

# Lab 3: Policy Recommendations for Reducing Crime

Ammara Essa, Curtis Lin, and David Wheeler

## Table of Contents:

- [Introduction, variable selection, and research question](#)
- [Descriptive analysis | exploratory data analysis](#)
- [Main analysis](#)
- [Diagnostics](#)
- [Robustness](#)
- [Conclusions](#)
- [Appendix](#)

## Introduction, variable selection & research question

[Back to Table of Contents](#)

### Introduction

Public safety is a major concern for constituents and elected officials, especially during elections. By identifying determinants of crime (represented by *crmte*: crimes committed per person) in North Carolina (NC), we can develop policy suggestions to help the campaign deliver on their promise of reducing crime rates.

### Variable selection

Of the 22 candidate explanatory variables ( $25 - \text{crmte} - \text{year} - \text{county} = 22$ ) only a subset can feasibly be altered with cheap and simple changes by politicians at the state level.

- By process of elimination the wage variables will likely not be simple or cheap to alter by politicians. Moreover, the potential links between crime and weekly wages are likely too nuanced and will be met with resistance if lowered.
- To change demographic variables like *pctmin80* and *pctymyle* without being accused of prejudice seems futile. To attempt such a task would not be cheap or simple.
- The *mix* variable is not responsive to change by politicians, unless the politicians are also criminals.

- Similarly, the geographic variables like *west*, *urban* and *central* are not amenable to change, unless politicians want to relocate people without infringing on their rights.
- Likewise, this applies to *density*- it seems unlikely for a politician to enforce laws that regulate the number of people that can live in a given space. However, long-term housing policies may alleviate population density but quickly or cheaply.
- Although tax revenues per capita (*taxpc*) can be changed by politicians, one needs only to recall the most recent federal attempts to change taxes to note that changes to our tax structure are not simple and cheap, but are often met with resistance.
- Finally, to ask politicians to change the average sentence length (*avglen*) and the 'probability' of prison sentences (*prbpris*) appears to be asking too much. Again, although there is precedence for these types of changes, they are certainly not simple and cheap and will likely require an arduous process punctuated by protests and episodes of public resistance.

Although these variables may not have potential to be changed (simply and cheaply) by politicians, we will still include some in our model building process as covariates and to control for their influences on crime. Finally, this leaves several variables which could be simply and cheaply altered by politicians to reduce crime: *polpc*, *prbarr*, and *prbconv*

## Research question

Can better policing, indicated by *polpc*, *prbarr*, *prbconv*, reduce crime rates (*crmte* = crimes committed per person) in NC?

**Importance:** This question is important because, if we can improve policing, then it follows that crime rates in NC should decline. Moreover, changes in policing are relatively simple and cheap to implement. Finally, if crime rates can be reduced by better policing then we can provide relatively cheap, simple, and actionable policy recommendations to the campaign.

### ***Hypothesized model that explains crimes rate (crimes committed per person):***

$$crmte = \beta_0 + \beta_1 polpc + \beta_2 prbarr + \beta_3 prbconv + \beta_4 prbarr \cdot prbconv + u$$

### **Null hypotheses:**

1. Police per capita (*polpc*) has no effect on *y*, crimes committed per person.
2. 'Probability' of arrest (*prbarr*) has no effect on *y*, crimes committed per person.
3. 'Probability' of conviction (*prbconv*) has no effect on *y*, crimes committed per person.
4. The 'probability' of conviction (*prbconv*) does not depend on the 'probability' of arrest (*prbarr*).

- $H_0^{(1)}: \frac{\partial crmte}{\partial polpc} = 0 \Rightarrow \beta_1 = 0$
- $H_0^{(2)}: \frac{\partial crmte}{\partial prbarr} = 0 \Rightarrow \beta_2 = 0$
- $H_0^{(3)}: \frac{\partial crmte}{\partial prbconv} = 0 \Rightarrow \beta_3 = 0$
- $H_0^{(4)}: \frac{\partial crmte}{\partial prbconv} \Big|_{arrested} = 0 \Rightarrow \beta_4 = 0$

### **Alternative hypotheses:**

1. Police per capita (*polpc*) has an effect on *y*, crimes committed per person.

2. 'Probability' of arrest (*prbarr*) has an effect on  $y$ , crimes committed per person.
3. 'Probability' of conviction (*prbconv*) has an effect on  $y$ , crimes committed per person.
4. The 'probability' of conviction (*prbconv*) depends on the 'probability' of arrest (*prbarr*).

- $H_0^{(1)}: \beta_1 \neq 0$
- $H_0^{(2)}: \beta_2 \neq 0$
- $H_0^{(3)}: \beta_3 \neq 0$
- $H_0^{(4)}: \beta_4 \neq 0$

### Covariates of interest

Although, the variables above will be key in providing advice and guidance to the political campaign, we need to include other variables to control for other sources of variation.

- geography (e.g. *west*, *urban*, and *central*) since crime is often associated with place & resultant models could be used to inform policies for specific locales.
- demographics (e.g. *density*, *pctmin80* and *pctymle*) since crime is often associated with demographics & resultant models could be used to help demographics of interest reduce crime.
- economics (e.g. *taxpc*, *wfed*, *wtuc*, and *wtrd*) since wages could conceivably be associated with crime rates & resultant models could be used to justify wage change policies to reduce crime rates.
- imprisonment (e.g. *prbpris* and *avglen*) since a high 'probability' of a prison sentence and a long average sentence length could reduce the likelihood that criminals reoffend.

### Context

- We recognize that the dataset used herein is a cross sectional sample and provides a snapshot of the population in time (1987). Thus, this limited dataset may not be best for causal inference. However, based on our understanding of the variables, we will attempt to draw conclusions and provide recommendations using this data. Moreover, we will discuss a few key sources of omitted variable bias that we believe have been potentially introduced since we can no longer control for them using multi-year panel data.
- NC has 100 counties, the latest incorporated in 1911 ([https://en.wikipedia.org/wiki/List\\_of\\_counties\\_in\\_North\\_Carolina](https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina)) ([https://en.wikipedia.org/wiki/List\\_of\\_counties\\_in\\_North\\_Carolina](https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina)). However, the dataset available to us only captures 90 unique counties. With that said, we will use this data to draw conclusions regarding crime rates in NC.
- Based on the summary of average prison sentence (Table 1), we see that the sentence length is approximately 10 days on average and at most approximately 21 days. This may be an indicator that law enforcement is dealing with more misdemeanor crimes than felony crimes.
- Finally, while most of the variables were collected in 1987, two of the variables, *pctmin80* and *pctymle* were collected in 1980!  $\therefore$  any effect that these variables have on crime rates may be confounded by time- we may not be able tease apart differences due to *pctmin80* and *pctymle* from differences in time.

# Descriptive analyses | exploratory data analysis

[Back to Table of Contents](#)

## Install and invoke packages

```
In [9]: 1 # install.packages("car")
2 # install.packages("GGally")
3 # install.packages("scatterplot3d")
4 # install.packages("lmtest")
5 # install.packages("sandwich")
6 # install.packages("corrplot", repos="http://cran.us.r-project.org")
7 library(Hmisc)
8 library(corrplot)
9 library(car)
10 library(stargazer)
11 library(GGally)
12 library(scatterplot3d)
13 library(lmtest)
14 library(sandwich)
```

## Load data

```
In [10]: 1 df = read.csv("crime_v2.csv", header=T)
```

## Check classes of each variable

Note that *prbconv* is treated as a factor, *county*, *west*, *central* and *urban* are treated as numeric variables, and *year* is a constant.

```
In [11]: 1 str(df)

'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 ...
 $ wmfgr : 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 ...
```

**Convert *prbconv* from factor to numeric since it expresses a *continuous* outcome**

```
In [74]: 1 df$prbconv = as.numeric(as.character(df$prbconv))
```

**Convert *county*, *west*, *central*, and *urban* from numerics to factors**

```
In [13]: 1 df$county= factor(df$county)
2 df$west= factor(df$west)
3 df$central= factor(df$central)
4 df$urban= factor(df$urban)
```

**Table 1: Descriptive statistics**

Note:

- *prbarr* and *prbconv* variables have values that exceed 1.
- the variables are expressed at different scales, from 0.002 to 2177.1
- the weekly wage variables look like they have been censored/bottom and or top-coded. Perhaps extreme values were ommitted...

```
In [14]: 1  stargazer(df, type="text",
          2          title="Descriptive statistics",
          3          digits= 1)
```

#### Descriptive statistics

=====							
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
-----							
year	91	87.0	0.0	87.0	87.0	87.0	87.0
crmrt	91	0.03	0.02	0.01	0.02	0.04	0.1
prbarr	91	0.3	0.1	0.1	0.2	0.3	1.1
prbconv	91	0.6	0.4	0.1	0.3	0.6	2.1
prbpris	91	0.4	0.1	0.2	0.4	0.5	0.6
avgsen	91	9.6	2.8	5.4	7.3	11.4	20.7
polpc	91	0.002	0.001	0.001	0.001	0.002	0.01
density	91	1.4	1.5	0.000	0.5	1.6	8.8
taxpc	91	38.1	13.1	25.7	30.7	40.9	119.8
pctmin80	91	25.5	17.0	1.3	9.8	38.1	64.3
wcon	91	285.4	47.5	193.6	250.8	314.8	436.8
wtuc	91	411.7	77.3	187.6	374.6	443.4	613.2
wtrd	91	211.6	34.2	154.2	190.9	225.1	354.7
wfir	91	322.1	53.9	170.9	286.5	345.4	509.5
wser	91	275.6	206.3	133.0	229.7	280.5	2,177.1
wmfg	91	335.6	87.8	157.4	288.9	359.6	646.8
wfed	91	442.9	59.7	326.1	400.2	478.0	598.0
wsta	91	357.5	43.1	258.3	329.3	382.6	499.6
wloc	91	312.7	28.2	239.2	297.3	329.2	388.1
mix	91	0.1	0.1	0.02	0.1	0.2	0.5
pctymle	91	0.1	0.02	0.1	0.1	0.1	0.2
-----							

From `str(df)`, a total of 97 observations are included in the dataset. However, in the `stargazer` table the number of observation is 91 ( $N = 91$ ). The difference of number of observations between `str(df)` and `stargazer` table indicates that there might be 6 missing values ( $97 - 91 = 6$ ) in the dataset. The missing values in the dataset will be identified in the following section.

#### Remove the *year* variable - it is a constant

```
In [15]: 1  df$year = NULL
```

#### Locate missing values, NAs

##### Table 2: Missing values

Note that rows 92-97 are all missing data, NAs. There are 6 missing values in the dataset.

```
In [16]: 1 apply(is.na(df), 2, which)
```

county	crmte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	...	wtuc	wtrd	wfi
92	92	92	92	92	92	92	92	92	92	...	92	92	92
93	93	93	93	93	93	93	93	93	93	...	93	93	93
94	94	94	94	94	94	94	94	94	94	...	94	94	94
95	95	95	95	95	95	95	95	95	95	...	95	95	95
96	96	96	96	96	96	96	96	96	96	...	96	96	96
97	97	97	97	97	97	97	97	97	97	...	97	97	97

### Remove rows with missing values, NAs, for analyses

```
In [17]: 1 crime.narm <- na.omit(df)
```

### Identify duplicate rows

Note two duplicate entries for county 193.

```
In [18]: 1 crime.narm[duplicated(crime.narm) == TRUE, ]
```

	county	crmte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west
89	193	0.0235277	0.266055	0.588859	0.423423	5.86	0.00117887	0.8138298	28.51783	

```
In [19]: 1 nrow(crime.narm)
```

91

### Delete duplicate row

```
In [20]: 1 crime.narm = crime.narm[-89, ]
```

```
In [21]: 1 nrow(crime.narm)
```

90

In the *prbarr* & *prbconv* variables, there are instances which are larger than 1.

```
In [22]: 1 paste("Summary of prbarr variable")
          2 summary(df$prbarr)
          3 paste("Summary of prbconv variable")
          4 summary(df$prbconv)
```

'Summary of prbarr variable'

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.09277	0.20568	0.27095	0.29492	0.34438	1.09091	6

'Summary of prbconv variable'

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.06838	0.34541	0.45283	0.55128	0.58886	2.12121	6

The *prbconv* has 10 cases with  $p > 1$

```
In [23]: 1 lenprbconv <- length(df[which(df$prbconv > 1),]$prbconv)
          2 paste("Number of prbconv > 1: ", lenprbconv)
```

'Number of prbconv > 1: 10'

The *prbarr* has 1 case with  $p > 1$

```
In [24]: 1 lenprbarr <- length(df[which(df$prbarr > 1),]$prbarr)
          2 paste("Number of prbarr > 1: ", lenprbarr)
```

'Number of prbarr > 1: 1'

### Justification for the inclusion of *prbarr* and *prconv* values $> 1$

As we can see, the two variables described by the study as *prbarr* and *prbconv* have values  $> 1$ . The first instinct may be to discard the values  $> 1$ , since the variable is supposedly a probability and probability  $p$  can only take valid values  $0 \leq p \leq 1$ . However, the research by C. Cornwell and W. Trumball (1994) study defines the variables as

- The probability of arrest ( $P_a$  or *prbarr* in the dataset) is the ratio of arrests to offenses
- The probability of conviction ( $P_c$  or *prbconv* in the dataset) is the ratio of convictions to arrests

In this case of *prbconv*, there are few reasons why the ratio of convictions to arrests may be greater than 1. The police could make a single arrest of multiple people, which could result in multiple convictions. Or a single person could be arrested but convicted for multiple crimes. Therefore, we do not consider *prbconv* to strictly indicate probability. Moreover, instead of truncating entries  $> 1$  to 1 or discarding the values completely, we will leave the entries in the dataset used for the analysis. We have used the same approach for *prbarr* and again did not discard values  $> 1$ .

More formally, in probability theory, all variables should follow  $f(x) \in [0, 1]$  for all  $x \in \Omega$ . However, the *prbarr* and *prbconv* indicated the ratio of arrests to offenses and the ratio to conviction, respectively. These two variables do not satisfy the axioms of probability theory. However, the *prbarr*  $> 1$  and *prbconv*  $> 1$  were kept for analysis because of the reasons described above.



## The dimensions of the final dataframe

```
In [25]: 1 dim(crime.narm)
```

```
90 24
```

## Transformations

As noted above, all continuous variables, except *avgsen* and the identifier *county*, were log-transformed because some of these variables are (i) expressed on different scales, from 0.002 to 2177.1 (**table 1**), (ii) not all linearly related (**figures A3-A6**), (iii) skewed (except: *prbpris*, *wtuc*, *wfir*, and *wloc*) (**figure A1 & A2**), and (iv) interpretation will be aided by log-log model specification. Categorical variables, such as *west*, *central*, etc...were not log-transformed. Moreover, *avgsen* was not log-transformed to aid interpretability. Here, we noted most variables are not normally distributed (**Figure A1**). Distributions of variables after log-transformation resolved most instances of non-symmetric distributions (**Figure A2**). Moreover, we verified that none of the transformed values were 0 or negative, therefore variables could transform without any undesirable modifications such as adding a constant value.

## Relationships among variables

Log-transform variables for correlation matrix (this will only be used for the matrix below)

```
In [26]: 1 log_df = log(crime.narm[,c(2:9,13:24)])
```

Bind log-transformed and untransformed data

```
In [27]: 1 cd = cbind(log_df,
2                 crime.narm$county,
3                 crime.narm$west,
4                 crime.narm$central,
5                 crime.narm$urban)
```

Correlation matrix

```
In [28]: 1 cm = rcorr(as.matrix(cd),type=c("spearman"))
```

**Figure 1a: Correlation matrix. Blue and red circles represent positive and negative Spearman correlations, respectively. Spearman was used instead of Pearson to accommodate categorical variables.**

Note:

- There is no perfect multicollinearity (**figures A3-A6**) among the explanatory variables - although *pctmin80* vs *west* is close.
- *ln(crmrte)* is correlated, when  $\alpha = 0.05$ , with all but 7 candidate explanatory variables. Note the *strong* correlations with several variables that politicians can actually influence, for example

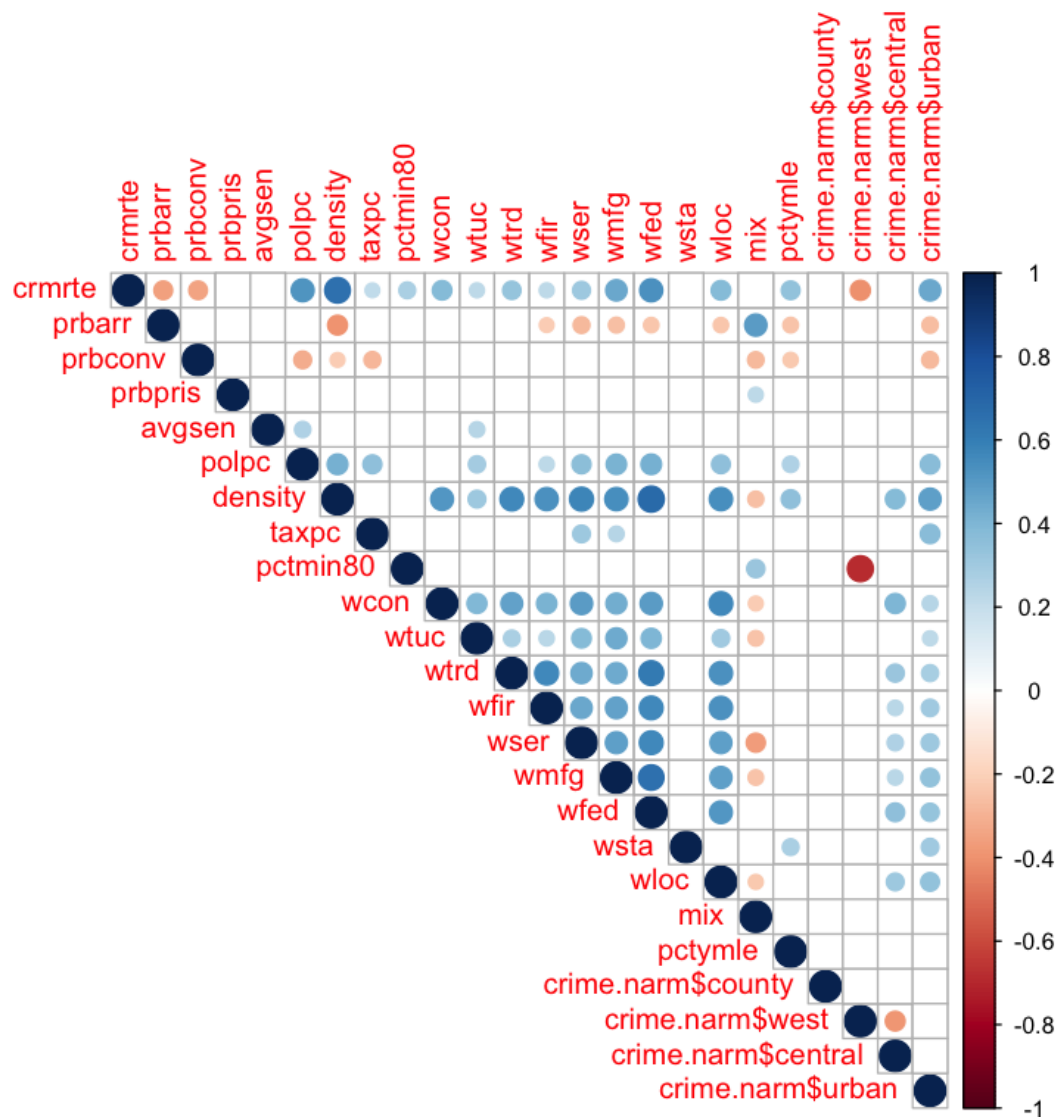
*polpc*. Although *wfed* is also *strongly* correlated with crime rates, it might be challenging for politicians to changes federal employee salaries without substantial resistance.

- See the plot below for the relationship between *crm rte* and the categorical variables.
- Finally, more detailed scatterplot matrices are presented in the appendix (**figures A3-A6**).

```

In [29]: 1 # Plot options
2 backup_options <- options()
3 options(backup_options)
4 # Continuous data
5 corrpplot(cm$r,
6           type = "upper",
7           #order='hclust',
8           p.mat = cm$P,
9           sig.level = 0.05,
10          insig = "blank")

```



**Figure 1b. Relationships between crime and continuous variables that are highly correlated with crime rates: police per capita and density.**

Note that crime rates are comparatively high in urban (red) areas and low in western (black). Therefore, *urban* might be an important covariate in subsequent models. There are clear and steep relationships between  $\ln(\text{crmte})$  &  $\ln(\text{density})$  and  $\ln(\text{crmte})$  &  $\ln(\text{polpc})$ . There are two outliers, *sensu lato*, from rows 51 and 79. The former is an outlier if  $\ln(\text{polpc})$  is regressed on  $\ln(\text{crmte})$ - it may *pull* the slope towards 0. The latter observation is likely influential (see Cook's distance discussion below) if  $\ln(\text{density})$  is regressed on  $\ln(\text{crmte})$ . Finally, when the outliers are removed the slopes appear more linear than when the outliers were included. These outliers are inspected in detail in the appendix.

```

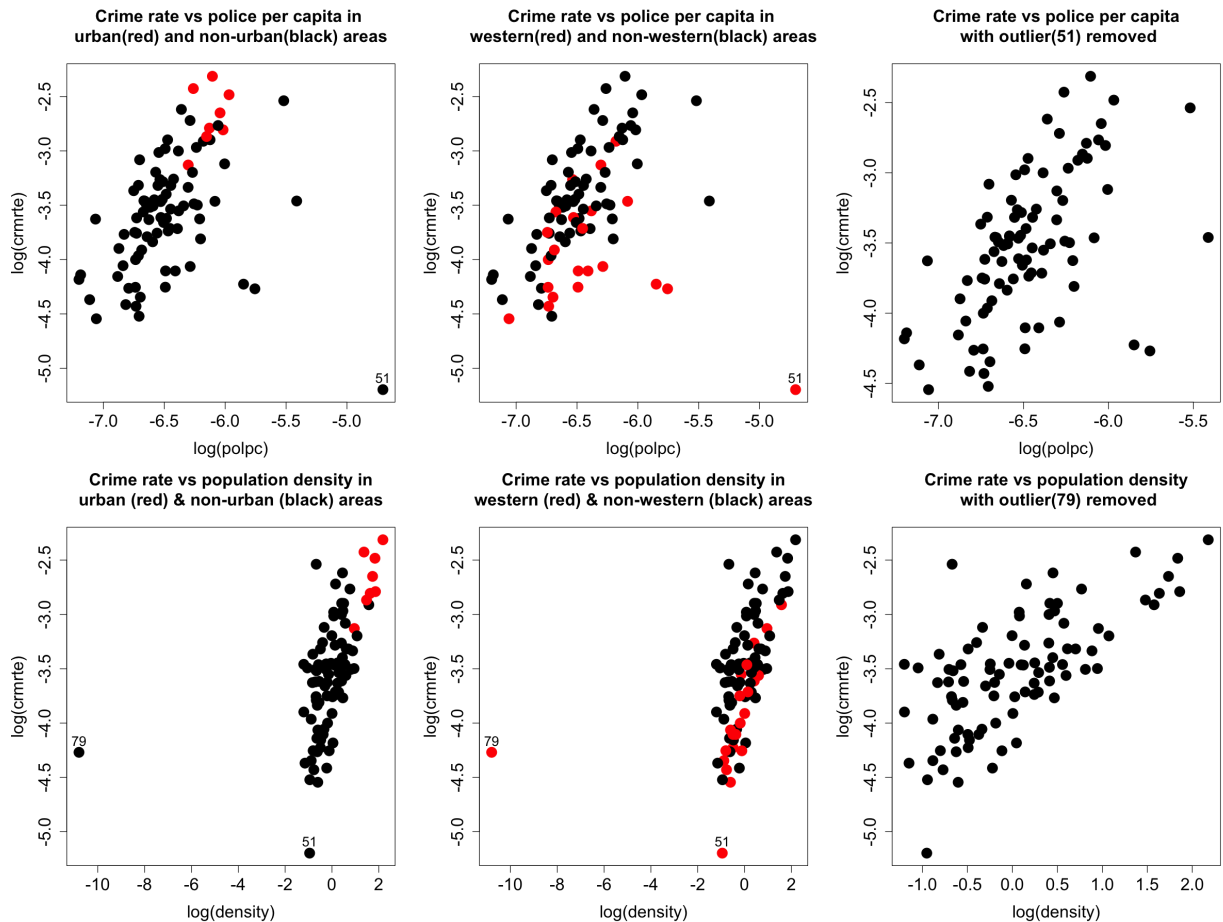
In [30]: 1 # Plot conditions
2 par(mfrow = c(2,3))
3 options(repr.plot.height = 15,
4         repr.plot.width = 20,
5         repr.plot.pointsize = 20)
6
7 # Crime rate vs polpc | urban
8 plot(log(crime.narm$crmrte)~log(crime.narm$polpc),
9      cex=2,pch=19,col=crime.narm$urban,
10     cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
11     main= "Crime rate vs police per capita in \
12 urban(red) and non-urban(black) areas\n",
13     ylab="log(crmrte)",xlab="log(polpc)")
14 text(log(crime.narm$crmrte)~log(crime.narm$polpc),
15      labels = ifelse(log(crime.narm$polpc) > -5,
16                      row.names(crime.narm), NA),
17      pos=3,cex=2)
18
19 # Crime rate vs polpc | west
20 plot(log(crime.narm$crmrte)~log(crime.narm$polpc),
21      cex=2,pch=19,col=crime.narm$west,
22     cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
23     main= "Crime rate vs police per capita in \
24 western(red) and non-western(black) areas\n",
25     ylab="log(crmrte)",xlab="log(polpc)")
26 text(log(crime.narm$crmrte)~log(crime.narm$polpc),
27      labels = ifelse(log(crime.narm$polpc) > -5,
28                      row.names(crime.narm), NA),
29      pos=3,cex=2)
30
31 # Crime rate vs polpc | no outlier
32 plot(log(crime.narm[-51,2])~log(crime.narm[-51,7]),
33      cex=2,pch=19,col="black",
34     cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
35     main= "Crime rate vs police per capita\n with outlier(51) removed
36     ylab="log(crmrte)",xlab="log(polpc)")
37
38 # Crime rate vs density | urban
39 plot(log(crime.narm$crmrte)~log(crime.narm$density),
40      cex=2,pch=19,col=crime.narm$urban,
41     cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
42     main= "Crime rate vs population density in\
43 urban (red) & non-urban (black) areas\n",
44     ylab="log(crmrte)",xlab="log(density)")
45 text(log(crime.narm$crmrte)~log(crime.narm$density),
46      labels = ifelse(log(crime.narm$density) < -8 |
47                      log(crime.narm$crmrte) < -5,
48                      row.names(crime.narm), NA),
49      pos=3,cex=2)
50
51 # Crime rate vs density | west
52 plot(log(crime.narm$crmrte)~log(crime.narm$density),
53      cex=2,pch=19,col=crime.narm$west,
54     cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
55     main= "Crime rate vs population density in\
56 western (red) & non-western (black) areas\n",

```

```

57     ylab="log(crmrte)", xlab="log(density)")
58     text(log(crime.narm$crmrate)-log(crime.narm$density),
59         labels = ifelse(log(crime.narm$density) < -8 |
60                         log(crime.narm$crmrate) < -5,
61                         row.names(crime.narm), NA),
62         pos=3, cex=2)
63
64     # Crime rate vs density | no outlier
65     plot(log(crime.narm[-79,2])-log(crime.narm[-79,8]),
66         cex=2, pch=19, col="black",
67         cex.main=2.5, cex.axis=2.5, cex.lab=2.5,
68         main="Crime rate vs population density\n with outlier(79) remove",
69         ylab="log(crmrte)", xlab="log(density)")
70
71

```



**Figure 1c: Relationship between crime and categorical variables.**

Note that crime rates differ between western & not-western and urban & not-urban regions in NC. Also note the unbalanced sampling sizes for urban vs not-urban regions. For central NC, crime rates look comparable within regions. Finally, note that the crime rate from county 51 is unexpectedly low.

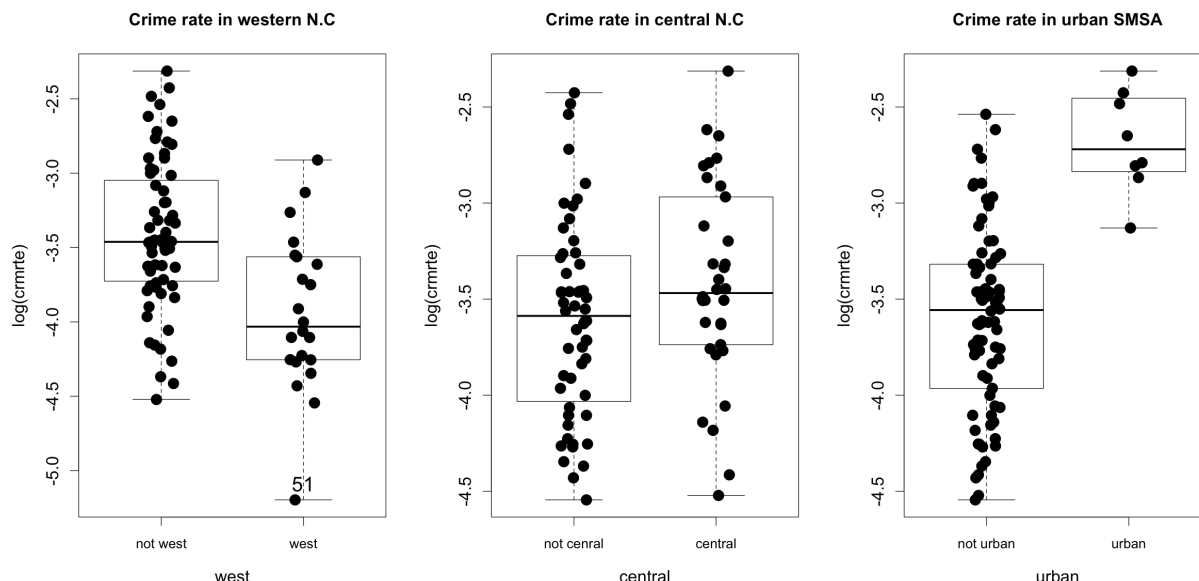


```

In [31]: 1 # Plot conditions
2 par(mfrow = c(1,3))
3 options(repr.plot.height = 10,
4         repr.plot.width = 20,
5         repr.plot.pointsize = 25)
6
7 # Crime in western NC
8 boxplot(log(crime.narm$scrmrte) ~ as.numeric(crime.narm$west),
9         cex=1.5, pch=19, xaxt='n', cex.main=2.5, cex.axis=2.5, cex.lab=1.5,
10         main= "Crime rate in western N.C",
11         ylab="log(crmrte)", xlab="west")
12 stripchart(log(crime.narm$scrmrte) ~ as.numeric(crime.narm$west),
13            vertical = TRUE, method="jitter", add=TRUE,
14            cex=2.5, pch=20, col='black')
15 text(log(crime.narm$scrmrte)~as.numeric(crime.narm$west),
16       labels = ifelse(log(crime.narm$scrmrte) < -5,
17                       row.names(crime.narm), NA),
18       pos=3, cex=3)
19 axis(1, at=c("1", "2"), labels=c("not west", "west"), cex.axis=2)
20
21 # Crime in central NC
22 boxplot(log(crime.narm$scrmrte) ~ as.numeric(crime.narm$central),
23         cex= 1.5, pch=19, xaxt = 'n', cex.main=2.5, cex.axis=2.5, cex.lab=1.5,
24         main= "Crime rate in central N.C",
25         ylab="log(crmrte)", xlab="central")
26 stripchart(log(crime.narm$scrmrte) ~ as.numeric(crime.narm$central),
27            vertical = TRUE, method="jitter", add=TRUE,
28            cex=2.5, pch=20, col='black')
29 text(log(crime.narm$scrmrte)~as.numeric(crime.narm$central),
30       labels = ifelse(log(crime.narm$scrmrte) < -5,
31                       row.names(crime.narm), NA),
32       pos=3, cex=3)
33 axis(1, at=c("1", "2"), labels=c("not cenral", "central"), cex.axis=2)
34
35 # Crime in urban NC
36 boxplot(log(crime.narm$scrmrte) ~ as.numeric(crime.narm$urban),
37         cex= 1.5, pch=19, xaxt='n', cex.main=2.5, cex.axis=2.5, cex.lab=1.5,
38         main= "Crime rate in urban SMSA",
39         ylab="log(crmrte)", xlab="urban")
40 stripchart(log(crime.narm$scrmrte) ~ as.numeric(crime.narm$urban),
41            vertical = TRUE, method="jitter", add=TRUE,
42            cex=2.5, pch=20, col='black')
43 text(log(crime.narm$scrmrte)~as.numeric(crime.narm$urban),
44       labels = ifelse(log(crime.narm$scrmrte) < -5,
45                       row.names(crime.narm), NA),
46       pos=3, cex=3)
47 axis(1, at=c("1", "2"), labels=c("not urban", "urban"), cex.axis=2)

```





## Summary

From the exploratory data analysis above we learned that, despite several anomalous observations, all but 7 candidate explanatory variables are correlated with crime rate. Of those variables, we will use (i) the 3 variables that we think politicians can influence as key variables in our models, (ii) covariates of interest that might improve model fit and enable use to control for other sources of variability, and (iii) all remaining variables to assess the robustness of our models.

# Main Analysis

[Back to Table of Contents](#)

**For inferential analysis, we would like to achieve 3 goals**

1. Model 1 will only include the explanatory variables of key interest.
2. Model 2 will include the covariates to find to balance between parsimony and accuracy and control for other sources of variation.
3. Model 3 will include all explanatory variables and demonstrate the robustness of our model specification.

**Model 1: Model with only key variables of interest (m\_key)**

$$\ln(\text{crmrte}) = \beta_0 + \beta_1 \ln(\text{polpc}) + \beta_2 \ln(\text{prbarr}) + \beta_3 \ln(\text{prbconv}) + \beta_4 \ln(\text{prbarr}) \cdot \ln(\text{prbconv}) + u$$

```
In [32]: 1 m_key <- lm(log(crmrte) ~ log(polpc) + log(prbarr) + log(prbconv)
          2           + log(prbconv)*log(prbarr), data = crime.narm)
```

**Test the hypotheses that the slopes,  $\beta_j$ , are equal to zero for *polpc*, *prbarr*, and *prbconv***

1.

$$H_0^{(1)} : \beta_1 \ln(polpc) = 0 \text{ vs. } H_A^{(1)} : \beta_1 \ln(polpc) \neq 0$$

In [33]:

```

1 # Hypothesis - H_0: Test if slopes are equal to zero
2 H0_p = c("log(polpc)=0")
3 # Wald/F test
4 linearHypothesis(
5     m_key, H0_p,
6     vcov=vcovHC(m_key),
7     singular.ok=TRUE
8 )

```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.541528	0.4638261

Fail to reject the  $H_0^{(1)} : \beta_1 \ln(polpc) = 0$ .  $polpc$  should stay in the model. The implications of this result will be discussed below.

2.

$$H_0^{(2)} : \beta_2 \ln(prbarr) = 0 \text{ vs. } H_A^{(2)} : \beta_2 \ln(prbarr) \neq 0$$

In [34]:

```

1 # Hypothesis - H_0: Test if slopes are equal to zero
2 H0_prba = c("log(prbarr) = 0")
3 # Wald/F test
4 linearHypothesis(
5     m_key, H0_prba,
6     vcov=vcovHC(m_key),
7     singular.ok=TRUE
8 )

```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.6383648	0.4265305

Fail to reject the  $H_0^{(2)} : \beta_2 \ln(prbarr) = 0$ .  $prbarr$  should stay in the model. The implications of this result will be discussed below.

3.

$$H_0^{(3)} : \beta_3 \ln(prbconv) = 0 \text{ vs. } H_A^{(3)} : \beta_3 \ln(prbconv) \neq 0$$

```
In [35]: 1 # Hypothesis - H_0: Test if slopes are equal to zero
2 H0_prbc = c("log(prbconv) = 0")
3 # Wald/F test
4 linearHypothesis(
5     m_key, H0_prbc,
6     vcov=vcovHC(m_key),
7     singular.ok=TRUE
8 )
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.001011247	0.974706

Fail to reject the  $H_0^{(1)} : \beta_3 \ln(\text{prbconv}) = 0$ .  $\text{prbconv}$  should stay in the model. The implications of this result will be discussed below.

4.

$H_0^{(4)} : \beta_4 \ln(\text{prbarr}) \cdot \ln(\text{prbconv}) = 0$  vs.  $H_A^{(4)} : \beta_4 \ln(\text{prbarr}) \cdot \ln(\text{prbconv}) \neq 0$

```
In [36]: 1 # Hypothesis - H_0:
2 # Test whether interactions between prbarr and prbconv could be dropped
3 H0 = c("log(prbarr):log(prbconv) = 0")
4 # Wald/F test
5 linearHypothesis(
6     m_key, H0,
7     vcov=vcovHC(m_key),
8     singular.ok=TRUE
9 )
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.2653271	0.6078207

Fail to reject the null hypothesis  $H_0 : \beta_4 \ln(\text{prbarr}) \cdot \ln(\text{prbconv}) = 0$ . Therefore, the interaction between  $\text{prbarr}$  and  $\text{prbconv}$  could not be dropped from the model.

## Model 2: A Model with key variables plus covariates (m\_associated)

$$\ln(\text{crmte}) = \beta_0 + \beta_1 \ln(\text{polpc}) + \beta_2 \ln(\text{prbarr}) + \beta_3 \ln(\text{prbconv}) + \beta_4 \ln(\text{prbarr}) \cdot \ln(\text{prbconv}) + \beta_5 \ln(\text{prbpris}) + \beta_6 \text{urban} + \beta_7 \ln(\text{taxpc}) + \beta_8 \ln(\text{wfed}) + \beta_9 \ln(\text{density}) + \beta_{10} \ln(\text{pctmin80}) + u$$

```
In [37]: 1 m_associated <- lm(log(crmrte) ~ log(polpc) + log(prbarr) + log(prbco
2          + log(prbarr)*log(prbconv)
3          + log(prbpris)
4          + log(taxpc) + log(wfed) + factor(urban
5          + log(density) + log(pctmin80), data =
```

### Model 3: A Model all variables (m\_all)

$$\ln(\text{crmrte}) = \beta_0 + \beta_1 \ln(\text{polpc}) + \beta_2 \ln(\text{prbarr}) + \beta_3 \ln(\text{prbconv}) + \beta_4 \ln(\text{prbpris}) \\ + \beta_5 \ln(\text{urban}) + \beta_6 \ln(\text{taxpc}) + \beta_7 \ln(\text{wfed}) + \beta_8 \ln(\text{density}) + \beta_9 \ln(\text{pctmin80}) + \beta_{10} \ln(\text{wcon}) \\ + \beta_{11} \ln(\text{wtuc}) + \beta_{12} \ln(\text{wtrd}) + \beta_{13} \ln(\text{wfir}) + \beta_{14} \ln(\text{wser}) + \beta_{15} \ln(\text{wmfg}) \\ + \beta_{16} \ln(\text{wsta}) + \beta_{17} \ln(\text{wloc}) + \beta_{18} \ln(\text{avgse}) + \beta_{19} \ln(\text{west}) + \beta_{20} \ln(\text{pctymle}) + \beta_{21} \ln(\text{mix}) \\ + u$$

```
In [38]: 1 m_all <- lm(log(crmrte) ~ log(polpc) + log(prbarr) + log(prbconv)
2          + log(prbpris) + avgse + log(taxpc)
3          + factor(west) + factor(urban)
4          + log(wfed) + log(wcon) + log(wtuc)
5          + log(wtrd) + log(wfir) + log(wser)
6          + log(wmfg) + log(wsta) + log(wloc)
7          + log(density) + log(pctmin80) + lo
8          + log(mix), data = crime.narm)
```

### Robust standard errors

Covariance matrix was applied for estimating a robust covariance matrix of variables. These are used because, as documented below, we violate the assumption of homoscedasticity.

```
In [39]: 1 # Compute robust standard errors
2 se.m_key = sqrt(diag(vcovHC(m_key)))
3 se.m_associated = sqrt(diag(vcovHC(m_associated)))
4 se.m_all = sqrt(diag(vcovHC(m_all)))
```

### Akaike Information Criteria (AIC)

This is used, in addition to the adjusted- $R^2$ , to assess model fit.

```
In [40]: 1 # Compute AIC values
2 m_key$AIC <- AIC(m_key)
3 m_associated$AIC <- AIC(m_associated)
4 m_all$AIC <- AIC(m_all)
```

**Table 3: Regression table. The coefficients, robust standard errors,  $p$ -values, adjusted- $R^2$  and AIC values are presented and discussed below.**

Coefficients and standard errors are interpreted below.

```
In [41]: 1 # We pass the standard errors into stargazer through the se argument.
2 stargazer(m_key, m_associated, m_all,
3           type = "text",
4           keep.stat=c("n", "adj.rsq", "aic"),
5           se = list(se.m_key, se.m_associated, se.m_all),
6           column.labels = c("key", "associated", "all"),
7           star.cutoffs=c(0.05, 0.01, 0.001)
8           )
```

=====			
	Dependent variable:		
	-----		
	log(crmrte)		
	key	associated	all
	(1)	(2)	(3)
-----			
log(polpc)	0.296 (0.403)	0.317 (0.307)	0.507* (0.235)
log(prbarr)	-0.542 (0.678)	-0.399 (0.391)	-0.512*** (0.136)
log(prbconv)	-0.032 (1.021)	-0.153 (0.576)	-0.293* (0.137)
log(prbpris)		-0.192 (0.231)	-0.315 (0.217)
avgsen			-0.035 (0.019)
log(taxpc)		0.048 (0.286)	0.049 (0.287)
factor(west)1			0.125 (0.121)
log(wfed)		0.679 (0.566)	0.555 (0.549)
log(wcon)			0.199 (0.224)
log(wtuc)			0.102 (0.359)
log(wtrd)			0.362 (0.348)
log(wfir)			-0.365 (0.417)
log(wser)			-0.337* (0.167)
log(wmfg)			0.031

			(0.225)
log(wsta)		0.007	(0.371)
log(wloc)		0.047	(0.694)
factor(urban)1	0.011	0.058	
	(0.250)	(0.240)	
log(density)	0.098	0.138	
	(0.179)	(0.151)	
log(pctmin80)	0.247***	0.273***	
	(0.052)	(0.060)	
log(prbarr):log(prbconv)	0.326	0.215	
	(0.633)	(0.355)	
log(pctymle)		0.156	
		(0.261)	
log(mix)		0.051	
		(0.106)	
Constant	-2.665	-7.525	-5.155
	(3.535)	(5.425)	(4.746)

Observations	90	90	90
Adjusted R2	0.474	0.772	0.790
Akaike Inf. Crit.	96.385	26.409	27.767

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

### Notes about model fit are in the appendix

#### Interpretation of coefficients of Model 1 (m\_key model):

- $\beta_1 \ln(polpc)$ : For each 1% increase in *polpc*, crime rate (*crmte*) increases by 0.296%. This coefficient is statistically insignificant and the standard error is large relative to the estimate. The *p*-value associated with  $H_0: \beta_1 = 0$  is  $p > 0.05$ - the slope is not different from 0. From a practical perspective, this may be significant - any increase in crime rates is significant from the victims perspective. Moreover, since the direction of causality is unknown, it seems likely that we could recommend more police per capita in areas with relatively high crime rates and reduce, not increase, crime rates. In other words, areas with more crime might need more police to maintain the safety.
- $\beta_2 \ln(prbarr)$  and  $\beta_3 \ln(prbconv)$ : For each 1% increase in *prbarr*, *crmte* decreases by 0.542%. For each 1% increase in *prbconv*, *crmte* decreases by 0.032%. *prbarr* and *prbconv* failed to satisfy the axioms of probability theory as described above ("Justification for the inclusion of *prbarr* and *prconv* values > 1"). These coefficients are not statistically significant. The *p*-values associated with  $H_0: \beta_2 = 0$  and  $\beta_3$  is  $p > 0.05$ - the slopes are not different from 0. From a practical perspective, these coefficient represent a small but perhaps important

decrease in crime rates/person. Perhaps people are less likely to commit a crime if they are more likely to be arrested or more afraid of being arrested. In addition, people may be scared to be convicted after they get arrested. Perhaps stauncher policing, arrest and conviction criteria may reduce crime slowly.

- $\beta_4 \ln(prbarr) \cdot \ln(prbconv)$ : There is not an interaction between  $prbarr$  and  $prbconv$ . The  $p$ -value associated with  $H_0: \beta_4 = 0$  is  $p > 0.05$ . In other words, the effect of  $\ln(prbarr)$  on  $\ln(crmrte)$  does not depend on the values of  $\ln(prbarr)$ . From a practical perspective this means that politicians need not consider the interaction of  $prbarr$  and  $prbconv$  when implementing policies to reduce crime- if it sufficient to  $prbarr$  and  $prbconv$  separately.
- Finally, the intercept,  $\beta_0$ , is -2.665 and not significant. The logarithm of  $crmte$  is -2.665 when  $polpc$ ,  $prbarr$ ,  $prbconv$ , and the interaction between  $prbarr$  and  $prbconv$  are 1 and  $u$  is 0. In other words, when transformed to the original units,  $e^{-2.665} = 0.069$ , crime rate per person is low, 0.069, even when there is one police officer per person, and  $prbarr$  and  $prbconv$  are high, at 1.

### Interpretation of coefficients of Model 2 (m\_associated model):

- $\beta_1 \ln(polpc)$ ,  $\beta_2 \ln(prbarr)$ ,  $\beta_3 \ln(prbconv)$  The effects and directions of these key explanatory variables and the interaction between  $prbarr$  and  $prbconv$  are similar to m\_key model as described above. For each 1% increase in  $polpc$ ,  $crmte$  increases by 0.317%. For each 1% increase in  $prbarr$ ,  $crmte$  decreases by 0.399%. For each 1% increase in  $prbconv$ , the  $crmte$  decreases by 0.153%. These coefficients of these key explanatory variables are not statistically and practically significant. The  $p$ -value associated with  $H_0: \beta_1 = 0$ ,  $\beta_2$  or  $\beta_3$  is  $p > 0.05$ - the slopes are not different from 0.
- $\beta_4 \ln(prbarr) \cdot \ln(prbconv)$ : As above, there is not an interaction between  $prbarr$  and  $prbconv$ . The  $p$ -value associated with  $H_0: \beta_4 = 0$  is  $p > 0.05$ . In other words, the effect of  $prbarr$  on  $crmte$  does not depend on the values of  $prbarr$ . Finally, the standard error for this interaction is less than for m\_key. From a practical perspective this means that politicians need not consider the interaction of  $prbarr$  and  $prbconv$  when implementing policies to reduce crime- if it sufficient to  $prbarr$  and  $prbconv$  separately.
- $\beta_5 \ln(prbpris)$  : For every 1% increase in  $prbpris$ , there is 0.192% decrease in  $crmte$ . While this coefficient may of limited practical significance in that prospect of prison time may lead to few crimes being the committed, in this model, this coefficient is statistically insignificant ( $p > 0.05$ ).
- $\beta_6 urban$  : Controlling for all other variables, on average, the crime rate in urban areas is 1.1% more compared to non urban area. This coefficient of this covariate is neither statistically nor practically significant. The  $p$ -value associated with  $H_0: \beta_6 = 0$  is  $p > 0.05$ . From a practical perspective, this coefficient may imply that the urban areas may have higher crime rates, although  $n$  is small and the effect is weak.
- $\beta_7 taxpc$  : For every 1% increase in  $taxpc$ ,  $crmte$  increases by 0.048%. This coefficient of this covariate is neither statistically nor practically significant. The  $p$ -value associated with  $H_0: \beta_7 = 0$  is  $p > 0.05$ - the slope is not different from 0. From a practical perspective, the county with high tax revenue may imply more rich people and/or bussiness activities which may, in turn, be associated with increased crime rates. However, the effect is week because the coefficient is small and statistically insignificant.
- $\beta_8 \ln(wfed)$  : For every 1% increase in  $wfed$ ,  $crmte$  increases 0.679%. This coefficient of this covariate is neither statistically nor practically significant. The  $p$ -value associated with  $H_0: \beta_8 = 0$  is  $p > 0.05$ - the slope is not different from 0. When converted back to the original

units, this coefficient represents a 1.97 unit increase in crime rates (crimes committed/person) for every unit increase in weekly wages for federal employees.

- $\beta_9 \ln(density)$  : For every 1% increase *density*, *crmte* increases by 0.098%. This coefficient of this covariate is neither statistically nor practically significant. The  $p$ -value associated with  $H_0: \beta_9 = 0$  is  $p > 0.05$ - the slope is not different from 0. From a practical perspective, this coefficient suggests that counties with more people have slightly hire crime rates. If it was ethical, we could relocate people and or encourage people to move to more rural, sparsely populated areas.
- $\beta_{10} \ln(pctmin80)$  : For every 1% increase in *pctmin80*, *crmte* increases by 0.247%. This coefficient of this covariate is statistically and practically significant. The  $p$ -value associated with  $H_0: \beta_{10} = 0$  is  $p < 0.001$ - the slope is different from 0. From a practical perspective, this coefficient seems to represent an important driver of crime rates/person.
- Finally, the intercept,  $\beta_0$ , is -7.527 and not significant. The logarithm of *crmte* is -7.527 when the covariates are 1 and  $u$  is 0. In other words, when transformed to the original units,  $e^{-7.527} = 0.0005$ , crime rate per person is low, 0.0005, even when covariates are fixed at 1.

### Interpretation of coefficients of Model 3 (m\_all model):

- $\beta_1 \ln(polpc)$ : For every 1% in increase in *polpc*, *crmte* increases by 0.507%. This is counter intuitive but can be indicative of the fact that when crime increases, more police resources per capita may be deployed and not that more police per capita causes a higher crime rate. The null hypothesis that  $H_0: \beta_1 = 0$  can be rejected, though it should be noted that the associated  $p$  value only satisfies ( $p < 0.05$ ).
- $\beta_2 \ln(prbarr)$  : For every 1% increase in *prbarr*, we see that *crmte* reduces by 0.512%. This coefficient is statically significant and thus reject the null hypothesis  $H_0: \beta_2 = 0$  since the associated  $p$  value  $p < 0.001$ . This decrease is also practically significant since there is the potential to impact the crime in an area in an actionable and positive manner. One reason for this negative relation between the independent and explanatory variable may be that once a person is arrested, their opportunity to commit crime is reduced.
- $\beta_3 \ln(prbconv)$  : For every 1% increase in *prbarr*, there is 0.293% decrease in  $\ln(crmte)$ . This coefficient is statistically significant and thus reject the null hypothesis  $H_0: \beta_3 = 0$  since the associated  $p$  value  $p < 0.05$ . This impact of this coefficient may be small but important practically significant in that perhaps people are less likely to commit a crime if they face the prospect of conviction.
- $\beta_4 \ln(prbpris)$  : For every 1% increase in  $\ln(prbpris)$ , there is 0.315% decrease in  $\ln(crmte)$ . While this coefficient may of limited practical significance in that prospect of prison time may lead to few crimes being the committed, in this model, this coefficient is statistically insignificant ( $p > 0.05$ ).
- $\beta_5 urban$  : Controlling for all other variables, on average, *crmte* in urban areas is 5.8% more compared to non urban area. This coefficient of this covariate is statistically insignificant. The  $p$ -value associated with  $H_0: \beta_5 = 0$  is  $p > 0.05$ . From a practical perspective, this coefficient may imply that the urban areas may have higher crime rates.
- $\beta_6 \ln(taxpc)$  : For every 1% increase in  $\ln(taxpc)$ , the  $\ln(crmte)$  increases by 0.049%. This coefficient of this covariate is neither statistically nor practically significant. The  $p$ -value associated with  $H_0: \beta_6 = 0$  is  $p > 0.05$ - the slope is not different from 0. From a practical perspective, the county with high tax revenue may imply more rich people and/or business activities; furthermore, potentially leading the increase of crime rate. However, the effect is weak because the coefficient is small and statistically insignificant.



- Wage coefficients:  $\beta_7 \ln(wfed)$ ,  $\beta_1 0 \ln(wcon)$ ,  $\beta_{11} \ln(wtuc)$ ,  $\beta_{12} \ln(wtrd)$ ,  $\beta_{15} \ln(wmfg)$ ,  $\beta_{17} \ln(wloc)$ ,  $\beta_{13} \ln(wfir)$ ,  $\beta_{14} \ln(wser)$ , and  $\beta_{16} \ln(wsta)$  : We see here that when all types of wages are included, any 10% increase in  $wfed$ ,  $wcon$ ,  $wtuc$ ,  $wtrd$ ,  $wmfg$ ,  $wloc$ , or  $wsta$  is associated with no more than a 6% increase in  $\ln(crmrte)$ , with the impact of federal wage being the highest at 5.6% and manufacturing wages having the lowest impact at 0.07% increase in crime rate. In many cases, the impact to crime rate is practically negligible, e.g in the case of state employees' wages. Any 10% increase in  $wfir$  is associated with a 3.65% decrease in crime rate but again the impact is practically insignificant. On the other hand, Any 10% increase in  $wser$  is associated with a 3.37% decrease in crime rate and the impact is statistically ( $p < 0.05$ ) and practically significant. Moreover, all the coefficients, except  $wser$ , related to wage are statistically insignificant (each associated  $p$ ,  $p > 0.05$ ) and we fail to reject the null hypothesis for each of these variables  $H_0: \beta_j = 0$  for  $j = wfed, wcon, wtuc, wtrd, wmfg, wloc, wfir, wser$  or  $wsta$ .
- We recognize that some of the interpretations for wage variables are counter intuitive e.g. higher wages in some categories are associated with higher crime rate whereas higher wages in other categories are associated with lower crime rates. As mentioned previously, the wage variables are most likely top and bottom coded and the extreme values have been censored. Combined with the various sources of omitted variable bias discussed in later sections, we believe that conclusions based on wages may be inaccurate.
- $\beta_8 \ln(density)$  : For every 1% increase in  $\ln(density)$ ,  $\ln(crmrte)$  increases by 0.138% which is practically insignificant. This is also statistically insignificant ( $p > 0.05$ ) and we fail to reject the null hypothesis that the slope is 0 i.e.  $H_0: \beta_{density} = 0$ .
- $\beta_9 \ln(pctmin80)$  : For every 1% increase in  $\ln(pctmin80)$ , the  $\ln(crmrte)$  increases by 0.273%. This coefficient of this covariate is statistically and practically significant. The  $p$ -value associated with  $H_0: \beta_9 = 0$  is  $p < 0.001$ - the slope is different from 0. From a practical perspective, this coefficient seems to represent an important driver of crime rates/person.
- $\beta_{18} avg\ sen$  : For 1 day increase in  $avg\ sen$ , there appears to be a 3.5% decrease in  $\ln(crmrte)$ . While this coefficient may be of practical significance in that perhaps longer the a prison sentences provide less opportunities to commit crime or even that the experience of prison deters recidivism, in this model  $\beta_{18} avg\ sen$  coefficient is statistically insignificant ( $p > 0.05$ ) and we fail to reject the null hypothesis that the slope is 0 i.e.  $H_0: \beta_{avg\ sen} = 0$ .
- $\beta_{19} \ln(west)$  : Controlling for all other variables, on average, the  $crmrte$  in the Western NC areas is 12.5% more compared to areas not categorized as west. This coefficient of this covariate is statistically insignificant. The  $p$ -value associated with  $H_0: \beta_{19} = 0$  is  $p > 0.05$ . From a practical perspective, this coefficient may imply that that western NC has a much higher crime rate than those areas not categorized as western NC, although none of the EDA supports this conclusion.
- $\beta_{20} \ln(pctmle)$  : For every 1% increase in  $\ln(pctmle)$ ,  $\ln(crmrte)$  increases by 0.156%. This coefficient of this covariate is statistically and practically insignificant. The  $p$ -value associated with  $H_0: \beta_{20} = 0$  is  $p > 0.05$ - the slope is different from 0.
- $\beta_{21} \ln(mix)$  : For every 1% increase in  $\ln(mix)$ ,  $\ln(crmrte)$  increases by 0.051%. This coefficient of this covariate is both statistically and practically insignificant. The  $p$ -value associated with  $H_0: \beta_{21} = 0$  is  $p > 0.05$ - the slope is different from 0.
- Finally, the intercept,  $\beta_0$ , is -5.5155 and not significant. The logarithm of  $crmrte$  is -5.5155 when the covariates are 1 and  $u$  is 0. In other words, when transformed to the original units,  $e^{-5.5155} = 0.004$ , crime rate per person is low, 0.004, even when covariates are fixed at 1.

# Diagnostics

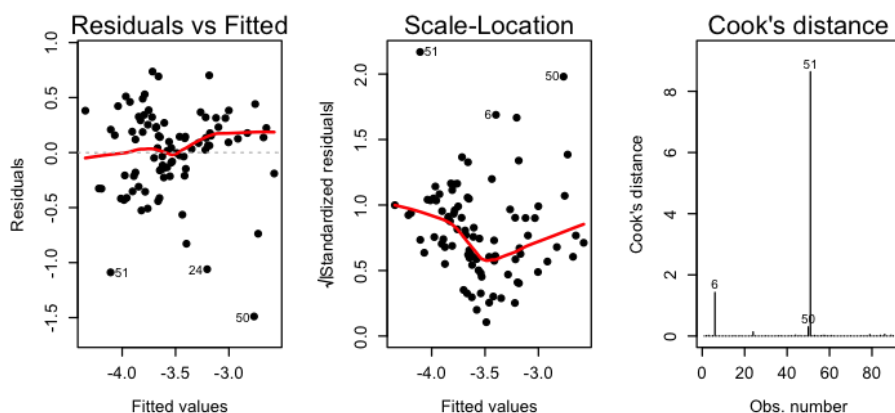
[Back to Table of Contents](#)

**Figure 2a: Diagnostic plots for model with only key explanatory variables (m\_key).**

Note that these plots will be discussed below.

In [42]:

```
1 options(repr.plot.height = 3, repr.plot.width = 8, repr.plot.pointsiz
2 par(mfrow = c(1,4))
3 plot(m_key, which=1, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
4 plot(m_key, which=3, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
5 plot(m_key, which=4, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=1)
```

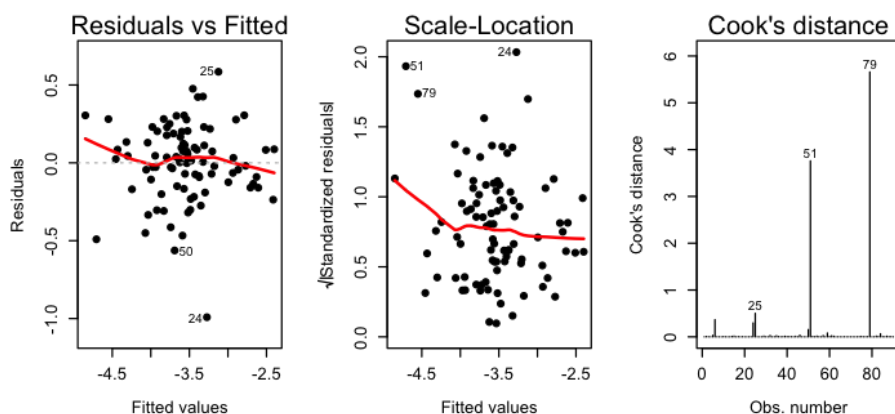


**Figure 2b: Diagnostic plots for model with key variables + covariates of interest (m\_associated).**

Note that these plots will be discussed below.

In [43]:

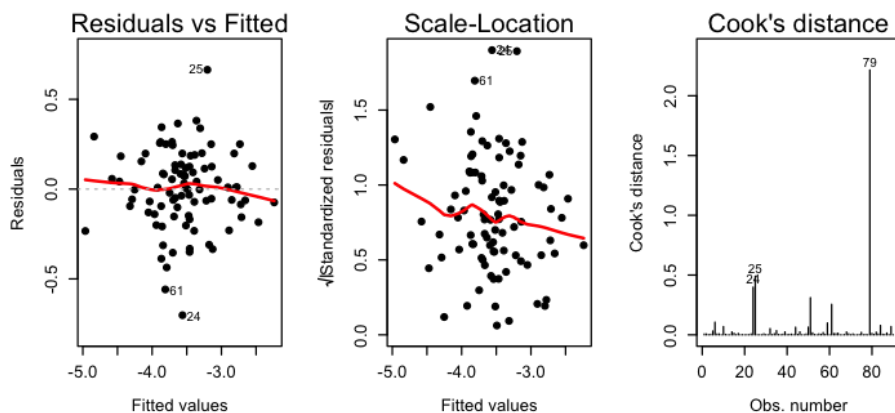
```
1 options(repr.plot.height = 3, repr.plot.width = 8, repr.plot.pointsiz
2 par(mfrow = c(1,4))
3 plot(m_associated, which=1, pch=19, cex=1, cex.lab=1, cex.main=4, lwd
4 plot(m_associated, which=3, pch=19, cex=1, cex.lab=1, cex.main=4, lwd
5 plot(m_associated, which=4, pch=19, cex=1, cex.lab=1, cex.main=4, lwd
```



**Figure 2c: Diagnostic plots for model with all explanatory variables (m\_all).**

Note that these plots will be discussed below.

```
In [44]: 1 options(repr.plot.height = 3, repr.plot.width = 8, repr.plot.pointsiz
2 par(mfrow = c(1,4))
3 plot(m_all, which=1, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
4 plot(m_all, which=3, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
5 plot(m_all, which=4, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=1)
```



### Summary of diagnostic plots and assumptions for all three models: m\_all, m\_associated, and m\_key

#### 1. Linear in parameters

- As seen in our models above, all of our models are linear and additive in the parameters, the  $\beta_0, \beta_1, \dots, \beta_k$

#### 2. Random sampling

- We assume that are  $n$  observations,  $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, 2, \dots, n\}$  are independent and identically distributed, *iid*. It is possible that this assumption is violated but it is difficult to know with certainty from the information provided.

#### 3. No perfect collinearity

- As documented in the scatterplot matrices below (**Figs A3-A6**) and those in the appendix, there is no *perfect* collinearity. Although, there are certainly explanatory variables that are correlated, these variables are not *perfectly* correlated.  $\therefore$  the explanatory variables are not redundant.

#### 4. Zero-conditional mean

- The zero-conditional mean assumption,  $E(u|x_1, x_2, \dots, x_k) = 0$  is not satisfied, *sensu stricto*, for any of the models presented above. The residual versus fitted values plots (**Figs 2a-c**) document the extent to which these models violate this assumption. If the zero mean assumption were satisfied then we should expect to see a horizontal red line, centered at zero (ASYN lecture 12.5). In contrast we see that our estimates of  $u$ , the residuals, deviate from this expectation, especially for m\_key and m\_associated.

#### 5. Homoskedasticity

- Homoskedasticity does not appear to be satisfied for all models presented above. Evidence for heteroskedasticity is provided by plots and Breusch-pagan tests, below. When **figures 2a-c** are inspected, we see minor to acute violations of this assumption in the residual versus fitted value plots (**Figs 2a-c**) and the scale-location plots. If variances were homoskedastic we should see uniform horizontal bands/scatters of points across these plots. However, note that we see non-constant variability in the residuals and standardized residuals across the range of fitted values. Likewise, Breusch-pagan tests for each model rejected the  $H_0$ : residuals are homoscedastic. Since we did not satisfy this assumption, robust standard errors were used.

### Note about outliers and their influence

- From **figures 2a-c** above we can see values with large residuals, large standardized residuals, and two observations, 51 and 79, with large Cook's distance (as noted in ASYNC lecture 11.10) that is greater than 1. Given that we have no evidence that these observations were recorded incorrectly, we will keep these observations in the dataframe, despite their influence on the regression surface.

### Breusch Pagan test for homoscedasticity

$H_0$ : residuals are homoscedastic

Squared residuals are regressed on explanatory variables

```
In [45]: 1 # m_key
2 bp_key = lm(m_key$residuals**2 ~ log(polpc) + log(prbarr) + log(prbco
3           + log(prbconv)*log(prbarr), data = crime.narm)
4 # m_associated
5 bp_assoc = lm(m_key$residuals**2 ~ log(polpc) + log(prbarr) + log(prb
6           + log(prbarr)*log(prbconv) +
7           + log(prbpris) + avgsgen +
8           + log(taxpc) + log(wfed) +
9           + factor(west) + factor(urban) + log(d
10 # m_all
11 bp_all = lm(m_key$residuals**2 ~ log(polpc) + log(prbarr) + log(prbco
12           + log(prbpris) + avgsgen + log(taxpc
13           + factor(west) + factor(urban) + fa
14           + log(wfed) + log(wcon) + log(wtuc)
15           + log(wtrd) + log(wfir) + log(wser)
16           + log(wmfg) + log(wsta) + log(wloc)
17           + log(density) + log(pctmin80) + lo
18           + log(mix), data = crime.narm)
```

Test statistics

```
In [46]: 1 # m_key
2 bpts_k = nobs(bp_key) * summary(bp_key)$r.squared
3 paste("Breusch-Pagan test-statistic:", bpts_k)
4 # m_associated
5 bpts_as = nobs(bp_assoc) * summary(bp_assoc)$r.squared
6 paste("Breusch-Pagan test-statistic:", bpts_as)
7 # m_all
8 bpts_al = nobs(bp_all) * summary(bp_all)$r.squared
9 paste("Breusch-Pagan test-statistic:", bpts_al)
```

'Breusch-Pagan test-statistic: 11.0058263780772'

'Breusch-Pagan test-statistic: 24.2752619331665'

'Breusch-Pagan test-statistic: 37.4205416315331'

df

```
In [47]: 1 # m_key
2 bpdf_k = length(bp_key$coefficients) - 1
3 # m_associated
4 bpdf_as = length(bp_assoc$coefficients) - 1
5 # m_all
6 bpdf_al = length(bp_all$coefficients) - 1
```

*p*-value

```
In [48]: 1 # m_key
2 paste("p-value:",
3       1-pchisq(bpts_k, bpdf_k))
4 # m_associated
5 paste("p-value:",
6       1-pchisq(bpts_as, bpdf_as))
7 # m_all
8 paste("p-value:",
9       1-pchisq(bpts_al, bpdf_al))
10
```

'p-value: 0.0264986118703746'

'p-value: 0.0116213319034999'

'p-value: 0.0150557449473256'

As noted above, since the assumption of homoscedasticity was not satisfied, robust standard errors were used (as recommended in ASYNC, unit 12).

## Robustness

[Back to Table of Contents](#)

## Omitted Variable Bias:

We recognize that crime is a complex issue, depending on a wide variety of factors. As we can see, in even the most aggressive model (m\_all), the Adjusted  $R^2 = 0.790$  indicates that nearly 80% of the variation in crime per person can be explained by the explanatory variables we have chosen for the model. Even in this all-inclusive model, 20% of the variation in crime rate remains unexplained due to variables we either cannot measure or have access to. Here we will discuss a few such possible sources of omitted variable bias,

### Poverty

In and of itself, higher poverty may not necessarily cause an increase in crime rates. However, poverty and the lack of steady income, especially at levels needed to cover basic living needs may drive individuals towards crime to make ends meet or acquire material items they may not be able to access otherwise. While the dataset provides information on wage, it does not provide a benchmark of the poverty line, which would allow us to compare how the poverty varies from across counties. Moreover, impoverished communities tend to have less access to quality education and thus better career opportunities which can otherwise lead to more financial stability. Thus, we estimate that

- Poverty has a positive effect on crime and the coefficient  $\beta_{poverty} > 0$
- Poverty is negatively correlated to wages i.e. higher poverty leads to lower wages. Given that  $\beta_{poverty} > 0$ , the wage coefficients are negatively biased.
- Poverty is negatively correlated to taxes i.e. people in poverty will pay less in taxes. Given that  $\beta_{poverty} > 0$ ,  $\beta_{taxpc}$  is negatively biased.
- Since the dataset shows that most NC counties are categorized as non-urban, poverty may be considered to be positively correlated to population density. The reason why context is important is because we may have dense urban areas that are not generally fiscally constrained e.g. San Francisco or Manhattan. Given that  $\beta_{poverty} > 0$ ,  $\beta_{density}$  is positively biased.

### Education

Education can have impact on crime rates. Better education may lead to better and steadier employment which can in turn impact income. By measuring the impact of education on crime rates, policies can be implemented to ensure proper funding and resources are given to improve education in disadvantaged communities. We should point out that in certain cases, highly educated people will still commit common crimes. Thus, we estimate that

- Education has a negative effect on crime and the coefficient  $\beta_{education} < 0$
- Education is positively correlated to wages i.e. wages may increase with better education. Given that  $\beta_{education} < 0$ , the wage coefficients are negatively biased due to lack of information regarding education.
- Education is positively correlated to taxes since we assume that better education leads to better wages and in turn results in more taxes being paid. Given that  $\beta_{education} < 0$ ,  $\beta_{taxpc}$  is negatively biased.

### Employment

Another socioeconomic factor (along with of poverty and education), employment status can impact crime rates. We believe that steady employment and subsequent income generated can help reduce the tendency for a person to commit crime. Thus, we estimate that

- Employment has a negative effect on crime and the coefficient  $\beta_{employment} < 0$
- Not surprisingly, employment is positively correlated to wages. Given that  $\beta_{employment} < 0$ , the wage coefficients are negatively biased due to lack of information regarding employment.
- Employment is positively correlated to taxes. Given that  $\beta_{employment} < 0$ ,  $\beta_{taxpc}$  is negatively biased.

### Family cohesiveness

We believe there is merit to the idea that individuals from stable families, especially where both parents are active in the child's upbringing and character building may be less inclined to break the law. While ambiguous to quantify, sometimes young individuals may feel that they don't receive enough support at home and use delinquency as a way to get their family's attention. Thus, we estimate that

- Greater family cohesiveness has a negative effect on crime and the coefficient  $\beta_{fam\_cohesiveness} < 0$
- Family cohesiveness is negatively correlated to  $pctmle$  in that a closer, possibly even a loving family dynamic, may encourage young males to stay away from crime and focus their energy on positive activities instead. Given that  $\beta_{fam\_cohesiveness} < 0$ ,  $\beta_{pctmle}$  is positively biased.

**Gang influence** Another important factor to consider is presence and degree of influence of gangs, especially dense environments. Youth, especially young males, may join gangs for a variety of reasons (e.g. fulfilling basic essentials, the promise of protection, and simply peer pressure to fit in). Thus we estimate,

- Higher gang influence has a positive effect on crime and the coefficient  $\beta_{gang} > 0$
- Gang influence is positively correlated to population *density* in that a densely populated area is more susceptible to criminal activity by gang members. Given that  $\beta_{gang} > 0$ ,  $\beta_{density}$  is positively biased.
- Gangs are positively correlated to police activity i.e. more gang presence may lead to more police and more arrests, convictions and prison time. Given that  $\beta_{gang} > 0$ ,  $\beta_{polpc}$ ,  $\beta_{prbarr}$ ,  $\beta_{prbconv}$ ,  $\beta_{prbpri}$  are positively biased.
- Lastly, gangs tend to have more males than females. As mentioned above, young males are often recruited and as such, we estimate that gang influence is positively correlated to percent of younger males. Given that  $\beta_{gang} > 0$ ,  $\beta_{pctmle}$  is positively biased.

### Type of crime

The type of crime is an important factor that has not been captured. The *mix* variable does not provide enough granular information. If we have captured the type of crime in a few indicator variables (e.g. Use of gun, sexual assault, trespassing etc), public officials could focus attention to specific solutions. e.g. if assaults involving a deadly weapon are most common, this could lead to policy changes involving better gun control or if sexual assault is the most pressing issue, then specific community outreach and educational programs may be introduced. Given the wide range of possibilities, estimating the direction of the bias is non-trivial.

### Day of week / Time of Day

Knowing when and where crimes occur most frequently can help public safety officials plan effectively and better allocate resources. This missing variable can potentially bias the *polpc* explanatory variable though the direction of the bias is a bit unclear.

## Location

Better knowledge of where the felonies are occurring (indoors / outdoors / type of location / intersection) can also help public safety officials curb crime by focusing on problem areas(e.g. if the data shows that crime is occurring in abandoned buildings more so than street intersections, policies can be implemented to better deal vacant properties, perhaps even convert them into spaces that positively impact the community. Again, this variable is potentially correlated to the *polpc* variable though the direction of the bias is a bit unclear.

# Conclusions

[Back to Table of Contents](#)

## Summary

The relationships between law enforcement (*polpc*), certainty of punishment (*prbarr* and *prbconv*), and crime rates (*crmte*) are not unprecedented. Our models weakly corroborate this precedence and intuition. Although we can not claim that *polpc*, *prbarr*, and *prbconv* are the main determinants of crime, since the slopes were not different than 0, we can not negate their importance, since we failed to reject the hypothesis that they can be removed from the model individually. Interestingly, these explanatory variables do not jointly contribute to our *m\_key* and could have been removed from the model (see appendix). Sources of variation that might have contributed to these results are proposed below.

- These variables do not exist in a vacuum- we need to control for other covariates, for example those in *m\_associated*. When covariates are added in *m\_associated* and *m\_all* the coefficient estimates for *polpc*, *prbarr*, *prbconv* change and the slopes become significantly different from zero ( $p$ -value < 0.05) (**table 3**). In all cases the standard errors decrease in size when covariates are added, from *m\_key* to *m\_associated* to *m\_all*. Thus, these explanatory variables are not robust to alternative model specifications. As noted in the omitted variable bias discussion, granular knowledge of time and location of crime may enable politicians to deploy more police in areas with high crime.
- Outlying observations with influence. Observation 51, from county 115, likely pulled the regression surface away from the cloud of observations that defined the relationship between *crmte* and *polpc* (**figure 1b**).
- Omitted variables may have diminished or negated the effects of *polpc*, *prbarr*, and *prbconv* on *crmte*.
- The effects of *polpc* on crime are likely not independent of *prbarr* and *prbconv*, as suggested above and illustrated weakly in **figure A6**.
- The small sample size may have decreased our power.

## Actionable Recommendations

While it might be tempting for politicians to increase police presence and mandate more arrests as a quick and easy solution, it is misguided to think that simply increasing police activity will decrease crime. We advise the politicians to consider crime in the context of location, demographics, and socio-economic factors (see the omitted variable bias section). With that said, in the context of police activity, we advise the following to the political campaign of NC:



- increase police presence in areas of high crime, like densely populated urban areas (**figure 1b**). Omitted variables like time and specific locations could also be used to deploy resources optimally.
- provide training to officers to enable **informed** arrests, since 'probability' of arrests is negatively related to crime rates (**figure A6**). Arrests should be fair and not influenced by prejudice or an artificial quota that needs to be met.

Finally we think it might be valuable to increase community engagement between law enforcement and citizens. This will help erode mistrust and establish partnerships between citizens and the police that serve their neighborhoods. This would be cheap and simple and would not exhaust county budgets. In conclusion, while overhauling the law enforcement presence and practices may not be simple and cheap, certain steps can be taken to use resources more efficiently and reduce crime.

## Appendix

[Back to Table of Contents](#)

### Supplementary figures.

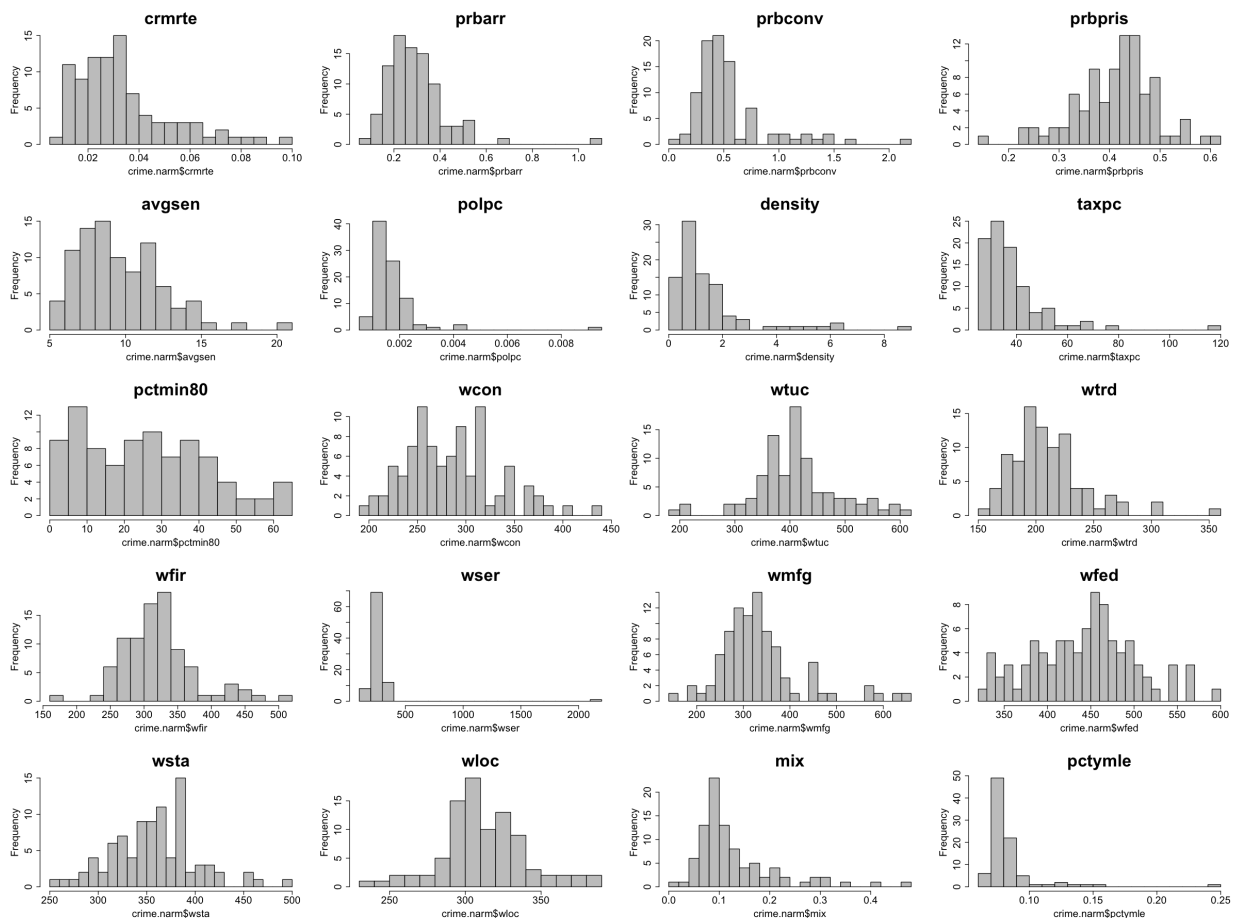
#### Figure A1: Distributions of variables

- Note that most variables are not normally distributed

```

In [49]: 1 options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
2 par(mfrow = c(5,4))
3 par(mar=c(6,5,4,2))
4 hist(crime.narm$crmrte,breaks=20, main = "crmrte",cex.lab=1.5,cex.mai
5 hist(crime.narm$prbarr,breaks=20, main = "prbarr",cex.lab=1.5,cex.mai
6 hist(crime.narm$prbconv,breaks=20, main = "prbconv",cex.lab=1.5,cex.m
7 hist(crime.narm$prbpris,breaks=20, main = "prbpris",cex.lab=1.5,cex.m
8 hist(crime.narm$avgsen,breaks=20, main = "avgsen",cex.lab=1.5,cex.mai
9 hist(crime.narm$polpc,breaks=20, main = "polpc",cex.lab=1.5,cex.main=
10 hist(crime.narm$density,breaks=20, main = "density",cex.lab=1.5,cex.m
11 hist(crime.narm$taxpc,breaks=20, main = "taxpc",cex.lab=1.5,cex.main=
12 hist(crime.narm$pctmin80,breaks=20, main = "pctmin80",cex.lab=1.5,cex
13 hist(crime.narm$wcon,breaks=20, main = "wcon",cex.lab=1.5,cex.main=2.
14 hist(crime.narm$wtuc,breaks=20, main = "wtuc",cex.lab=1.5,cex.main=2.
15 hist(crime.narm$wtrd,breaks=20, main = "wtrd",cex.lab=1.5,cex.main=2.
16 hist(crime.narm$wfir,breaks=20, main = "wfir",cex.lab=1.5,cex.main=2.
17 hist(crime.narm$wser,breaks=20, main = "wser",cex.lab=1.5,cex.main=2.
18 hist(crime.narm$wmfg,breaks=20, main = "wmfg",cex.lab=1.5,cex.main=2.
19 hist(crime.narm$wfed,breaks=20, main = "wfed",cex.lab=1.5,cex.main=2.
20 hist(crime.narm$wsta,breaks=20, main = "wsta",cex.lab=1.5,cex.main=2.
21 hist(crime.narm$wloc,breaks=20, main = "wloc",cex.lab=1.5,cex.main=2.
22 hist(crime.narm$mix,breaks=20, main = "mix",cex.lab=1.5,cex.main=2.5,
23 hist(crime.narm$pctymle,breaks=20, main = "pctymle",cex.lab=1.5,cex.m

```



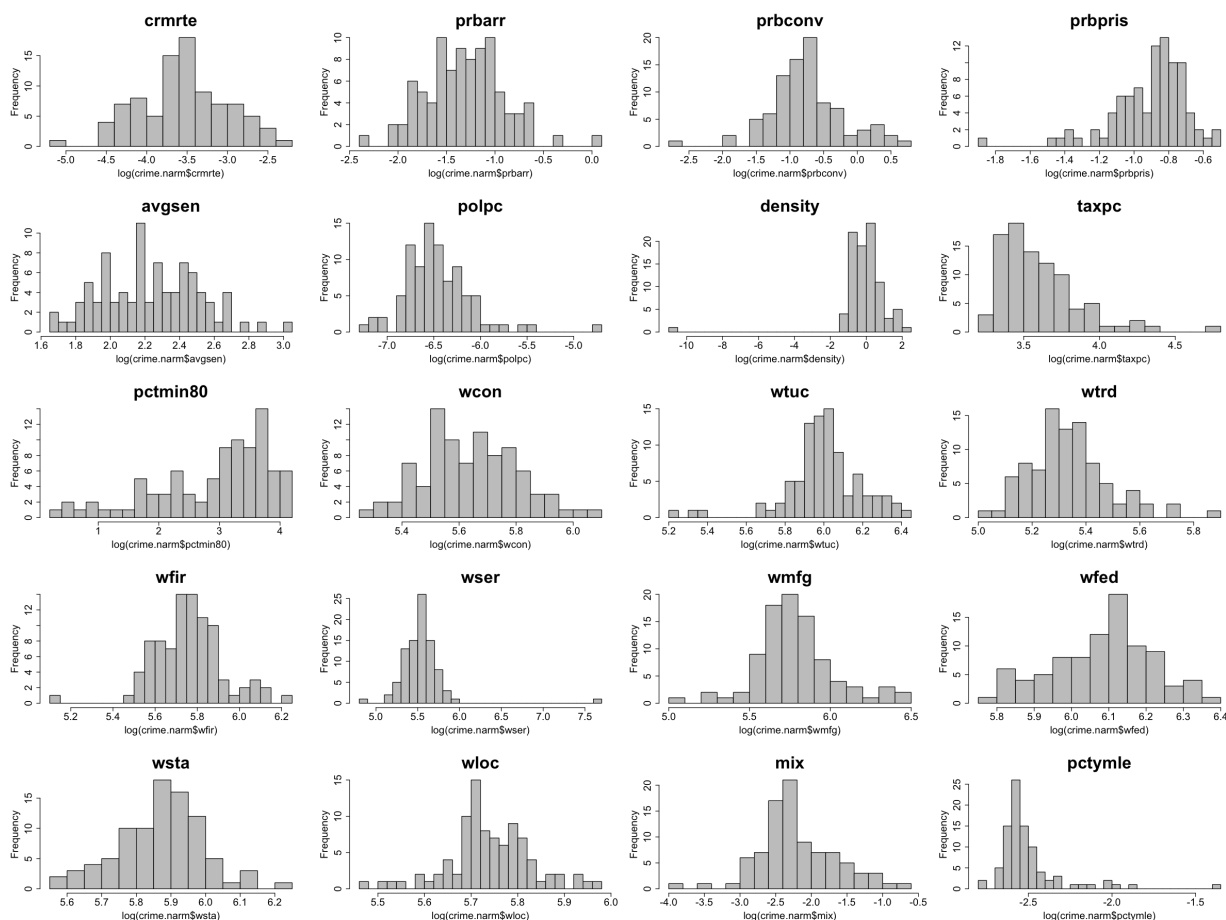
**Figure A2: Distributions of variables after log-transformation**

- Note that log-transformation resolved most instances of non-symmetric distributions

```

In [50]: 1 options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
2 par(mfrow = c(5,4))
3 par(mar=c(6,5,4,2))
4 hist(log(crime.narm$crmrte),breaks=20, main = "crmrte",cex.lab=1.5,cex.ma
5 hist(log(crime.narm$prbarr),breaks=20, main = "prbarr",cex.lab=1.5,cex.ma
6 hist(log(crime.narm$prbconv),breaks=20, main = "prbconv",cex.lab=1.5,cex.ma
7 hist(log(crime.narm$prbpris),breaks=20, main = "prbpris",cex.lab=1.5,cex.ma
8 hist(log(crime.narm$avgsen),breaks=20, main = "avgsen",cex.lab=1.5,cex.ma
9 hist(log(crime.narm$polpc),breaks=20, main = "polpc",cex.lab=1.5,cex.ma
10 hist(log(crime.narm$density),breaks=20, main = "density",cex.lab=1.5,cex.ma
11 hist(log(crime.narm$taxpc),breaks=20, main = "taxpc",cex.lab=1.5,cex.ma
12 hist(log(crime.narm$pctmin80),breaks=20, main = "pctmin80",cex.lab=1.5,cex.ma
13 hist(log(crime.narm$wcon),breaks=20, main = "wcon",cex.lab=1.5,cex.ma
14 hist(log(crime.narm$wtuc),breaks=20, main = "wtuc",cex.lab=1.5,cex.ma
15 hist(log(crime.narm$wtrd),breaks=20, main = "wtrd",cex.lab=1.5,cex.ma
16 hist(log(crime.narm$wfir),breaks=20, main = "wfir",cex.lab=1.5,cex.ma
17 hist(log(crime.narm$wser),breaks=20, main = "wser",cex.lab=1.5,cex.ma
18 hist(log(crime.narm$wmfg),breaks=20, main = "wmfg",cex.lab=1.5,cex.ma
19 hist(log(crime.narm$wfed),breaks=20, main = "wfed",cex.lab=1.5,cex.ma
20 hist(log(crime.narm$wsta),breaks=20, main = "wsta",cex.lab=1.5,cex.ma
21 hist(log(crime.narm$wloc),breaks=20, main = "wloc",cex.lab=1.5,cex.ma
22 hist(log(crime.narm$mix),breaks=20, main = "mix",cex.lab=1.5,cex.ma
23 hist(log(crime.narm$pctymle),breaks=20, main = "pctymle",cex.lab=1.5,cex.ma

```



### Correlations between dependent variable (*crm rte*) and independent variables

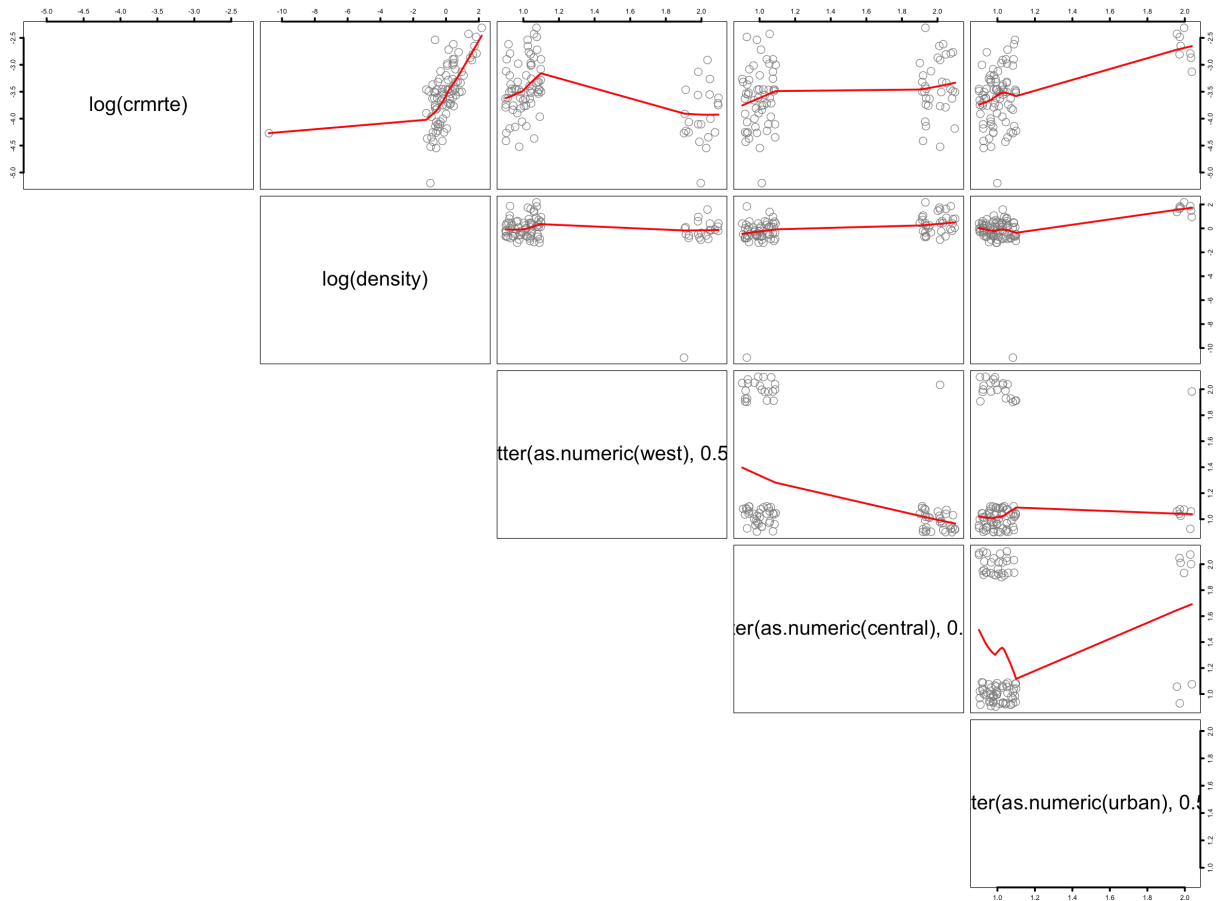
#### Figure A3: Crime and geography

- Note that (i) there is no perfect collinearity between the explanatory variables (ii)  $\ln(\text{density})$  is clearly related to  $\ln(\text{crm rte})$ .

```

In [51]: 1 options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
2 pairs(~log(crmrte)
3         + log(density)
4         + jitter(as.numeric(west),0.5)
5         + jitter(as.numeric(central),0.5)
6         + jitter(as.numeric(urban),0.5),
7         data = crime.narm,
8         cex.labels=3, lower.panel = NULL,
9         upper.panel=panel.smooth,
10        pch=1,cex=3,lwd=3,col="grey55")

```



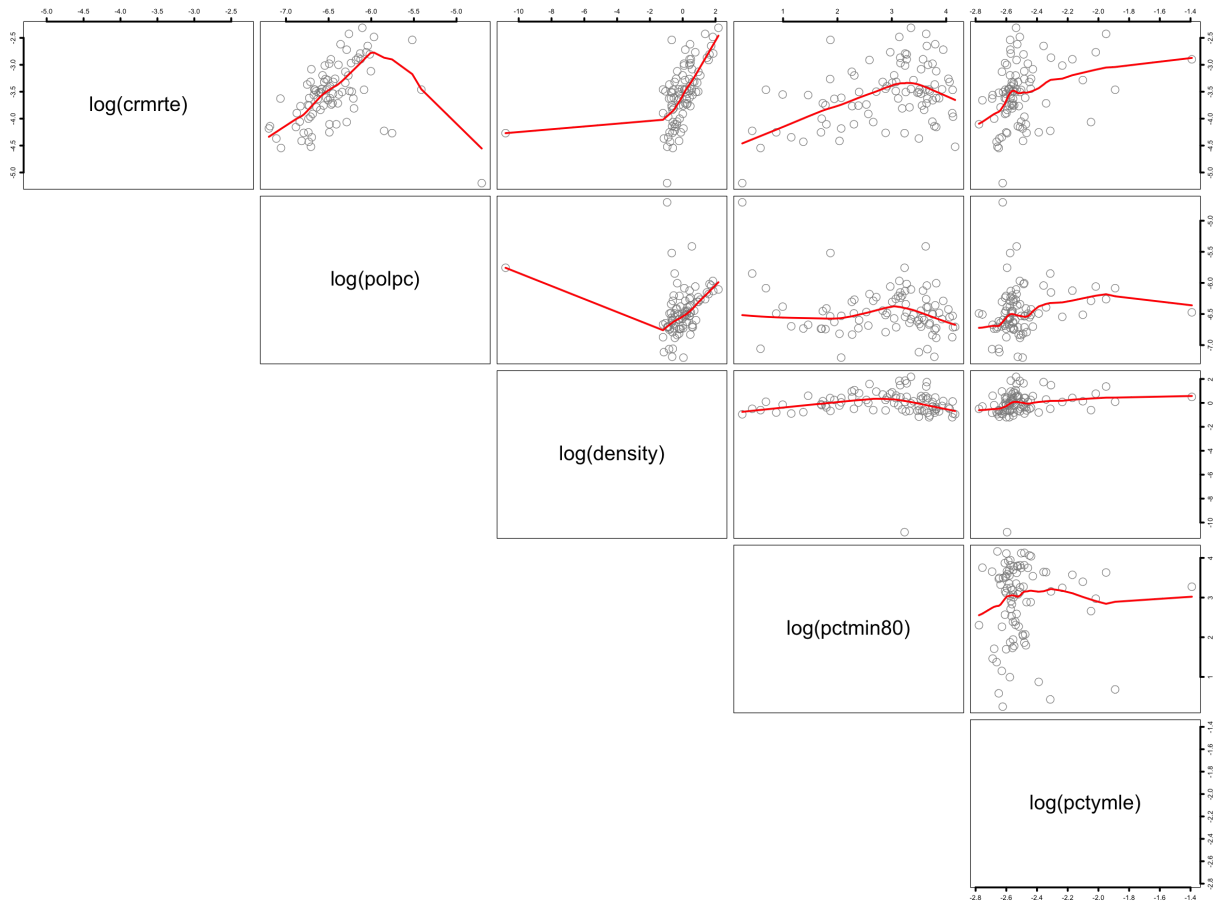
**Figure A4: Crime and demographics**

- Note that (i) there is no perfect collinearity between the explanatory variables (ii)  $\ln(polpc)$  is clearly related to  $\ln(crmrte)$ .

```

In [52]: 1 options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
2 pairs(~log(crmrte)
3         + log(polpc)
4         + log(density)
5         + log(pctmin80)
6         + log(pctymle),
7         data = crime.narm,
8         cex.labels=3, lower.panel = NULL,
9         upper.panel=panel.smooth,
10        pch=1, cex=3, lwd=3, col="grey55")

```



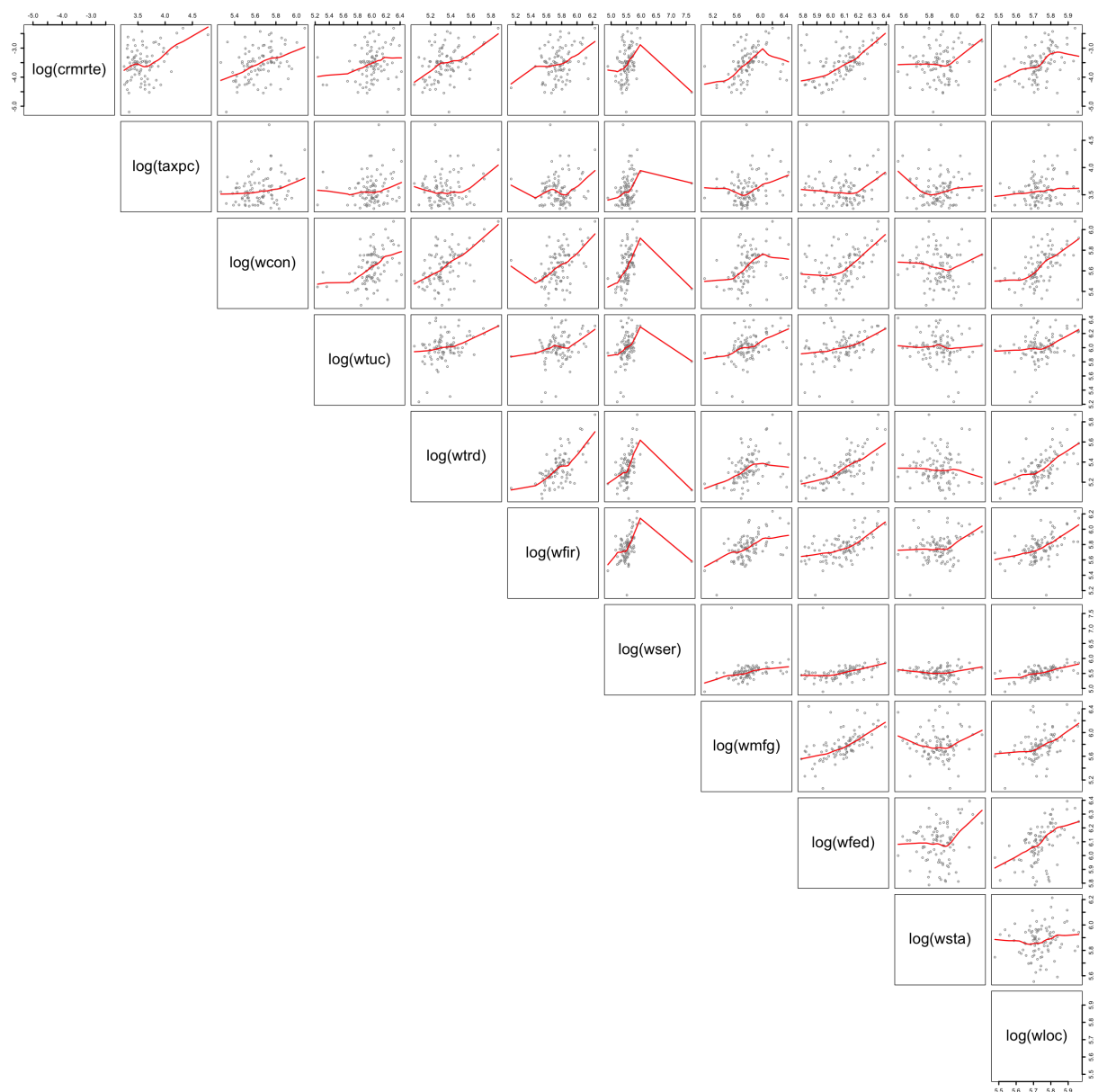
**Figure A5: Crime and economics**

- Note that (i) there is no perfect collinearity between the explanatory variables (ii)  $\ln(wfed)$ ,  $\ln(wcon)$ , and  $\ln(wmfg)$  are clearly related to  $\ln(crmrte)$ .

```

In [53]: 1 options(repr.plot.height = 20, repr.plot.width = 20, repr.plot.points
2 pairs(~log(crmrte)
3         + log(taxpc)
4         + log(wcon)
5         + log(wtuc)
6         + log(wtrd)
7         + log(wfir)
8         + log(wser)
9         + log(wmfg)
10        + log(wfed)
11        + log(wsta)
12        + log(wloc),
13        data = crime.narm,
14        cex.labels=2.5, lower.panel = NULL,
15        upper.panel=panel.smooth,
16        pch=1,cex=0.8,lwd=2,col="grey55")

```





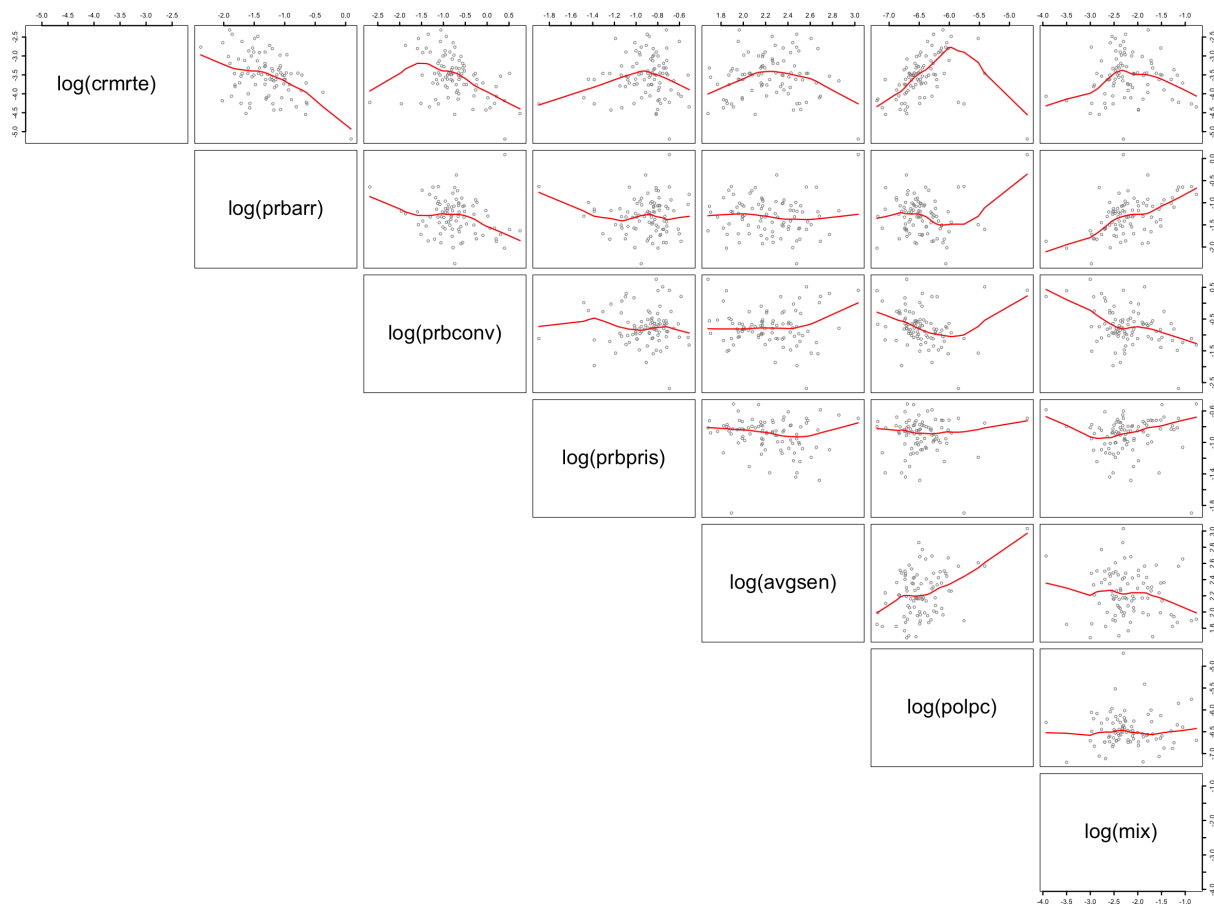
**Figure A6: Crime and propensity to commit crime, be arrested, and sentence lengths**

- Note that (i) there is no perfect collinearity between the explanatory variables (ii)  $\ln(prbarr)$  and  $\ln(prconv)$  are clearly related to  $\ln(crmrte)$ .

```

In [54]: 1 options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
2 pairs(~log(crmrte)
3         + log(prbarr)
4         + log(prbconv)
5         + log(prbpris)
6         + log(avgsen)
7         + log(polpc)
8         + log(mix),
9         data = crime.narm,
10        cex.labels=3, lower.panel = NULL,
11        upper.panel=panel.smooth,
12        pch=1,cex=1,lwd=2,col="grey55")

```



## Outliers

Given that these observations may be influential in future models, future investigation is in order.

Observation 79: note that the *density* is extremely low. There are  $\frac{0.00002 \text{ people}}{\text{sq mile}}$  in county 173!

```
In [55]: 1 subset(crime.narm, log(crime.narm$density) < -8)
```

	county	crmte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west
79	173	0.0139937	0.530435	0.327869	0.15	6.64	0.00316379	2.03422e-05	37.72702	1

Observation 51: note that *polpc* is very high. There are 0.009 police per capita in county 115.

```
In [56]: 1 subset(crime.narm, log(crime.narm$polpc) > -5)
```

	county	crmte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	..
51	115	0.0055332	1.09091	1.5	0.5	20.7	0.00905433	0.3858093	28.1931	1	..

## Notes about model fit

The model fit, indicated by  $R^2$  and AIC values, increased when all covariates were incorporated. The OLS model with covariates shows adjusted ( $R^2 = 0.772$ , AIC = 26.409) in the comparison to the model with only key exploratory variables ( $R^2 = 0.474$ , AIC = 96.385). In addition, even though the model with all variables provided the best prediction ( $R^2 = 0.790$ , AIC = 27.767), the risk of model overfitting and the cost (e.g. computing power, data collection, etc.) are increased at the same time.

The discrepancy between the coefficient of determination,  $R^2$ , for the nearly saturated model with all explanatory variables included ( $R^2 = 0.790$ ) and the model with only the variables that politicians can influence simply and cheaply ( $R^2 = 0.474$ ) is likely due to variety of sources and demands reconciliation. The full model, *m\_all*, includes all explanatory variables while the simple model, *m\_key*, includes 3 variables. It follows then that we accrue a  $0.790 - 0.474 = 0.316 * 100 = 31.6\%$  increase in our ability to explain crime rate with the addition on 19 explanatory variables. Given that we can explain approximately 47% of the variation in crime rate with only 3 explanatory variables, the addition of 19 more explanatory variables to absorb 31.6% more variation in crime rate seems potentially wasteful.

From a broader perspective, it's more reasonable to communicate policy recommendations to the campaign using a model comprised of 3 versus 22 explanatory variables. The media and public have limited cognitive bandwidth. To ask the politicians to reduce crime rate by juggling 3 variables will be difficult enough, to include 19 more is ludicrous and beyond the scope of our goal.

Finally, from an economical perspective, why use the resources to explain crime rate with all variables when a model, `m_associated`, with comparable performance can be built with 10 variables? More importantly, by using the 3 variables from `m_all`, we could not only reduce the crime rate but also save the government resources to improve welfare, public services, senior care etc. If this study is to be repeated or reproduced elsewhere, we can save those investigators time, money, and ease of interpretation by only including the most salient explanatory variables.

## Joint hypothesis test

$$H_0 : \beta_1 \ln(polpc) = \beta_2 \ln(prbarr) = \beta_3 \ln(prbconv) = 0 \Rightarrow E(y|x_1, x_2, \dots, x_n) = E(y)$$

$$H_A : \beta_1 \ln(polpc) = \beta_2 \ln(prbarr) = \beta_3 \ln(prbconv) \neq 0 \Rightarrow \exists \beta_j \neq 0$$

```
In [57]: 1 # Hypothesis - H_0: Test if slopes are equal to zero
2 H0s = c("log(polpc) = 0", "log(prbarr) = 0", "log(prbconv) = 0")
3 # Wald test
4 linearHypothesis(
5     m_key, H0s,
6     vcov=vcovHC(m_key),
7     singular.ok=TRUE
8 )
```

Res.Df	Df	F	Pr(>F)
88	NA	NA	NA
85	3	11.35223	2.462692e-06

Interestingly, we reject the null of the joint hypothesis described above. While we can not reject the hypotheses that these slopes are 0 alone, together they do not contribute towards describing  $y$ , *crmrte*.

## Identify the independent variables which show significant contribution to crime rate

Does population density, tax revenue, and police force affect crime rate?

- Population, government income

```
In [58]: 1 # the explanatory variables of key interest
2 mp1 <- lm(log(crmrte) ~ log(density), data = crime.narm)
3 mp2 <- lm(log(crmrte) ~ log(polpc), data = crime.narm)
4 mp3 <- lm(log(crmrte) ~ log(taxpc), data = crime.narm)
```

```
In [59]: 1 # Compute robust standard errors
2 se.mp1 = sqrt(diag(vcovHC(mp1)))
3 se.mp2 = sqrt(diag(vcovHC(mp2)))
4 se.mp3 = sqrt(diag(vcovHC(mp3)))
5 # We pass the standard errors into stargazer through the se argument.
6 stargazer(mp1, mp2, mp3,
7           type="text", keep.stat=c("n", "adj.rsq"),
8           se = list(se.mp1, se.mp2, se.mp3),
9           star.cutoffs=c(0.05, 0.01, 0.001)
10          )
```

```
=====
                        Dependent variable:
                        -----
                                log(crmrte)
                                (1)      (2)      (3)
                        -----
log(density)      0.197
                  (0.291)

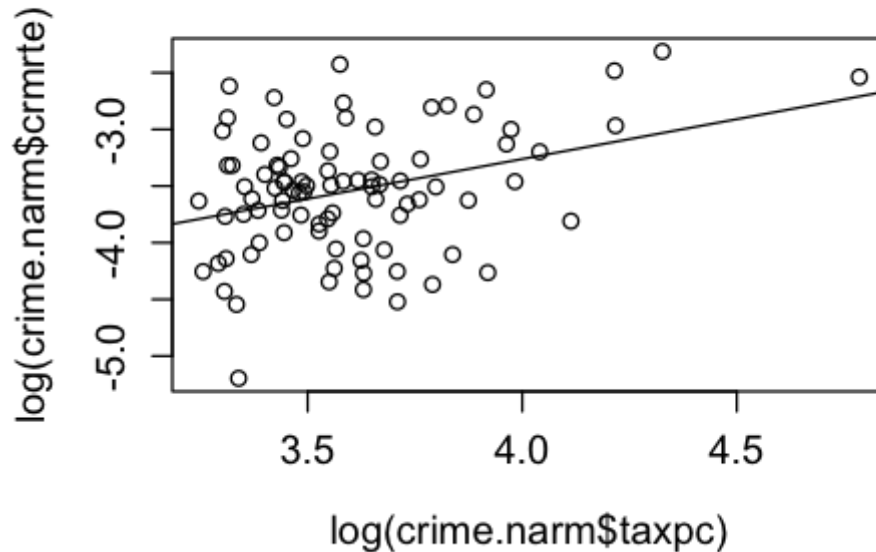
log(polpc)                  0.417
                           (0.475)

log(taxpc)                  0.704***
                           (0.207)

Constant      -3.525***    -0.847    -6.076***
               (0.055)     (3.104)   (0.750)

                        -----
Observations      90        90        90
Adjusted R2       0.235     0.071     0.105
=====
Note:             *p<0.05; **p<0.01; ***p<0.001
```

```
In [60]: 1 # check correlation between dependent variable and independent variab
2 options(repr.plot.height = 3, repr.plot.width = 4, repr.plot.pointsiz
3 plot(log(crime.narm$taxpc), log(crime.narm$crmrte))
4 abline(mp3)
```



In the category of population and government income, *taxpc* significantly affects crime rate with positive correlation.

### Does income affect crime rate?

- Employee incomes

```
In [61]: 1 me1 <- lm(log(crmrte) ~ log(wcon), data = crime.narm)
2 me2 <- lm(log(crmrte) ~ log(wtuc), data = crime.narm)
3 me3 <- lm(log(crmrte) ~ log(wtrd), data = crime.narm)
4 me4 <- lm(log(crmrte) ~ log(wfir), data = crime.narm)
5 me5 <- lm(log(crmrte) ~ log(wser), data = crime.narm)
6 me6 <- lm(log(crmrte) ~ log(wmfg), data = crime.narm)
7 me7 <- lm(log(crmrte) ~ log(wfed), data = crime.narm)
8 me8 <- lm(log(crmrte) ~ log(wsta), data = crime.narm)
9 me9 <- lm(log(crmrte) ~ log(wloc), data = crime.narm)
```

```

In [62]: 1 # Compute robust standard errors
2 se.me1 = sqrt(diag(vcovHC(me1)))
3 se.me2 = sqrt(diag(vcovHC(me2)))
4 se.me3 = sqrt(diag(vcovHC(me3)))
5 se.me4 = sqrt(diag(vcovHC(me4)))
6 se.me5 = sqrt(diag(vcovHC(me5)))
7 se.me6 = sqrt(diag(vcovHC(me6)))
8 se.me7 = sqrt(diag(vcovHC(me7)))
9 se.me8 = sqrt(diag(vcovHC(me8)))
10 se.me9 = sqrt(diag(vcovHC(me9)))
11 # We pass the standard errors into stargazer through the se argument.
12 stargazer(me1, me2, me3, me4, me5, me6, me7, me8, me9,
13           type="text", keep.stat=c("n", "adj.rsq"),
14           se = list(se.me1, se.me2, se.me3, se.me4, se.me5, se.me6, se.me7, se.me8, se.me9),
15           star.cutoffs=c(0.05, 0.01, 0.001)
16           )

```

```

=====
=====
Dependent variable:
-----
log(crmrte)
(1) (2) (3) (4) (5) (6)
(7) (8) (9)
-----
log(wcon) 1.322***
(0.396)
log(wtuc) 0.574
(0.334)
log(wtrd) 1.432***
(0.331)
log(wfir) 0.959**
(0.352)
log(wser) 0.093
(0.933)
log(wmfg) 0.800**
(0.305)
log(wfed) 2.010***
(0.473)
log(wsta) 0.686
(0.459)
log(wloc)

```

1.846\*

(0.907)

Constant	-10.999***	-6.986***	-11.191***	-9.068***	-4.060	-8.170***
	-15.768***	-7.573**	-14.136**			
	(2.237)	(1.995)	(1.770)	(2.020)	(5.142)	(1.747)
	(2.898)	(2.695)	(5.181)			

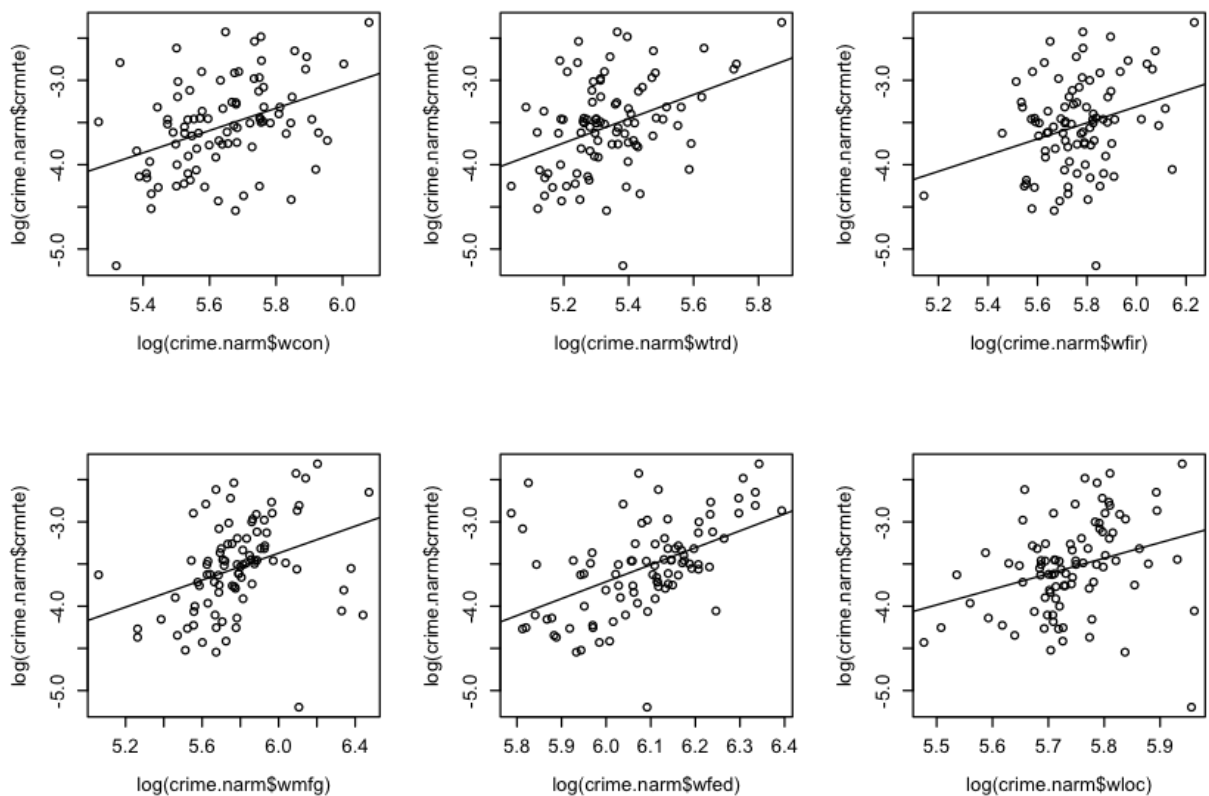
Observations	90	90	90	90	90	90
90	90	90				
Adjusted R2	0.147	0.034	0.142	0.073	-0.009	0.116
0.243	0.012	0.081				

Note:

\*p&lt;0.05; \*\*p&lt;0.01; \*\*\*p&lt;0.001



```
In [63]: 1 options(repr.plot.height = 5, repr.plot.width = 7, repr.plot.pointsiz
2 par(mfrow = c(2,3))
3 plot(log(crime.narm$wcon), log(crime.narm$crmrte))
4 abline(me1)
5 plot(log(crime.narm$wtrd), log(crime.narm$crmrte))
6 abline(me3)
7 plot(log(crime.narm$wfir), log(crime.narm$crmrte))
8 abline(me4)
9 plot(log(crime.narm$wmfg), log(crime.narm$crmrte))
10 abline(me6)
11 plot(log(crime.narm$wfed), log(crime.narm$crmrte))
12 abline(me7)
13 plot(log(crime.narm$wloc), log(crime.narm$crmrte))
14 abline(me9)
```



In the category of employee income, *wcon*, *wtrd*, *wfir*, *wmfg*, *wfed*, and *wloc* significantly affect crime rate with positive correlation.

**Do young males and the percent of population that is minority affect crime rate?**

- minority and young male

```
In [64]: 1 mm1 <- lm(log(crmrte) ~ pctmin80, data = crime.narm)
2 mm2 <- lm(log(crmrte) ~ pctymle, data = crime.narm)
```

```
In [65]: 1 # Compute robust standard errors
2 se.mm1 = sqrt(diag(vcovHC(mm1)))
3 se.mm2 = sqrt(diag(vcovHC(mm2)))
4 # We pass the standard errors into stargazer through the se argument.
5 stargazer(mm1, mm2,
6           type="text", keep.stat=c("n", "adj.rsq"),
7           se = list(se.mm1, se.mm2),
8           star.cutoffs=c(0.05, 0.01, 0.001)
9           )
```

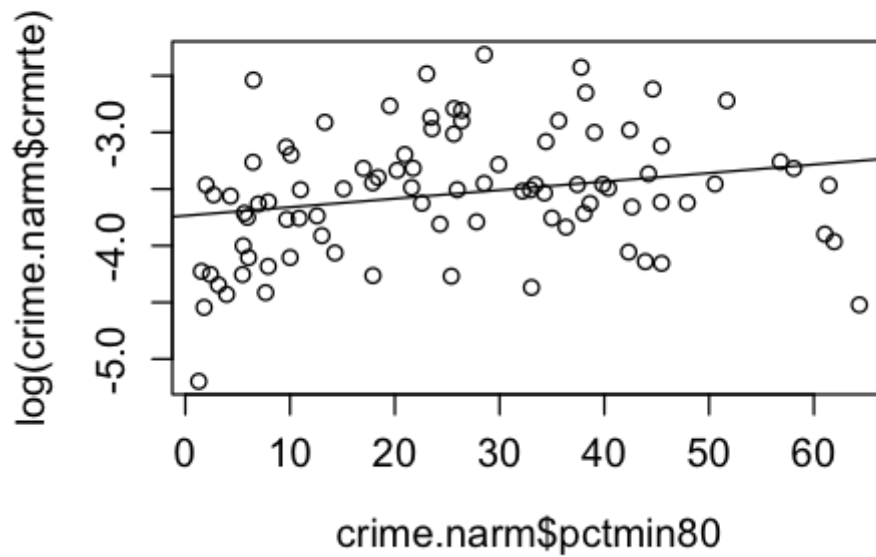
```
=====
Dependent variable:
-----
log(crmrte)
(1) (2)
-----
pctmin80      0.008
              (0.004)

pctymle                      6.509
                              (3.739)

Constant      -3.735***      -4.089***
              (0.115)        (0.311)

-----
Observations      90          90
Adjusted R2       0.044        0.067
=====
Note:      *p<0.05; **p<0.01; ***p<0.001
```

```
In [66]: 1 options(repr.plot.height = 3, repr.plot.width = 4, repr.plot.pointsiz  
2 plot(crime.narm$pctmin80, log(crime.narm$crmrte))  
3 abline(mml)
```



In the category of minority and young male, *pctmin80* significantly affects crime rate with positive correlation.

### Does geographic location affect crime rate?

- geographic factors

```
In [67]: 1 mg1 <- lm(log(crmrte) ~ factor(west), data = crime.narm)
2 mg2 <- lm(log(crmrte) ~ factor(central), data = crime.narm)
3 mg3 <- lm(log(crmrte) ~ factor(urban), data = crime.narm)
```

```
In [68]: 1 # Compute robust standard errors
2 se.mg1 = sqrt(diag(vcovHC(mg1)))
3 se.mg2 = sqrt(diag(vcovHC(mg2)))
4 se.mg3 = sqrt(diag(vcovHC(mg3)))
5 # We pass the standard errors into stargazer through the se argument.
6 stargazer(mg1, mg2, mg3,
7           type="text", keep.stat=c("n", "adj.rsq"),
8           se = list(se.mg1, se.mg2, se.mg3),
9           star.cutoffs=c(0.05, 0.01, 0.001)
10          )
```

```
=====
Dependent variable:
-----
log(crmrte)
(1)      (2)      (3)
-----
factor(west)1  -0.526***
                (0.129)

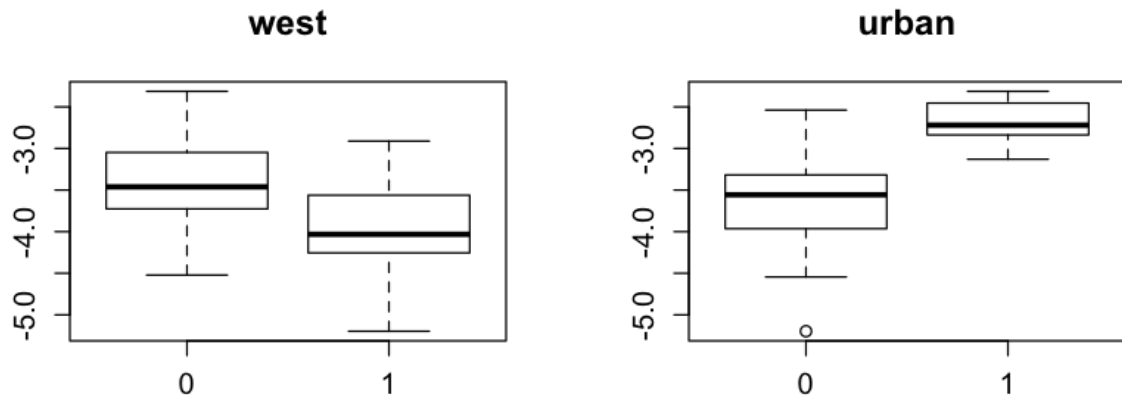
factor(central)1      0.208
                    (0.118)

factor(urban)1      0.942***
                   (0.115)

Constant      -3.413***  -3.620***  -3.625***
              (0.061)   (0.075)   (0.055)

-----
Observations      90      90      90
Adjusted R2      0.162    0.023    0.233
=====
Note:      *p<0.05; **p<0.01; ***p<0.001
```

```
In [69]: 1 options(repr.plot.height = 3, repr.plot.width = 7, repr.plot.pointsiz
2 par(mfrow = c(1,2))
3 plot(factor(crime.narm$west), log(crime.narm$crm rte), main = "west")
4 plot(factor(crime.narm$urban), log(crime.narm$crm rte), main = "urban")
```



In the category of geographic factors, *west* significantly affects crime rate with negative correlation. On the other hand, *urban* significantly affects crime rate with positive correlation

### Do outcomes of crime affect crime rate?

- crime outcomes

```
In [70]: 1 mc1 <- lm(log(crm rte) ~ log(prbarr), data = crime.narm)
2 mc2 <- lm(log(crm rte) ~ log(prbconv), data = crime.narm)
3 mc3 <- lm(log(crm rte) ~ log(prbpris), data = crime.narm)
4 mc4 <- lm(log(crm rte) ~ log(avgsen), data = crime.narm)
5 mc5 <- lm(log(crm rte) ~ log(mix), data = crime.narm)
```

```

In [71]: 1 # Compute robust standard errors
2 se.mc1 = sqrt(diag(vcovHC(mc1)))
3 se.mc2 = sqrt(diag(vcovHC(mc2)))
4 se.mc3 = sqrt(diag(vcovHC(mc3)))
5 se.mc4 = sqrt(diag(vcovHC(mc4)))
6 se.mc5 = sqrt(diag(vcovHC(mc5)))
7 # We pass the standard errors into stargazer through the se argument.
8 stargazer(mc1, mc2, mc3, mc4, mc5,
9           type="text", keep.stat=c("n", "adj.rsq"),
10          se = list(se.mc1, se.mc2, se.mc3, se.mc4, se.mc5),
11          star.cutoffs=c(0.05, 0.01, 0.001)
12          )

```

```

=====
                        Dependent variable:
-----
                        log(crmrte)
-----
                (1)          (2)          (3)          (4)          (5)
-----
log(prbarr)    -0.594***
                (0.156)

log(prbconv)           -0.366*
                      (0.158)

log(prbpris)                0.172
                      (0.287)

log(avgsen)                0.046
                      (0.294)

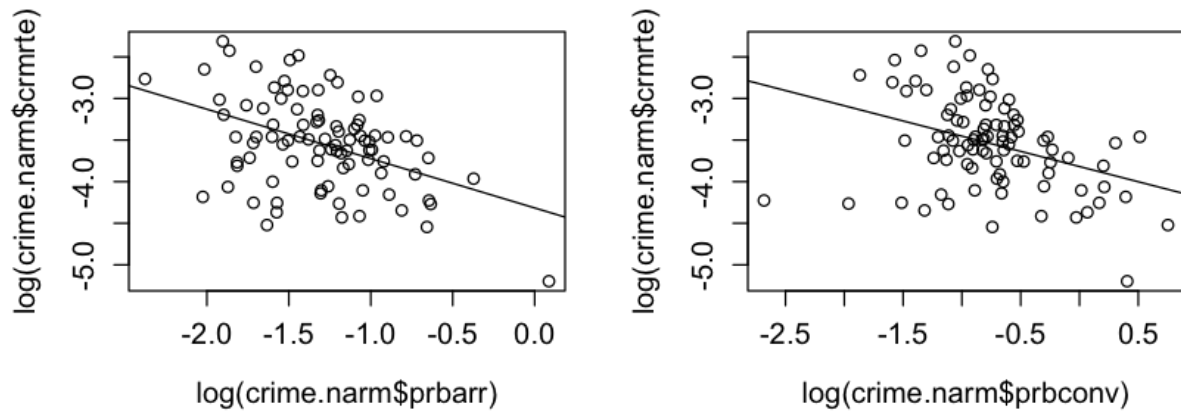
log(mix)                                0.0004
                                      (0.114)

Constant      -4.316*** -3.818*** -3.385*** -3.644*** -3.541***
                (0.199)  (0.113)  (0.276)  (0.644)  (0.252)

-----
Observations    90          90          90          90          90
Adjusted R2      0.181      0.129      -0.006      -0.011      -0.011
=====
Note:                                *p<0.05; **p<0.01; ***p<0.001

```

```
In [72]: 1 options(repr.plot.height = 3, repr.plot.width = 7, repr.plot.pointsiz  
2 par(mfrow = c(1,2))  
3 plot(log(crime.narm$prbarr), log(crime.narm$crmrte))  
4 abline(mc1)  
5 plot(log(crime.narm$prbconv), log(crime.narm$crmrte))  
6 abline(mc2)
```



In the category of geographic factors, *prbarr* and *prbconv* significantly affect crime rate with negative correlation.

**Test the regression model based on the variables with related characteristics**

```
In [73]: 1  mp4 <- lm(log(crmrte) ~ log(density) + log(polpc)
2          + log(taxpc), data = crime.narm)
3  me10 <- lm(log(crmrte) ~ log(wcon) + log(wtuc) + log(wtrd) +
4            log(wfir) + log(wser) + log(wmfg) + log(wfed) +
5            log(wsta) + log(wloc), data = crime.narm)
6  mm3 <- lm(log(crmrte) ~ pctmin80 + pctymle, data = crime.narm)
7  mg4 <- lm(log(crmrte) ~ factor(west) + factor(central) + factor(urban)
8  mc6 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) +
9            log(prbpris) + log(avgsen) + log(mix), data = crime.narm)
10
11  paste("adj.r.squared of population and goverment income: ", summary(m
12  paste("adj.r.squared of employee income: ", summary(me10)$adj.r.squar
13  paste("adj.r.squared of minority and young male: ", summary(mm3)$adj.
14  paste("adj.r.squared of geographic factors: ", summary(mg4)$adj.r.squ
15  paste("adj.r.squared of crime outcomes: ", summary(mc6)$adj.r.squared
```

'adj.r.squared of population and goverment income: 0.335445808769382'

'adj.r.squared of employee income: 0.255168337418916'

'adj.r.squared of minority and young male: 0.114261587697344'

'adj.r.squared of geographic factors: 0.360644623908643'

'adj.r.squared of crime outcomes: 0.396489655078683'

The OLS model of population and goverment income varibales shows adjusted  $R^2 = 0.335$ . The OLS model of employee income varibales shows adjusted  $R^2 = 0.255$ . The OLS model of minority and young male varibales shows adjusted  $R^2 = 0.114$ . The OLS model of geographic factors varibales shows adjusted  $R^2 = 0.361$ . The OLS model of crime outcomes varibales shows adjusted  $R^2 = 0.396$ . Therefore, the OLS models based on the categories do not provide competitive explanations of the variation in *crmrte*.

In [ ]: 1