Lab 3: Reducing Crime (Stage 3)

w203 Summer 2018

Madeleine Bulkow, Kim Darnell, Alla Hale, Emily Rapport

August 1, 2018

1 Introduction

As advisors to political campaigns for state and local office, 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 in order to assure constituents that their communities are safe, and rigorous data analysis is needed to understand the factors that contribute to crime in different parts of the state. This report examines data related to crime rates provided by a current campaign in North Carolina (NC) and attempts to answer the following research question: *What variables are associated with crime rates across counties in North Carolina?*

Guided by this research question, we first examine, clean, and transform the provided data. We then propose a framework for understanding crime rates motivated by the concepts of *means*, *motive*, and *opportunity*, and link these concepts to specific variables in the data set. We then specify a model motivated by this three-part framework and examine its implications and stastical significance. Based on the findings of our primary analysis and other data exploration, we propose and interpret alternative models that provide additional predictive power. Our analysis is used to generate policy suggestions applicable to candidates seeking or defending office in North Carolina. In addition, we suggest a range of omitted variables that could form the foundation of future research on crime rates across the state.

2 Variable Definitions and Assumptions

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

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 data set, this report will focus exclusively on the opensource data from 1987. As such, our findings and recommendations apply best to the North Carolina of the late 1980s, but may have some general application to current circumstances in the state. The list of NC counties in 1987 is identical to that for 2018, and has been consistent since roughly 1779.

The data set 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 data file 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 rows. This reflects that the current data are a subset of a larger, multiyear data set.

For the remaining values, we define them here and provide distribution summaries in **Table 1**, after we we apply transformations.

crmrte: The ratio of crimes committed per person, based on reports from the Federal Bureau of Investigation's (FBI) Uniform Crime Reporting (UCR) program (see https://ucr.fbi.gov/).

prbarr: The ratio of arrests to offenses, based on reports from the FBI:UCR program.

prbconv: The ratio of convictions to arrests. Conviction data are taken from the prison and probation files of the North Carolina Department of Correction (NCDOC). Arrest data are based on reports from the FBI:UCR program.

prbpris: The ratio of prison sentences to convictions, taken from NCDOC.

avgsen: The average prison sentence in days; although the source is not specificed, we assume these values come from the NCDOC.

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

density: The number of people per square mile of land; we assume each unit reflects 100 people.

taxpc: The tax revenue per capita; we assume that this refers to taxes assessed in units of $100 U.S. at the state level or lower (e.g., individual taxes, sales taxes, and property taxes).

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 according to the U.S. Census (see https://www.encyclopedia.com/finance/finance-and-accounting-magazines/standard-metropolitan-statistical-areas).

pctmin80: The percentage of population that identified as belonging to a race/ethnicity other than "White" according to the 1980 U.S. Census. We note that in 1980, respondents were not permitted to select more than one racial/ethnic identification (see http://www.pewsocialtrends.org/interactives/multiracial-timeline/ for the specific categories by census year).

mix: The ratio of face-to-face offenses (e.g., crimes targeting a person or people) to other offenses (e.g., crimes targeting property).

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

The remaining variables in the provided data represent weekly wages in particular industries, as provided by the North Carolina Employment Security Commission (NCESC).

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

The NCESC wage data do not include details regarding the number or proportion of people whose wages were included overall or by category. Moreover, it is unclear how jobs were assigned to each wage category or if there are omitted employment categories (e.g., educators, athletes, entertainers, artists). As such, we are unable to determine the representativeness or accuracy of the provided wage data and elect not to include them in our analysis.

Next, we evaluate the available data, clean it by removing anomalous values, and transform relevant variables.

2.1 Data Adjustments and Anomalies

The data set has several ratio variables, including **prbarr**, **prbpris**, **pctymle**, and **mix**, that are presented as decimal values between 0-1. To facilitate comparing the coefficients for these variables more easily with other numerical values in the data set, we converted their scale to 0-100, as in percentages.

The exception to the percentage conversion was **prbconv**, which reflects the ratio of convictions to arrests. This variable has several values that are greater than 1, which seems anomalous, if not erroneous. However, it is also the case that the **prbconv** values were calculated using numerators and denominators from different sources (i.e., the NCDOC and the the FBI:UCR program, respectively). Thus, we suspect that the extreme cases reflect a common preference among states to report criminal convictions in enumerative detail, even

if all of the crimes for which individuals are convicted do not correspond with specific arrests that would be reported to the federal government. Larger numbers of convictions make state and local criminal justice systems appear "tough on crime," which tends to be popular with residents. We chose to leave the scale of this variable unmodified, as converting it to 0-100 did not seem to improve its interpretability.

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 person." To address this, we multiplied the scores for this variable by 1000, permitting descriptions such as "4 police officers per 1000 people."

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 individuals in this county to be arrested with greater frequency than they commit specific offenses. Given that the point of concern represents such a small part of the overall data set and is unlikely to have an untoward impact on our analysis, we chose not remove, replace, or adjust this score.

The data for Wilkes Country (FIPS 193) are given twice in the provided data. We removed one set of these values so that they would not unduly affect the overall analysis. In addition, there are six rows in the data set that have no values for any variable. We assume 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 data set, nor can we say for certain the extent to which the omission of 1/10 counties across the state might affect the effectiveness of our analysis or 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. This pattern of omission draws our attention to the variable of **density**, which we will consider with particular care.

```
# 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), ]
```

```
stargazer(df_clean[,3:25], title = "Variable Distribution Summaries",
          table.placement = "!h")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Aug 01, 2018 - 22:01:24

Table 1: Variable Distribution Summaries

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
crmrte	90	0.034	0.019	0.006	0.021	0.040	0.099
prbarr	90	29.524	13.767	9.277	20.495	34.487	109.091
prbconv	90	0.551	0.354	0.068	0.344	0.585	2.121
prbpris	90	41.063	8.067	15.000	36.422	45.756	60.000
avgsen	90	9.689	2.834	5.380	7.375	11.465	20.700
polpc	90	1.708	0.991	0.746	1.238	1.886	9.054
density	90	1.436	1.522	0.00002	0.547	1.569	8.828
taxpc	90	38.161	13.112	25.693	30.735	41.010	119.761
west	90	0.244	0.432	0	0	0	1
central	90	0.378	0.488	0	0	1	1
urban	90	0.089	0.286	0	0	0	1
pctmin80	90	25.713	16.985	1.284	10.024	38.183	64.348
wcon	90	285.353	47.753	193.643	250.754	314.979	436.767
wtuc	90	410.907	77.355	187.617	374.331	440.679	613.226
wtrd	90	210.921	33.870	154.209	190.710	224.282	354.676
wfir	90	321.621	53.999	170.940	285.560	342.628	509.466
wser	90	275.338	207.396	133.043	229.338	277.650	2,177.068
wmfg	90	336.033	88.231	157.410	288.598	359.895	646.850
wfed	90	442.619	59.951	326.100	398.785	478.255	597.950
wsta	90	357.740	43.294	258.330	329.272	383.155	499.590
wloc	90	312.280	28.132	239.170	297.228	328.775	388.090
mix	90	12.905	8.176	1.961	8.060	15.206	46.512
pctymle	90	8.403	2.345	6.216	7.437	8.352	24.871

3 Understanding Crime Rate

Our central goal for this analysis is to determine what variables are most clearly predictive of crime across the different counties of North Carolina. For this reason, we will use **crmrte** as our primary outcome variable.

To begin, we examine the distribution of **crmrte** to determine its center and variability, based on data from 90 counties; there were no missing cases. This reveals that value of **crmrte** ranges from approximately 0.006 to .099, with a mean of approximately .034. In practical terms, this means that the 1987 crime rate in North Carolina varies from approximately 0.6 to 9.9 crimes per 100 people, with an average of 3.4 crimes per 100 people.

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

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

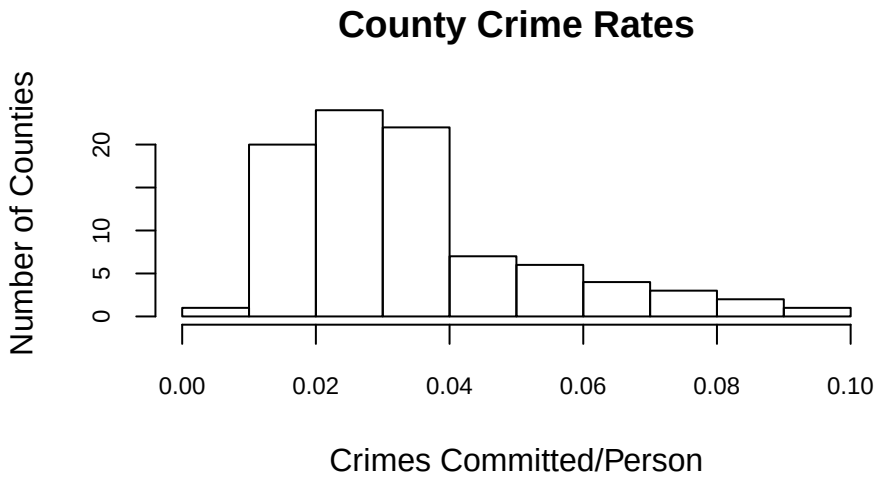
```
length(df_clean$crmrte)
```

```
## [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% (i.e., 1-4 crimes per 100 people). The extended right tail indicates that a few counties have substantially higher crime rates, with some between 8-10% (i.e., 8-10 crimes per 100 people).

```
hist(df_clean$crmrte,
     main="County Crime Rates",
     xlab= "Crimes Committed/Person",
```

```
    ylab= "Number of Counties", cex.axis=.75)
```



Although constituents are concerned with the crime rate in general, our experience suggests that they are often more concerned about crimes that result in personal physical harm to people (i.e., "personal crime", with a focus on "violent crime") than those that simply result in loss or damage to property (i.e., "property crime"). For this reason, the effective political candidate must not simply focus on policies for reducing the general crime rate, but must consider how to perceptibly reduce the personal crime rate – and especially the violent crime rate – for their constituents.

In the current data set, we may generally access the distinction between personal and property crime by examining the different rates of "face-to-face crime"", which reflects crime directly involving people, and"other" crime, which includes crime not directly involving people. This permits us to address a corollary to our primary research question, namely: *What are the variables associated with the* ***face-to-face*** *crime rates across counties in North Carolina?*

Extracting the face-to-face crime rate involves manipulations on **crmrte** involving **mix**, which the reader will recall is the ratio of face-to-face crimes to other crimes. Specifically, we begin by assuming that the total crime rate equals the face-to-face crime rate + the other crime rate. A somewhat tortuous manipulation, detailed below, allows us to calculate the ratio of face-to-face crimes among all crimes committed for each country in the data set.

$$\frac{\text{face-to-face}}{\text{total}} = 1 - \frac{\text{other}}{\text{total}} \tag{1}$$

$$= 1 - \frac{\text{other}}{\text{face-to-face} + \text{other}} \tag{2}$$

$$= 1 - \frac{1}{\frac{\text{face-to-face}+\text{other}}{\text{other}}} \tag{3}$$

$$= 1 - \frac{1}{\frac{\text{face-to-face}}{\text{other}} + 1} \tag{4}$$

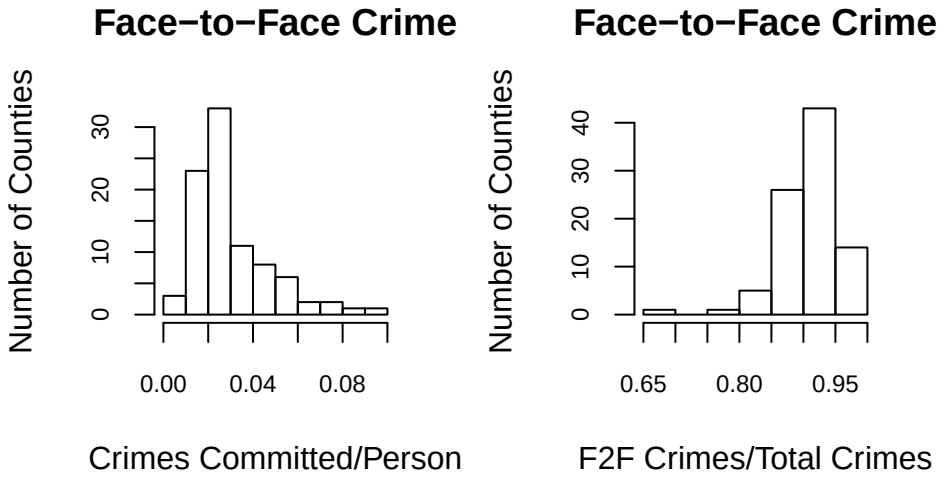
$$= 1 - \frac{1}{\text{mix} + 1} \tag{5}$$

We use the resulting fraction, multiplied with the value for **crmrte** for a given county, to generate the face-to-face crime rate (i.e., **f2fcrmrte**) for that county. Below, we summarize the distribution of face-to-face crime across the counties of North Carolina in two histograms. On the left, the number of face-to-face crimes per person across counties. On the right, the proportion of face-to-face crimes to total crimes across counties.

```r
# 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
```

```r
# Plot the results
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
hist(df_clean$ftfcrmrte,
     main="Face-to-Face Crime",
     xlab= "Crimes Committed/Person",
     ylab= "Number of Counties", cex.axis=.75)
df_clean$crmrte_ratio <- 1-1/(df_clean$mix+1)
hist(df_clean$crmrte_ratio,
     main= "Face-to-Face Crime",
     xlab= "F2F Crimes/Total Crimes",
     ylab= "Number of Counties", cex.axis=.75)
```



We note that the distribution of the face-to-face crime rate is similar, but not identical to, the general crime rate. In particular, **f2rcrmrte** has a slightly lower mean of 3% (i.e., an average of 3 personal crimes per 100 people), and a range of 0.5% to 9.3% (i.e., 0.5-9.3 personal crimes per 100 people). As these numbers suggest, the similarity in **crmrte** and **f2fcrmrte** reflects that in most of North Carolina, face-to-face crime makes up 80% or more of all reported crime. In turn, this means that an analysis of crime rates across the state is, in many regards, also an analysis of face-to-face crime rates. For this reason, we choose to focus our modeling primarily on the more inclusive **crmrte** variable, but consider face-to-face crime in the context of policy recommendations.

4 Models and Assumptions

For analyses of this type, we use ordinary least squares (OLS) regression, also known as multiple linear regression, to determine what associations, if any, exist among the variables in the data set. We identify a variable to be explained – the "outcome" variable – and then perform analyses involving different combinations of "explanatory" (or "predictor") variables from the data set to see the degree to which those variables and combinations can predict the observed outcomes. Each set of calculations, involving the outcome variable and a specific combination of explanatory variables, is referred to as a "model". Models typically build on each other, starting with a few explanatory variables that are determined to be particularly important, and

are subsequently extended by adding more explanatory variables in order to increase the predictive value relative to the outcome variable.

In addition to the outcome and explanatory variables, each model contains some degree of statistical error. This error represents an unknown value of the difference between the true value of the variables in the world and the values for those variables given in the data set, combined with the unknown influence of variables that are not in the data on the variables that are included. For example, there is some difference between the actual crime rate across counties in North Carolina and the measures we have for that crime rate in the current data set, because the data set does not reflect every single crime that was committed in every single county across the entire state during the time period of interest, or every single variable that could have influenced the occurrence of crime. To the extent that the variables in the data set were well selected, designed, and/or implemented, and the values for those variables accurately and thoroughly gathered, the statistical error will be smaller. All statistical models involve some degree of statistical error.

Other statistical terms that may be useful for the reader to be familiar with include:

- "coefficient": A value multiplied by a variable (e.g., in 5x, 5 is a coefficient)
- "fitted value": A predicted value for a variable that is generated when trying to find the best fit for the value into a particular regression equation.
- "residual": The difference between the observed value of a variable and the predicted value for that same variable. Each data point has one residual.

For the current analysis, we model the variables contributing to crime rate across the counties of North Carolina in five stages, resulting in five models. An overview of each model is provided below.

- Model 1 includes only the variables we believe to be the main predictors of crime rate, guided by a framework of *means*, *motive*, and *opportunity*: (**crmrte**): percentage of young males in the population (**pctymle**), tax per capita (**taxpc**), and population density (**density**).
- 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 people who do not identify racially or ethnically as White (**pctmin80**).
- Model 3 builds on Model 2 by adding more information about the location of the county (**central**), the ratio of prison sentences to convictions (**prbpris**), and the average length of prison sentence (**avgsen**).
- Model 4 builds on Model 3 and adds all other explanatory variables in the data set that are not covariant with any explanatory variables already included.
- Model 5 explores whether predictors we identify for general crime rate are comparably effective for explaining the rate of face-to-face crime across counties in North Carolina by using the secondary outcome variable **f2fcrmrte**, based on **crmrate** and **mix**.

Each of our models will be assessed to determine its consistency with the following assumptions, which are standard for classic linear regression models. The statistical quality of our findings is dependent on these assumptions being met.

- 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.

For each of our models, we expect a classical linear model meeting Assumption 1; this is supported by the nature of our variables and the type of analysis we employ.

We also expect all of our models to meet Assumption 2, regarding random sampling, given that our data set reflects 90 of North Carolina's 100 counties, which is very close to the overall population. As a caveat, we note that 9/10 of the omitted counties are those for which 2012 population estimates are in the lowest 15% of population totals for the state. Although it is possible that these counties could have been omitted via an appropriately executed random sampling procedure, the pattern inherent in these omissions may require additional explanation or analysis, particularly if population density emerges as a useful predictor. It is relevant to note that in Northeast cities in the early 1980s, the rate of violent crime, but not property crime, was correlated with population density (see https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=99314).

Assumptions 3-6, which are dependent on the explanatory variables involved, will be tested independently for each model. Related to this, we note that, in some cases, it may be necessary to mathematically transform some portion of the data (e.g., convert the values for an explanatory variable to their logarithmic form) in order to facilitate an assumption being met. For example, transformations are commonly used to linearize the relationship between variables, improve homoskedasticity, or make the data more consistent with expected practice in a given scientific discipline. Beyond cases that have specific statistical or theoretical motivations, we will use the data in its original form.

For each fit model, we also calculate the statistical significance of each of its coefficients. This allows us to use our models for hypothesis testing. The null hypothesis for each model coefficient, β_j, is $\beta_j = 0$, meaning that the predictor variable, x_j, has no association with the outcome variable. High statistical significance, defined as t-values with probabilities of less than .05, allow us to reject the null hypothesis, and state that the given predictor has a statistically significant relationship with the outcome. Although our models may fulfill the assumption of homoskedasticity, we use heteroskedasticity-robust standard errors as a conservative best practice.

4.1 Model 1

We create our initial model guided by the concept of *means*, *motive*, and *opportunity*, which is a concise summary of the elements that typically need to be present for an individual to be found guilty of a crime. We extend this concept to the community level, proposing that communities that have more means, motives, and opportunities for crime to occur will also be those with higher crime rates.

We conceptualize the relationship between our tripartite framework and key variables from our data set as follows:

- Means: The ability and/or resources to commit a crime and likely avoid detection. We associate this element of our framework with **pctymle**, or the percentage of young males between 15-24. These are the members of the community who are most likely to have physical abilities relevant to committing crime, such as strength, speed, and agility. Where there are higher percentages of young males, we expect higher crime rates.
- Motive: Reason(s) to commit a crime, particularly against a given target. We associate this element of the framework with **taxpc**, or the tax rate per capita. As crime often involves an involuntary redistribution of resources, we anticipate the availability of resources – and perceived inequity thereof – to encourage criminal activity. Higher tax rates are typically associated with higher incomes, and we anticipate seeing more criminal activity where tax rates are higher.
- Opportunity: Occasions when crime can occur, especially when the risk of being caught is low. We associate this element of the framework with **density**, or the concentration of people in a given area.

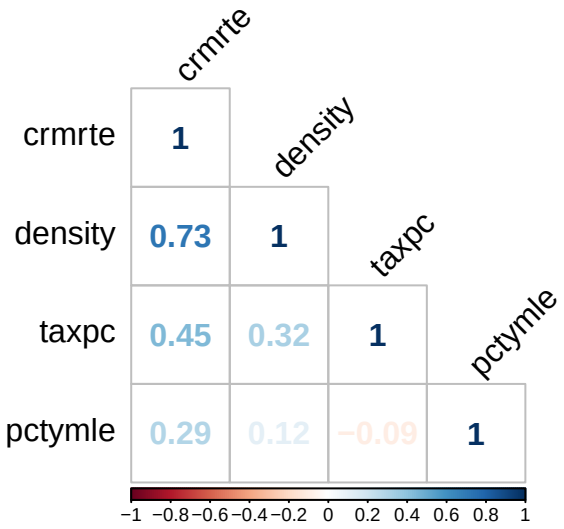


Figure 1: Correlation grid comparing crime rate, density, tax revenue per capita, and percent young male.

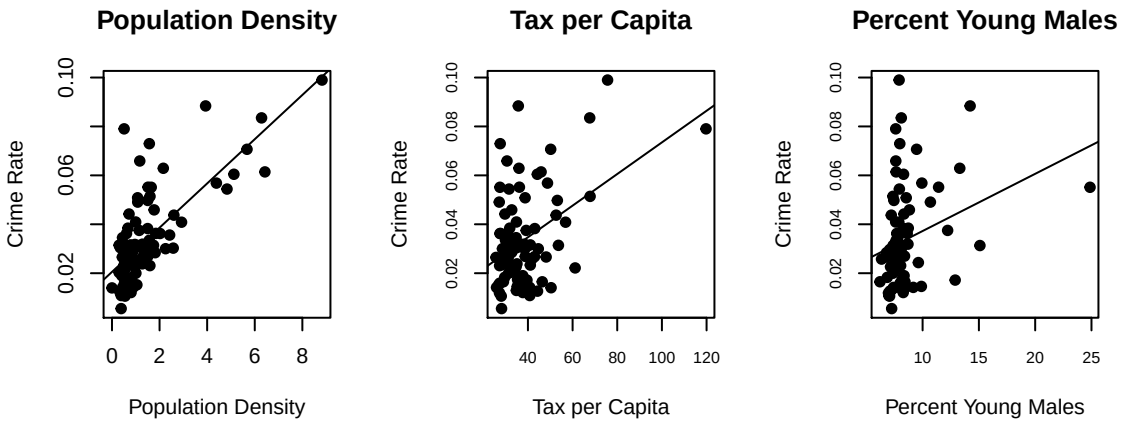
The more people live in a place, the more chances there are for conflicts to develop, inequities to be observed, and self-protective subgroups (e.g., "us" and "them") to form. Moreover, with increased population comes increased likelihood of anonymity, which reduces the likelihood of identification during criminal activity and potentially lowers the social barriers for committing crime. As such, we expect areas with higher concentrations of people to have higher crime rates.

4.1.1 Evaulating the Variables

An exploratory examination of the data suggests that this conceptualization of means, motive, and opportunity is useful with regard to crime rates in North Carolina. All three of our selected variables are strong predictors of the general crime rate, and shown in the correlation grid and scatterplots below.

```
m1vars <- df_clean[, c(3,9,10,25)]
cormat1 <- cor(m1vars)
#cormat1
?corrplot
corrplot(cormat1, type = "lower", method = "number",
         tl.col = "black", tl.srt = 45, cl.cex=.6)
```

```
# scatterplots of Model 1 variables
attach(df_clean)
layout(matrix(c(1,2,3), 1, 3, byrow = TRUE))
plot(density, crmrte, main = "Population Density",
     xlab = "Population Density", ylab = "Crime Rate", pch = 19)
abline(lm(crmrte~density), cex.axis=.8)
plot(taxpc, crmrte, main = "Tax per Capita",
     xlab = "Tax per Capita", ylab = "Crime Rate", pch = 19, cex.axis = .8)
abline(lm(crmrte~taxpc))
plot(pctymle, crmrte, main = "Percent Young Males",
     xlab = "Percent Young Males", ylab = "Crime Rate", pch = 19, cex.axis=.8)
abline(lm(crmrte~pctymle))
```



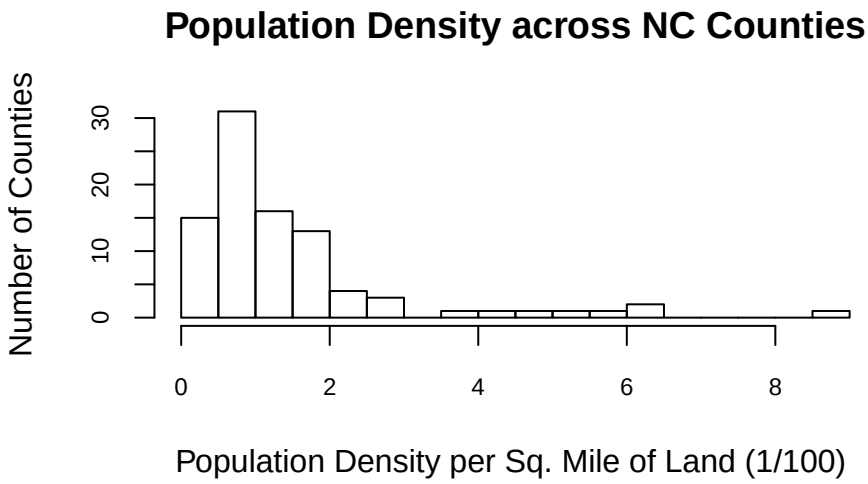
We begin by evaluating and describing each of these predictor variables in turn.

Population density (**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 Density across NC Counties",
     xlab= "Population Density per Sq. Mile of Land (1/100) ",
     ylab= "Number of Counties",
     breaks = 15, cex.axis = .75)
```



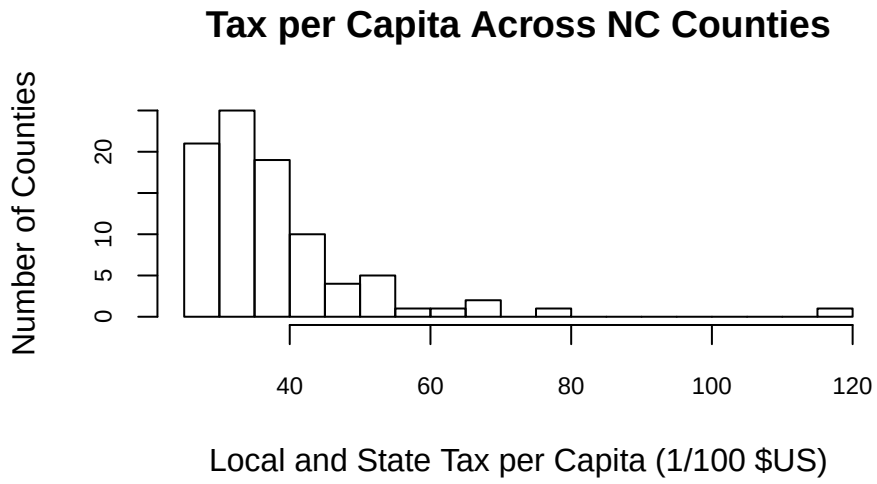
The value of **density** ranges from a score of approximately 0.002 to 880 people per square mile of land, with a mean of 145. The distribution of county densities is right skewed, with most counties having a score of 200 or fewer people per square mile of land. We note that 9/10 counties omitted from the data set are among those with lower populations, so it is likely that our descriptive statistics are somewhat elevated.

Tax per Capita (**taxpc**):

```
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 $US)",
     ylab= "Number of Counties",
     breaks = 30, cex.axis=.75)
```



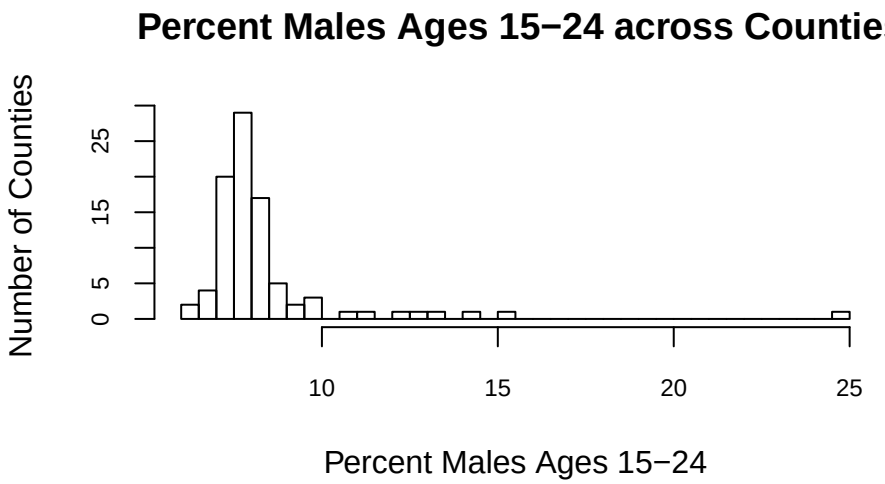
The cumulative value of taxes assessed at the local and state levels per capita ranges from $2,569 to $11,976 per year. Once again, we see a distribution that is right skewed, with revenue in most counties below the mean of $3,813 per year. The maximum value, the value for Dare county (FIPS 55) is nearly 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 tourism industry and features a number of the state's most 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 county can be explained by taxes on activities related to tourism, such as those appended to hotel, rental car, and park entrance fee costs. As such, Dare county generates substantial tax-related income from people who do not live in the county, but the value of these taxes is still incorporated into the tax per capita rate for those who *do* live in the county.

Percentage of Young Males (**pctymle**):

```
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, cex.axis=.75)
```



Across the counties of North Carolina, the percentage of males between 15-24 years of age ranges from 6.2 to 24.9. Once again, we see a distribution that is right skewed, with the majority of counties having fewer than the mean of 8.4% young males. There is one extreme value: that for Onslow county (FIPS 133). This reflects that Onslow county includes the city of Jacksonville, which contains the United States Marine Corps' Camp Lejeune and the New River Air Station, both of which are inhabited predominately by males between 18-25 years of age.

4.1.2 Evaluating Assumptions

Before we build our model, we check the correlations between our independent and predictor variables to assess the variables' consistency with Assumption 3 regarding no perfect collinearity. We can reference the correlation matrix above to see that all the correlations among our variables are less than 1. Specifically, **crmrte** is positively predicted, in decending order, by **density** (r = .73), **taxpc** (r = .45), and **pctymle** (r = .29). Additionally, **density** shows a strong positive correlation with **taxpc** (r = .32) and a weaker positive correlation with **pctymle** (r = .12). None of the variables are perfectly collinear with any of the others, so Assumption 3 is validated.

Given the skewed nature of **crmrte**, we explore a log transform of the dependent variable and build two versions of Model 1: one with untransformed **crmrte** and one with a log transform of **crmrte**. A comparison of the two versions and their consistency with the statistical assumptions will allows us to select the better foundational model.

```
# Build Model 1 (untransformed) and Model 1 (log)
(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.squared
```

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

```
(model_1_log = lm(log(crmrte) ~ density + taxpc + pctymle, data = df_clean))
```

```
##
## Call:
```

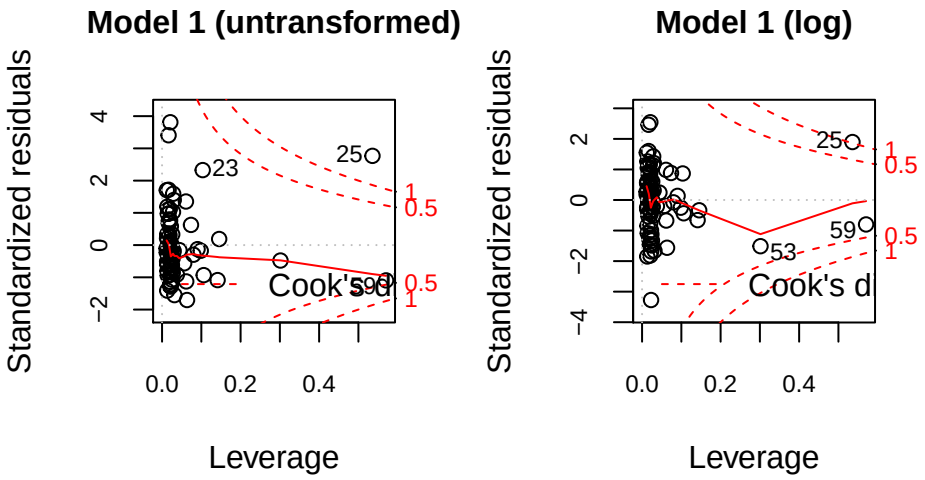


Figure 2: Residuals vs Leverage Plots for candidate versions of Model 1.

```
## lm(formula = log(crmrte) ~ density + taxpc + pctymle, data = df_clean)
##
## Coefficients:
## (Intercept)      density        taxpc      pctymle
##   -4.613815     0.194629     0.008667     0.054976
```

```
summary(model_1_log)$r.squared
```

```
## [1] 0.4811686
```

The version of the model with an untransformed **crmrte** has an the adjusted R^2 value which indicates that 62.79% of the variance in crime rate can be explained by our combination of three variables.

In contrast, in the version of the model with the log transformed **crmrte**, the adjusted R^2 has dropped to 46.31%, reflecting a substantial reduction in predicted variance.

Next, we evaluate the Cook's Distance for the residuals.

```
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(model_1, which= 5,
     caption = "", main = "Model 1 (untransformed)", cex.axis = .75, cex.main = 1)
plot(model_1_log, which = 5,
     caption = "", main = "Model 1 (log)", cex.axis = .75, cex.main = 1)
```

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

Assumption 4, the assumption that the mean error conditional on the dependent variable equals zero, is not met in a strict causal sense for either version of Model 1. That is, we cannot observe "true relationships" among the underlying quantities, due to the likely presence of other factors that may influence the outcome variable. For this reason, we can only validate a weaker, but still sufficient, form of the assumption: exogenity of the error. This means that our interpretations of the relationships among the variables in our data will be strictly associative. From this point forward, Assumption 4 and Assumption 6 will be discussed in this more limited associative manner.

The plots below contrast the residuals for the two version of Model 1 with their corresponding fitted values. We see that the untransformed version shows close to a zero-conditional mean for the errors, whereas the

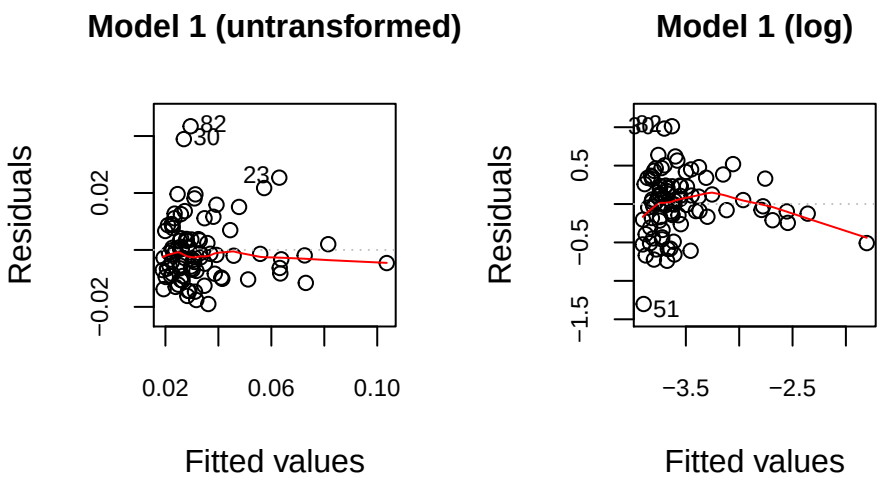


Figure 3: Residuals vs Fitted Values Plots for candidate versions of Model 1.

log transformed version has a marked curvature in the errors. This outcome preferences the untransformed version of the model.

```
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(model_1, which= 1,
     caption = "", main = "Model 1 (untransformed)", cex.axis = .75, cex.main = 1)
plot(model_1_log, which = 1,
     caption = "", main = "Model 1 (log)", cex.axis = .75, cex.main = 1)
```

To validate the model in terms of Assumption 5 regarding homoskedasticity, we plot the residuals vs. fitted values for each version. For the untransformed version, we find that the range of errors is relatively constant throughout the range of fitted values. For the transformed model, however, the range of errors is less constant. We note that there are fewer data points at the higher values of **crmrte** than the lower values, so validity in this case may be somewhat weaker than with other assumptions. Nonetheless, this outcome also preferences the untransformed version of the model.

```
## Emily comment: is the block of text above describing these plots or the one above it? currently no cl
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(model_1, which= 3,
     caption = "", main = "Model 1 (untransformed)", cex.main = 1, cex.axis = .75)
plot(model_1_log, which = 3,
     caption = "", main = "Model 1 (log)", cex.main = 1, cex.axis = .75)
```

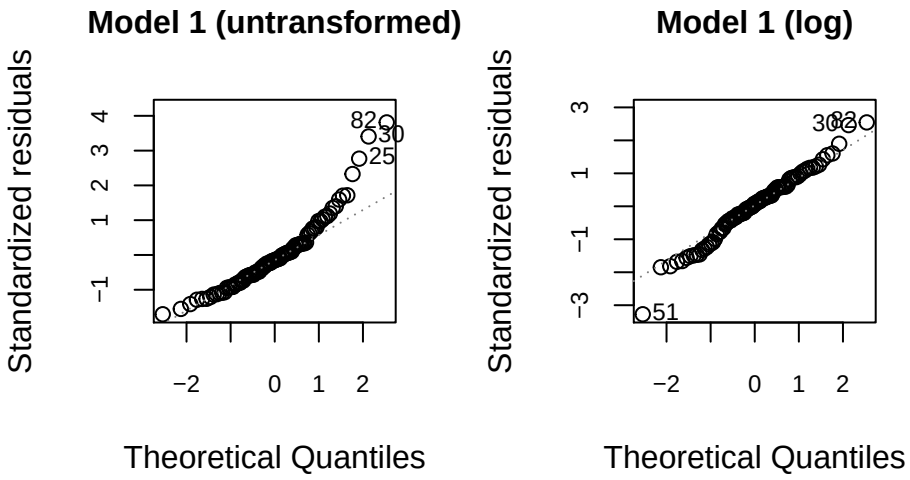
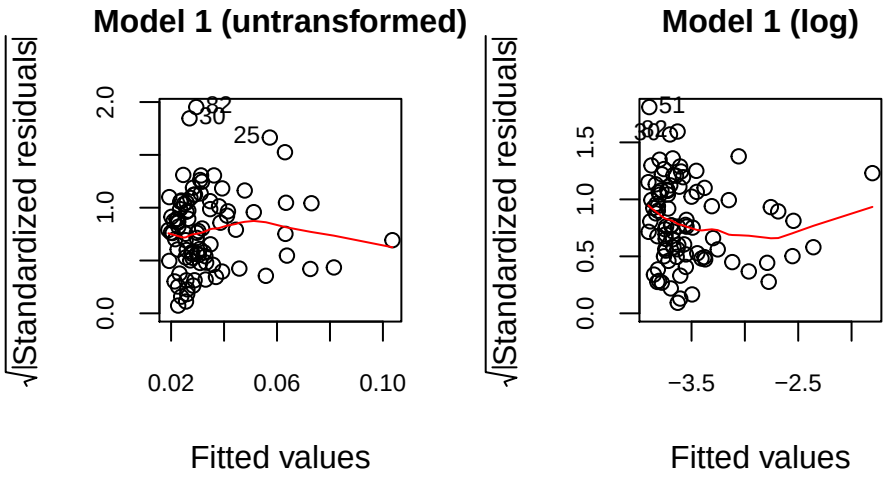


Figure 4: Theoretical quantiles vs standardsized residuals for candidate versions of Model 1.



To validate Assumption 6, the normality of the residuals, we examine a Q-Q plot of the standardized residuals for each version of the model. In general, we observe a fairly straight line for the exceptions of the tails in both versions, with the untransformed model showing substantially more deviation from the fit line at higher quantiles. Normally distributed residuals, as evidenced by consistency to the fit line, indicate that variation in the outcome variable not predicted by the model is likely due to random noise. In this case, the pattern for both versions of the model suggest that there are explanatory variables outside of Model 1 that influence crime rate, and these other variables are biasing our coefficients. Log transformation is specifically intended to normalize the distribution of residuals, so it is expected that this evaluation would preference the log transformed version of the model; it does.

```
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(model_1, which= 2,
     caption = "", main = "Model 1 (untransformed)", cex.main = 1, cex.axis = .75)
plot(model_1_log, which = 2,
     caption = "", main = "Model 1 (log)", cex.main = 1, cex.axis = .75)
```

Overall, our comparison of untransformed and transformed versions of Model 1 suggests that the untransformed version is superior with regard to meeting assumptions of the classic linear model and demonstrating predictive power. That is, although the log transformed version better addresses the demands of Assumption 6, it also appears to violate Assumptions 4 and 5. In conjuction with the reduction in adjusted R^2 from 63% (untransformed) to 48% (log transformed), we prefer the version of Model 1 without a log transform of

crmrte; this version of the model is summarized in **Table 2**.

4.1.3 Statistical Inference for Model 1

The table below gives us the heteroskedasticity-robust standard errors, t-values, and p-values for the parameters in Model 1. The coefficient on **density** is statistically significant at the .0001 level; the one for **pctymle** is statistically significant at a .05 level. As such, we can reject the null hypotheses that 1) there is no association between **density** and **crmrte** and 2) there is no association between **pctymle** and **crmrte**. We fail to reject the null hypothesis that there is no association between **taxpc** and **crmrte**.

Our interpretation of Model 1 will consider both the practical and statistical significance of each coefficient.

```
coeftest(model_1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -0.00910806  0.01266244 -0.7193    0.4739
## density      0.00759718  0.00124324  6.1108 2.805e-08 ***
## taxpc        0.00039646  0.00029917  1.3252    0.1886
## pctymle      0.00197340  0.00082687  2.3866    0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.1.4 Interpreting Model 1

In Model 1, shown in Table 2, the coefficients are all positive. This indicates that as the population density, tax per capita, or the percentage of young males increase, an associated increase occurs in the crime rate. We interpret the coefficient on a predictor as describing the magnitude of the associated change in **crmrte** given a one-unit increase in the predictor, holding everything else equal. Our model predicts that a one-unit increase in **density** (an additional hundred people per square mile of land) is associated with approximately a .0076 increase in the **crmrte**, holding all else equal. The same interpretation can be applied to the other variables: a one-unit (hundred dollar) increase in **taxpc** is associated with approximately a .0004 increase in **crmrte**, and a one-unit (one percent) increase in **pctymle** is associated with a .002 increase in **crmrte**, holding all else equal in both cases.

As the coefficients show, not all of our predictor variables are equally influential 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 about four times that generated by higher percentages of young males. The statistical significance of the coefficients maps neatly to their size; **density**'s coefficient is the highest in magnitude and has the highest statistical significance, and **taxpc**'s coefficient is the lowest in magnitude and has the lowest statistical significance. As such, this model indicates that, although 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 the proportion of young males, then by the tax revenue per capita. We will address each of the variables in turn, highlighting possible omitted variables related to each in **bold**.

To begin, the **density** of an area contributes to the *opportunity* individuals have to commit crime, and there are a number of reasons why increases in population density could facilitate increases the crime rate. For example, the more people live in a particular space, the more opportunities there are for them to come into conflict with one another, to interact with others who have different access to desirable resources, and to be unfamiliar with the 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 opportunities for conflict and perceptions of inequity. For example, **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 management, 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 inequity**, such as access to adequate educational and employment opportunities, social support programs, and affordable housing should also result in a reduction in crime. Last but not least, there is the issue of anonymity. 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 recognized and/or 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 opportunities 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 **pctymle** in an area is connected with the proportion of the population that have the physical *means* to commit crime. There is certainly immense social pressure linking 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 prioritize traditional social values, which is likely to be true in a historically conservative state such as North Carolina. To the extent that a candidate's constituency includes communities with large percentages of young men, it could be fruitful to consider the role that local **culture** contributes to young men committing crimes and how providing alternative (i.e., 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 activities that contribute positively to the community and encourage them to develop rather than damage it. Religious institutions could play an important role in this effort, as it is common for religious doctrine across faiths to discourage violence against others and other types of criminal behavior.

The **taxpc** in an area is connected to potential *motive* for committing crime, particularly in socioeconomically diverse communities where some "have" – sometimes to a conspicuous degree – and others "have not." 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 desirable items and services. This suggests developing policies that address socioeconomic disparity should reduce the impetus for individuals with limited access to resources to engage in criminal activity to secure basic needs 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."

Another policy direction suggested by our analysis relates to **tourism**. In areas where there is a lot of tourism, there may be special circumstances that contribute to the crime rate. If tourism is seen by a community as an opportunity for good employment and additional funding for community infrastructure and beneficial social programs, we would expect this to lower the crime rate, particularly against visitors. If, on the other hand, it is viewed as an influx of distracted strangers with an abundance of extra money and little familiarity with the local environment, the crime rate would likely go up. Tourism contributes positive to the economy of North Carolina and reduces the direct tax burden on residents, while encouraging improvements in infrastructure and community facilities. To the extent a candidate can frame tourism and the support thereof as a strong positive for the their constituency (e.g., as one means to reduce local socioeconomic disparity), constituents are less likely to want to engage in criminal activity that negatively influences the tourism industry.

4.2 Model 2

For Model 2, we include additional variables that seem to fit into our means, motives, and opportunities framework, albeit less neatly, and that have substantial correlations with the variable of interest, crime rate.

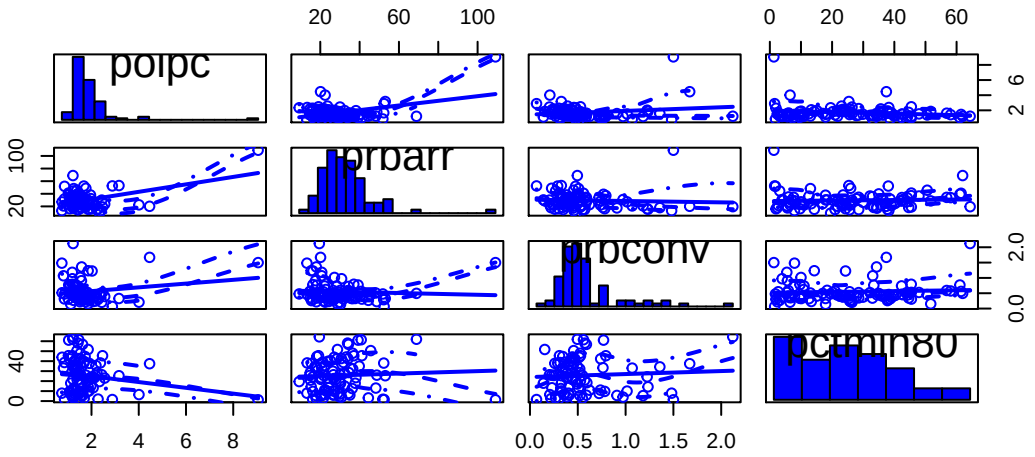


Figure 5: Scatterplot matrix showing relationships between **polpc**, **prbarr**, **prbconv**, and **pctmin80**.

We expect **prbarr**, **prbconv**, and **polpc** to detract from the opportunity to commit crime, as it seems feasible that the more likely one is to encounter police, or get arrested or convicted, the less likely one would be to commit crime. We also anticipate some relationship with **pctmin80**, perhaps because there is more opportunity for crime in a more diverse neighborhood, as there may be social friction caused by different racial groups living in close proximity. It is challenging to slot this variable neatly into the means, motives, and opportunities framework, due to the potential omitted variables that we think might be correlated with this one, which we will discuss in greater detail in the interpretation section. We also include the **west** variable, noting its strong correlation with **crmrte**, but without a clear theory as to what might explain that relationship; we will explore potential cultural factors and omitted variables related to **west** in the interpretation section.

We summarize the relationships among the new explanatory variables, with the exception of **west** (coded effectively as "in the western part of the state" or "not in the western part of the state"). For the data set as a whole, 24.4 % of all counties were coded as being **west**.

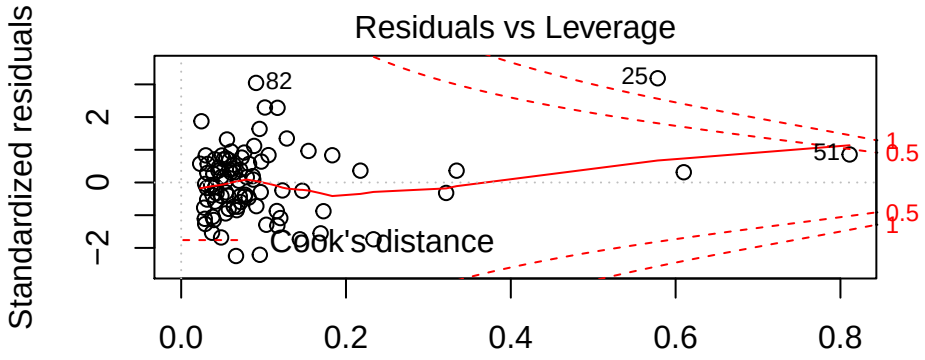
```
vars <- c("polpc", "prbarr", "prbconv","pctmin80")
suppressWarnings(scatterplotMatrix(df_clean[,vars],
                                   diagonal = list(method = "histogram")))
```

The matrix plot shows little to no collinearity among the considered variables, validating Assumption 3.

```
# 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
```

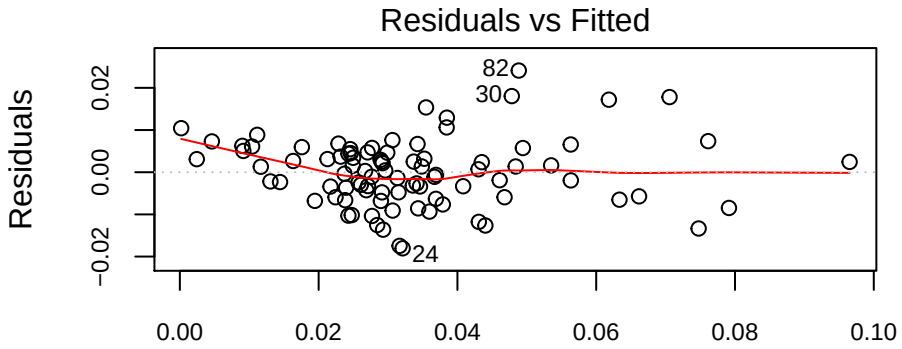
```
plot(model_2, which = 5)
```



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.

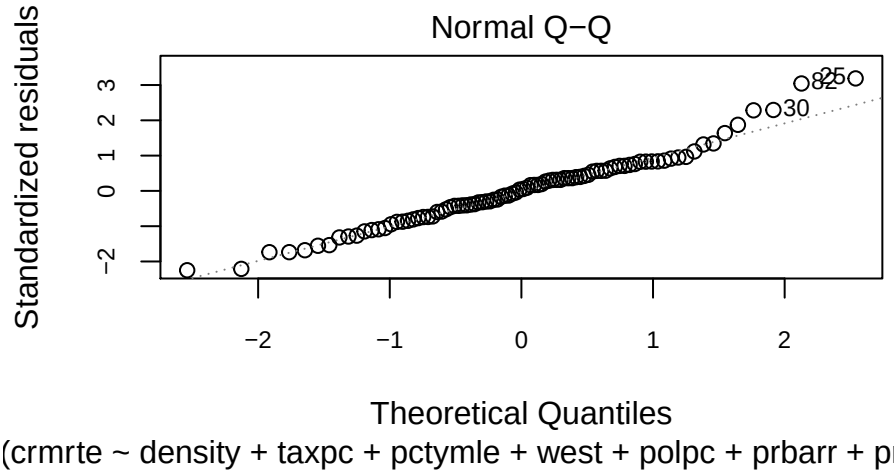
Assumption 4, exogeneity, is verified below, by checking the plot of residuals against fitted values. We see a fairly flat mean across the entire range of fitted values.

```
plot(model_2, which = 1, cex.axis = .75)
```



Assumptions 5 and 6 were validated for this model as they were for Model 1. We note that the Q-Q plot of the standardized residuals appears much closer to linear than for Model 1, indicating that we likely have most of the significant sources of variation described by Model 2.

```
plot(model_2, which= 2, cex.axis = .75)
```



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.

Statistical Inference for Model 2

```
coeftest(model_2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.8840e-02  8.1882e-03  2.3009 0.0239671 *
## density      5.4313e-03  1.3879e-03  3.9134 0.0001886 ***
## taxpc        1.7816e-04  2.3922e-04  0.7447 0.4585871
## pctymle      8.7469e-04  3.9818e-04  2.1967 0.0309008 *
## west        -1.6253e-03  2.5319e-03 -0.6419 0.5227198
## polpc        6.5622e-03  2.0968e-03  3.1296 0.0024336 **
## prbarr      -5.4952e-04  1.2699e-04 -4.3272 4.279e-05 ***
## prbconv     -1.8753e-02  3.9228e-03 -4.7806 7.702e-06 ***
## pctmin80     3.2930e-04  7.8636e-05  4.1876 7.123e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The heteroskedasticity-robust t test for Model 2 shows that all of the new variables we added to the model, other than **west**, are statistically significant at a .05 level (and even at a .01 level). The low p-values of **polpc**, **pbarr**, **pbconv**, and **pctmin80** allow us to reject the null hypotheses that each of those variables has no association with **crmrte**. Like with Model 1, we need to rely on both statistical significance and practical significance in order to interpret the model's coefficients.

Interpreting Model 2

Like with model 1, we interpret the coefficient on a given predictor as representing an increase the size of the coefficient on y for a one-unit increase in the predictor. The associations between the predictors and the crime rate 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 are some omitted variables associated with crime that are more prevalent in Western counties or those with larger non-white minority populations. Examples of these are socioeconomic status, household income,

and education rate, all of which are likely to be significantly different for minorities than whites in a Southern state at the end of the 20th century. Also, western North Carolina has a thriving tourism sector, due to the attractive nature, and we suspect that **tourism** is a relevant omitted variable.

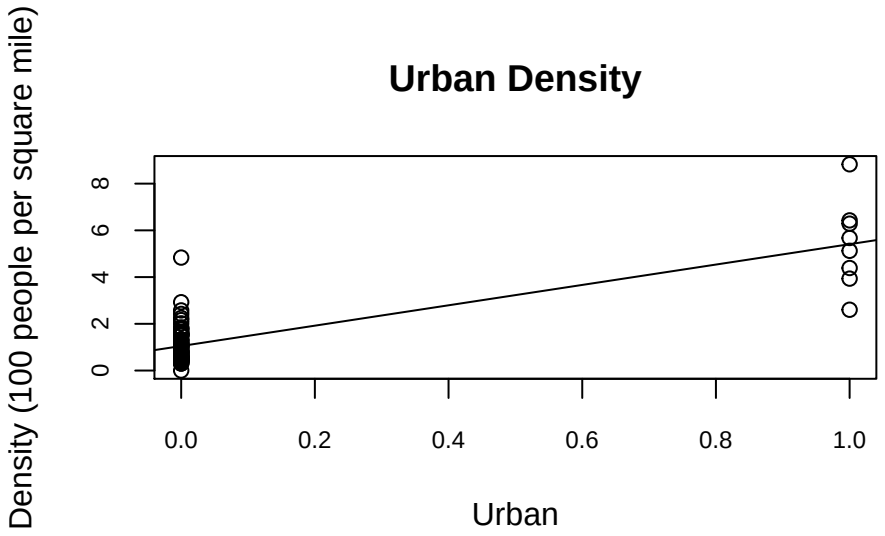
Additionally, the positive association between police per capita and crime is noteworthy. This association should be studied further, ideally with causal analysis, as there are plausible causal theories going in either direction. Perhaps heightened police presence creates an antagonistic relationship between officers and citizens, which leads to a distrust of authority and an increase in crime; the ideal way to test that would be to find counties with similar crime rates and other demographics where one county changes a policing policy and the other one does not, a natural paired experiment. However, it also seems possible that a county that experiences more crime would choose to up the size and activity 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 scatter plot:

```
plot(df_clean$urban , df_clean$density,
     main= "Urban Density",
     ylab= "Density (100 people per square mile)",
     xlab= "Urban", cex.axis = .75)
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.

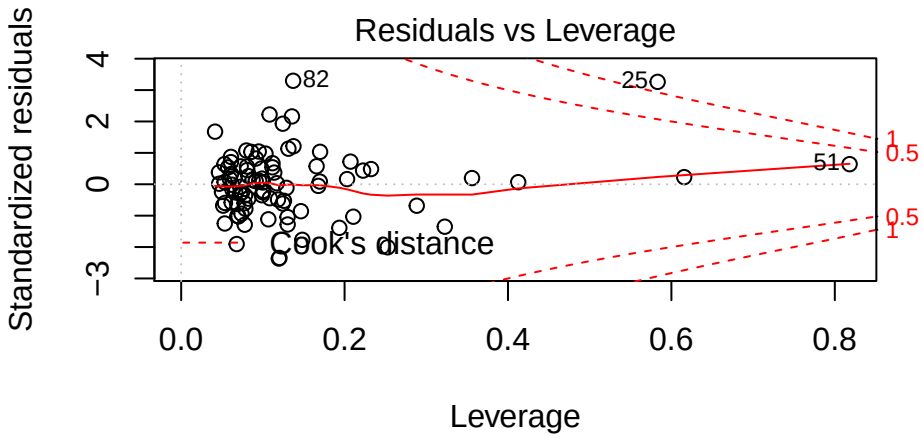
With that, we build Model 3:

```
# 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
```

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



We note that point 25, representing Dare county, 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:

```
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.

Statistical Inference and Interpretation for Model 3

```
coeftest(model_3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  2.4830e-02  9.8357e-03  2.5245  0.013617 *
## density      5.8503e-03  1.4385e-03  4.0670  0.000113 ***
## taxpc        1.4705e-04  2.4494e-04  0.6004  0.550007
## pctymle      7.5521e-04  3.6326e-04  2.0790  0.040909 *
## west        -4.9285e-03  3.4884e-03 -1.4128  0.161693
## polpc        6.8611e-03  2.0912e-03  3.2809  0.001548 **
## prbarr      -5.4372e-04  1.3004e-04 -4.1812 7.520e-05 ***
## prbconv     -1.8259e-02  4.1137e-03 -4.4385 2.942e-05 ***
```

```
## pctmin80     2.6353e-04  8.9263e-05  2.9523  0.004166 **
## central     -3.9980e-03  2.6039e-03 -1.5354  0.128735
## avgsen      -2.9651e-04  3.7968e-04 -0.7809  0.437200
## prbpris      3.7196e-05  1.3020e-04  0.2857  0.775875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our heteroskedasticity-robust t-test shows that none of the added variables are statistically significant at a .05 level. This means we cannot reject the null hypothesis that each variable in **central**, **avgsen**, and **prbpris** has no association with **crmrte**. For this reason, in confunction with the low practical significance of the small coefficients, we opt not to make policy recommendations based on these variables.

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 for policy decisions. However, it is interesting to note that the R^2 rises to 0.85, which is not much higher than Model 2. Additionally, many points exceeding a Cook's distance of 1 are observed.

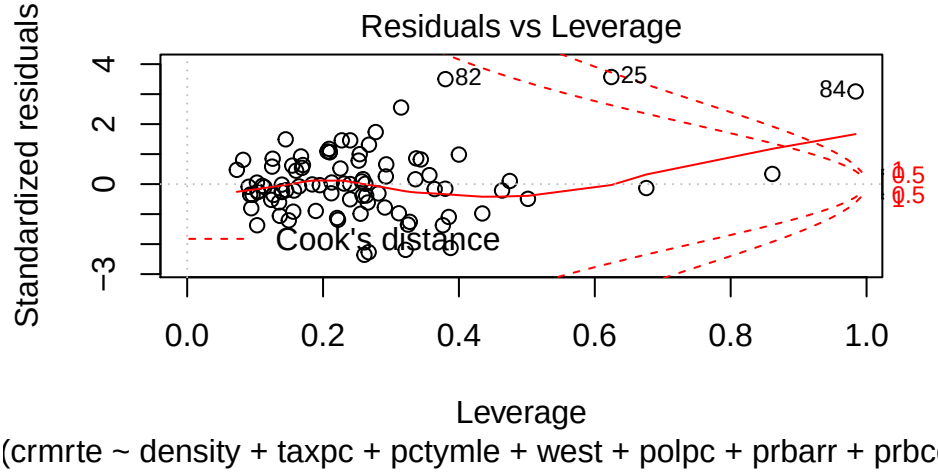
```
# Build Model 4
# model 4: kitchen sink. urban, wage.
model_4 = lm(crmrte ~ density + taxpc + pctymle
             + west + polpc + prbarr + prbconv + pctmin80
             + central + avgsen + prbpris
             + urban + wcon + wtuc + wtrd + wfir + wser
             + wmfg + wfed + wsta + wloc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



(crmrte ~ density + taxpc + pctymle + west + polpc + prbarr + prbcor

Statistical Inference and Interpretation for Model 4

```
coeftest(model_4, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.3853e-02  3.0755e-02  0.4504 0.6538622
## density      5.3314e-03  1.4895e-03  3.5793 0.0006464 ***
## taxpc        1.6240e-04  2.8408e-04  0.5717 0.5694537
## pctymle      1.0125e-03  4.7826e-04  2.1170 0.0379748 *
## west        -2.5652e-03  4.4698e-03 -0.5739 0.5679579
## polpc        6.9679e-03  2.9536e-03  2.3591 0.0212406 *
## prbarr      -5.1466e-04  1.5689e-04 -3.2805 0.0016467 **
## prbconv     -1.8633e-02  6.5853e-03 -2.8295 0.0061464 **
## pctmin80     3.2542e-04  1.3849e-04  2.3497 0.0217429 *
## central     -4.2416e-03  3.7423e-03 -1.1334 0.2610725
## avgsen      -3.9858e-04  5.5361e-04 -0.7200 0.4740570
## prbpris      3.1727e-05  1.3586e-04  0.2335 0.8160642
## urban       -9.6498e-05  8.2752e-03 -0.0117 0.9907307
## wcon         2.3025e-05  3.2876e-05  0.7004 0.4861334
## wtuc         6.1914e-06  1.9862e-05  0.3117 0.7562178
## wtrd         2.8767e-05  8.7294e-05  0.3295 0.7427756
## wfir        -3.5455e-05  3.5699e-05 -0.9932 0.3242068
## wser        -1.7158e-06  9.9447e-05 -0.0173 0.9862856
## wmfg        -8.9675e-06  1.7469e-05 -0.5133 0.6094087
## wfed         2.9075e-05  3.7780e-05  0.7696 0.4442480
## wsta        -2.2302e-05  3.6828e-05 -0.6056 0.5468431
## wloc         1.4456e-05  8.5367e-05  0.1693 0.8660410
## mix         -1.8693e-04  2.2922e-04 -0.8155 0.4176761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As with Model 3, we see that none of the new variables added in this model are statistically significant at a .05 level. However, it is worth noting that the statistical significance of several of our Model 1 and 2 variables have decreased as a result of these new variables being in the model. While **polpc**, **prbarr**, **prbconv**, **pctmin80**, **pctymle**, and **density** all have coefficients that remain statistically significant in this model, the absolute values of each of their t-scores is lower than it was in prior models, which suggests that some of the effect of these variables in previous models has been absorbed by the new variables.

Though we have many predictors in this model that are not statistically significant on their own, it is possible that they could have joint significance, which we can determine using an F-test. To do this, we create a restricted model with only our statistically significant covariates, and compare this restricted model to the unrestricted Model 4 to see if the combined effect of Model 4's individually insignificant variables is statistically significant.

```
restricted_model <- lm(crmrte ~ density + pctymle + polpc
                       + prbarr + prbconv + pctmin80, data = df_clean)
waldtest(model_4, restricted_model, vcov=vcovHC)
```

```
## Wald test
##
## Model 1: crmrte ~ density + taxpc + pctymle + west + polpc + prbarr +
##     prbconv + pctmin80 + central + avgsen + prbpris + urban +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     mix
```

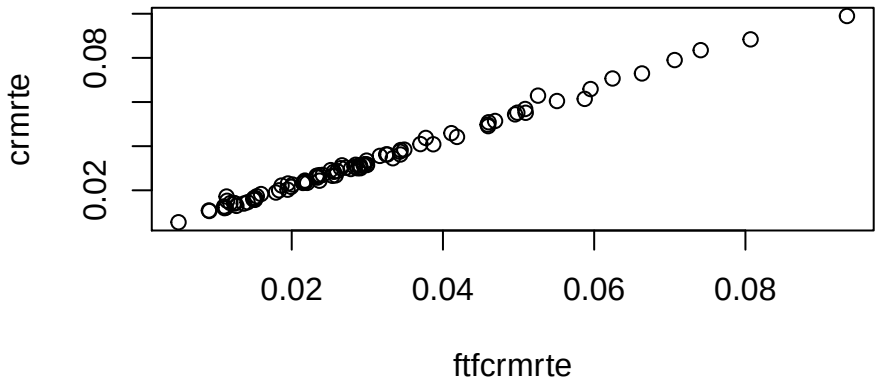


Figure 6: Face-to-Face Crime Rate vs Crime Rate.

```
## Model 2: crmrte ~ density + pctymle + polpc + prbarr + prbconv + pctmin80
##   Res.Df  Df      F Pr(>F)
## 1     67
## 2     83 -16 0.8621 0.6131
```

Our F-test shows that even the combined statistical significance of the 16 additional variables in the unrestricted model is not enough to argue for the joint statistical significance of these additional variables. With a p-value of 0.6131, we are far from the 0.05 critical value we would need to reject the null hypothesis that these additional values have no association with crime rate.

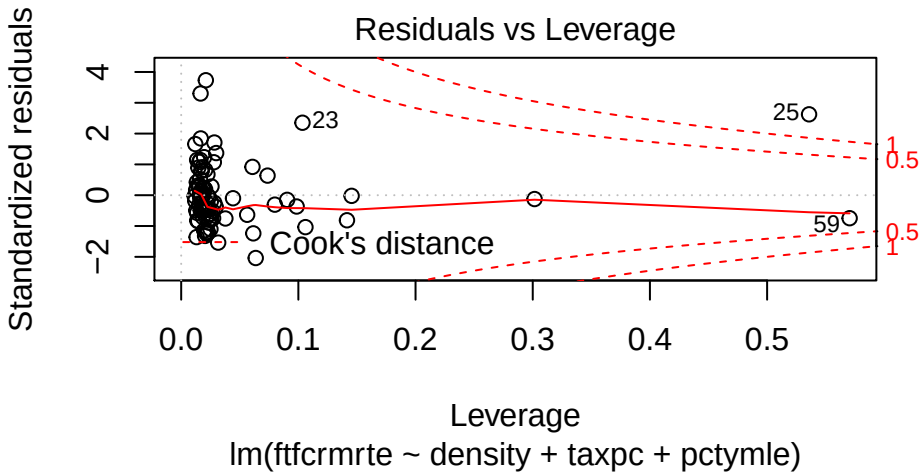
4.5 Model 5

We now return to our face-to-face crime rate and attempt a model using it as the dependent variable. As previously noted, the purpose of this model is to determine the extent to which Model 1, intended to predict the general crime rate, effectively extends to predict the face-to-face crime rate. Our expectation is the the results for Model 5 and Model 1 will be very similar, as **crmrte** and **f2fcrmrte** are closely associated:

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

Implementation of Model 5 shows that population density, tax per capita, and the percentage of young males in the population explain approximately 63% of the variation in the face-to-face crime rate; this is almost identical to their predictive power relative to general crime rate.

```
# 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
```

Statistical Inference and Interpretation of Model 5

```
coeftest(model_5, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -0.00731872  0.01078922 -0.6783  0.499380
## density      0.00698061  0.00107721  6.4803 5.497e-09 ***
## taxpc        0.00035398  0.00026346  1.3436  0.182615
## pctymle      0.00169822  0.00062913  2.6993  0.008364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The t test on coefficients for Model 5 closely resembles that of Model 1, with **density** and **pctymle** showing statistical significance at the .05 level and **taxpc** not demonstrating statistical significance (though it is, perhaps, worth noting that **pctymle** has greater statistical significance in this model than it did in Model 1). Given that we see similar practical and statistical significance for the covariates in this model that we do in Model 1, the answer to our corollary question, *What are the variables associated with the* ***face-to-face*** *crime rates across counties in North Carolina?*, appears to be "The same variables that are associated with the general crime rate, and the impact is of a comparable scale."

As noted previously, our experience indicates that constituents are interested in policies and programs intended to reduce crime, but are especially concerned with those directed at reducing personal crime. Model 5 demonstrates to the candidate that policies and programs intended to reduce the general crime rate, about which we have made a number of recommendations, are also likely to reduce the personal crime rate. What matters most in this case is constituents' *perception*: The candidate is likely to get more support for policies and programs that are framed in terms of reducing personal crime – particularly violent crime. Moreover, the candidate can promote policies and programs in terms of reducing personal crime while knowing that implementation of these policies should also reduce the general crime rate. This analysis provides a foundation for the candidate to pursue this strategy without sacrificing veracity or efficiency.

4.6 Model Summary

The models built above are summarized in **Table 2**. Of note, Model 2 appears to be the best balance of predictive ability and parsimony of these five judging by the highest adjusted R^2 and the lowest AIC value.

```
#TODO: Can anyone write a more efficient loop here for assigning these AICs?
model_1$AIC <- AIC(model_1)
model_2$AIC <- AIC(model_2)
model_3$AIC <- AIC(model_3)
model_4$AIC <- AIC(model_4)
model_5$AIC <- AIC(model_5)

stargazer(model_1, model_2, model_3, model_4, model_5, type = "latex", se = list(sqrt(diag(vcovHC(model
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("aic", "rsq", "n", "adj.rsq"),
          no.space = TRUE, table.placement = "!", star.cutoffs = c(0.05, 0.01, 0.001))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Aug 01, 2018 - 22:01:41

4.7 Future Statistical Tests

The data we have been provided are not ideal for statistical inference, because they do not comprise a simple random sample of any sufficiently large broader population. There are some issues with calling this a random sample of counties in North Carolina in 1987, both because we do not know if the criteria of their selection were truly random, nor their measurements independent (geographically close counties, for instance, could see mutually dependent crime rates, etc.) Moreover, the most commonly used statistical methods are designed for samples that comprise less than 10% of the population, whereas our data comprise 90% of the population they were drawn from. Our estimates would at best be extremely conservative, and at worst glaringly inaccurate.

Furthermore, since we have nearly sampled the entire population of North Carolinian counties in 1987, the usual business of statistical inference – trying to learn about a population based on a sample – is not really necessary in order to make meaningful claims about relationships between crime rate and other variables in North Carolina in 1987. Even supposing the relationships we discovered only hold for the 90 counties for which we have data (out of the 100 total in North Carolina), constituents will likely be persuaded that methods to reduce crime in these counties alone would be a sufficient goal in and of itself.

If we insist on proceeding with statistical inference on our current data, the best we can do is imagine them as a random sample of counties drawn from some theoretical infinite population of possible sets of measurements of each variable - although even then, we run the risk of clustering geographically similar counties. We could then infer confidence intervals for the regression between variables in this hypothetical population. There is a temptation to make inferences across multiple states or years, but in this case we must consider that the counties we have available are far from randomly selected.

Finally, since we have already used all of our data for exploratory data analysis, it would not be appropriate to conduct actual hypothesis testing on the data we have. Were we given new data from another year or another state, or even data that were drawn randomly from multiple years or states, we could test hypotheses based on our current findings (e.g., that change in population density has a nonzero effect on crime rate). Even better would be the possibility to implement policy changes in some North Carolina counties but not others, creating an experiment of sorts. However, we advise that the general public might not take kindly to this approach.

Table 2: 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***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.00001)
taxpc	0.0004	0.0002	0.0001	0.0002	0.0004***
	(0.0003)	(0.0002)	(0.0002)	(0.0003)	(−0.00000)
pctymle	0.002*	0.001*	0.001*	0.001*	0.002***
	(0.001)	(0.0004)	(0.0004)	(0.0005)	(−0.00000)
west		−0.002	−0.005	−0.003	
		(0.003)	(0.003)	(0.004)	
polpc		0.007**	0.007**	0.007*	
		(0.002)	(0.002)	(0.003)	
prbarr		−0.001***	−0.001***	−0.001**	
		(0.0001)	(0.0001)	(0.0002)	
prbconv		−0.019***	−0.018***	−0.019**	
		(0.004)	(0.004)	(0.007)	
pctmin80		0.0003***	0.0003**	0.0003*	
		(0.0001)	(0.0001)	(0.0001)	
central			−0.004	−0.004	
			(0.003)	(0.004)	
avgsen			−0.0003	−0.0004	
			(0.0004)	(0.001)	
prbpris			0.00004	0.00003	
			(0.0001)	(0.0001)	
urban				−0.0001	
				(0.008)	
wcon				0.00002	
				(0.00003)	
wtuc				0.00001	
				(0.00002)	
wtrd				0.00003	
				(0.0001)	
wfir				−0.00004	
				(0.00004)	
wser				−0.00000	
				(0.0001)	
wmfg				−0.00001	
				(0.00002)	
wfed				0.00003	
				(0.00004)	
wsta				−0.00002	
				(0.00004)	
wloc				0.00001	
				(0.0001)	
mix				−0.0002	
				(0.0002)	
Constant	−0.009	0.019*	0.025*	0.014	−0.007***
	(0.013)	(0.008)	(0.010)	(0.031)	(0.0001)
Observations	90	90	90	90	90
R^2	0.640	0.824	0.830	0.855	0.631
Adjusted R^2	0.628	0.807	0.806	0.807	0.619
Akaike Inf. Crit.	−542.122	−596.4438	−593.615	−585.586	−556.107

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

5. Omitted Variables

Although our Models 1 and 2 do provide some useful information that can inform government policy, there are a number of omitted variables that we did not have access to in this analysis that we suspect have meaningful associations 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 resources and those who have very little, the disparity can create social tension. Furthermore, if those who do not have resources live in close proximity to those who have many, those without may be motivated to engage in crime to secure the resources they need from those who have them. We could measure socioeconomic 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. This would diminish the effect of density, bringing the coefficient closer to zero.

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. This would lower the effect of percent young male, bringing the coefficient closer to zero.

Additionally, we anticipate that **educational attainment** 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 meaningfully 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. A negative omitted variable bias on a positive coefficient would only increase the coefficient, making the effect of density even greater.

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 be 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. As with mean education value, the negative omitted variable bias on a factor with a positive coefficient would only make the effect of density greater.

Infrastructure quality would be negatively correlated with crime since areas with well developed infrastructure leave people less desperate to commit crime. We would expect a positive correlation between infrastructure quality and tax per capita since taxes are used to fund the infrastructure development. Therefore, the omitted variable bias for infrastructure quality is negative, and the fitted values for a given value of tax per capita would be higher if we controlled for infrastructure quality.

We expect **community membership** to be negatively correlated with crime since a sense of belonging is presumed to reduce the desire to commit crime. We could expect a negative correlation between community membership and the percent of young males since young males are typically moving through a city and not planting roots. Perhaps the more obvious correlation is a negative one with density, since community membership is harder when there are too many people to know. Either way, the omitted variable bias for community membership would be positive. We expect that the fitted values for a given value of percent

young male or population density could be higher if we controlled for community membership, making the effect of percent young male more extreme.

Further, we believe **negative police interaction** to be positively correlated with crime, and positively correlated with police per capita, leading to a positive omitted variable bias. This leads to fitted values for a given value of police per capita to be lower, if we controlled for negative police interaction, making the effect of police per capita less extreme.

Tourism is a surprising, but very interesting, omitted variable. We noted it by way of an anomalous data point in Dare county. We expect tourism to be negatively correlated with crime since it will increase the strength of a local economy, leading to less crime. We also expect tourism to be positively correlated with **west** since western counties have lots of tourist activity in the way of the National Park and other attractions. This would lead to a negative omitted variable bias. If we controlled for tourism, we expect the fitted values for a given value of west to be higher, making the effect of west closer to zero.

Lastly, we believe cultural values like **religious faith** may play a role in the crime rate. For example, if the people living in a particular region report a strong religious faith or high levels of religious activity, they may be less likely to engage in crime, because many faiths have specific prohibitions against engaging in violence, stealing, damaging what belongs to others, and other illegal behavior. In other words, higher levels of religious beliefs or activity should negatively correlated with crime rate. Moreover, we suspect that the more uniform the religious faith in a region, the more robust this effect should be. This is because religious differences in a community can create social conflict, and such conflict may motivate crime against those perceived as "other". We note that shared cultural values, such as religious faith, may represent a form of **community membership**, discussed above.

The discussion of omitted variables, however, is speculative, and should be reinforced with research. Ideally, this research would include randomized, controlled trials where possible.

6. Conclusion

Based on our regression analysis of the available data, we find that population density, tax per capita, and the percentage of young males are responsible for the majority of the variation in crime rate in North Carolina in the late 1980s. Some measures of crime management such as the number of police per capita, the ratio of arrests to offenses, and the ratio of convictions to arrests also appear to influence the variation in crime rate. The location of a county within the state and the percentage non-White minorities in a county also appear to influence the crime rate, but this may be primarily through the effect of unmeasured or omitted variables, such as community membership, unemployment rate, and educational attainment .

Based on this analysis, we generate several policy suggestions applicable to candidates seeking or defending office in North Carolina. Specifically, candidates are advised to pursue programs that increase public access to key resources, encourage community involvement, and decrease perceptions of social injustice. Moreover, we encourage the candidates to discuss their crime reduction plans from the perspective of reducing personal crime, including violent crime, because such crimes are typically of greater concern to constituents than property crimes and are likely to generate voting behavior that favors the candidate.

Although we believe our analysis supports our recommendations well, it does have some important limitations. First, the data are reported by county, which is a rather coarse measure for examining crime rate and the variables that influence it. It seems unlikely, for example, that population density and tax per capita are evenly distributed across a single county, so potentially meaningful variations in these variables are not transparent at the county level. Similarly, some major metropolitan areas stretch across multiple counties, and their police forces and court systems are accustomed to interacting with and accommodating one another. As such, the behaviors of the police and courts might appear more similar in these counties, when this is actually an artifact of the city size. Second, our data set is limited to North Carolina counties in 1987. Not only is this data more than 30 years old, but it is restricted to a specific state at a very specific time. As such, we cannot generalize our findings to other states or years, or even test any hypotheses we might generate. Third, we have limited information about how the data were collected, recorded, or prepared prior to being

delivered to us. In the absence of this information, we cannot assess the quality of the data or develop clear strategies for addressing in weaknesses that may have resulted from the collection methodology. Lastly, the counties that are omitted from the data set appear to be those with lower population. Given that population density has proven to be an important variable in our models, it would be useful to have data from these counties to make sure our understanding of this variable is as complete and accurate as possible.