Lab 3: Reducing Crime

w203 Summer 2018

Madeleine Bulkow, Kim Darnell, Alla Hale, Emily Rapport

7/19/2018

1. Introduction

As advisees to political campaigns for statewide office in North Carolina (NC), we believe that the crime rate across the state should be of central concern to any candidate. Local governments desire to control the crime rate, and rigorous data analysis is needed to understand the role of crime in different parts of the state. This report examines the available crime data and attempts to answer the following research question: What variables are associated with crime rates across counties in North Carolina? Based on this analysis, we generate several policy suggestions applicable to government officials in North Carolina for the late 1980s.

2. Data Definitions and Data Cleaning

The data analyzed in this report were collected as part of a multi-year study on crime by Cornwell and Trumball, originally published in 1994. The data include various factors potentially related to crime for 90 of the 100 counties in North Carolina. Because of legal limitations on access to the full dataset, this report will focus exculsively on the opensource data from 1987.

The dataset includes the following variables, which we present with definitions and assumptions:

county: An integer code indicating which NC county the row in the datafile represents. Review of relevant factors suggests that these are FIPS codes, which are standard county identification codes generated by the Environmental Protection Agency (see http://enacademic.com/dic.nsf/enwiki/49697 for details on FIPS codes for the state, including detailed maps).

year: A value of 1987 for all data points.

crmrte: The ratio of crimes committed per person, taken from the FBI's Uniform Crime Reports.

prbarr: The ratio of arrests to offenses, taken from the FBI's Uniform Crime Reports.

prbconv: The ratio of convictions to arrests. Arrest data is taken from the FBI's Uniform Crime Reports. Conviction data is taken from the North Carolina Department of Correction.

prbpris: The ratio of prison sentences to convictions, taken from the North Carolina Department of Correction.

avgsen: The average prison sentence in days. Although it is unspecified in the materials we received, we assume these data come from the North Carolina Department of Correction.

polpc: The police per capita, computed using the FBI's police agency employee counts.

density: The number of 100 people per square mile.

taxpc: The tax revenue per capita; we assume that this refers only to taxes assessed in units of $100 dollars at the state level or lower.

west: An indicator code specifying whether county is in Western North Carolina (1 if yes, 0 if no).

central: An indicator code specifying whether county is in Central North Carolina (1 if yes, 0 if no).

urban: An indicator code specifying whether county is urban, defined by whether the county is in a Standard Metropolitan Statistical Area as defined by the U.S. Census (see https://www.encyclopedia.com/finance/finance-and-accounting-magazines/standard-metropolitan-statistical-areas).

pctmin80: The percentage of population that belongs to a non-White racial group according to the 1980 U.S. Census.

mix: The ratio of face-to-face offenses (e.g., physical assault) to other offenses (e.g., automobile theft).

pctymle: The percentage of young males, defined as proportion of population that is male between the ages of 15 and 24, according to the 1980 U.S. Census data.

The remaining variables represent weekly wages in particular industries, as provided by the North Carolina Employment Security Commission:

wcon: construction

wtuc: transit, utilities, and communication

wtrd: wholesale, retail trade

wfir: finance, insurance, real estate

wser: service industry

wmfg: manufacturing

wfed: federal employees

wsta: state employees

wloc: local government employees

We start by evaluating the available data, cleaning it by removing anomolous values, and and transforming relevant variables.

```
# Import the data
df = read.csv("crime_v2.csv")
#summary(df)
```

In the original dataset, there are several variables, including *prbarr*, *prbpris*, and *pctymle* that represent probabilities and are given as decimal values between 0-1. To facilitate comparing the coefficients for these variables more easily with other numerical values in the dataset, we converted these to percentages between 0-100. We note that there is one county, Madison County (FIPS code 115), for which the *prbarr* value post-conversion is greater than 100%. This could reflect an error in data gathering or recording, but it may also reflect that it is common for indviduals in this county to be arrested with greater frequency than they commit specific offenses.

There is another "probability" factor, namely the probability of conviction after arrest (*prbconv*), for which several of the original values are greater than 1. Although this may seem anomalous, it reflects the fact that this factor is more effectively described as a ratio of the number of convictions per arrest, rather than as a probability of being convicted following arrest. Given that one might not be arrested for all crimes that one is subsequently convicted of, it makes sense that these numbers have a greater range than a traditional probabilities.

The variable *mix*, related to whether a crime is "face-to-face" or not, is also scaled in the original dataset with values between 0-1 in the original dataset. To facilitate the discussion of these data in tangible terms, we converted the values to values between 0-100 to make their scale more consistent with the rest of the data.

The variable *polpc* represents the number of police officers per known resident in a county, which is somewhat intangible on an individual scale. To address this, we converted the scores for this variable from number of police per capita in a county to number of police per 1000 residents of a county. That is, it is simply clearer to say, “There are 4 police officers per 1000 residentsin X county” than “There is .004 of a police officer per resident of X county”.

In the original dataset, the values for Wilkes Country (FIPS code 193) are given twice. We removed one set of these values so that they would not affect the overall analysis. In addition, there were six rows in the dataset that had no values for any variable. We assumed these rows were unintentionally included and removed all of them.

Data were not provided for the following counties (FIPS county codes are provide in parentheses): Camden (29), Carteret (31), Clay (43), Gates (73), Graham (75), Hyde (95), Jones (103), Mitchell (121), Tyrrell (177), and Yancey (199). We do not know why these cases were omitted from the original dataset, nor can we say the extent to which the omission of 1/10 counties across the state might affect the effectiveness of our recommendations. However, a review of 2012 population estimates for the omitted counties (see http://us-places.com/North-Carolina/population-by-County.htm) indicate that 9/10 are ranked between 86-100 of the 100 counties in overall population. The remaining omitted county is ranked 37th overall in population in the state and is close to several major metropolitan areas in the Northeast.

```
# Clean the data

## Reassign the dataframe to a working variable
df_calc <- df

# Convert the prbarr, prbpris, and pctymle variables from decimals to percentages
df_calc$prbarr <- df$prbarr * 100
df_calc$prbpris <- df$prbpris * 100
df_calc$pctymle <- df$pctymle * 100

# Convert the mix variable from decimals to percentage
df_calc$mix <- df$mix * 100

# Convert the polpc variable from decimals to number of police per 1000 people
df_calc$polpc <- df$polpc * 1000

# Convert the prbconv variable from integer to numeric
df_calc$prbconv <- as.numeric(levels(df$prbconv)[df$prbconv])
```

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

```
#remove row 89, which is a duplicate of row 88 (Madison County, FIPS 193)
df_clean <- df_calc[-c(89), ]

#remove rows with no data (i.e., all NA values)
df_clean <- df_clean[-c(91:97), ]
```

3. Building Models

Our central goal for this analysis is to determine what variables are most associated with crime across the state of North Carolina. For this reason, we will use the *crmrte* variable as the outcome variable for all of our models.

To begin, we examine the distribution of *crmrte* to determine its center and variability. This reveals that value of *crmrte* ranges from approximately .6% to 9.9%, with a mean of approximately 3.4%.There are 90 total cases and no missing cases.

```
summary(df_clean$crmrte)
```

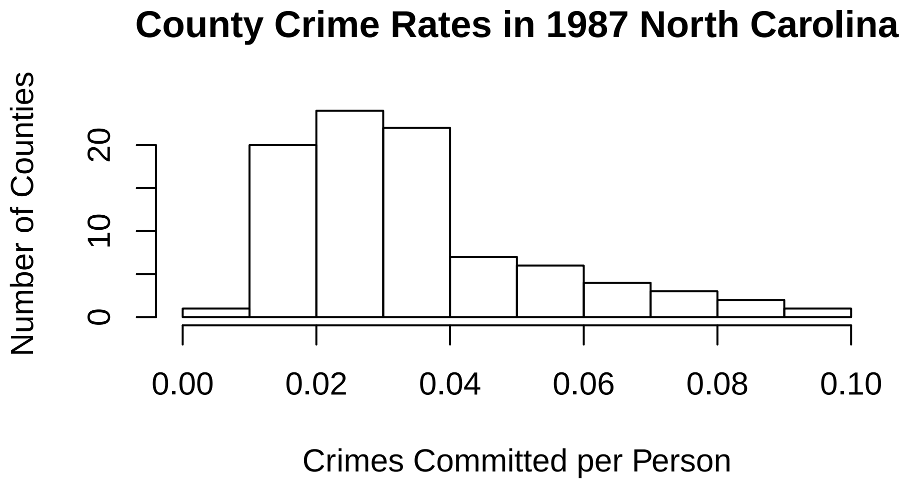
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966
```

```
length(df_clean$crmrte)
```

```
## [1] 90
```

A histogram of the data reveals that the crime rate data are positively skewed, with the majority of counties having a crime rate between 1-4%. The extended right tail indicates that a few counties have substantially higher crime rates, with some between 8-10%.

```
hist(df_clean$crmrte,
     main="County Crime Rates in 1987 North Carolina",
     xlab= "Crimes Committed per Person",
     ylab= "Number of Counties")
```



4. The Models

We model the factors contributing to crime rate in North Carolina in five stages, resulting in five related, but distinct, models.

The first four of our models are linear regressions of crime rate against an increasing numbers of predictive variables.

- Model 1 includes only the variables we believe to be the main predictors of crime rate: population density (*density*), tax per capita (*taxpc*), and percentage of young males in the population (*pctymle*).
- Model 2 includes the factors from Model 1 as well as several others that we believe contribute meaningfully to crime rate, including location in the state, the number of police per 1000 residents (*polpc*), the probability of being arrested (*prbarr*), and the pro bability of being convicted if arrested (*prbconv*).
- Model 3 builds on Model 2 by adding information about the percentage of minorties (*pctmin80*), the probability of getting a prison sentence (*prbpris*), and the average length of sentence (*avgsen*).
- Model 4 includes all available variables, even those of questionable merit.

For each of these models, we expect a classic linear model with the following assumptions:

- Assumption 1: Linearity in parameters, such that each fit model has slope coefficients that are linear multipliers of the associated predictor variables.
- Assumption 2: Random sampling, such that the data points are independent and identically distributed. We have data for 90

The remaining model, Model 5 is a regression of *crmrte* multiplied by *mix*. NEED MADELEINE'S INPUT HERE.

4.1 Model 1

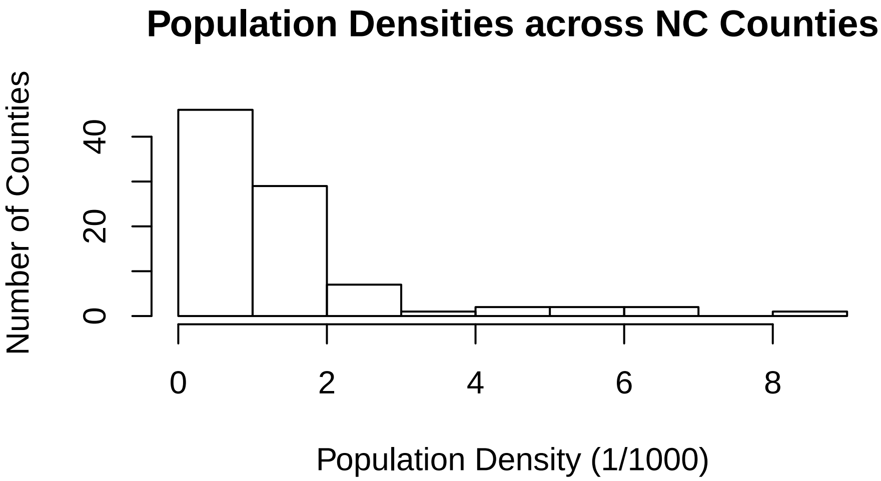
The current data suggest that population density, local tax per capita, and the percentage of young males in the population are strong predictors of crime. We evaluate each of these three predictor variables in turn.

Density:

```r
summary(df_clean$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.54718 0.97925 1.43567 1.56926 8.82765
```

```r
hist(df_clean$density,
     main="Population Densities across NC Counties",
     xlab= "Population Density (1/1000)",
     ylab= "Number of Counties",
     breaks = 10)
```



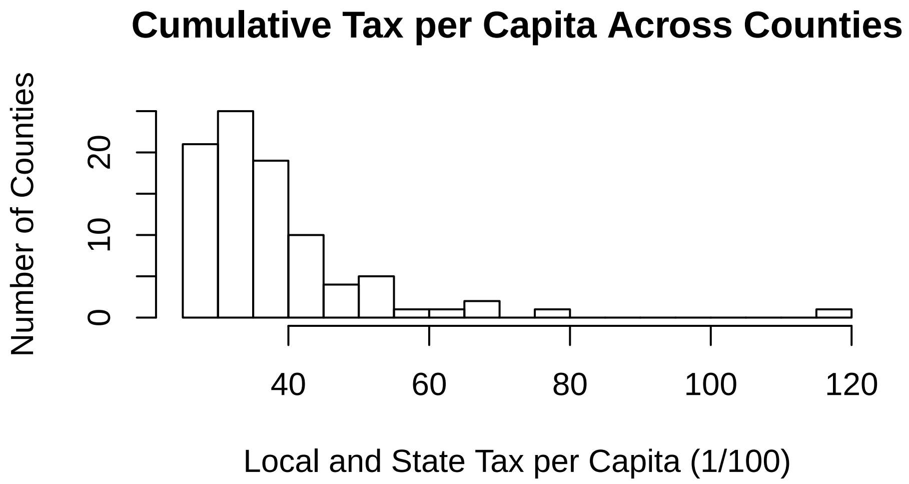
The value of density ranges from a score of approximately .002 to 880 people per square mile, with a mean of 145. The distribution of county densities is right skewed, with most counties having a score of 200 or fewer people per square mile.

Tax per Capita:

```r
summary(df_clean$taxpc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.73   34.92   38.16   41.01  119.76
```

```
hist(df_clean$taxpc,
     main="Cumulative Tax per Capita Across Counties",
     xlab= "Local and State Tax per Capita (1/100)",
     ylab= "Number of Counties",
     breaks = 30)
```



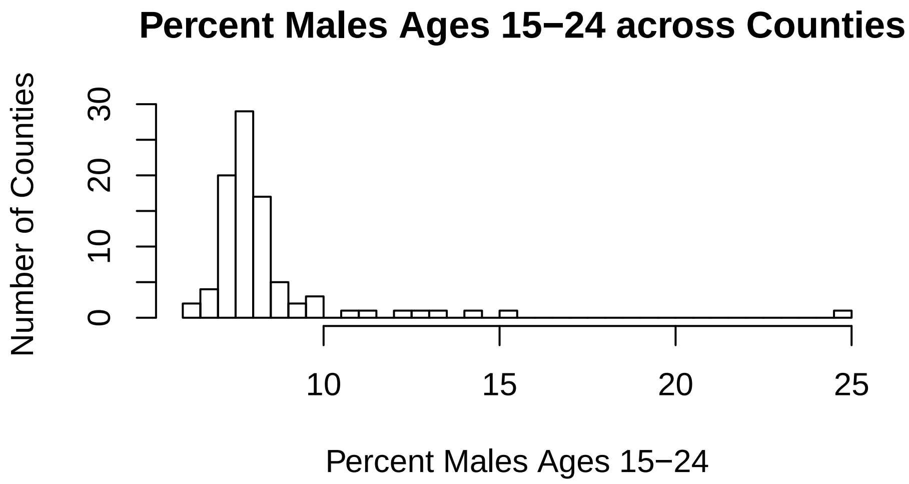
The cumulative value of taxes assessed at the local and state levels per capita ranges from $2,569 to $11,976 per year. Once again, we see a distribution that is right skewed, with revenue in most counties below the mean of $3,813 per year. The maximum value, the value for Dare county (FIPS 55) is nearly 50% higher than the next closest value, suggesting that this county has an anomalously high tax rate for the state.

Percent Young Male:

```
summary(df_clean$pctymle)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.216   7.437   7.770   8.403   8.352  24.871
```

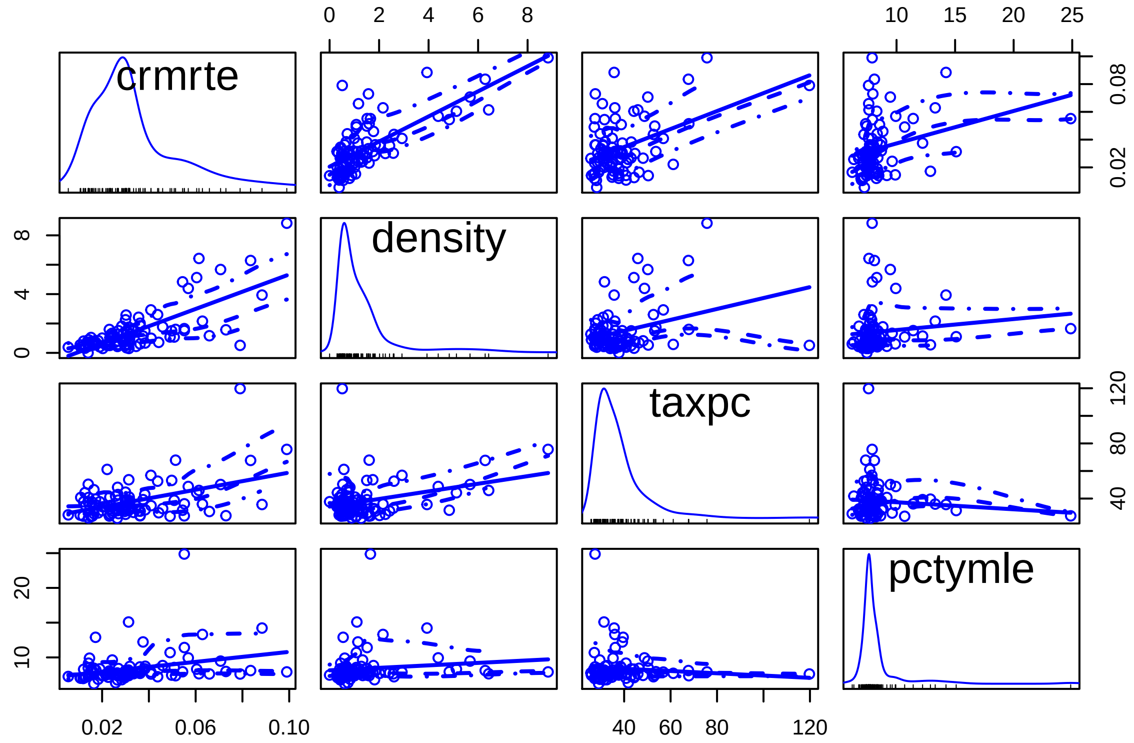
```
hist(df_clean$pctymle,
     main= " Percent Males Ages 15-24 across Counties",
     xlab= "Percent Males Ages 15-24",
     ylab= "Number of Counties",
     breaks = 30)
```



The percent of males between 15-24 years of age ranges from 6.2 to 24.9% across counties. Once again, we see a distribution that is right skewed, with the majority of counties having fewer than the average distribution of 8.4%. As with other primary variables, we see one extreme value: that for Onslow county (FIPS 133). This reflects that Onslow county includes the city of Jacksonville, recognized as the youngest county in the U.S. (see https://en.wikipedia.org/wiki/Jacksonville,_North_Carolina) in large part because it contains both the United States Marine Corps' Camp Lejeune and the New River Air Station, both of which are inhabited predominately by males under 25 years of age.

Before we build our model, we review the matrix of scatterplots of crime rate and the three variables evaluated above to identify any potential collinearity.

```r
vars <- c("crmrte", "density", "taxpc","pctymle")
suppressWarnings(scatterplotMatrix(df_clean[,vars], diagonal = "histogram"))
```



As we suspected, crime rate looks well predicted by each of the three primary variables selected as evidenced by the fairly strong positive slopes in the bivariate regressions in the scatterplot matrix. Additionally, though density and taxpc appear to have a positive correlation, none of the variables are collinear with any of the others.

With the evaluation of the variables complete, we build model 1, and evaluate the Cook's Distance for the residuals:

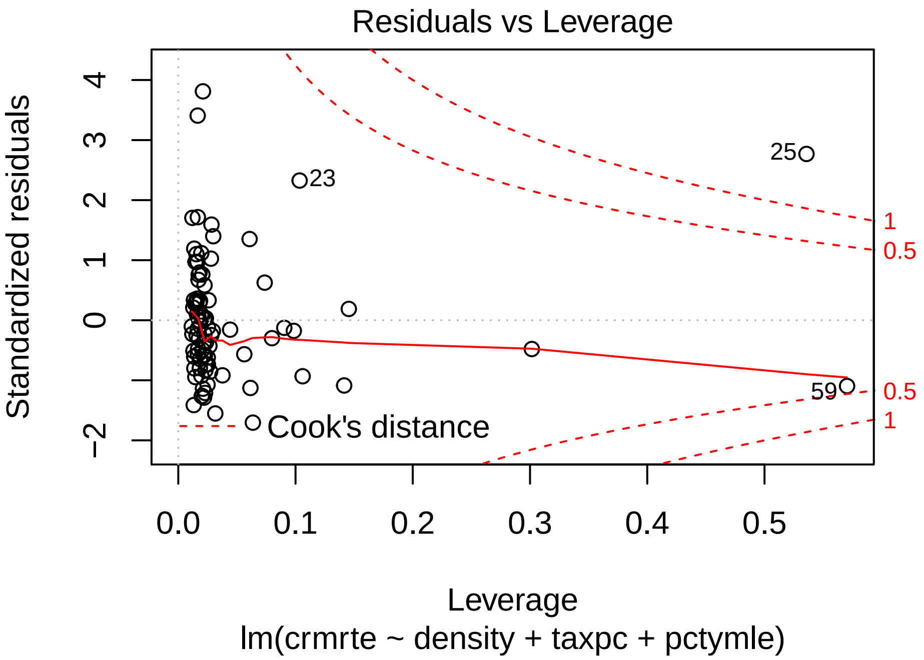
```
# Build Model 1
(model_1 = lm(crmrte ~ density + taxpc + pctymle, data = df_clean))
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle, data = df_clean)
##
## Coefficients:
## (Intercept)      density        taxpc      pctymle
##  -0.0091081    0.0075972    0.0003965    0.0019734
```

```
summary(model_1)$r.square
```

```
## [1] 0.6404252
```

```
plot(model_1, which = 5)
```



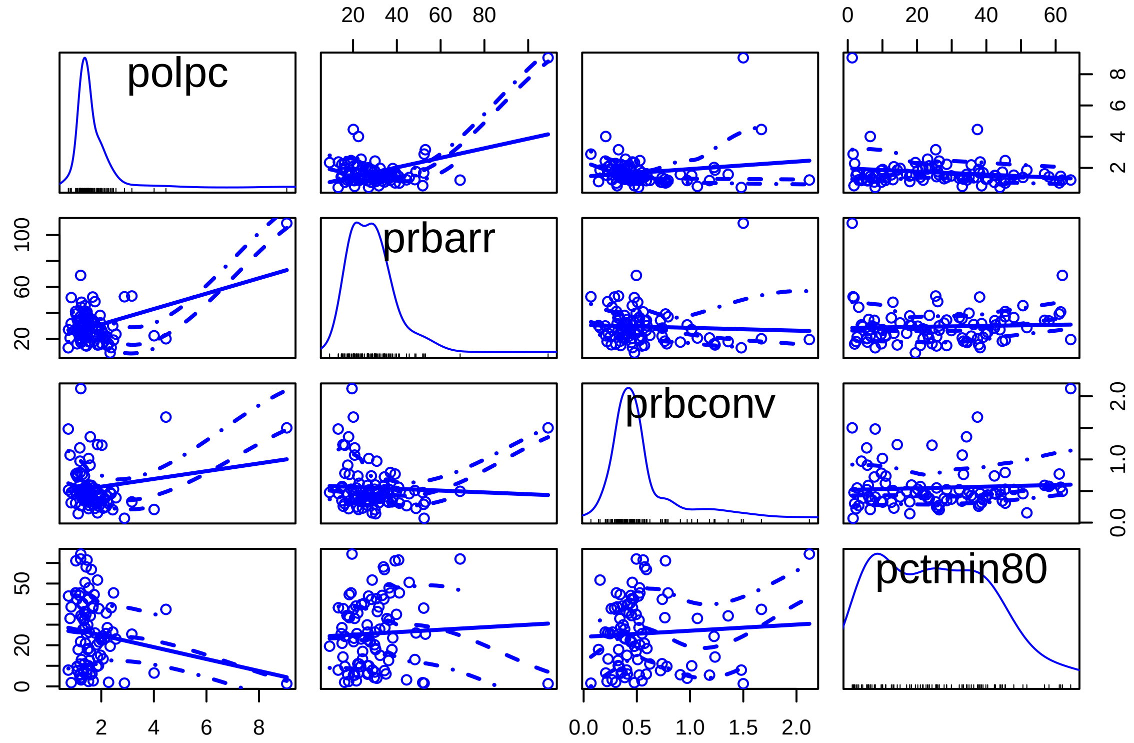
We find one point that has a Cook's Distance greater than 1, but with no justification to remove it from the dataset, we simply note it.

4.2 Model 2

Model 2 includes west, polpc, prbarr, and prbconv in addtion to the three variables from Model 1. During our EDA, we found that each of these had interesting correlations with the variable of interest, crime rate.

We conducted a full EDA on each of the explanatory variables, but for the sake of space, a simple matrix plot of the additional variables, other than west, is shown below.

```
vars <- c("polpc", "prbarr", "prbconv","pctmin80")
suppressWarnings(scatterplotMatrix(df_clean[,vars], diagonal = "histogram"))
```



The matrix plot shows little correlation between most of the additional variables in this model.

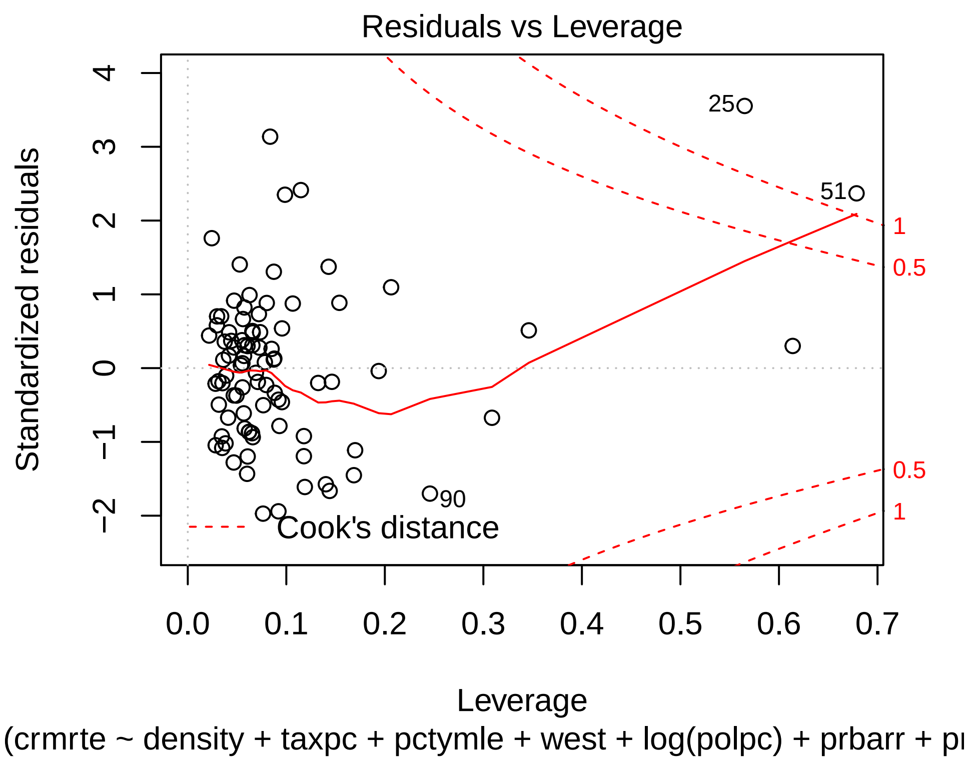
```
# Build Model 2
# model 2: other things that are explanatory but maybe questionable: west, polpc, arrest/conviction, pp
(model_2 = lm(crmrte ~ density + taxpc + pctymle
              + west + log(polpc) + prbarr + prbconv + pctmin80,
              data = df_clean))
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle + west + log(polpc) +
##     prbarr + prbconv + pctmin80, data = df_clean)
##
## Coefficients:
## (Intercept)      density        taxpc      pctymle         west
##   0.0191705    0.0054129    0.0001725    0.0008912   -0.0027764
##  log(polpc)       prbarr      prbconv     pctmin80
##   0.0147147   -0.0004212   -0.0150239    0.0002830
```

```
summary(model_2)$r.square
```

```
## [1] 0.8151506
```

```
plot(model_2, which = 5)
```



Interestingly, the R^2 increased from 0.63 to 0.79 with these additional 5 variables included.

4.3 Model 3

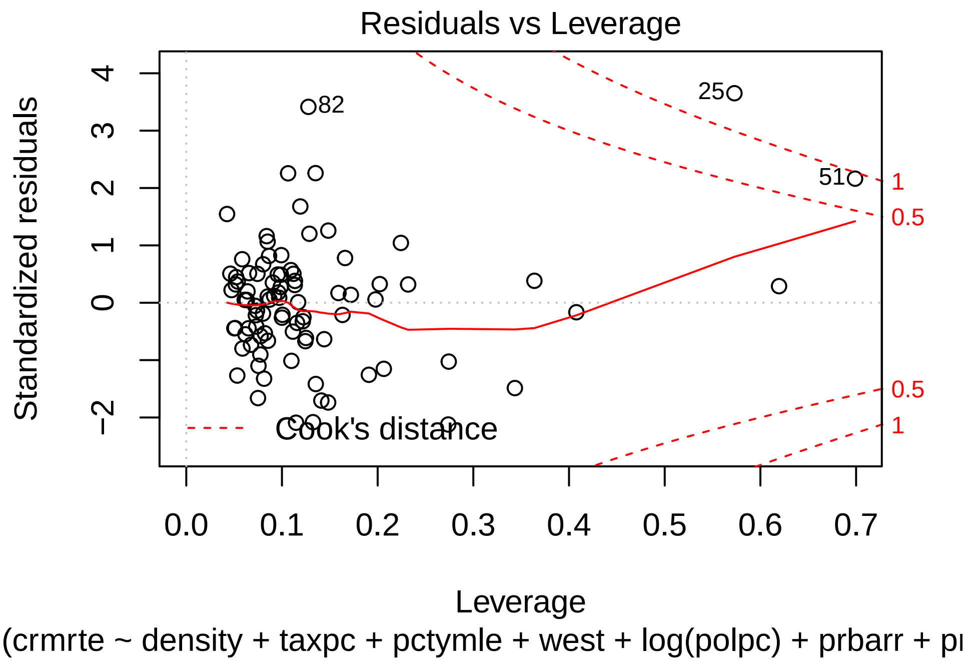
```
# Build Model 3
#model 3: not necessarily explanatory, but not problematic: central, avgsen, prison.
(model_3 = lm(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + prbconv + pctmin80 + ce
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle + west + log(polpc) +
##     prbarr + prbconv + pctmin80 + central + avgsen + prbpris,
##     data = df_clean)
##
## Coefficients:
## (Intercept)      density        taxpc      pctymle         west
##   2.496e-02    5.862e-03    1.312e-04    7.258e-04   -6.947e-03
##  log(polpc)       prbarr      prbconv     pctmin80      central
##   1.571e-02   -4.165e-04   -1.447e-02    2.011e-04   -4.968e-03
##      avgsen      prbpris
##  -2.530e-04    7.541e-05
```

```
summary(model_3)$r.square
```

```
## [1] 0.824118
```

```r
plot(model_3, which = 5)
```



### 4.4 Model 4

```r
# Build Model 4
# model 4: kitchen sink. urban, wage.
(model_4 = lm(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + prbconv + pctmin80 + cer
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle + west + log(polpc) +
##     prbarr + prbconv + pctmin80 + central + avgsen + prbpris +
##     urban + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
##     wsta + wloc, data = df_clean)
##
## Coefficients:
## (Intercept)      density        taxpc      pctymle         west
##   2.125e-03    5.307e-03    1.565e-04    1.038e-03   -4.494e-03
##  log(polpc)       prbarr      prbconv     pctmin80      central
##   1.444e-02   -4.248e-04   -1.324e-02    2.532e-04   -5.039e-03
##      avgsen      prbpris        urban         wcon         wtuc
##  -3.050e-04    4.287e-05   -1.472e-04    1.223e-05    9.789e-06
##        wtrd         wfir         wser         wmfg         wfed
##   3.468e-05   -4.130e-05   -3.190e-06   -4.202e-06    2.495e-05
##        wsta         wloc
##  -1.293e-05    4.626e-05
```

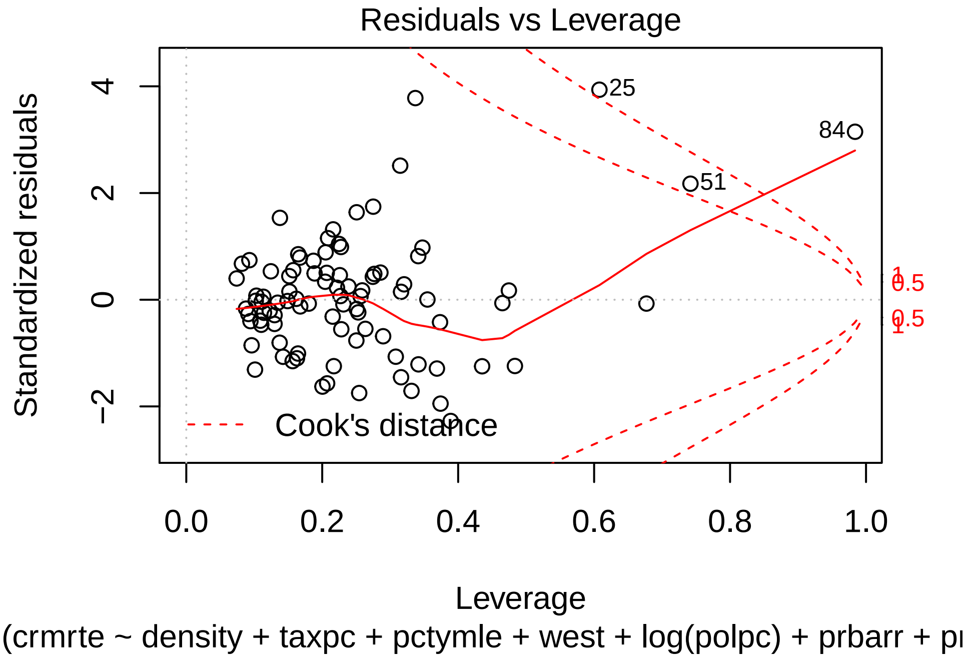
```r
summary(model_4)$r.square
```

```
## [1] 0.8420878
```

```
plot(model_4, which = 5)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



4.5 Model 5

```
# Build Model 5
# model 5: the model 1 version of a model for this dependent variable - crmrate*mix
```

4.2 Model Summary

This is where we put our model summary table.

```
stargazer(model_1, model_2, model_3, model_4, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "n"),
          omit.table.layout = "n") # Omit more output related to errors
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 19, 2018 - 3:56:27 PM

Table 1: Linear Models Predicting Crime Rate

	Dependent variable:			
	crmrte			
	(1)	(2)	(3)	(4)
density	0.008	0.005	0.006	0.005
taxpc	0.0004	0.0002	0.0001	0.0002
pctymle	0.002	0.001	0.001	0.001
west		−0.003	−0.007	−0.004
log(polpc)		0.015	0.016	0.014
prbarr		−0.0004	−0.0004	−0.0004
prbconv		−0.015	−0.014	−0.013
pctmin80		0.0003	0.0002	0.0003
central			−0.005	−0.005
avgsen			−0.0003	−0.0003
prbpris			0.0001	0.00004
urban				−0.0001
wcon				0.00001
wtuc				0.00001
wtrd				0.00003
wfir				−0.00004
wser				−0.00000
wmfg				−0.00000
wfed				0.00002
wsta				−0.00001
wloc				0.00005
Constant	−0.009	0.019	0.025	0.002
Observations	90	90	90	90
R^2	0.640	0.815	0.824	0.842

5. Omitted Variables

6. Conclusion

- might be worth making a point about county as unit : might make sense since county likely determines different police/judicial jurisdictions, but certain elements in our model might not be consistent across whole county (i.e. density, tax rate in cities, etc.)