# Lab 3: Reducing Crime

w203 Summer 2018

*Madeleine Bulkow, Kim Darnell, Alla Hale, Emily Rapport*

*July 17, 2018*

## 1. Introduction

As advisees to candidates running for statewide office in North Carolina, we believe that the crime rate across the state should be of central concern to any candidate. Local governments across the state 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 local government in North Carolina for the late 1980s.

## 2. Data Definitions and Data Cleaning

The data in this report was collected by researchers Cornewell and Trumball. They collected data related to crime, demographics, and the economy for 97 counties in North Carolina. While the authors collected data over a number of years, we will focus on the data from the year 1987.

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

county: integer code representing which county the row represents. We received the data with these identifier codes in place of county names, so we cannot identify the individual counties in the dataset.

year: 1987 for all data points.

crmrte: ratio of crimes permitted to population, taken from the FBI's Uniform Crime Reports.

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

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

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

avgsen: average prison sentence in days, which we believe is taken from the North Carolina Department of Correction.

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

density: people per square mile

taxpc: tax revenue per capita.

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

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

urban: indicator code specifying whether county is urban, defined by whether the county is in a Standard Metropolitan Statistical Area as defined by the US Census.

pctmin80: percentage of population that belongs to minority racial group, as taken by the 1980 US Census.

mix: ratio of face-to-face offenses to other offenses.

pctymle: percent young male, defined as proportion of population that is male between the ages of 15 and 24, as taken by US 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 perhaps transforming the data.

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

Clean up the apostrophe.

It appears that probconv is in percent, while the other two probability estimates (prbarr and prbpris) are fractions. To be able to compare coefficients more easily, let's get all percentage values in percent (0-100).

Remove the points where probabilities exceed 100 %.

```
# Clean the data

## NOTE FROM ALLA: This is just what I did to clean the data.  I am sure this can be done in a more effi
df_calc <- df
df_calc$prbconv <- as.numeric(as.numeric(df$prbconv))
df_calc$prbarr <- df$prbarr * 100
df_calc$prbpris <- df$prbpris * 100
df_calc$pctymle <- df$pctymle * 100
#(df_calc$county)
#summary(df_calc)
df_clean <-df_calc[with(df_calc, prbarr <= 100 & wser <= 2000),]
```

## 3. Building Models

Our central goal for this analysis is to determine what variables are most associated with crime in North Carolina. For this reason, we will use the crmrte variable as the outcome variable in most of our models. Before we begin modeling, we need to examine our outcome variable.

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

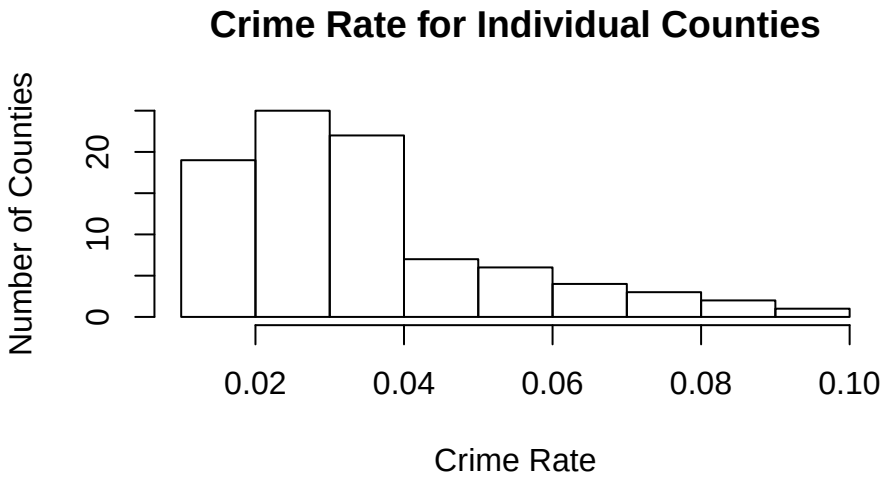
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01062 0.02216 0.03002 0.03397 0.04086 0.09897       6
```

The value of crmrte ranges from approximately .011 to .099, with a mean of approximately .034. Six observations have NA values in place of their crmrte. Since our analysis primarily seeks to discover variables associated with crmrte, we will remove these rows with NA's, as they will not serve our analysis.

```
df_clean <- df_clean[complete.cases(df_clean$crmrte), ]
```

We are left with 89 observations. A histogram of the observations shows us the shape of their distribution:

```
hist(df_clean$crmrte,
     main="Crime Rate for Individual Counties",
     xlab= "Crime Rate",
     ylab= "Number of Counties")
```




We see in this histogram that the data is right skewed; the majority of counties have crime rates below .04, while the long right tail demonstrates that a smaller number of counties have substantially higher crime rates.

## 4. The Models

We complete the model building process in 5 stages, resulting in 5 separate models.

The first four models are linear regressions of crime rate against increasing numbers of predictors. The first model includes only variables we believe to be the main predictors of crime rate, density, tax per capita, and percent young male. The second model includes several other factors we believe may be explanatory. The third model adds the variables we have available, which are not problematic. Finally, we show the fourth model, which includes the balance of variables, even those of questionable merit.

The fifth model, is a regression of the crime rate multiplied by the mix. NEED MADELEINE'S INPUT HERE.

For each model, we use the classic linear model assumptions:

Assumption 1: Linear in Parameters Assumption 2: Random (i.i.d) Sampling Assumption 3: Multicollinearity Assumption 4: Zero Conditional Mean/Exogeneity Assumption 5: Homoskedasticity Assumption 6: Normality of Variance

Assumption 1, linearity in parameters, holds, as each fit model has slope coefficients that are linear multipliers of the associated predictor variables. Assumption 2, random sampling, says that are data points must be independent and identically distributed. We have data for 97 of North Carolina's 100 counties (89 after remove NA values). While this represents something closer to the population of counties than a sample, we should be okay on the random sampling assumption if we assume that there is no pattern to the counties that weren't included in the data or that had NA values. We will validate the other 4 assumptions for each subsequent model.

### 4.1 Model 1

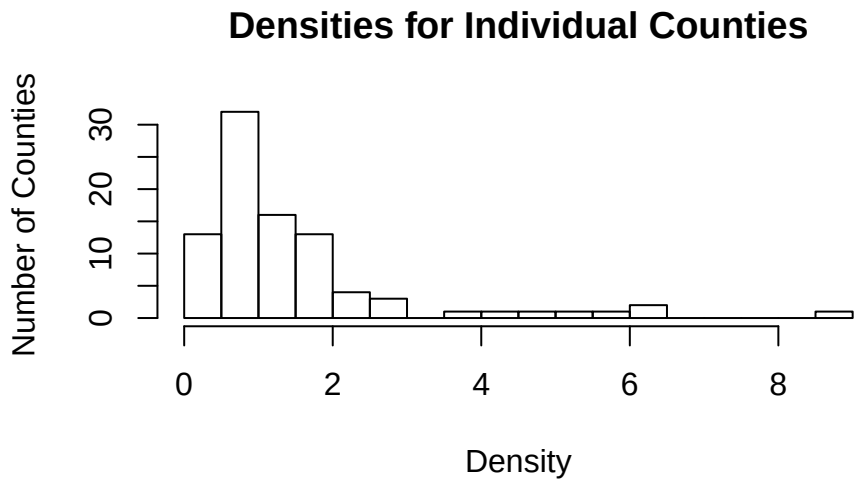
Causes of crime have been debated, but we suspect that density, tax per capita, and percent young male are strong predictors of crime. The dependent variable, crime rate, has already been assessed, so we evaluate these three predictor variables.

Density:

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

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.56397 0.99623 1.45224 1.57028 8.82765
```

```
hist(df_clean$density,
     main="Densities for Individual Counties",
     xlab= "Density",
     ylab= "Number of Counties",
     breaks = 30)
```



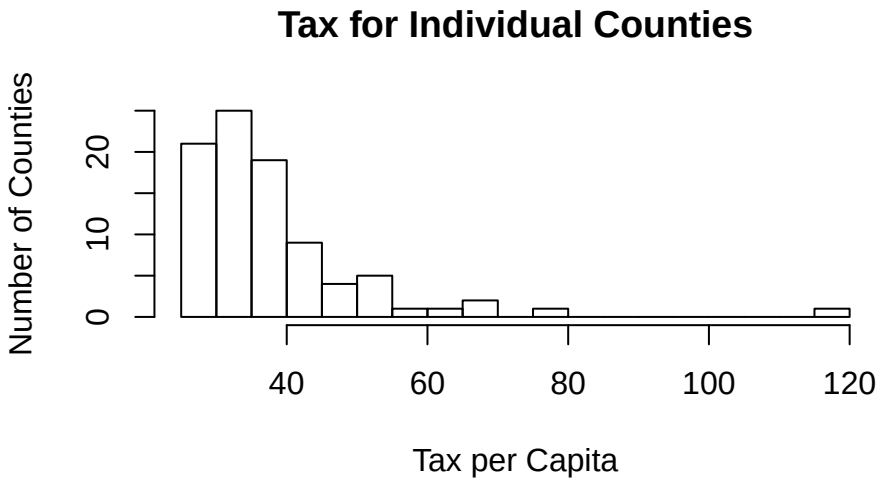

The value of density ranges from approximately .00002 to 8.8 people per square mile, with a mean of 1.45. The distribution of county densities is right skewed, with most counties being sparse and a long tail of more populated counties. After reviewing the census data, it is logical to conclude that these numbers are in hundreds of people per square mile, but for consistency we will continue to use the people per square mile unit.

Tax per Capita:

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

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    25.69   30.70   34.87   38.13   41.07  119.76
```

```
hist(df_clean$taxpc,
     main="Tax for Individual Counties",
     xlab= "Tax per Capita",
     ylab= "Number of Counties",
     breaks = 30)
```



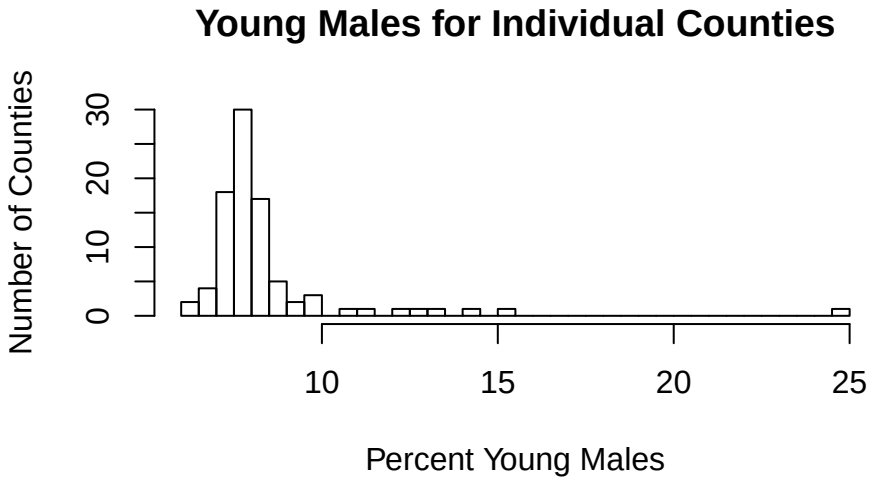

The value of tax per capita ranges from 25.69 to 119.76. Once again, we see a distribution that is right skewed, with revenue in most counties below the mean of 38.13. The maximum value is much higher than the next highest value. Though this is an interesting note, evaluating the row, we have no reason to doubt this data point.

Percent Young Male:

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.216   7.463   7.787   8.425   8.354  24.871
```

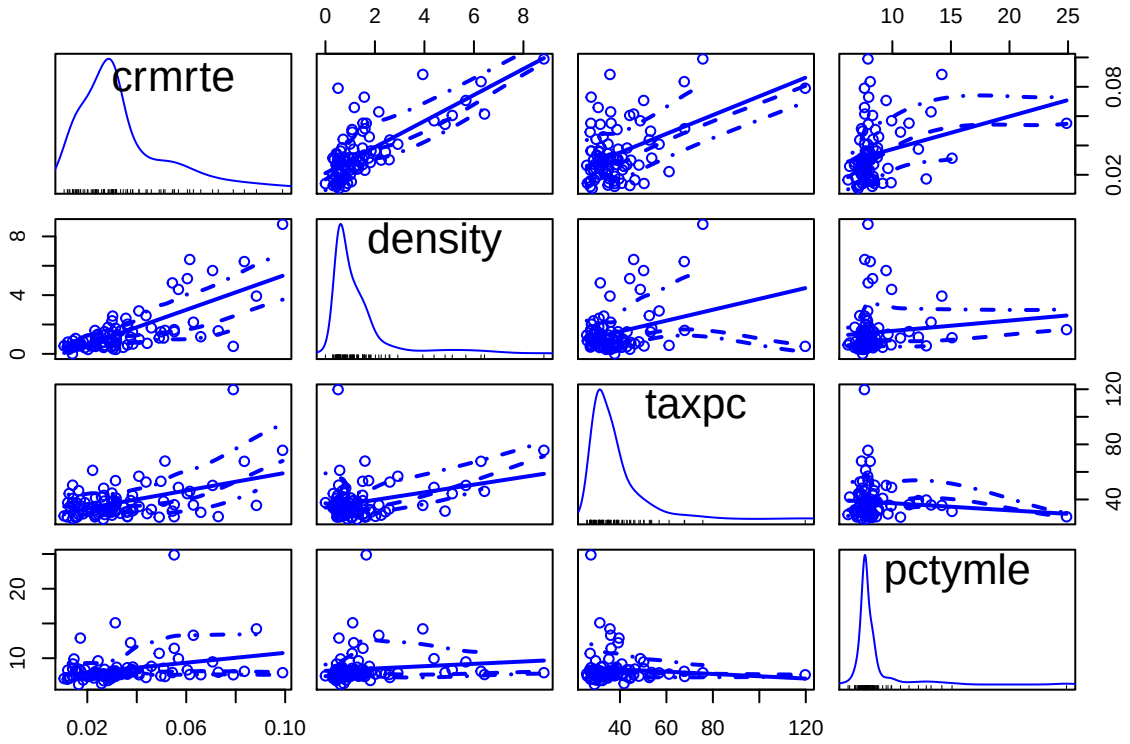
```
hist(df_clean$pctymle,
     main="Young Males for Individual Counties",
     xlab= "Percent Young Males",
     ylab= "Number of Counties",
     breaks = 30)
```




The percent of young males ranges from 6.2 to 24.9 %. Once again, we see a distribution that is right skewed, with revenue in most counties below the mean of 8.4 %. The skew of this distribution leads us to take the log of this variable within our model, to spread out the data in the lower end of the range. Again, we see the maximum value is much higher than the next highest value, and again, we have no reason to doubt this data point.

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, and validate assumption MLR.3, multicollinearity.

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




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

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

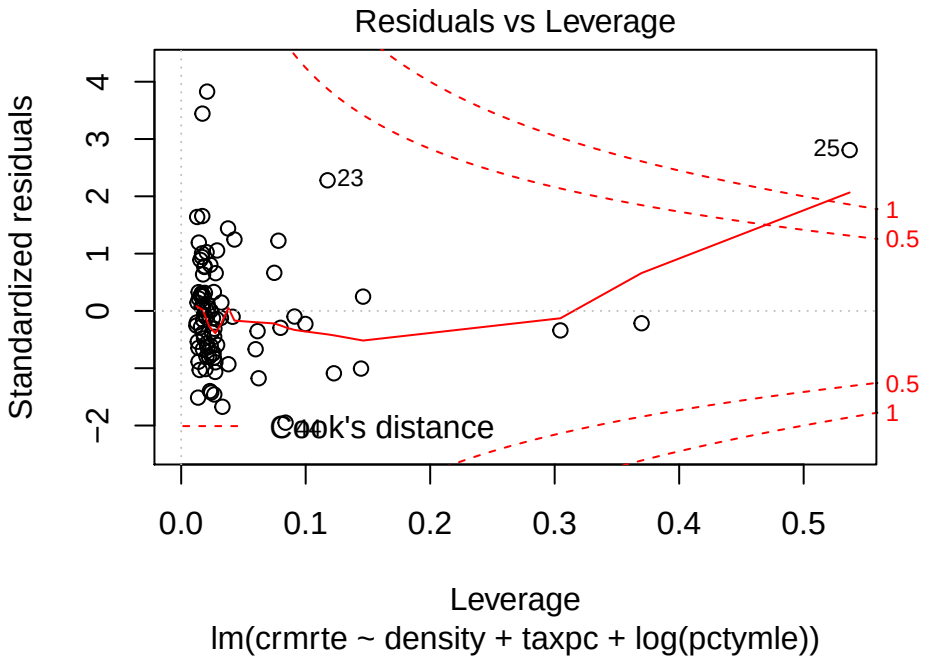
```
# Build Model 1
(model_1 = lm(crmrte ~ density + taxpc + log(pctymle), data = df_clean))
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + log(pctymle), data = df_clean)
##
## Coefficients:
##  (Intercept)       density         taxpc  log(pctymle)
##   -0.0409401     0.0073688     0.0003934     0.0233503
```

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

```
## [1] 0.6415746
```

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




We find one point that has a Cook's Distance greater than 1, Manteo county, but with no justification to remove it from the dataset, we simply note it. Of note, this county has the highest tax per capita, which could stem from its tourist destination status as the location of the Wright brothers' first flight.

Now that the model is built, we can validate assumption 4, the exogeneity assumption. To do this, we check that the sum of the residuals is 0.

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

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

We find that the residuals sum to 0, validating assumption 4.

To validate assumption 5, homoskedasticity, we took a look at the residuals vs. fitted values plot and noted that the error range was relatively constant throughout the range of fitted values. This was difficult to validate because we have fewer data points at the higher values of crime rate than the lower values of crime rate.

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.

From the equation for model 1 (Equation 1), we can see that all of these model coefficients are positive, indicating that for an increase in density, taxpc, or pctymle, there is an associated increase in crime rate.

$$crmrte = -0.0409 + 0.007 \cdot density + 0.0004 \cdot taxpc + 0.023 \cdot log(pctymle) \qquad \text{(Equation 1)}$$

The log transformation of pctymle masks the magnitude of the effect, but density has a larger effect than tax per capita.

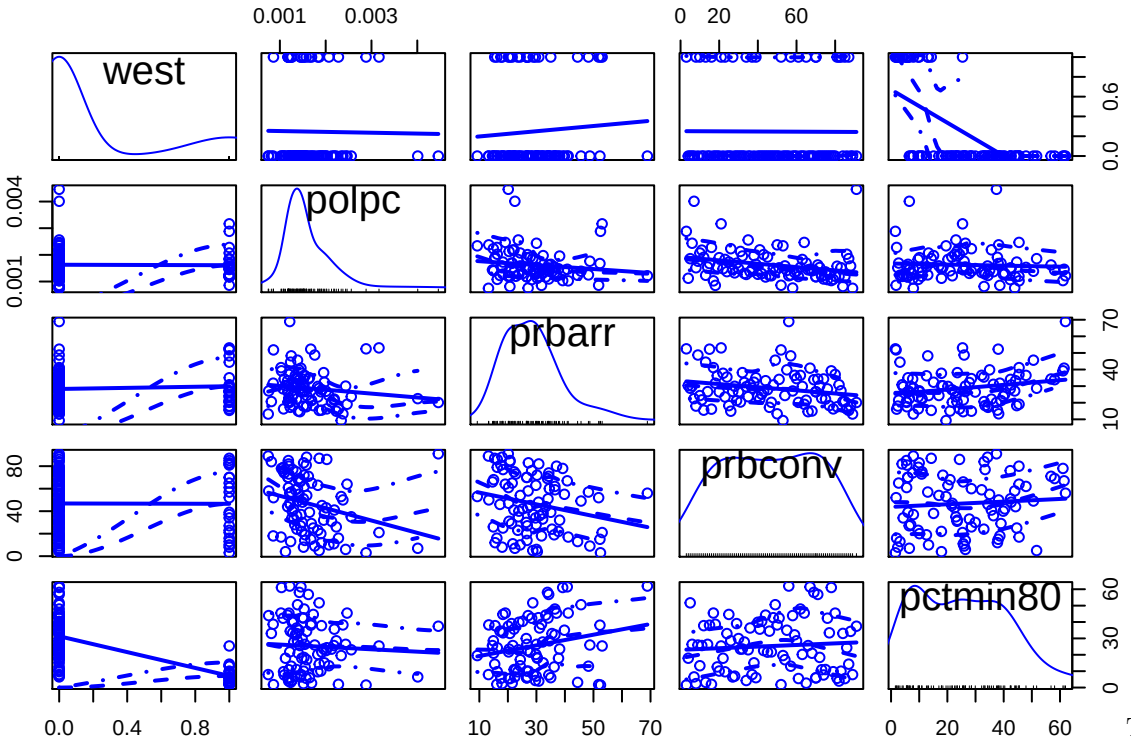
KIM: THIS WOULD BE A GOOD PLACE TO TALK ABOUT WHY WE SUSPECT THE EFFECTS ARE IN THE DIRECTION THEY ARE IN AND WHAT SORTS OF POLICIES WE MIGHT SUGGEST TO DETER CRIME.

### 4.2 Model 2

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

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

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




The matrix plot shows little correlation, and certainly no perfect multicollinearity, between the additional variables in this model. One noteworthy observation is that pctmin80 appears correlated with west, which may absorb some of the effect. To validate assumption 3, collinearity, we also checked the scatterplots of these variables against the original three included in model 1. No perfect multicollinearity was found.

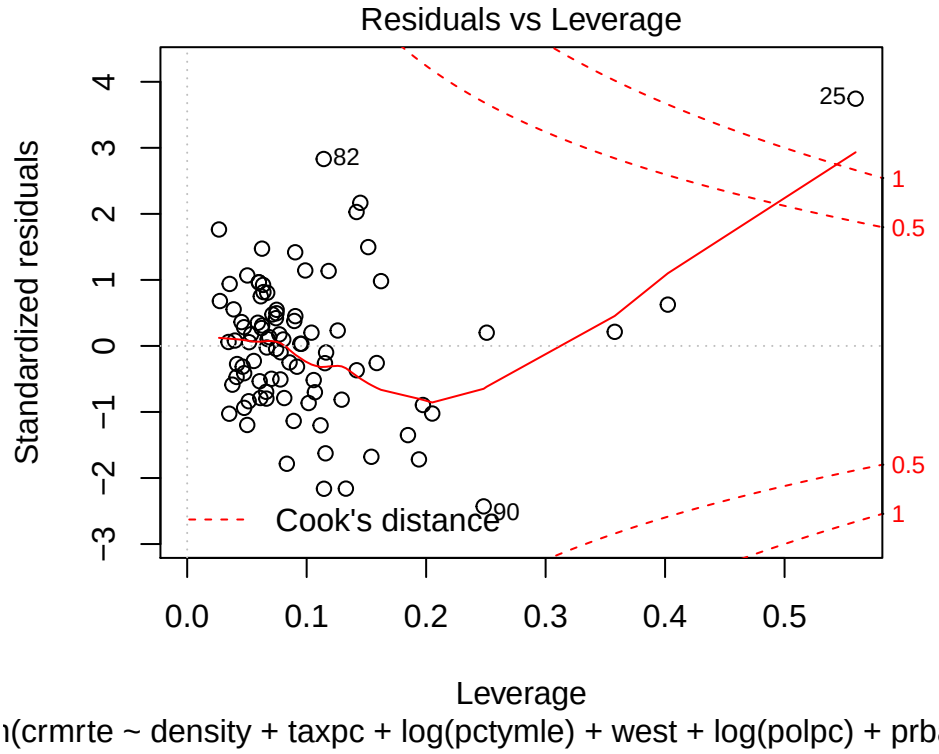
```
# Build Model 2
(model_2 = lm(crmrte ~ density + taxpc + log(pctymle)
              + west + log(polpc) + prbarr + prbconv + pctmin80,
              data = df_clean))
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + log(pctymle) + west +
##     log(polpc) + prbarr + prbconv + pctmin80, data = df_clean)
##
## Coefficients:
##  (Intercept)       density         taxpc  log(pctymle)          west
##    0.0704871     0.0055627     0.0002045     0.0099548    -0.0014126
##  log(polpc)        prbarr       prbconv      pctmin80
##    0.0092700    -0.0005056    -0.0001554     0.0003482
```

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

```
## [1] 0.7883347
```

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




ı(crmrte ~ density + taxpc + log(pctymle) + west + log(polpc) + prbar

Unsurprisingly, the $R^2$ increased from 0.64 to 0.79 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.

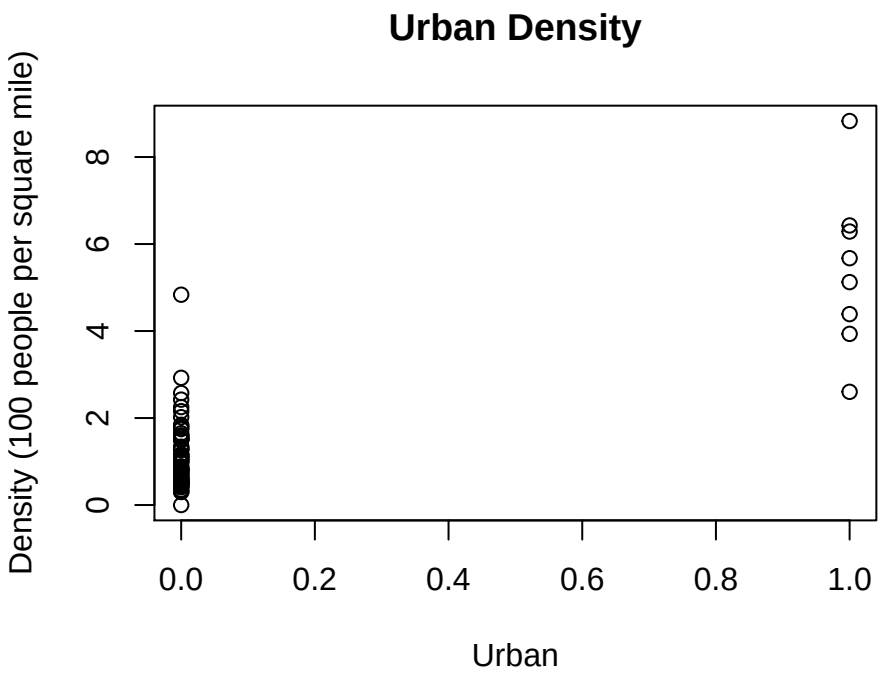
KIM: THIS WOULD BE A GOOD PLACE TO TALK ABOUT WHY WE SUSPECT THE EFFECTS ARE IN THE DIRECTION THEY ARE IN AND WHAT SORTS OF POLICIES WE MIGHT SUGGEST TO DETER CRIME.

### 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")
```




We also excluded all of the wage variables because we cannot make any meaningful conclusions without a breakdown of what portions of each county are involved in each profession.

With that, we build model 3:

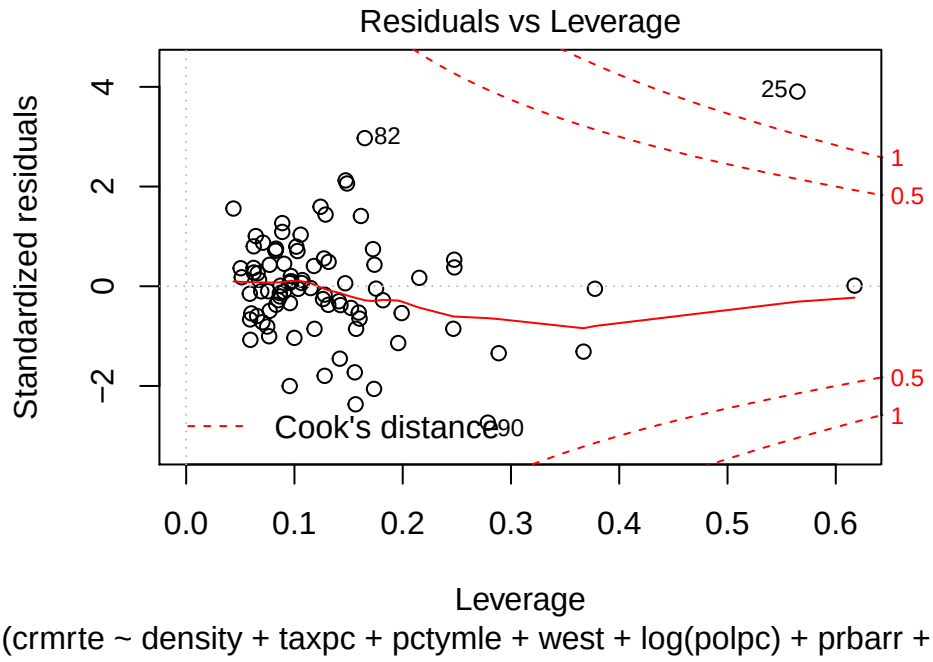
```
# Build Model 3
(model_3 = lm(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + prbconv + pctmin80 + ce
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle + west + log(polpc) +
##     prbarr + prbconv + pctmin80 + central + avgsen + prbpris,
##     data = df_clean)
##
## Coefficients:
## (Intercept)      density        taxpc      pctymle         west
##   1.096e-01    5.902e-03    1.671e-04    7.857e-04   -4.863e-03
##  log(polpc)       prbarr      prbconv     pctmin80      central
##   1.177e-02   -4.935e-04   -1.477e-04    2.744e-04   -3.656e-03
##      avgsen      prbpris
##  -6.368e-04    4.461e-05
```

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

```
## [1] 0.7995326
```

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




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

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

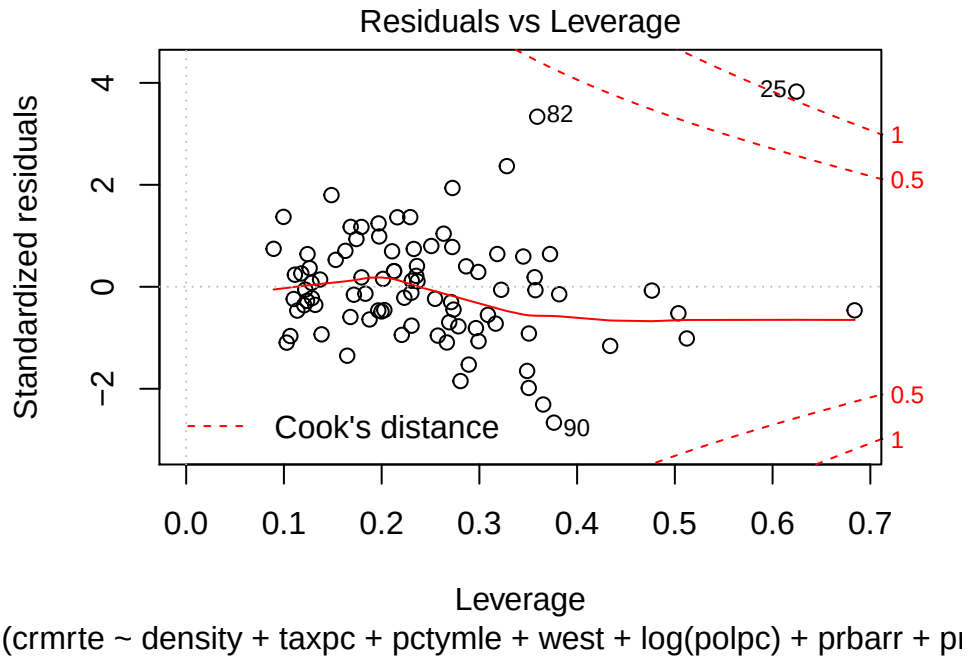
### 4.4 Model 4

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 0.84, which is not much higher than model 3.

```
# Build Model 4
model_4 = lm(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + prbconv + pctmin80 + cen
paste("The R.square is", round(summary(model_4)$r.square, 2))
```

```
## [1] "The R.square is 0.84"
```

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




### 4.5 Model 5

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

### 4.6 Model Summary

This is where we put our model summary table.

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

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 17, 2018 - 21:53:07

## 5. Omitted Variables

household earnings: we would expect coefficient on hh_earnings to be negative, since wealthier areas = less crime. We would expect correlation between household earnings and density to positive, because in cities people have higher salaries. Therefore, omitted variable bias is *negative*, and if we had household earnings, the fitted values for a given density value would likely be higher.

socioeconomic disparity:

unemployment rate: we would expect coefficient on unemployment rate to be positive, since more people without work = more people in desperate situations that lead to them breaking the law. We would expect correlation between unemployment and percent young male to be positive, since young men are most likely to be unemployed (I'm REALLY not sure about this one - if there's a better claim to be made with density, let's

Table 1: Linear Models Predicting Crime Rate

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | crmrte | | | |
| | (1) | (2) | (3) | (4) |
| density | 0.007 | 0.006 | 0.006 | 0.005 |
| taxpc | 0.0004 | 0.0002 | 0.0002 | 0.0002 |
| log(pctymle) | 0.023 | 0.010 | | |
| pctymle | | | 0.001 | 0.001 |
| west | | −0.001 | −0.005 | −0.003 |
| log(polpc) | | 0.009 | 0.012 | 0.011 |
| prbarr | | −0.001 | −0.0005 | −0.0005 |
| prbconv | | −0.0002 | −0.0001 | −0.0001 |
| pctmin80 | | 0.0003 | 0.0003 | 0.0003 |
| central | | | −0.004 | −0.005 |
| avgsen | | | −0.001 | −0.001 |
| prbpris | | | 0.00004 | −0.00000 |
| urban | | | | 0.0001 |
| wcon | | | | 0.00003 |
| wtuc | | | | 0.00001 |
| wtrd | | | | 0.00004 |
| wfir | | | | −0.00003 |
| wser | | | | −0.0001 |
| wmfg | | | | −0.00000 |
| wfed | | | | 0.0001 |
| wsta | | | | −0.00001 |
| wloc | | | | 0.0001 |
| Constant | −0.041 | 0.070 | 0.110 | 0.076 |
| Observations | 89 | 89 | 89 | 89 |
| $R^2$ | 0.642 | 0.788 | 0.800 | 0.842 |

do it). Therefore, omitted variable bias is *positive*, and the fitted values would be lower for a given pctymle value if we had unemployment rate.

education level: We would expect coefficient on education level to be negative, as when people have more education, they probably have more job opportunities –> less likelihood to commit crime. We would expect coefficient between education level and density to be positive, since people tend to be more educated in urban areas. Therefore, omitted variable bias is *negative*, and the fitted values for a given value of density would likely be lower if we could control for education level.

## 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.)