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.

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)
```

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

```
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), ]
## 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.

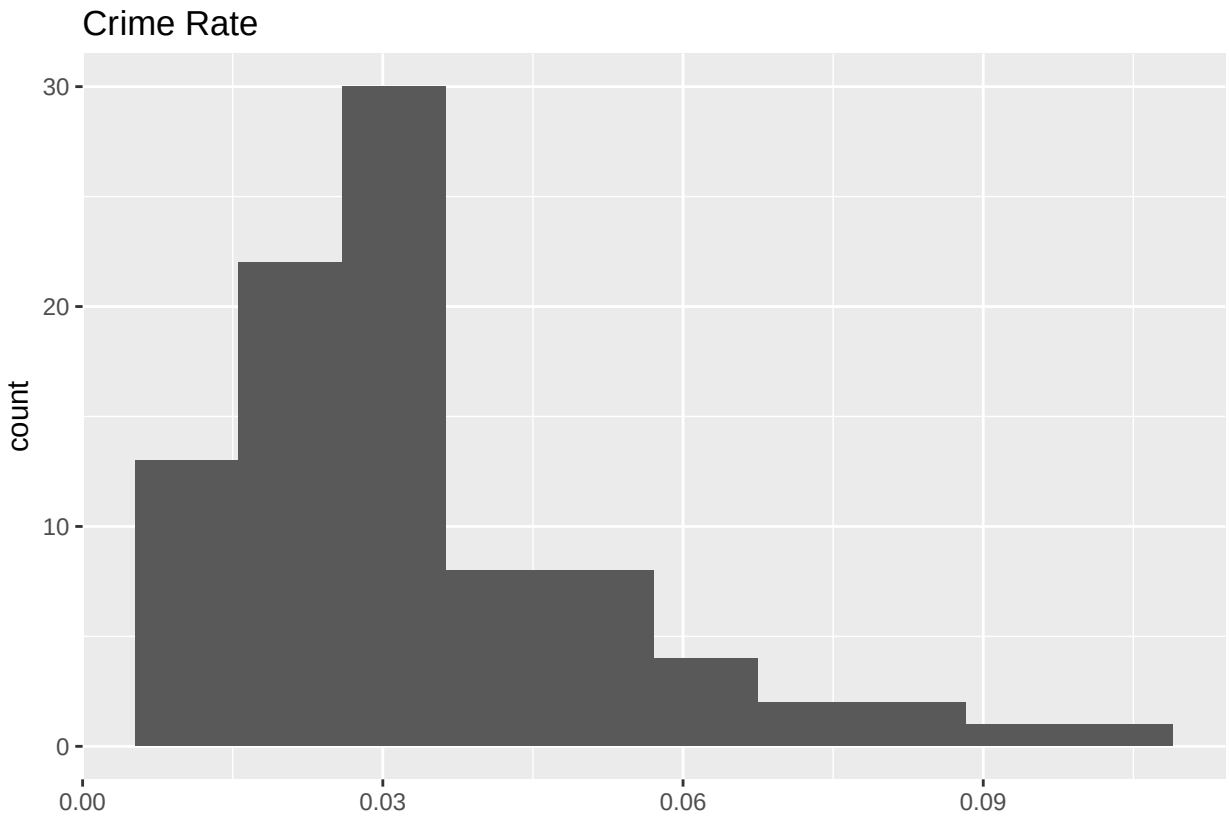
Intial Exploratory Data Analysis

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

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020927 0.029986 0.033400 0.039642 0.098966
```

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



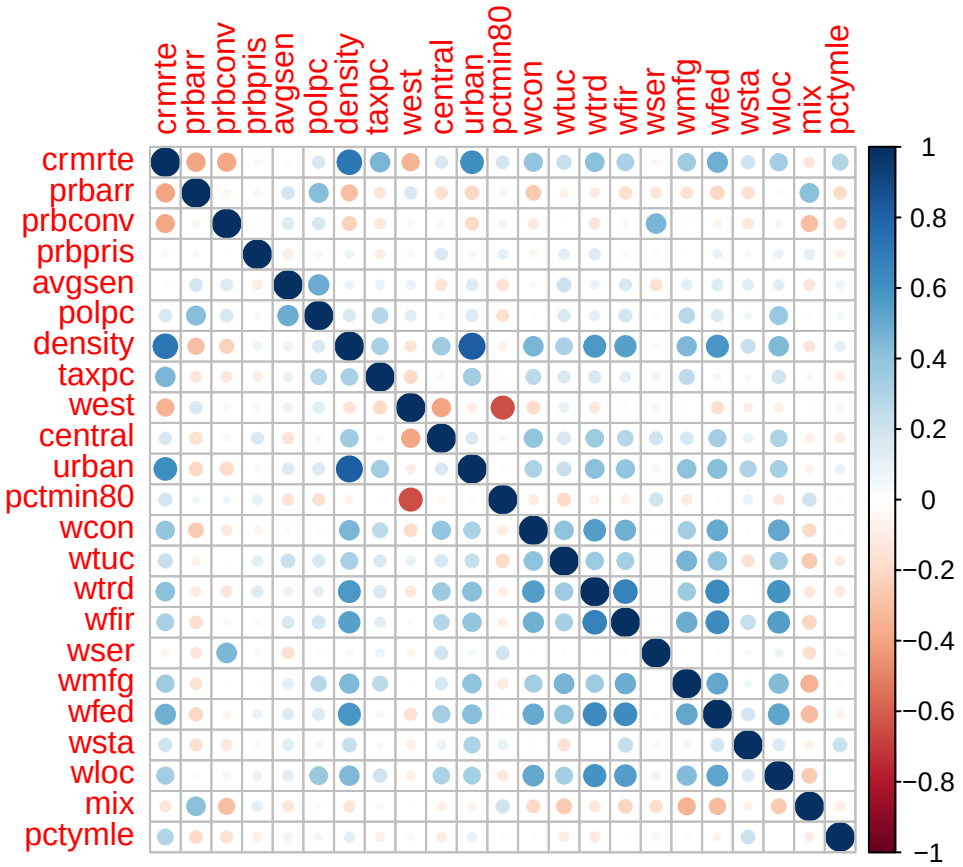
We note a range of 0.5% to 9.9% and a right skew in the data. This distribution is reasonable, with most counties experiencing a relatively low crime rate, and a handful of counties that are more crime-ridden. A quick cross check validates that in 2016, crime rate in North Carolina is 3109.7 per 100000 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.

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 pushment 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. Punshiment 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 and 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)
p1 <- 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)
p2 <- 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)
p3 <- 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
p4 <- 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)
```

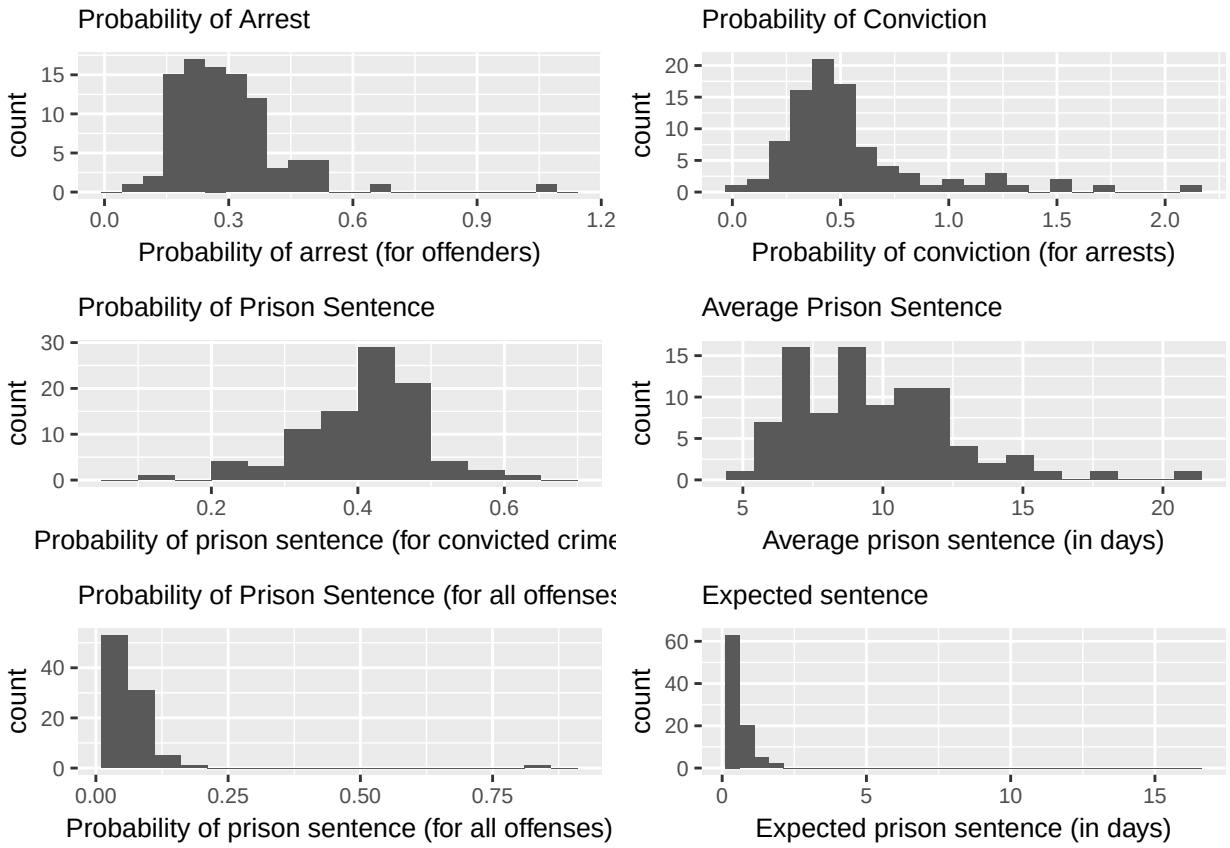
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01064 0.03465 0.05289 0.06668 0.07283 0.81818
```

```
p5 <- 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  0.3151  0.4549  0.7276  0.6555 16.9364
```

```
p6 <- 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(p1, p2, p3, p4, p5, p6, nrow = 3)
```



Of the four provided punishent 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. We will take note of these entries and in later regressions see whether these entries have high influence on our regression model.

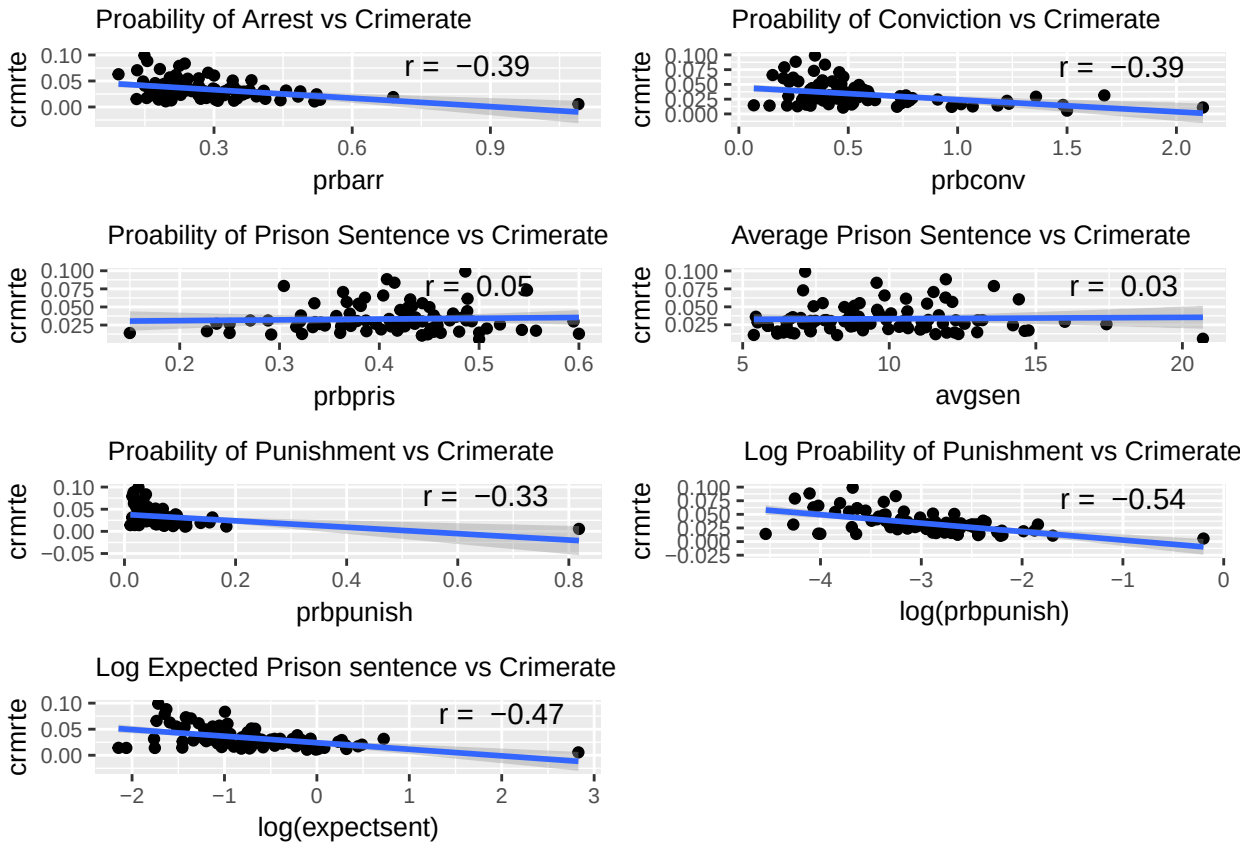
We also note that the three probability measures each capture one stage of the legal process, and none capture the full story from arrest to sentencing. While it is possible that arrest probability, conviction probability, and sentencing probability each have a unique impact on crime rate, it is also 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 detterence 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.

```
p13 <- ggplot(data = f, aes(prbarr, crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Arrest vs Crimerate") +
    annotate(x = 0.85, y = 0.08, label = paste("r = ", round(cor(f$prbarr,
        f$crmrte), 2)), geom = "text") + theme(text = element_text(size = 10),
```

```
    plot.title = element_text(size = 10))

p14 <- ggplot(data = f, aes(prbconv, crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Conviction vs Crimerate") +
    annotate(x = 1.75, y = 0.08, label = paste("r = ", round(cor(f$prbconv,
        f$crmrte), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p15 <- ggplot(data = f, aes(prbpris, crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Prison Sentence vs Crimerate") +
    annotate(x = 0.5, y = 0.08, label = paste("r = ", round(cor(f$prbpris,
        f$crmrte), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p16 <- ggplot(data = f, aes(avgsen, crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Average Prison Sentence vs Crimerate") +
    annotate(x = 18, y = 0.08, label = paste("r = ", round(cor(f$avgsen,
        f$crmrte), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p17 <- ggplot(data = f, aes(prbpunish, crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Proability of Punishment vs Crimerate") +
    xlab("prbpunish") + annotate(x = 0.65, y = 0.08, label = paste("r = ",
    round(cor(f$prbpunish, f$crmrte), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

p18 <- ggplot(data = f, aes(log(prbpunish), crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Log Proability of Punishment vs Crimerate") +
    xlab("log(prbpunish)") + annotate(x = -1, y = 0.08, label = paste("r = ",
    round(cor(log(f$prbpunish), f$crmrte), 2)), geom = "text") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))

p19 <- ggplot(data = f, aes(log(expectsent), crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Log Expected Prison sentence vs Crimerate") +
    xlab("log(expectsent)") + annotate(x = 2, y = 0.08, label = paste("r = ",
    round(cor(log(f$expectsent), f$crmrte), 2)), geom = "text") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))

grid.arrange(p13, p14, p15, p16, p17, p18, p19, 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.

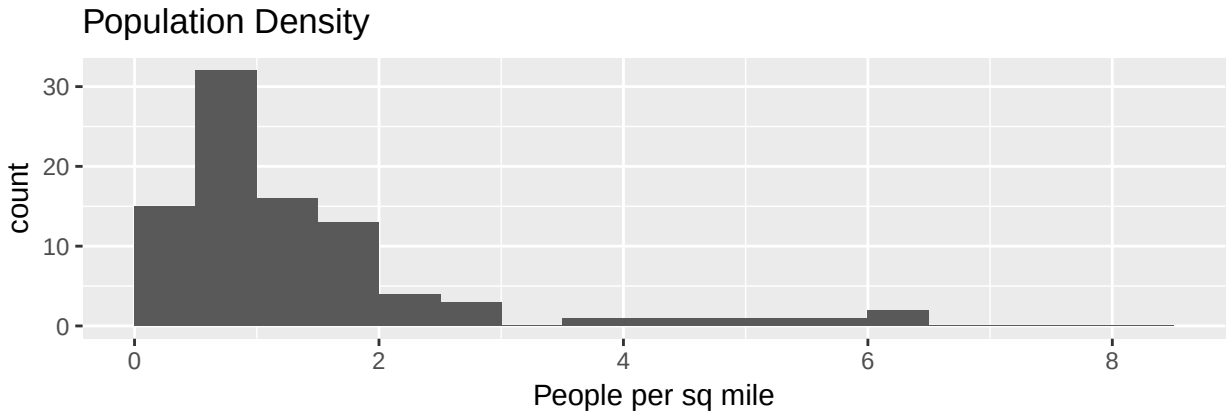
For our newly created *prbpunish* variable, there does seem to be a strong relationship with crime rate, however an outlier to the far right is obscuring the relationship. Taking the logarithm helps to clarify this relationship. Similarly our newly created *expectsent* variable also exhibits a strong 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.

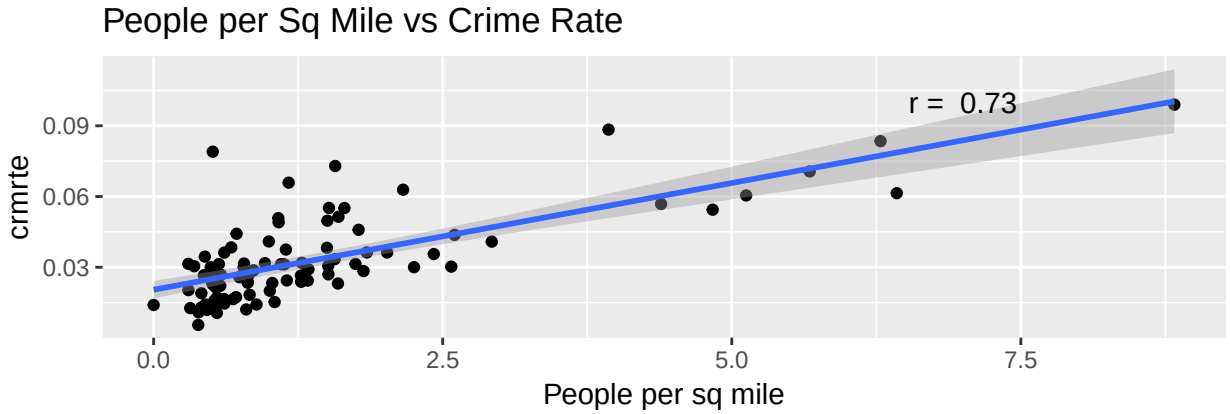
For now, we see that *prbpunish* (certainty of prison sentence for all offenses) seems to have the highest correlation to crime rate, and the log form seems to be a promising variable for our base model.

Density

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

```
# Density
p21 <- ggplot(data = f, aes(x = f$density)) + geom_histogram(breaks = seq(min(f$density),
    max(f$density) + 0.1, 0.5)) + xlab("People per sq mile") +
    ggtitle("Population Density")
p22 <- ggplot(data = f, aes(density, crmrte)) + geom_point() +
    geom_smooth(method = lm) + xlab("People per sq mile") + ggtitle("People per Sq Mile vs Crime Rate")
    annotate(x = 7, y = 0.1, label = paste("r = ", round(cor(f$density,
        f$crmrte), 2)), geom = "text")
grid.arrange(p21, p22, nrow = 2)
```





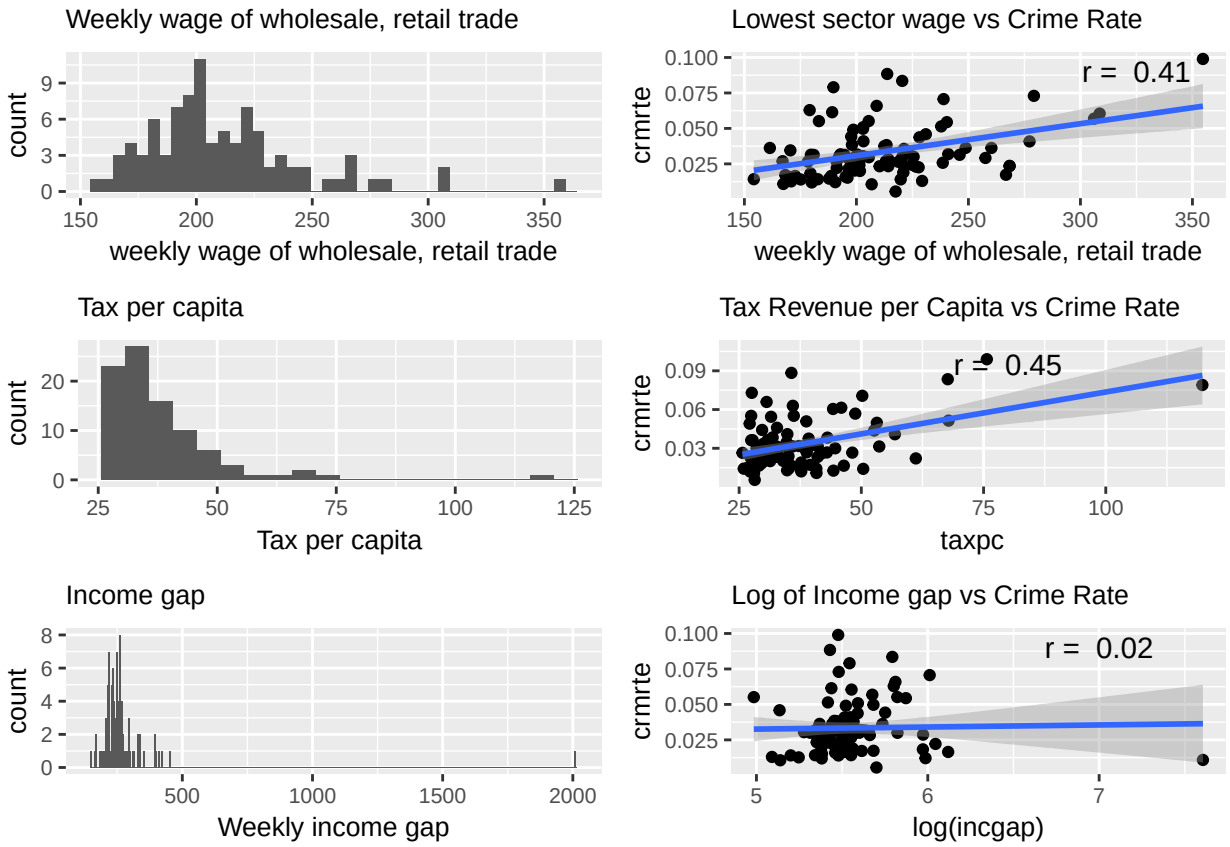
As briefly mentioned in our summary stats analysis, we find that the density values in the dataset extremely low, with a max value of merely 8.8 people per square mile. This is very hard to believe, especially given that the average population density of North Carolina in 1990 was 136 people per square 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 91 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 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. Lastly, 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.

```r
# 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$wtrd) + 10, 5)) + xlab("weekly wage of wholesale, retail trade") +
    ggtitle("Weekly wage of wholesale, retail trade") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
p24 <- ggplot(data = f, aes(wtrd, crmrte)) + geom_point() + geom_smooth(method = lm) +
    xlab("weekly wage of wholesale, retail trade") + ggtitle("Lowest sector wage vs Crime Rate") +
    annotate(x = 325, y = 0.09, label = paste("r = ", round(cor(f$wtrd,
        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$taxpc) + 10, 5)) + xlab("Tax per capita") + ggtitle("Tax per capita") +
    theme(text = element_text(size = 10), plot.title = element_text(size = 10))
p28 <- ggplot(data = f, aes(taxpc, crmrte)) + geom_point() +
    geom_smooth(method = lm) + xlab("taxpc") + ggtitle("Tax Revenue per Capita vs Crime Rate") +
    annotate(x = 80, y = 0.095, label = paste("r = ", round(cor(f$taxpc,
        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), crmrte)) + geom_point() +
    geom_smooth(method = lm) + xlab("log(incgap)") + ggtitle("Log of Income gap vs Crime Rate") +
    annotate(x = 7, y = 0.09, label = paste("r = ", round(cor(log(f$incgap),
        f$crmrte), 2)), geom = "text") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))

grid.arrange(p23, p24, p27, p28, p29, p30, nrow = 3)
```



From the analysis, we see that both of the wage factors *wtrd* and *wfed* as well as *taxpc* have a positive relationship with 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. The income gap 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. Given all of this we will add *taxpc* as a variable to our model.

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. We construct a model with them and examine the output.

```
# Simple Model with Key Explanatory Variables
model_base <- lm(crmrte ~ log(prbpunish) + density + taxpc, data = f)
stargazer(model_base, header = FALSE, type = "latex", omit.stat = "f",
    float = FALSE, star.cutoffs = c(0.05, 0.01, 0.001))
```

	Dependent variable:
	crmrte
log(prbpunish)	-0.008^{***}
	(0.002)
density	0.007^{***}
	(0.001)
taxpc	0.0003^{**}
	(0.0001)
Constant	-0.011
	(0.006)
Observations	91
R^2	0.648
Adjusted R^2	0.636
Residual Std. Error	0.011 (df = 87)
Note:	*p<0.05; **p<0.01; ***p<0.001

The coefficients of the estimators are all statistically significant. *Log(prbpunish)* and *density* are significant at the 0.1% level and *taxpc* is significant at the 1% level. The signs of the coefficients are also in accordance with our hypothesis above: certainty of punishiment is a deterrant for crime and has a negative coefficient, while density and taxpc increase criminal opportunity and reward and have positive coefficients.

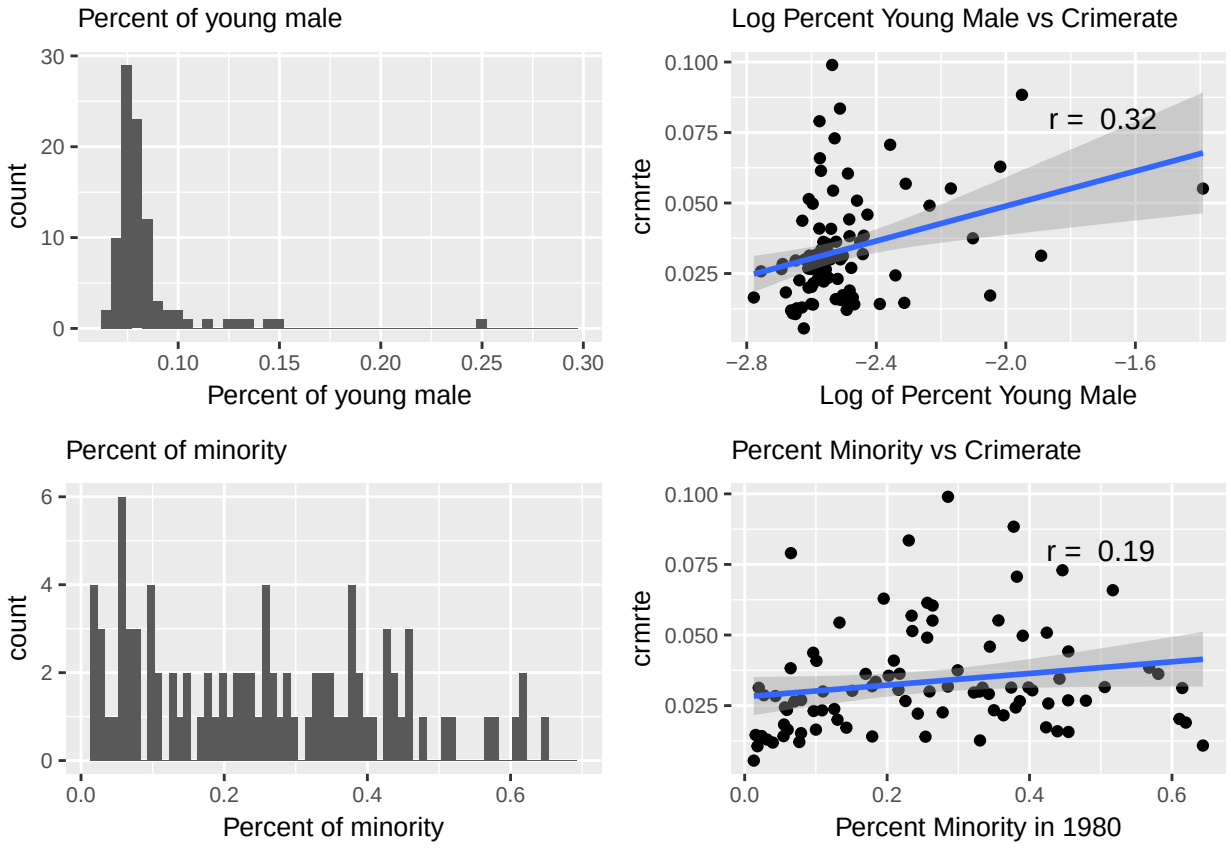
Statistical significance is a good sign, but we must also consider the practical significance of these coefficients. When we look at the coefficients in detail, ceteris paribus, we see that a 1 percent increase in probability of prison sentence (for all offenses), is associated with 0.8 percentage point decrease in 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 0.7 percentage point increase in crime rate. 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. Holding other factors constant, a 1 dollor increase in tax revenue per capita is associated with an increase of 0.03 percentage point in crime rate. The causal link here is a little harder to establish though, because we use *taxpc* as an indirect proxy for income. Without controlling for income, it is rather unclear whether higher income or tax (or both) contibute to the positive association with crime.

Considering the model as a whole, the adjusted R^2 is 0.636 so the model explains about 64% 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 demographics 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$pctymle) + 0.05, 0.005)) + xlab("Percent of young male") +
    ggtitle("Percent of young male") + theme(text = element_text(size = 10),
    plot.title = element_text(size = 10))
p33 <- ggplot(data = f, aes(log(pctymle), crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Log Percent Young Male vs Crimerate") +
    xlab("Log of Percent Young Male") + annotate(x = -1.7, y = 0.08,
    label = paste("r = ", round(cor(log(f$pctymle), 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, crmrte)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Percent Minority vs Crimerate") +
    xlab("Percent Minority in 1980") + annotate(x = 0.5, y = 0.08,
    label = paste("r = ", round(cor(f$pctmin80, f$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 very 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 a lot more spread out and there does seem to be a mild linear relationship between *pctmin80* and the crime rate. We should emphasize that this does not imply that minorities 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. We now show this new model compared to the intial, base model.

```
model_plus <- lm(crmrte ~ log(prbpunish) + density + taxpc +
    log(pctymle) + pctmin80, data = f)
stargazer(model_base, model_plus, header = FALSE, type = "latex",
    omit.stat = "f", float = FALSE, star.cutoffs = c(0.05, 0.01,
        0.001))
```

	Dependent variable:	
	crmrte	
	(1)	(2)
log(prbpunish)	-0.008^{***}	-0.008^{***}
	(0.002)	(0.002)
density	0.007^{***}	0.007^{***}
	(0.001)	(0.001)
taxpc	0.0003^{**}	0.0003^{***}
	(0.0001)	(0.0001)
log(pctymle)		0.015^{**}
		(0.006)
pctmin80		0.030^{***}
		(0.006)
Constant	-0.011	0.020
	(0.006)	(0.017)
Observations	91	91
R^2	0.648	0.749
Adjusted R^2	0.636	0.734
Residual Std. Error	0.011 (df = 87)	0.010 (df = 85)
Note:	*p<0.05; **p<0.01; ***p<0.001	

The new model exhibits a lower Akaike Information Criteria -577.6299044 compared to the base model -550.9676273, 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 0.734, about 10 percentage points higher than the base model. This is a nontrivial improvement, and means that our 5 independent variables can explain more than 70% of the variability in crime rate. Further, both *pctmin80* and *log(pctymle)* exhibit statistical significance of 0.1% and 1% respectively.

In practical terms, a 1 percent increase in the percent of young males in a county is associated with a 1.5 percentage point increase in the crime rate. This is of course, ceteris paribus. For local governments, this

relationship may be of particular use as policies to engage young males in healthy noncriminal activity or work could successfully lower the crime rate. For percent minority, holding other factors constant, a 1 percentage point increase in the percent minority in a county is associated with a 3 percentage point increase in the crime rate. 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.

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.

```
model_control <- lm(crmrte ~ log(prbpunish) + density + taxpc +
    log(pctymle) + pctmin80 + avgsen + west + urban + wcon +
    wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc + mix,
    data = f)
stargazer(model_base, model_plus, model_control, header = FALSE,
    float = FALSE, type = "latex", omit.stat = "f", font.size = "tiny",
    star.cutoffs = c(0.05, 0.01, 0.001))
```

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
log(prbpunish)	−0.008***	−0.008***	−0.008***
	(0.002)	(0.002)	(0.002)
density	0.007***	0.007***	0.006***
	(0.001)	(0.001)	(0.001)
taxpc	0.0003**	0.0003***	0.0003***
	(0.0001)	(0.0001)	(0.0001)
log(pctymle)		0.015**	0.018**
		(0.006)	(0.006)
pctmin80		0.030***	0.036***
		(0.006)	(0.009)
avgsen			−0.0002
			(0.0004)
west			0.001
			(0.003)
urban			−0.001
			(0.007)
wcon			−0.00001
			(0.00003)
wtuc			0.00001
			(0.00002)
wtrd			0.00001
			(0.0001)
wfir			−0.00004
			(0.00003)
wser			−0.00001*
			(0.00001)
wmfg			−0.00000
			(0.00002)
wfed			0.0001
			(0.00003)
wsta			−0.00003
			(0.00003)
wloc			0.0001
			(0.0001)
mix			−0.017
			(0.014)
Constant	−0.011	0.020	0.005
	(0.006)	(0.017)	(0.025)
Observations	91	91	91
R^2	0.648	0.749	0.801
Adjusted R^2	0.636	0.734	0.751
Residual Std. Error	0.011 (df = 87)	0.010 (df = 85)	0.009 (df = 72)
Note:			*p<0.05; **p<0.01; ***p<0.001

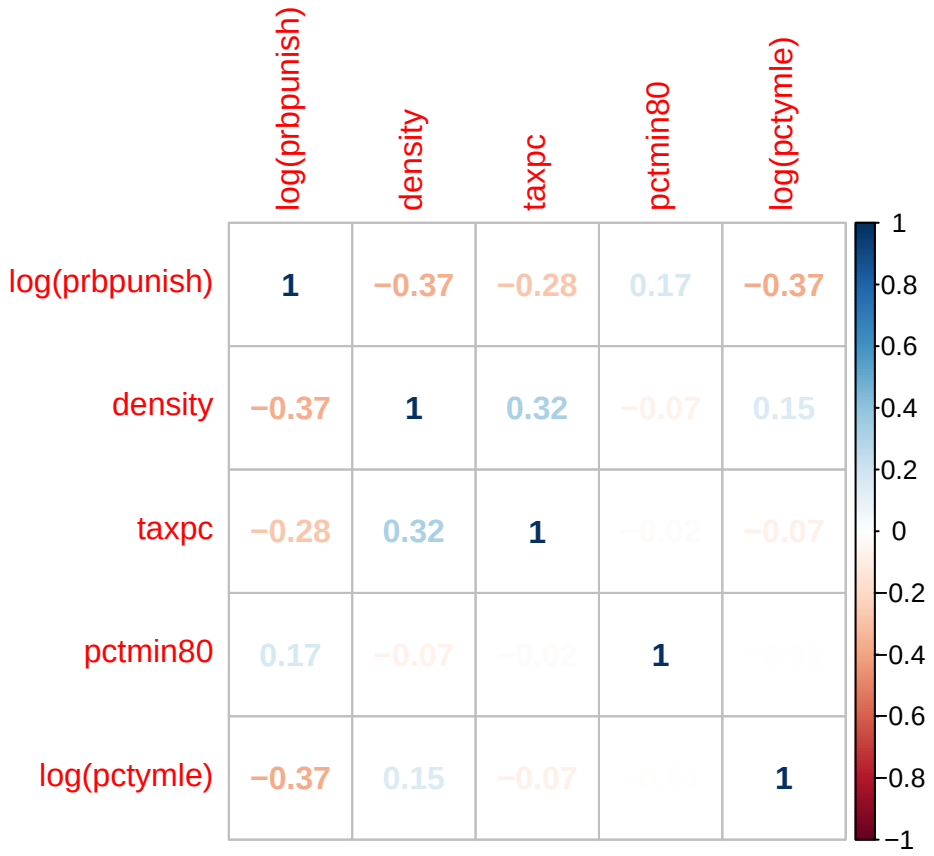
From the regression table above, we see that adding the full range of covariates did not have a large impact on the sign or size of the coefficients in our first and second models. This indicates robustness in our modeling choice. Moreover, we note the full model has an AIC of -572.6175317, which is larger than -577.6299044 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.

```
g <- f[, c("prbpunish", "density", "taxpc", "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")
```



From the correlation graph, we see that fortunately, our independent variables are not very strongly correlated. 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 *taxpc* variable. We focus on taxpc because it is strongly correlated with income level, and wealthy residents are more likely to afford more years of education. To be able to deduct the direction of this the omitted variable bias on taxpc 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 taxpc and education (i.e. $\delta_{taxpc} > 0$), and therefore expect a negative bias on β_{taxpc}.

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 *taxpc*. 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 less tax per capita. Therefore we expect $\delta_{taxpc} < 0$ and thus a negative bias on β_{taxpc}.

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 $\beta_{density}$.

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 *taxpc*, 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 taxpc, we expect $\delta_{homeownership} > 0$, and therefore homeownership will have a negative bias on β_{taxpc}.

Alcohol Consumption

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

For this variable, we want to focus on the impact on $\beta_{pctymle}$, because young males are likely key consumers of alcohol. Assuming all other variables are uncorrelated, we expect a positive relationship between alcohol consumption and percent of young male, i.e. $\delta_{alcohol} > 0$, and therefore, we believe that alcohol consumption will have a positive bias on $\beta_{pctymle}$

Conclusion

We believe that the probability of punishment, population density, and income explain a much of the variability in the crime rate. Accounting for young males and minorities in the population improves this initial model. These variables also suggest a variety of possible policy responses by local governments to lower their crime rate, which we discussed above. The standard errors of the model still must be examined in detail, which we plan to do in subsequent drafts.