Lab 3: Reducing Crime

w203 Summer 2018

Madeleine Bulkow, Kim Darnell, Alla Hale, Emily Rapport

7/21/2018

1. Introduction

As advisees to political campaigns for state and local office in North Carolina (NC), we believe that the crime rates across the state should be of central concern to any candidate. State and 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 candidates and government officials in North Carolina for the late 1980s.

2. Variable Definitions and Assumptions

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 North Carolina county a given row in the datafile represents. Review of relevant factors suggests that these integers 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; we assume these data come from the North Carolina Department of Correction.

polpc: The number of police officers 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 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 transforming relevant variables.

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

Data Adjustments and Anomalies

The dataset has several ratio variables, including **prbarr**, **prbpris**, **pctymle**, and **mix**, that recorded 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 their scale to 0-100, as in percentages. The exception to this approach was *prbconv*, which reflects the ratio of convictions to arrests. This variable has several values that are greater than 1, indicating that there are counties where individuals are convicted of more crimes than they were intially arrested for. Modifying the scale of this variable did not seem to improve its interpretablity, so it was unchanged.

The variable *polpc* represents the number of police officers per known resident in a county, which is somewhat intangible on an individual scale. That is, it is awkward to refer to ".004 police officers per resident." To address this, we multiplied the scores for this variable by 1000, permitting descriptions such as "4 police officers per 1000 residents."

There is one county, Madison County (FIPS 115), for which the *prbarr* value is greater than 100%. This anomaly 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. We did not remove, replace, or adjust this score.

The data for Wilkes Country (FIPS 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. Understanding Crime Rate

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 crime rate as our primary outcome variable.

To begin, we examine the distribution of *crmrte* to determine its center and variability. This reveals that value of *crmrte* ranges from approximately 0.6% to 9.9%, with a mean of approximately 3.4%. There are 90 total cases, with no missing values among them.

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

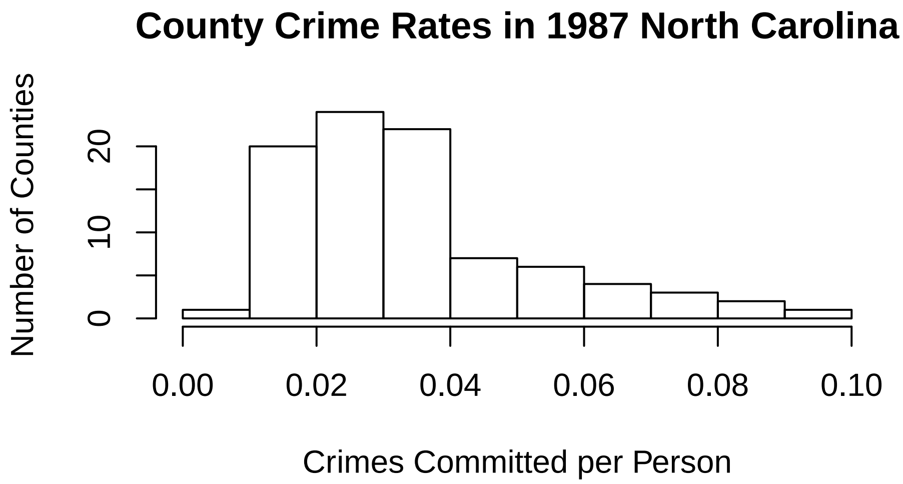
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966
```

```
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")
```



```
## Note from Kim: I don't understand the point of this section on face-to-face crime as it is currently

## Describing the creation of a new variable as "torturous" seems out of place in a statistical report 

## Also, we need to be sure to make all of our figure titles and labels consistent in terms of wording,
```

Due to the perceived severity of face-to-face crimes over non-face-to-face- crimes, we determined that a useful addendum to our research question would be: what are the factors associated with the face-to-face crime rate in North Carolina? We can find the face-to-face crime rate using the overall crime rate and the mix variable, which we recall contains the fraction of face-to-face crimes to other crimes. After making the fairly safe assumption that face-to-face + other = total crime, a somewhat tortuous manipulation gives us what we need: the ratio of face-to-face crimes among all crimes committed in a county.

$$\begin{aligned}
\frac{\text{face-to-face}}{\text{total}} &= 1 - \frac{\text{other}}{\text{total}} && (1)\\
&= 1 - \frac{\text{other}}{\text{face-to-face} + \text{other}} && (2)\\
&= 1 - \frac{1}{\frac{\text{face-to-face}+\text{other}}{\text{other}}} && (3)\\
&= 1 - \frac{1}{\frac{\text{face-to-face}}{\text{other}} + 1} && (4)\\
&= 1 - \frac{1}{\text{mix} + 1} && (5)
\end{aligned}$$

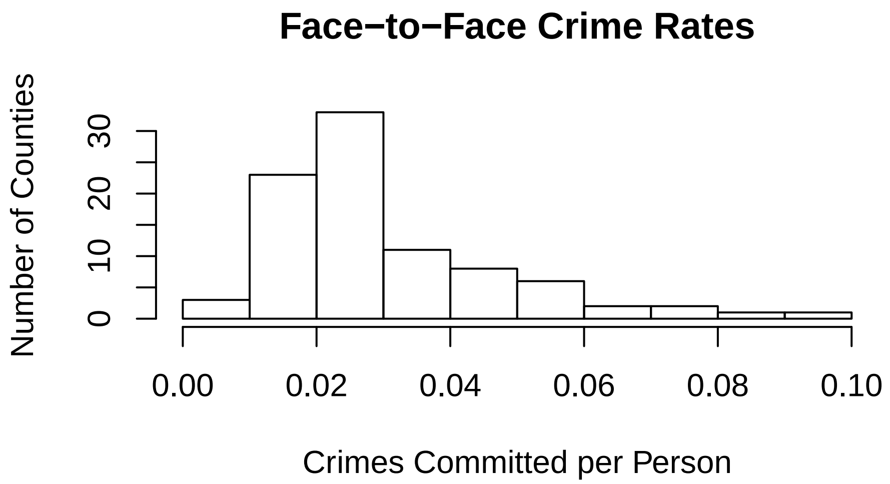
Now when multiplied with the overall crime rate, this gives the face-to-face crime rate in each county.

```
# Calculate the face-to-face crime rate
df_clean$ftfcrmrte <- df_clean$crmrte * (1-1/(df_clean$mix+1))
# Examine the distribution
```

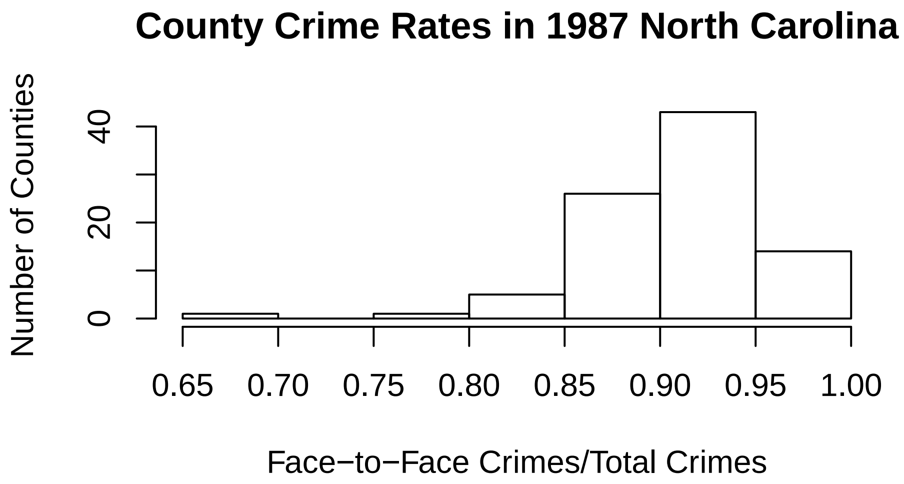
```
summary(df_clean$ftfcrmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.00503 0.01886 0.02692 0.03048 0.03649 0.09343
```

```
hist(df_clean$ftfcrmrte,
     main="Face-to-Face Crime Rates",
     xlab= "Crimes Committed per Person",
     ylab= "Number of Counties")
```



```
df_clean$crmrte_ratio <- 1-1/(df_clean$mix+1)
hist(df_clean$crmrte_ratio,
     main= "County Crime Rates in 1987 North Carolina",
     xlab= "Face-to-Face Crimes/Total Crimes",
     ylab= "Number of Counties")
```



The distribution of the face-to-face crime rate is similar, but not identical to, the base crime rate. The mean is lower, as we would expect, but only slightly lower, and values now range from 0.5% to 9.3%. The effect of this manipulation appears to be fairly small, so we will focus on the unadjusted crime rate for the majority of our analysis.

4. Models and Assumptions

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 number 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 (*west*), the number of police per 1000 residents (*polpc*), the ratio of arrests to offenses (*prbarr*), the ratio of convictions to arrests (*prbconv*), and the proportion of non-White minorities (*pctmin80*).
- Model 3 builds on Model 2 by adding more information about the location of the county (*central*), the ration of prison sentences to convictions (*prbpris*), and the average length of prison sentence (*avgsen*).
- Model 4 builds on Model 3 and adds all other variables in the dataset that are not covariant with any predictor variables that are already included.
- Model 5 explores whether predictors we identify for general crime rate are comparably effective for predicting face-to-face crime by creating a modified outcome variable using *crmrate* and *mix*.

Each of our models will be assess to determine it's consistency with the following assumptions, which are standard for classic linear regression models like these.

```
##Note from Kim: I think we really need to move the general descriptions of all of our CLM 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.
- Assumption 3: No perfect collinearity, such that none of the variables in the sample is a constant and there is no exact linear relationship among the predictor variables.
- Assumption 4: Zero conditional mean, such that the statistical error in the model has an expected value of 0 given any values of the predictor variables.
- Assumption 5: Homoskedasticity, such that the statistical error in the model has the same variance given any value of the predictor variables.
- Assumption 6: Normality, such that the statistical error in the population is independent of the predictor variables and is normally distributed with zero mean and variance sigma-squared.

```
## The paragraph below is important, but it should go after the assumptions have all been described in
```

Relevant to Assumption 2 regarding random sampling, we remind the reader that our dataset reflects 90 of North Carolina's 100 counties, which is very close to the overall population. However, 9/10 of the omitted counties are those for which current population estimates are in the low 15% for the state. It is relevant to note that in Northeast cities in the early 1980s, violent (i.e., face-to-face), but not property (i.e., non-face-to-face), crime rates were correlated with population density (see https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=99314/).

```
## The statement below seems out of place. If we are going to transform variables to when it improves pr
```

To maintain interpretability, we did not transform the variables unless significant predictive gains were made.

4.1 Model 1

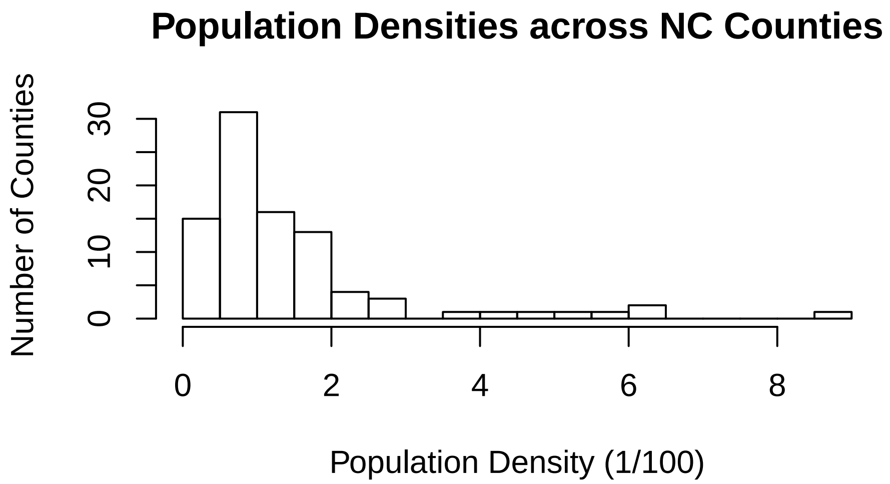
An exploratory examination of the data suggest that population density, local and state tax per capita, and the percentage of young males in the county are strong predictors of the general crime rate. We begin by evaluating and describing each of these predictor variables in turn.

Density:

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

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

```
hist(df_clean$density,
     main="Population Densities across NC Counties",
     xlab= "Population Density (1/100)",
     ylab= "Number of Counties",
     breaks = 15)
```



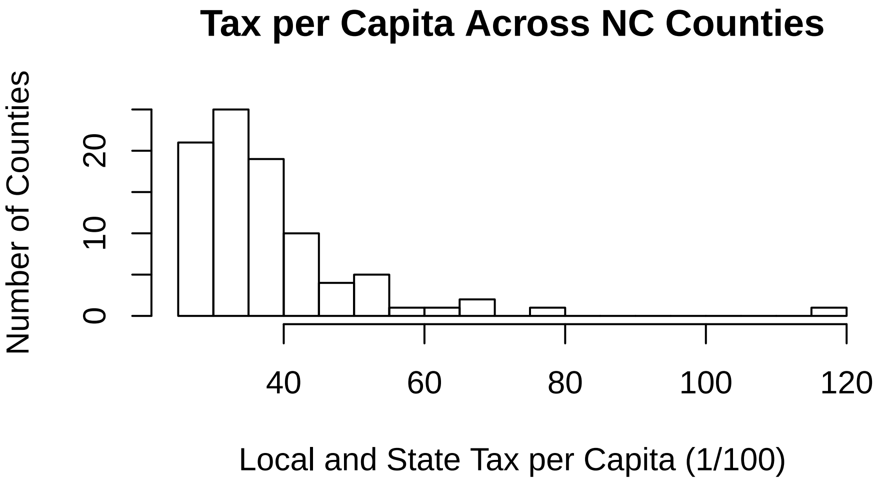
The value of density ranges from a score of approximately 0.002 to 880 people per square mile, with a mean of 145. The distribution of county densities is right skewed, with most counties having a score of 200 or fewer people per square mile.

Tax per Capita:

```
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="Tax per Capita Across NC Counties",
     xlab= "Local and State Tax per Capita (1/100)",
     ylab= "Number of Counties",
     breaks = 30)
```



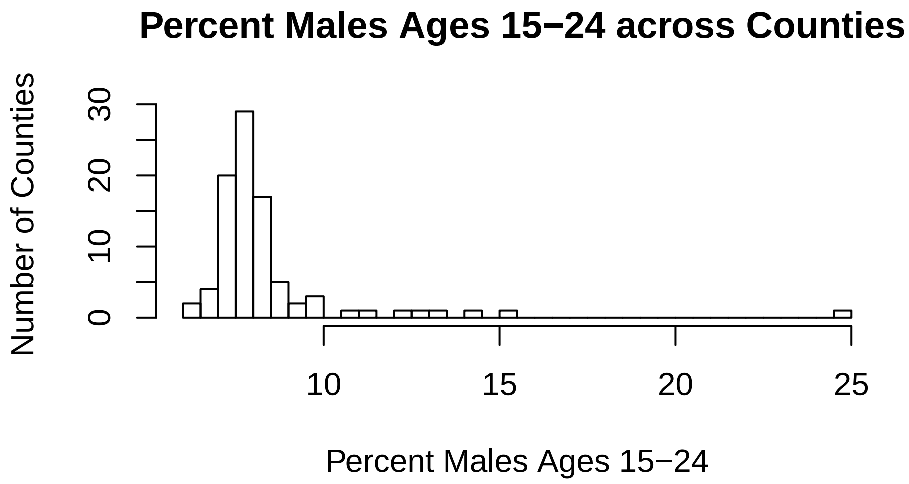
The cumulative value of taxes assessed at the local and state levels per capita ranges from $2,569 to $11,976 per year. Once again, we see a distribution that is right skewed, with revenue in most counties below the mean of $3,813 per year. The maximum value, the value for Dare county (FIPS 55) is nearly 50% higher than the next closest value, suggesting that this county has an anomalously high tax rate for the state. In fact, a review of the official website for Dare county (https://www.darenc.com/) reveals that it has an extremely active toursim industry and features a number of popular attractions, including the Outer Banks beach resort area, the Wright Brothers National Monument, the North Carolina Aquarium, and a number of other historic and recreational sites. The high rate of tax per capita for this country can easily be explained by taxes on activities related to toursim, such as those appended to hotel, rental car, and park entrance fee costs.

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
```

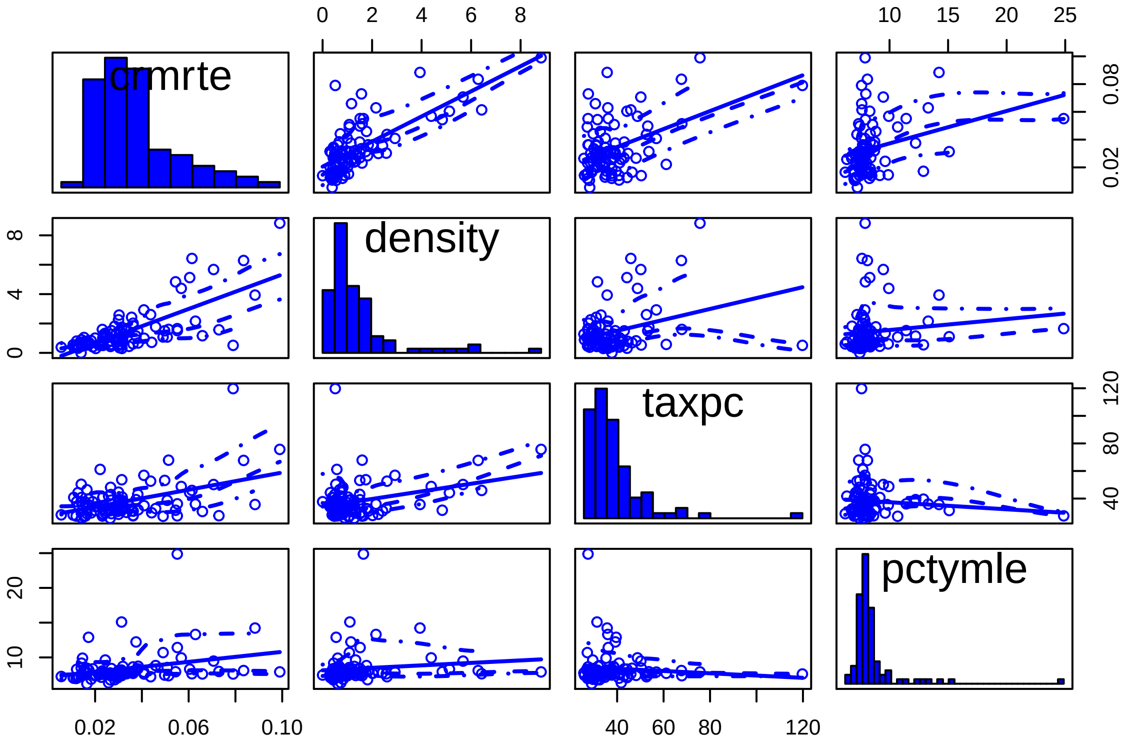
```
hist(df_clean$pctymle,
     main= " Percent Males Ages 15-24 across Counties",
     xlab= "Percent Males Ages 15-24",
     ylab= "Number of Counties",
     breaks = 30)
```



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.

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



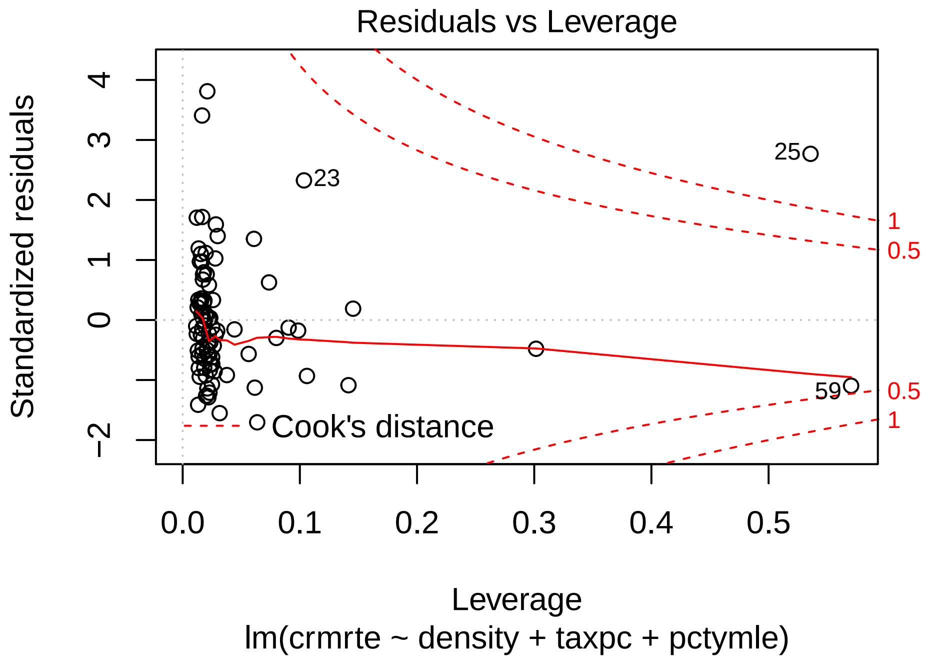
As previously indicated, crime rate appears predicted by each of the three primary variables selected as evidenced by the fairly strong positive slopes in the bivariate regressions in the scatterplot matrix. Additionally, though *density* and *taxpc* appear to have a positive correlation, none of the variables are collinear with any of the others, so we can make Assumption 3, no perfect collinearity.

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

```r
# Build Model 1
model_1 = lm(crmrte ~ density + taxpc + pctymle, data = df_clean)
summary(model_1)$r.square
```

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

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



We find one point that has a Cook's Distance greater than 1, corresponding to Dare county. As noted previously, Dare country has substantially higher tax per capita than other North Carolina counties because of tax revenue from tourism. As such, the deviation of this single case is understandable and does not warrant its removal.

Now that the model is built, we can validate Assumption 4 regarding zero conditional mean. To do this, we check that the expectation of the fitted values times the residuals is 0 (the denominator does not matter for this calculation, so we just take the sum of fitted values times residuals).

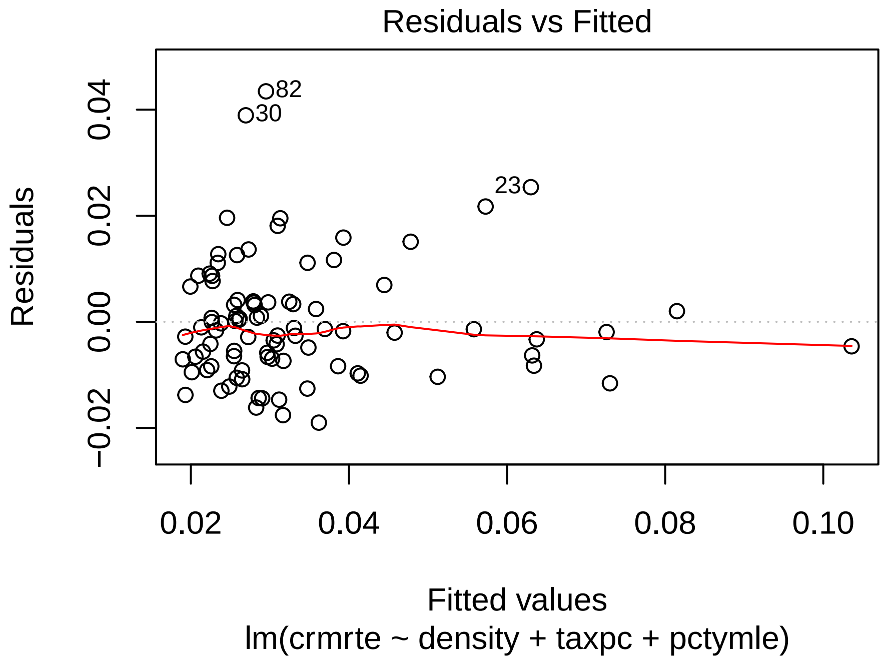
```
round(sum(model_1$residuals * model_1$fitted.values), 15)
```

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

We find that the residuals times the fitted values sum to 0, indicating the model has the desired property of exogeneity.

To validate the model in terms of Assumption 5 regarding homoskedasticity, we create a plot of the residuals vs. fitted values and note that the error range was relatively constant throughout the range of fitted values. Although the data appear to demonstrate the desired characteristic, we note that we have fewer data points at the higher values of crime rate than the lower values of crime rate, so validity in this case may be somewhat weaker than with other assumptions.

```
plot(model_1, which = 1)
```



To validate Assumption 6, the normality of the residuals, we looked at a Q-Q plot of the residuals, and noted the fairly straight line.

```
## We need the plot here that goes with the claim above.  Otherwise it seems like we are hiding somethi
```

Interpreting Model 1

In the equation for Model 1, the model coefficients are positive, indicating that as population density, tax per capita, or the percentage of young males increase, an associated increase occurs in crime rate.

```
## I don't see the equation for Model anywhere.  We need to insert that here so it makes sense for us t
```

However, it is not the case that all of our predictor variables are equally influencial when it comes to crime rate. Specifically, increases in population density result in an increase of the crime rate that is an order of magnitude greater than that for increases in tax per capita and almost four times that generated by higher percentages of young males. As such, this model indicates that while all of these predictors are useful to understanding the crime rate, the candidate's energy may be best spent on addressing crime-related concerns connected to population density first, followed by those related to tax rate and the proportion of young males. In the comments to follow, we highlight factors that could be influencing our findings through correlation with our predictor variables in **bold**.

There are a number of reasons why increases in population density could facilitate increases the crime rate. As more people live in a particular space, there are more opportunities for them to come into conflict with one another, to interact with others who have different access to desireable resources an items, and to be unfamiliar with others with whom one comes in contact day by day. As such, candidates with constituencies in high population areas should consider addressing the crime rate by developing policies that improve the ease with which large numbers of people can live and move in the same space, while reducing opportunties for conflict. **Infrastructure** projects that increase the livability and communal nature of high population areas, such as well-maintained public parks and recreational areas, effective public transportation, and improved traffic and parking managment may make it easier for residents live in close quarters with others and reduce the number of negative experiences that might lead to criminal behavior. Similarly, addressing problems related to **socioeconomic inequality**, such as access to quality **education**, employment opportunies, social support programs, and affordable housing should also result in a reduction in crime. Last but not least, there is the issue of anonomity. Certainly it is easier to commit a crime against a stranger than it is a neighbor or a friend, if only because there are fewer personal costs and a lower likelihood of being caught. So, investing in events, facilities, and services that encourage people to get to know and develop positive relationships with those around them, take pride in their joint **community membership**, and have opportunties to get to know one another as people should also reduce crime. These might include cultural celebrations, neighborhood vegetable gardens, or fundraising activities for an important local cause.

The finding regarding the influence of tax per capita on crime suggests two directions for policy to reduce crime. First, it makes sense that areas where residents make more money would pay higher taxes *and* be more tempting targets for crime, because their higher income affords them more access to desireable items and services. Again, this suggests developing policies that address socioeconomic disparity so that individuals with limited access to resources are less inspired to engage in criminal activity to secure basic needs (e.g., money, credit cards, items that can be fenced or pawned) from those who have more. What we do not encourage is simply increasing the police presence in high income areas or encouraging the police to engage in discriminatory profiling of members of communities that are stereotypically not associated with high socioeconomic status. These sorts of policies foment distrust among different status communities and are likely to result in unjustified harassment, mistreatment, and arrest of members of marginalized groups. In fact, such policies might increase criminal activity, by discouraging people from reporting crimes for fear of **negative police interaction** or community backlash for "snitching."

The other policy direction suggested by the tax per capita result relates to **tourism**. That is, in areas where there is a lot of tourism, this can be seen by the community as an opportunity for good employment and additional funding for community infrastructure and beneficial social programs, or it can be seen as an influx of distracted strangers with an abundance of extra money and little familiarity with the local environment. Obviously, to the extent that tourism can be framed as the having the former set of positive qualities for the communities in which it occurs, the more likely crime involving tourists is to be reduced.

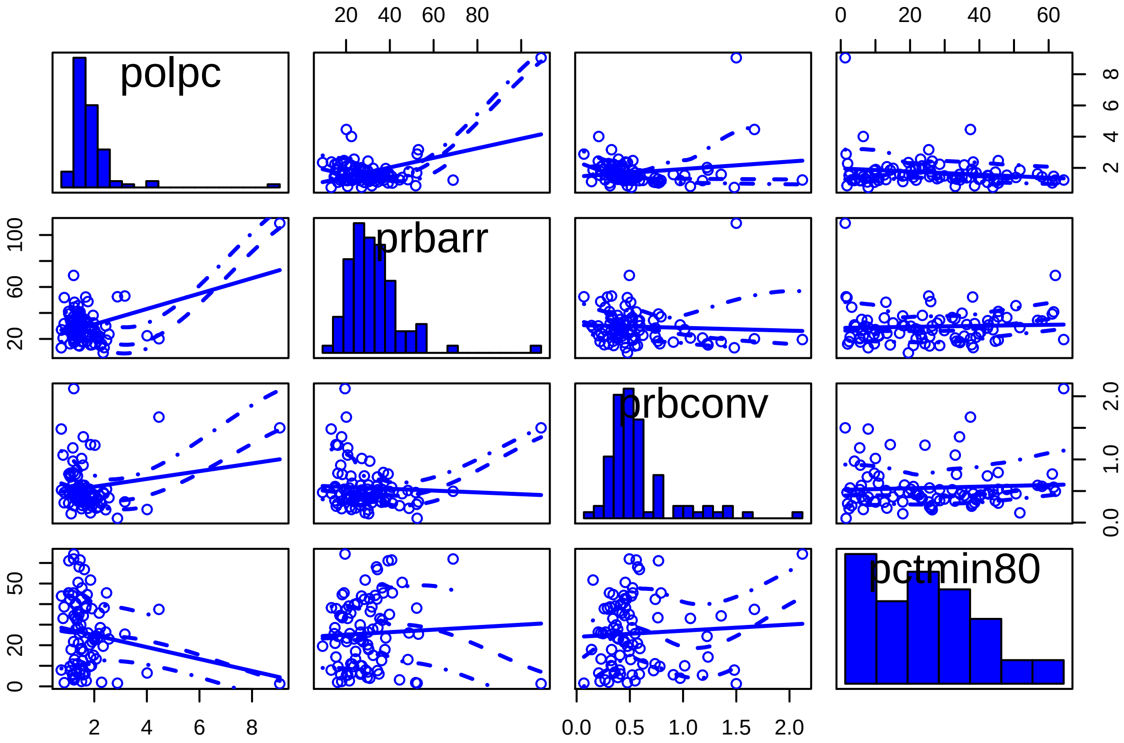
With regard to the increase in crime seen with an increase in the percentage of young males, there are a number of possibly solutions. There is certainly immense social pressure connecting masculinity with wealth and the ability to provide for a family, as well as factors that socialize men to be more aggressive or violent when their needs and wants are not immediately met. In fact, these sorts of pressures may be even more common in communities that priortize traditional and/or conservative social values. To the extend that a candidate's constituency includes such communities, it could be fruitful to consider the role that local **culture** contributes to young men committing crimes and how providing alternative, as well as socially and personally constructive outlets to demonstrate their masculinity could reduce crime. Relevant policies could support educational, vocational, and athletic programs, as well as involve young men in activies that contribute positively to the community and encourage them to develop rather than damage it.

4.2 Model 2

Model 2 includes *west*, *polpc*, *prbarr*, *prbconv*, and *pctmin80* in addition to the three variables from Model 1. During our EDA, we found that each of these had substantial 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. 24.4 % of all counties were labeled *west*.

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

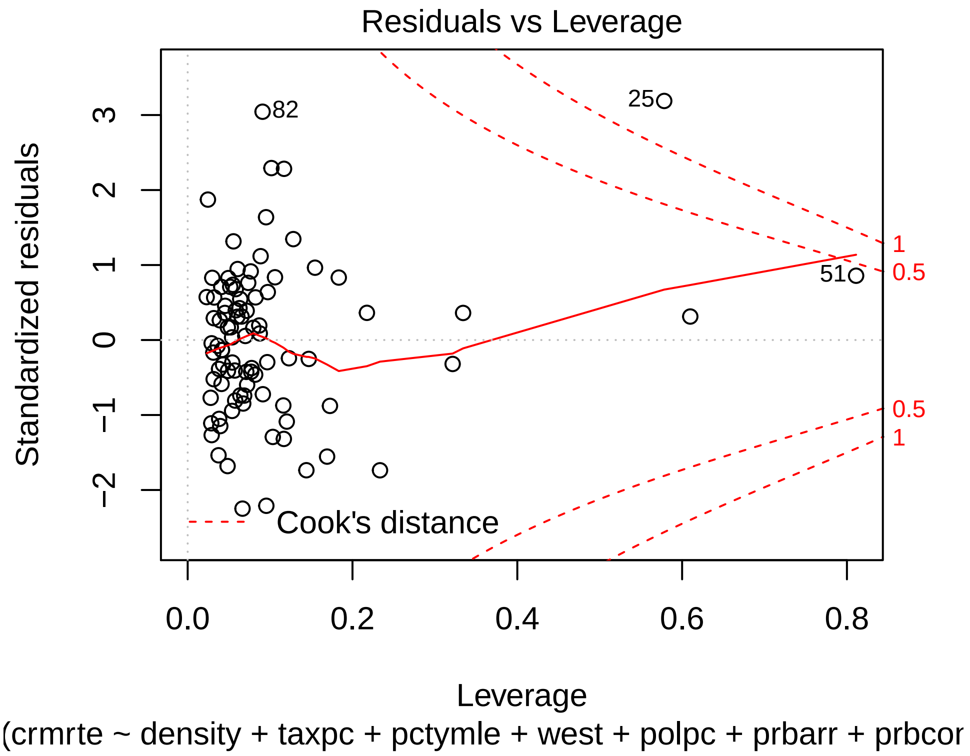


The matrix plot shows little correlation between most of the additional variables in this model, validating assumption 3, no perfect multicolinearity.

```
# Build Model 2
model_2 = lm(crmrte ~ density + taxpc + pctymle
              + west + polpc + prbarr + prbconv + pctmin80,
              data = df_clean)
summary(model_2)$r.square
```

```
## [1] 0.8240404
```

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



Unsurprisingly, the R^2 increased from 0.64 to 0.82 with these additional 5 variables included. We also note that point 25 still has high leverage, just as in model 1. Perhaps we should study that county a bit more closely.

We also check Assumption 4, exogeneity, by summing the product of the residuals and fitted values and finding the sum of 0.

```
round(sum(model_2$residuals * model_2$fitted.values), 15)
```

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

Assumptions 5 and 6 were validated for this model as they were for model 1.

Model 2, shown in the table in section 4.6 has positive coefficients for *density*, *taxpc*, *pctymle*, *polpc*, and *pctmin80* indicating that crime rate increases and these variables increase. On the other hand, the coefficients for *west*, *prbarr*, and *prbconv* are negative, indicating that crime rate decreases as these increase.

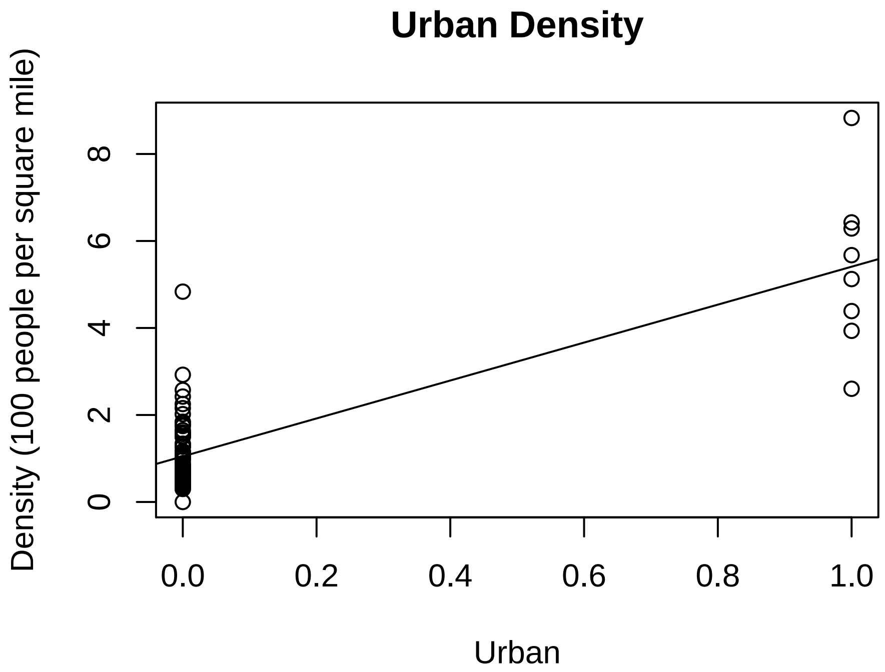
The additional coefficients in this model are somewhat more challenging to interpret than those in model 1. It seems unlikely that the longitude of a county would have a direct impact on its crime rate, and more likely that there is some omitted variable associated with crime that is more prevalent in Western counties. Additionally, the positive association between police per capita and crime is noteworthy. This association should be studied further, ideally with causal analysis, as there are plausible causal theories going in either direction. Perhaps heightened police presence creates an antagonistic relationship between officers and citizens, which leads to a distrust of authority and an increase in crime; the ideal way to test that would probably be to find counties with similar crime rates and other demographics where one county changes a policing policy and the other one doesn't, a natural paired experiment. However, it also seems possible that a county that experiences more crime would choose to up the size and acitivity of its police force in order to combat said crime; in this case, police records and government policy could probably help uncover this relationship. Local officials should pursue this line of research further to make informed policy decisions about policing.

The negative correlation between crime rate and both the probability of arrest and probability of conviction should also be studied further, with causal analysis as described above. It could be hypothesized that higher arrest and conviction rates deter crime. Alternatively, it could be hypothesized that when crime rate is lower, and fewer overall crimes are committed, it is easier to fully pursue all of the cases.

4.3 Model 3

For model 3, in addition to the variables from model 2, we added the remainder of the variables that we did not find problematic: *central*, *avgsen*, *prison*. These variables do not necessarily explain the crime rate well, but serve to show that model 2 gives a reasonable explanation of the observed crime rate. We excluded the urban variable because it is too closely related to density, as can be seen in this scatterplot:

```
plot(df_clean$urban , df_clean$density,
     main= "Urban Density",
     ylab= "Density (100 people per square mile)",
     xlab= "Urban")
abline(lm(density ~ urban, data = df_clean))
```



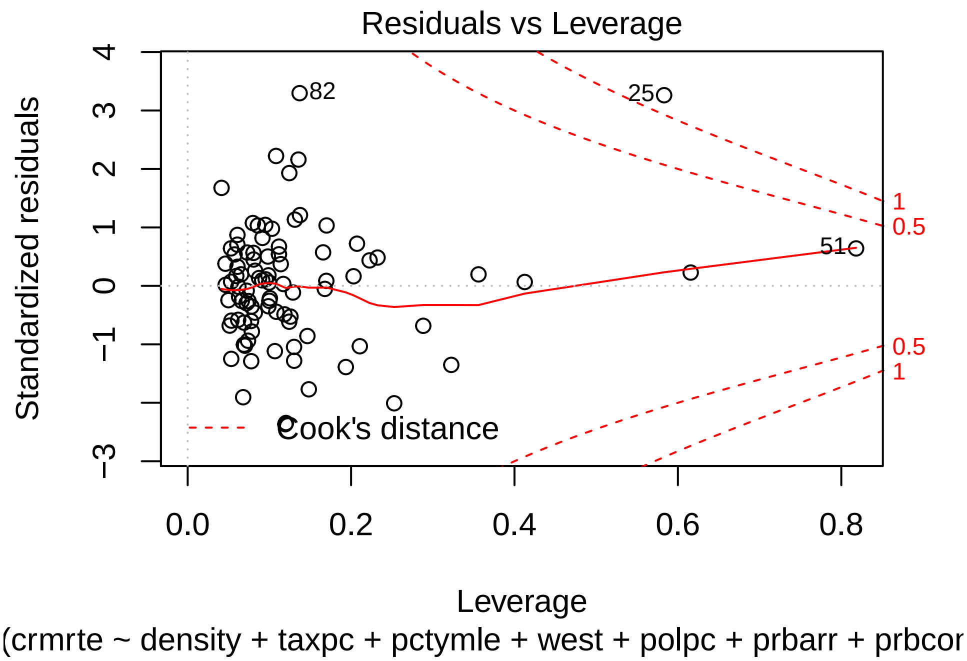
Excluded, are all of the wage variables because we cannot make any meaningful conclusions without a breakdown of what fraction of each county are involved in each profession.

With that, we build model 3:

```
# Build Model 3
model_3 = lm(crmrte ~ density + taxpc + pctymle
             + west + polpc + prbarr + prbconv + pctmin80
             + central + avgsen + prbpris,
             data = df_clean)
summary(model_3)$r.square
```

```
## [1] 0.8301355
```

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



We note that point 25 is still exhibiting a Cook's distance of greater than 1.

Assumption 3 was tested by evaluating and eliminating the chance of any perfect collinearity between these variables.

To justify Assumption 4, we show that the sum of the residuals times the fitted values is 0:

```r
round(sum(model_3$residuals * model_3$fitted.values), 15)
```

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

Assumptions 5 and 6 were validated for this model as they were for models 1 and 2.

We note that the R^2 for this model, at 0.83, is negligibly better than the R^2 for model 2. This model, while interesting as an upper bound on what can reasonably be included in a model, should not be used to influence policy decisions.

4.4 Model 4

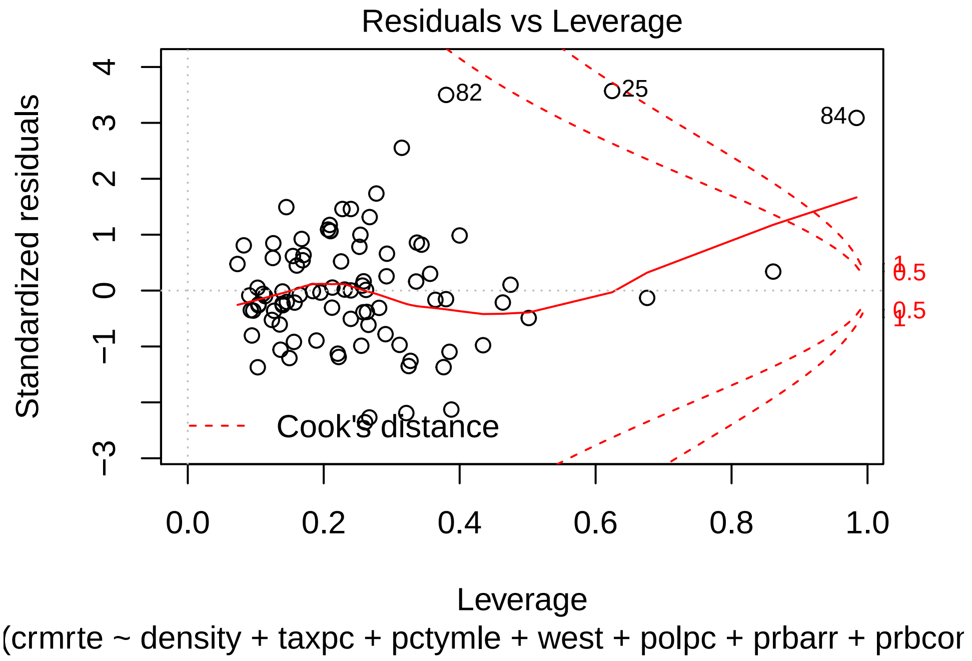
For this model, we included every variable available to us, simply to set an upper limit on the possible R^2. The resulting model is not a parsimonious one, and as such, we should not use it. However, it is interesting to note that the R^2 rises to 0.85, which is not much higher than model 3. Additionally, many points exceeding a Cook's distance of 1 are observed.

```r
# Build Model 4
# model 4: kitchen sink. urban, wage.
model_4 = lm(crmrte ~ density + taxpc + pctymle
             + west + polpc + prbarr + prbconv + pctmin80
             + central + avgsen + prbpris
             + urban + wcon + wtuc + wtrd + wfir + wser
             + wmfg + wfed + wsta + wloc + mix,
             data = df_clean)
summary(model_4)$r.square
```

```
## [1] 0.8545586
```

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

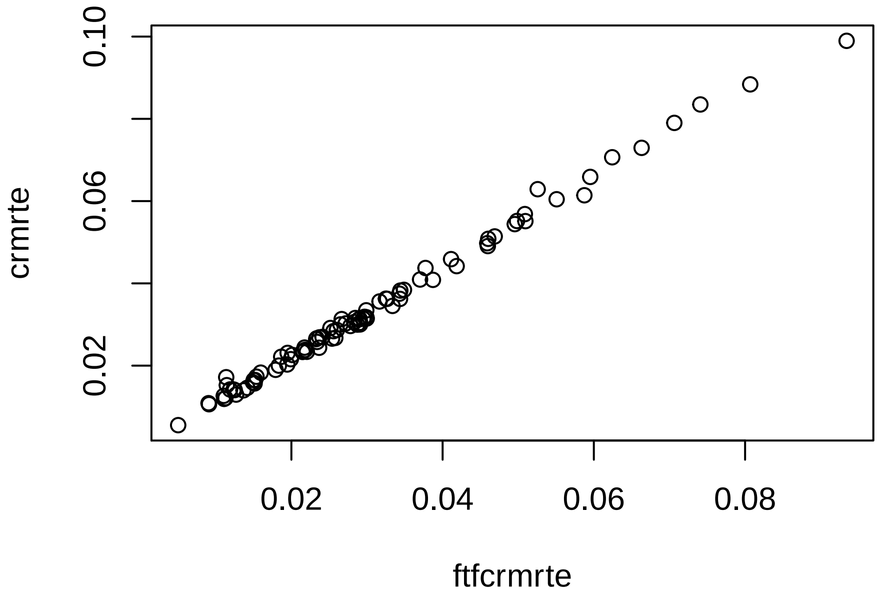
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

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



4.5 Model 5

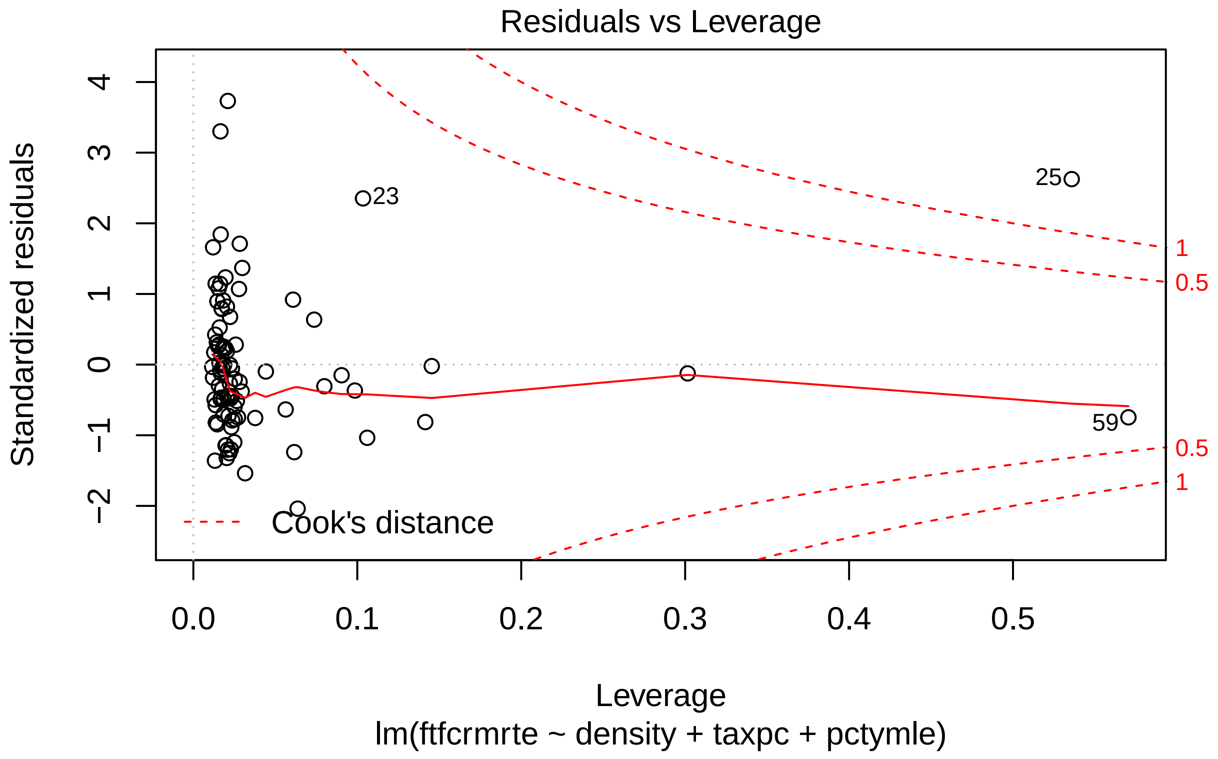
We repeated the analysis with the face to face crime rate as the dependent variable. However, results were not directionally different than the models shown above. This is unsurprising when the scatterplot of ftfcrmrte vs. crmrte is evaluated:

```
plot(crmrte ~ ftfcrmrte, data = df_clean)
```



Therefore, we only share model 5, and do not use it for any policy recommendations.

```
# Build Model 5
model_5 <- lm(ftfcrmrte ~ density + taxpc + pctymle, data=df_clean)
plot(model_5, which = 5)
```



```
summary(model_5)$r.squared
```

```
## [1] 0.6314511
```

This model is analogous to model 1, with face to face crime rate as the dependent variable. The coefficient magnitudes and signs are similar to model 1, and thus the interpretation and policy suggestions will all mirror section 4.1, and will not be restated here.

4.6 Model Summary

The models built above are summarized in **Table 4.6.1**.

```
stargazer(model_1, model_2, model_3, model_4, model_5, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "4.6.1 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: Sat, Jul 21, 2018 - 7:03:20 PM

5. Omitted Variables

While our models 1 and 2 do provide some useful information that can inform policy, there are a number of omitted variables that we did not have access to in this analysis that we suspect have meaningful associations

Table 1: 4.6.1 Linear Models Predicting Crime Rate

	Dependent variable:				
	crmrte				ftfcrmrte
	(1)	(2)	(3)	(4)	(5)
density	0.008	0.005	0.006	0.005	0.007
taxpc	0.0004	0.0002	0.0001	0.0002	0.0004
pctymle	0.002	0.001	0.001	0.001	0.002
west		−0.002	−0.005	−0.003	
polpc		0.007	0.007	0.007	
prbarr		−0.001	−0.001	−0.001	
prbconv		−0.019	−0.018	−0.019	
pctmin80		0.0003	0.0003	0.0003	
central			−0.004	−0.004	
avgsen			−0.0003	−0.0004	
prbpris			0.00004	0.00003	
urban				−0.0001	
wcon				0.00002	
wtuc				0.00001	
wtrd				0.00003	
wfir				−0.00004	
wser				−0.00000	
wmfg				−0.00001	
wfed				0.00003	
wsta				−0.00002	
wloc				0.00001	
mix				−0.0002	
Constant	−0.009	0.019	0.025	0.014	−0.007
Observations	90	90	90	90	90
R^2	0.640	0.824	0.830	0.855	0.631

with the crime rate. It is imperative that we name these variables and deduce the impact we believe they would have, or else we risk biasing our conclusions by considering only the variables we can measure.

We believe that **socioeconomic diversity** is likely to have a strong association with crime rate. If a county has a mix of people who have ample money and resources and people who have very little, there is likely to be social tension, and there is large opportunity for crime when populations who have abundant resources are in close proximity to others who need those resources desperately. We could measure socieconomic diversity by measuring the gap between the 1st and 3rd quartiles of household income. We would expect a large gap to be associated with a high crime rate, and we would also expect a positive correlation between our measured income gap and density, as dense urban areas tend to have both wealthy and impoverished people living in close proximity. In this case, omitted variable bias is positive, and the fitted values would be lower for a given density value if we had socioeconomic diversity as a variable.

We believe that the **unemployment rate**, as well as the **rate of citizens not participating in the labor force**, in a county would likely have a positive association with crime. When people are unable to earn a living, they may not have meaningful ways to spend their time, and they might struggle to pay their basic living expenses, both of which are scenarios that could be associated with crime. We might expect the correlation between unemployment and percent young male to be positive, as many young people are students or otherwise not participating in the labor force. Therefore, omitted variable bias is positive, and the fitted values would be lower for a given percent young male value if we had unemployment rate.

Additionally, we anticipate that **mean education level** for a county would likely have an impact on crime. If we added education level to a model, we would expect its coefficient to be negative, since when people have more education, they are more likely to have incomes and to contribute meaninfully to society, which seem like conditions that are unlikely to be related to crime. We anticipate a positive correlation between education level and density, since urban areas tend to have higher education levels due to job opportunities and presence of higher education institutions. Therefore, omitted variable bias for education level is negative, and the fitted values for a given value of density would likely be higher if we could control for education level.

We expect that **household earnings** would be negatively correlated with crime, since wealthier areas typically have less crime. We would expect correlation between household earnings and density to positive, because in salaries are typically higher in cities. Therefore, omitted variable bias is negative, and if we controlled for household earnings, the fitted values for a given density value would likely be higher.

The discussion of ommitted variables, however, is speculative, and should be reinforced with research, ideally randomized, controlled trials where possible.

WE SHOULD ALSO talk about the variables in our set we wanted to know more about: percentages of folks who fell in those different wage categories, better racial demographic detail, etc.

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.)