. It could also be that a same person arrested was convicted on multiple crimes. We will take note of these entries and in later regressions see whether these entries have high influence on our regression model.

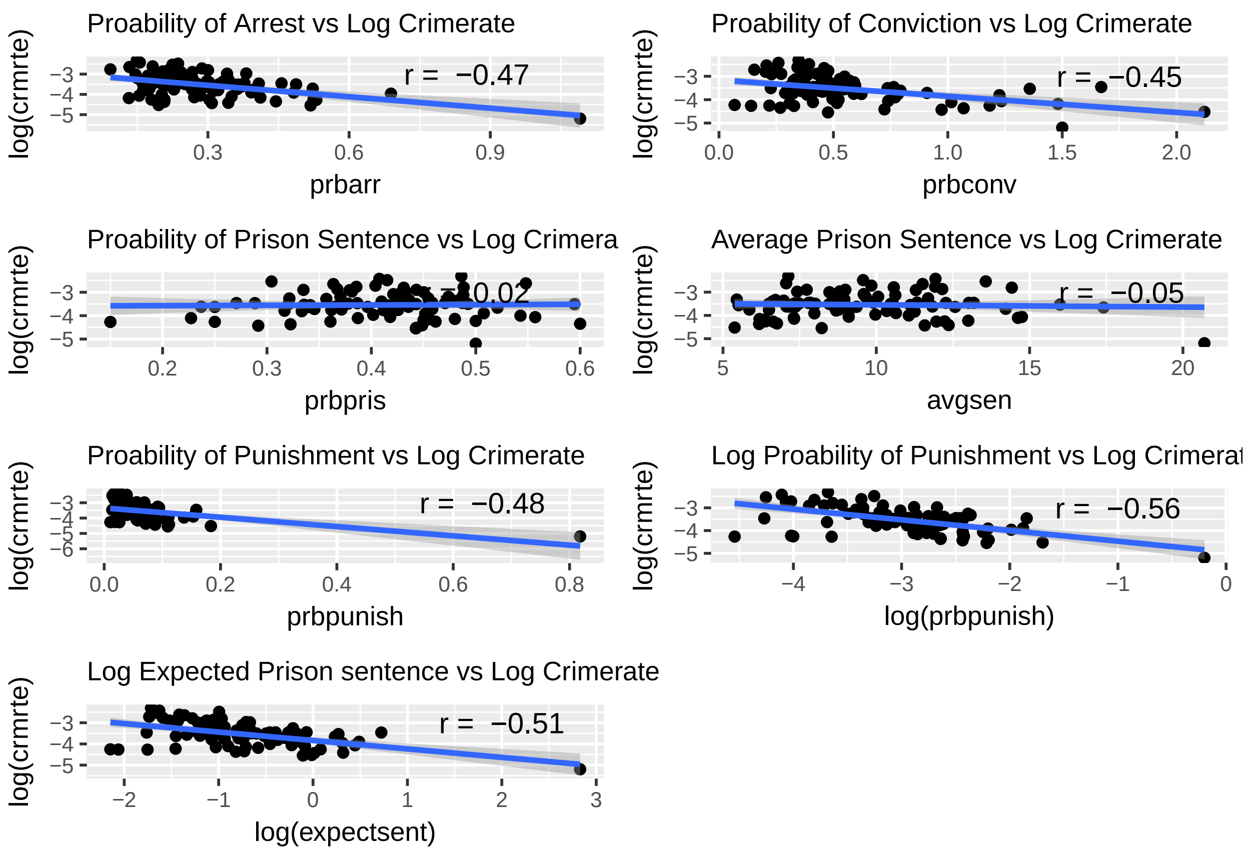
We also note that the three probability measures each capture one stage of the legal process, and none capture the full story from arrest to sentencing. While it is possible that arrest probability, conviction probability, and sentencing probability each have a unique impact on crime rate, it is also possible that the deterrant to crime is the general punishement certainty. Therefore, we create a new variable *prbpunish* which is simply the product of these three policing variables, to represent the probability of receiving a prison sentence per offense.

Moreover, we explore the deterrence of the expectation of prison sentence. This is simply the probability of receiving a prison sentence (i.e. *prbpunish*) multiplied by the average length of prison sentences (i.e. *avgsen*).

Now we plot these policing variables against the crime rate to see if a relationship exists and whether the relationship seems linear.

```
p8 <- ggplot(data = f, aes(prbarr, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Arrest vs Log Crimerate") +
    annotate(x = 0.85, y = -3, label = paste("r = ", round(cor(f$prbarr,
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
```

```
p9 <- ggplot(data = f, aes(prbconv, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Conviction vs Log Crimerate") +
    annotate(x = 1.75, y = -3, label = paste("r = ", round(cor(f$prbconv,
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p10 <- ggplot(data = f, aes(prbpris, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Prison Sentence vs Log Crimerate") +
    annotate(x = 0.5, y = -3, label = paste("r = ", round(cor(f$prbpris,
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p11 <- ggplot(data = f, aes(avgsen, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Average Prison Sentence vs Log Crimerate") +
    annotate(x = 18, y = -3, label = paste("r = ", round(cor(f$avgsen,
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p12 <- ggplot(data = f, aes(prbpunish, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Punishment vs Log Crimerate") +
    xlab("prbpunish") + annotate(x = 0.65, y = -3, label = paste("r = ",
    round(cor(f$prbpunish, log(f$crmrte)), 2)), geom = "text") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))

p13 <- ggplot(data = f, aes(log(prbpunish), log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Log Proability of Punishment vs Log Crimerate") +
    xlab("log(prbpunish)") + annotate(x = -1, y = -3, label = paste("r = ",
    round(cor(log(f$prbpunish), log(f$crmrte)), 2)), geom = "text") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))

p14 <- ggplot(data = f, aes(log(expectsent), log(crmrte))) +
    geom_point() + geom_smooth(method = lm) + ggtitle("Log Expected Prison sentence vs Log Crimerate") +
    xlab("log(expectsent)") + annotate(x = 2, y = -3, label = paste("r = ",
    round(cor(log(f$expectsent), log(f$crmrte)), 2)), geom = "text") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))

grid.arrange(p8, p9, p10, p11, p12, p13, p14, nrow = 4)
```



We observe that *prbarr* and *prbconv* both seem to have a negative relationship to crime rate. *Prbpris* and *avgsen* seems to have little relation to crime rate.

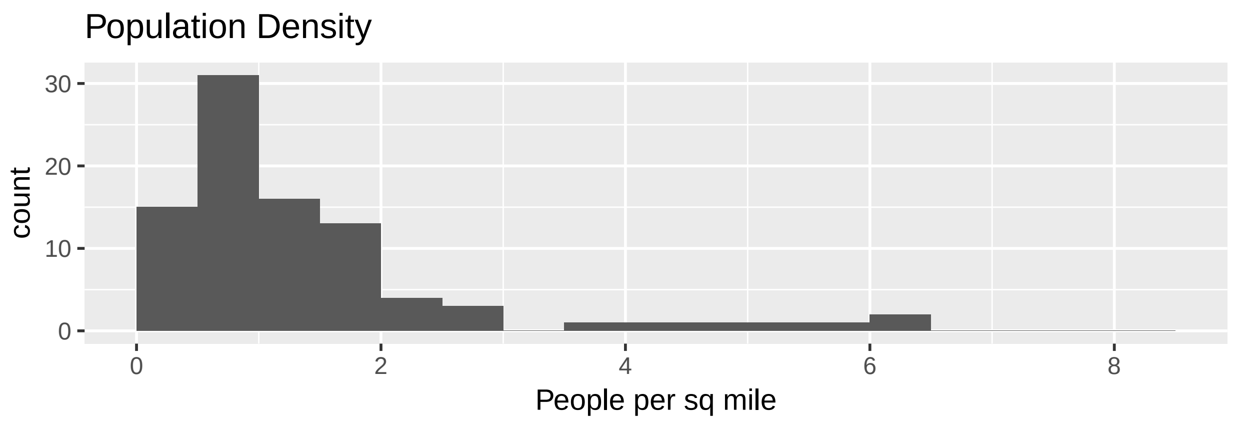
For our newly created *prbpunish* variable, there does seem to be a strong relationship with crime rate, however an outlier to the far right is obscuring the relationship. Taking the logarithm helps to clarify this relationship. Similarly our newly created *expectsent* variable also exhibits a strong negative relationship with crime once we apply the log transformation. This relationship is not as strong as *prbpunish*, potentially because the overall probability of punishment is so low, that average sentence did not add much deterrence.

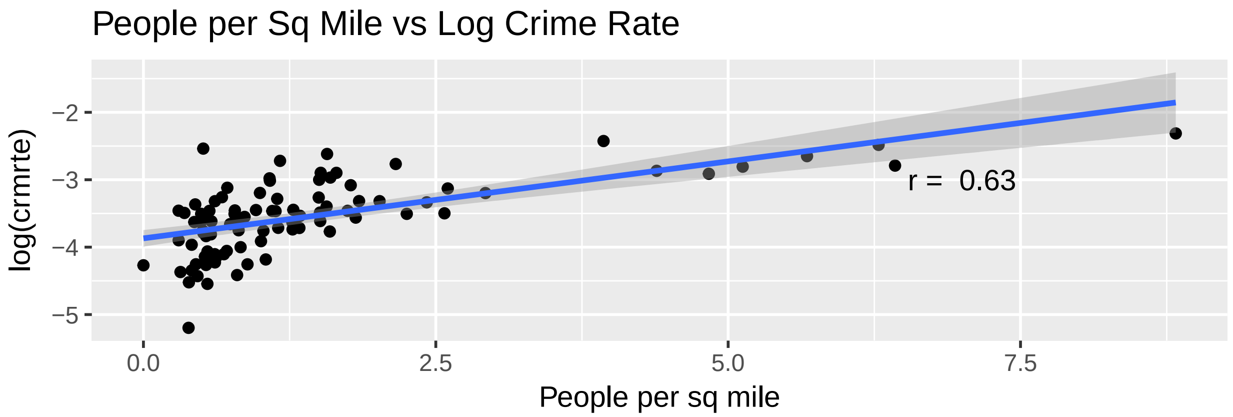
For now, we see that *prbpunish* (certainty of prison sentence for all offenses) seems to have the highest correlation to crime rate, and the log form seems to be a promising variable for our base model.

Density

Next we examine the density variable and its relationship to crime rate.

```r
# Density
p15 <- ggplot(data = f, aes(x = f$density)) + geom_histogram(breaks = seq(min(f$density),
    max(f$density) + 0.1, 0.5)) + xlab("People per sq mile") +
    ggtitle("Population Density")
p16 <- ggplot(data = f, aes(density, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + xlab("People per sq mile") + ggtitle("People per Sq Mile vs Log Crime Ra
    annotate(x = 7, y = -3, label = paste("r = ", round(cor(f$density,
        log(f$crmrte)), 2)), geom = "text")
grid.arrange(p15, p16, nrow = 2)
```





As briefly mentioned in our summary stats analysis, we find that the density values in the dataset extremely low, with a max value of merely 8.8 people per square mile. This is very hard to believe, especially given that the average population density of North Carolina in 1990 was 136 people per square mile according to the 2010 census data (https://web.archive.org/web/20111028061117/http://2010.census.gov/2010census/data/apportionment-dens-text.php). A discrepency of this size is also unlikely due to selection bias, since our dataset contains 90 out of 100 counties in North Carolina. Looking at the values, the most probable explanation may be that this variable is off by orders of mangnitude. However, this is just our speculation. Without more information from the data author, we need to be treat this variable with caution and be extremely careful using and interpreting this in our model.

When looking at the relationship between density and log of crime rate, the strong positve correlation is quite apparent. The relationship also seems relatively linear, therefore no transformation is required

Income and Wealth

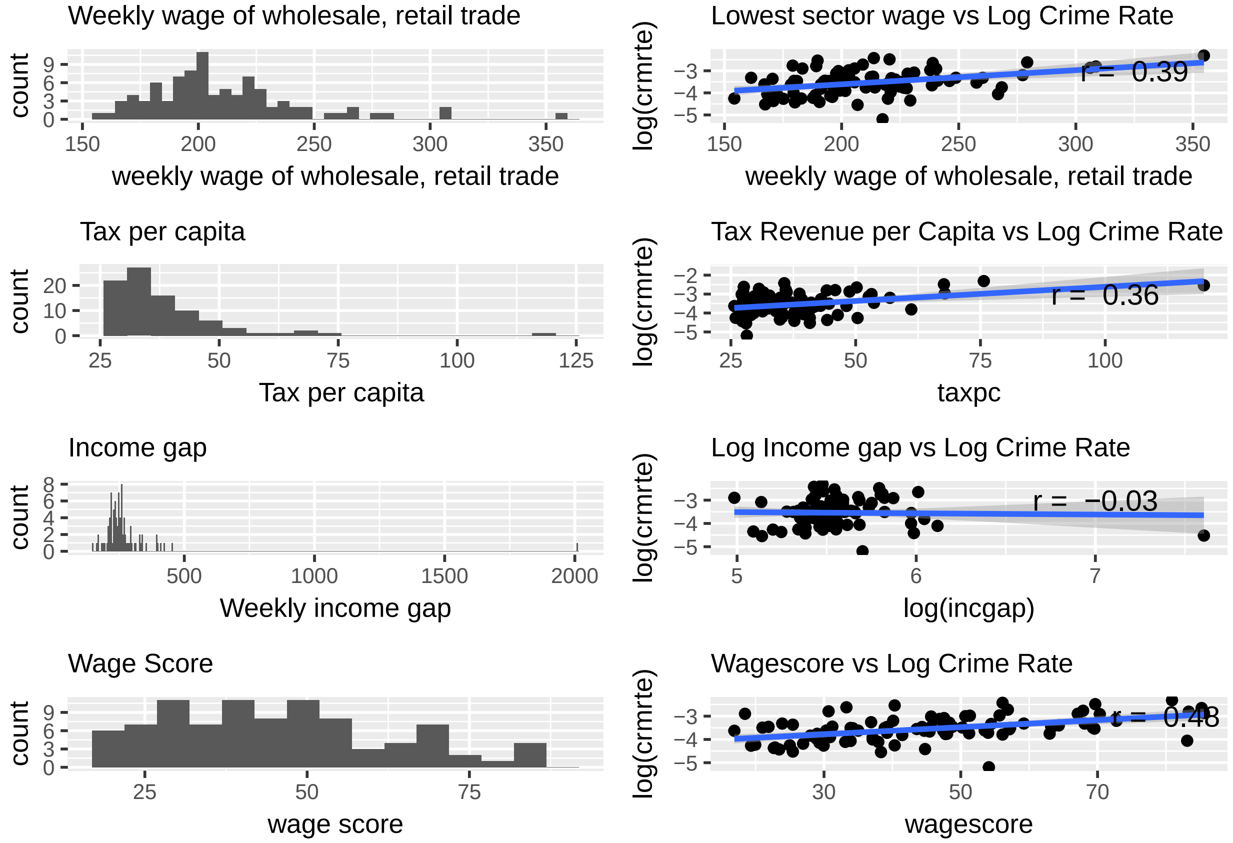
Our last base model factor is income. As explained in our reasoning for selecting this factor, we aim to find proxies for proverty and affluence. However, we do not have a perfect representation of income in our data. For imperfect proxies, we can use the wage of the lowest earning sector to represent poverty. This variable is *wtrd* (weekly wage of wholesale, retail trade). For affluency, *taxpc* is also a potential proxy because tax is tied to income and expenditure. We expect that financially well-off counties generate more tax payments, since a large portion of the tax is a porpotional to residents' income and expenses. We believe affluency creates a high reward for criminal behavior and therefore contributes to high crime rates. So, we expect *taxpc* to have a positive impact on crime rate.

Moreover, we can also create a proxy for pay-gap, by taking the difference of highest paying sector and the lowest paying sector of each county. Our hypothesis is that higher income gaps lead to more crime.

Lastly, in an attempt to proxy general income level more closely we also create a variable called "*wagescore*". Here we rank the counties for each wage variable, and take an average of the 9 ranks as the "wagescore". Without any information about how labor force is distributed across the different sectors, this is at least a rough proxy for the general income level of the county relative to other counties. The higher the score, the higher the income level.

```
# Lowest earning sector: weekly wage of wholesale, retail
# trade
p23 <- ggplot(data = f, aes(x = f$wtrd)) + geom_histogram(breaks = seq(min(f$wtrd),
    max(f$wtrd) + 10, 5)) + xlab("weekly wage of wholesale, retail trade") +
    ggtitle("Weekly wage of wholesale, retail trade") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
p24 <- ggplot(data = f, aes(wtrd, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + xlab("weekly wage of wholesale, retail trade") +
    ggtitle("Lowest sector wage vs Log Crime Rate") + annotate(x = 325,
    y = -3, label = paste("r = ", round(cor(f$wtrd, log(f$crmrte)),
        2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

## Proxy for affluence
p27 <- ggplot(data = f, aes(x = f$taxpc)) + geom_histogram(breaks = seq(min(f$taxpc),
    max(f$taxpc) + 10, 5)) + xlab("Tax per capita") + ggtitle("Tax per capita") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))
p28 <- ggplot(data = f, aes(taxpc, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + xlab("taxpc") + ggtitle("Tax Revenue per Capita vs Log Crime Rate") +
    annotate(x = 100, y = -3, label = paste("r = ", round(cor(f$taxpc,
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

## Proxy for income gap
f$incgap <- apply(f[, 15:23], 1, max) - apply(f[, 15:23], 1,
    min)
p29 <- ggplot(data = f, aes(x = f$incgap)) + geom_histogram(breaks = seq(min(f$incgap),
    max(f$incgap) + 10, 5)) + xlab("Weekly income gap") + ggtitle("Income gap") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))
p30 <- ggplot(data = f, aes(log(incgap), log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + xlab("log(incgap)") + ggtitle("Log Income gap vs Log Crime Rate") +
    annotate(x = 7, y = -3, label = paste("r = ", round(cor(log(f$incgap),
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

## Proxy for general wage level (higher rank = lower wage)
f$wagescore <- (rank(f$wcon) + rank(f$wtuc) + rank(f$wtrd) +
    rank(f$wfir) + rank(f$wser) + rank(f$wmfg) + rank(f$wfed) +
    rank(f$wsta) + rank(f$wloc))/9
p31 <- ggplot(data = f, aes(x = f$wagescore)) + geom_histogram(breaks = seq(min(f$wagescore),
    max(f$wagescore) + 10, 5)) + xlab("wage score") + ggtitle("Wage Score") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))
p32 <- ggplot(data = f, aes(wagescore, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + xlab("wagescore") + ggtitle("Wagescore vs Log Crime Rate") +
    annotate(x = 80, y = -3, label = paste("r = ", round(cor(f$wagescore,
        log(f$crmrte)), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
```

```r
grid.arrange(p23, p24, p27, p28, p29, p30, p31, p32, nrow = 4)
```



From the analysis, we see that both of the wage factors *wtrd* and *wfed* as well as *taxpc* have a positive relationship with log of crime rate. The relationships seem loosely linear and therefore we consider including them in our base models without transformation. However, both wage factors are highly correlated with density. *Taxpc* is still correlated with density but much less so than the wage variables, and may be a better explanatory variable in our model. Further *taxpc* seems to be a more generalizable as a causal variable.

For the variables we created, *incgap* variable does not seem to exhibit a significant relationship with crime rate at all. We looked at the log transformation to try to control the outlier impact, but even then the correlation is virtually non-existant. We therefore do not proceed to use this variable in our base model. On the other hand, *wagescore* showed rather significant correlation with log of crime rate (highest correlation among all income variables), therefore we can also consider adding this to our model

Between *taxpc* and *wagescore*, we decided to first use *taxpc* in our base model, because result would be much easier to interpret. However, we will consider using *wagescore* instead if *taxpc* does not turn out to be statistically significant.

Base model

Guided by this initial exploratory analysis, we begin to build our model. To recap, we select *log(prbpris)*, *density*, and *taxpc* as key explanatory variables:

$$log(crmrte) = \beta_0 + \beta_1 log(prbpunish) + \beta_2 density + \beta_3 taxpc + u$$

We now examine the output. Note: We will omit a detailed discussion of the CLM assumptions for this base model and save it for our more elaborate model with covariates. However, in our CLM assumption assemssment, we observed evidence for heteroskedasticity (MLR5), and therefore we opt to use heteroskedasticity-robust standard errors for our regression table.

```
# Simple Model with Key Explanatory Variables
model_base <- lm(log(crmrte) ~ log(prbpunish) + density + taxpc,
    data = f)
(se.model_base = sqrt(diag(vcovHC(model_base))))
```

(Intercept) log(prbpunish) density taxpc 0.370423732 0.113892615 0.040307983 0.005721095

```
## plot(model_base)
stargazer(model_base, header = FALSE, type = "latex", omit.stat = "f",
    float = FALSE, se = list(se.model_base), star.cutoffs = c(0.05,
        0.01, 0.001))
```

	Dependent variable:
	log(crmrte)
log(prbpunish)	-0.301^{**}
	(0.114)
density	0.168^{***}
	(0.040)
taxpc	0.005
	(0.006)
Constant	-4.853^{***}
	(0.370)
Observations	90
R^2	0.536
Adjusted R^2	0.520
Residual Std. Error	0.380 (df = 86)
Note:	*p<0.05; **p<0.01; ***p<0.001

The coefficients of some of the estimators are statistically significant. *Log(prbpunish)* and *density* are significant at the 0.1% level and *taxpc* is not statistically significant. The signs of the coefficients are also in accordance with our hypothesis above: certainty of punishment is a deterrant for crime and has a negative coefficient, while density and taxpc increase criminal opportunity and reward and have positive coefficients.

Because *taxpc* did not turn out to be statistically significant, we decided to try using *wagescore* as a control for general income level instead. Our alternate base model becomes:

$$log(crmrte) = \beta_0 + \beta_1 log(prbpunish) + \beta_2 density + \beta_3 wagescore + u$$

We now examine the of this alternative base model:

```
# Simple Model with Key Explanatory Variables
model_base_2 <- lm(log(crmrte) ~ log(prbpunish) + density + wagescore,
    data = f)
(se.model_base_2 = sqrt(diag(vcovHC(model_base_2))))
```

(Intercept) log(prbpunish) density wagescore 0.305148483 0.112337171 0.030721909 0.002926853

```
## plot(model_base)
stargazer(model_base_2, header = FALSE, type = "latex", omit.stat = "f",
    float = FALSE, se = list(se.model_base_2), star.cutoffs = c(0.05,
        0.01, 0.001))
```

	Dependent variable:
	log(crmrte)
log(prbpunish)	−0.335**
	(0.112)
density	0.130***
	(0.031)
wagescore	0.006*
	(0.003)
Constant	−5.015***
	(0.305)
Observations	90
R^2	0.550
Adjusted R^2	0.534
Residual Std. Error	0.375 (df = 86)
Note:	*p<0.05; **p<0.01; ***p<0.001

In this alternative base model, we are glad to see that all of our explanatory variables are significant, with heteroskedasticity-robust standard errors. We also found that our adjusted r2 actually increased from our original base model, which shows that we are able to account for more of the variations on our dependent variable.

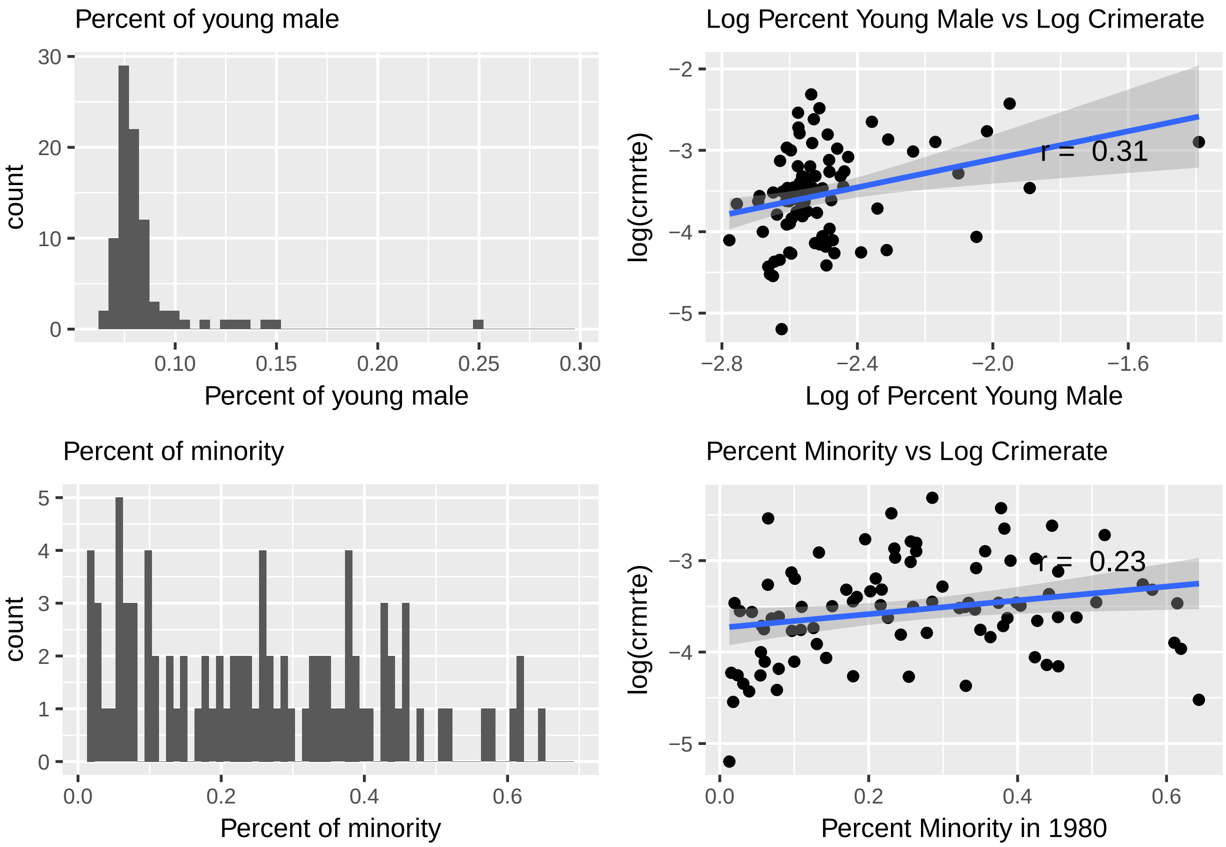
Statistical significance is a good sign, but we must also consider the practical significance of these coefficients. When we look at the coefficients in detail, ceteris paribus, we see that a 1 percent increase in probability of prison sentence (for all offenses), is associated with 0.335 percent decrease in crime rate, or in other word, 10% increase in probability of punishment is associated with 3.35% decrease in the crime rate. If proven causal, tightening law enforcement and increasing the probabiity of prison sentences would be a very relevant policy recommendation for local governments. For *density*, our model suggest that, again ceteris paribus, an increase of 1 person per square mile is associated with a 13% percent increase in crime rate. If the crime rate were 10% and the density increased by 1 person per square mile, we would expect, all else held constant, the crime rate would increase 13% or 1.3 percentage points, to 11.3%. Note that from our EDA we have doubts about the density variable and speculated that it might instead be in units of 100 people per square mile. In which case, the effect size would be smaller. The effect of *wagescore* is a lot more obscure. Here we see that ceteris paribus, a wagescore increase of 1 (i.e. on average, the wage rank of a county goes up by 1) is associated with a 0.6% increase in crime rate. This is not very practically significant, and since the wage score is a relative measure, the policy implication is also rather obscure. We use this variable primarily as a control variable so that we can make causal inferences on other variables.

Considering the model as a whole, the adjusted R^2 is 0.534 so the model explains about 53% of the variability in crime rate. This seems like a good first attempt but there are several covariates that could be added to improve the model's accuracy and reduce omitted variable bias.

Adding Covariates

We now expand our model to include some other covariates to control the variation across counties and get a better understanding of potential causal effects. We did not include demographic factors in our base model because we believed that the causal link with crime rate was unclear. To control for the variation brought about by demographic differences, we are interested in including *pctymle* (percent of young male) and *pctmin80* (percent of minority in 1980). In our initial correlation matrix, both of these factors exhibit some positive correlation with crime rate, and virtually no correlation with the predictor variables in our base model. We will examine these now.

```r
# Percent Young Male
p31 <- ggplot(data = f, aes(x = f$pctymle)) + geom_histogram(breaks = seq(min(f$pctymle),
    max(f$pctymle) + 0.05, 0.005)) + xlab("Percent of young male") +
    ggtitle("Percent of young male") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
p33 <- ggplot(data = f, aes(log(pctymle), log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Log Percent Young Male vs Log Crimerate") +
    xlab("Log of Percent Young Male") + annotate(x = -1.7, y = -3,
    label = paste("r = ", round(cor(log(f$pctymle), log(f$crmrte)),
        2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
# Percent Minority
p34 <- ggplot(data = f, aes(x = f$pctmin80)) + geom_histogram(breaks = seq(min(f$pctmin80),
    max(f$pctmin80) + 0.05, 0.01)) + xlab("Percent of minority") +
    ggtitle("Percent of minority") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
p35 <- ggplot(data = f, aes(pctmin80, log(crmrte))) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Percent Minority vs Log Crimerate") +
    xlab("Percent Minority in 1980") + annotate(x = 0.5, y = -3,
    label = paste("r = ", round(cor(f$pctmin80, log(f$crmrte)),
        2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
grid.arrange(p31, p33, p34, p35, nrow = 2)
```



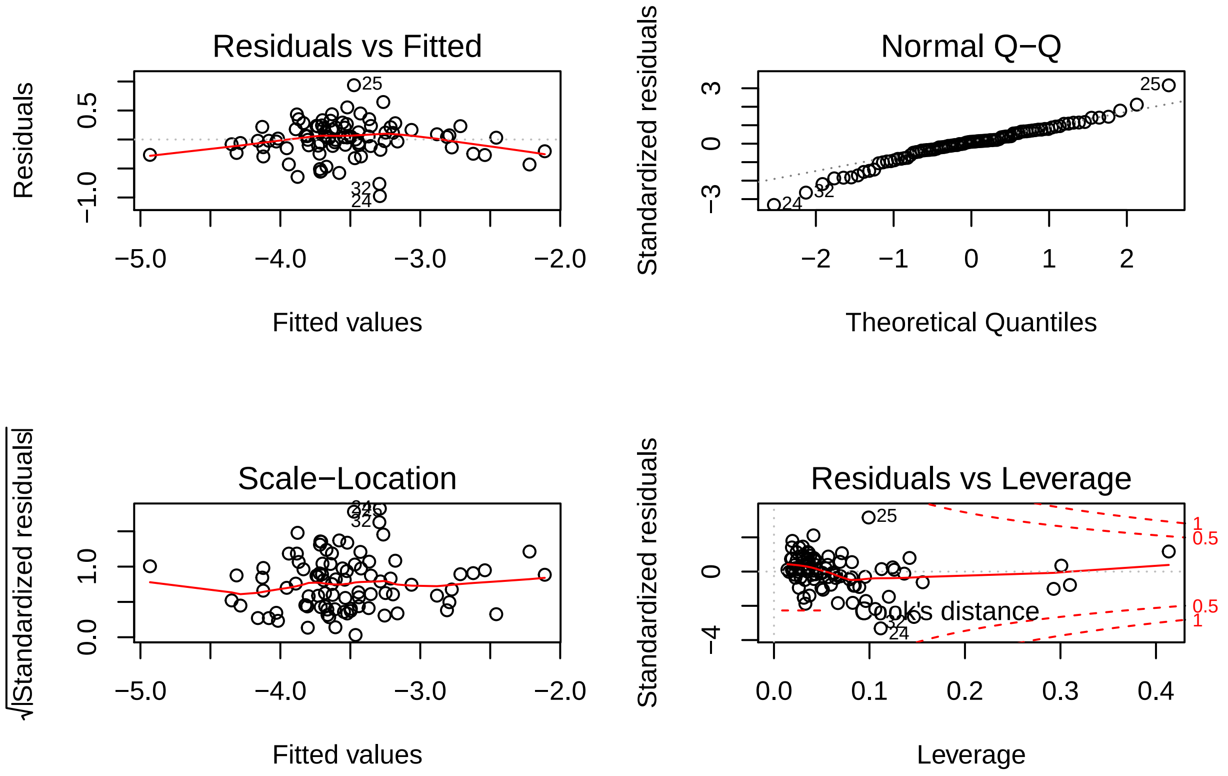
We can see that the distribution of percent young male is concentrated between 0 and ~6% with only a handful of counties skewing far to the right. When we look at the relationship between this variable and crime rate, we see that the handful of counties with a high percent of young males is driving the positive correlation. For the rest of the data points, the correlation seems rather unclear. Taking the log of the variable clarified the relationship further, and it does seem that a small positive relationship exists.

For percent minority, the distribution is more spread out and there does seem to be a mild linear relationship between *pctmin80* and the log of crime rate. We should emphasize that this does not imply that minorities themselves have a higher crime rate. There could be many ommited variables that drive the demographic composition of a county. The relationship of these omitted variables with the crime rate may be what cause the positive correlation we see here. Additionally, different demographics may have different reactions to policing measures. Therefore, it is something that the model should account for and may prove of use to local governments. Our new model is now specified as

$$log(crmrte) = \beta_0 + \beta_1 log(prbpunish) + \beta_2 density + \beta_3 wagescore + \beta_4 log(pctymle) + \beta_5 pctmin80 + u$$

We now show this new model compared to the intial, base model.

```r
model_plus <- lm(log(crmrte) ~ log(prbpunish) + density + wagescore +
    log(pctymle) + pctmin80, data = f)
model_plus_2 <- lm(log(crmrte) ~ log(prbpunish) + density + wagescore +
    pctmin80, data = f)
par(mfrow = c(2, 2))
plot(model_plus, which = 1)
plot(model_plus, which = 2)
plot(model_plus, which = 3)
plot(model_plus, which = 5)
```



```
se.model_plus = sqrt(diag(vcovHC(model_plus)))
stargazer(model_base_2, model_plus, header = FALSE, type = "latex",
    omit.stat = "f", float = FALSE, se = list(se.model_base_2,
        se.model_plus), star.cutoffs = c(0.05, 0.01, 0.001))
```

	Dependent variable:	
	log(crmrte)	
	(1)	(2)
log(prbpunish)	-0.335^{**}	-0.367^{***}
	(0.112)	(0.099)
density	0.130^{***}	0.113^{**}
	(0.031)	(0.035)
wagescore	0.006^{*}	0.009^{**}
	(0.003)	(0.003)
log(pctymle)		0.264
		(0.206)
pctmin80		1.213^{***}
		(0.209)
Constant	-5.015^{***}	-4.843^{***}
	(0.305)	(0.696)
Observations	90	90
R^2	0.550	0.695
Adjusted R^2	0.534	0.677
Residual Std. Error	0.375 (df = 86)	0.312 (df = 84)
Note:	*p<0.05; **p<0.01; ***p<0.001	

The new model exhibits a lower Akaike Information Criteria 53.5851223 compared to the base model 84.5524878, indicating the new model is of better quality. This is in spite of adding more parameters to the model; the tradeoff between complexity and quality seems to be worthwhile. Further, the new model exhibits an adjusted R^2 of 0.677, about 14 percentage points higher than the base model. This is a nontrivial improvement, and means that our 5 independent variables can explain more than 67% of the variability in crime rate. *pctmin80* exhibits statistical significance at the 0.1% level. *log(pctymle)* does not exhibit any significance. We proceeded to see if we should then drop the variable from our model and created model_plus_2. Here we see that after dropping the variable, we actually have a higher AIC of 53.8688575, which shows that dropping this variable made the model worse. Therefore, we continue use our orginal model_plus as our main model. We will treat *log(pctymle)* as a control variable.

Now we look at practical significance of our covariate model. Since an increase of 1 unit in pctmin80 does not make realistic sense, we look at what happens when we increase pctmin80 by 1 percentage point (i.e. 0.01 unit). Holding other factors constant, a 1 percentage point increase in the percent minority in a county would be associated with a 1.213% percent increase in the crime rate. In other word, if the crime rate were 10% and the percentage minority in 1980 were to increase by 1 percentage point in the county, all else equal, we would expect that the crime rate would increase 1.213% or 0.1213 percentage points, to 10.1213%. As we mentioned when first exploring this variable, there could be several omitted variables that explain this relationship. However, for local governments this could still be a useful guidepost. By engaging their minority communities and further studying their needs, local governments could help to lower the crime rate. We omit the examination of practical significance of *log(pctmle)* here since it is not statistically significant. Also, the practical significance of the other variables (i.e. log(prbpunish), density and wagescore) did not change significantly from the base model, so we omit the discussion here.

Now, we dive into the discussion to assess all 6 CLM assumptions, and see whether any of these assumptions are violated in the current model, and whether we need to change the model in response.

- MLR.1. Linear population model: We don't have to check the linear population model, because we haven't constrained the error term. As long as we do not constrain the error term, we can always fit a linear model to the population.
- MLR.2. Random Sampling: To check random sampling, we need background knowledge of how the data was collected. Unfortunately, we don't have much information on how the data was selected. Further, we do not have information about how the counties were chosen. There are 100 counties in North Carolina and we have 90, but we do not know how they were selected.
- MLR.3. No perfect multicollinearity: We do not have perfect multicollinearity in our model, otherwise our statistical software will show a warining. Further, we have specified the model such that none of the independent variables is a linear combination of another. This ensures that we do not have perfect multicollinearity.
- MLR.4. Zero-conditional mean: We see in the residual vs. fitted values plot that for most of X values, residual seems to be averaging around 0, with the exeption of values on the very left and very right, where there are very few data points. For the middle area where most of the data is, the red line is also rather flat and at 0. Therefore, it seems that this assumption is not violated.
- MLR.5. Homoskedasticity: According to the residual vs. fitted values plot, we do not have homoskedasticity because the data band is considerably thicker towards the middle range, and thiner on both ends, indicating that variance of the residuals changes with X. The assumption seems violated. We come to this same conclusion by examining the scale-location plot. We see that the red line is not flat. To address heteroskedasticity, we will use heteroskedasticity-robust standard errors for all regression tables and tests.
- MLR.6. Normality of errors: The QQ plot shows deviation from normal distribution towards the right hand side. However, we do have a large dataset (n=90), which means we can likely rely on Central Limit Theorm, which tells us that our estimators will have a normal sampling distribution.

Full model with all covariates

Next, to assess the robustness of this new model we compare it to a control model that utilizes almost all the other available variables. It is worth noting the few variables that we intentionally excluded:

1. *polpc* (police per capita). We excluded this because we believe that more police are often assigned to counties with higher crime rate. Therefore, factors that lead to high crime rates may also lead to high police per capita. Having this variable as an explanatory variable will absorb some of the causal effect that we are studying.

2. *prbarr* (probability of arrest), *prbconv* (probability of conviction), *prbpris* (probability of prison sentence). We excluded these three because our *prbpunish* is calculated from these three variables. Including all four variables would introduce perfect multicollinearity.

3. All of the wage variables. Because *wagescore* is calculated from these variables we exluded them to avoid issues of collinearity.

We now examine the model. Note: We also examined this control model for violations of the CLM assumptions. We identified a violation of homoskedasticity as was the case with the other models. We proceed using robust standard errors to correct for this.

```
model_control <- lm(log(crmrte) ~ log(prbpunish) + density +
    wagescore + pctmin80 + taxpc + log(pctymle) + avgsen + west +
    urban + mix, data = f)
se.model_control = sqrt(diag(vcovHC(model_control)))
stargazer(model_base_2, model_plus, model_control, header = FALSE,
    float = FALSE, type = "latex", omit.stat = "f", se = list(se.model_base_2,
        se.model_plus, se.model_control), star.cutoffs = c(0.05,
        0.01, 0.001))
```

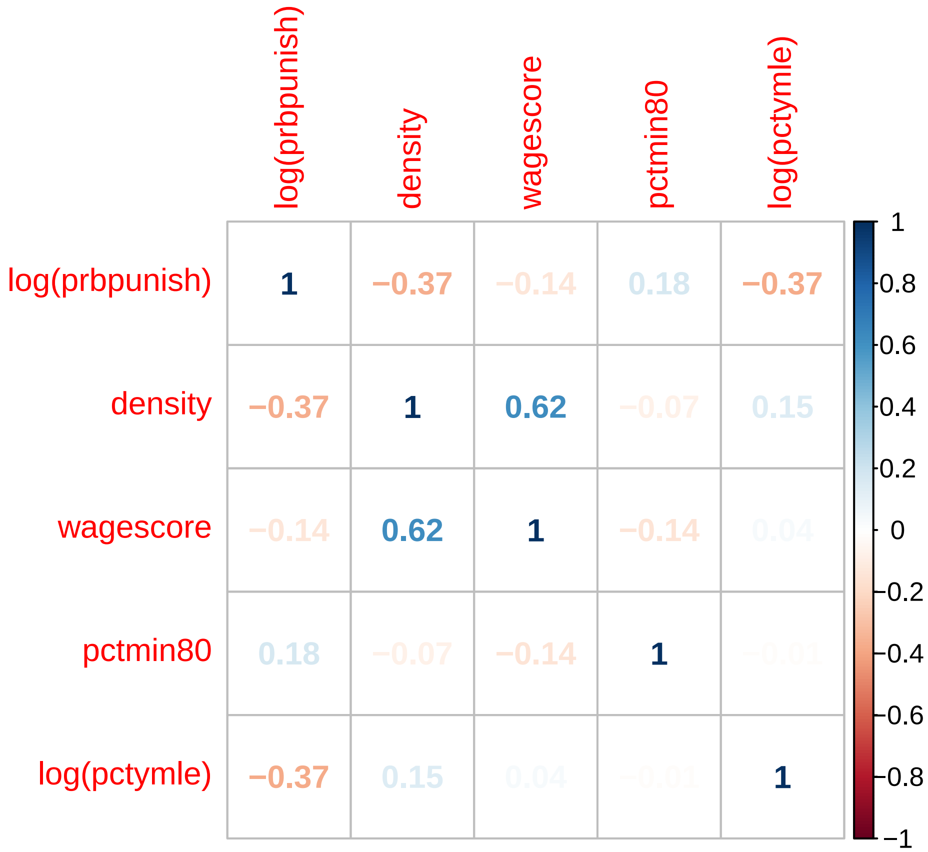
	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
log(prbpunish)	−0.335** (0.112)	−0.367*** (0.099)	−0.330*** (0.095)
density	0.130*** (0.031)	0.113** (0.035)	0.152* (0.066)
wagescore	0.006* (0.003)	0.009** (0.003)	0.007 (0.004)
log(pctymle)		0.264 (0.206)	0.331* (0.168)
avgsen			−0.007 (0.014)
west			−0.118 (0.108)
urban			−0.289 (0.253)
mix			−0.228 (0.590)
pctmin80		1.213*** (0.209)	1.026*** (0.299)
taxpc			0.005 (0.008)
Constant	−5.015*** (0.305)	−4.843*** (0.696)	−4.554*** (0.665)
Observations	90	90	90
R^2	0.550	0.695	0.723
Adjusted R^2	0.534	0.677	0.688
Residual Std. Error	0.375 (df = 86)	0.312 (df = 84)	0.306 (df = 79)
Note:	*p<0.05; **p<0.01; ***p<0.001		

From the regression table above, we see that adding the full range of covariates did not have a large impact on the sign or size of the coefficients in our first and second models. This indicates robustness in our modeling choice. Moreover, we note the full model has an AIC of 54.7572719, which is larger than 53.5851223 in our second model. This shows that adding all the covariates was not efficient and the improvment in the model's explanatory abilities did not justify the increase in complexity.

Omitted Variables:

Although our model is improved, it is not without bias. There are many variables omitted either because they are immeasurable or because they were not included in the data. We must attempt to account for the bias induced by these omitted variables. To do so, we must consider the correlation among our predictor variables. We show this below in a correlogram.

```r
g <- f[, c("prbpunish", "density", "wagescore", "pctmin80", "pctymle")]
g$prbpunish <- log(g$prbpunish)
g$pctymle <- log(g$pctymle)
names(g)[names(g) == "prbpunish"] <- "log(prbpunish)"
names(g)[names(g) == "pctymle"] <- "log(pctymle)"
N <- cor(g)
# corrplot.mixed(N, lower.col = 'black', number.cex = .7)
corrplot(N, method = "number")
```



	log(prbpunish)	density	wagescore	pctmin80	log(pctymle)
log(prbpunish)	1	−0.37	−0.14	0.18	−0.37
density	−0.37	1	0.62	−0.07	0.15
wagescore	−0.14	0.62	1	−0.14	0.04
pctmin80	0.18	−0.07	−0.14	1	−0.01
log(pctymle)	−0.37	0.15	0.04	−0.01	1

From the correlation graph, we see that fortunately, our independent variables are not very strongly correlated, with the exeption of wagescore and density. For the purpose of our reasoning on direction below, we need assume that these correlations are minimal, and regard them as zero.

Education

We believe that the education level of county residents is an omitted variable in our model. This could be measured by metrics such as mean or median years of education. We expect education level to be negatively related to the crime rate, i.e. $\beta_{educ} < 0$.

We believe that the most important bias that education has is on the *wagescore* variable. We focus on *wagescore* because it is a proxy for income level, and wealthy residents are more likely to afford more years of education. To deduce the direction of the omitted variable bias on *wagescore* due to education, we need to assume that the other independent variables are uncorrelated with each other, and uncorrelated with education. After these assumptions, we expect a strong positive correlation between *wagescore* and education (i.e. $\delta_{wagescore} > 0$), because a higher wagescore means higher income, and thus potentially higher education level.Therefore we expect a negative bias on $\beta_{wagescore}$.

Unemployment Rate

We also believe that unemployment rate is an important omitted variable in our model because it is a better proxy for the prevalence of poverty. We expect that poverty may induce criminal behavior to satisfy basic needs, and therefore $\beta_{unemploy} > 0$.

For this omitted variable, we believe that the biases also mainly lies with *wagescore*. Similar to our approach with education, we assume that other independent variables are uncorrelated with each other and with unemployment rate. After these assumptions, we believe that higher unemployment rate means lower average income, which should be associated with a lower *wagescore*. Therefore we expect $\delta_{wagescore} < 0$ and thus a negative bias on $\beta_{wagescore}$.

Social Cohesion

Communities with more cohesion are less likely to commit crimes against their neighbors and are likely more vigilant of criminal behavior. This could be proxied by measures like church attendance or number of homeowner's associations. We expect social cohesion to be negatively related to the crime rate, i.e. $\beta_{social} < 0$.

For this measure, we would like to focus our bias analysis on density. Assuming that all other predictor variables are uncorrelated with each other and with social cohesion, we expect that density and social cohesion to be positively related (i.e. $\delta_{density} > 0$). This is because we believe that residents are more likely to know each other and have more opportunities to interact with each other (increase social cohesion) in more densely populated counties. Therefore, we expect a negative bias on $\beta_{density}$ towards zero.

Home Ownership

The rate of home ownership by county would be of interest, since home ownership maybe a proxy for the stability of county residents. Homeowners have a greater sense of belonging to a community and so will be more vigilant and less likely to commit crimes against that community. Therefore we expext counties with more homeowners, rather than renters, to have a lower crime rate, i.e. $\beta_{homeownership} < 0$.

Here, we again want to focus on the bias on *wagescore*, because home ownership is highly tied to income. Assuming that all other predictor variables are uncorrelated with each other and with home ownership, we believe that wealthier counties will have higher home ownership because residents are more likely to be able to afford to buy homes instead of rent. Since wealth is positively tied to *wagescore*, we expect $\delta_{wagescore} > 0$, and therefore homeownership will have a negative bias on $\beta_{wagescore}$.

Drug Consumption

We also believe that drug consumption is an important omitted variable in our model. Many crimes are drug-related, and drug abuse often reduces inhibitions, which may encourage criminal behavior. This could be a measure like a drug arrests per capita. We expect drug consumption to be positively related to the crime rate $\beta_{drugs} > 0$.

For this variable, we want to focus on the impact on $\beta_{pctmin80}$, because minority communities historically have problems with drug abuse. Assuming all other variables are uncorrelated, we expect a positive relationship between drug consumption and percent minority, i.e. $\delta_{pctmin80} > 0$, and therefore, we believe that drug use will have a positive bias on $\beta_{pctmin80}$

In light of all these considerations, we reassess the following about our models: - the effect of *wagescore* is likely understated due to negative bias exerted by several omitted variables - the effects of *density* and *pctmin80* may be slightly overstated as there are some omitted variables that exert upward bias on them

Conclusion

Based on the analysis above, we offer the following conclusions to the campaign and local governments:

- A high probability of punishment lowers the crime rate, potentially acting as a deterrent to would be criminals. Communities experiencing higher levels of crime should expand their law enforcement efforts, especially with an eye toward punishing criminals not merely arresting them. However, we would note that we found the length of prison sentence unrelated to the crime rate. This indicates that receiving the punishment, not necesarilly the severity of the punishment, deters criminals. Communities should still seek to punish criminals, but perhaps with shorter sentences.
- Population density seems strongly related to the crime rate. Law enforcement and policy efforts should focus more on cities. Local governments would be well served to study the needs of residents in these densely populated areas. Addressing their concerns and needs may help to lower the crime rate "at the source."
- Income (as proxied by *wagescore*) seems to be related to the crime rate. Areas with high wagescores may want to consider taking measures to mitigate criminals. For eaxmple, they could likely afford higher taxes for more law enforcement. Further, there could be a geographic component not yet considered. Poorer communities may go commit crimes in adjacent wealthier communities.
- Minorities seem to drive the crime rate in communities. Engaging with these often marginalized group to integrate them into the broader community may help to lower the crime rate. This is an area for further study by local governments.