W203 Section 0904 Lab 3: Reducing Crime

Jason Baker, John Boudreaux, Alex West

11/23/2018

Introduction

How do you reduce crime? Leaders have wrestled with this question since the dawn of civilization, focusing on all elements of society from education, to economics, to criminal punishment. The candidate released a platform that includes public safety and reduction of crime as a core component, and hired our firm to analyze the data and present policy recommendations. The dataset includes variables describing multiple facets of the North Carolina population, including demographics, law enforcement, criminal punishment, population density, wages, and more. Our approach is to examine the dependent variable, crime rate, against only those variables that we believe can be specifically affected by public sector resources, which therefore have public policy solutions.

Research question:

Can the crime rate be reduced with public sector resources? Do we have direct levers to influence crime rate?

Data Loading and Cleaning

Our data come from a 1994 study from Cornwell and Trumball, who collected various panel data from counties across North Carolina. We will use R (>= 3.4.3) in order to analyze our data and create models.

We first load in the data to our session, and run some basic summary commands to get a broad understanding of the data.

```r
data <- read.csv("../data/crime_v2.csv")
# summary(data) # alternate means to explore data
str(data)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

We can see that our data primarily has numerical fields, some of which are binary categorical variables (west, central, urban) with values of 0 and 1. Because we will be performing a linear regression, it will be useful to keep these as numerical variables rather than discrete factors. Since the 'county' and 'year' variables only act as identifying labels on our data, we can remove these from our data frame to reduce its size with no adverse effects to our working data. We will save these into vectors that we can reference later, should the need arise.

```
county <- data$county
data$county <- NULL
year <- data$year
data$year <- NULL
```

A next logical step for us is to look into the 'prbconv' variable, and why it is being treated as a factor instead of a numeric.

```
data$prbconv
```

```
##  [1] 0.527595997 1.481480002 0.267856985 0.525424004 0.476563007
##  [6] 0.068376102 0.520606995 0.769231021 0.436441004 1.225610018
## [11] 0.334701002 0.403780013 0.406780005 0.352941006 0.515464008
## [16] 0.325300992 0.385495991 0.972972989 0.452829987 0.450567007
## [21] 0.763333023 0.371879011 0.259833008 0.140350997 0.207830995
## [26] 0.736908972 0.62251699  0.493438005 0.459215999 0.154451996
## [31] 0.248275995 0.739394009 0.229589999 0.528302014 0.308411002
## [36] 0.203724995 0.457210004 0.549019992 0.548494995 0.386925995
## [41] 0.589905024 0.573943973 0.595077991 1.234380007 0.571429014
## [46] 0.384236008 0.364353001 0.781608999 0.522387981 0.220339
## [51] 1.5         0.793232977 0.347799987 0.226361006 0.438960999
## [56] 1.358139992 0.393413007 0.495575011 0.271946996 0.477732986
## [61] 1.068969965 0.28947401  0.412698001 0.314606994 0.340490997
## [66] 0.426777989 1.015380025 0.36015299  0.520709991 0.559822977
## [71] 0.443681002 0.492940009 0.50819701  0.401198    0.468531013
## [76] 0.322580993 0.722972989 0.909090996 0.327868998 0.410596013
## [81] 0.328664005 0.343023002 0.381908    2.121210098 0.443114012
## [86] 0.300577998 0.449999988 0.588859022 0.588859022 1.670519948
## [91] 1.182929993
## [96]             `
## 92 Levels:  ` 0.068376102 0.140350997 0.154451996 ... 2.121210098
```

We can see that there are entries that are not numeric, with commas, apostrophes, and other characters. Unfortunately, considering we expect this field to be numerical values, we should treat these as missing data since it is likely entered incorrectly. For our analysis, we will simply replace them with NA values. We can do this while converting all of the numeric values into R-numeric format with the following command, which will coerce all the non-numerics to NA.

```
data$prbconv <- as.numeric(as.character(data$prbconv))
```

```
## Warning: NAs introduced by coercion
```

At this point, we should look at the missing values throughout our data. We will do this by searching for the missing rows in each column of the data frame.

```r
na.rows <- lapply(data, function(x){which(is.na(x))})
na.rows
```

```
## $crmrte
## [1] 92 93 94 95 96 97
##
## $prbarr
## [1] 92 93 94 95 96 97
##
## $prbconv
## [1] 92 93 94 95 96 97
##
## $prbpris
## [1] 92 93 94 95 96 97
##
## $avgsen
## [1] 92 93 94 95 96 97
##
## $polpc
## [1] 92 93 94 95 96 97
##
## $density
## [1] 92 93 94 95 96 97
##
## $taxpc
## [1] 92 93 94 95 96 97
##
## $west
## [1] 92 93 94 95 96 97
##
## $central
## [1] 92 93 94 95 96 97
##
## $urban
## [1] 92 93 94 95 96 97
##
## $pctmin80
## [1] 92 93 94 95 96 97
##
## $wcon
## [1] 92 93 94 95 96 97
##
## $wtuc
## [1] 92 93 94 95 96 97
##
## $wtrd
## [1] 92 93 94 95 96 97
##
## $wfir
## [1] 92 93 94 95 96 97
##
## $wser
## [1] 92 93 94 95 96 97
##
```

```
## $wmfg
## [1] 92 93 94 95 96 97
##
## $wfed
## [1] 92 93 94 95 96 97
##
## $wsta
## [1] 92 93 94 95 96 97
##
## $wloc
## [1] 92 93 94 95 96 97
##
## $mix
## [1] 92 93 94 95 96 97
##
## $pctymle
## [1] 92 93 94 95 96 97
```

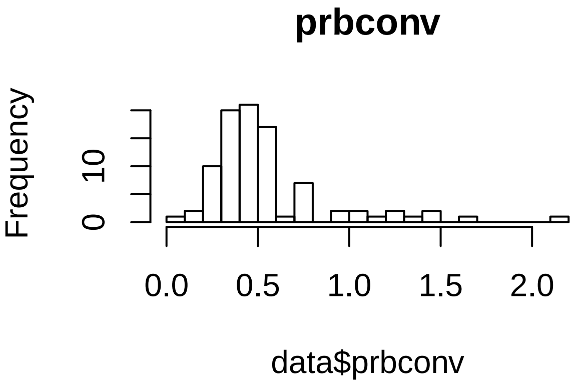
Rows 92 through 97 are missing values for nearly every column in our data. Given this, we should be skeptical about the information that the existing values give us in these rows. For our analysis, we will drop all of these rows entirely since we do not know the exact methods in which these data were collected.

```
data <- data[-c(92:97),]
```

While our group analyzed boxplots and histograms for all variables in the data, we will only highlight a few for the sake of brevity. We should point out the 'prbconv' variable, which is supposed to be the probability of a conviction given an arrest. Because this is a probability, it does not make sense to have any values above 1. We manually set these values to NA.

```
# command for running all boxplots, histograms for all variables
# for(i in 1:ncol(data)){
#   if(is.numeric(data[[i]])){
#     hist(data[[i]], breaks = 15, main = colnames(data)[i])
#     boxplot(data[[i]], main = colnames(data)[i]
#   }
# }

# commands for exploring prbconv
hist(data$prbconv, breaks = 20, main = "prbconv")
```



```
data$prbconv[data$prbconv > 1] <- NA
```

While there are statistical outliers in nearly all of our variables according to the boxplots, which calculate outliers as 1.5 +- IQR, we cannot simply eliminate all statistical outliers because we do not have a grasp on the realistic boundaries of these data. Given we do not have much information about the collection methods for this data set, we choose to keep the majority of these "outliers" considering we do not have information that says they are not reflective of reality.

There is one exception to the comments above, however. With the 'wser' variable, we see that there is a single outlier that lies extremely far away from the rest of the data. Our group finds this point to be very suspect, and will remove it from further analysis. Performing a hypothesis test might provide further justification, but we will keep this out of the report for the sake of brevity.

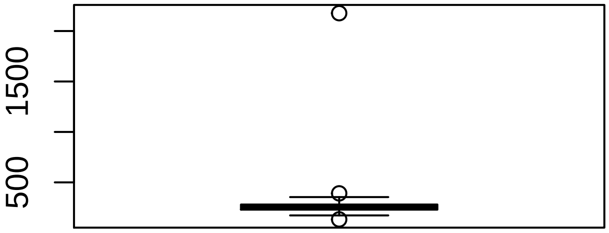
```
mean(data$wser, na.rm = TRUE)
```

```
## [1] 275.5642
```

```
median(data$wser, na.rm = TRUE)
```

```
## [1] 253.2281
```

```
boxplot(data$wser)
```

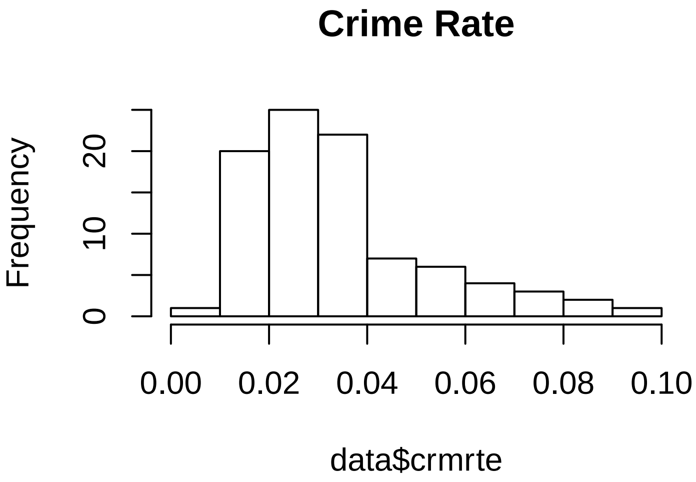


```
# let's set our major outlier to NA just for wser
data$wser[data$wser > 1500] <- NA
```

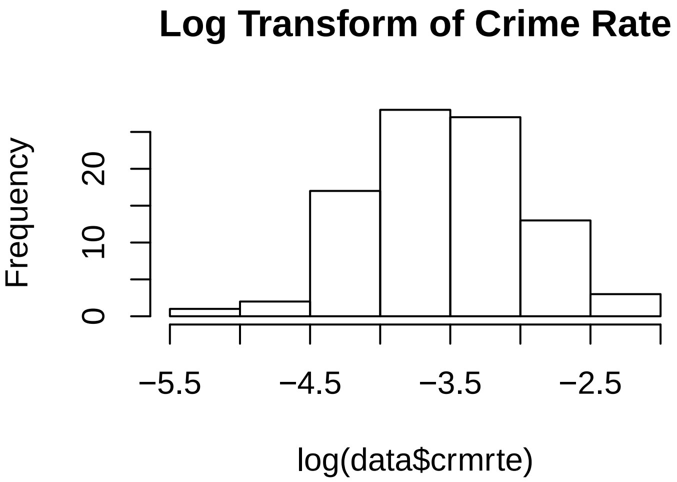
Model Building

With our primary objective looking towards the impact of public sector resources on current crime rates our initial step was to investigate the Crime Rate ('crmrte'). A cursory look at the summary and histogram plots indicate the 'crmrte' is rightly skewed. In an effort to reduce skewness, preserve linear relationships, and allowing for comparisons of relative differences as opposed to absolute differences, it was determined to conduct a log transformation of the 'crmrte' variable. The resulting histogram was more normally distributed and led to better fit regression models.

```
hist(data$crmrte, main = "Crime Rate")
```



```r
hist(log(data$crmrte), main = "Log Transform of Crime Rate")
```

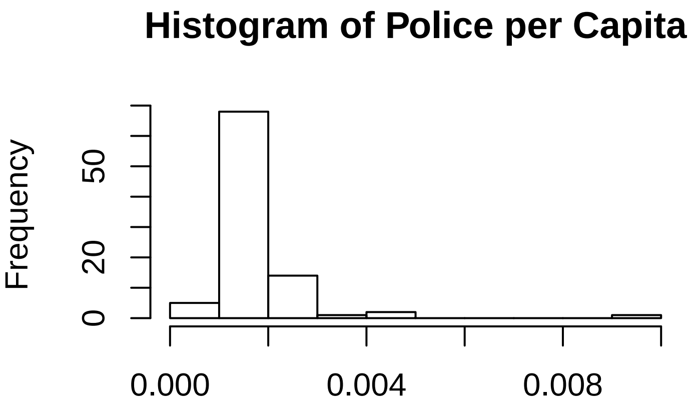


```r
data$log.crmrte <- log(data$crmrte)
```

We continued investigating the remainder of the data, searching for variables the public sector could potentially influence in the hopes of reducing crime within the state of North Carolina. We deemed the Police per Capita ('polpc') as a possible primary explanatory variable due to the perceived effect higher police presence has on the reduction of crime. Furthermore, additional Tax Revenue per Capita ('taxpc') was noted as an additional variable of interest due to the ability to strengthen police forces or the funding of programs directed towards education, job creation, and other programs aimed specifically at reducing crime. Population 'density', while not within control or influence of local governments, was considered a valid regressor as it could highlight the rate of crime with respect to higher populations thereby directing governments to geographic areas where resources could be allocated.

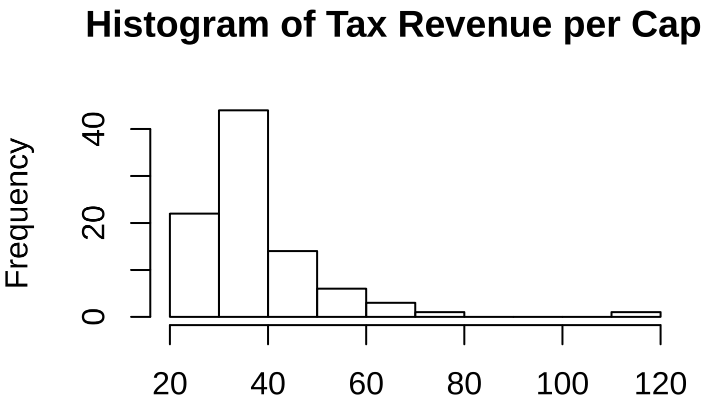
With variables of interest having been determined, we further inspected the chosen regressors. In looking at histograms for each variable, police per capita was found to be rightly skewed with a median value of 0.0014853 and a slightly higher mean of 0.0017022, likely due to the maximum value of 0.0090543. This maximum value warranted additional attention as it was significantly larger than the median value; this is addressed in the base model discussion below.

```r
hist(data$polpc, main = "Histogram of Police per Capita",
     xlab = NULL)
```



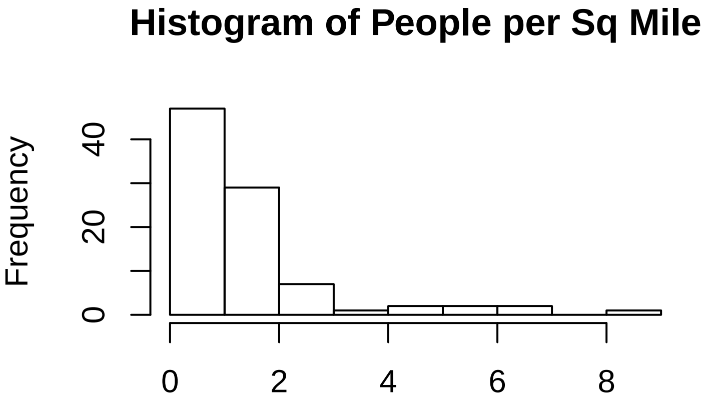
The histogram for tax revenue per capita is rightly skewed as well, having a median value of 34.87, a mean of 38.06, and a maximum of 119.76.

```r
hist(data$taxpc, main = "Histogram of Tax Revenue per Capita",
     xlab = NULL)
```



We also find the density histogram is rightly skewed with a median value of 0.96226, a mean of 1.4288 and a maximum of 8.82765.

```
hist(data$density, main = "Histogram of People per Sq Mile",
     xlab = NULL)
```



Having applied the log transformation to the crime rate variable, and completing a thorough exploratory investigation of the remaining data, the team began constructing multiple regression models aimed at addressing the level of impact of our selected explanatory variables.

Our approach looked initially at a single variable, police per capita:

$$log(Crimerate) = \beta_0 + \beta_1(Policepercapita) + u$$

Our second model took tax revenue per capita and density into account:

$$log(Crimerate) = \beta_0 + \beta_1(Policepercapita) + \beta_2(density) + \beta_3(Taxpercapita) + u$$

Our final model accounted for the geographic region, the weekly wage for state employees, and the average sentence length.

$$log(Crimerate) = \beta_0 + \beta_1(Policepercapita) + \beta_2(density) + \beta_3(Taxpercapita) +$$
$$\beta_4(west) + \beta_5(central) + \beta_6(urban) + \beta_7(statewage) + \beta_8(averagesentence) + u$$

In order to infer anything from this data analysis, it is necessary to articulate the assumptions that allow the models to function. These assumptions are known as the classical linear model assumptions.

1. Linear in Parameters

- The true relationship between our explanatory variables and the crime rate is linear (not parabolic, or exponential, or any other shape). In order for our ordinary linear regression analysis to be valid, we are assuming this to be true.
- Possible obstacles: the relationship between the variables may not be linear in nature. Our models in this analysis may not capture the nuance.

2. Random Sampling

- We are assuming that the sample is independent and identically distributed, meaning that each data point is independent and does not affect any other data points (one draw does not affect any other draws).
- Possible obstacles: With the nature of an external dataset, it's impossible to know this for certain.
- We can be reasonably sure because this data has been used with some success in other research, however that is not always a good indicator of random sampling.

3. No Perfect Collinearity

- We are assuming that there is no exact linear relationship among the independent variables. In other words, the variable measuring police per capita is not also measuring tax revenue, density, etc. In this particular dataset we are not comparing items with the same units, and we are relatively confident that there is no perfect collinearity amongst the independent variables.
- Possible obstacles: We expect some of the variables to be correlated (such as tax revenue and density) but not perfectly correlated. If they are, the model cannot be estimated by ordinary least squares regression, however upon dropping one of the linearly dependent terms we could perform OLS regression.
- We also note that the R regression function, lm(), automatically checks for perfect collinearity and will return a warning specifying a rank-deficient matrix should this be the situation with the data. Given our code generates no warnings, we can confirm CLM assumption 3.

4. Zero Conditional Mean

- There is no functional relationship between our explanatory variables (police per capita, tax revenue, density, etc) and the error term, u.
- Possible obstacles: This is a difficult assumption to assert since we are working with one year of data and may be subject to omitted variable bias. We will discuss this in greater detail later on in the analysis.

5. Homoskedasticity

- Variance of the error term does not depend on the levels of the explanatory variables. In other words, the variance in the error term, u, conditional on any of our explanatory variables, is the same for all combinations of outcomes.
- Possible obstacles: If this assumption does not hold and the error term varies differently with each explanatory variable, or even within one variable, the results of the regression should not be trusted.

6. Normality

- The error is independent of the explanatory variables and is normally distributed. This assumption is much stronger than the previous assumptions, and if true, automatically includes assumptions 4 and 5. With a large sample size we can implicitly assume normality by invoking the Central Limit Theorem.

Base Model

The base model involves the most visible public resource in law enforcement and crime prevention: police. In this model the dependent variable is 'crime rate' and the independent variable is 'police per capita.'

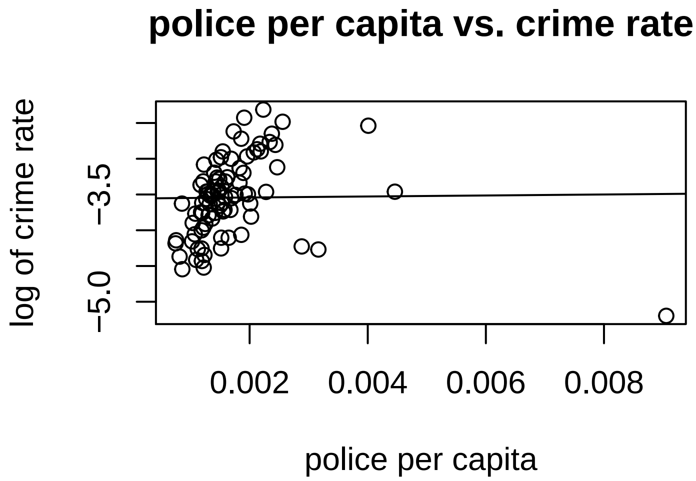
The model looks like this:

$$log(Crimerate) = \beta_0 + \beta_1(Policepercapita) + u$$

```
plot(data$polpc, log(data$crmrte), xlab = "police per capita",
         ylab = "log of crime rate", main = "police per capita vs. crime rate")
(linear.model.1 <- lm(log(crmrte) ~ polpc, data = data))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ polpc, data = data)
##
## Coefficients:
## (Intercept)        polpc
##      -3.556        6.985
```

```
abline(linear.model.1)
```



```
summary(linear.model.1)$r.square
```

```
## [1] 0.0001593452
```

There is just one coefficient in this model, and it represents the relationship between the ratio variable of 'police per capita' and the log transformation of the crime rate. According to this model, each extra member of the police force is associated with 6.9(check) change in the log of crime rate. This is surprising considering the generally held belief that police help to prevent crime. However, it is important to realize that the police variable here is normalized for population, so this model may just be measuring that the more people are located in an area, the higher the crime rate.

We did notice a potential outlier in the data and explored it a bit more.

First, we explored the entire row housing the high police per capita data point - does the county have a particularly dense population?

```
# row for outlier
print("Data from the row with the outlier:")
```

```
## [1] "Data from the row with the outlier:"
```

```
print(data[data$polpc > 0.006,])
```

```
##       crmrte  prbarr prbconv prbpris avgsen      polpc   density   taxpc
## 51 0.0055332 1.09091      NA     0.5   20.7 0.00905433 0.3858093 28.1931
##    west central urban pctmin80     wcon     wtuc     wtrd     wfir
## 51    1       0     0  1.28365 204.2206 503.2351 217.4908 342.4658
##         wser   wmfg  wfed    wsta    wloc mix    pctymle log.crmrte
## 51 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495  -5.196989
```

```
# medians for comparison
print("The median values for our data:")
```

```
## [1] "The median values for our data:"
```

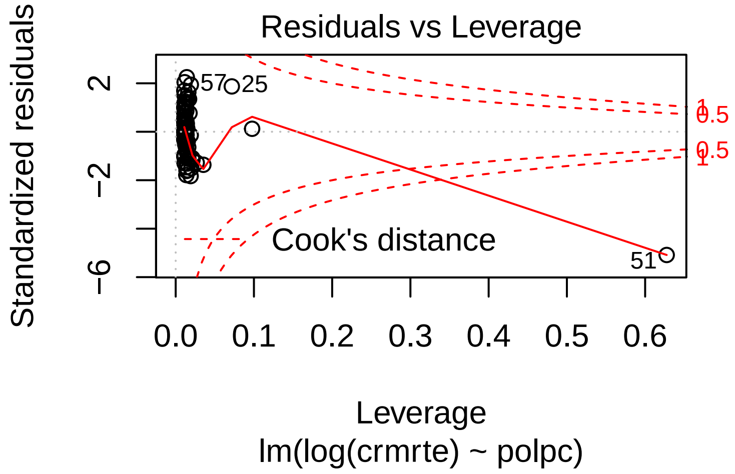
```
print(apply(data, 2, function(x){median(x, na.rm = TRUE)}))
```

```
##        crmrte        prbarr       prbconv       prbpris        avgsen
##    0.02998560    0.27094999    0.43896100    0.42342299    9.10000038
##         polpc       density         taxpc          west       central
##    0.00148532    0.96226412   34.87021255    0.00000000    0.00000000
##         urban      pctmin80          wcon          wtuc          wtrd
##    0.00000000   24.31170082  281.42590330  406.50405880  203.01623540
##          wfir          wser          wmfg          wfed          wsta
## 317.30767820  253.11882020  320.20001220  449.83999630  357.69000240
##          wloc           mix       pctymle    log.crmrte
## 308.04998780    0.10186080    0.07771273   -3.50703805
```

The density at that is 0.38 (lower than the median) - so no, the county is not particularly dense.

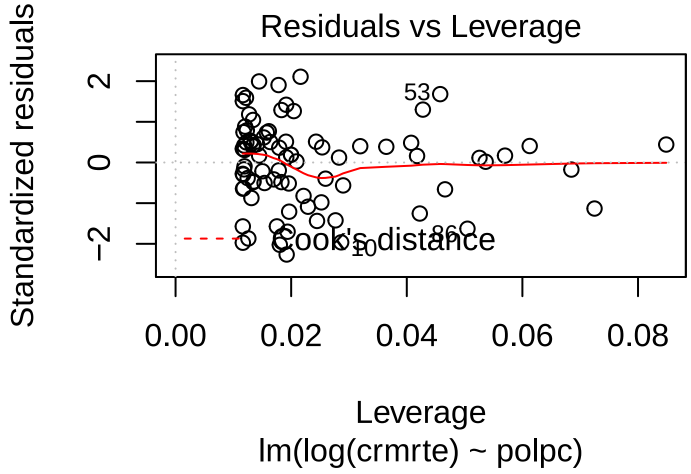
Next, we performed Cook's test to determine if the outlier truly affects the data.

```
plot(linear.model.1, which = 5)
```

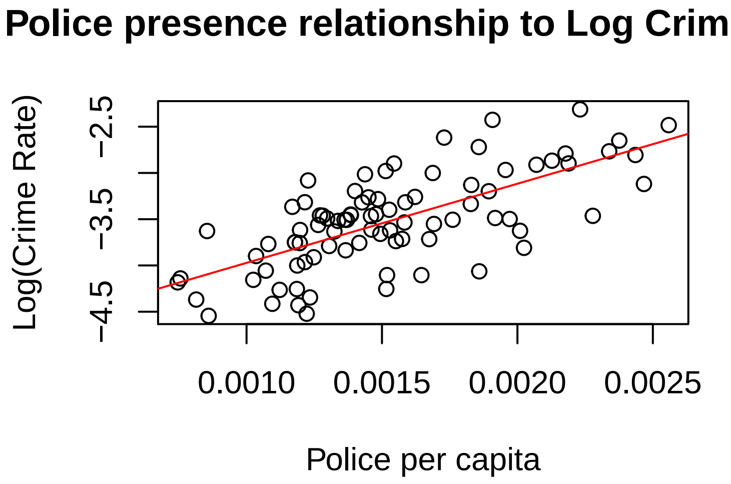


Cook's test proved that the point is a true outlier and affects the model significantly. We then removed the data point to observe the effect on the data:

```
# it appears a few points are weighting our model.
# let's take them out and see how much better we can get
data2 <- data[data$polpc < 0.0027,]
linear.model.1.2 <- lm(log(crmrte) ~ polpc, data = data2)
plot(linear.model.1.2, which = 5)
```



```
plot(x = data2$polpc, y = data2$log.crmrte,
     main = "Police presence relationship to Log Crime Rate",
     xlab = "Police per capita", ylab = "Log(Crime Rate)")
abline(linear.model.1.2, col = "red")
```



If removed, the coefficient jumps to 14, meaning a change in one police officer per capita (which would also signify a change in population) would result in a change in the log of the crime rate of 14.

It is clear that this univariate model is not enough to measure the effect of increasing the number of police per capita. First of all, the model is measuring multiple variables within one variable, and may in fact be a better predictor of population density than police presence. It will be necessary to add other elements to the model to help control for these effects.

Second Model

Our second model incorporates a few new covariates: population density and tax revenue per capita. The rationale to incorporate both of these has been discussed previously, as these are both directly related the resources available to implement new policy and also the potential policy impact. Population density in particular is an important variable to include, as it may reduce some of the bias introduced by 'police per capita.'

Our new linear model will take the form:

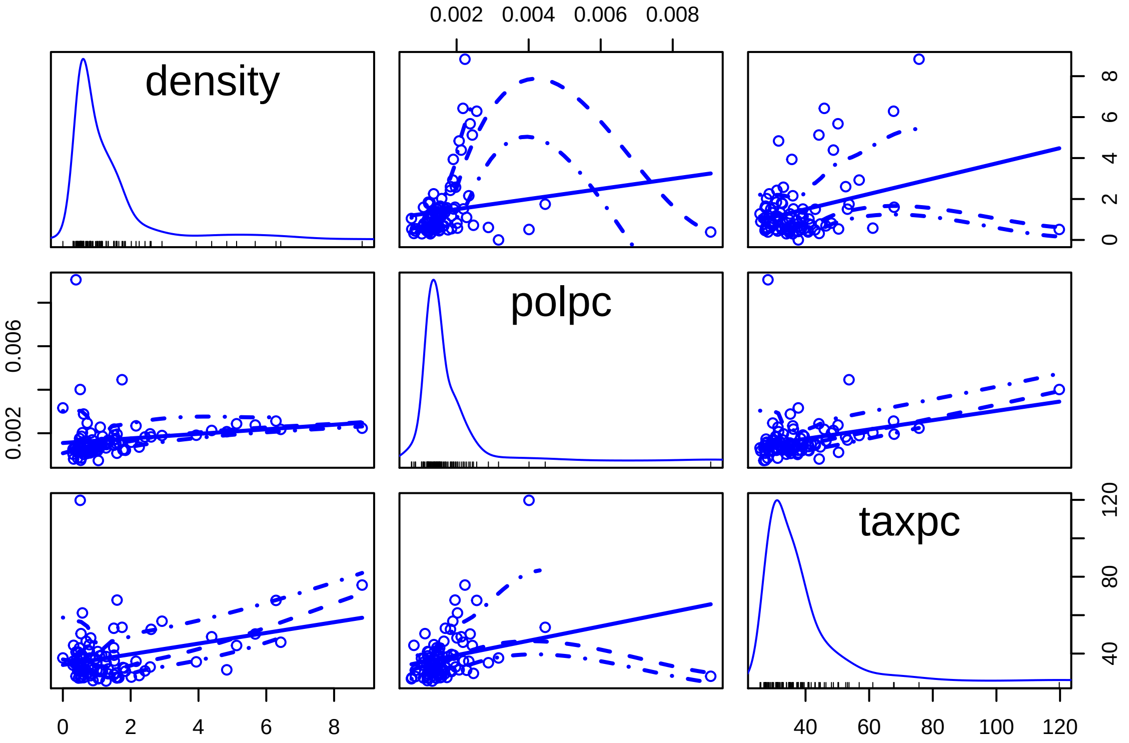
$$log(Crimerate) = \beta_0 + \beta_1(Policepercapita) + \beta_2(density) + \beta_3(Taxpercapita) + u$$

Before computing this linear model, we should first take a look at all of our regressors to understand their relationships and especially to see if any might be completely linearly dependent. If we found this, we would violate CLM assumption 3.

```
library(car)
```

```
## Loading required package: carData
```

```
scatterplotMatrix(~ density + polpc + taxpc, data = data)
```



While police per capita and tax per capita appear to be positively correlated, there is no concern for perfect collinearity. We assume that the remainder of the CLM assumptions hold true, and will check the validity of these upon calculation of the model.

```
linear.model.2 <- lm(log(crmrte) ~ density + polpc + taxpc, data = data)
summary(linear.model.2)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ density + polpc + taxpc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81901 -0.25850 -0.01256  0.24586  0.99017
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.048369   0.140003 -28.916  < 2e-16 ***
## density       0.212269   0.030457   6.969 5.82e-10 ***
## polpc       -78.559969  46.138606  -1.703   0.0922 .
## taxpc         0.008797   0.003630   2.423   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

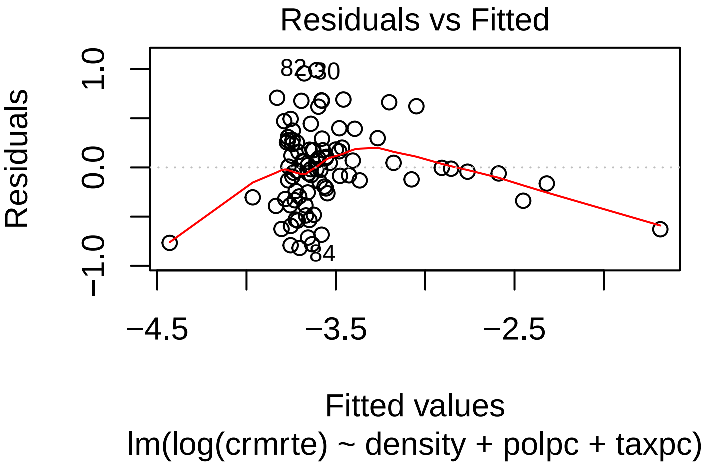
```
##
## Residual standard error: 0.413 on 87 degrees of freedom
## Multiple R-squared:  0.4471, Adjusted R-squared:  0.428
## F-statistic: 23.45 on 3 and 87 DF,  p-value: 3.258e-11
```

It is particularly interesting that our coefficient for the police per capita has now changed from a positive value to a larger negative value. We will discuss this in more detail later in our report when we compare all of our linear models together, as this tells us useful information about omitted variable bias.

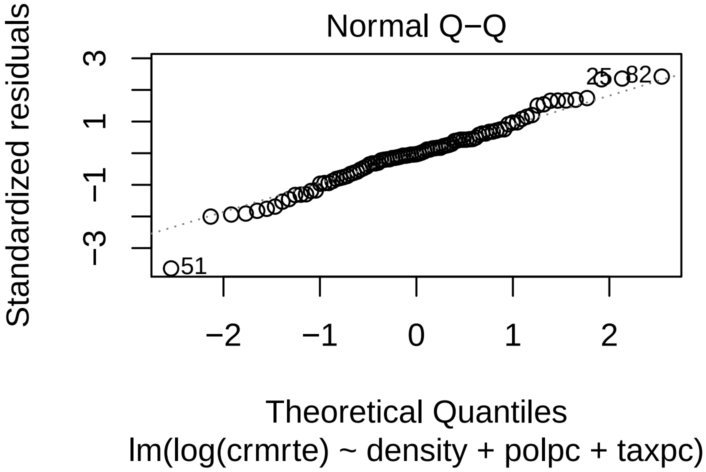
The interpretation of density is intuitive; as density increases, we expect the crime rate to increase at a rate of 0.21% per unit of density increase (people per square mile). Our value for tax per capita, however, is less intuitive. The coefficient suggests that crime rate increases with an increase in tax per capita. Our group proposes that this is again due to bias effects from covariates and omitted variables. It may be useful to note that a univariate regression of tax per capita with the log transformed crime rate yields a positive coefficient of 0.015, so the inclusion of our new variables has accounted for some of the bias here.

Our multiple R Squared value is 0.4471, which suggests that our model explains about 44.7% of the variance in the data. The more useful statistic to observe, however, is the adjusted R squared of 0.428. This statistic also measures the variance in the data explained by the model, but also has corrections for including additional terms.

```r
plot(linear.model.2, which = 1)
```

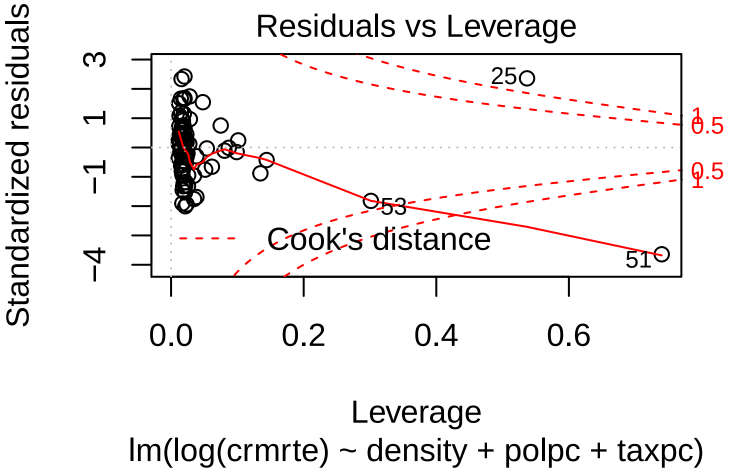


```r
plot(linear.model.2, which = 2)
```



Upon looking at our residuals vs fitted value plot we see that our values remain scattered relatively evenly around zero, with some extreme values on the high and low ends that could possibly be outliers. This validates CLM assumption 4. We can see from that our Normal Q-Q plot for residuals falls closely to the expected line for normally distributed residuals, with the exception of an extreme point on the higher and lower end. Given that the majority of our data fall close to expected values for normally distributed values, we can say that CLM assumption 6 holds which implies CLM assumption 5 holds as well.

```
plot(linear.model.2, which = 5)
```



When looking further into the Cooks distance we see that we have two points that are identified as highly influential, with a Cooks distance greater than 1. These were also both the points that were the furthest away from our Normal Q-Q plot. We note that while excluding these values would likely lead to a more explanatory model, we do not have enough information about the data to remove these from the analysis.

Third Model

Our third model incorporates as many covariates as we possibly can, in order to demonstrate the "kitchen sink" approach to ordinary least squares regression. On top of our original explanatory variables of police per capita, tax revenue per capita, and population density, we are including average sentence in days, geographic indicators like west, central, and urban, and wages of state employees. The rationale to incorporate all of these in addition to the others is to illustrate that while we can create a model that raises the statistical significance of certain variables, it may be misleading and suggest relationships where there are none. The extra variables chosen fit the research question - public policy solutions can work to affect sentence time, where resources are focused geographically, and wages of state employees. However, in the original model building process we felt other variables were more appropriate to focus on.

Our third linear model will take the form:

$$log(Crimerate) = \beta_0 + \beta_1(Policepercapita) + \beta_2(density) + \beta_3(Taxpercapita) + \beta_4(west) + \beta_5(central) +$$

$$\beta_6(urban) + \beta_7(statewage) + \beta_8(averagesentence) + u$$

Before computing this linear model, we first took a look at all of our regressors to understand their relationships and especially to see if any might be completely linearly dependent. If we found this, we would violate CLM assumption 3. We performed a scatterplot matrix but did not include in the analysis in the interest of brevity.

While police per capita and tax per capita again appear to be positively correlated, there is no concern for perfect collinearity. Additionally, density and urban appear to be positively correlated, which is a good sanity check on our data. Tax per capita and urban also appear to be positively correlated. There do not appear to be any relevant negative correlations.

We will assume that CLM assumptions 1 - 3 hold true, and will check the validity of CLM assumptions 4 - 6 upon calculation of the model.

```
linear.model.3 <- lm(log(crmrte) ~ density + polpc + taxpc + west + central + wsta + avgsen, data = data
summary(linear.model.3)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ density + polpc + taxpc + west + central +
##     wsta + avgsen, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85069 -0.21730  0.00705  0.18329  0.94774
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.815e+00  3.875e-01  -9.846 1.33e-15 ***
## density      2.245e-01  3.084e-02   7.281 1.73e-10 ***
## polpc       -1.972e+01  4.908e+01  -0.402   0.6889
## taxpc        4.822e-03  3.466e-03   1.391   0.1679
## west        -4.626e-01  1.038e-01  -4.458 2.57e-05 ***
## central     -2.312e-01  9.699e-02  -2.383   0.0194 *
## wsta         4.747e-04  9.632e-04   0.493   0.6234
## avgsen      -1.720e-02  1.638e-02  -1.050   0.2969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3769 on 83 degrees of freedom
## Multiple R-squared:  0.5608, Adjusted R-squared:  0.5238
## F-statistic: 15.14 on 7 and 83 DF,  p-value: 1.35e-12
```

In comparison with our second model, the coefficient for police has increased while the coefficient for tax revenue has reduced, and density has remained relatively unchanged. The R squared values have improved with the new model but with the inclusion of additional variables, we must be wary of overfitting or spurious correlations. The addition of more independent variables within the regression leads to a greater probability that one or more will be found to be statistically significant, yet having no causal effect on the dependent variable. Although this model incorporates many new variables, it still does not include many important omitted variables from the other models that will be referenced further in the report. We should be wary of the improved accuracy that this model provides since there are many factors that do not directly measure our dependent variable.

Regression Table

```
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(linear.model.1, linear.model.2, linear.model.3,
          title = "Results",
          align = TRUE)
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Sun, Nov 25, 2018 - 21:08:32
## % Requires LaTeX packages: dcolumn
## \begin{table}[!htbp] \centering
##   \caption{Results}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lD{.}{.}{-3} D{.}{.}{-3} D{.}{.}{-3} }
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{3}{c}{\textit{Dependent variable:}} \\
## \cline{2-4}
## \\[-1.8ex] & \multicolumn{3}{c}{log(crmrte)} \\
## \\[-1.8ex] & \multicolumn{1}{c}{(1)} & \multicolumn{1}{c}{(2)} & \multicolumn{1}{c}{(3)}\\
## \hline \\[-1.8ex]
##  density &  & 0.212^{***} & 0.225^{***} \\
##   &  & (0.030) & (0.031) \\
##   & & & \\
##  polpc & 6.985 & -78.560^{*} & -19.715 \\
##   & (58.649) & (46.139) & (49.076) \\
##   & & & \\
##  taxpc &  & 0.009^{**} & 0.005 \\
##   &  & (0.004) & (0.003) \\
##   & & & \\
##  west &  &  & -0.463^{***} \\
##   &  &  & (0.104) \\
##   & & & \\
##  central &  &  & -0.231^{**} \\
##   &  &  & (0.097) \\
##   & & & \\
##  wsta &  &  & 0.0005 \\
##   &  &  & (0.001) \\
##   & & & \\
##  avgsen &  &  & -0.017 \\
##   &  &  & (0.016) \\
##   & & & \\
##  Constant & -3.556^{***} & -4.048^{***} & -3.815^{***} \\
##   & (0.115) & (0.140) & (0.388) \\
##   & & & \\
## \hline \\[-1.8ex]
## Observations & \multicolumn{1}{c}{91} & \multicolumn{1}{c}{91} & \multicolumn{1}{c}{91} \\
## R$^{2}$ & \multicolumn{1}{c}{0.0002} & \multicolumn{1}{c}{0.447} & \multicolumn{1}{c}{0.561} \\
## Adjusted R$^{2}$ & \multicolumn{1}{c}{-0.011} & \multicolumn{1}{c}{0.428} & \multicolumn{1}{c}{0.524
## Residual Std. Error & \multicolumn{1}{c}{0.549 (df = 89)} & \multicolumn{1}{c}{0.413 (df = 87)} & \mu
## F Statistic & \multicolumn{1}{c}{0.014 (df = 1; 89)} & \multicolumn{1}{c}{23.447$^{***}$ (df = 3; 87
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{3}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

We can see from our regression table that our last model has the best predictive power, based upon the comparatively high adjusted R Squared value of 0.524. Our original model, only utilizing police per capita, is extremely poor for predictive power with a negative adjusted R squared value.

Although police per capita is perhaps the most direct policy lever we have available, the poor R squared value and the fact that the 'polpc' variable changes drastically in each of our models implies that there is insufficient proof of causality. The changing coefficients show the effects of the omitted variables that were included in each subsequent model; this will be discussed further in the following section on omitted variables.

It is interesting to note that the coefficient for density increased in significance after the inclusion of other effects pertaining to locality, such as west and central. This would suggest that geography is a better predictor of crime rate than police per capita, as we originally thought. Although it may be a decent predictor, we are inclined to say that this is a case of correlation rather than causation since it is more likely that certain omitted variables such as unemployment may be correlated to geography. It is these omitted variables that we are more interested in for a causal analysis rather than density or geography themselves.

Omitted Variables

As we are limited to one year of data and are using ordinary least squares regression, it is possible that these findings may be heavily influenced by omitted variable bias.

Possible variables that have been omitted from this dataset may include: - Speed of Sentencing/Conviction - Severity of Punishment ("Harshness", fines, types…jail/community service) - Educational attainment level of population (% with HS diploma, Assoc degree, bach degree or higher) - Unemployment rate by county - Happiness and fulfillment

The crime rate has a cause, and if we could just write all of the causes correctly, we would have a causal model. The central problem is that even though these causes exist, we can't measure all of them. Some of the possible omitted variables are measurable, like educational attainment, and some are not, like happiness.

We will discuss the possible effects of omitted variables on the base model, as police per capita is the most direct and visible public policy avenue for reducing crime.

Base model:

$$Crimerate = \beta_0 + \beta_1(policepercapita) + u$$

Our base model determined that the β_1 coefficient is positive. This factors heavily into the analysis of the omitted variable bias. Though not what we would like to see and not what we see in subsequent models, this coefficient remains positive throughout the analysis in order to show the effects of omitted variables on this base model specifically.

We first write down both equations (expressing the first equations in terms of the omitted variable (for the purposes of demonstration, we'll choose the first omitted variable on the list, speed of sentencing):

Omitted: speed of sentencing/conviction

$$Crimerate = \beta_0 + \beta_1(policepercapita) + \beta_2(speedofsentencing/conviction) + u$$

$$speedofsentencing/conviction = \alpha_0 + \alpha_1(policepercapita) + u$$

Then, we apply background knowledge to estimate whether omitted variable bias will drive the slope coefficient towards zero or away from zero:

In this case, we believe the β_2 coefficient will be less than 0 (or negative), and the α_1 coefficient is difficult to pinpoint (does more police presence increase the speed of sentencing or is that purely the realm of the courts?). If it is related at all, the relationship is likely slightly positive ($\alpha_1 > 0$). Therefore the omitted variable bias (OMVB) = $\beta_2\alpha_1 < 0$, and we've already calculated β_1 to be greater than 0 (the effect of police per capita on crime rate is positive according to the data). As a result of the omitted variable bias, the OLS coefficient on police per capita will be scaled toward zero (less positive), losing statistical significance.

Given that the perceived omitted variable bias for speed of sentencing/conviction is negative, the OLS estimates that we performed in the base model will underestimate the marginal effect of police per capita on crime rate. Furthermore it will scale the coefficient closer to zero, making it harder to reject the null hypothesis, and lose statistical significance.

We can apply this same technique to our multiple omitted variables. β_1, or our coefficient of police per capita, is always positive (in our base model), and each coefficient analysis requires background knowledge and estimation. Our analysis determines that the ordinary least squares regression in our base model underestimates the effects of most of the possible omitted variables, indicating that our original base model has very little statistical significance.

Omitted Variable	β_2 +/-	α_1 +/-	β_1 +/- (from base model)	OMVB $\beta_2\alpha_1$ +/-
Speed of Sentencing/Conviction	-	+(a little)	+	-
Severity of Punishment	-	+(a little)	+	-
Educational attainment	-	No correlation (or minimal)	+	0
Unemployment	+	No correlation (or minimal)	+	0
Happiness	-	+(a little)	+	-

This is expected given our original analysis, and the reason for including multiple variables in subsequent models.

Conclusion

As with most data analysis, we are left with some insights and more questions rather than absolute answers.

The candidate is building a platform on public safety and crime reduction. We do not have any variables from our analysis that can be used effectively as direct policy levers, such as police presence, however this analysis allows us to conclude that the following public policy solutions could be applied to affect the crime rate via proxy metrics: 1. Density is the best "lever" we have available - perhaps we look into incentivizing people to move to less populated areas. 2. Incentivize businesses to create jobs in less dense areas. 3. Do not apply the same solution indiscriminately around the state. Given that our most significant coefficients were related to density and geographic location, it makes sense to learn more about the drivers of crime specific to each location.

Further analysis especially focused on these areas may yield clearer results.

High level concerns of a political campaign are different than the high level concerns of an elected politician. It is a different question to pose: will reducing crime rate get our candidate elected? Or the appearance of "being tough on crime"? Increasing police presence may actually be better for the purposes of getting elected. We need to ask more about what will get a candidate elected, and this would likely require different data than what is given.