Lab 3 Report

Priscilla Burity, Oscar Linares, Alissa Stover

Due: 4/14/2020

Table of Contents

# Introduction

**Which policies are most promising in reducing crime rate: those that target punishing crimes (criminal justice policy) or reducing the need to commit crimes (economic policy)?**

To answer this question, we use a cross-section of data on crime statistics and related factors (e.g., economic indicators) for a selection of counties in North Carolina. These data were first used in a study by Cornwell and Trumball, researchers from the University of Georgia and West Virginia University (C. Cornwell and W. Trumball (1994), “Estimating the Economic Model of Crime with Panel Data,” Review of Economics and Statistics 76, 360-366.). Most of the data are from 1987, except for some demographic variables from the 1980 Census. Our task to help a political campaign understand the determinants of crime and generate local policy suggestions to curb it. As such, we focus on variables that can be influenced by policy.

We also include a range of contextual variables that control for local characteristics influencing crime and the efficacy of these policies. In contrast to our variables of main focus, these are generally not actionable (e.g., demographic and geographic indicators).

The data were limited in the following ways: (1) they only contained one year of cross-sectional data, (2) there are important omitted variables, and (3) they provide little variation in the distribution of policy variables. Having data from one year alone means that the context of that specific year is an uncontrolled factor. For example, one specific event that could have affected both crime and economic variables was the [Black Monday crash in October of that year](https://en.wikipedia.org/wiki/Black_Monday_(1987)). Issues (2) and (3) made it difficult to isolate the true effect of policies. Firstly, omitted variables like unemployment would likely affect crime as well as other explanatory variables. Secondly, variation in policies across counties is needed to identify their impact on crime and some policies, such as as minimum wage and tax levels, were similar across all counties in the 1980s. Thus, in general, these issues limits the causal claims we can make.

That said, we believe produced a model that help guide policy decisions aimed to reduce crime across these counties.

The main finding was that that acting on crime deterrents within the criminal justice system (like arrests and convictions) is important when it comes to preventing crime. In fact, for a 10% rise in the probability of arrests and convictions individually, the negative impact on crime rate hovers between 5-7% for arrests and between 4-5% for convictions. This result has a material practical significance.

Even though we could not find much support for the importance of economic policy variables to prevent crimes, this could be due to this dataset’s lack of a good metric for this kind of policy.

## Actionable Variables

### Criminal Justice Policy

Many of these criminal justice policies are used as deterrents of crime, with the assumption that people will be less likely to commit crimes because they want to avoid the punishment.

The 3 variables below are dubbed “probabilities”. However, this is a misnomer as these variables are truly ratios. We interpret these as measures of the assertiveness of the criminal justice system in a county.

#### Probability of arrest

If we find that the assertiveness of arrest is associated with crime rate (with higher punishment associated with lower crime), we may consider stricter policies around police practice that encourage more arrests.

* prbarr: number of arrests for every crime reported

#### Probability of conviction

If we find that the assertiveness of conviction is associated with crime rate (with higher conviction rates associated with lower crime), we may consider stricter policies around court practices so that more arrests lead to convictions.

* prbconv: number of convictions for every arrest made

#### Probability of sentencing

If we find that the assertiveness of sentencing is associated with crime rate (with higher sentencing rates associated with lower crime), we may consider stricter policies around judicial practices so that more convictions lead to sentencing.

A prison sentence depends on the severity of the crime and as such counties with more severe crime would see elevated numbers for this variable (and vice versa). Thus, this variable may be more limited in its explanatory power than arrests and convictions and will not be a prime candidate for our model.

* prbpris: number of prison sentences for every conviction

#### Severity of punishment

If we find that the severity of punishment is associated with crime rate (with greater severity associated with lower crime), we may consider stricter policies around judicial practices so that people receive more severe punishments. Similarly to prison sentence rate, the length of someone’s sentence depends on the severity of the crime. Thus, this variable may also be more limited in its explanatory power than arrests and convictions and will not be a prime candidate for our model.

* avgsen: average sentence in days

#### Number of police officers per capita

The number of police officers per capita (polpc) is a very intuitive example of crime deterrent, as it in theory increases the likelihood of being caught in the act of the crime and is also a measure of the capacity of the State to enforce the law. However, cross-sectional data limitations prevent us from understanding the relationship between police officers per capita and crime rate because the impact of increasing the number of police per capita might not be realized in the same period. One can expect that counties with larger crime rates in a year *t* would consider necessary to have a larger number of police officers on the streets in year *t*, which can imply a positive correlation between presence of police and crime rates that does not necessarily go one way in terms of causality.

As our job is to advise a policy maker, it’s very important that we feel safe about the causal interpretation of our findings. We need to avoid potential reverse causality effects between independent and dependent variables. Thus, in this exercise we opted not to include police per capita as a candidate policy variable nor covariate.

### Economic Policy

We will use the wage variables to operationalize economic policy as these variables could be reshaped with economic policy. For example, if increased wages are found to be related significantly to crime, our candidate could consider strategies such as raising the minimum wage in an attempt to lower crime.

Additionally, empirical research, such as [this study](https://www.journals.uchicago.edu/doi/abs/10.1086/449290), suggests that increasing a worker’s wage can deter the worker from committing crimes. Completely omitting these variables could introduce bias on the other estimated relationships.

The following are weekly wages in different sectors:

* wcon: construction
* wtuc: transportation, utilities, communication
* wtrd: wholesale and retail trade
* wfir: finance, insurance, real estate
* wser: services
* wmfg: manufacturing
* wsta: state employees
* wloc: local government employees

The federal wage variable will be grouped with other wages in much of our exploratory data analyses for easier comparison but it is not related to minimum wage policy but rather to cost of living. Federal employee wages are controlled by the federal government and are [adjusted for a locality’s cost of living](https://www.opm.gov/policy-data-oversight/pay-leave/salaries-wages/fact-sheets/). If we find a relationship between federal wages and crime, one could predict that changing the cost of living may affect crime.

Policy recommendations that could be made from this variable are limited by the fact that we are missing important economic variables such as inequality. Raising the cost of living alone would not have the intended effect on crime without a combined effort to address inequality. For example, areas with high cost of living and high rates of inequality would expect to see more crime, as those who are *in need* would find more reason to take from those who have *more than they need*.

Although it may be complicated to interpret our quantification of the relationship between crime and this variable, we imagine that omitting this variable could introduce bias in other estimated relationships. Thus, it is important to attempt to measure the relationship between federal wages and crime to test the robustness of other variables.

* wfed: federal employees

The following variable represents tax revenue per capita, but does not differentiate between different revenue streams (income, sales, property, or businesses). Many taxes are set at the state and federal level and so the level of variation at the county-level may be limited. Since this variable could represent a mix of effects, the causal relationship would be difficult to interpret but we will include this variable as an important covariate to test other relationships.

* taxpc: tax revenue per capita

## Contextual Variables

The following variables describe county characteristics that provide important information about how policies could work across different contexts. Although they are not actionable, they strengthen our ability to make causal inferences.

### Types of Crime

If an area has mostly petty crimes, one would imagine that some of the punishment (e.g., fines) could be less visible and perhaps less of a deterrent than arresting people. This difference could change the relationship between our crime policy variables and crime rate.

* mix: offense mix: face-to-face/other

### Demographics

#### Urban/rural dwellers

Higher density areas would see more interaction between people, which could drive up crime. Also, social bonds and norms in small communities could act as deterrents of crime.

* density: people per square mile

#### Minority status

It is well known that people of minority status (compared to Caucasians) are more likely to be involved in the criminal justice system (this doesn’t mean they actually commit more crimes, just that they are arrested more often) and as such this is a key covariate in our analysis.

* pctmin80: percent minority, 1980

#### Gender & Age

Young males are more likely to enter the criminal justice system; this is another demographic variable that we would expect relates to crime.

* pctymle: percent young male, 1980

### Geography

The following variables identifies a county’s region within the state, with the assumption that counties cluster geographically in terms of culture and other characteristics we are not explicitly measuring. Western North Carolina is along the Appalachian mountains and qualitatively different from the rest of the state. For example, [it has fewer universities and colleges](https://www.google.com/maps/search/universities+colleges+north+carolina/@35.6210094,-82.5051642,7z).

* west: =1 if in western N.C.
* central: =1 if in central N.C.
* urban: =1 if in SMSA

# Model Building Process

## Data Cleaning & Exploratory Data Analysis

# import libraries   
library(tidyverse) # for data import, manipulation, viz  
library(corrplot) # for correlation matix  
library(stargazer) # visualize model fit  
library(skimr) # generate summary statistics  
library(car) # statistics   
library(lmtest) # linear modeling  
library(olsrr) # evaluating OLS regression   
library(sandwich) # correcting heteroskedasticity violation  
library(grid) # to arrange ggplot figures into a grid   
library(gridExtra) # to arrange ggplot figures into a grid   
library(wesanderson) # color palettes

# import data  
data <- read\_csv("crime\_v2.csv", n\_max = 91)

#### File-level data checks

We have 91 observations and 25 columns in the raw data. All 25 columns were read in as numeric columns.

print(paste("Number of records:", dim(data)[1]))

## [1] "Number of records: 91"

print(paste("Number of variables:", dim(data)[2]))

## [1] "Number of variables: 25"

##### The Where & When of these data

It looks like we have 1 observation for each of the counties – except for county #193. We noticed that these rows appear to be exact duplicates.

data %>% filter(county == 193)

## # A tibble: 2 x 25  
## county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 193 87 0.0235 0.266 0.589 0.423 5.86 0.00118 0.814 28.5  
## 2 193 87 0.0235 0.266 0.589 0.423 5.86 0.00118 0.814 28.5  
## # … with 15 more variables: west <dbl>, central <dbl>, urban <dbl>,  
## # pctmin80 <dbl>, wcon <dbl>, wtuc <dbl>, wtrd <dbl>, wfir <dbl>,  
## # wser <dbl>, wmfg <dbl>, wfed <dbl>, wsta <dbl>, wloc <dbl>, mix <dbl>,  
## # pctymle <dbl>

The attributes associated with each record led us to believe that the [county IDs are county FIPS codes](https://www.lib.ncsu.edu/gis/countfips). We supplement some of our data quality checks with internet searches on counties according to their FIPS code, without assuming that there is perfect alignment with the county IDs.

The year column is an expected constant (1987). We can drop this column in future data processing steps.

table(data$year, useNA = "always")

##   
## 87 <NA>   
## 91 0

We create a new data frame dropping our superfluous column (year and polpc) and row (duplicate of county 193). Now we have 90 rows and 24 numeric variables. Since we have one row per county, we appear to be missing information from 10 of the 100 counties in North Carolina.

data2 <- data %>% distinct() %>% select(-c("year", "polpc"))

##### Missingness & Data Validity

We can see from the following table that we don’t have any missing data.

However, we appear to have a few values that are extremely high based on maximum values (p100). For example, one county has weekly wages in the service sector reported to be $2177 (almost 10x the median of $253).

Note that pctmin80 values are not in the [0,1] scale despite being a percentage of total metric. The scale for this variable does not match the other percentage metric in our data set (pctymle) and will be adjusted to facilitate coefficient interpretation.

#generate skim table  
data2\_skim <- skim(data2)  
data2\_skim

## Skim summary statistics  
## n obs: 90   
## n variables: 23   
##   
## ── Variable type:numeric ────────────────────────────────────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## avgsen 0 90 90 9.69 2.83 5.38 7.38 9.11   
## central 0 90 90 0.38 0.49 0 0 0   
## county 0 90 90 100.6 58.32 1 51.5 103   
## crmrte 0 90 90 0.034 0.019 0.0055 0.021 0.03   
## density 0 90 90 1.44 1.52 2e-05 0.55 0.98   
## mix 0 90 90 0.13 0.082 0.02 0.081 0.1   
## pctmin80 0 90 90 25.71 16.98 1.28 10.02 24.85   
## pctymle 0 90 90 0.084 0.023 0.062 0.074 0.078  
## prbarr 0 90 90 0.3 0.14 0.093 0.2 0.27   
## prbconv 0 90 90 0.55 0.35 0.068 0.34 0.45   
## prbpris 0 90 90 0.41 0.081 0.15 0.36 0.42   
## taxpc 0 90 90 38.16 13.11 25.69 30.73 34.92   
## urban 0 90 90 0.089 0.29 0 0 0   
## wcon 0 90 90 285.35 47.75 193.64 250.75 281.16   
## west 0 90 90 0.24 0.43 0 0 0   
## wfed 0 90 90 442.62 59.95 326.1 398.78 448.85   
## wfir 0 90 90 321.62 54 170.94 285.56 317.13   
## wloc 0 90 90 312.28 28.13 239.17 297.23 307.65   
## wmfg 0 90 90 336.03 88.23 157.41 288.6 321.05   
## wser 0 90 90 275.34 207.4 133.04 229.34 253.12   
## wsta 0 90 90 357.74 43.29 258.33 329.27 358.4   
## wtrd 0 90 90 210.92 33.87 154.21 190.71 202.99   
## wtuc 0 90 90 410.91 77.36 187.62 374.33 404.78   
## p75 p100 hist  
## 11.47 20.7 ▆▇▅▅▂▁▁▁  
## 1 1 ▇▁▁▁▁▁▁▅  
## 150.5 197 ▇▆▇▇▆▇▇▇  
## 0.04 0.099 ▆▇▇▃▂▁▁▁  
## 1.57 8.83 ▇▃▁▁▁▁▁▁  
## 0.15 0.47 ▃▇▂▁▁▁▁▁  
## 38.18 64.35 ▇▅▅▅▆▃▁▂  
## 0.084 0.25 ▇▁▁▁▁▁▁▁  
## 0.34 1.09 ▆▇▃▁▁▁▁▁  
## 0.59 2.12 ▃▇▂▁▁▁▁▁  
## 0.46 0.6 ▁▁▂▅▇▇▂▁  
## 41.01 119.76 ▇▃▁▁▁▁▁▁  
## 0 1 ▇▁▁▁▁▁▁▁  
## 314.98 436.77 ▂▇▇▇▅▃▁▁  
## 0 1 ▇▁▁▁▁▁▁▂  
## 478.26 597.95 ▃▅▅▇▆▃▂▁  
## 342.63 509.47 ▁▁▅▇▃▁▁▁  
## 328.78 388.09 ▁▁▃▇▅▂▁▁  
## 359.89 646.85 ▁▃▇▅▁▁▁▁  
## 277.65 2177.07 ▇▁▁▁▁▁▁▁  
## 383.15 499.59 ▂▃▆▇▆▂▁▁  
## 224.28 354.68 ▃▇▆▂▂▁▁▁  
## 440.68 613.23 ▁▁▂▆▇▂▂▁

# convert % minority variable  
data2$pctmin80 <- data2$pctmin80 \* 0.01

# save clean data file   
save(data2, file = "data\_clean.rda")

## Variable-level data checks & Exploratory Data Analysis

### Dependent Variable: Crime Rate

In Figure 1, we used a histogram to visualize the shape of crmrte’s distribution and can see that it is unimodal and approximately normal (with a slight skew). We used a boxplot to better visualize potential outliers. Most values are between 0.02 - 0.04 with some to 0.10. We don’t appear to have any spurious values.

ggplot(data2, aes(crmrte)) +  
 geom\_histogram(bins = 35, fill = "seashell4", color = "grey28") +  
 theme\_minimal() +  
 xlab("Crime Rate") +  
 ylab("Frequency") +  
 ggtitle("Histogram of Dependent Variable: Crime Rate")

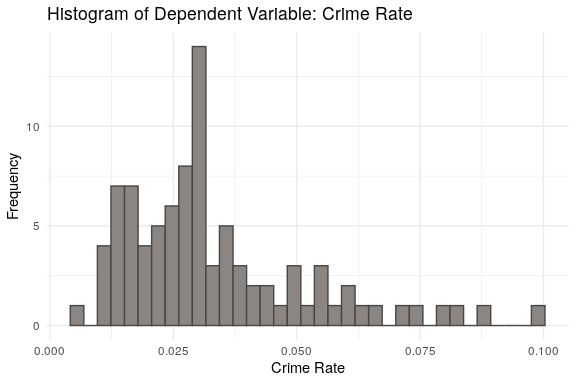


Figure 1

### Independent Variables

#### Actionable Variables

##### Criminal Justice Policy

From our violin plots in Figure 2, we can see that we have different distributions for the 3 core criminal justice policy variables (prbarr, prbconv, prbpris), but that they have a similar shape (unimodal with different levels of skew). The embedded boxplots reveal some extreme values for conviction rate, with multiple counties having a rate above 1. This values imply that for each arrest, there could be multiple convictions - one county has over 2 convictions for every arrest. Only 1 county has more than 1 arrest per crime committed - most counties in fact put someone under arrest for about one-quarter of crimes. When it comes to sentencing, counties show a much lower spread and cluster around 1 conviction, resulting in a prison sentence for every other conviction.

# format data for subplots   
data\_cj\_long <- data2 %>% select(prbarr, prbconv, prbpris) %>%  
 gather(key = var, value = value)  
  
# convert to factor for subplot labelling   
data\_cj\_long$var <- factor(data\_cj\_long$var, labels = c("Prob Arrest", "Prob Convicted", "Prob Sentenced"))

# plot violin/boxplot   
ggplot(data\_cj\_long, aes(x = var, y = value, fill = var)) +  
 geom\_violin(fill = "thistle3", color = "grey28") +   
 geom\_boxplot(width=0.1, fill = "white") +   
 theme\_minimal() +  
 xlab("") +  
 ylab("Ratio") +   
 ggtitle("Criminal Justice Policy Variables: Assertiveness Variables")

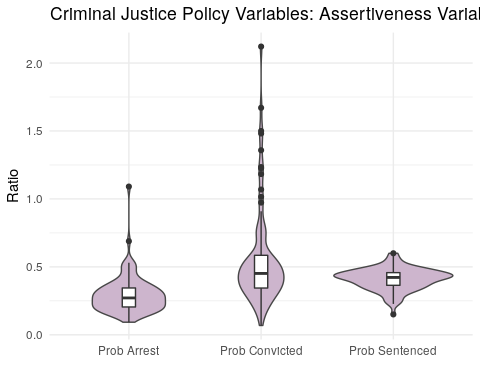


Figure 2

Which county had the lowest conviction rate? This is a county in County #11. It has a very high arrest rate given its conviction rate but we could not spot anything that would indicate this is an erroneous value.

data2 %>% filter(prbconv == min(prbconv)) %>% select(county, prbarr, prbconv, prbpris, avgsen, pctmin80)

## # A tibble: 1 x 6  
## county prbarr prbconv prbpris avgsen pctmin80  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 11 0.525 0.0684 0.5 13 0.0154

Which county had the highest conviction rate? This is a county in central North Carolina. It also has the highest percentage of minorities.

data2 %>% filter(prbconv == max(prbconv)) %>% select(county, prbarr, prbconv, prbpris, avgsen, pctmin80)

## # A tibble: 1 x 6  
## county prbarr prbconv prbpris avgsen pctmin80  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 185 0.195 2.12 0.443 5.38 0.643

The fact that the conviction rate is so high seems odd for county #185. We can see from Figure 3 that the other measures of assertiveness (for arrest and sentencing) seem close to their medians.

data\_prb <- data2 %>% as\_tibble() %>% select(prbarr, prbconv, prbpris) %>% summarise\_all(median) %>%   
 gather(var, value) %>% mutate(Group = "Median")  
  
data\_ep\_long <- data2 %>%   
 select(wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc) %>%  
 gather(key = var, value = value)  
  
prb\_185 <- data2 %>% as\_tibble() %>% filter(county == 185) %>% select(prbarr, prbconv, prbpris) %>%   
 gather(var, value) %>% mutate(Group = "County #185")  
  
prbs <- bind\_rows(data\_prb, prb\_185)  
  
ggplot(data = prbs, aes(x = var, y = value, color = Group)) +   
 geom\_point(size = 4, alpha = .7) +  
 xlab('Criminal Justice Assertiveness Category') +  
 ylab('Assertiveness') +  
 theme\_minimal() +  
 ggtitle("Criminal Justice Assertiveness: County 185 vs. Median")

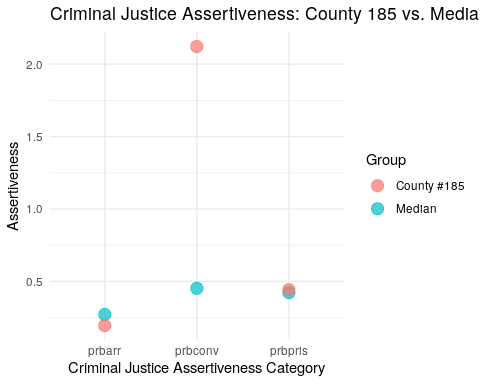


Figure 3

According to the FIPS codes, #185 stands for Warren County. [During the 1980s, its citizens conducted a series of demonstrations against a newly built landfill](https://en.wikipedia.org/wiki/Warren_County_PCB_Landfill), which led to multiple arrests and convictions associated with disobedience (not crime). This was an extraordinary event outside the process of interest, so we will remove this county.

We can see from Figure 4 that average sentence length follows an approximately normal distribution with one extremely high value, which corresponds to Madison County (according to the FIPS code). This western county is home to the state’s oldest jail, [still in operation in 1987](https://mountainx.com/news/murky-future-for-madisons-historic-jailhouse/). The fact that this county is in one of the less prosperous areas of the USA and home to this jail could explain its average sentence length. We retain this outlier in the data since it probably doesn’t indicate spurious data and would be an interesting edge case to study.

ggplot(data2, aes(avgsen)) +  
 geom\_histogram(binwidth = 1, fill = "thistle3", color = "grey28") +  
 theme\_minimal() +  
 xlab("Days of Imprisonment") +  
 ylab("Frequency") +  
 ggtitle("Criminal Justice Policy Variables: Assertiveness Variables")

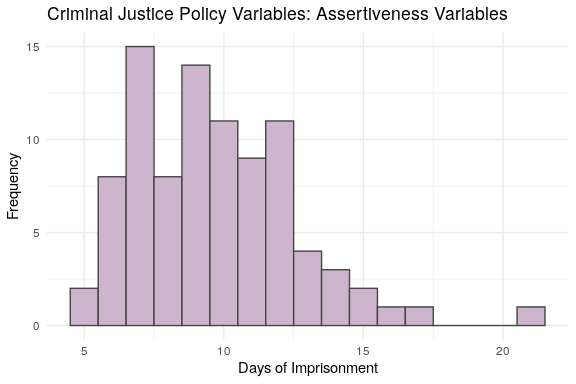


Figure 4

data2 %>% filter(avgsen == max(avgsen)) %>% select(county, avgsen, west, urban)

## # A tibble: 1 x 4  
## county avgsen west urban  
## <dbl> <dbl> <dbl> <dbl>  
## 1 115 20.7 1 0

##### Economic Policy

Minimum wage is one of the few economic factors that state/local politicians can change. In North Carolina in 1987, the minimum hourly wage was $3.35 ($134 per week, given a 40 hour week).

In Figure 5, we plot weekly wage by sector with a red vertical line indicating this minimum weekly wage.

The violin plots with embedded boxplots make it clear that Wholesale & Retail Trade wages are closest to the minimum wage - suggesting that wages in this sector are most under the influence of minimum wage policy. The next sector that would be most influenced is Services. Note that wages in the Services industry may be an inaccurate of income because many of its workers make additional and often unreported income through tips. Thus, wtrd would be more sensitive to a minimum wage policy.

Another aspect of wage our candidate could influence are state and local employee wages. Raising the wages of government employees could increase the quality of social services, reducing crime. Conversely, it could cut into budgets for other crime-cutting efforts and may increase crime. We will include these in our models in order to test this policy, but implementation may be challenging.

# format data for subplots   
data\_ep\_long <- data2 %>%   
 select(wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc) %>%  
 gather(key = var, value = value)  
# convert to factor for better labeling   
data\_ep\_long$var <- factor(data\_ep\_long$var,  
 levels = c("wcon", "wtuc", "wtrd", "wfir",  
 "wser", "wmfg", "wfed", "wsta", "wloc"),  
 labels = c("Construction",   
 "Utilities/Transport",   
 "Wholesale & Retail Trade",  
 "Finance/Insurance/Real Estate",  
 "Services",  
 "Manufacturing",  
 "Federal Emp",  
 "State Emp",  
 "Local/State Emp"))

# plot violin & boxplot subplots   
ggplot(data\_ep\_long, aes(x = var, y = value)) +  
 geom\_violin(fill = "lightseagreen") +   
 geom\_boxplot(width=0.1) +  
 coord\_flip() +   
 theme\_minimal() +  
 xlab("") +   
 ylab("Weekly Wage") +  
 ggtitle("Economic Variables: Wages")+  
 geom\_hline(aes(yintercept = 134, color = "red")) +  
 theme(legend.position = "none")

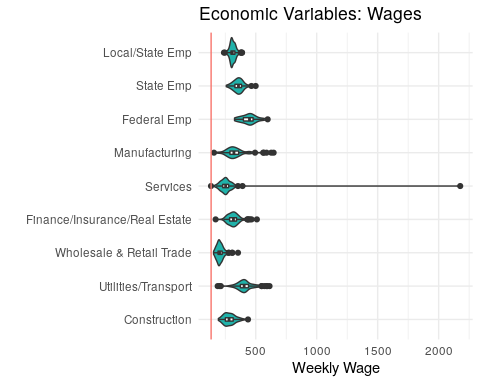


Figure 5

Another variable influenced by economic policy is tax revenue per capita. Figure 6 demonstrates that the distribution of taxpc is skewed. This distribution coupled with the fact that taxpc is measured in dollars makes the variable a top candidate for a log transformation.

# plot histogram & boxplot  
ggplot(data2, aes(taxpc)) +  
 geom\_histogram(binwidth = 2, fill = "lightseagreen", color = "grey28") +  
 theme\_minimal() +  
 xlab("Tax Revenue Per Capita") +  
 ylab("Frequency") +  
 ggtitle("Economic Variables: Tax Revenue Per Capita")

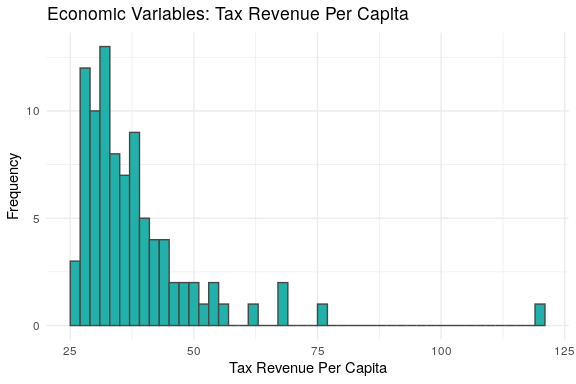


Figure 6

We appear to have an extreme value at $119.76, which is for county #55. This is likely Dare County according to its FIPS code, a small island that is a popular vacation spot.

One would expect that counties with high tax revenue also have high wages of their populace. We checked the wages of this county against the median and noticed that many were *below or close to* the median. This pattern (shown in Figure 7) could be explained by the fact that in small touristy counties, an important share of taxes results from tourism activities of non-residents, which boosts the region’s tax revenue per local citizen. Given that this row probably does not have a data quality issue, we retained it.

# reformat data for plotting  
data\_tax <- data2 %>% as\_tibble() %>% select(taxpc, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc) %>% summarise\_all(median) %>%   
 gather(var, value) %>% mutate(Group = "Median")  
taxes\_55 <- data2 %>% as\_tibble() %>% filter(county == 55) %>% select(taxpc, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc) %>%   
 gather(var, value)%>% mutate(Group = "County #55")  
taxes <- bind\_rows(data\_tax, taxes\_55)  
  
# plot  
ggplot(data = taxes, aes(x = var, y = value, color = Group)) +   
 geom\_point(size = 4, alpha = .7) +  
 xlab('Category') +  
 ylab('$') +  
 theme\_minimal() +  
 ggtitle("Tax Revenue and Wages: County 55 vs. Median")

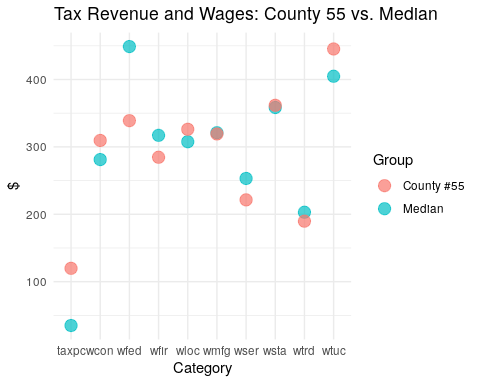


Figure 7

#### Contextual Variables

Figure 8 reveals that the distributions for mix, density and pctymle approximate a normal distribution but the that of pctmin80 approximates a uniform distribution. Minorities seem to be present in most counties in similar proportions.

# hist  
p5 <- ggplot(data2, aes(mix)) +  
 geom\_histogram(bins = 35, color = "grey28", fill = wes\_palette("BottleRocket2", n = 1)) +  
 theme\_minimal() +  
 ggtitle("Mix of Offense Types") +  
 xlab("offense mix: face-to-face/other")  
  
p7 <- ggplot(data2, aes(density)) +  
 geom\_histogram(bins = 35, color = "grey28", fill = wes\_palette("BottleRocket2", n = 1)) +  
 theme\_minimal() +  
 xlab("Density") +  
 ylab("Count") +  
 ggtitle("Population Density")  
  
p9 <- ggplot(data2, aes(pctmin80)) +  
 geom\_histogram(bins = 25, color = "grey28", fill = wes\_palette("BottleRocket2", n = 1)) +  
 theme\_minimal() +  
 xlab("% Minority") +  
 ylab("Count") +  
 ggtitle("Percentage Minority (1980)")  
  
  
p11 <- ggplot(data2, aes(pctymle)) +  
 geom\_histogram(bins = 25, color = "grey28", fill = wes\_palette("BottleRocket2", n = 1)) +  
 theme\_minimal() +  
 xlab("% Young Males") +  
 ylab("Count") +  
 ggtitle("Percentage Young Male (1980)")  
  
  
grid.arrange(p5, p9, p7, p11, ncol = 2, nrow = 2, top = "Contextual Variables")

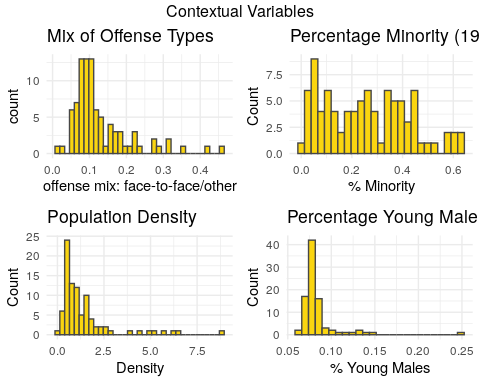


Figure 8

County #119 has the highest density. This county probably is Mecklenburg County, home to city of Charlotte (the state capital and one of the most populated cities).

data2 %>% filter(density == max(density)) %>% select(county, density, central, urban)

## # A tibble: 1 x 4  
## county density central urban  
## <dbl> <dbl> <dbl> <dbl>  
## 1 119 8.83 1 1

The smallest density county would correspond to Swain County, a very rural county in the Western part of the state. It straddles two national parks/forests, which would explain why it has such low density.

data2 %>% filter(density == min(density)) %>% select(county, density, west, urban)

## # A tibble: 1 x 4  
## county density west urban  
## <dbl> <dbl> <dbl> <dbl>  
## 1 173 0.0000203 1 0

We retain these rows, which don’t have data quality issues and may be edge cases of interest.

County #133 (Onslow County based on FIPS code) is the main outlier for pctymle. The deviation in percent of young males could be due to Camp Lejeune, a marine corps base. Most recruits are young males and so this base would impact the the county’s share of young males, especially if its population is small. As recruits would most likely not commit crimes, this can interfere in our expected positive relationship between the percent of young males and crime. In fact, this is not the only county in North Carolina with a military base, so we will keep this observation and monitor the estimated coefficient of pctymle in our crime model.

data2 %>% filter(pctymle == max(pctymle)) %>% select(county, pctymle)

## # A tibble: 1 x 2  
## county pctymle  
## <dbl> <dbl>  
## 1 133 0.249

##### Geography

We were given 3 variables to track geography. These are dummy variables that code for west versus central versus east and urban versus rural.

The majority of North Carolina’s population is located in the Central and Eastern parts of the state. Approximately one-quarter of the population is in Western counties. Almost 90% of North Carolina’s inhabitants live in a rural area.

# create new east variable  
geo\_supp <- data2 %>% mutate(east = ifelse(west == 0 & central == 0, 1, 0), rural = ifelse(urban == 0, 1, 0))

geo\_pop\_tbl <- bind\_cols("Location" = c("East", "Central", "West"),  
 "% of Population" = c(round(100\*sum(geo\_supp$east)/90, 2),  
 round(100\*sum(geo\_supp$central)/90, 2),  
 round(100\*sum(geo\_supp$west)/90, 2)))  
  
type\_pop\_tbl <- bind\_cols("Type" = c("Urban", "Rural"),  
 "% of Population" = c(round(100\*sum(geo\_supp$urban)/90, 2),  
 round(100\*sum(geo\_supp$rural)/90, 2)))  
knitr::kable(geo\_pop\_tbl)

|  |  |
| --- | --- |
| Location | % of Population |
| East | 38.89 |
| Central | 37.78 |
| West | 24.44 |

knitr::kable(type\_pop\_tbl)

|  |  |
| --- | --- |
| Type | % of Population |
| Urban | 8.89 |
| Rural | 91.11 |

Figure 9 demonstrates that the vast number of counties are rural across central, western, and eastern counties. Central counties have the largest share of urban geographies.

# reformat data for plotting  
geo\_long <- geo\_supp %>%  
 mutate\_at(vars(west, central, east, rural, urban), ~ ifelse(. == 0, NA, .)) %>%  
 gather("location", "present1", west, central, east, na.rm = TRUE) %>%  
 gather("type", "present2", rural, urban, na.rm = TRUE) %>%  
 select(-present1, -present2)

# plot barplots  
ggplot(geo\_long, aes(x = location, fill = type)) +  
 geom\_bar(position = "fill") +  
 scale\_fill\_manual(values = wes\_palette("GrandBudapest1", n = 3)) +   
 theme\_minimal() +  
 ggtitle("Contextual Variables: Geography") +  
 xlab("")

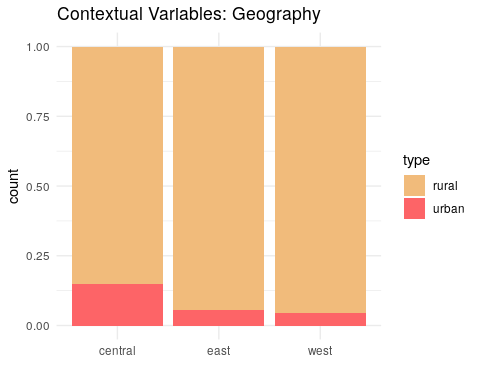


Figure 9

### Relationships between Variables

# remove row with erroneous values (#185) before investigating relationships   
data2 <- data2 %>% filter(county != 185)  
print(paste("Number of records:", dim(data2)[1]))

## [1] "Number of records: 89"

Figure 10 is a correlation matrix of all of our variables, using Spearman’s due to the fact that not all variables follow a normal distribution.

As expected, prbarr and prbconvshow a negative correlation with crmrte. However, prbpris and avgsen do not show a strong correlation with cmrte. The fact that in general the number of imprisonments is low likely weakens the relationship between prbpris and cmrte. Also, given that most crimes committed are petty crimes with relatively low sentence lengths, average sentence length might not be a strong deterrent of crimes.

Additionally, mix does not seem to have a strong correlation with cmrte. mix values are very low, indicating that face-to-face crimes are almost non-existent across counties. Thus, the low variability of mix does not allow for a material relationship between it and crime rate.

Wages are mostly positively correlated with crime, opposite to our prediction. This unexpected direction could be explained in part by the strong positive relationship between the wage variables and density and between crmrte and density. The bigger the city, usually the higher the wages but also the higher the crime rate. This unexpected relationship could also be due to our points above about inequality - if inequality is positively correlated with both to crime and average wages, it could be a hidden factor that is impacting this relationship. This pattern bolsters our decision to *not* include the wage variables in early models; when we do include it, it’s crucial to include additional controls.

As expected, population density has a strong positive relationship with crime rate. It is also positively correlated with other geographical variables. For example, urban and density are positively correlated because cities will have higher density. We will pay special attention to these strongly correlated variables as they can increase the variance of models’ residuals if they are cast as independent variables in the same equation.

All wages as well as taxpc are positively correlated to each other. This behavior is expected as wages and taxable income are both determined by analogous economic factors. Similarly as above, we will need to pay close attention as the joint impact of including multiple variables that are highly correlated in models.

# omit county to create correlation matrix   
data\_corr <- data2 %>% select(-county)   
# create matrix of correlations  
matrix\_corr <- round(cor(data\_corr, method = "spearman"), 1)

# visualize correlation matrix   
corrplot(matrix\_corr, type = "lower", method = "ellipse", order = "alphabet")

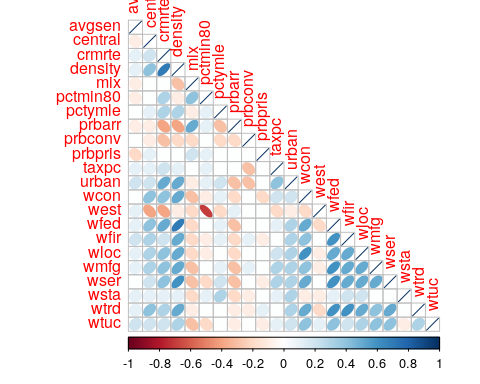


Figure 10

## Model Results

In order select the optimal model in a certain range of model specification options, we rely on information criteria. Two popular IC are AIC (Akaike information criterion) and and BIC (Bayesian information criterion). AIC and BIC for a model is usually written in the form , where is the likelihood function, is the number of parameters in the model, and is for AIC and for BIC (=number of observations). Despite subtle theoretical differences, their only difference in practice is the size of the penalty: BIC penalizes model complexity more heavily. To maintain parsimony of our models, we opted to use BIC for model selection.

### Model set #1

For our first model, we aim to evaluate the relationship between crime policy and crime. We expect that more assertive crime policies will be associated with lower crime rates.

We used the following independent variables:

1. log(prbarr)
2. log(prbconv)
3. log(prbpris)
4. log(avg\_sen)

#### Data Transformations

We log transformed crime policy measures and crmrte to improve our model fit and interpretability. With the log-log specification, the coefficients of the criminal justice system variables can be interpreted as the expected relative change in crime rate given a relative change in assertiveness of the criminal justice system.

#### Discussion

In the previous section, we found that prbarr and prbconvhave a strong negative correlation with crmrte, but prbpris and avgsen don’t. With that in mind, in the table below we compare the output of a model featuring only arrest and conviction rates with the output of a model featuring all the crime policy variables.

# create models  
mod1\_1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgsen), data = data2)  
mod1\_2 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) , data = data2)  
mod1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) , data = data2)  
  
# generate robust standard errors  
se\_mod1\_1 <- sqrt(diag(vcovHC(mod1\_1)))  
se\_mod1\_2 <- sqrt(diag(vcovHC(mod1\_2)))  
se\_mod1 <- sqrt(diag(vcovHC(mod1)))  
  
# produce regression table  
stargazer(  
 mod1\_1  
 , mod1\_2  
 , mod1  
 , type = "text"  
 , se = list(se\_mod1\_1, se\_mod1\_2, se\_mod1)  
 , add.lines=list(c("BIC", round(BIC(mod1\_1),1), round(BIC(mod1\_2),1), round(BIC(mod1),1)))  
 , notes = "Robust SE"  
 , star.cutoffs = c(0.05, 0.01, 0.001)  
)

##   
## ========================================================================================  
## Dependent variable:   
## --------------------------------------------------------------------  
## log(crmrte)   
## (1) (2) (3)   
## ----------------------------------------------------------------------------------------  
## log(prbarr) -0.728\*\*\* -0.728\*\*\* -0.730\*\*\*   
## (0.121) (0.118) (0.115)   
##   
## log(prbconv) -0.443\*\* -0.441\*\* -0.442\*\*   
## (0.151) (0.144) (0.140)   
##   
## log(prbpris) 0.164 0.160   
## (0.241) (0.239)   
##   
## log(avgsen) 0.033   
## (0.195)   
##   
## Constant -4.743\*\*\* -4.673\*\*\* -4.822\*\*\*   
## (0.525) (0.289) (0.176)   
##   
## ----------------------------------------------------------------------------------------  
## BIC 123.2 118.8 114.9   
## Observations 89 89 89   
## R2 0.405 0.404 0.400   
## Adjusted R2 0.376 0.383 0.386   
## Residual Std. Error 0.428 (df = 84) 0.425 (df = 85) 0.424 (df = 86)   
## F Statistic 14.271\*\*\* (df = 4; 84) 19.234\*\*\* (df = 3; 85) 28.671\*\*\* (df = 2; 86)  
## ========================================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
## Robust SE

The table above shows that approximately 38-40% of the variation in log(crime rate) can be explained only by the criminal justice policy variables. Coefficients on log(prbpris) and log(avg\_sen) were not statistically significant in column (1). Also note that excluding criminal justice indicators that are weakly related to crime - log(avgsen) and log(prbpris) - did not affect R2 nor changed the estimated coefficients on the criminal justice indicators that are statistically significant, log(prbarr) and log(prbconv). Finally, BIC is smaller for the most parsimonious model (column 3).

At this point, we believe we have enough information to conclude that log(prbpris) and log(avg\_sen) do not have explanatory power over log(crmrte) that goes beyond that of log(prbarr) and log(prbconv). Thus, our preferred specification at this stage is that of column 3.

##### Conclusion

Politicians could leverage criminal justice policies to curb crime rate. In fact, this first model set featuring only criminal justice policies as independent variables suggest that, ceteris paribus, a 10% increase in prbarr would lead to a ~7% decrease in crime rate, while a 10% increase in prbconv would lead to a ~5% decrease in crime rate. These figures also have an important practical meaning as a 5-7% change in crime rate may have material impact on people’s lives.

### Model set #2

For our second model, we aimed to evaluate the relationship between crime and a county’s economic policies and context. As mentioned in the EDA section, an important economic variable that state and local politicians control is minimum wage. As such, we include wages of industries whose average wage is close to the minimum wage (wtrd followed by wser). We include wloc and wsta as a test of policies related to local and state government employees. Finally, wfed and taxpc will be included as important economic context covariates.

We are aware that there are several determinants of wages that go beyond political interference, so wages are only proxies of economic policies. As such, conclusions from these exercises should be taken as hints for future research rather than strong evidence.

We used the following independent variables to predict crime:

*Crime policy*

1. log(prbarr)
2. log(prbconv)

*Economic policy*

1. log(wtrd)
2. log(wser)
3. log(wloc)
4. log(wsta)

*Economic context*

1. log(wfed)
2. log(taxpc)

#### Data Transformations

We decided to take the log of our wage variables to show the relationship between relative changes in these to crime. This improves our interpretability since we are making policy recommendations around changing these relative to the existing levels in these counties.

#### Economic policy

From Figure 11, we can see that we have a fairly linear relationship between the log of Trade Wages and log(crmrte). So is the relationship between the log of Service Wages and log(crmrte), although there are some outliers that seems to have strong influence on the relationship. The relationship between the log of Federal Wages and log(crmrte) also seems fairly linear.

Tax per capita is log-transformed because it is measured in dollars, and thus an elasticity-like interpretation of the coefficient of this variable would be simpler. Even after log-transforming the variable, many of its points cluster together, leaving little variability with which to predict crime.

wtrdwg<- ggplot(data = data2, aes(x = log(wtrd), y = log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Wholesale & Retail Trade Wages") +   
 xlab("Log Average Weekly Wage") + ylab("Log Crime Rate") +  
 geom\_smooth(method='lm', formula= y~x, color = "lightseagreen", fill = "grey")  
  
serwg <- ggplot(data = data2, aes(x = log(wser), log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Service Wages") +   
 xlab("Log Average Weekly Wage") + ylab("Log Crime Rate") +  
 geom\_smooth(method='lm', formula= y~x, color = "lightseagreen", fill = "grey")  
  
fedwg <- ggplot(data = data2, aes(x = log(wfed), log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Federal Employee Wages") +   
 xlab("Log Average Weekly Wage") + ylab("Log Crime Rate") +  
 geom\_smooth(method='lm', formula= y~x, color = "lightseagreen", fill = "grey")  
  
taxpc <- ggplot(data = data2, aes(x = log(taxpc), log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Tax Per Capita") +   
 xlab("Log Tax Per Capita") + ylab("Log Crime Rate") +  
 geom\_smooth(method='lm', formula= y~x, color = "lightseagreen", fill = "grey")  
  
grid.arrange(wtrdwg, serwg, fedwg, taxpc, ncol = 2, nrow = 2, top = "Relationship: Crime Rate and Economics")



Figure 11

#### Discussion

The first iteration of this model includes all crime justice variable as well as all key economic variables. In this specification, none of the economic policy variables had statistically significant relationships to crime. The coefficient estimates for economic context variables (taxpc and wfed) were statistically significant and positive. This indicates that crime rate is larger in counties with more tax revenue and higher cost of living, controlling for the other variables in the model. At this point, these variables could be capturing part of the crime rate variation that should be captured by demographic variables such as density. We expect the estimated coefficients for these variables to fluctuate as we add more demographic variables.

After adding economic policy and context to the model, the estimated impact of log(prbarr) on log(crmrte) reduces only slightly versus model set #1 (from ~0.7 to ~0.6). We conducted an F-test to determine if the economic policy variables were jointly statistically significant at explaining variation on crime rate. In this test, the null hypothesis was that economic policies, proxied by average wages at lower paying industries and at local administrations, do not have explanatory power over crime rate. The p-value of this test was well over our alpha of .05 and thus we could not reject the hypothesis.

The exclusion of economic policy variables from our models is advisable also according to BIC, which is smaller for the most parsimonious model shown in column 2, our preferred specification in this set. Notice though that we decided to keep one proxy for wage policy, Wholesale & Retail Trade sector wages because these are the closest to the minimum wage and most likely to be affected by minimum wage policy. Keeping this variable will give it a chance to be tested with other covariates, such as density and percentage of minority, in our later model specifications.

mod2\_1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wser) + log(wloc) + log(wsta) + log(wfed) + log(taxpc), data = data2)  
# fit models  
#mod2\_2 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wfed) + log(taxpc), data = data2) # no economic policy variables  
#mod2\_3 <- lm(log(crmrte) ~ log(prbarr)\*log(wfed) + log(prbconv)\*log(wfed) + log(wser)\*log(wfed) + log(wsta)\*log(wfed) + log(taxpc), data = data2) # taxpc interactions  
#mod2\_4 <- lm(log(crmrte) ~ log(prbarr)\*log(taxpc) + log(prbconv)\*log(taxpc) + log(wser)\*log(taxpc) + log(wsta)\*log(taxpc) + wfed, data = data2) # wfed interactions  
mod2 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc), data = data2) # some economic policy variables, no interactions  
  
# generate robust standard errors  
se\_mod2\_1 <- sqrt(diag(vcovHC(mod2\_1)))  
se\_mod2 <- sqrt(diag(vcovHC(mod2)))  
  
# produce regression table  
stargazer(  
 mod2\_1  
 , mod2  
 , type = "text"  
 , se = list(se\_mod2\_1, se\_mod2)  
 , add.lines=list(c("BIC", round(BIC(mod2\_1),1), round(BIC(mod2))))  
 , notes = "Robust SE"  
 , star.cutoffs = c(0.05, 0.01, 0.001)  
 )

##   
## =================================================================  
## Dependent variable:   
## ---------------------------------------------  
## log(crmrte)   
## (1) (2)   
## -----------------------------------------------------------------  
## log(prbarr) -0.615\*\*\* -0.567\*\*\*   
## (0.130) (0.119)   
##   
## log(prbconv) -0.427\*\*\* -0.402\*\*\*   
## (0.123) (0.121)   
##   
## log(wtrd) 0.007 0.074   
## (0.420) (0.417)   
##   
## log(wser) -0.512   
## (0.315)   
##   
## log(wloc) 0.640   
## (0.699)   
##   
## log(wsta) -0.266   
## (0.313)   
##   
## log(wfed) 1.769\*\* 1.559\*\*   
## (0.573) (0.539)   
##   
## log(taxpc) 0.404 0.374   
## (0.207) (0.202)   
##   
## Constant -16.205\*\*\* -15.810\*\*\*   
## (3.414) (2.360)   
##   
## -----------------------------------------------------------------  
## BIC 102.7 93   
## Observations 89 89   
## R2 0.614 0.597   
## Adjusted R2 0.575 0.573   
## Residual Std. Error 0.353 (df = 80) 0.354 (df = 83)   
## F Statistic 15.875\*\*\* (df = 8; 80) 24.614\*\*\* (df = 5; 83)  
## =================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
## Robust SE

# F-test: joint significance of economic policy variables  
print("F-test results for joint significance of economic policies")

## [1] "F-test results for joint significance of economic policies"

linearHypothesis(mod2\_1, c("log(wser)=0", "log(wsta)=0", "log(wtrd)=0", "log(wloc)=0"), vcov=vcovHC)

## Linear hypothesis test  
##   
## Hypothesis:  
## log(wser) = 0  
## log(wsta) = 0  
## log(wtrd) = 0  
## log(wloc) = 0  
##   
## Model 1: restricted model  
## Model 2: log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wser) +   
## log(wloc) + log(wsta) + log(wfed) + log(taxpc)  
##   
## Note: Coefficient covariance matrix supplied.  
##   
## Res.Df Df F Pr(>F)  
## 1 84   
## 2 80 4 0.7892 0.5356

##### Conclusion

We fail to reject hypothesis that economic policy variables have no impact on crime rate. It seems that politicians might not be able to use economic policy variables to impact crime rate, at least as captured by the available dataset.

The results from this model set though reinforces our previous section’s findings: Politicians could leverage criminal justice policies to curb crime rate. In fact, this second model suggests that, ceteris paribus, a 10% increase in prbarr would lead to a ~6% decrease in crime rate, while a 10% increase in prbconv would lead to a ~4% decrease in crime rate. These estimated effects are statistically significant and have important practical meaning.

### Model set #3

For our third model, we added demographic control variables to evaluate the robustness of the relationship between crime policy, economic policy, and crime rate when controlling for key demographic contexts in addition to the economic context variables. We retained the one wage variable we thought was the best proxy for minimum wage (wtrd) so that we can confirm that the lack of relationship was not due to omitted demographic variables biasing our estimates. We limit wage variables to one to avoid overfitting.

We used the following independent variables to predict crime:

*Crime policy*

1. log(prbarr)
2. log(prbconv)

*Economic policy*

1. log(wtrd)

*Economic context*

1. log(wfed)
2. log(taxpc)

*Demographic*

1. density
2. pctymle
3. pctmin80

#### Data Transformations

Many counties had similar demographics (density, mix, pctymle). This may be due to the fact that North Carolina is a small state that may have limited demographic variation. Logarithmic transformation is a convenient mean of transforming a highly skewed variable into a more normalized dataset. Still, plotting the log of demographic versus the log of crime rate didn’t seem to improve our fitting and we decided not to log-transform these variables.

Taking the log of density or pctymle does not improve either of their fits to crime rate. The percent minority in 1980 does not have a strong relationship with crime rate. Counties appear to have a wide range of this variable and there is only a slight positive correlation with crime rate. We noticed a slight concave form on the scatterplot so we will try a quadratic transformation of the variable in the model specifications below. These patterns are shown in Figure 12.

den1 <- ggplot(data = data2, aes(x = density, log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Density") +   
 ylab("Log Crime Rate") + xlab("People per Sq. Mile") +  
 geom\_smooth(method='lm', formula= y~x, color = wes\_palette("BottleRocket2", n = 1), fill = "grey")  
  
den2 <- ggplot(data = data2, aes(x = log(density), log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Log Density") +   
 ylab("Log Crime Rate") + xlab("Log People per Sq. Mile") +  
 geom\_smooth(method='lm', formula= y~x, color = wes\_palette("BottleRocket2", n = 1), fill = "grey")  
  
ymle1 <- ggplot(data = data2, aes(x = pctymle, log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Percent Young Male") +   
 ylab("Log Crime Rate") + xlab("% Young Male") +  
 geom\_smooth(method='lm', formula= y~x, color = wes\_palette("BottleRocket2", n = 1), fill = "grey")  
  
ymle2 <- ggplot(data = data2, aes(x = log(pctymle), log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Log Percent Young Male") +   
 ylab("Log Crime Rate") + xlab("Log % Young Male") +  
 geom\_smooth(method='lm', formula= y~x, color = wes\_palette("BottleRocket2", n = 1), fill = "grey")  
  
min1 <- ggplot(data = data2, aes(x = pctmin80, log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Percent Minority") +   
 ylab("Log Crime Rate") + xlab("% Minority") +  
 geom\_smooth(method='lm', formula= y~x, color = wes\_palette("BottleRocket2", n = 1), fill = "grey")  
  
min2 <- ggplot(data = data2, aes(x = pctmin80\*pctmin80, log(crmrte))) +  
 geom\_point() +  
 theme\_minimal() +  
 ggtitle("Percent Minority^2") +   
 ylab("Log Crime Rate") + xlab("% Minority^2") +  
 geom\_smooth(method='lm', formula= y ~x + I(x^2), color = wes\_palette("BottleRocket2", n = 1), fill = "grey")  
  
grid.arrange(den1, den2, ymle1, ymle2, min1, min2, ncol = 2, nrow = 3, top = "Relationship: Crime Rate and Demographics")

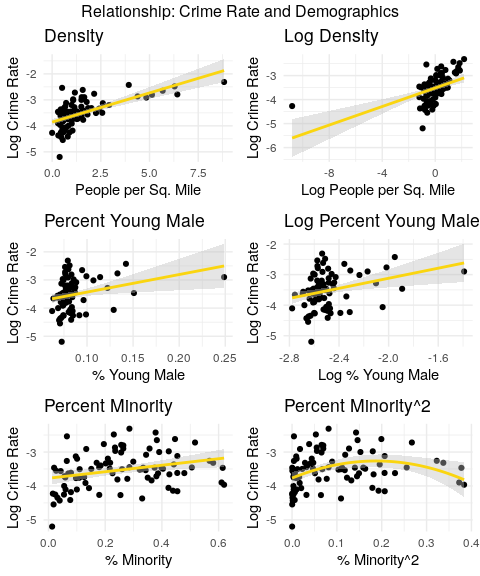


Figure 12

#### Discussion

The first iteration (column 1) of this model set includes all the variables listed above. Including all factors related to the demographics improves our model (lower BIC and larger adjusted R2 vs. model set #2).

Among all demographic variables included, only pctymle is not individually significant. An F-test was conducted with results leading us to reject the null that the demographic variables are not jointly significant.

In the following iteration (column 2), we added a quadratic transformation of percent minority, in an attempt to capture the seemingly concave relationship of this variable with log(crmrte) discussed above. However, the coefficient on the quadratic transformation of percent minority was not statistically significant and the model’s BIC is slightly larger than in column 1, leading us to give up on column 2 specification. In column 3, we dropped pctymle, leading to a larger BIC vs. column 1. This fact and our understanding that pctymle is an important control in crime rate models made us decide to keep the variable.

Given the above, our preferred equation in the model set #3 is column 1. The above also led us to conclude that that demographic factors should be part of of our model. Higher rates of any of these are related to higher levels of crime.

Our economic policy proxy does not seem to be statistically significant in explaining variation in crime rate when modeled with or without the addition of demographic control variables. As we noted above, this could be due to omitted variable bias and/or the fact that (even lower-income) average wages are imperfect proxies for minimum wage.

data2$pctmin802 = data2$pctmin80\*data2$pctmin80  
  
# fit models  
mod3\_1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = data2) # all vars  
mod3\_2 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctmin802 + pctymle, data = data2) # pctmin80 quadratic transform  
mod3\_3 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 , data = data2) # no pctymle  
#mod3\_4 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = data2) # no wages  
  
# generate robust standard errors  
se\_mod3\_1 <- sqrt(diag(vcovHC(mod3\_1)))  
se\_mod3\_2 <- sqrt(diag(vcovHC(mod3\_2)))  
se\_mod3\_3 <- sqrt(diag(vcovHC(mod3\_3)))  
#se\_mod3\_4 <- sqrt(diag(vcovHC(mod3\_4)))  
  
# produce regression table  
stargazer(  
 mod3\_1  
 , mod3\_2  
 , mod3\_3  
 #, mod3\_4  
 , type = "text"  
 , se = list(se\_mod3\_1, se\_mod3\_2, se\_mod3\_3)  
 , star.cutoffs = c(0.05, 0.01, 0.001)  
 , notes = c("Robust SE")  
 , add.lines=list(c("BIC", round(BIC(mod3\_1),1), round(BIC(mod3\_2),1) , round(BIC(mod3\_3),1)   
 #, round(BIC(mod3\_4),1)  
 )  
 )  
 )

##   
## ========================================================================================  
## Dependent variable:   
## --------------------------------------------------------------------  
## log(crmrte)   
## (1) (2) (3)   
## ----------------------------------------------------------------------------------------  
## log(prbarr) -0.504\*\*\* -0.478\*\*\* -0.559\*\*\*   
## (0.110) (0.126) (0.089)   
##   
## log(prbconv) -0.329\*\* -0.326\*\* -0.363\*\*\*   
## (0.100) (0.101) (0.094)   
##   
## log(wtrd) 0.075 0.032 -0.008   
## (0.373) (0.376) (0.341)   
##   
## log(wfed) 1.072\* 1.096\* 1.002\*   
## (0.427) (0.431) (0.444)   
##   
## log(taxpc) 0.289 0.269 0.230   
## (0.317) (0.362) (0.322)   
##   
## density 0.079 0.077 0.089\*   
## (0.043) (0.043) (0.045)   
##   
## pctmin80 1.127\*\*\* 1.841 1.153\*\*\*   
## (0.214) (0.978) (0.210)   
##   
## pctmin802 -1.295   
## (1.551)   
##   
## pctymle 3.087 3.008   
## (3.491) (3.279)   
##   
## Constant -13.064\*\*\* -12.921\*\*\* -11.835\*\*\*   
## (2.256) (2.283) (2.650)   
##   
## ----------------------------------------------------------------------------------------  
## BIC 62.6 65.7 63.1   
## Observations 89 89 89   
## R2 0.754 0.758 0.740   
## Adjusted R2 0.729 0.730 0.717   
## Residual Std. Error 0.282 (df = 80) 0.282 (df = 79) 0.288 (df = 81)   
## F Statistic 30.642\*\*\* (df = 8; 80) 27.426\*\*\* (df = 9; 79) 32.863\*\*\* (df = 7; 81)  
## ========================================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
## Robust SE

mod3 <- mod3\_1  
se\_mod3 <- se\_mod3\_1

# F-test: joint significance of demographic variables  
print("F-test results for joint significance of demographic variables")

## [1] "F-test results for joint significance of demographic variables"

linearHypothesis(mod3\_1, c("density=0", "pctmin80=0", "pctymle=0"), vcov=vcovHC)

## Linear hypothesis test  
##   
## Hypothesis:  
## density = 0  
## pctmin80 = 0  
## pctymle = 0  
##   
## Model 1: restricted model  
## Model 2: log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) +   
## log(taxpc) + density + pctmin80 + pctymle  
##   
## Note: Coefficient covariance matrix supplied.  
##   
## Res.Df Df F Pr(>F)   
## 1 83   
## 2 80 3 9.6385 1.673e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### Outliers

Figure 13 shows that there are 2 observations labeled as outliers and high leverage points (which is to be expected since they were at the extremes of some of our predictor variables). These labels correspond to the row numbers but we list their county numbers below for easier comprehension. We will not evaluate their influence as we had already decided to retain them during our EDA.

* observation 6 - county 11
* observation 25 - county 55

ols\_plot\_resid\_lev(mod3)

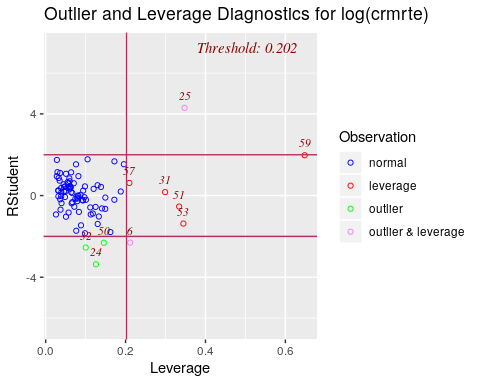


Figure 13

There are 3 additional counties that are outliers but not high leverage points and were not spotted in our EDA:

* observation 24 - county 53
* observation 32 - county 69
* observation 50 - county 113

Do these three counties influence our model fits? We can see that without these observations, we now explain a bit more of the variation. Although eliminating these have improved our fit, we do not have any reason to suspect that they have data quality issues. Dropping any of these do not change our conclusions around our main predictor variables.

# county 53  
mod3\_out53 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = filter(data2, county != 53))   
  
se\_mod3\_out53 <- sqrt(diag(vcovHC(mod3\_out53)))  
  
# county 69  
mod3\_out69 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = filter(data2, county != 69))   
  
se\_mod3\_out69 <- sqrt(diag(vcovHC(mod3\_out69)))  
  
# county 113  
mod3\_out113 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = filter(data2, county != 113))   
  
se\_mod3\_out113 <- sqrt(diag(vcovHC(mod3\_out113)))  
  
stargazer(  
 mod3  
 , mod3\_out53  
 , mod3\_out69  
 , mod3\_out113  
 , type = "text"  
 , se = list(se\_mod3, se\_mod3\_out53, se\_mod3\_out69, se\_mod3\_out113)  
 , star.cutoffs = c(0.05, 0.01, 0.001)  
 , column.labels = c("With", "Omit 53", "Omit 69", "Omit 113")   
 , notes = c("Robust SE")  
)

##   
## ===============================================================================================================  
## Dependent variable:   
## -------------------------------------------------------------------------------------------  
## log(crmrte)   
## With Omit 53 Omit 69 Omit 113   
## (1) (2) (3) (4)   
## ---------------------------------------------------------------------------------------------------------------  
## log(prbarr) -0.504\*\*\* -0.529\*\*\* -0.505\*\*\* -0.541\*\*\*   
## (0.110) (0.103) (0.110) (0.105)   
##   
## log(prbconv) -0.329\*\* -0.377\*\*\* -0.318\*\* -0.356\*\*\*   
## (0.100) (0.102) (0.099) (0.107)   
##   
## log(wtrd) 0.075 0.215 0.222 -0.020   
## (0.373) (0.349) (0.351) (0.357)   
##   
## log(wfed) 1.072\* 0.964\* 1.121\*\* 1.013\*   
## (0.427) (0.403) (0.421) (0.423)   
##   
## log(taxpc) 0.289 0.332 0.302 0.294   
## (0.317) (0.286) (0.319) (0.301)   
##   
## density 0.079 0.064 0.066 0.078   
## (0.043) (0.041) (0.042) (0.041)   
##   
## pctmin80 1.127\*\*\* 1.111\*\*\* 1.174\*\*\* 1.075\*\*\*   
## (0.214) (0.215) (0.211) (0.217)   
##   
## pctymle 3.087 2.995 3.351 2.815   
## (3.491) (3.287) (3.432) (3.423)   
##   
## Constant -13.064\*\*\* -13.332\*\*\* -14.196\*\*\* -12.239\*\*\*   
## (2.256) (2.264) (2.060) (2.088)   
##   
## ---------------------------------------------------------------------------------------------------------------  
## Observations 89 88 88 88   
## R2 0.754 0.780 0.770 0.765   
## Adjusted R2 0.729 0.758 0.747 0.741   
## Residual Std. Error 0.282 (df = 80) 0.265 (df = 79) 0.273 (df = 79) 0.275 (df = 79)   
## F Statistic 30.642\*\*\* (df = 8; 80) 35.044\*\*\* (df = 8; 79) 33.081\*\*\* (df = 8; 79) 32.094\*\*\* (df = 8; 79)  
## ===============================================================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
## Robust SE

##### Conclusion

The most important takeaways from model set #3 are (1) the estimated effect of criminal justice policies on crime rate is robust to the inclusion of demographic controls, and (2) even controlling for demographics, the coefficient on our proxy for economic policy (log(wtrd)) remains statistically not significant. Besides, we rejected the null hypothesis that demographic covariates are not related to crime rate.

To summarize, this third model set featuring criminal justice, economic policies, economic context as well demographic variables suggests that, ceteris paribus, a 10% increase in prbarr would lead to a ~5% decrease in crime rate, while a 10% increase in prbconv would lead to a ~4% decrease in crime rate. These estimated effects are statistically significant and have important practical meaning. Politicians should adjust criminal justice system policies depending on demographic variables.

### Model set #4

For our fourth model, we aim to evaluate the robustness of the relationship between crime policy and crime when controlling for economic, demographic, and geographic context variables. In addition, we will also reassess the impact of economic policy on crime rate.

We used the following independent variables to predict crime:

*Crime policy*

1. log(prbarr)
2. log(prbconv)

*Economic policy*

1. log(wtrd)

*Economic context*

1. log(wfed)
2. log(taxpc)

*Demographic*

1. density
2. pctymle
3. pctmin80

*Geographic*

1. west
2. central
3. urban

#### Discussion

Figure 14 shows that different geographies show the same pattern when it comes to urban versus rural: rural counties have a lower crime rate compared to urban ones (as expected). However, urban and density are highly correlated as seen early, so adding urban to the model may not improve fits. We can also see that central and eastern counties are more similar in terms of crime rate versus western counties, which have lower crime overall. Thus, we don’t expect the covariate for central to be significant.

# plot barplots  
ggplot(geo\_long, aes(x = location, y = crmrte, fill = type)) +  
 geom\_bar(position="dodge", stat="summary") +  
 scale\_fill\_manual(values = wes\_palette("GrandBudapest1", n = 3)) +   
 theme\_minimal() +  
 ggtitle("Barplot of Contextual Variables: Geography") +  
 xlab("") +  
 ylab("Average Crime Rate")

## No summary function supplied, defaulting to `mean\_se()

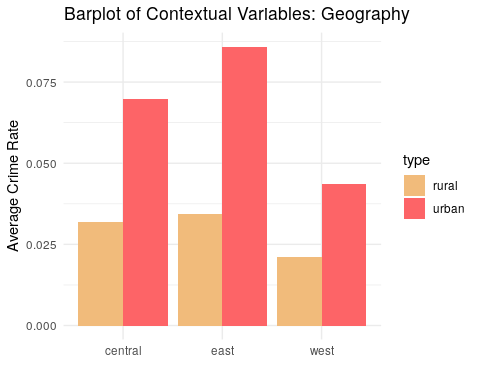


Figure 14

*Statistics Interpretation*

mod4\_1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle + west + central + urban, data = data2)  
mod4\_2 <- lm(log(crmrte) ~ log(prbarr)\*west + log(prbconv)\*west + log(prbarr)\*central + log(prbconv)\*central + log(wtrd) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = data2) # west and central interaction  
mod4\_3 <- lm(log(crmrte) ~ log(prbarr)\*urban + log(prbconv)\*urban + log(`wtrd`) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = data2) # urban interactions  
mod4\_4 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(`wtrd`) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle + west , data = data2) # west only, no interactions  
mod4\_5 <- lm(log(crmrte) ~ log(prbarr)\*west + log(prbconv)\*west + log(`wtrd`) + log(wfed) + log(taxpc) + density + pctmin80 + pctymle, data = data2) # west only, interactions  
  
# generate robust standard errors  
se\_mod4\_1 <- sqrt(diag(vcovHC(mod4\_1)))  
se\_mod4\_2 <- sqrt(diag(vcovHC(mod4\_2)))  
se\_mod4\_3 <- sqrt(diag(vcovHC(mod4\_3)))  
se\_mod4\_4 <- sqrt(diag(vcovHC(mod4\_4)))  
se\_mod4\_5 <- sqrt(diag(vcovHC(mod4\_5)))  
  
# produce regression table  
stargazer(  
 mod4\_1  
 , mod4\_2  
 , mod4\_3  
 , mod4\_4  
 , mod4\_5  
 , type = "text"  
 , se = list(se\_mod4\_1, se\_mod4\_2, se\_mod4\_3, se\_mod4\_4, se\_mod4\_5)  
 , star.cutoffs = c(0.05, 0.01, 0.001)  
 , notes= c("Robust SE")  
 , add.lines=list(c("BIC", round(BIC(mod4\_1),1), round(BIC(mod4\_2),1) , round(BIC(mod4\_3),1) , round(BIC(mod4\_4),1) , round(BIC(mod4\_5),1)  
 )  
))

##   
## ===========================================================================================================================================  
## Dependent variable:   
## ----------------------------------------------------------------------------------------------------------------------  
## log(crmrte)   
## (1) (2) (3) (4) (5)   
## -------------------------------------------------------------------------------------------------------------------------------------------  
## log(prbarr) -0.473\*\*\* -0.661\*\*\* -0.510\*\*\* -0.482\*\*\* -0.508\*\*\*   
## (0.109) (0.127) (0.103) (0.111) (0.147)   
##   
## log(prbconv) -0.333\*\* -0.354\* -0.321\*\* -0.327\*\* -0.408\*\*   
## (0.104) (0.180) (0.099) (0.100) (0.149)   
##   
## log(wtrd) 0.020 -0.043 0.038 0.009 0.008   
## (0.368) (0.381) (0.368) (0.366) (0.362)   
##   
## log(wfed) 1.083\*\* 1.165\*\* 1.013\* 1.071\* 1.067\*   
## (0.407) (0.405) (0.450) (0.421) (0.421)   
##   
## log(taxpc) 0.260 0.183 0.333 0.260 0.260   
## (0.291) (0.334) (0.328) (0.302) (0.298)   
##   
## density 0.136\* 0.084 0.135\* 0.081\* 0.077   
## (0.060) (0.045) (0.066) (0.041) (0.046)   
##   
## pctmin80 0.837\* 0.862 1.194\*\*\* 0.931\*\* 0.951\*\*\*   
## (0.398) (0.442) (0.220) (0.306) (0.289)   
##   
## pctymle 2.667 2.738 3.214 2.988 2.790   
## (2.956) (3.152) (3.222) (3.263) (3.230)   
##   
## log(prbarr):west 0.167 0.006   
## (0.219) (0.217)   
##   
## west:log(prbconv) 0.117 0.177   
## (0.238) (0.217)   
##   
## log(prbarr):central 0.317   
## (0.234)   
##   
## log(prbconv):central -0.182   
## (0.214)   
##   
## west -0.175 0.129 -0.113 0.035   
## (0.157) (0.359) (0.107) (0.342)   
##   
## central -0.112 0.196   
## (0.106) (0.443)   
##   
## log(prbarr):urban 0.418   
## (1.055)   
##   
## urban:log(prbconv) -0.105   
## (0.987)   
##   
## urban -0.290 0.231   
## (0.232) (2.432)   
##   
## Constant -12.554\*\*\* -12.677\*\*\* -12.748\*\*\* -12.486\*\*\* -12.539\*\*\*   
## (2.261) (2.320) (2.515) (2.335) (2.350)   
##   
## -------------------------------------------------------------------------------------------------------------------------------------------  
## BIC 70.6 77.9 71.7 65.6 71.9   
## Observations 89 89 89 89 89   
## R2 0.769 0.784 0.765 0.758 0.765   
## Adjusted R2 0.735 0.743 0.732 0.730 0.732   
## Residual Std. Error 0.279 (df = 77) 0.275 (df = 74) 0.280 (df = 77) 0.281 (df = 79) 0.281 (df = 77)   
## F Statistic 23.239\*\*\* (df = 11; 77) 19.185\*\*\* (df = 14; 74) 22.848\*\*\* (df = 11; 77) 27.487\*\*\* (df = 9; 79) 22.803\*\*\* (df = 11; 77)  
## ===========================================================================================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
## Robust SE

mod4 <- mod4\_4  
se\_mod4 <- se\_mod4\_4

We can see that the addition of geographic variables did not reduce BIC nor improve our model fit vs model set #3. Further, we found from the F-test that geographical variables are not jointly significant (p-value well above 0.05).

Across equations of this model set, our preferred one is column 4, which features only west as geographic variables. Column 4 has the lowest BIC, and our external research suggests that this area of the state is in fact different from the others.

When we added geographic variables to the model, the point estimates of the impact of probability of arrest and probability of conviction on crime remains virtually unchanged versus model #3.

# F-test: joint significance of geographic characteristics  
print("F-test results for joint significance of geographic characteristics")

## [1] "F-test results for joint significance of geographic characteristics"

linearHypothesis(mod4\_1, c("west=0", "central=0", "urban=0"), vcov=vcovHC)

## Linear hypothesis test  
##   
## Hypothesis:  
## west = 0  
## central = 0  
## urban = 0  
##   
## Model 1: restricted model  
## Model 2: log(crmrte) ~ log(prbarr) + log(prbconv) + log(wtrd) + log(wfed) +   
## log(taxpc) + density + pctmin80 + pctymle + west + central +   
## urban  
##   
## Note: Coefficient covariance matrix supplied.  
##   
## Res.Df Df F Pr(>F)  
## 1 80   
## 2 77 3 0.8147 0.4896

*Conclusion*

The most important takeaways from model set #4 are (1) the estimated effect of criminal justice policies on crime rate is robust to the inclusion of geographic controls, and (2) even controlling for demographics, the coefficient on our proxy for economic policy (log(wtrd)) remains statistically not significant. Also importantly, we fail to reject the null hypothesis that geographic covariates are not related to crime rate.

Our preferred specification in this set (column 4) suggests that, ceteris paribus, a 10% increase in prbarr would lead to a ~5% decrease in crime rate, while a 10% increase in prbconv would lead to a ~4% decrease in crime rate - these figures are the same of model set #3 and in fact not much different from model sets #1 (only criminal justice system variables) and #2 (criminal justice system + economic variables). Again, it’s fair to argue that this figures have important practical meaning, as a 4-5% change in crime rate may have material impact on people’s lives.

## Regression Table

Across all equations estimated, we prefer equation 3 in the table below. It’s from model set #3, which uses criminal justice, economic policies, economic context and demographic variables (but not geographic variables) to predict crime rate. In fact, geographic determinants explain little of the cross-county variation of crime rate, probably because North Carolina is a small state. Besides, Equation 3 has the lowest BIC, which justifies our choice.

The coefficients of both criminal justice system policy variables (pbarr and pconv) are statistically significant and have the expected (negative) sign - for a 10% rise in each probability individually, the negative impact on crime rate hovers around 5% for the probability arrest and around 3% for the probability of conviction. Our estimates are robust to different model specifications.

Not all our covariates are significant and so we are not confident on their relationship with crime.

We could not find a significant impact of economic policies (proxied by wages in lower-paying industries) on crime. The main contextual variables around the economy - wfed and taxpc - show different patterns. The latter is not statistically significant across any of the models below. The former has a positive relationship with crime rate but its effects are partially absorbed when we add in covariates. This is aligned with our hunch that the relationship between this variable and crime rate would be greatly influenced by other factors and thus we shy away from any causal interpretation.

Across demographic determinants, the percentage of minorities, with a positive impact on crime, is the only significant variable.

stargazer(  
 mod1  
 , mod2  
 , mod3  
 , mod4  
 , type = "text"  
 , se = list(se\_mod4, se\_mod4, se\_mod4, se\_mod4)  
 , star.cutoffs = c(0.05, 0.01, 0.001)  
 , notes = c("Robust SE")  
 , add.lines=list(c("BIC", round(BIC(mod1),1), round(BIC(mod2),1) , round(BIC(mod3),1) , round(BIC(mod4),1)  
 )))

##   
## ===============================================================================================================  
## Dependent variable:   
## -------------------------------------------------------------------------------------------  
## log(crmrte)   
## (1) (2) (3) (4)   
## ---------------------------------------------------------------------------------------------------------------  
## log(prbarr) -0.730\*\*\* -0.567\*\*\* -0.504\*\*\* -0.482\*\*\*   
## (0.111) (0.111) (0.111) (0.111)   
##   
## log(prbconv) -0.442\*\*\* -0.402\*\*\* -0.329\*\* -0.327\*\*   
## (0.100) (0.100) (0.100) (0.100)   
##   
## log(wtrd) 0.074 0.075 0.009   
## (0.366) (0.366) (0.366)   
##   
## log(wfed) 1.559\*\*\* 1.072\* 1.071\*   
## (0.421) (0.421) (0.421)   
##   
## log(taxpc) 0.374 0.289 0.260   
## (0.302) (0.302) (0.302)   
##   
## density 0.079 0.081\*   
## (0.041) (0.041)   
##   
## pctmin80 1.127\*\*\* 0.931\*\*   
## (0.306) (0.306)   
##   
## pctymle 3.087 2.988   
## (3.263) (3.263)   
##   
## west -0.113   
## (0.107)   
##   
## Constant -4.822\* -15.810\*\*\* -13.064\*\*\* -12.486\*\*\*   
## (2.335) (2.335) (2.335) (2.335)   
##   
## ---------------------------------------------------------------------------------------------------------------  
## BIC 114.9 92.9 62.6 65.6   
## Observations 89 89 89 89   
## R2 0.400 0.597 0.754 0.758   
## Adjusted R2 0.386 0.573 0.729 0.730   
## Residual Std. Error 0.424 (df = 86) 0.354 (df = 83) 0.282 (df = 80) 0.281 (df = 79)   
## F Statistic 28.671\*\*\* (df = 2; 86) 24.614\*\*\* (df = 5; 83) 30.642\*\*\* (df = 8; 80) 27.487\*\*\* (df = 9; 79)  
## ===============================================================================================================  
## Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
## Robust SE

# get coefficients  
mod3\_coefs = coeftest(mod3, vcov. = vcovHC)  
  
# calculate impact of policy variables on crime rate for a 10% relative increase  
print(paste0("% increase on crime rate given a 10% increase on probability of arrest: ", round(100\*((1.10^mod3\_coefs[2]) - 1), 2), "%"))

## [1] "% increase on crime rate given a 10% increase on probability of arrest: -4.69%"

print(paste0("% increase on crime rate given a 10% increase on probability of conviction: ", round(100\*((1.10^mod3\_coefs[3]) - 1), 2), "%"))

## [1] "% increase on crime rate given a 10% increase on probability of conviction: -3.08%"

print(paste0("% increase on crime rate given a 10% increase on federal employees' wages: ", round(100\*((1.10^mod3\_coefs[5]) - 1), 2), "%"))

## [1] "% increase on crime rate given a 10% increase on federal employees' wages: 10.75%"

print(paste0("% increase on crime rate given a 1% increase on people % minority: ", round(100\*(exp(mod3\_coefs[8])-1), 2), "%"))

## [1] "% increase on crime rate given a 1% increase on people % minority: 208.77%"

## Checking the 6 Classical Linear Model assumptions for our preferred model (Model #3)

### Linear population model

This assumption is automatically fulfilled because we haven’t constrained the error term, i.e we haven’t required it to be normal. So there’s nothing to check at this point.

### Random Sampling

North Carolina has 100 counties and this dataset has only 90. To check random sampling, we need more background knowledge of why these remaining counties were omitted. In general, we are concerned about possible problems with independence. For example, maybe omitted counties are the poorest ones where it’s more difficult to gather data. If these counties were included, we could perhaps capture better the effect of economic policy variables.

### No perfect collinearity

There is no need to explicitly check for perfect collinearity, because R will alert us if this rare condition happens.

### Zero Conditional Mean

Let’s take a look at the diagnostic plots:

plot(mod3, which = 1)

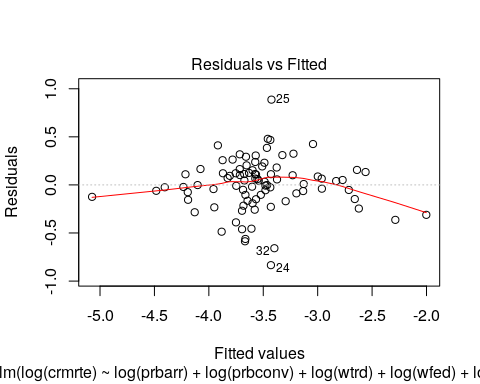


Figure 15

Notice that there is only a slight deviation from zero conditional mean, indicated by a not perfectly flat curve. This means that our coefficients maybe be biased. Next section, we discuss some omitted variables candidates and the bias direction they would imply.

### Homoskedasticity

Our residuals versus fitted values plot seems to indicate heteroskedasticity based on the scale location plot. We see a bump for values close to the mean and high end of the distribution.

plot(mod3, which = 3)

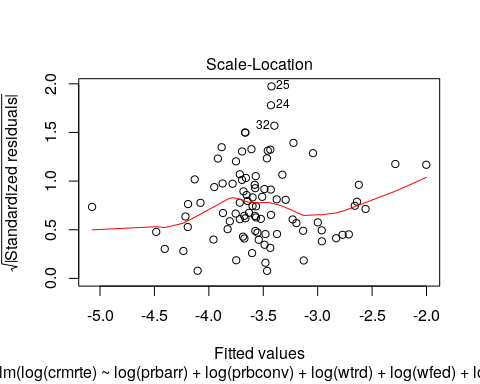


Figure 16

We used robust standard error to evaluate all our models, which is the indicated strategy for dealing with heteroskedasticity.

### Normality of Errors

To check normality of errors, we can look at the qqplot that’s part of R’s standard diagnostics.

plot(mod3, which = 2)

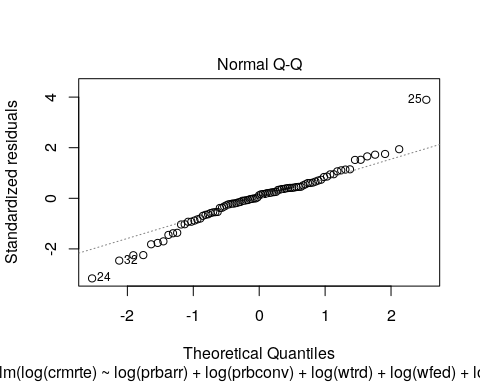


Figure 17

We can also visually look at the residuals directly.

hist(mod3$residuals, breaks = 20, main = "Residuals from Linear Model Predicting log(Crime Rate)")

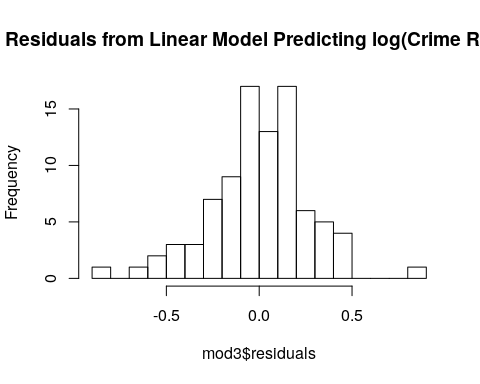


Figure 18

Both methods suggest we have a leftward skew. However, we have a large sample size, so the CLT tells us that our estimators will have a normal sampling distribution. The histogram confirms that we aren’t in a situation with an extreme skew, so n=90 should be sufficient for the CLT.

# Discussion of Possible Omitted Variables (OV)

As discussed throughout the text - and hinted by “Residual vs Fitted” and “Scale-Location” plots - we have reason to believe that we lack many variables that we expect relate to crime and can affect the quantification of the relationships estimated above.

In order to access the direction of omitted variable bias (OVB) on the estimated coefficients of criminal justice system variables resulting from each of the OV, we analyze the product of two correlations:

* = correlation between the omitted variable and crmrte
* = correlation between the omitted variable and prbarr/prbconv

As we assume the true relationship between crmrte and prbarr/prbconv () is negative, a positive OVB biases the estimated effect of prbarr/prbconv on crmrte towards zero, while a negative OVB biases the estimated effect away from zero.

**Unemployment** ( and -> towards zero) Unemployment rate is likely positively related to crime rate (so ), people in need are more prone to commit crimes. Unemployment rate is also arguably positively related to prbarr/prbconv (so ) due to the negative stigma that unemployment people have in society. Thus this OV should bias the estimated effect of prbarr/prbconv on crmrte away from zero.

**Education levels of the populace** ( and -> towards zero)

Although we may be able to proxy education levels by looking at which region a county is in, we would not have a very good measure of overall education levels. A county can have a university, but still have low education levels overall. One would predict that education is negatively related to crime - the more educated, the less crime one would expect to commit (so ). The relationship between education levels and prbarr/prbconv is arguably negative due to the negative stigma that less educated people have in society (so ).

**Public works & investments** ( and -> towards zero)

One would predict that the relationship between and public works and investments with crime would be negative (so ). More nonprofits and community organizations could strengthen ties in communities, thus lowering crime rates. This factor would probably be inversely related to arrest and conviction rates, since a more punitive society may invest less in rewards to the community (so ).

**Individual Health (mental and physical)** ( and -> towards zero)

Health - mental and physical - would likely be related to less crime. The healthier you are, the more mental and physical resources you have to solve problems in ways that do not necessitate committing a crime (so ). The direction of the relationship to arrest or conviction rates is fuzzier but one could imagine that communities that have a focus on health may tend to be less punitive (so ).

**Community Health** ( and -> towards zero)

Community health is about the cohesiveness of a community. If you have good supports in your community, you would be less likely to commit crime (so ). This factor would probably be inversely related to arrest and conviction rates, since a more loving, cohesive community may be less likely to turn to punishment (so ).

**Inequality** ( and -> away from zero)

With more inequality, one would expect to see more crime - more anger & therefore reason to commit crime combined with greater levels of deprivation and need to commit crime would naturally lead to more crime (so ). A more egalitarian society may have less punitive practice, since people would see each other as closer to their own group. This can help humanize other people and make it harder to commit cruelties (so ).

We cannot exactly determine the size of each of the biases above, however since most of those are in the same direction the net effect likely have led us to underestimate the (negative) relationship between crime rate and these criminal justice policies.

What about our economic policy variables? Assuming that wtrd is a good proxy for minimum wage, omitted variables could have caused the coefficient on wtrd to not be significant across our models. Recall that, as we assume the true relationship between crmrte and minimum wage is negative, a positive OVB biases the estimated effect of wtrd on crmrte towards zero, while a negative OVB biases the estimated effect away from zero.

The unemployment rate is an important candidate of OV that can be biasing the estimated effect of wtrd on crmrte *towards zero*. As argued above, the unemployment rate is likely positively related to crime rate, as people in need are more prone to commit crimes. Meanwhile, even though there has been mixed evidence around increases in minimum wage and unemployment, many people believe that unemployment and minimum wage increases are positively correlated, so we will make that assumption here. Thus, including an unemployment variable could increase the estimated impact of economic policy. We cannot determine the size of this relationship but since evidence is mixed we would guess that it would be a medium impact.

# Conclusion

Our analyses suggest that acting on crime deterrents within the criminal justice system (like arrests and convictions) is an important policy target for preventing crime. In fact, for a 10% rise in the probability of arrests and convictions individually, the negative impact on crime rate hovers between 4-7% for the arrests and between 3-5% for convictions. Helpful follow-ups to this investigation would be to leverage variation across time and possibly test for large changes in policy that, for example, would help in designing a regression discontinuity model. However, our findings can be used to hone in on future policy changes that can be evaluated with more rigorous causal analyses.

Most of the omitted variables we considered as potentially biasing the estimated effect of arrests and convictions rates on crime rate would cause a bias towards zero - i.e., including them in the model would likely lead to an even stronger estimated coefficient of arrests and convictions rates in our crime model. This finding made us more confident about the call that politicians could leverage criminal justice policies to curb crime rate.

On the other hand, we could not find much support for the importance of economic policy variables to prevent crimes. In the absence of a better measure, we used average wages in the trade industry as a proxy for a minimum wage and found no statistically significant impact of this wage on crime. We argued that there are omitted factors (such as the unemployment rate) that would likely bias the estimated coefficient of average wage towards zero.

Real life variations in these counties lead to some extreme values - we did not remove these, as they were likely edge cases rather than erroneous data. Any policies should take the local context into consideration for these extreme cases. A stronger follow-up study would control for more of these within-county factors to produce more robust causal findings. The distribution of minorities is a good example of one of these important contextual variable that is not (at least ethically) under the influence of a politician. Communities with higher rates of minorities or crime may require additional resources or different types of policies to prevent crime. For example, if the actual number of crimes is different from the number of reported crimes - in a community with a lot of racial conflict some groups over-report others to the police - we would see elevated measures of crime. If policies aim to reduce actual crime, we would need more nuanced measures to track and address this crime issue (e.g., hosting workshops to reduce racism).

In conclusion, we recommend policies that increase arrests and convictions alongside those that help increase the cost of living in an area. However, further analysis with the inclusion of omitted variables is needed to guarantee the actual impact on crime rate of changing these factors. An important tactic for any politician would be to fund research that can more rigorously tease out these causal factors.