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.

1

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.

```r
# 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 %.

```r
# 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 eff
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.

```r
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.

```r
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:

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

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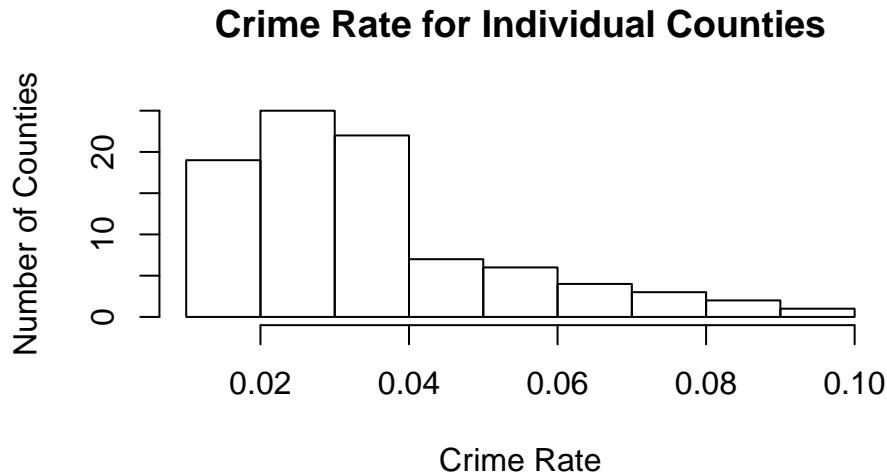
We see in this histogram that the data is right skewed; the majority of counties have crime rates below .04, while the long right tail demonstrates that a smaller number of counties have substantially higher crime rates.

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.

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

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

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.
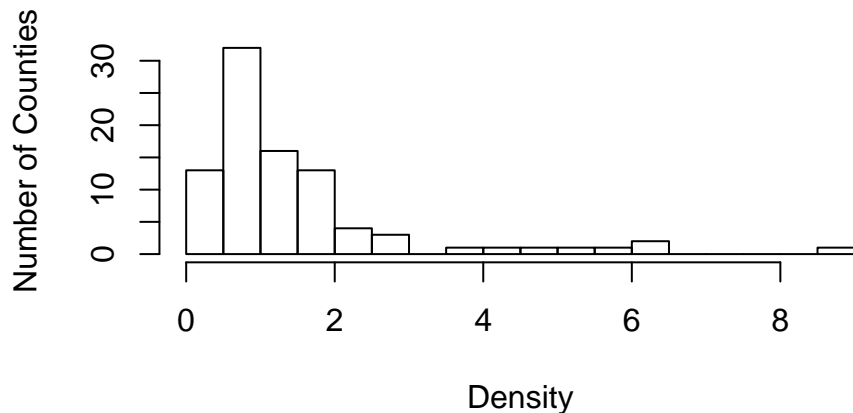
Density:

```
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)
```

Densities for Individual Counties

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.
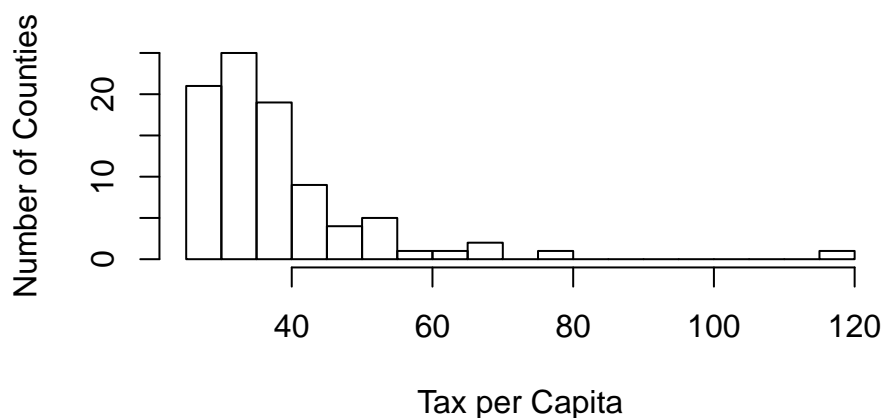
Tax per Capita:

```
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)
```

Tax for Individual Counties

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
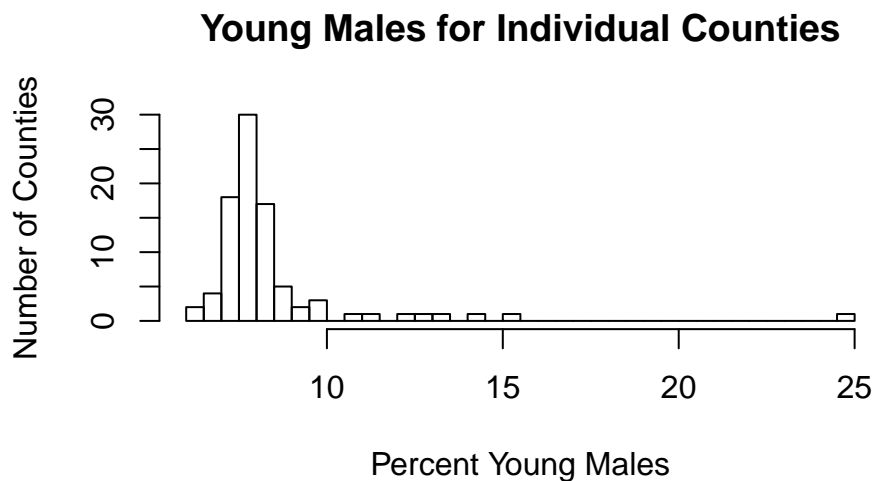
this data point.

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)
```

Young Males for Individual Counties

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 %. 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.

```
vars <- c("crmrte", "density", "taxpc","pctymle")
suppressWarnings(scatterplotMatrix(df_clean[,vars], diagonal = "histogram"))
```
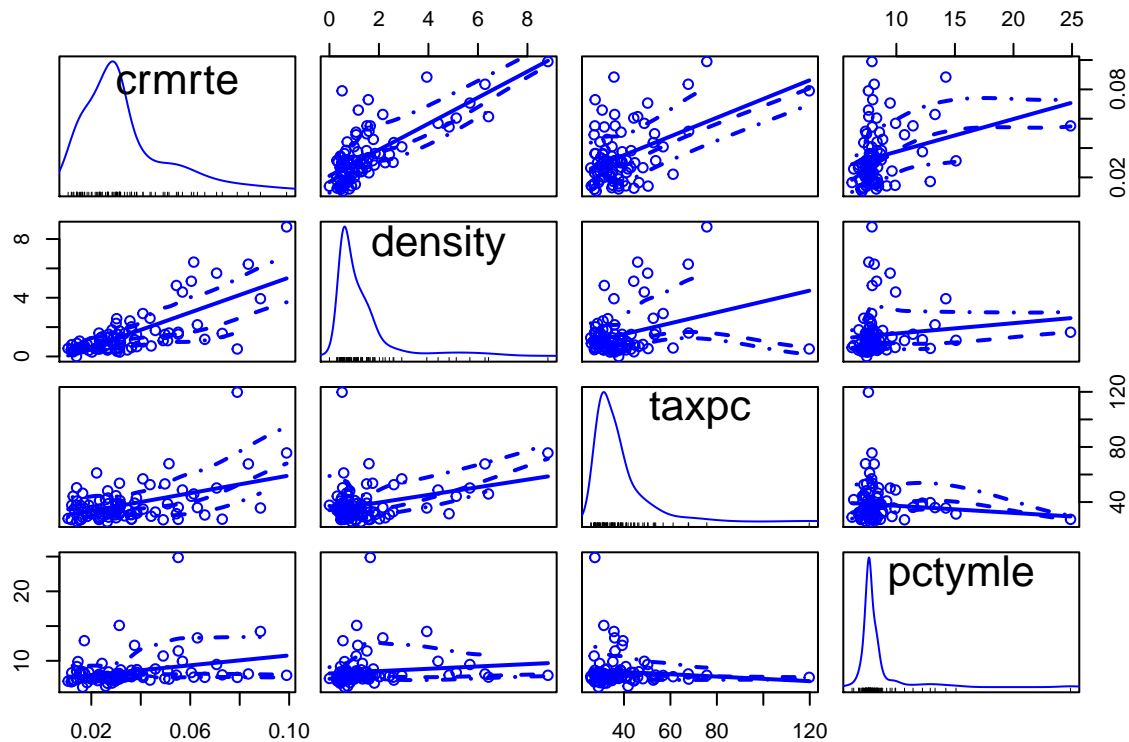
As we suspected, crime rate looks well predicted by each of the three primary variables selected as evidenced by the fairly strong positive slopes in the bivariate regressions in the scatterplot matrix. Additionally, though density and taxpc appear to have a positive correlation, none of the variables are collinear with any of the others.

With the evaluation of the variables complete, we build model 1, and evaluate the Cook's Distance for the residuals:
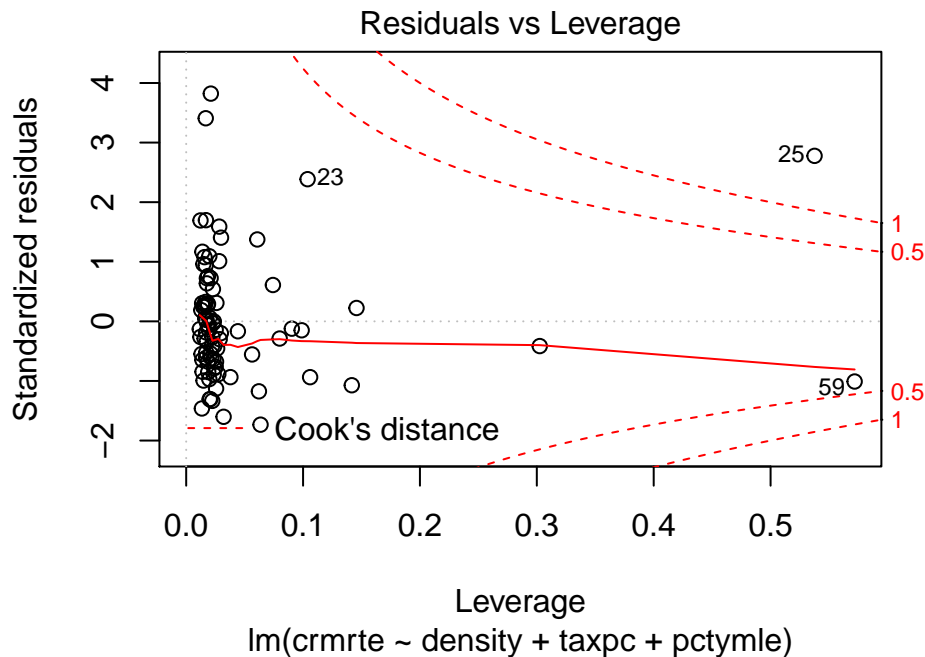
```
# Build Model 1
(model_1 = lm(crmrte ~ density + taxpc + pctymle, data = df_clean))

##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle, data = df_clean)
##
## Coefficients:
## (Intercept)       density         taxpc       pctymle
##  -0.0079797     0.0074778     0.0003933     0.0019095
```

```
summary(model_1)$r.square

## [1] 0.638322
```

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

Residuals vs Leverage
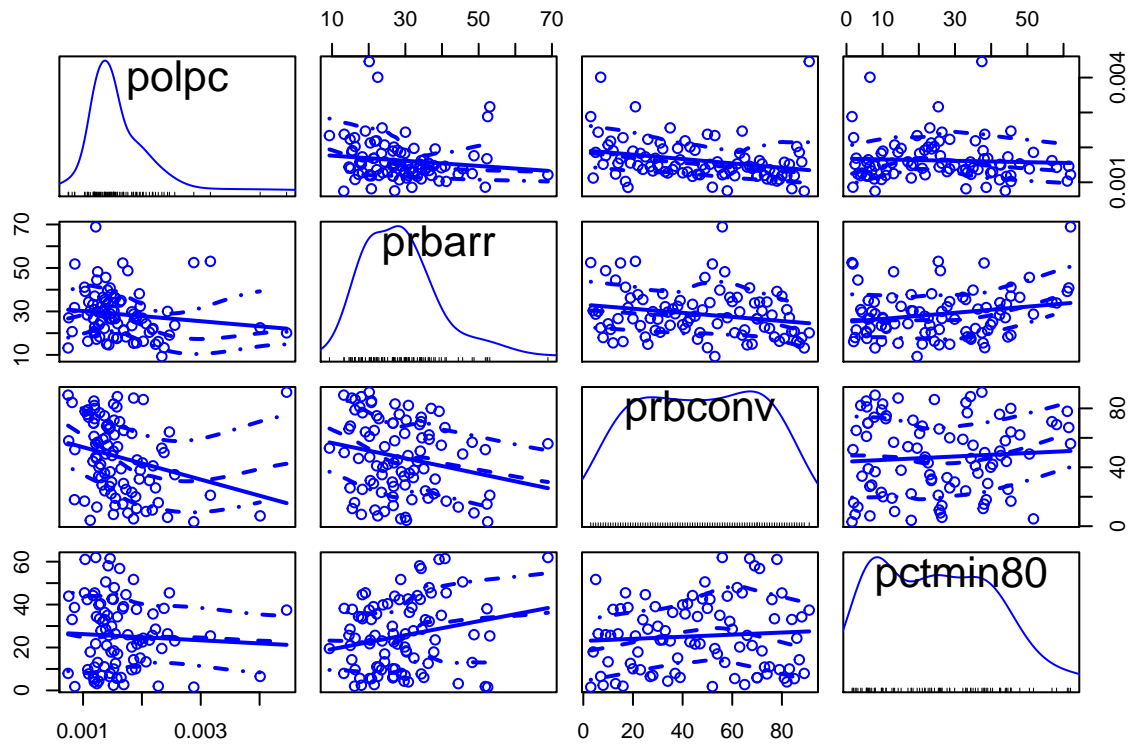

lm(crmrte ~ density + taxpc + pctymle)

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

4.2 Model 2

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

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

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

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

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

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle + west + log(polpc) +
##     prbarr + prbconv + pctmin80, data = df_clean)
##
## Coefficients:
## (Intercept)       density         taxpc        pctymle          west
##   0.0851380     0.0055851     0.0002049      0.0008898    -0.0014142
##  log(polpc)         prbarr        prbconv       pctmin80
##   0.0094749    -0.0005058    -0.0001538      0.0003497
```
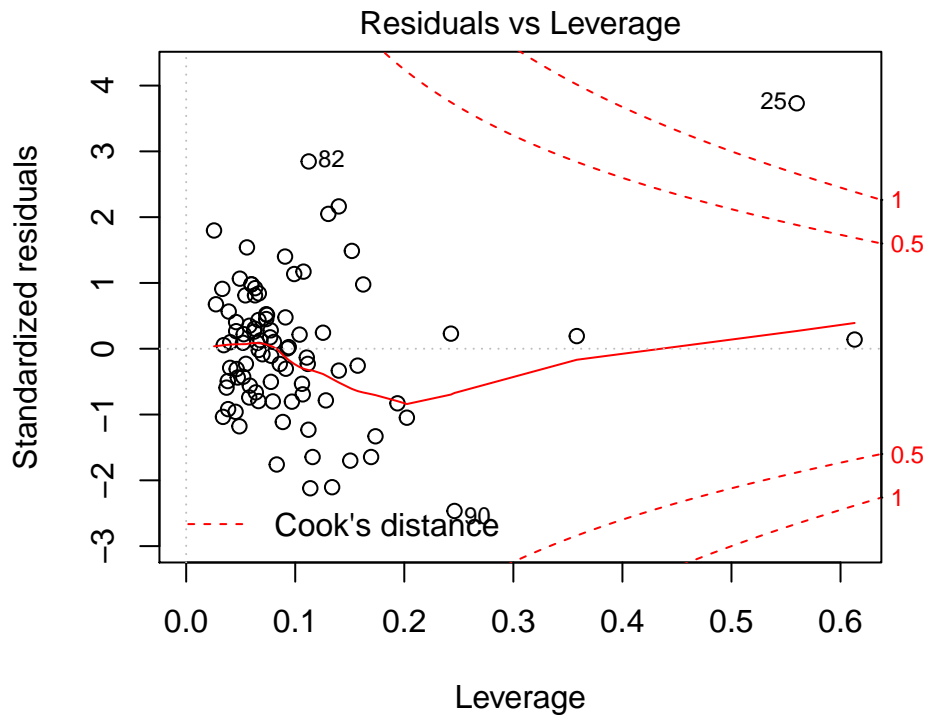
```r
summary(model_2)$r.square
```

```
## [1] 0.7898102
```

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

Residuals vs Leverage

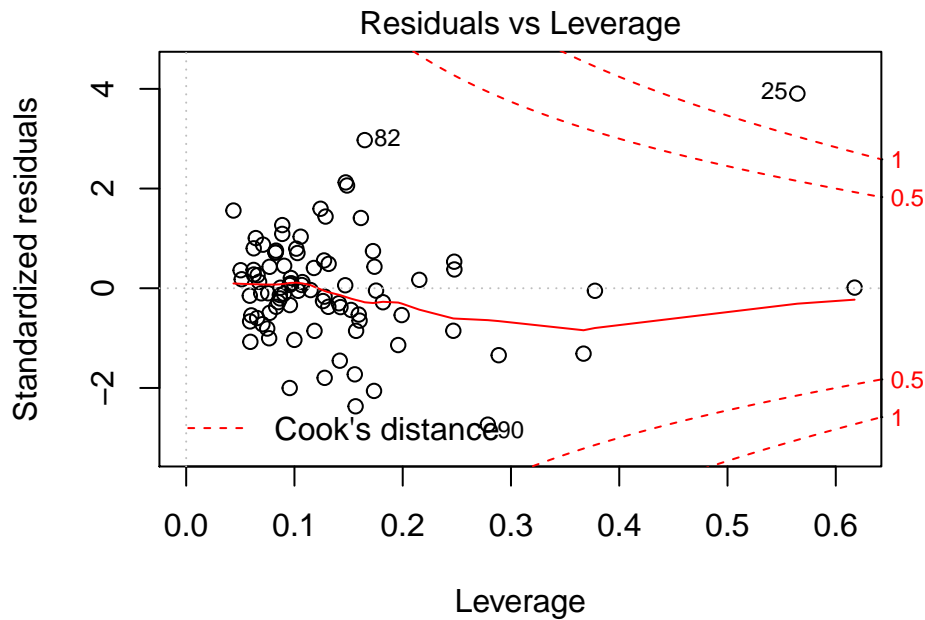(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + pr

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

4.3 Model 3

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

```
##
## 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)
```

Residuals vs Leverage

(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + p

4.4 Model 4

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

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc + pctymle + west + log(polpc) +
##     prbarr + prbconv + pctmin80 + central + avgsen + prbpris +
##     urban + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
##     wsta + wloc, data = df_clean)
##
## Coefficients:
## (Intercept)       density          taxpc        pctymle          west
##   7.644e-02     5.288e-03      2.285e-04      1.335e-03     -2.939e-03
##  log(polpc)        prbarr         prbconv       pctmin80        central
##   1.090e-02    -4.951e-04     -1.422e-04      2.595e-04     -4.802e-03
##      avgsen        prbpris           urban           wcon           wtuc
##  -8.445e-04    -5.032e-07      5.164e-05      2.799e-05      1.411e-05
##        wtrd           wfir            wser           wmfg           wfed
##   3.674e-05    -3.240e-05     -9.835e-05     -4.298e-07      6.090e-05
##        wsta           wloc
##  -9.292e-06     5.039e-05
```
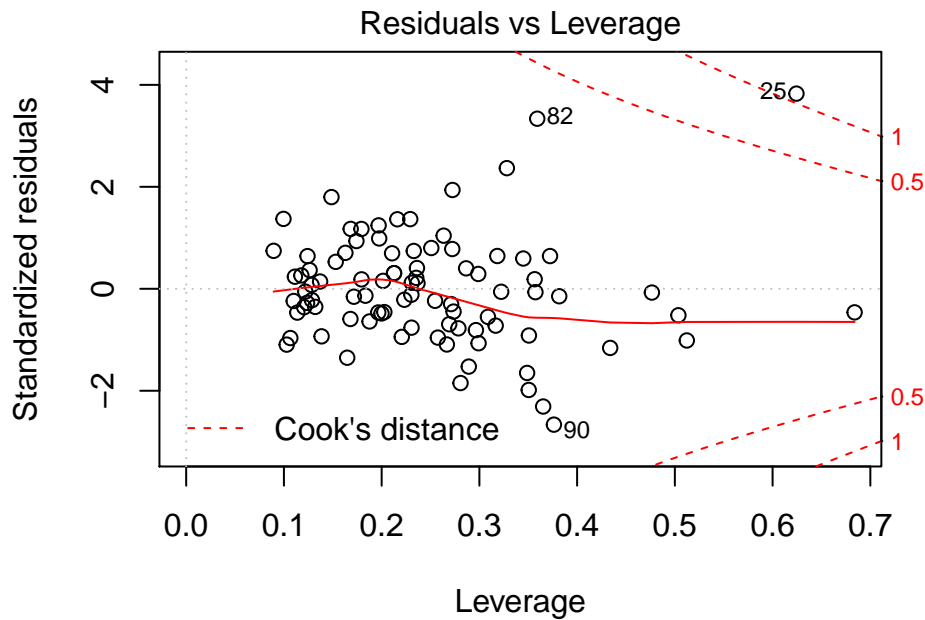
```
summary(model_4)$r.square
```

```
## [1] 0.8417722
```

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

Residuals vs Leverage

(crmrte ~ density + taxpc + pctymle + west + log(polpc) + prbarr + pr

4.5 Model 5

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

4.2 Model Summary

This is where we put our model summary table.

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

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 17, 2018 - 19:38:03

5. Omitted Variables

6. Conclusion

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

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
pctymle	0.002	0.001	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	−0.008	0.085	0.110	0.076
Observations	89	89	89	89
R^2	0.638	0.790	0.800	0.842