Lab 3: Reducing Crime (DRAFT)

W203 Statistics

Luke Evans, Daniel Rasband, and Yulia Zamriy

April 3, 2018

An Analysis of Crime in North Carolina to Support Policy Decisions

Introduction

Crime is expected to be a significant issue during the upcoming election in North Carolina. Using statistical techniques, this report attempts to provide data-driven insights into the determinants of crime in the state. A mixture of both long- and short-term policy suggestions will be included to address the factors that exacerbate crime, and to capitalize on those factors that act as suppressors.

Exploratory Data Analysis

The data utilized to conduct this statistical analysis generally comes from the year 1987, with a single variable from 1980 (percent minority). Data is provided for most counties in North Carolina, and can be further grouped by region (West, Central and Other). Granularity below the county level is not available.

While averages and rates are presented for many variables, the absolute numbers, for example of population, are not. This can generate some challenges when discussing practical significance.

Data Cleaning

Our initial exploration of the data has revealed several notable features. The information below provides the dimensions of the raw data: 25 variables and 97 rows.

```
crime <- read.csv("crime_v2.csv", stringsAsFactors = FALSE)
dim(crime)
```

```
## [1] 97 25
```

There are a total of 91 observations in the dataset; 6 rows are completely devoid of data and can be excluded. It should be noted that there are 100 counties in North Carolina; therefore this dataset contains data for 91% of them. It is not possible to tell if the excluded counties are randomly excluded or share specific features that may bias this data set.

```
crime <- na.omit(crime)
```

Counties range in population numbers from 15,000 people to over 1 million. The data provided has many abstracted values such as ratios and averages, but without the actual numbers relating to those abstractions, it can be hard to draw practical significance from conclusions as each county will be considered equal to any other. As electoral representation in general does not follow population density, there may be advantages to analyzing data at a county level only, but this limitation should be considered depending on the inference that is being generated.

From the summary, the probability of conviction dimension, `prbconv`, needed to be transformed to a more appropriate data type in R:

```
crime$prbconv <- as.numeric(as.character(crime$prbconv))
```

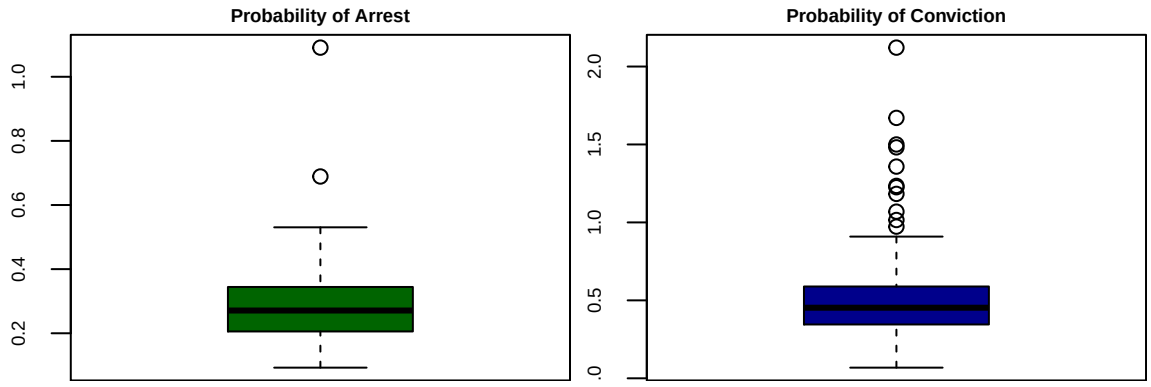
The variables are made up of ratios, specifically the probability of arrest, conviction and prison sentence, the percent minority, young male, police and tax revenue per capita, and the ratio (mix) of face to face crimes to other types of crime. The means of several variables are provided for each county: a series of weekly wages in different business segments, and prison sentences in days. Finally, an indicator of the location of the county in the state is also provided, indicating West and Central regions. An "Other" region can be identified by difference. The `urban` variable also indicates whether the county is a "Standard Metropolitan Statistical Area." Below is a summary of variables including some summary statistics:

```
crime_summary <- data.frame(t(mapply(summary, crime)))
crime_summary <- crime_summary[,c("Min.", "Mean", "Max.")]
crime_summary$Min. <- round(crime_summary$Min.,5)
crime_summary$Mean <- round(crime_summary$Mean,4)
crime_summary$Max. <- round(crime_summary$Max.,4)
kable(crime_summary, booktabs = TRUE) %>%
  kable_styling(font_size = 7)
```

	Min.	Mean	Max.
county	1.00000	101.6154	197.0000
year	87.00000	87.0000	87.0000
crmrte	0.00553	0.0334	0.0990
prbarr	0.09277	0.2949	1.0909
prbconv	0.06838	0.5513	2.1212
prbpris	0.15000	0.4108	0.6000
avgsen	5.38000	9.6468	20.7000
polpc	0.00075	0.0017	0.0091
density	0.00002	1.4288	8.8277
taxpc	25.69287	38.0551	119.7615
west	0.00000	0.2527	1.0000
central	0.00000	0.3736	1.0000
urban	0.00000	0.0879	1.0000
pctmin80	1.28365	25.4955	64.3482
wcon	193.64316	285.3585	436.7666
wtuc	187.61726	411.6680	613.2261
wtrd	154.20900	211.5529	354.6761
wfir	170.94017	322.0982	509.4655
wser	133.04306	275.5642	2177.0681
wmfg	157.41000	335.5887	646.8500
wfed	326.10001	442.9007	597.9500
wsta	258.32999	357.5220	499.5900
wloc	239.17000	312.6808	388.0900
mix	0.01961	0.1288	0.4651
pctymle	0.06216	0.0840	0.2487

From the above table it can be seen that in several counties, the probability of arrest or the probability of conviction variables are greater than one, indicating that more arrests were carried out than crimes committed, or more convictions than those arrested.

```
par(mar=c(0,2,1,0))
par(mfrow=c(1,2))
boxplot(crime$prbarr,
        col = "darkgreen", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Probability of Arrest")
boxplot(crime$prbconv,
        col = "darkblue", cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
        main = "Probability of Conviction")
```



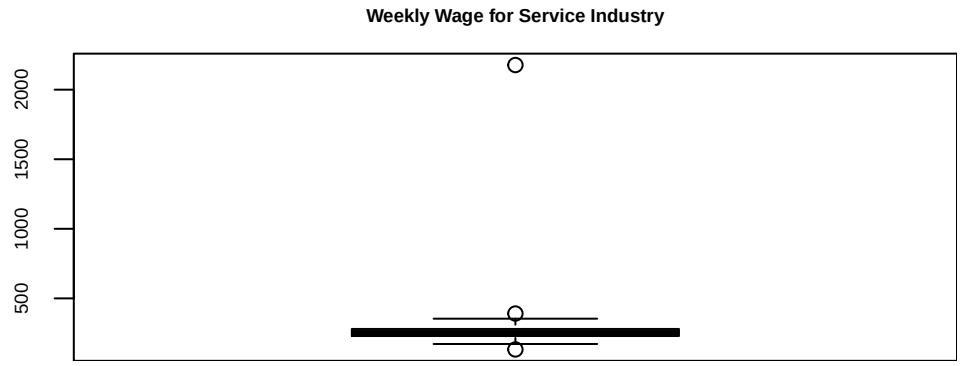
In case of probability of arrests, there is only one observation where the value is above 1, and it is significantly higher than the next closest value. It indicates that there have been more arrests than there have been crimes in a county. As this is time-limited data covering a single year, it is possible that crimes committed in the previous year and not recorded as a 1987

crime actually generated an arrest in 1987. Similarly, convictions may also have occurred in 1987, with the arrest relating to that conviction occurring in a prior period. It is not unfeasible that convictions for prior period arrests occur as the waiting time between being charged with an offense and a court date can be lengthy. Higher rates are an indication that a county is moving faster through its backlog.

Additionally, the table identifies some unusual features in some of the variables, including some significant outliers. Some of these outliers clearly appear to be inconsistent with the data and will be mentioned and corrected here; others may be more subtle and will be discussed as they are considered in models. Service industry wages and police per capita will be addressed in this section.

In the series of variables noting the weekly wages in a county, there is an exceptional value in one of the counties average wage, as seen in the below box-plot.

```
par(mar=c(0,2,2,0))
boxplot(crime$wser, cex.main = 0.6, cex.lab = 0.6, cex.axis = 0.6,
        main = "Weekly Wage for Service Industry",
        ylab = "Wage in $")
```



This one value is not only over over 9 standard deviations from the mean (as seen below) of wser wages, but greater than any other weekly wage value in the state.

```
 (max(crime$wser) - mean(crime$wser)) / sd(crime$wser)
```

```
## [1] 9.21935
```

The observation will need to be maintained, and therefore only the service weekly wage value will be replaced by an imputed value.

The result of a predictive model using the total of average weekly wages, with which the wages of the service sector are strongly correlated, is $211 per week. This is not dissimilar from the mean of $254 per week and therefore use of the mean as an imputed value is reasonable and simple. A new variable `wser_imp` is populated so that we do not lose the original values.

```
crime$wser_imp <- ifelse(crime$wser > 2000, mean(crime[crime$wser < 2000,]$wser), crime$wser)
summary(crime$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.7   253.2   275.6   280.5  2177.1
```

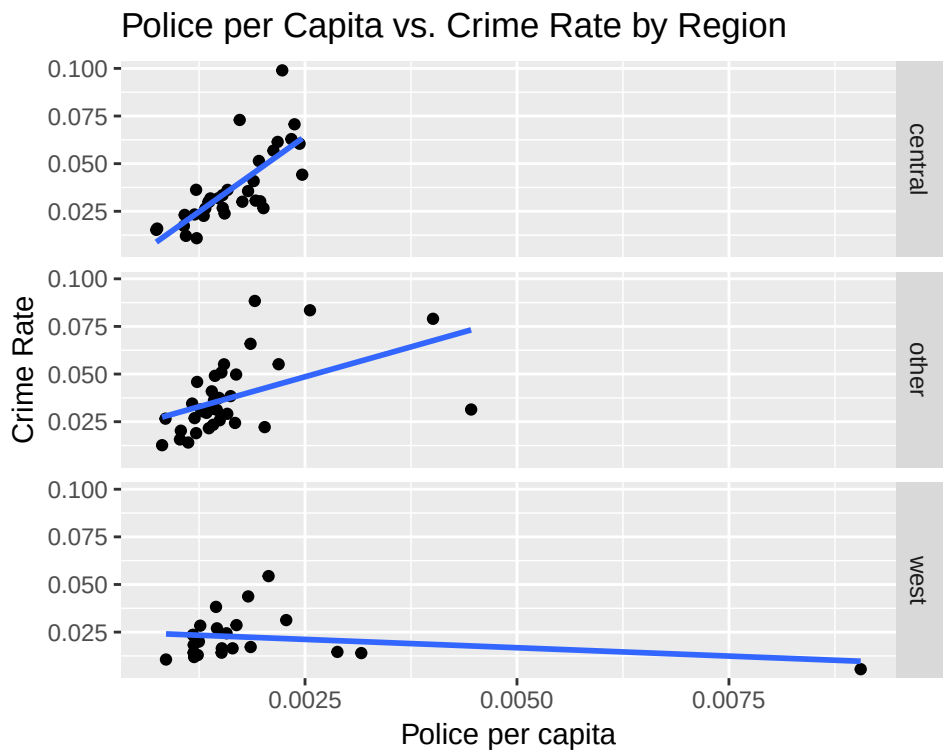
```
summary(crime$wser_imp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.7   253.2   254.4   277.2   391.3
```

The variable for police per capita (`polpc`) also has a notable outlier. This has generated some incongruent results with the rest of the dataset when segmented by region, as seen in the regression plots below.

```
crime$region <- ifelse(crime$west == 1, "west", ifelse(crime$central == 1, "central", "other"))
ggplot(crime, aes(polpc, crmrte)) +
geom_point() +
facet_grid(region~.) +
geom_smooth(method = "lm", se = FALSE) +
  xlab("Police per capita") +
```

```
  ylab("Crime Rate") +
  ggtitle("Police per Capita vs. Crime Rate by Region")
```



It is clear that the impact of this observation is significant to the trend of police per capita on crime rate. And, while perhaps possible, it is not representative of the rest of the population. Additionally, according to governing.com[1], police per population in Washington DC (where the highest concentration of police force might be expected) is 0.0065, significantly lower than this outlier.

Based on this analysis, the outlier will be recoded with the mean of `polpc` in the West region:

```
crime$polpc_imp <-
  ifelse(crime$polpc == max(crime$polpc),
         mean(crime[crime$west == 1 & crime$polpc < 0.009,]$polpc),
         crime$polpc)
summary(crime$polpc)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012308 0.0014853 0.0017022 0.0018768 0.0090543
```

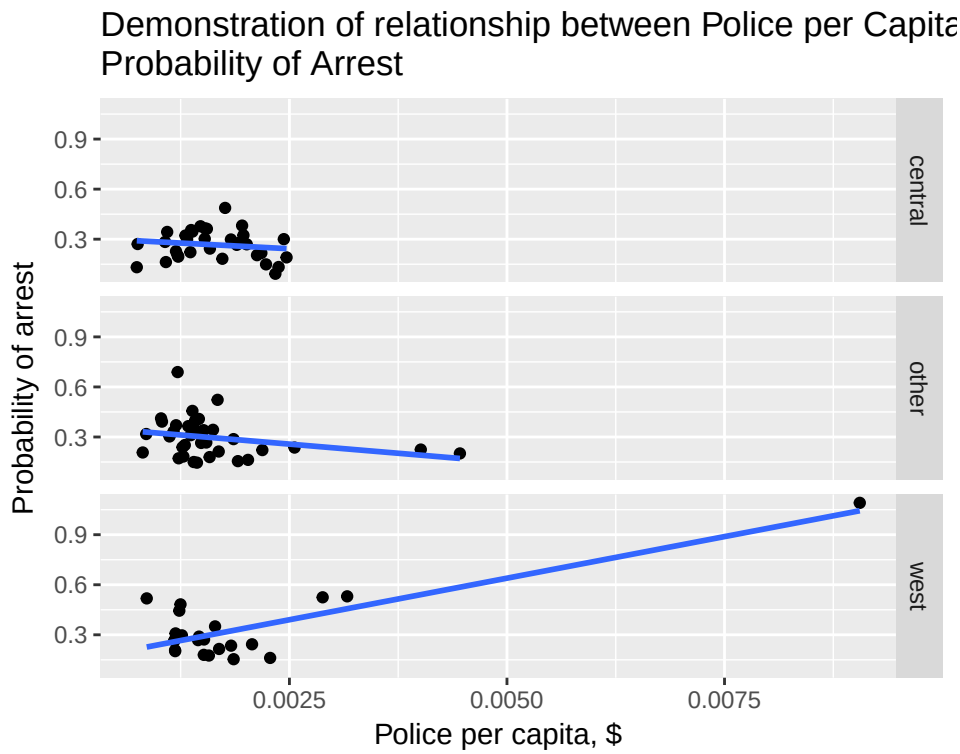
```
summary(crime$polpc_imp)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012308 0.0014853 0.0016204 0.0018583 0.0044592
```

Furthermore, this outlier in police per capita occurs in the same observation as the upper outlier in probability of arrest.

```
ggplot(crime, aes(polpc, prbarr)) +
  geom_point() +
  facet_grid(region~.) +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Police per capita, $") +
  ylab("Probability of arrest") +
  ggtitle(paste("Demonstration of relationship between Police per Capita and",
                "Probability of Arrest", sep = "\n"))
```

[1]The Offending, Crime and Justice Survey (2003); RECIDIVISM AMONG FEDERAL OFFENDERS: A COMPREHENSIVE OVERVIEW



The unrepresentativeness of these outliers can be demonstrated when correlation between two variables is compared. The correlation changes from positive to negative if the observation with the outlier is excluded:

```
cat("Correlation with the outlier inlcuded:",
    cor(crime$polpc, crime$prbarr), "\n")
```

```
## Correlation with the outlier inlcuded: 0.4264409
```

```
cat("Correlation with the outlier excluded:",
    cor(crime[-51,]$polpc, crime[-51,]$prbarr), "\n")
```

```
## Correlation with the outlier excluded: -0.1241811
```

Therefore, a new variable has been created for the probability of arrest with the mean of the variable in the west becoming the imputed value for the outlier:

```
crime$prbarr_imp <-
  ifelse(crime$prbarr > 1,
         mean(crime[crime$west == 1 & crime$prbarr < 1,]$prbarr),
         crime$prbarr)
summary(crime$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20568 0.27095 0.29492 0.34438 1.09091
```

```
summary(crime$prbarr_imp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20568 0.27095 0.28622 0.34323 0.68902
```

Though other variables appear to have exceptional values or outliers (particularly probability of arrest and the percent young male), none are as clear. These outliers will be addressed during the development of the models as appropriate and with due consideration for the practical significance and the leverage and influence they have on the models developed.

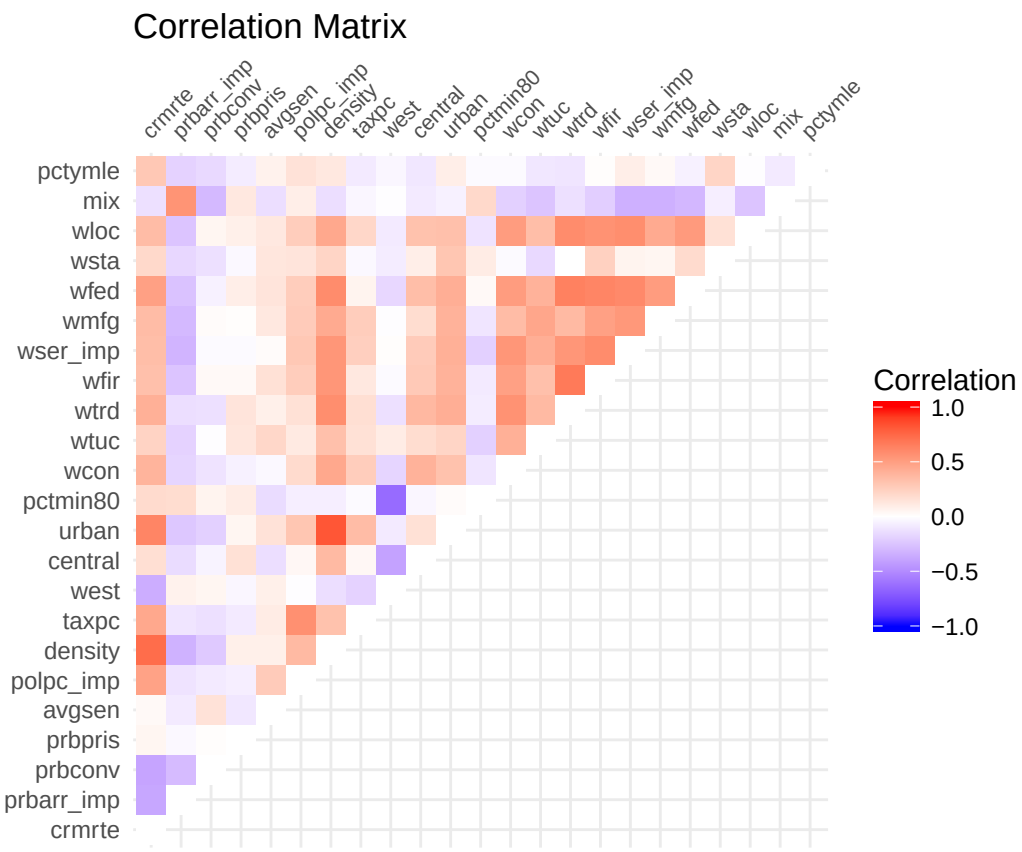
Correlations

To conclude our initial data exploration, an easy-to-reference correlation heatmap has been developed for quick identification of positive or negative correlations between variables in the data set.

```r
ind_variables <- c(
  "crmrte", "prbarr_imp", "prbconv", "prbpris", "avgsen", "polpc_imp",
  "density", "taxpc", "west", "central", "urban", "pctmin80", "wcon", "wtuc",
  "wtrd", "wfir", "wser_imp", "wmfg", "wfed", "wsta", "wloc", "mix", "pctymle"
)

cor_mat <- round(cor(crime[,ind_variables]),2)
get_upper_tri <- function(cor_mat){
    cor_mat[lower.tri(cor_mat)]<- NA
    return(cor_mat)
}
cor_mat_upper <- get_upper_tri(cor_mat)
cor_mat_upper2 <- melt(cor_mat_upper, na.rm = TRUE)
cor_mat_upper2[cor_mat_upper2$value == 1,]$value <- 0

ggplot(data = cor_mat_upper2, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  scale_x_discrete(position = "top") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 8, hjust = 0),
        axis.title.x=element_blank(),
        axis.title.y=element_blank()) +
  coord_fixed() +
  ggtitle("Correlation Matrix")
```



Based on the above matrix a few important patterns have been identified:

- `density` has the strongest correlation with `crmrte`. Hence, it would be one of the most important variables to test in the models.

- All wages variables are positively correlated with each other. Hence, it would create challenges for keeping multiple variables in the model.
- `density` has a strong negative correlation with urban indicator. This relationship is as expected. Given that `density` has higher correlation with `crmrte`, it might be enough to control for the level of urbanization in the area in our models.

Summary of variables

The table below summarizes all variables in the dataset, and includes the expected impact of each on the dependent variable, crime rate, along with the actual correlation. Also included, as a framework for the analysis, is an assessment of the rapidity at which policy could be enacted and be effective. The support and lobbying for judges with perspectives that would support policies relating to custodial terms and their length, could be implemented quickly. However developing incentives and strategies to reduce population density, for example, will take a longer time to generate results.

```
var_labels <- c("crimes committed per person", "probability of arrest",
  "probability of conviction", "probability of prison sentence",
  "avg. sentence, days", "police per capita", "people per sq. mile",
  "tax revenue per capita", "=1 if in western N.C.", "=1 if in central N.C.",
  "=1 if in SMSA", "perc. minority, 1980", "weekly wage, construction",
  "wkly wge, trns, util, commun", "wkly wge, whlesle, retail trade",
  "wkly wge, fin, ins, real est", "wkly wge, service industry",
  "wkly wge, manufacturing", "wkly wge, fed employees",
  "wkly wge, state employees", "wkly wge, local gov emps",
  "offense mix: face-to-face/other", "percent young male")

impact <- c("Dependent", "Negative" , "Negative", "Negative", "Negative",
  "Negative", "Positive", "Negative", "Unclear", "Unclear", "Unclear",
  "Unclear", "Negative","Negative","Negative", "Negative", "Negative",
  "Negative", "Negative", "Negative", "Negative", "Unclear","Positive")

control <- c("NA", "Medium Term", "Medium Term", "Short Term", "Short Term",
  "Medium Term", "Long Term", "Long Term",
  "No", "No", "No", "Long Term",
  "Medium Term", "Medium Term", "Medium Term",
  "Medium Term", "Medium Term", "Medium Term", "Medium Term",
  "Short Term", "Medium Term", "No", "Long Term")

cor_w_crimerate <- round(cor(crime[,ind_variables])[1,],2)
desc <- data.frame(ind_variables, var_labels, impact, cor_w_crimerate, control,
                   row.names = NULL)
colnames(desc) <- c("Explanatory Variables", "Explanation",
  "Expected Impact on Crime Rate", "Correlation w/ Crime Rate",
  "Potential Policy Impact Timeframe")

kable(desc, booktabs = TRUE, align = c("llccc")) %>%
  kable_styling(latex_options = c("scale_down"), full_width = FALSE) %>%
  row_spec(0, bold = TRUE) %>%
  column_spec(1, width = "7em") %>%
  column_spec(3, width = "10em") %>%
  column_spec(4, width = "8em") %>%
  column_spec(5, width = "10em")
```

Explanatory Variables	Explanation	Expected Impact on Crime Rate	Correlation w/ Crime Rate	Potential Policy Impact Timeframe
crmrte	crimes committed per person	Dependent	1.00	NA
prbarr_imp	probability of arrest	Negative	-0.38	Medium Term
prbconv	probability of conviction	Negative	-0.39	Medium Term
prbpris	probability of prison sentence	Negative	0.05	Short Term
avgsen	avg. sentence, days	Negative	0.03	Short Term
polpc_imp	police per capita	Negative	0.48	Medium Term
density	people per sq. mile	Positive	0.73	Long Term
taxpc	tax revenue per capita	Negative	0.45	Long Term
west	=1 if in western N.C.	Unclear	-0.35	No
central	=1 if in central N.C.	Unclear	0.17	No
urban	=1 if in SMSA	Unclear	0.62	No
pctmin80	perc. minority, 1980	Unclear	0.19	Long Term
wcon	weekly wage, construction	Negative	0.39	Medium Term
wtuc	wkly wge, trns, util, commun	Negative	0.23	Medium Term
wtrd	wkly wge, whlesle, retail trade	Negative	0.41	Medium Term
wfir	wkly wge, fin, ins, real est	Negative	0.33	Medium Term
wser_imp	wkly wge, service industry	Negative	0.34	Medium Term
wmfg	wkly wge, manufacturing	Negative	0.35	Medium Term
wfed	wkly wge, fed employees	Negative	0.49	Medium Term
wsta	wkly wge, state employees	Negative	0.20	Short Term
wloc	wkly wge, local gov emps	Negative	0.35	Medium Term
mix	offense mix: face-to-face/other	Unclear	-0.13	No
pctymle	percent young male	Positive	0.29	Long Term

The Model Building Process

Overview

As we are moving into model building section of the report, let's outline our objective: identify the impact of causal variables on crime rate to build crime-fighting policies. What are the causal variables of interest in this case? We hypothesize that in this dataset there are two variables that cause crime rate to increase/decrease: probability of arrest and probability of conviction. The third probability variable, `prbpris`, has a weak correlation with crime rate. Most likely this is due to the fact that prison sentence is far enough from the act of a crime to be ineffective in altering criminal behavior.

Our first model will be developed with these two variables along with two control variables that will help us to get unbiased estimates of our main variables of interest (explained in the appropriate section).

Our second model will expand on the first one. We will add variables that help us improve the fit of the model without interacting significantly with our main causal effects. The added variables also make sense in term of interpretability.

The third model will contain all provided variables (except county and year). This model will be used to demonstrate that our model #2 is robust.

The last part of this report will focus on residuals of all three models.

Dependent variable

Our dependent variable is crime rate (`crmrte`), which is defined as "Crimes committed per person."

After careful consideration, in order for us to understand the impact of our main causal effects (probability of arrest and probability of conviction) onto crime rate, we decided to transform our dependent variable by taking a natural log.

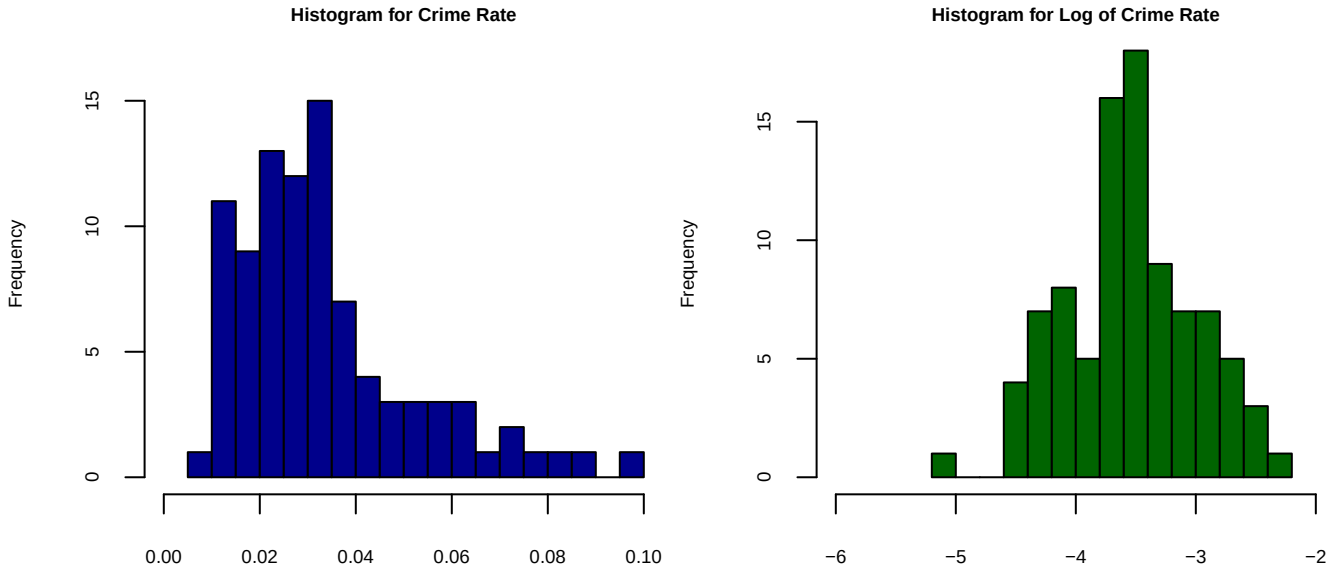
Since this variable is a ratio (crimer per person), hypothetically it can vary between 0 and 1 (though it's highly unlikely to find a county with such a high crime rate). This makes it not very suitable for OLS because this method can predict values outside 0 to 1 range. Natural log will help us only with part of the problem (avoiding negative values in prediction of actual crime rate). Caveat: in our dataset crime rate variable is never equal to zero. Hence, transformation is straight forward.

However, since zero is a real possible value, we would need to watch out for those values while transforming crime rate in different datasets.

This transformation would also allow us to interpret the coefficients of predictive factors as semi-elasticities: if probability of arrest goes up by one point, then the crime rate decreases by `100*y%` (assuming our stated hypothesis is true and the probability of arrest `prbarr` has a negative effect). If we were to keep the variable as is, we would interpret the coefficient for `prbarr` as: if probability of arrest goes up by one point, then the crime rate decreases by `y` crimes per person. However, this interpretation does not allow us to judge the practical significance of the effect (is that `y` big or small?).

Let's take a look at histograms for `crmrte` (as it is and transformed):

```
par(mar=c(2,4,1,0))
par(mfrow=c(1,2))
hist(crime$crmrte,
     breaks = 15, xlim = c(0,0.1), ylim = c(0,17), col = "darkblue",
     cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
     xlab = "Crime Rate",
     main = "Histogram for Crime Rate")
hist(log(crime$crmrte),
     breaks = 15, xlim = c(-6,-2), col = "darkgreen",
     cex.main = 0.6, cex.axis = 0.6, cex.lab = 0.6,
     xlab = "Log of Crime Rate",
     main = "Histogram for Log of Crime Rate")
```



Based on the above charts, `crmrte` is skewed towards the right tail (a number of counties has large crime rates). The log of `crmrte`, on the other hand, looks normally distributed. This definition of the dependent variables will help us build a model with a better fit.

Main control variables

Our primary focus in this analysis is two variables: `prbarr` and `prbconv`. These two variables, the probability of arrest and the probability of conviction respectively, have relatively high correlation with crime rate and have potential to be influenced by political action. We will try to understand how probability of arrest `prbarr` and probability of conviction `prbconv` impact crime rate. If they are strong causal factors, we can define policies that influence these two factors and, hence, help us lower crime rates across North Carolina.

Earlier in this report, we hypothesised that these two variables will have a negative impact on our dependent variable: the higher the probabilities of arrest and conviction, the lower the crime rate. Before building a model with these two variables, however, we want to make a case of including two more variables in our first model: `density` and `west`.

First, let's consider crime rate by region (we recoded the third region as "other" for analysis purposes):

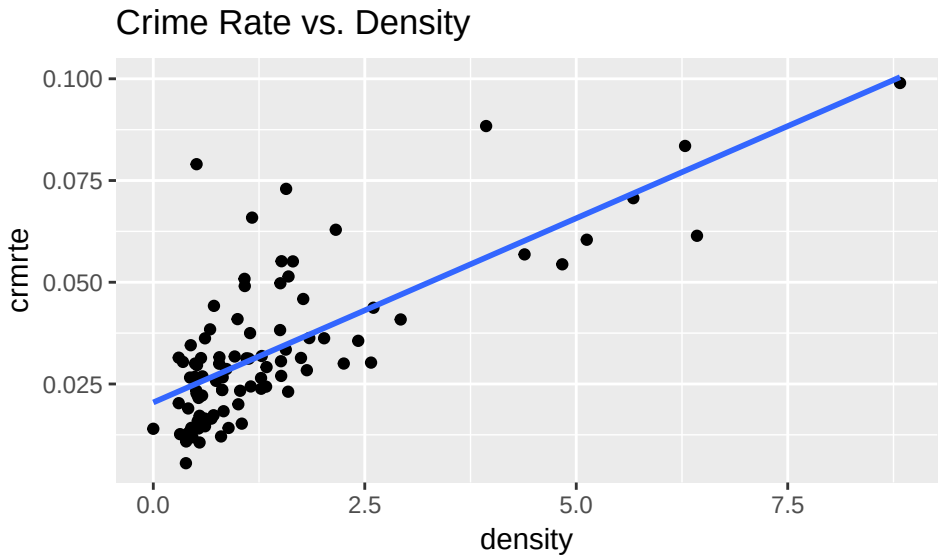
```
crime$region <- ifelse(crime$west == 1, "west",
                       ifelse(crime$central == 1, "central", "other"))
aggregate(crmrte ~ region, data = crime, mean)
```

```
##    region     crmrte
## 1 central 0.03699627
## 2   other 0.03739491
## 3    west 0.02216183
```

Based on the table above, crime rate in the West region is lower than in the Central and Other regions. We therefore need to control for regionality in order to get an unbiased read on the two selected probability variables.

On the other hand, density has the highest correlation with crime rate (0.73). And the chart below shows clear support for a strong linear relationship between the two variables:

```
ggplot(crime, aes(density, crmrte)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Crime Rate vs. Density")
```



We also know that `west` and `density` have a different relationship with `crmrte`; even though crime rate is the lowest in the West, density is the highest in the Central region. Hence, we need both `west` and `density` in our initial model to get unbiased estimates of `prbarr` and `prbconv`.

```
aggregate(density ~ region, data = crime, mean)
```

```
##    region  density
## 1 central 2.047960
## 2   other 1.085503
## 3    west 1.062994
```

Note: we tested *central* and *urban* in our models and they were not significant predictors for crime rate.

Model #1

Our first model contains four variables: `density`, `west`, `prbarr`, `prbconv`.

However, first, since the two probability variables are on the scale from 0 and 1 (except the outliers), we will multiply them both by 100 to change the scale to 0 to 100. This will allow the interpretation to be: the percent change in crime rate per one point change in probability.

```
crime$prbarr_imp100 <- 100*crime$prbarr_imp
crime$prbconv100 <- 100*crime$prbconv
```

```r
model1.ind_vars <- c("density", "west", "prbarr_imp100", "prbconv100")
model1.formula <- as.formula(paste("log(crmrte) ~",
                                   paste(model1.ind_vars, collapse = " + "),
                                   sep = " "))
model1 <- lm(model1.formula, data = crime)
```

```r
interpret1 <- c("","For each person per square mile increase in density,
                crime rate increases by 13.7% when everything else stays the same",
                "Crime rate in the West is 39.4% lower than in Central and Other
                regions on average (and controlling for all other included variables)",
                "For approximately each percentage increase in probability of arrest,
                crime rate decreases by 1.69%",
                "For approximately each percentage increase in probability of arrest,
                crime rate decreases by 0.69%")

coef1 <- data.frame("Model 1 Coefficients" = round(model1$coefficients,4),
                    "Interpretation" = interpret1)

kable(coef1, booktabs = TRUE) %>%
  kable_styling(font_size = 8, full_width = FALSE) %>%
      column_spec(3, width = "35em")
```

	Model.1.Coefficients	Interpretation
(Intercept)	-2.7790	
density	0.1370	For each person per square mile increase in density, crime rate increases by 13.7% when everything else stays the same
west	-0.3939	Crime rate in the West is 39.4% lower than in Central and Other regions on average (and controlling for all other included variables)
prbarr_imp100	-0.0169	For approximately each percentage increase in probability of arrest, crime rate decreases by 1.69%
prbconv100	-0.0069	For approximately each percentage increase in probability of arrest, crime rate decreases by 0.69%

Our model is consistent with our initial hypothesis: both probability variables have a negative impact on crime rate. Moreover, a one-point change in probability of arrest has almost three times larger impact on crime rate than one-point change in probability of conviction. This confirms our hypothesis that probability of arrest has a stronger effect on crime rate because it's closer to the act of crime (being arrested is easier to relate to than being convicted).

The adjusted R^2 for this model is 67.2%:

```r
summary(model1)$adj.r.squared
```

```
## [1] 0.6716899
```

All of the coefficients are highly statistically significant when we look at heteroskedastic-robust errors:

```r
coeftest(model1, vcov = vcovHC, level = 0.05)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)   -2.7789646  0.2021670 -13.7459 < 2.2e-16 ***
## density        0.1369895  0.0269657   5.0801 2.167e-06 ***
## west          -0.3939265  0.0748157  -5.2653 1.018e-06 ***
## prbarr_imp100 -0.0168814  0.0037203  -4.5376 1.834e-05 ***
## prbconv100    -0.0068575  0.0014410  -4.7589 7.789e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: we will analyze the residuals later on, after we develop all three models.

Model #2

We tested most of other variables in the dataset that potentially could be related to crime rate (using correlations). At the end we decided to add the following variables: `polpc`, `pctmin80`, and an interaction of `west` and `polpc`. Police per capita is not only highly correlated with crime rate, but it does seem to have direct link to crime. What we see from correlation analysis is counter intuitive at first glance: the higher police per capita, the higher crime rate. Logically, we would expect that increasing police presence would decrease crime rate over time. However, this dataset is panel data at one moment in time, not a time series. Hence, counties with higher police per capita require more police presence. Therefore, this variable is necessary for control purposes and it improves the fit of the model.

Percent minorities also helps with model fit. It is hard to hypothesize why correlation with crime rate is positive. Do poorer counties have larger minority populations? In this case, personal income would be a confounding variable that we don't have. Or counties with more minorities have more gangs (on the ethnic basis)? This would also be a confounding variable. In any case, percent minorities will be used as representative of omitted factors.

Before adding these variables, though, we decided to transform the `polpc` variable by taking its log. We perform this transformation for two reasons: there is a stronger correlation between the log of police per capita and the log of the crime rate variable than that of the raw values or of the log of the crime rate and the raw police per capita. This tells us that there is a better correlation between percent changes in the two variables than the raw changes. Performing this transformation also improves the quality of our regression model. The transformation is performed here:

```
crime$polpc_imp.ln <- log(crime$polpc_imp)
```

Correlation is strongest between the logs of both variables:

```
descriptions <- c(
  "log(crmrte), log(polpc)",
  "log(crmrte), polpc",
  "crmrte, polpc"
)
correlations <- c(
  round(cor(crime$polpc_imp.ln, log(crime$crmrte)), 4),
  round(cor(crime$polpc_imp, log(crime$crmrte)), 4),
  round(cor(crime$polpc_imp, crime$crmrte), 4)
)
crmrte.polpc.cor <- data.frame(descriptions, correlations)
kable(crmrte.polpc.cor, col.names = c("Variables", "Correlation"),
      booktabs = TRUE)
```

Variables	Correlation
log(crmrte), log(polpc)	0.5099
log(crmrte), polpc	0.4338
crmrte, polpc	0.4791

We also combine `west` and the transformed `polpc` (`polpc_imp.ln`) by multiplying them. Why do we add this interaction? Because the police per capita in the West region has much lower correlation with crime rate than police per capita in the other regions, as is shown here:

```
region <- c("West",
            "Central",
            "Other")

region_cor <- c(
  round(cor(log(crime[crime$region=="west",]$crmrte),
            crime[crime$region=="west",]$polpc_imp.ln),2),
  round(cor(log(crime[crime$region=="central",]$crmrte),
            crime[crime$region=="central",]$polpc_imp.ln),2),
  round(cor(log(crime[crime$region=="other",]$crmrte),
            crime[crime$region=="other",]$polpc_imp.ln),2)
)
cor.by.region <- data.frame(region, region_cor)
kable(cor.by.region, col.names = c("Region", "Correlation"),
      booktabs = TRUE)
```

Region	Correlation
West	0.17
Central	0.80
Other	0.59

Now, to the actual model:

```
model2.ind_vars <- c("density", "west", "prbarr_imp100", "prbconv100",
                     "polpc_imp.ln", "pctmin80", "west*polpc_imp.ln")
model2.formula <- as.formula(paste("log(crmrte) ~ ",
                                   paste(model2.ind_vars, collapse = " + "),
                                   sep = ""))
model2 <- lm(model2.formula, data = crime)
```

```
interpret2 <- c(
  "",
  "(Before: 0.14): The effect of density has decreased as we are controlling for
  more factors. For each person per square mile increase in density,
  crime rate increases by 9.9%",
  "(Before: -0.39): This coefficient cannot be interpreted by itself
  as it is now part of interaction with police per capita.
  See explanation below",
  "(Before: -0.0169): The probability of arrest has a
  stronger effect.
  A single percentage increase in the probability of arrest results
  in a 2.02% decrease in the crime rate",
  "(Before: -0.0069): The effect of the probability of conviction
  has also increased slightly. For approximately each percentage increase in the
  probability of arrest, crime rate decreases by 0.74%",
  "This coefficient indicates that a 1% increase in police per capita is associated
  with a 0.63% increase in crime rate in all regions but West",
  "This coefficient indicates that 1 percent point increase in the minority
  population means a 0.99% increase in crime per capita",
  "See explanation below")

coef2 <- data.frame("Model 2 Coefficients" = round(model2$coefficients, 4),
                    "Interpretation" = interpret2)

kable(coef2, booktabs = TRUE) %>%
  kable_styling(font_size = 8, full_width = FALSE) %>%
      column_spec(3, width = "35em")
```

	Model.2.Coefficients	Interpretation
(Intercept)	1.2282	
density	0.0889	(Before: 0.14): The effect of density has decreased as we are controlling for more factors. For each person per square mile increase in density, crime rate increases by 9.9%
west	-4.2426	(Before: -0.39): This coefficient cannot be interpreted by itself as it is now part of interaction with police per capita. See explanation below
prbarr_imp100	-0.0202	(Before: -0.0169): The probability of arrest has a stronger effect. A single percentage increase in the probability of arrest results in a 2.02% decrease in the crime rate
prbconv100	-0.0074	(Before: -0.0069): The effect of the probability of conviction has also increased slightly. For approximately each percentage increase in the probability of arrest, crime rate decreases by 0.74%
polpc_imp.ln	0.6360	This coefficient indicates that a 1% increase in police per capita is associated with a 0.63% increase in crime rate in all regions but West
pctmin80	0.0099	This coefficient indicates that 1 percent point increase in the minority population means a 0.99% increase in crime per capita
west:polpc_imp.ln	-0.6296	See explanation below

The interaction term (`west:polpc_imp.ln`) is harder to interpret. It applies only to the West region. That means that

`polpc_imp.ln` coefficient of 0.636 applies only to Central and Other regions. In the West the coefficient for `polpc_imp.ln` is actually 0.0064 (0.6360 - 0.6296) or per each percent change in police per capita in the West, there is only 0.0064% change in crime rate. This value is close to zero and implies no significant relationship between two variables in that region. This is supported by correlations we examined earlier in this section.

As for the `west` coefficient, it also cannot be interpreted in isolation because setting `polpc_imp.ln` to zero does not make practical sense. If on the other hand, we use the mean of `polpc_imp.ln` for the West region then the partial effect for `west` is as follows:

```
mean_polpc <- mean(crime[crime$region=="west",]$polpc_imp.ln)
coef_west_polpc <- model2$coefficients["west:polpc_imp.ln"]
coef_west <- model2$coefficients["west"]
coef_west + coef_west_polpc*mean_polpc
```

```
##       west
## -0.1644222
```

This second model remains consistent with our initial hypothesis. The overall predictive strength of the model has also increased. The adjusted R^2 for this model is 80.1%, which is 12.9 percentage points higher than our first model:

```
summary(model2)$adj.r.squared
```

```
## [1] 0.8009718
```

All of the coefficients are statistically significant when we look at heteroskedasticity-robust errors:

```
coeftest(model2, vcov = vcovHC, level = 0.05)
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         1.2282470  0.8071212  1.5218 0.1318694
## density             0.0889342  0.0247597  3.5919 0.0005551 ***
## west               -4.2425746  1.2678084 -3.3464 0.0012317 **
## prbarr_imp100      -0.0202374  0.0032741 -6.1810 2.275e-08 ***
## prbconv100         -0.0074262  0.0011860 -6.2617 1.602e-08 ***
## polpc_imp.ln        0.6360198  0.1208091  5.2647 1.083e-06 ***
## pctmin80            0.0099089  0.0021410  4.6282 1.347e-05 ***
## west:polpc_imp.ln -0.6295556  0.1970078 -3.1956 0.0019741 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model #3 - All Variables

Finally, our last model includes all variables, including our imputed variables. We transform all wage variables by taking their natural log. This will allow us to interpret the coefficients as elasticities instead of using absolute wage changes.

```
wage.vars <- c("wcon", "wtuc", "wtrd", "wfir", "wser_imp", "wmfg", "wfed",
               "wsta", "wloc")
wage.vars.ln <- mapply(function(var.name) paste(var.name, ".ln", sep=""),
                       wage.vars)
crime[, wage.vars.ln] <- log(crime[, wage.vars])
```

And then we create our third model, which includes all of variables, transformed as needed:

```
model3.ind_vars <- c("prbarr_imp100", "prbconv100", "prbpris", "avgsen",
                     "polpc_imp.ln", "density", "taxpc", "west", "central",
                     "urban", "pctmin80", "wcon.ln", "wtuc.ln", "wtrd.ln",
                     "wfir.ln", "wser_imp.ln", "wmfg.ln", "wfed.ln", "wsta.ln",
                     "wloc.ln", "mix", "pctymle")
model3.formula <- as.formula(paste("log(crmrte) ~ ",
                                   paste(model3.ind_vars, collapse = " + "),
                                   sep = ""))
```

```
model3 <- lm(model3.formula, data = crime)
summary(model3)$adj.r.squared
```

```
## [1] 0.8233825
```

Despite adding a lot more variables, the R^2 of the all-inclusive regression model went up only to 82.3% (from 80.1% in model 2). That means that additional variables did not contribute much to the model fit.

Now let's compare the coefficients in this model to the other two models:

```
se.model1 <- sqrt(diag(vcovHC(model1)))
se.model2 <- sqrt(diag(vcovHC(model2)))
se.model3 <- sqrt(diag(vcovHC(model3)))
stargazer(model1, model2, model3,
          type = "text", omit.stat = "f",
          se = list(se.model1, se.model2, se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          no.space = TRUE, align = TRUE)
```

```
##
## ========================================================================
##                                  Dependent variable:
##                      ---------------------------------------------------
##                                        log(crmrte)
##                            (1)             (2)             (3)
## ------------------------------------------------------------------------
## density                 0.137***        0.089***         0.107
##                          (0.027)         (0.025)         (0.056)
## taxpc                                                    0.001
##                                                          (0.007)
## west                    -0.394***       -4.243***        -0.200
##                          (0.075)         (1.268)         (0.116)
## central                                                  -0.161*
##                                                          (0.076)
## urban                                                    -0.120
##                                                          (0.224)
## prbarr_imp100           -0.017***       -0.020***       -0.017***
##                          (0.004)         (0.003)         (0.003)
## prbconv100              -0.007***       -0.007***       -0.007***
##                          (0.001)         (0.001)         (0.001)
## prbpris                                                  -0.120
##                                                          (0.423)
## avgsen                                                   -0.024
##                                                          (0.015)
## polpc_imp.ln                            0.636***         0.515**
##                                          (0.121)         (0.189)
## pctmin80                                0.010***         0.008**
##                                          (0.002)         (0.003)
## west:polpc_imp.ln                       -0.630**
##                                          (0.197)
## wcon.ln                                                  0.299
##                                                          (0.235)
## wtuc.ln                                                  0.163
##                                                          (0.287)
## wtrd.ln                                                  0.252
##                                                          (0.314)
## wfir.ln                                                  -0.164
##                                                          (0.341)
## wser_imp.ln                                              -0.526
##                                                          (0.310)
## wmfg.ln                                                  -0.042
```

```
##                                                          (0.164)
## wfed.ln                                                   0.795
##                                                          (0.420)
## wsta.ln                                                  -0.324
##                                                          (0.323)
## wloc.ln                                                   0.096
##                                                          (0.643)
## mix                                                      -0.595
##                                                          (0.536)
## pctymle                                                   2.484
##                                                          (1.325)
## Constant                -2.779***         1.228          -2.880
##                          (0.202)         (0.807)         (4.196)
## ------------------------------------------------------------------
## Observations               91              91              91
## R2                        0.686           0.816           0.867
## Adjusted R2               0.672           0.801           0.823
## Residual Std. Error 0.313 (df = 86) 0.244 (df = 83) 0.230 (df = 68)
## ==================================================================
## Note:                                *p<0.05; **p<0.01; ***p<0.001
```

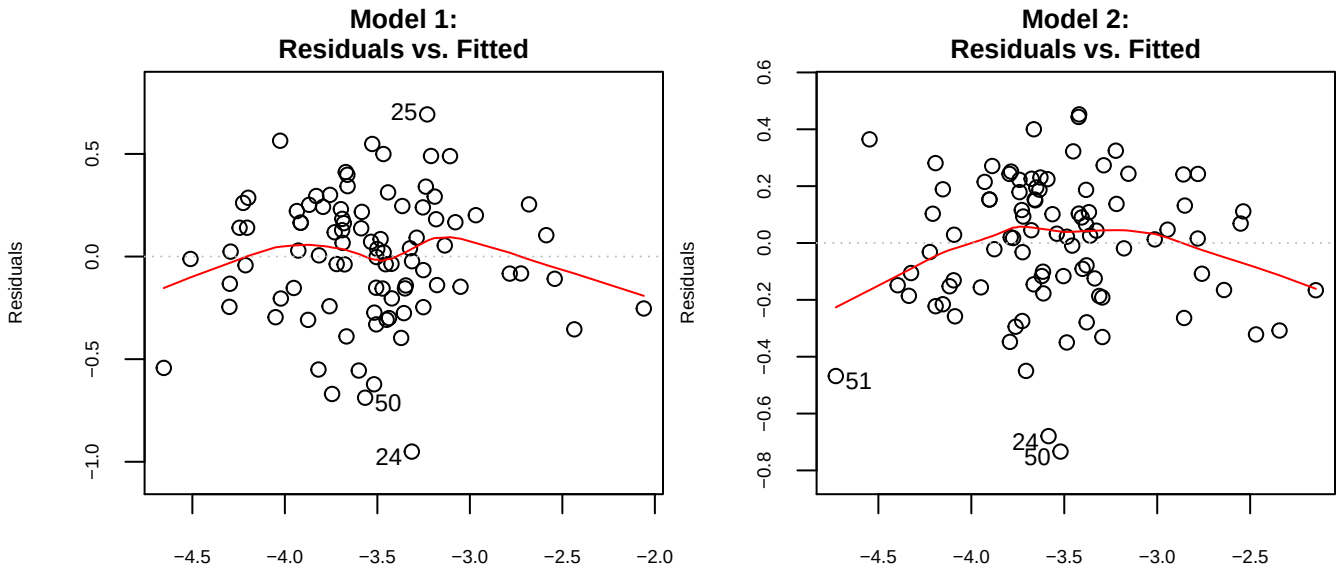
This table demonstrates the following:

- Our main variables of interest `prbarr` and `prbconv` have robust estimates in all models. Moreover, `prbarr` coefficient is consistently higher than for `prbconv`. Hence, it is more likely to cause crime rate to change.
- The coefficient for `density` also remains robust and doesn't change between model 2 and 3 significantly
- `polpc` remains a strong variable even without interaction in model #3. However, its interaction with *west* helps model #2 significantly.
- `pctmin80` is also robust as its coefficient stays statistically significant in model #3
- All other variables are not statistically significant in model #3. The only exception is *central*. However, when we tested this variable in model #2 it didn't maintain its significance. Hence, it is highly sensitive to model definition

Residual Analysis

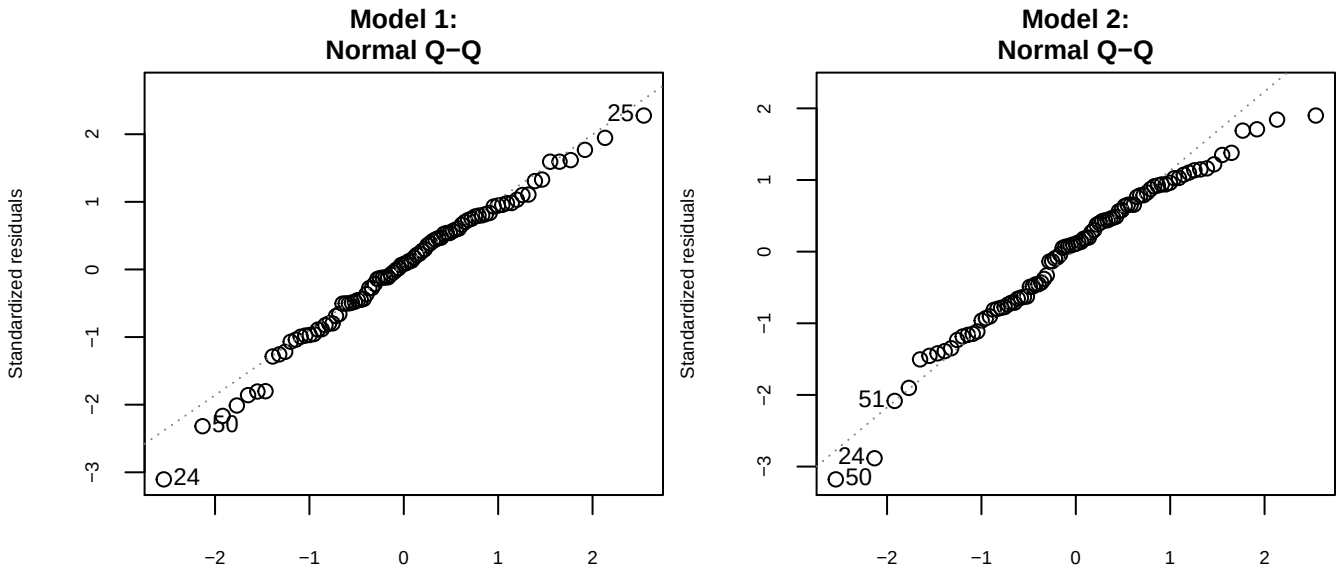
From the first plot (residuals vs. fitted values), it is evident that there are some instances with very low and very high crime rates that make residuals deviate from 0. Interestingly enough, observation #51 still appears as an outlier (that's where we recoded *polpc* and *prbarr*). It might be that this observation needs to be excluded from the anlaysis in future itirations.

```
par(mar=c(2,4,2,0))
par(mfrow=c(1,2))
plot(model1, which = 1, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 1:\nResiduals vs. Fitted", caption = "")
plot(model2, which = 1, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 2:\nResiduals vs. Fitted", caption = "")
```



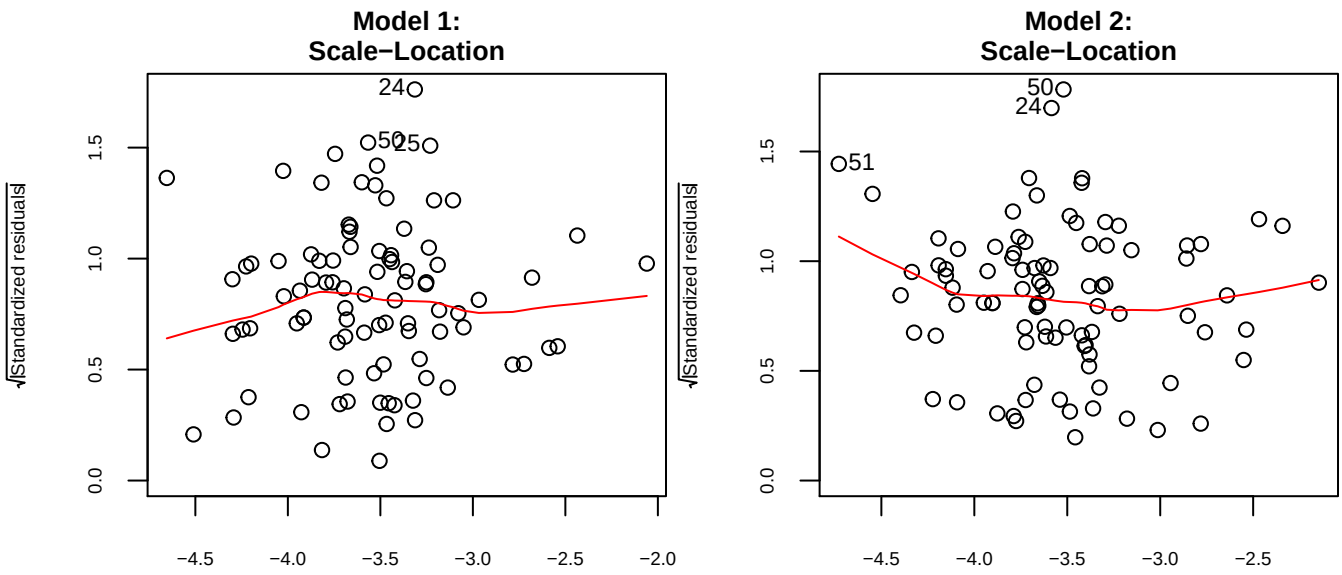
Similar findings apply to Normal Q-Q plot below. However, most significant deviation from normality is observed for counties with high crime rates in model #2. This pattern is not as evident in model #1. Hence, more investigation need to go into which variables in model #2 contribute to this deviation.

```
par(mar=c(2,4,2,0))
par(mfrow=c(1,2))
plot(model1, which = 2, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 1:\nNormal Q-Q", caption = "")
plot(model2, which = 2, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 2:\nNormal Q-Q", caption = "")
```



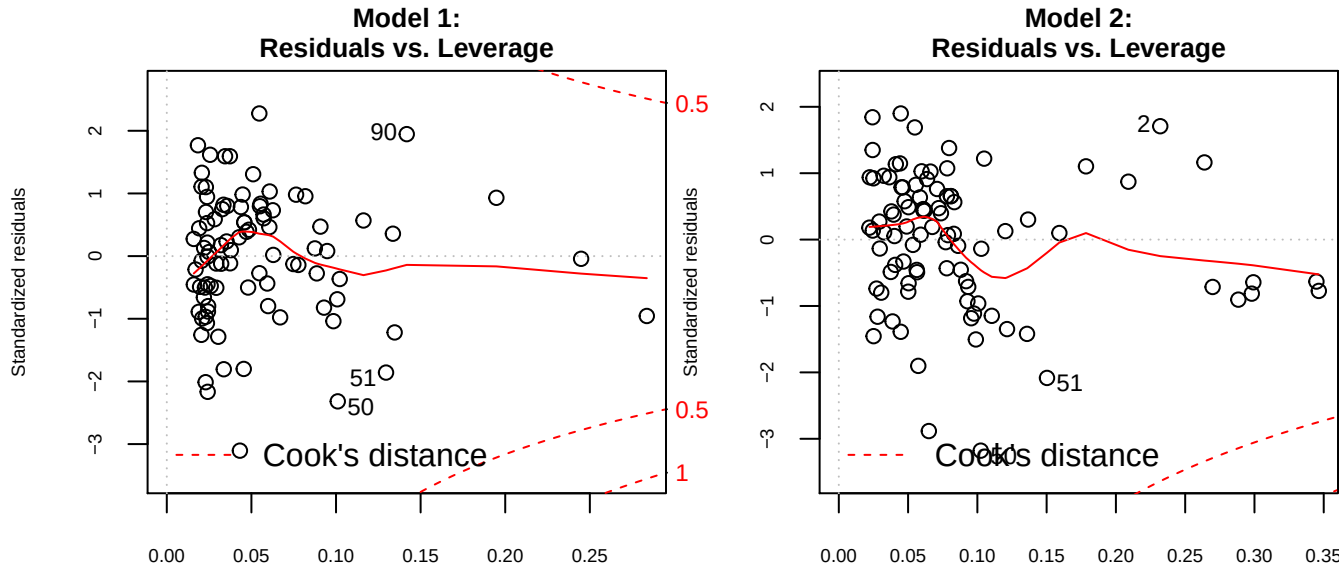
In Scale-Location plot, again observation #51 is causing some heteroskedasticity in residuals for model #2. In model #1 standard deviation of residuals, on the other hand, looks constant.

```
par(mar=c(2,4,2,0))
par(mfrow=c(1,2))
plot(model1, which = 3, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 1:\nScale-Location", caption = "")
plot(model2, which = 3, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 2:\nScale-Location", caption = "")
```



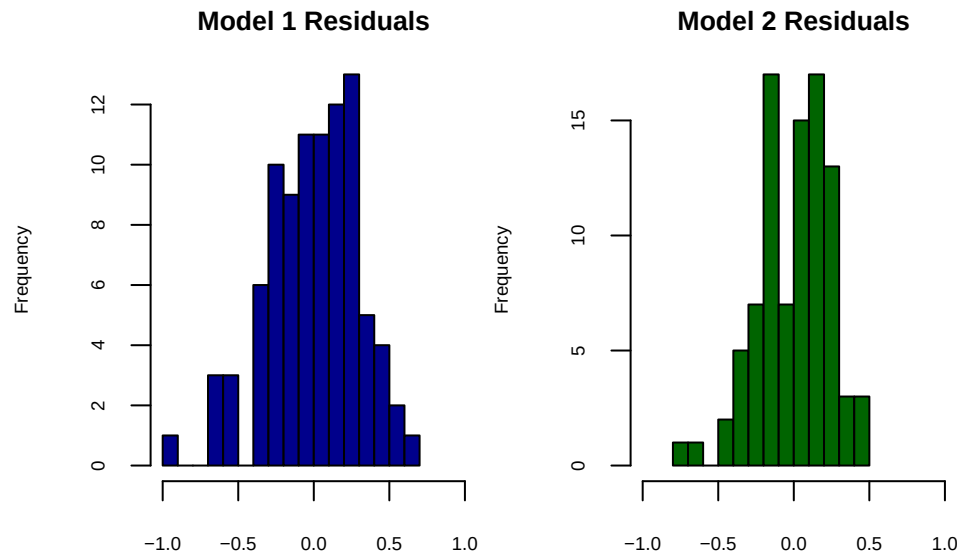
However, none of the observations stands out as an outlier, even #51 in the Leverage charts below.

```
par(mar=c(2,4,2,0))
par(mfrow=c(1,2))
plot(model1, which = 5, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 1:\nResiduals vs. Leverage", caption = "")
plot(model2, which = 5, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 2:\nResiduals vs. Leverage", caption = "")
```



Our dataset contains more than 30 observations (91 to be precise). Hence, we can apply CLT to its residuals and assume they are normal. However, we still investigate the histograms for model #1 and #2 below. Both of them do not resemble normal distribution particularly well.

```
par(mar=c(2,4,2,0))
par(mfrow=c(1,2))
hist(model1$residuals,
     breaks = 15,  col = "darkblue", xlim = c(-1,1),
     cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6, xlab = "",
     main = "Model 1 Residuals")
hist(model2$residuals,
     breaks = 15, col = "darkgreen", xlim = c(-1,1),
     cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6, xlab = "",
     main = "Model 2 Residuals")
```



Overall, the residual analysis showed that our models have some issues with fitting certain values of crime rate (on the high end). Also, we can see some heteroskedasticity in residuals caused by a few observations. Further investigation will be conducted to analyze the reasons and potentially remove them.

Omitted Variable Analysis

Drug and alcohol abuse levels

The presence of drug and alcohol problems in a community is a significant contributing factor to crime rates in many areas. There is an expectation that less affluent communities in urban areas would be most impacted, which may explain some of the higher rates of crime in higher density populated counties and bias the coefficients of density in our model. These coefficients may have factors related to drug and alcohol abuse included in them.

Unreported crime

The stigma of some crimes for victims within a community, the feeling that nothing will be done to catch the perpetrators or perhaps vigilante justice may lead to crimes in some areas being under reported. Sexual assaults specifically can be difficult for victims to report for fear of community isolation or reprisals in smaller communities. The presence of gangs, undocumented immigrants or local judicial services being overwhelmed and unavailable may be a cause in some more urban areas. The presence of unreported crimes will impact the probability of arrest as the total number of crimes that have occured is understated.

Recidivism

There have been several studies that suggest that someone who has committed a crime in the past is more likely to commit crimes in the future[2]. The proportion of people with prior convictions in a county could be an additional driver that would impact crime rates. It is unclear where this may bias, but is unlikely to be in the most affluent areas with the higher wages and taxes per capita. Were that to be the case, the absence of a population with prior convictions in wealthy areas may artificially lower coefficients for tax per capita or the weekly wages vs crime rate.

Unemployment levels

The employment levels in a county is likely to have an impact on crime rates. Unemployment is usually higher in the young and minorities. The higher positive coefficients of percent minority variable in our models will include some amount of bias from the impact of unemployment which will prove a problem to the model.

[2]The Offending, Crime and Justice Survey (2003); RECIDIVISM AMONG FEDERAL OFFENDERS: A COMPREHENSIVE OVERVIEW

Education levels

The level of education in a county could be an indicator of some crimes. There may be covariance between lower levels of education and lower wages, along with unemployment. Rural areas, with less density, and some inner city areas may have populations that are less well educated resulting in a complex impact accross the state which will be hard to predict without data

Strength of community

Strong community ties, generally in rural areas, can have a suppressing effect on crime. This is a counter-weight to education levels and unemployment, both of which may be higher in such communities. This is not necessarily in all areas of low population as there are areas, on the outer banks, where many second homes are located and which can be victim to burglury and theft. The strength of community can also cause crimes to be unreported and dealt with through informal means. This is another complex factor, but would be expected to explain part of the coefficient for density meaning density would be biased and negativly impacting models

Income inequality

Local inequality of wealth can be a driver of crime. Regardless of average wages, and if wages are generally high, if there are disparities in the distribution amongst the population, crimes will tend to increase. This is likely to impact higher density areas, as the population is greater and the the probability of disparities exist. Therefore, this is another factor that would detrimentally impact the density coefficient in models

Conclusion

Through our regression analysis, it is apparent that there are aspects of the criminal justice system that could be leveraged in the short term to decrease crime throughout North Carolina: increasing the probability of arrest for committing a crime and increasing the probability of conviction once arrested. Even single percentage increases in those areas can have a substantial impact on the number of crimes committed. We hypothesize that this is due to a general feeling within a population about whether or not you will get caught for committing a crime, and whether or not you will get convicted after getting caught. In short, the higher the number of arrests and convictions, the greater the fear is of getting caught, and fewer crimes are committed. Since judges are elected in North Carolina, campaigns can be formulated to attempt elect judges that are hard on crime and have higher precedent of conviction. Police training could also be implemented or campaigned for to increase the number of arrests in counties where things have gotten somewhat lax. More research is necessary to determine what is most appropriate in terms of training and implementation.

Besides arrests and convictions, it is apparent that counties with higher density populations and higher percentages of minorities are prone to greater amounts of crime. Additional research is needed to determine why there is more crime in communities with higher minority populations, but this could very well be due to poverty and/or gangs. Further research into crime amongst minority groups would likely unearth much useful information that could be used to inform additional policies to help decrease crime in these areas.

Lastly, due to the high number of omitted variables that are highly relevant to levels of crime, we recommend further research in order to uncover other factors that could be leveraged to decrease crime throughout North Carolina.