W203 Lab 3 - Crime

Eduard Gelman, Jennifer Mahle, Alena McLucas, Atit Wongnophadol

December 10, 2018

Introduction

We are tasked with understanding factors that cause higher crime rates at the county-level in North Carolina, with the aim of supporting political campaigns in choosing effective policy platforms. Crime rates can play an important role in political campaigns, allowing candidates to demonstrate how they will manage crime and enforce laws - an attractive outcome for local voters. We believe that a detailed, statistical analysis will identify policies that appeal to voters during elections and could also be implemented in order to affect positive change in these counties. Consequently, we focus our analysis on variables that could reasonably be impacted by elected state or local government officials. We analyze crime rates and related data for North Carolina counties in 1987 using ordinary least squares regression to test and explore factors that contribute to higher or lower crime rates.

1.0 Data Understanding

With the goal of understanding how various crime-related factors in North Carolina impact crime rates, we begin with a description of the data set and available variables. The data set contains a single cross section of data for the year 1987 at the county-level. The outcome variable we analyze in this report is "crimes committed per person", which we refer to as the crime rate. The independent variables include the following types:

1. *Punishment Variables*: These variables give proxies for the probability of arrest, probability of conviction, probability of prison sentence and average prison sentence in days. The probability proxy variables come from the FBI's Uniform Crime Reports and are calculated as the number of events (arrest, conviction, prison sentence) divided by the number of offenses.
2. *County Description Variables*: This variable set describes the county including population density, tax revenue per capita, percent of minorities in the county in 1980, and the percent of the county population that are male ages 15-24. The latter two variables are drawn from census data. This also includes police per capita, which uses the FBI's police agency employee counts.
3. *Geography Variables*: They indicate where the county is located, if it is located in western or central NC. Another variable indicates if the county is located in a Metropolitan Statistical Area (MSA), a more urban location.
4. *Wage Variables*: These variables give wages for different job types, like construction, state employees, and finance/insurance/real estate, and come from the North Carolina Employment Security Commission.

We focus on metrics with potential for improvement under new governmental policies. We believe that police per capita and some of the wage variables have the highest potential to be influenced by state or local governmental policies. State/local governments have control over income because they have the ability to alter local tax rates and minimum wages. Accordingly, the greatest wage impacts may be on wages at the lower end of the spectrum, like within the service industry. Wages for government jobs can also be impacted by government budgets, and are therefore something that can be impacted by elected officials. Additionally, police staffing is controlled by tax-funded local budgets and can thus be increased or decreased by funding allocated by state/local governments. Therefore, we use police per capita and wages for lower income jobs, like the service industry, as well as state or local government wages as our key variables of interest. We expect that, holding all else constant, increases in police per capita and in lower wage jobs or state/local government wages would decrease crime rates.

That said, we also expect density and punishment variables impact crime rates but are less actionable by local governments; therefore, we will look at density and punishment variables as our "control" variables. We expect that density and wage will be positively correlated and that density may be one of the underlying causes for higher crime rates in areas with higher wages. Punishment variables may have a similar correlation with density because inner cities may have stricter sentencing and the defendants arrested may have less ability to afford a private attorney. Although government may have some control to manipulate these metrics, they are not as easily impacted as our primary variables and may give less actionable recommendations. For example, population density could be influenced by incentives to build certain housing types, which can impact density, but these changes happen slowly over the course of years or even decades and are not easily attributed to a single politician or campaign.

1.1 Research Question

Next, we pose a research question using the data documentation and problem statement, before performing any exploratory analyses of the data. This is a formal step chosen to help reduce bias in research design and statistical analysis, since "peeking" at data may affect how we specify models and reduce the validity of our results.

We investigate the following research question to help guide policy recommendations:

> **Research Question:** Do crimes committed per person decrease with higher police per capita and with higher wages for lower wage jobs or public sector jobs?

While we believe there are many other factors that contribute to crime rate, we endeavor to specifically evaluate whether increasing police per capita and raising wages for lower wage and public sector jobs are factors that can both help a candidate win an election and serve as impactful policy recommendations for government officials.

Ultimately, the analysis challenged our assumptions and expectations, leading to specific, evidence-based policy recommendations, which are detailed in the "Conclusions and Policy Recommendations" section at the end of the report.

1.2 Regression Specification Strategy

We used the above research question, focusing on the variables of key interest, to come up with specific regressions to run prior to exploratory data analysis. It is well known that wage variables tend to be skewed and generally lend themselves to transformations; however, in this data set we have the average wage by county for specific job types, so these variables may tend to have lower skew than a general wage variable including individuals' wages. That being said, it is still appropriate to log transform these variables because a given dollar amount change can have dramatically different impacts depending on starting salary; therefore, the percentage change in wage is more meaningful than a dollar amount change even in this context. In all regressions outlined below, we use crime rate per capita as the target or dependent variable.

For the first regression, we would like to test the effects of key independent variables on crime rates:

- Police per Capita (*polpc*)
- Weighted average wage for lower wage workers (*wser* (68.5%), *wcon* (5.3%), *wmfg* (26.2%)), with log transform
- Weighted average public sector wages (*wfed* (11.3%), *wsta* (30.3%), *wloc* (58.4%)), with log transform

We collected the population-proportion weights of each labor sector for year 1990 from the Labor Data Search Tool available in the North Carolina's Department of Commerce website. The earliest data was from 1990, and given only 3 years difference we believe it provides good approximation of true 1987 proportions.

Next, we check if controlling for density impacts the results of the above regression because we believe density affects both wages and crime rates. Then, we control for punishment variables, including probabilities of

conviction and arrest, as well as average prison sentence. We will evaluate if this mix of variables interacts well and run additional specifications to refine as needed.

Following the addition of secondary variables of interest, we will run a regression with all the appropriate, available variables. We will exclude the year variable because it is a constant, the offense mix variable because it is too closely related to the crime rate and should be treated as an outcome variable, and the unique county identifier.

In the following section, we start with a brief exploratory data analysis. Additional data analysis will be interspersed with regression evaluation. We evaluate all regression models based on a variety of factors, including goodness of fit metrics like adjusted R^2 and Akaike Information Criterion (AIC). We also conduct various diagnostic plots to test CLM (classical linear model) assumptions and Cook's distance to identify any negative affects impacting our model estimation. After analyzing all preliminary models, we select one of the models to fine-tune, testing additional specifications before analyzing the final model in detail. We then conduct an analysis of omitted variables and our expectations of their impacts. Finally, we conclude with policy recommendations.

2.0 Dataset Overview

In this section, we will look at descriptive statistics, identify any obvious anomalies, and correct any apparent data errors. Additional EDA and data transformations will be performed as necessary in the model specifications section. For now, we proceed with basic data exploratory analysis.

2.1 Load and Inspect Dataset

```
# load libraries
library(car)
library(dplyr)
library(stargazer)
library(lmtest)
library(sandwich)

# load data
dataset <- read.csv("crime_v2.csv", sep=",", header = TRUE)

# visually inspect data type, row count, and data samples
str(dataset)
```

```
## '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 ...
```

```
# show detailed summaries of the first few columns
summary(dataset)[1:7,1:4]
```

```
##      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
```

2.2 Explore and Clean Data

We identify and correct (if applicable) the following data issues:

Issue 1: A visual scan of the data in a tabular format, using Microsoft Excel, tells us to expect 90 county sample observations, but the `read.csv()` function returns 97 rows, with 6 rows showing all `NA` values reported by `summary()`. The culprit is an errant apostrophe character (‘) in the *prbconv* column. This may have been caused by an uncareful analyst - or MIDS employee - saving over the raw data set with the accidentally inserted character, and the error was propagated by saving the file to a .csv format.

There are 6 observations with ‘NA’ values. We removed these excessive rows.

```
# inspect last 6 rows to verify NAs, only show the first columns for brevity
dataset[92:97,1:4]
```

```
##    county year crmrte prbarr
## 92     NA   NA     NA     NA
## 93     NA   NA     NA     NA
## 94     NA   NA     NA     NA
## 95     NA   NA     NA     NA
## 96     NA   NA     NA     NA
## 97     NA   NA     NA     NA
```

```
# delete the last 6 rows
dataset <- dataset[1:91,]

# count NA values, confirm no NA values remain
sum(is.na(dataset))
```

```
## [1] 0
```

Issue 2: The data type for *prbconv*, the probability of conviction, was loaded as a factor. With the other probability variables being numeric, we coerce this variable into numeric type.

```r
# make the probability of conviction variable numeric type.
dataset$prbconv <- as.numeric(as.character(dataset$prbconv))
str(dataset$prbconv)
```

```
##  num [1:91] 0.528 1.481 0.268 0.525 0.477 ...
```

Issue 3: Since the number of unique records in the table is one less than the total number of observations, we know one row was duplicated. We observe and remove the duplicate row.

```r
# count unique and total rows
paste("Number of rows: ", nrow(dataset), "   Number of unique rows: ",
      nrow(unique(dataset)))
```

```
## [1] "Number of rows:  91    Number of unique rows:  90"
```

```r
# preview duplicated rows
dataset[,1:9] %>% filter(county == 193)
```

```
##   county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc   density
## 1    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887 0.8138298
## 2    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887 0.8138298
```

```r
# drop duplicates
dataset = unique(dataset)
```

Issue 4: *prbarr* and *prbconv* have a "top-coding" issue - their maximum values exceed 1. We believe this is due to several factors, including that these features are not true probabilities; rather, they are proxies computed from FBI data. Because there may be more than one arrest per crime, or more convictions than crimes, we do not change the data. Moving forward, we are careful in reasoning about these probability-based variables as they are not completely intuitive.

```r
# notice a top-coded issue for these two variables
summary(dataset$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20495 0.27146 0.29524 0.34487 1.09091
```

```r
summary(dataset$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34422 0.45170 0.55086 0.58513 2.12121
```

Issue 5: There is an anomalous value in *wser*, weekly wages in the service industry.

```r
# show the two highest values in variable wser
round(sort(dataset$wser, decreasing = TRUE)[1:2], 2)
```

```
## [1] 2177.07  391.31
```

The maximum value is over five times as large as the second highest value, which may affect results calculated using this variable. We would expect this observation to have high leverage in future regressions; however, it is not clear this is a data error, so we retain the data point as it may be informative.

2.3 Wage Transformations

Given that we want to test the relationship between crime rate and public sector wages, we construct a variable for the weighted average weekly wage of the three public sector wage variables. As mentioned in the previous section, we collected weight information from an official government source. The three wage

variables of interest include those of federal, state, and local employees, who comprise 11.3%, 30.3%, and 58.4% of the workforce, respectively.

```
# create weighted average public sector wage variable
dataset$avg_public <- dataset$wfed*0.113 + dataset$wsta*0.303 + dataset$wloc*0.584
```

Similarly, policy can make an impact to change wages for lower wage jobs. We construct a variable for the weighted average weekly wage for lower wage jobs, including service, construction, and manufacturing jobs, which comprise 68.5%, 5.3%, and 26.2% of the lower wage workforce, respectively.

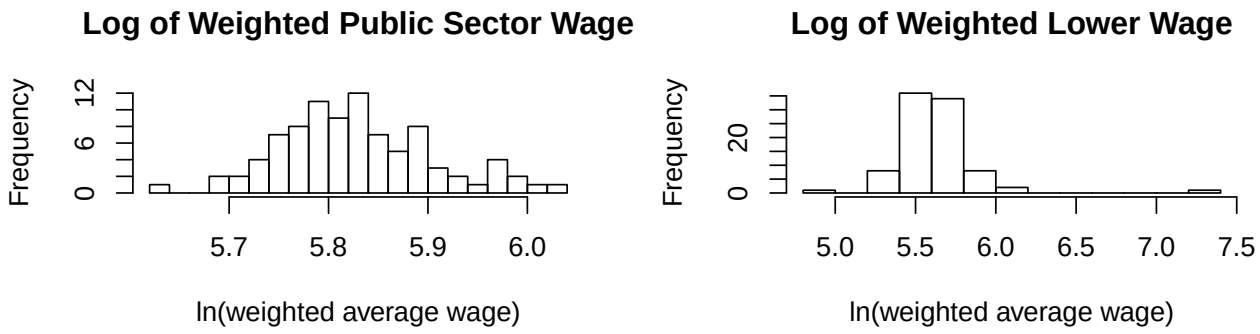
```
# create weighted average lower wage variable
dataset$avg_lwage <- dataset$wser*0.685 + dataset$wcon*0.053 + dataset$wmfg*0.262
```

Then we apply log transformations on these two weighted average variables, associating crime rate change with a marginal percentage change in wage.

```
# perform log transformations on the new weighted average variables
dataset$ln_avg_public = log(dataset$avg_public)
dataset$ln_avg_lwage = log(dataset$avg_lwage)
```

We then visually inspect the histogram of the two log wage variables and show that both exhibit reasonable distributions for modeling purposes.

```
# plot 2 tiled histograms
par(mfrow = c(1,2))
hist(dataset$ln_avg_public, breaks = 15, main = "Log of Weighted Public Sector Wage",
     xlab = "ln(weighted average wage)")
hist(dataset$ln_avg_lwage, breaks = 15, main = "Log of Weighted Lower Wage",
     xlab = "ln(weighted average wage)")
```



3.0 Model Building Process

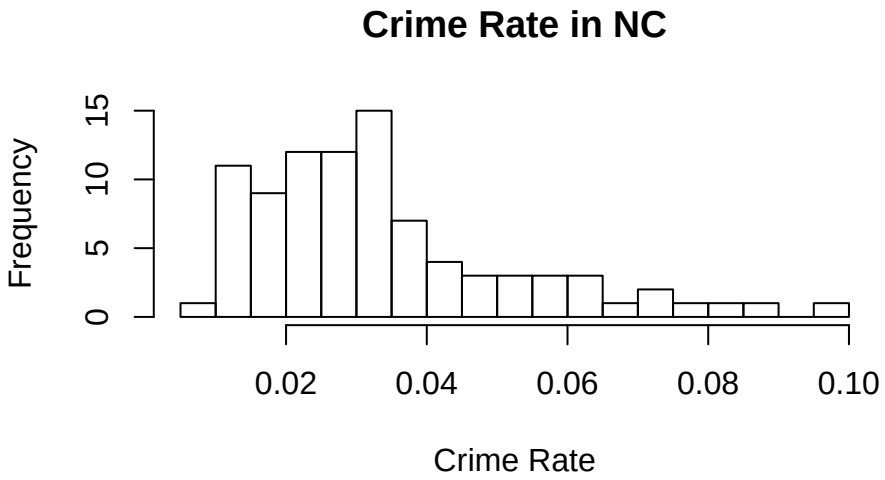
3.1 Outcome Variable

In this section, we conduct a univariate analysis of our outcome variable, crime rate.

```
# summarize crimerate variable
summary(dataset$crmrte)
```

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

```
# plot histogram
hist(dataset$crmrte, breaks = 20, main = "Crime Rate in NC", xlab = "Crime Rate")
```



Most counties have crime rates of around 1 in 30, but there are still a fair number with higher crime rates. While we considered transformations to crime rate that could improve our model, we found no substantive justification in this exploratory analysis.

3.2 Key Independent Variables

Now we examine the three independent variables our research question focused on: police per capita, weighted average wage of lower wage workers, and the weighted average wage of public sector workers. We check the summary statistics for these variables.

```
# create summary statistic table of key variables
stargazer(select(dataset, polpc, ln_avg_lwage, ln_avg_public), type = 'latex',
          median = TRUE, iqr = FALSE, title = 'Summary Statistics - Key Variables',
          digits = 4, star.cutoffs = c(0.05, 0.01, 0.001), header = FALSE)
```

Table 1: Summary Statistics - Key Variables

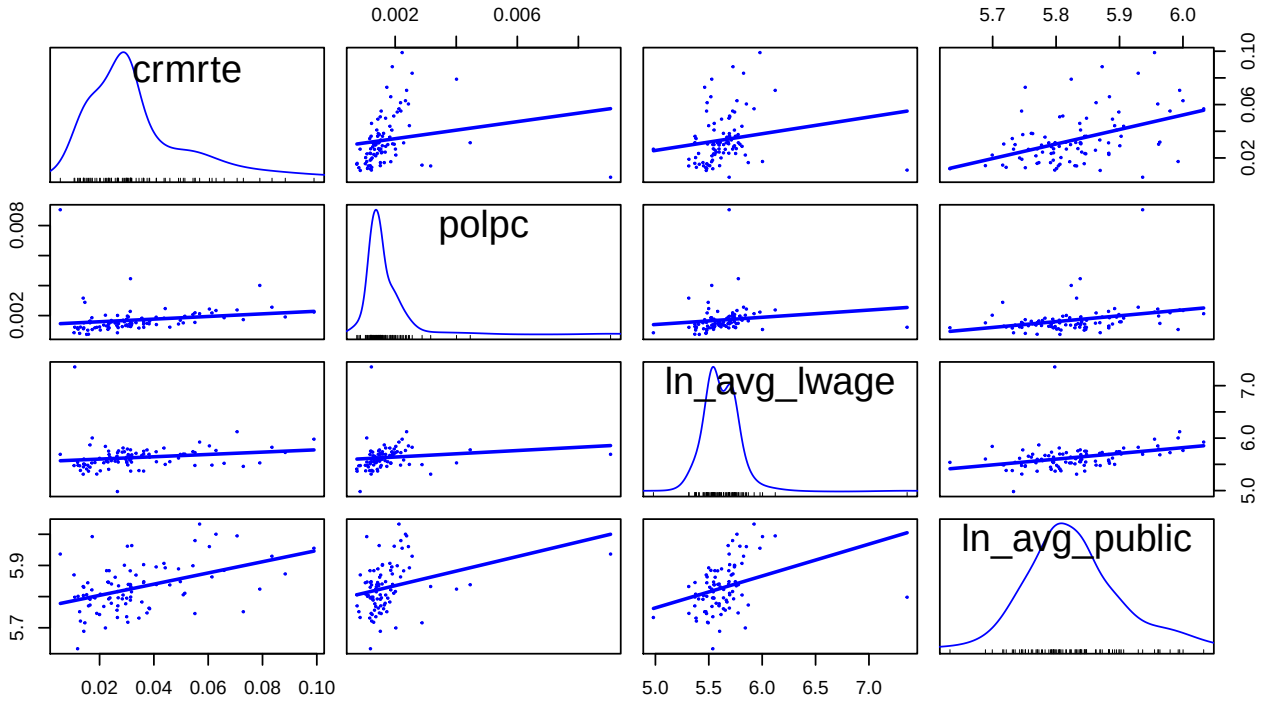
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
polpc	90	0.0017	0.0010	0.0007	0.0012	0.0015	0.0019	0.0091
ln_avg_lwage	90	5.6304	0.2514	4.9813	5.5018	5.5996	5.7261	7.3577
ln_avg_public	90	5.8283	0.0768	5.6328	5.7775	5.8222	5.8720	6.0327

Examining the independent variables:

- Police per capita shows that counties in NC range between 7 and 91 police officers per 10,000 people. The mean and median are pretty close, so we expect the distribution to be fairly symmetric.
- Logged weighted average weekly wage for lower wage jobs contains an outlier, so we will check symmetry. We expect that if we remove it, the distribution will be more symmetric.
- Logged weighted average of public sector weekly wages' mean is very close to the median, so we expect a symmetric distribution.

To examine these variables further, we build scatter plots and histograms.

```
# tiled histograms and pairwise scatterplots
scatterplotMatrix( ~ crmrte + polpc + ln_avg_lwage + ln_avg_public, data = dataset,
                   smooth= FALSE, cex= .2, diagonal=c("density"),
                   main='Relationships between Key Variables')
```



Police per capita has a couple outliers, but the rest of the data points are fairly symmetrically distributed. Aside from one outlying point in logged weighted average weekly wages for lower wage jobs, this variable is also approximately symmetrically distributed. Logged weighted average weekly wage for public sector jobs is a little flatter than expected but has a symmetric distribution and no distinct outliers.

The overall relationship between these three variables and crime rate appears to be positive. This is reasonably expected, since areas with higher crime may have lower wages and more police assigned to address crime.

4.0 Regression Model

4.1 Assumption Testing

Before running the regression models, we consider the following six Classical Linear Model (CLM) assumptions, which apply equally to all models in this report:

1. **Linearity in Parameters**: We believe that we can linearly model the effects of our independent variables on crime rate and reason about associations using unit-coefficients additively. This justifies the use of the multiple regression model specification.
2. **Random Sampling**: This assumption requires that samples be independently and identically distributed. Given that North Carolina has 100 total counties, and we have data on 90 of them, we will assume that there are no systematic differences between the 10 excluded counties and our sample, meaning that "identical distribution" is plausible. On the other hand, we should reasonably suspect that there are clusters in the data, as neighboring counties have more similar characteristics than distant counties. Therefore, we fail the independence assumption. We can correct for these clustering issues with demographic and geographic characteristics in the data set.
3. **No Perfect Multicollinearity**: In the pairwise scatter plots above, we see that there is no perfect multicollinearity between any of our key variables. Additionally, R checks for violations of this assumption when models are built and fails if this assumption is not met.

The remaining three assumptions will be evaluated at the model level:

4. **Zero Conditional Mean / Exogeneity**
5. **Homoskedasticity**
6. **Normality of the Error Term**

4.2 Regression Analysis: Base Model

Given the three variables of key interest and the above EDA, we proceed with our base linear model in accordance with our research question:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 ln\_avg\_lwage + \beta_3 ln\_avg\_public + \epsilon$$

```
# fit a linear model
model_1 <- lm(crmrte ~ polpc + ln_avg_lwage + ln_avg_public, data = dataset)

# create robust standard errors for model
model_1_se <- sqrt(diag(vcovHC(model_1, type = "HC")))

# print the estimated coefficients
stargazer(model_1, title = "Model 1: Base Model with Police Per Capita and Wages",
          align = TRUE, no.space=TRUE, dep.var.labels=c("Crime Rate"),
          se = list(model_1_se), add.lines=list(c("AIC",round(AIC(model_1),1))),
          omit.stat=c("LL","ser","f"), type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Model 1: Base Model with Police Per Capita and Wages
## ===========================================
##                     Dependent variable:
##                 ---------------------------
##                          Crime Rate
## -------------------------------------------
## polpc                      0.700
##                           (4.215)
## ln_avg_lwage               0.001
##                           (0.009)
## ln_avg_public            0.105***
##                           (0.026)
## Constant                 -0.585***
##                           (0.151)
## -------------------------------------------
## AIC                       -469.9
## Observations                90
## R2                         0.197
## Adjusted R2                0.169
## ===========================================
## Note:          *p<0.05; **p<0.01; ***p<0.001
```

The regression results from the Base Model indicate that there may be some relationship between police per capita and crime rate and some relationship between log of weighted average weekly public sector wage and crime rate. However, only logged weighted average public sector wages are reported as statistically significant. The Adjusted R^2 reports a modest 17% of variation in crime rate being explained by the three independent variables, with an AIC of -470.

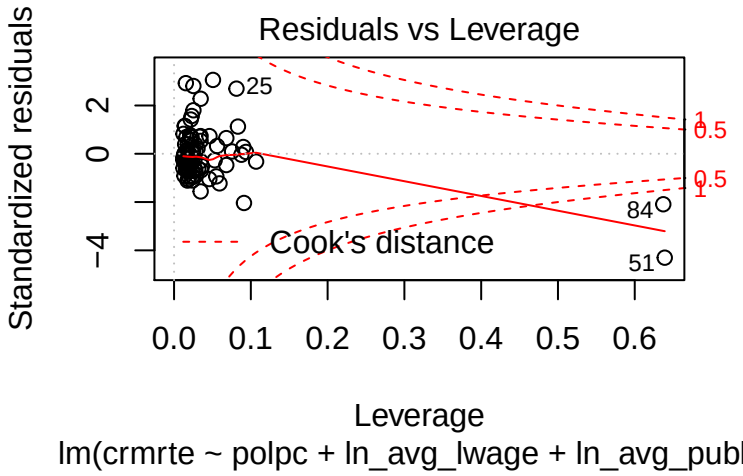
Interestingly, the model seems to indicate that there is a positive association between police per capita and crime rate. This is the opposite of the effect we predicted in our research question, as we thought more police would make areas safer. However, the positive relationship between police per capita and crime rate seems reasonably intuitive - areas with higher crime rates may actively recruit more police officers in response, leading to this positive relationship.

Similarly, the positive relationship between the log of weighted average weekly public sector wage and crime rate is the opposite effect we predicted in our research question, as we thought lower wage would be associated with higher crime rates. However, the relationship in this model suggests that the variation due to wage difference may have been confounded by other effects. For example, it is possible that areas with higher wages tend to have higher density which is positively associated with crime rate.

The log of weighted average weekly wage for lower wage jobs has a small coefficient and is also not statistically significant. We may have to control for omitted variables, like density or the urban indicator, to see the ceteris paribus relationship between lower wage jobs and crime rate.

Cook's Distance Plot

```
plot(model_1, which = 5) # residuals vs leverage
```

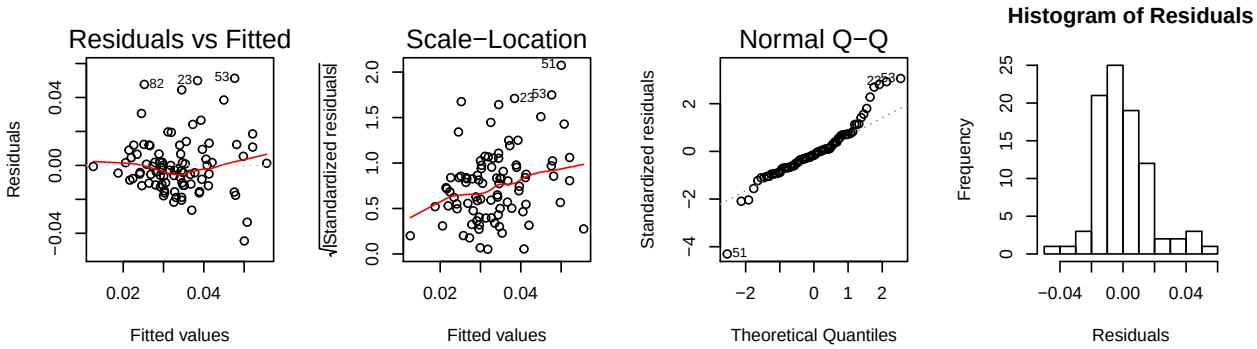


The observations 51 and 84 have high leverage and are influential, as their Cook's distance measures exceed 1. We take note that the estimates of our model coefficients may be influenced by these two data points.

Assumptions

Before evaluating our next model, we check assumptions 4-6 of the CLM:

```
par(mfrow = c(1,4))
plot(model_1, which = 1) # residuals vs fitted plot
plot(model_1, which = 3) # scale-location plot
plot(model_1, which = 2) # normal qq plot
hist(model_1$residuals, breaks = 7, main = "Histogram of Residuals",
     xlab = "Residuals") # histogram of model residuals
```



4. **Zero Conditional Mean / Exogeneity**: The residuals fluctuate around 0, with data points being a little more scattered at higher fitted values. This assumption is reasonably satisfied.

5. **Homoskedasticity**: Conduct a Breusch-Pagan test to check for heteroskedasticity, in conjunction with a scale-location plot.

```
# conduct a Breusch-Pagan test
paste("Test for Homoskedasticity: p-value = ", round(bptest(model_1)$p.value,5))
```

```
## [1] "Test for Homoskedasticity: p-value =  0.00067"
```

With $p < 0.05$, the null hypothesis is rejected with some evidence that heteroskedasticity is present. Further, the scale-location plot shows dispersion that tends to increase with higher fitted values. Therefore, the assumption for homoskedasticity does not hold. We use robust standard errors in response to this violation.

6. **Normality of the Error Term**: The relationship between standardized residuals and theoretical quantiles is not linear, indicating a non-symmetrical distribution of the error terms. The normality assumption is therefore not met.

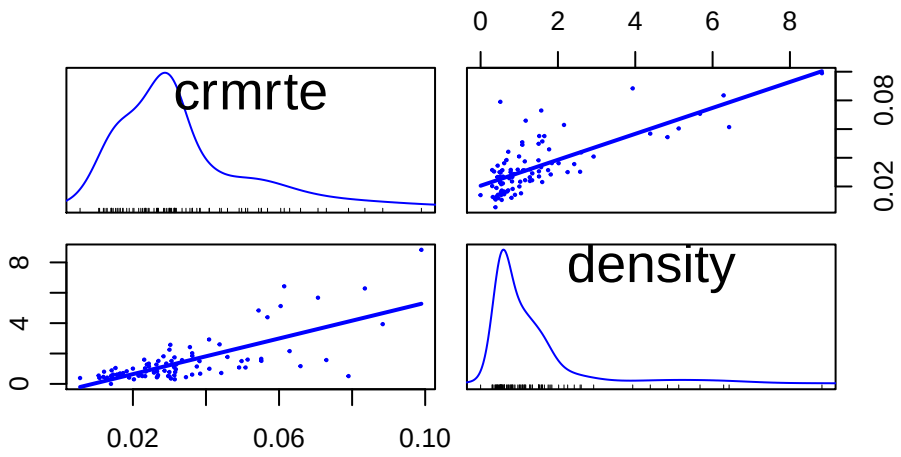
Since there is a clear need to include control variables to address the obvious confounds, we will not yet interpret practical significance of these estimates. In the next models, we iteratively include independent variables to help control for confounding effects.

4.3 Regression Analysis: Base Model + Density Control

Next, we check how controlling for population density changes the results of the above regression. We use density as the first, major control variable because we believe that population density is a main confounder of crime rates and can help us reason about adjustments to police presence and wage variables while holding population density constant.

Let's examine the relationship between crime rate and population density.

```
# plot density and bivariate scatterplot of crime rate vs density
scatterplotMatrix( ~ crmrte + density, data = dataset,
                   smooth= FALSE, cex= .2, diagonal=c("density"),
                   main='Crime Rate and Population Density')
```



We note that population density is strongly, positively correlated with crime rate. This justifies its inclusion as a control variable. Now, we fit a linear model with the addition of density:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 ln\_avg\_lwage + \beta_3 ln\_avg\_public + \beta_4 density + \epsilon$$

```
# fit a linear model
model_2 <- lm(crmrte ~ polpc + ln_avg_lwage + ln_avg_public + density, data = dataset)

# create robust standard errors for model
model_2_se <- sqrt(diag(vcovHC(model_2, type = "HC")))

# print the estimated coefficients
stargazer(model_2, title = "Model 2: Base Model + Density",
          align = TRUE, no.space=TRUE, dep.var.labels=c("Crime Rate"),
          se = list(model_2_se), add.lines=list(c("AIC",round(AIC(model_2),1))),
          omit.stat=c("LL","ser","f"), type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Model 2: Base Model + Density
## ===========================================
##                     Dependent variable:
##                 ---------------------------
##                         Crime Rate
## -------------------------------------------
## polpc                     0.784
##                          (2.808)
## ln_avg_lwage             -0.007*
##                          (0.003)
## ln_avg_public             0.020
##                          (0.021)
## density                  0.009***
##                          (0.001)
## Constant                 -0.060
##                          (0.125)
## -------------------------------------------
## AIC                      -518.4
## Observations               90
```

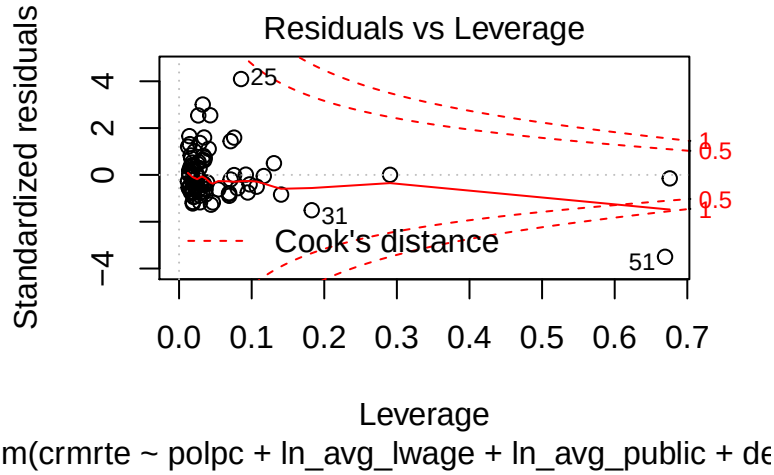
```
## R2                          0.542
## Adjusted R2                 0.521
## ============================================
## Note:          *p<0.05; **p<0.01; ***p<0.001
```

From this model, we find that density has some positive estimated impact on crime rate. We also find that it lessens the impact of police per capita and public sector wages, as both coefficients are reduced in magnitude. Interestingly, the coefficient on lower wage jobs becomes negative, satisfying our original intuition about this variable having a negative relationship with crime rate.

Meanwhile, the model fitness increased from 17% to 52% and AIC is reduced from -476 to -518, indicating that density has a large role as an explanatory variable for crime rate and adding control variables from this data set in order to refine our model is reasonable. This approach could yield a better model if applied iteratively with similar goodness-of-fit increases.

Cook's Distance Plot

```
plot(model_2, which = 5) # residuals vs leverage
```

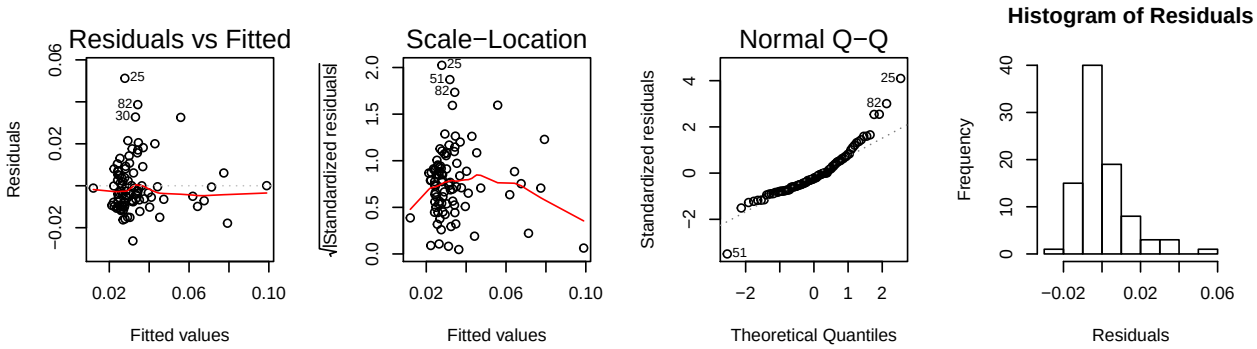


We found that observation 51 has high leverage and is influential, as its Cook's distance measure exceeds 1. We take note that the estimates of our model coefficients may be influenced by this data point.

Assumptions

Before interpretation and choosing next steps, we will test CLM 4-6 of the second model:

```
par(mfrow = c(1,4))
plot(model_2, which = 1) # residuals vs fitted plot
plot(model_2, which = 3) # scale-location plot
plot(model_2, which = 2) # normal qq plot
hist(model_2$residuals, breaks = 7, main = "Histogram of Residuals",
     xlab = "Residuals") # histogram of model residuals
```



4. **Zero Conditional Mean / Exogeneity**: Despite a few lower fitted residuals above 0, most fluctuate around 0. Therefore, zero conditional mean is approximately satisfied.

5. **Homoskedasticity**: Conduct a Breusch-Pagan test to check for heteroskedasticity, in conjunction with a scale-location plot.

```
# Breusch-Pagan test
paste("Test for Homoskedasticity: p-value = ", round(bptest(model_2)$p.value,5))
```

```
## [1] "Test for Homoskedasticity: p-value =  0.01308"
```

With $p < 0.05$, the null hypothesis is rejected with some evidence that heteroskedasticity is present. Further, the scale-location plot shows dispersion that is quite different across the fitted values. Therefore, the assumption for homoskedasticity does not hold and we should use robust standard errors for modeling.

6. **Normality of the Error Term**: The error terms are not normal and have not changed much from the previous model, exhibiting similar residual skew.

Interpreting the coefficients remains difficult since the magnitude and significance of both coefficients on wage variables are very small, limiting the practical significance of these estimates. Similarly, police per capita has a fairly large coefficient, but has a very poor standard error. While it is tempting to stop here and conclude that increased police presence may actually increase crimes rate, and that increasing wages for lower paid jobs decreases crime rate, we believe that additional controls will improve our model.

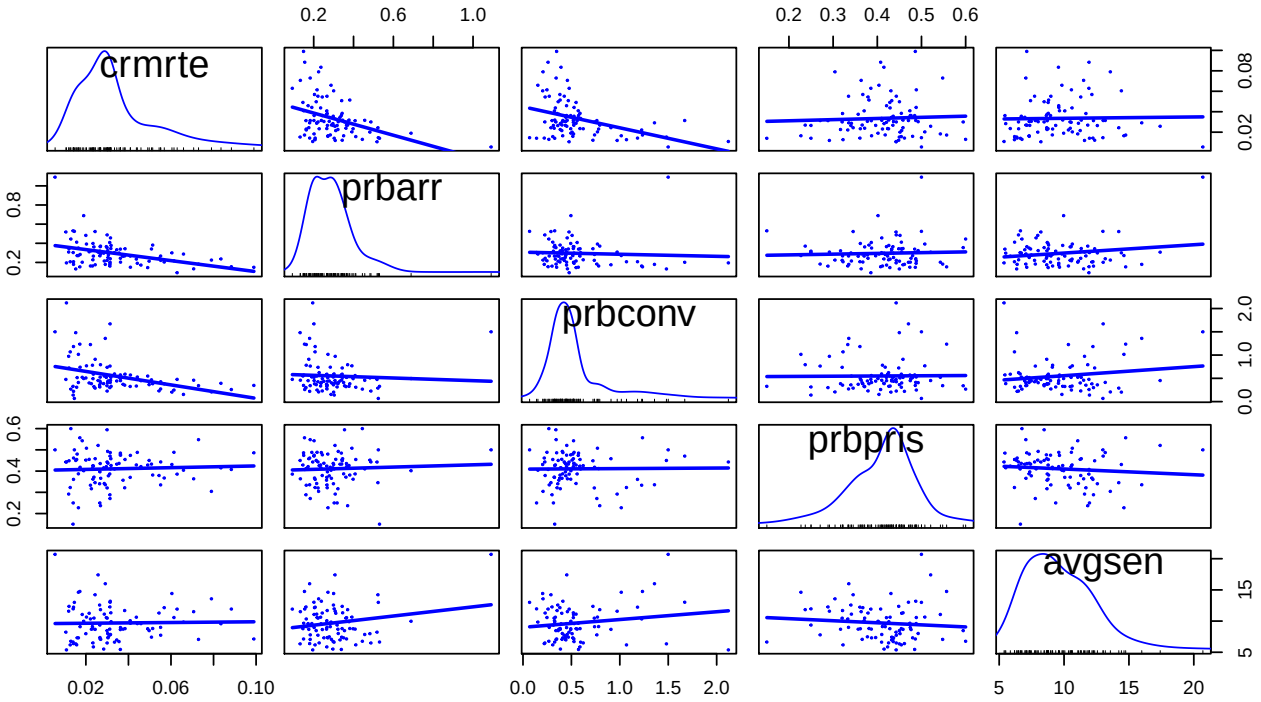
We proceed by creating the third model with more robust control variables in order to refine our estimates.

4.4 Regression Model: Base + Density Control + Punishment Control

In the third model, we want to look at punishment variables because they are proxies for more aggressive policing that may also have confounding effects with police per capita. We will include independent variables for probabilities of conviction and arrest, as well as average prison sentence.

We evaluate the potential control variables in histograms and pairwise scatter plots.

```
# plot density and bivariate scatterplots
scatterplotMatrix( ~ crmrte + prbarr + prbconv + prbpris + avgsen, data = dataset,
                   smooth= FALSE, cex= .2, diagonal=c("density"),
                   main='Relationships between Crime Rates and Punishment Variables')
```



We notice crime rate seems to be negatively associated with increasing proxy-probabilities of arrest and conviction likelihoods. This implies that more aggressive policing and legal climates seem to lower crime rate. This is intuitive - if residents perceive that there are more arrests and convictions, this may deter further criminal behavior. The same is not observed for the prison sentencing proxy or for the length of sentences, perhaps because these events are more temporally and geographically removed from everyday life - that is, arrests and convictions are more immediate and visible than sentencing and sentence lengths.

Fitting this model with new control variables:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 ln\_avg\_lwage + \beta_3 ln\_avg\_public +$$
$$+\beta_4 density + \beta_5 prbarr + \beta_6 prbconv + \beta_7 prbpris + \beta_8 avgsen + \epsilon$$

```
# fit a linear model
model_3 <- lm(crmrte ~ polpc + ln_avg_lwage + ln_avg_public + density + prbarr +
                prbconv + prbpris + avgsen, data = dataset)

# create robust standard errors for model
model_3_se <- sqrt(diag(vcovHC(model_3, type = "HC")))

# print the estimated coefficients
stargazer(model_3, title = "Model 3: Base Model + Density + Punishment Variables",
          align = TRUE, no.space=TRUE, dep.var.labels=c("Crime Rate"),
          se = list(model_3_se), add.lines=list(c("AIC",round(AIC(model_3),1))),
          omit.stat=c("LL","ser","f"), type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Model 3: Base Model + Density + Punishment Variables
## ========================================
##                       Dependent variable:
##                   ---------------------------
```

```
##                         Crime Rate
## ------------------------------------------
## polpc                     6.731**
##                           (2.049)
## ln_avg_lwage              0.0001
##                           (0.005)
## ln_avg_public              0.006
##                           (0.019)
## density                  0.006***
##                           (0.001)
## prbarr                   -0.057***
##                           (0.014)
## prbconv                  -0.019***
##                           (0.004)
## prbpris                    0.003
##                           (0.014)
## avgsen                    -0.0004
##                          (0.0004)
## Constant                   0.007
##                           (0.103)
## ------------------------------------------
## AIC                       -549.3
## Observations                90
## R2                         0.703
## Adjusted R2                0.673
## ==========================================
## Note:         *p<0.05; **p<0.01; ***p<0.001
```

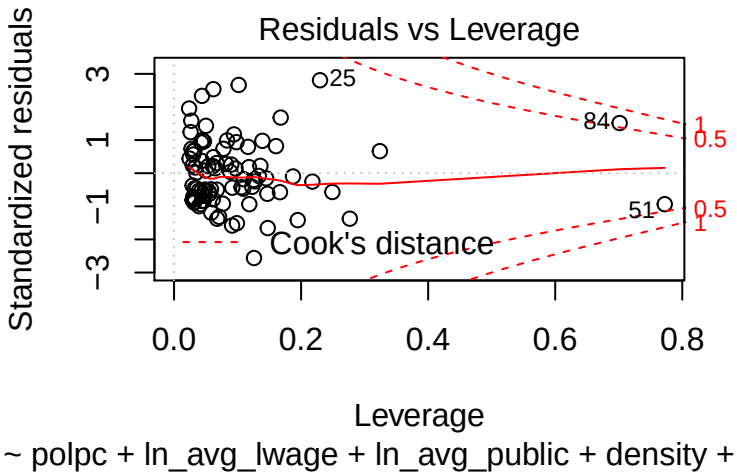
The magnitude of the police per capita coefficient jumps up to 6.73 while the impact of wages remains negligible. Interestingly, the sign of weekly average wage for lower wage jobs flipped back to positive, which is contrary to our expectation that wages for lower wage jobs have a negative relationship to crime rate. Though, it does so with no statistical significance.

The impact of density is less than in Model 2, as the coefficient was reduced slightly, but the direction and significance remain stable. The probability of arrest and conviction have a negative impact on crime rate. The probability of prison sentence has a small but positive coefficient, where we would have expected a higher probability of prison sentence to be a crime deterrent. The average sentence in days has some impact on reducing crime rate (i.e., an additional year in sentence length, 365x0.0004, is related to a reduction in crime rate by 0.146). This is a practically significant result.

By including additional punishment variables as measures of deterrence, the model fit improved with adjusted R^2 of 67% compared to 52% in the second model. We see another drop in AIC from -518 to -549, and while we are still seeing positive returns from our model, they appear to be diminishing. We will see how our model is impacted if we add all other viable explanatory variables in the next model.

Cook's Distance Plot

```r
plot(model_3, which = 5) # residuals vs leverage
```

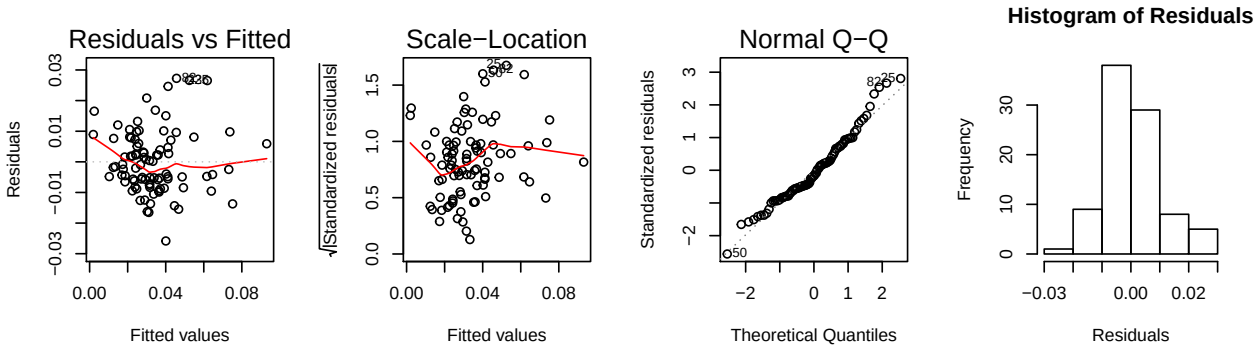


All data points are within the Cook's distance of 1, which doesn't present concern for our coefficient estimates.

Assumptions

Check assumptions 4-6 of the CLM:

```
par(mfrow = c(1,4))
plot(model_3, which = 1) # residuals vs fitted plot
plot(model_3, which = 3) # scale-location plot
plot(model_3, which = 2) # normal qq plot
hist(model_3$residuals, breaks = 7, main = "Histogram of Residuals",
     xlab = "Residuals") # histogram of model residuals
```



4. **Zero Conditional Mean / Exogeneity**: There is a departure near the low end of fitted values, but this may be due to a low number of lower fitted values. The residuals fluctuate around 0, so the zero conditional mean is approximately satisfied.

5. **Homoskedasticity**: Conduct a Breusch-Pagan test to check for heteroskedasticity, in conjunction with a scale-location plot.

```
# Breusch-Pagan test
paste("Test for Heteroskedasticity: p-value = ", round(bptest(model_3)$p.value,5))
```

```
## [1] "Test for Heteroskedasticity: p-value =  0.00498"
```

With $p < 0.05$, the null hypothesis is rejected with some evidence that heteroskedasticity is present. The scale-location plot shows dispersion that is varied across the fitted values, so we use robust standard errors for model testing.

6. **Normality of the error term**: Looking at the normal q-q plot, the error terms are approximately normal with a bit more fluctuation in the left-most values. The histogram of residuals agrees. This is

the best fit so far, with the normality assumption being approximately met.

4.5 Regression Model: Kitchen Sink

To demonstrate the robustness of our previous models, we will include most other variables in a final model. We want to see if the magnitude and direction of our coefficients, as well as their statistical significance, remains consistent after adding additional explanatory variables. We exclude the following variables from this regression:

- *county* because it is a unique, arbitrary identifier
- *year* because it is constant, and it therefore doesn't add anything explanatory
- *mix* because it is a secondary outcome variable

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 ln\_avg\_lwage + \beta_3 ln\_avg\_public + \beta_4 density + \beta_5 prbarr + \beta_6 prbconv + \beta_7 prbpris + \beta_8 avgsen +$$

$$\beta_9 wtuc + \beta_{10} wtrd + \beta_{11} wfir + \beta_{12} taxpc + \beta_{13} west + \beta_{14} central + \beta_{15} urban + \beta_{16} pctmin80 + \beta_{17} pctymle + \epsilon$$

```
# fit a linear model for almost all of our variables
model_all <- lm(crmrte ~ polpc + ln_avg_lwage + ln_avg_public + density + prbarr +
                  prbconv + prbpris + avgsen + wtuc + wtrd + wfir + taxpc + west +
                  central + urban + pctmin80 + pctymle, data = dataset)

#create robust standard errors
model_all_se <- sqrt(diag(vcovHC(model_all, type = "HC")))

# print the estimated coefficients
stargazer(model_all, title = "Kitchen Sink Model", align = TRUE, no.space=TRUE,
          dep.var.labels=c("Crime Rate"), se = list(model_all_se),
          add.lines=list(c("AIC",round(AIC(model_all),1))), omit.stat=c("LL","ser","f"),
          type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Kitchen Sink Model
## ===========================================
##                     Dependent variable:
##                 ---------------------------
##                           Crime Rate
## -------------------------------------------
## polpc                     6.816***
##                            (1.673)
## ln_avg_lwage              -0.002
##                            (0.003)
## ln_avg_public              0.010
##                            (0.019)
## density                   0.006***
##                            (0.001)
## prbarr                    -0.055***
##                            (0.011)
## prbconv                   -0.017***
##                            (0.003)
## prbpris                    -0.001
##                            (0.010)
## avgsen                     -0.0004
```

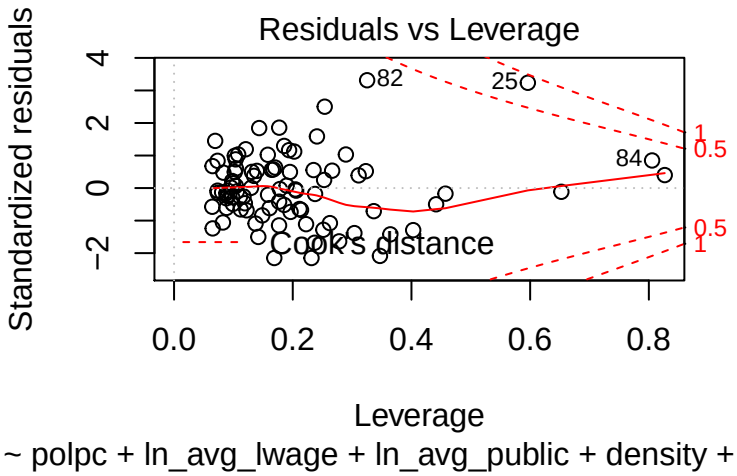
```
##                          (0.0004)
## wtuc                      0.00002
##                          (0.00001)
## wtrd                      0.0001
##                          (0.0001)
## wfir                     -0.00004
##                          (0.00002)
## taxpc                     0.0002
##                          (0.0001)
## west                      -0.004
##                           (0.003)
## central                   -0.004
##                           (0.002)
## urban                     -0.002
##                           (0.006)
## pctmin80                 0.0003***
##                          (0.0001)
## pctymle                  0.098***
##                           (0.028)
## Constant                  -0.030
##                           (0.109)
## ------------------------------------------
## AIC                       -588.6
## Observations                90
## R2                         0.843
## Adjusted R2                0.806
## ==========================================
## Note:         *p<0.05; **p<0.01; ***p<0.001
```

Comparing our linear regression model that includes controls for density and punishment variables, we see that the there is still some room to explain more variation in crime rates by controlling for more variables. The adjusted R^2 for this model is 0.81, as compared to the previous model's 0.67. AIC has also received another boost from -550 to -589.

However, coefficients are very similar to the previous model, suggesting that the additional variables in this model are not changing the magnitude of our key explanatory variables or control variables. Police per capita, population density, and two of our punishment control variables remain statistically significant as well. More discussion follows after a comparison of all 4 models.

Cook's Distance Plot

```r
plot(model_all, which = 5) # residuals vs leverage
```

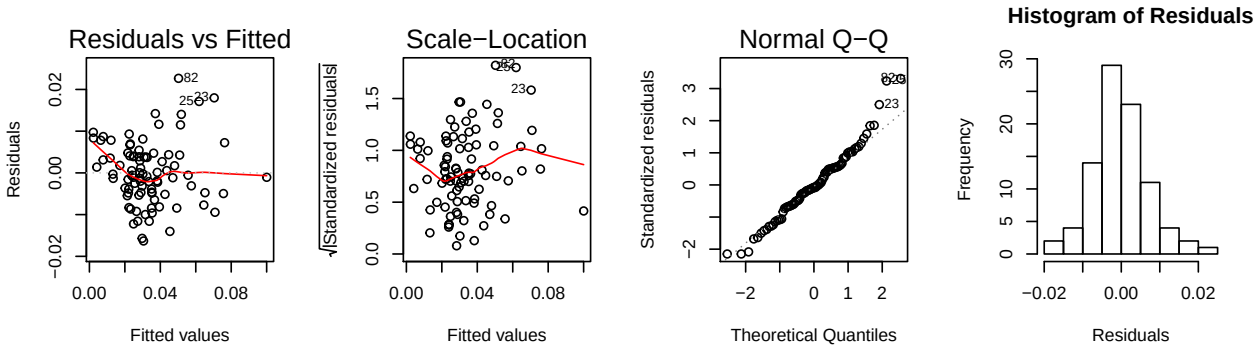


All data points are within the Cook's distance of 1, so they don't present concerns for our estimates.

Assumptions

Check assumptions 4-6 of the CLM:

```
par(mfrow = c(1,4))
plot(model_all, which = 1) # residuals vs fitted plot
plot(model_all, which = 3) # scale-location plot
plot(model_all, which = 2) # normal qq plot
hist(model_all$residuals, breaks = 7, main = "Histogram of Residuals",
     xlab = "Residuals") # histogram of model residuals
```



4. **Zero Conditional Mean / Exogeneity**: The residuals fluctuate around 0, with a slight upturn towards 0 in the plot with lower fitted values. Overall, the zero conditional mean is satisfied in this case.

5. **Homoskedasticity**: Conduct a Breusch-Pagan test to check for heteroskedasticity, in conjunction with a scale-location plot.

```
# Breusch-Pagan test
paste("Breusch-Pagan for Heteroskedasticity: p-value = ",
      round(bptest(model_all)$p.value,5))
```

```
## [1] "Breusch-Pagan for Heteroskedasticity: p-value =  0.07035"
```

With p > 0.05, we fail to reject the null hypothesis of homoskedasticity. The scale-location plot has a high amount of standard error fluctuation, and so we conclude that heteroskedasticity is present. This assumption does not hold, and we use robust standard errors as best practice.

6. **Normality of the Error Term**: Looking at the normal q-q plot, the error terms are approximately normal and the best fit as compared to previous models.

4.6 The Regression Table

We create a table to compare the results of our four regression models:

```
# nicely formatted table to compare regression models
stargazer(model_1, model_2, model_3, model_all, title="Regression Comparison",
          align = TRUE, no.space=TRUE, dep.var.labels=c("Crime Rate"),
          add.lines=list(c("AIC",round(AIC(model_1),1), round(AIC(model_2),1),
                           round(AIC(model_3),1), round(AIC(model_all),1))),
          omit.stat=c("LL","ser","f"), type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Regression Comparison
## ================================================
##                         Dependent variable:
##                -----------------------------------
##                              Crime Rate
##                   (1)      (2)       (3)       (4)
## ------------------------------------------------
## polpc           0.700    0.784   6.731***  6.816***
##                (1.932)  (1.467)   (1.622)   (1.456)
## ln_avg_lwage    0.001    -0.007   0.0001    -0.002
##                (0.008)  (0.006)   (0.006)   (0.005)
## ln_avg_public 0.105***   0.020     0.006     0.010
##                (0.026)  (0.023)   (0.019)   (0.017)
## density                 0.009*** 0.006***  0.006***
##                         (0.001)   (0.001)   (0.001)
## prbarr                           -0.057*** -0.055***
##                                   (0.011)   (0.009)
## prbconv                          -0.019*** -0.017***
##                                   (0.004)   (0.003)
## prbpris                            0.003    -0.001
##                                   (0.014)   (0.012)
## avgsen                            -0.0004   -0.0004
##                                  (0.0005)  (0.0004)
## wtuc                                        0.00002
##                                            (0.00001)
## wtrd                                        0.0001
##                                            (0.00004)
## wfir                                       -0.00004
##                                            (0.00003)
## taxpc                                       0.0002
##                                            (0.0001)
## west                                        -0.004
##                                             (0.004)
## central                                     -0.004
##                                             (0.003)
## urban                                       -0.002
##                                             (0.006)
## pctmin80                                   0.0003***
##                                            (0.0001)
## pctymle                                     0.098*
##                                             (0.044)
## Constant      -0.585***  -0.060    0.007    -0.030
##                (0.144)  (0.128)   (0.107)   (0.095)
```

```
## ------------------------------------------------
## AIC            -469.9    -518.4   -549.3    -588.6
## Observations     90        90       90        90
## R2              0.197    0.542     0.703     0.843
## Adjusted R2     0.169    0.521     0.673     0.806
## ================================================
## Note:                  *p<0.05; **p<0.01; ***p<0.001
```

We examine coefficients across models:

- Police per capita is only statistically significant in the last two models where the key difference in significance was the inclusion of probability of arrest and conviction. The latter two models suggest that, holding those probabilities constant, higher police per capita coincides with a higher crime rate. It is important to note that the police per capita coefficients are positive across models. It does not make sense that hiring more police officers leads to more crime, so our results may imply that high crime rates drive higher police per capita. Alternatively, this may start to build reasonable case for the existence of "over-policing" in these counties - where a high rate of police officers causes local antagonism and negatively feeds back in the form of increased crime. The available data do not sufficiently capture evidence in favor of, or against, over-policing. More research is needed, perhaps gathering information from town hall meetings that may reveal local discussion on the topic of police antagonism.
- The weighted-averaged and unaveraged wage variables generally do not have statistical significance, with the one exception in our first model. They do not have practical significance, as in Model 4, holding all else constant a 1% increase in average weekly wage for lower waged workers decreases crime rate by 0.002%. However, they are not consistently positive or negative and tend to have high standard errors. Therefore, we expect that wages - alone, or including available controls - are not good predictors for understanding crime rates.
- Population density is statistically significant for all of our models and is therefore a key explanatory variable. The coefficient suggests that a 1 unit increase in population density per square mile increases crime rate by about 0.006%, which is a practically significant change. The coefficient is consistently positive, suggesting there is a positive relationship between people per square mile and crime rate.
- Probability of arrest and conviction are both statistically significant in both models to which they have been added. They consistently have a negative relationship with crime rate. The coefficients have larger magnitude, which in part makes up for the fact that these variables are percent proxies in decimal form, but they are also very practically significant. The coefficient for probability of arrest implies that a 1% increase in probability of arrest decreases crime rate by about 0.055%, which makes sense that higher chance of arrest would deter crime.

Comparing each adjusted R^2 value, each subsequent model explains more of the variation in the data set, which is expected as more variables are included. The model fitness increases with diminishing margin by 0.35, 0.15, and 0.14 respectively. A continuously decreasing AIC supports this conclusion.

5.0 Model Selection and Fine-Tuning

Based on our model building process, analysis, and regression table above, we select Model 3 as our initial "best model". We use this section to build on our analysis to enhance the model.

5.1 Problems and Tuning Choices

Leaning on the model building and interpretation thus far, we have the following problems:

1. The least visible and immediate punishment variables seemed to be redundant in terms of having other more powerful variables included in the regression.
2. Including wage variables, having controlled for other effects, does not yield significant coefficients.

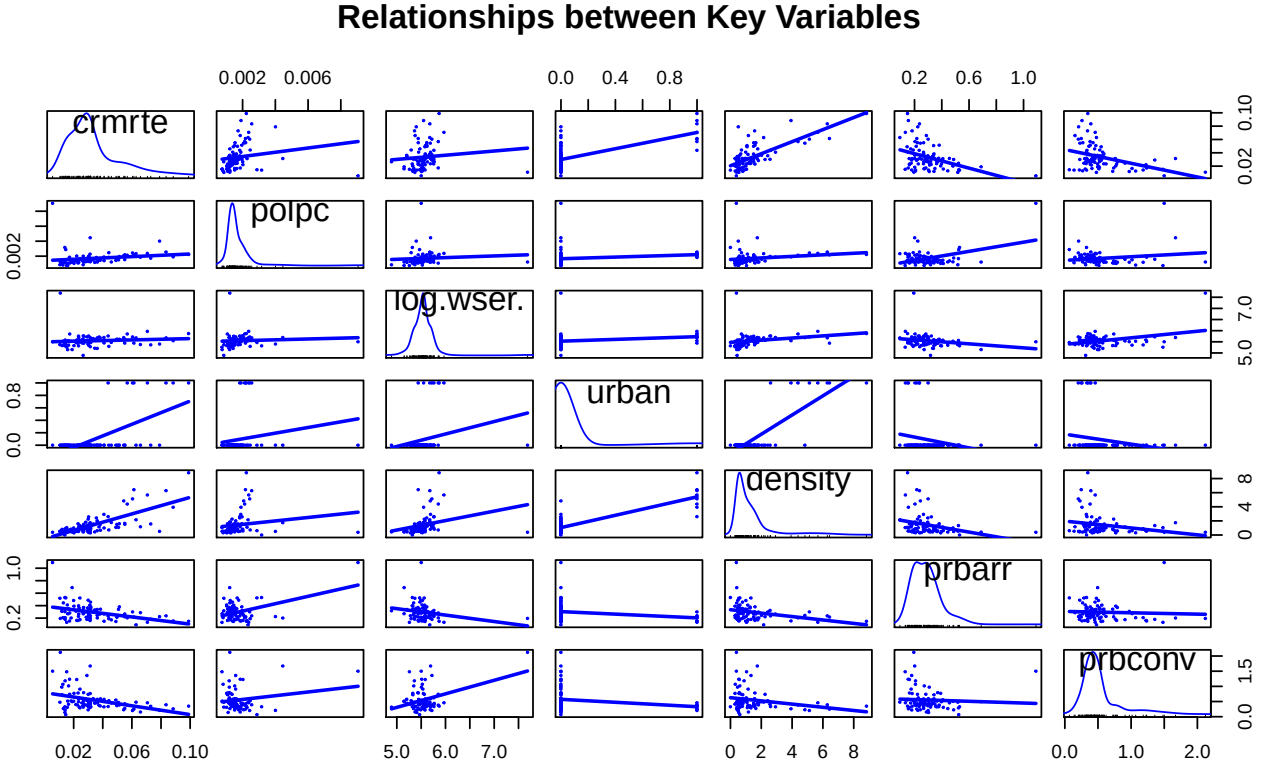
We address these concerns by:

1. Dropping *prbpris* and *avgsen* as redundant variables, keeping *prbarr* and *prbconv*, since they seem to measure "deterrence from the criminal justice system process" reasonably well on their own.
2. Trying an interaction term with one lower income variable and the urban indicator, as we suspect that the wage variables may be insignificant because we used a weighted average. The statistical insignificance may also be because wages in urban areas tend to be higher to account for higher costs of living, but crime rates in cities also tend to be higher. We use a log transform of service industry wages because it is representative of lower income wages that could be increased with higher minimum wage legislation.

5.2 Model Tuning

First, we visualize the relationships of all variables to be included in the Augmented version of Model 3.

```
# tiled histograms and pairwise scatterplots
scatterplotMatrix( ~ crmrte + polpc + log(wser) + urban + density + prbarr + prbconv,
                   data = dataset, smooth= FALSE, cex= .2, diagonal=c("density"),
                   main='Relationships between Key Variables')
```



We observe that there are no obvious departures from linearity between any pair of variables that includes *crmrte*, and no large correlations between independent variables. We expect the urban indicator may capture a similar effect as density, so we will only use the urban indicator in the following specification.

We test the following specification, with the log-transformed wage variable, the urban indicator and an interaction between them:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 log(wser) + \beta_3 urban + \beta_4 log(wser) * urban + \beta_5 prbarr + \beta_6 prbconv + \epsilon$$

```r
# fit a linear model
model_3a <- lm(crmrte ~ polpc + log(wser) + urban + log(wser)*urban + prbarr + prbconv,
               data = dataset)

# create robust standard errors for model
model_3a_se <- sqrt(diag(vcovHC(model_3a, type = "HC")))

# print the estimated coefficients
stargazer(model_3a, title = "Model 3(a): Wage, Urban Indicator, and Interaction Term",
          align = TRUE, no.space=TRUE, dep.var.labels=c("Crime Rate"),
          se = list(model_3a_se), add.lines=list(c("AIC",round(AIC(model_3a),1))),
          omit.stat=c("LL","ser","f"), type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Model 3(a): Wage, Urban Indicator, and Interaction Term
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              Crime Rate
## ------------------------------------------------
## polpc                         7.581***
##                               (1.443)
## log(wser)                      0.003
##                               (0.003)
## urban                          0.040
##                               (0.151)
## prbarr                        -0.069***
##                               (0.012)
## prbconv                       -0.023***
##                               (0.004)
## log(wser):urban               -0.003
##                               (0.026)
## Constant                      0.034*
##                               (0.016)
## ------------------------------------------------
## AIC                            -539.5
## Observations                     90
## R2                              0.654
## Adjusted R2                     0.629
## ================================================
## Note:              *p<0.05; **p<0.01; ***p<0.001
```

Despite parsing out the effects of differences in wages between urban and rural areas, and trying to use the log of a single lower income wage variable (*log(wser)*) rather than a weighted average, none of the wage related variables are significant. In addition, the R^2 in model 3 is 70.3%, whereas this specification has an R^2 that is around 5% lower, indicating this model explains approximately 5% less of the variation in crime rate.

Further, we look at the practical significance of the wage variables in this model and interpret the meaning of the coefficients. Including the interaction term between *log(wser)* and *urban* gives different intercepts and different slopes for urban and rural areas. The specification indicates that non-urban areas have a 1% increase in service industry wages that results in a 0.003% increase in crime rate; whereas in urban areas, a 1% increase in wages results in a 0% change in crime rates since the coefficients for *log(wser)* and the interaction with *urban* cancel each other out. This demonstrates there is a lack of practical significance in the wage variables as well as a lack of statistical significance, so we will not select this as our final model.

Because we will not use this as our final model, we forego checking assumptions.

For our next test, we will make the following changes in an attempt to improve and finalize our model:

1. Remove the insignificant wage variables, including the urban indicator and the interaction term.
2. Add density back in, as this was highly significant in Model 3 and it better captures the county level differences in density that are indicative of crime rates than the urban indicator.

5.3 Final Model Fit and Diagnostics

Accordingly, we proceed to fit the following regression:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \epsilon$$

```
# fit a linear model
model_3b <-  lm(crmrte ~ polpc + density + prbarr + prbconv, data = dataset)

# create robust standard errors for model
model_3b_se <- sqrt(diag(vcovHC(model_3b, type = "HC")))

# print the estimated coefficients
stargazer(model_3b, title = "Model 3(b): Final Model",
          align = TRUE, no.space=TRUE, dep.var.labels=c("Crime Rate"),
          se = list(model_3b_se), add.lines=list(c("AIC",round(AIC(model_3b),1))),
          omit.stat=c("LL","ser","f"), type = "text", star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Model 3(b): Final Model
## ===========================================
##                     Dependent variable:
##                 ---------------------------
##                           Crime Rate
## -------------------------------------------
## polpc                     6.311***
##                            (1.811)
## density                   0.006***
##                            (0.001)
## prbarr                    -0.057***
##                            (0.013)
## prbconv                   -0.019***
##                            (0.004)
## Constant                  0.042***
##                            (0.005)
## -------------------------------------------
## AIC                        -556.3
## Observations                 90
## R2                          0.699
## Adjusted R2                 0.685
## ===========================================
## Note:         *p<0.05; **p<0.01; ***p<0.001
```

First, we see that all variable coefficients show expected directional effect on crime rates and have high statistical significance. The adjusted R^2 shows that the 68.5% of variation in crime rate can be explained by these variables. The AIC value is also reasonable compared to all the previous models.

We also see that an increase in police per capita coincides with an increase in the crime rate, which is likely due to the fact that higher crime areas require more police per capita. In addition, with higher density, we expect higher crime rates, as urban areas are known to have more crime. The punishment variables, probability of arrest and probability of conviction, both have negative coefficients, indicating that counties with higher rates of arrest and conviction have lower crime rates.

To evaluate practical significance, we examine summary statistics.

```
# summary statistic table of the dependent and independent variables
stargazer(select(dataset, crmrte, polpc, density, prbarr, prbconv), type = 'text',
          median = TRUE, iqr = FALSE, digits = 4, star.cutoffs = c(0.05, 0.01, 0.001),
          title = 'Summary Statistics - Dependent and Independent Variables')
```

```
##
## Summary Statistics - Dependent and Independent Variables
## ===========================================================================
## Statistic N   Mean  St. Dev.   Min   Pctl(25) Median Pctl(75)  Max
## ---------------------------------------------------------------------------
## crmrte    90 0.0335  0.0189  0.0055   0.0206  0.0300  0.0402  0.0990
## polpc     90 0.0017  0.0010  0.0007   0.0012  0.0015  0.0019  0.0091
## density   90 1.4357  1.5216  0.00002  0.5472  0.9792  1.5693  8.8277
## prbarr    90 0.2952  0.1377  0.0928   0.2049  0.2715  0.3449  1.0909
## prbconv   90 0.5509  0.3542  0.0684   0.3442  0.4517  0.5851  2.1212
## ---------------------------------------------------------------------------
```

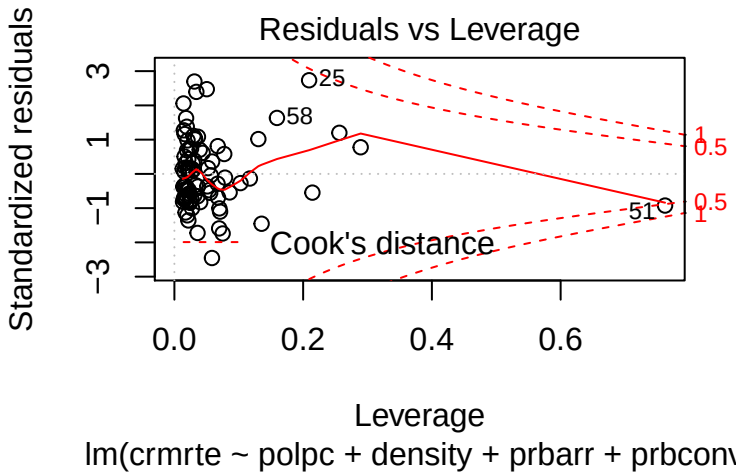
Based on our final model and summary statistics of key variables, we attempt to understand practical significance of explanatory variables.

- The model suggests that 10 additional police officers per 10,000 people, about one standard deviation away from the mean, is associated with a 0.0063 (6.31 x 0.0010) increase in crimes committed per person. Given the median crime rate of 0.03, 300 crimes per 10,000 people, the police per capita factor is not only statistically significant but also practically significant.
- In terms of density, given one additional person per square mile, less than one standard deviation away from the mean, crime rate increases by 0.006. This translates to 60 more crimes per 10,000 people. Given the median crime rate of 0.0300, 300 crimes per 10,000 people, the density factor is both statistically and practically significant.
- A 10% increase in the probability of arrest is associated with a 0.0057 (0.057 x 0.1) reduction in crime rate, or 57 fewer crimes per 10,000 people. A 10% increase in the probability of conviction is associated with a 0.0019 (0.019 x 0.1) reduction in crimes committed per person, or 19 fewer crimes per 10,000 people. Therefore, both deterrent factors *prbarr* and *prbconv* are practically significant.

Now, we turn to the analysis of Cook's distance and the assumptions 4-6 of the CLM:

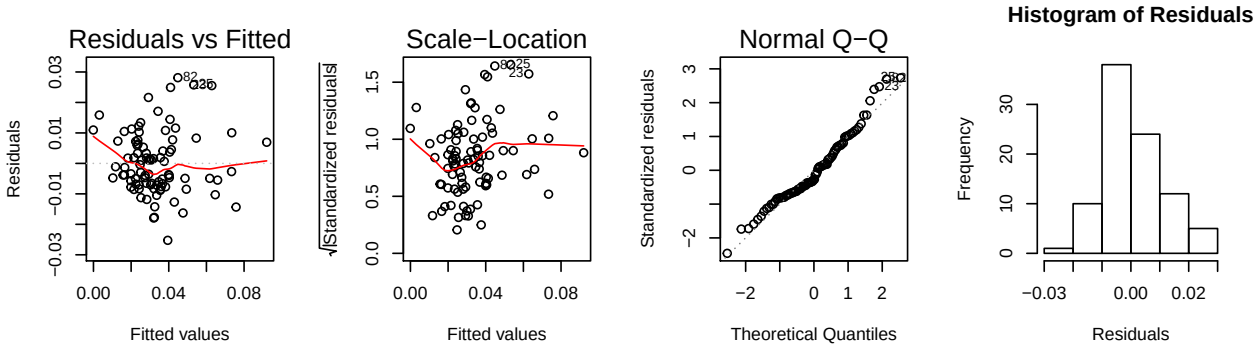
Cook's Distance Plot

```
plot(model_3b, which = 5) # residuals vs leverage
```



All data points are within the Cook's distance of 1, so they don't present concern on our estimates.

```
par(mfrow = c(1,4))
plot(model_3b, which = 1) # residuals vs fitted plot
plot(model_3b, which = 3) # scale-location plot
plot(model_3b, which = 2) # normal qq plot
hist(model_3b$residuals, breaks = 7, main = "Histogram of Residuals",
     xlab = "Residuals") # histogram of model residuals
```



4. **Zero Conditional Mean / Exogeneity**: Overall, the residuals fluctuate around 0, so the zero conditional mean assumption is satisfied. There is some departure towards lower fitted values, but this may be due to a low volume of points in that region. There is the additional effect of crime rates needing to be positive, so a model would ideally over-predict when crime rate is near 0 since a prediction less than 0 would make no sense. Finally, the model will mostly be used to suggest policy for areas with medium or high crime rates, so slightly worse performance in low-crime counties is less concerning. All-together, this assumption is approximately satisfied, and the regions of concern for violations are the least important in our policy recommendations.

5. **Homoskedasticity**: Conduct a Breusch-Pagan test to check for heteroskedasticity, in conjunction with a scale-location plot.

```
# Breusch-Pagan test
paste("Test for Heteroskedasticity: p-value = ", round(bptest(model_3b)$p.value,5))
```

```
## [1] "Test for Heteroskedasticity: p-value =  0.00181"
```

With $p < 0.05$, the null hypothesis is rejected with some evidence that heteroskedasticity is present. The scale-location plot shows dispersion that fluctuates across the bulk of the fitted values, with the main departure appearing with mid-range fitted values. To remain conservative, we reject the assumption of homoskedasticity and proceed with heteroskedasticity-robust standard errors.

6. **Normality of the Error Term**: The error terms appear approximately normal when evaluating the normal q-q plot and residual histogram, with the main departure from normality occurring for a small number of observations on the high end. This may be an artifact of the truncation of crime rate at 0, which will favor larger positive residuals. The normality of the error term assumption is approximately satisfied.

6.0 Omitted Variables

It is reasonable to believe there are omitted factors that drive or deter crimes. Based on debate and discussion, we came up with the top 5 variables we will examine as potential confounding factors on our final model.

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \epsilon$$

6.1 **Unemployment rate**:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \beta_5 unemployment + \epsilon$$
$$unemployment = \alpha_0 + \alpha_1 polpc + \alpha_2 density + \alpha_3 prbarr + \alpha_4 prbconv + v$$

By adding unemployment as $\beta_5 unemployment$ to our regression, we expect that $\alpha_1 > 0$ (*polpc*), $\alpha_2 < 0$ (*density*), $\alpha_3 > 0$ (*prbarr*) and $\alpha_4 > 0$ (*prbconv*). We believe unemployment rate has a positive relationship with police presence due to joblessness leading to crime. Larger presence of law enforcement units increase arrest and conviction as well. The relationship between unemployment rate and density is less clear, but we believe low density areas tend to associate with higher unemployment due to having less job opportunities.

If $\beta_5 > 0$, where high crime rate is associated with high unemployment rate, then OMVB $\beta_5\alpha_1 > 0$, $\beta_5\alpha_2 < 0$, $\beta_5\alpha_3 > 0$, and $\beta_5\alpha_4 > 0$. This means (i) the positive coefficient on *polpc* is more extreme (moving away from zero) than it should be, (ii) the positive coefficient on *density* is less important (moving closer to zero) than it should be, and (iii) the negative coefficients on *prbarr* and *prbconv* are less important (moving closer to zero) than they should be.

In other words, if unemployment is a strong driver behind crime rate, then it would have lessened the importance of police presence to some degree. On the contrary, it would have made density a more critical factor on crime, and the chances of arrest and conviction even stronger deterrents to crime.

6.2 **Education levels**:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \beta_5 education\_level + \epsilon$$
$$education\_level = \alpha_0 + \alpha_1 polpc + \alpha_2 density + \alpha_3 prbarr + \alpha_4 prbconv + v$$

By adding education level as $\beta_5 education\_level$ to our regression, we expect that $\alpha_1 < 0$, $\alpha_2 > 0$, $\alpha_3 < 0$ and $\alpha_4 < 0$. We believe educational level has a negative relationship with police presence as we believe higher educated citizens require less policing. The relationship between education level and density tends to be positive, as many higher educational institutions are located in urban areas. With higher education levels and more employment opportunities, chances of arrest and conviction get lower.

If $\beta_5 < 0$, lower crime associates with higher education levels, then $\beta_5\alpha_1 > 0$, $\beta_5\alpha_2 < 0$, $\beta_5\alpha_3 > 0$, and $\beta_5\alpha_4 > 0$. This means (i) the positive coefficient on *polpc* is more extreme (moving away from zero) than its population value, (ii) the positive coefficient on *density* is less important (moving closer to zero) than it should be, and the negative coefficients on *prbarr* and *prbconv* are less important (moving closer to zero) than they should be.

In other words, if education level is a strong driver behind crime rate, then it would have lessened the importance of police presence on crime rate to some degree. On the contrary, it would have made density a

stronger driver for crime. And lastly it would have made the chances of arrest and conviction even stronger crime deterrents.

6.3 **Income to cost of living ratio**:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \beta_5 income\_to\_cost + \epsilon$$
$$income\_to\_cost = \alpha_0 + \alpha_1 polpc + \alpha_2 density + \alpha_3 prbarr + \alpha_4 prbconv + v$$

By adding income to cost of living ratio as $\beta_5 income\_to\_cost$ to our regression, we expect that $\alpha_1 < 0$, $\alpha_2 < 0$, $\alpha_3 < 0$ and $\alpha_4 < 0$. We believe higher income to cost ratio has a negative relationship with police presence as higher earners will commit less crime and require less policing. The relationship between income ratio and density is assumed to be negative because we believe wealthier citizens tend to live in suburban, less dense, areas. And the higher the income to cost ratio, the lower the chances of being arrested and convicted as wealthier citizens tend to have better job prospects and can better prepare to defend themselves against criminal charges.

If $\beta_5 < 0$, lower crime rate associates with higher income to cost ratio, then $\beta_5\alpha_1 > 0$, $\beta_5\alpha_2 > 0$, $\beta_5\alpha_3 > 0$, and $\beta_5\alpha_4 > 0$. The positive coefficients on *polpc* and *density* are more extreme (moving away from zero) than they should be, and the negative coefficients on *prbarr* and *prbconv* are less important (moving closer to zero) than they should be.

In other words, if income to cost of living ratio is a strong driver behind crime rate, then it would have lessened the importance of police presence and density on crime rate to some degree. On the contrary, it would have made the chances of arrest and conviction even stronger crime deterrents.

6.4 **Illiteracy levels**:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \beta_5 illiteracy + \epsilon$$
$$illiteracy = \alpha_0 + \alpha_1 polpc + \alpha_2 density + \alpha_3 prbarr + \alpha_4 prbconv + v$$

By adding income to cost of living ratio as $beta_5 illiteracy$ to our regression, we expect that $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_3 > 0$ and $\alpha_4 > 0$. We believe illiteracy tends to be associated with higher chance for committing crime and hence requiring more policing. We also believe that illiterate citizens tend to reside in high density area in which job opportunities are more likely for illiterate people. The more literacy level, the higher the chances of being arrested and convicted as illiterate citizens are more likely to commit crime due to financial need and will be less prepared to defend themselves against criminal charges.

If $\beta_5 > 0$, higher crime rate is associated with higher illiteracy level, then $\beta_5\alpha_1 > 0$, $\beta_5\alpha_2 > 0$, $\beta_5\alpha_3 > 0$, and $\beta_5\alpha_4 > 0$. The positive coefficients on *polpc* and *density* are more extreme (moving away from zero) than they should be, and the negative coefficients on *prbarr* and *prbconv* are less important (moving closer to zero) than they should be.

If illiteracy level is a strong driver behind crime rate, then it would have lessened the importance of police presence and density on crime rate to some degree. On the contrary, it would have made the chances of arrest and conviction even stronger crime deterrents.

6.5 **Homelessness**:

$$\widehat{crmrte} = \beta_0 + \beta_1 polpc + \beta_2 density + \beta_3 prbarr + \beta_4 prbconv + \beta_5 homeless + \epsilon$$
$$homeless = \alpha_0 + \alpha_1 polpc + \alpha_2 density + \alpha_3 prbarr + \alpha_4 prbconv + v$$

By adding homelessness as $beta_5 homeless$ to our regression, we expect that $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_3 > 0$ and $\alpha_4 > 0$. Similar to our reasoning in the analysis of illiteracy level, we believe homelessness has the same directional impacts on our model's variables. Homelessness tends to associate with crime, requiring more

police presence. Homeless people tend to reside in high density areas to sustain their living, and with less income opportunities they are more likely to commit crime and face arrests and convictions.

If $\beta_5 > 0$, higher crime rate is associated with higher homelessness, then $\beta_5\alpha_1 > 0$, $\beta_5\alpha_2 > 0$, $\beta_5\alpha_3 > 0$, and $\beta_5\alpha_4 > 0$. The positive coefficients on *polpc* and *density* are more extreme (moving away from zero) than they should be, and the negative coefficients on *prbarr* and *prbconv* are less important (moving closer to zero) than they should be.

In other words, if homelessness is a strong driver behind crime rate, then it would have lessened the importance of police presence and density on crime rate to some degree. On the contrary, it would have made the chances of arrest and conviction even stronger crime deterrents.

7.0 Conclusion and Policy Recommendations

In building our models, we expected to see that counties with higher police per capita, higher average weekly wage for lower wage workers, and higher average weekly income for public sector workers would have lower crime rates. However, initial results from the model building process revealed that (i) police per capita has the opposite effect to what we hypothesized and (ii) wages do not have a significant impact on crime rates. We then further refined the model and identified the final model to include density, probability of arrest, probability of conviction and police per capita as the explanatory variables. Our final model is as follows:

$$\widehat{crmrte} = 0.042 + 6.311polpc + 0.006density - 0.057prbarr - 0.019prbconv + \epsilon$$

The model suggests that police per capita and density have positive relationship with crime rate, whereas the probabilities of arrest and conviction have a deterrent effect on crime. All explanatory variables are statistically and practically significant, and can explain variation of crime rate with an adjusted R^2 of 68.5%

Based on the model results, we believe the following policy recommendations can help a political candidate win an election, and if implemented once in office, they can help reduce crime rate:

- Because increasing the probability of arrest decreases crime rate, we recommend supporting an increase in arrests, holding the police force size constant. Additionally, marketing increased arrests may, in itself, help to decrease crime rate.
- Likewise, an increase in conviction rate also impacts crime rate. There may be certain crimes that have higher conviction rates to prioritize, which should be further researched.
- Because density of people per square mile has a positive relationship with crime rate, political candidates may consider focusing on suburban and rural development projects as a long-term strategy to decrease density in urban areas.
- An increase in the number of police officers may not yield an intended effect given its positive relationship with crime rate. Based on the analysis of omitted variables, we believe favorable socioeconomic factors such as better employment rate, higher education and lower homelessness may have confounding effects on police per capita. If these variables truly have impact on crime rate, they would also lessen the importance of police per capita on crime rate. If this is the case, then elected officials should look into making these socioeconomic conditions better for the citizens. All of the omitted variables explored here would be compelling areas for further analysis.