Lab 3: Reducing Crime

W203 Statistics

Luke Evans, Daniel Rasband, and Yulia Zamriy

April 15, 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.

```r
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.

```r
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.

After additional investigation, it appeared that there is duplicate record in this dataset:

```r
nrow(crime)
```

```
## [1] 91
```

```r
length(unique(crime$county))
```

```
## [1] 90
```

The county #193 has two identical rows in the dataset:

```r
crime[duplicated(crime$county),"county"]
```

```
## [1] 193
```

We will remove one of the rows:

```r
crime <- crime[!duplicated(crime),]
dim(crime)
```

```
## [1] 90 25
```

As a result our final dataset contains 90 observations and 25 variables.

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

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

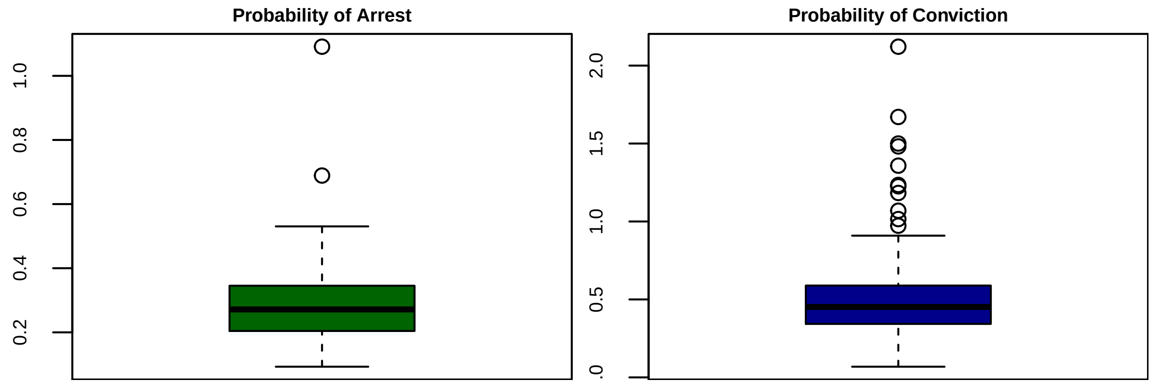
The variables are made up of ratios, specifically the probability of arrest, conviction and prison sentence, the percent minority, young male, police and tax revenue per capita, and the ratio (mix) of face to face crimes to other types of crime. The means of several variables are provided for each county: 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:

```r
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	100.6000	197.0000
year	87.00000	87.0000	87.0000
crmrte	0.00553	0.0335	0.0990
prbarr	0.09277	0.2952	1.0909
prbconv	0.06838	0.5509	2.1212
prbpris	0.15000	0.4106	0.6000
avgsen	5.38000	9.6889	20.7000
polpc	0.00075	0.0017	0.0091
density	0.00002	1.4357	8.8277
taxpc	25.69287	38.1610	119.7615
west	0.00000	0.2444	1.0000
central	0.00000	0.3778	1.0000
urban	0.00000	0.0889	1.0000
pctmin80	1.28365	25.7129	64.3482
wcon	193.64316	285.3532	436.7666
wtuc	187.61726	410.9065	613.2261
wtrd	154.20900	210.9214	354.6761
wfir	170.94017	321.6213	509.4655
wser	133.04306	275.3379	2177.0681
wmfg	157.41000	336.0327	646.8500
wfed	326.10001	442.6189	597.9500
wsta	258.32999	357.7402	499.5900
wloc	239.17000	312.2801	388.0900
mix	0.01961	0.1290	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.

```r
par(mfrow=c(1,2), mar=c(0,2,1,0))
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")
```

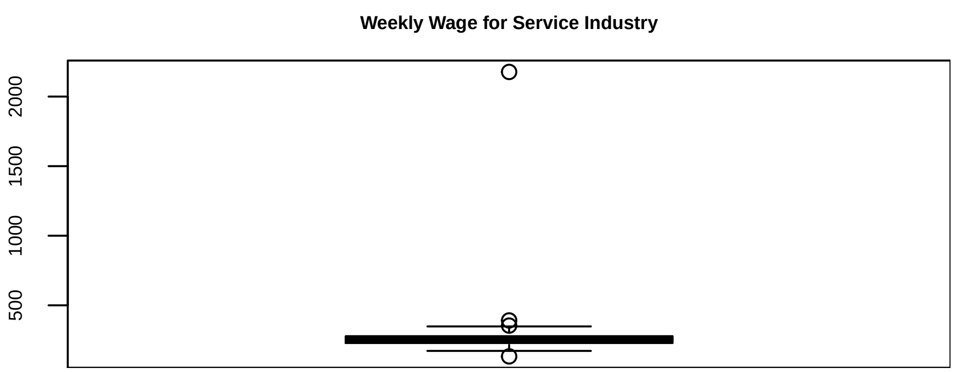


In case of probability of arrests, there is only one observation where the value is above 1, and it is significantly higher than the next closest value. It indicates that there have been more arrests than there have been crimes in a county. As this is time-limited data covering a single year, it is possible 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.

```r
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. This outlier may be an effect of the metric. The data does not provide a number of employees in each sector, and therefore it is impossible to compare these figures effecivly. This service business may be small, niche, highly paid staff such as a investment bank, and in a county with few other service industries, driving this sectors wages up. It is also not clear if this is the county where the wage is earned or the counter where the wage earner lives. The value is clearly not representative of the North Carolina population, however, and will be removed from our models

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

```
## [1] 9.169582
```

Only the service weekly wage value will be replaced by an imputed value; the rest of the observation will need to be maintained

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.

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

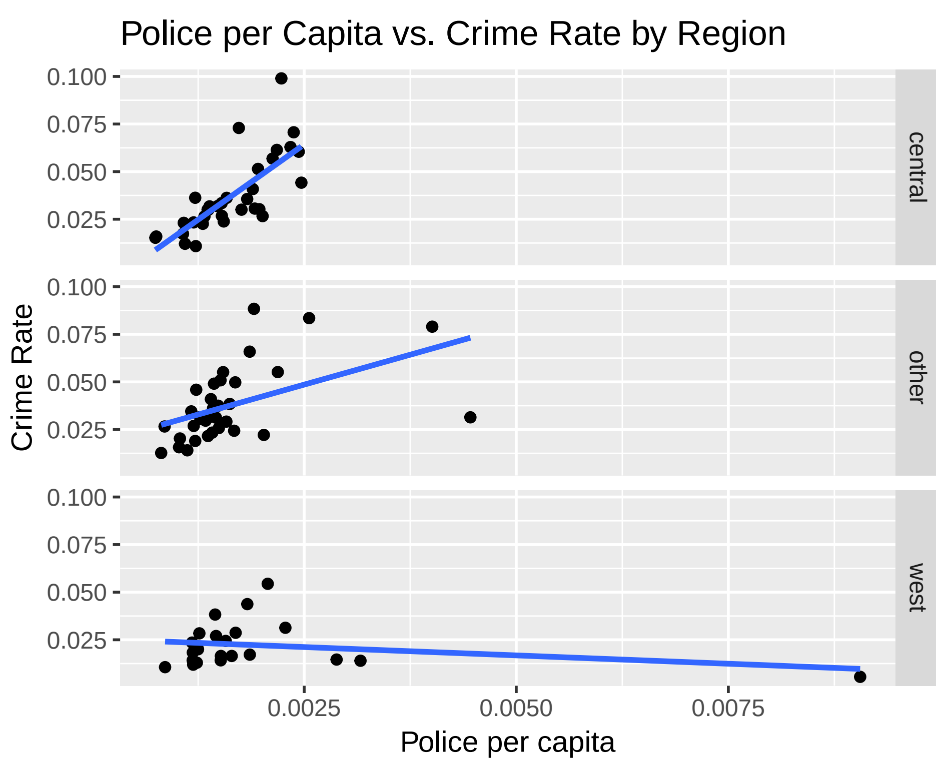
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   275.3   277.6  2177.1
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   254.0   275.9   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.

```r
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. There may be a valid reason for so many police per capita in one specific county, but as it is not representative of the rest of the population it will be removed. 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, which is still a lot fewer than this outlier.

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

```r
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.
```

[1] http://www.governing.com/gov-data/safety-justice/law-enforcement-police-department-employee-totals-for-cities.html

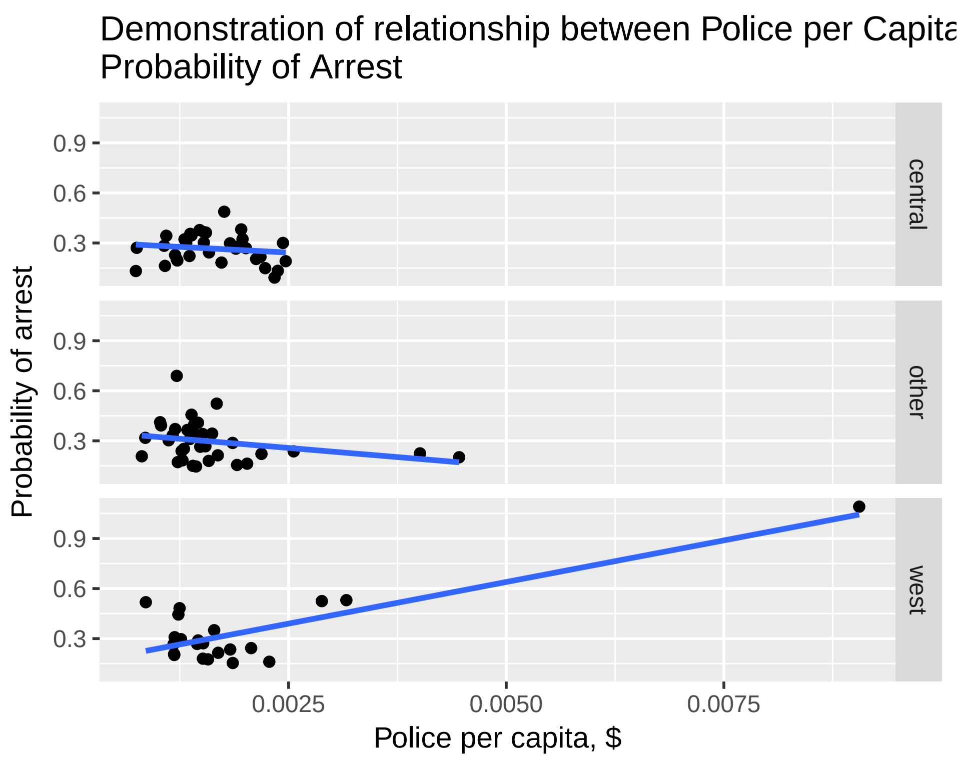
```
## 0.0007459 0.0012378 0.0014897 0.0017080 0.0018856 0.0090543
```

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

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012378 0.0014897 0.0016255 0.0018587 0.0044592
```

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

```r
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"))
```



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 #51 with the outlier is excluded:

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

```
## Correlation with the outlier inlcuded: 0.4259648
```

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

```
## Correlation with the outlier excluded: -0.126103
```

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:

```r
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 0.20495 0.27146 0.29524 0.34487 1.09091
```

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.09277 0.20495 0.27146 0.28646 0.34331 0.68902
```

The observation in question is county 115, observation 51. This county has the probability of arrest of over 1, and the exceptional police per capita. Furthermore, this county has the lowest crime rate, the highest average sentence, the lowest minority percentage and a conviction to arrest rate of over 1. It is an interesting county to be aware of as it does not appear to be representative of the state as a whole, and therefore some of the policy reccomendations may not be as effective there

An additional data issue was identified for a county where both *west* and *central* variables are equal to 1:

```r
table(crime$west, crime$central)
```

```
##
##       0  1
##    0 35 33
##    1 21  1
```

Is it possible that this county (#71) belongs to two regions? Unfortunately, we don't have enough information to answer this question. Moreover, we investigated this county and compared it to the averages of all variables for $west = 1$, $central = 1$ and $west = central = 0$. The available characteristics of this county did not make it obvious where this county belongs. However, we do not want to exclude it entire from our analysis as it seems like a valid data point in all other respects. Hence, we'll create a flag variable that will allow us to exclude it if necessary.

```r
crime$exclude <- ifelse(crime$county %in% c(71), 1, 0)
table(crime$exclude)
```

```
##
##  0  1
## 89  1
```

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

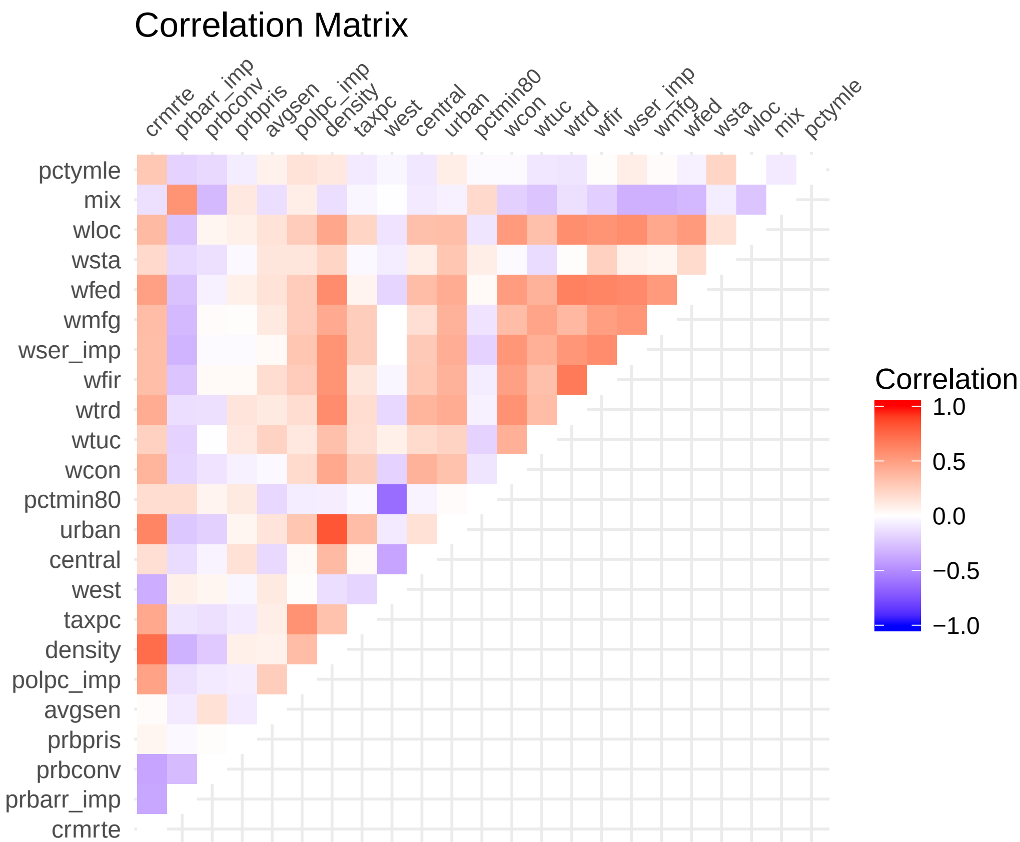
To conclude our initial data exploration, an easy-to-reference correlation heatmap has been developed for quick identification of positive or negative correlations between variables in the data set.

```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 crime rate. It is one of the most important variables to test in the models as there are a number of factors, discussed in omitted variables, that are included in this factor. Its impact must be considered.
- All wages variables are positively correlated with each other. Hence, it would create challenges for keeping multiple variables in the model. Additionally, the understanding of the interpretation of the wages variables poses challenges, as discussed above. More information is required before these could be transformed and utilized in any model
- `density` has a strong positive correlation with urban indicator. This relationship is as expected. Given that `density` has higher correlation with `crmrte`, including density will be satisfactory to control for the level of urbanization in the county in the models.
- The level of policing, or the police per capita, is positivly correclated with crime. This impact is not expected as an increase in policing might logically be associated with a reduction in crime. However, the reaction to a high crime rate would be to increase policing to help control crime and it is unclear from this single year based dataset the increase in policing is reducing the crime rate as compared to prior periods. Furthermore, the presence of police may make it more likely to report crime, increasing the crimes reported above other counties
- The Western region of the state exhibits a lower prevelance of crime per person than the other regions. This is an interaction that may require some further investigation, but will be controled for in the models presented. Some comparison of group means reveals a 25% difference in percent minority between the Western region and others. While this may be related to the presence of minorities, it is more likely to be the product of a bivariate relationship with a 3rd factor such as poverty. Tax per capita is not very different from other parts of the state, but the western regions of the outer banks are attractive for retirement communities which may be driving this impact

```
pctmin_per_region <- lm(pctmin80~west, data = crime)
(summary(pctmin_per_region)$coefficients)
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  31.79991   1.602533 19.843531 3.045441e-34
## west        -24.90159   3.241282 -7.682636 2.051258e-11
```

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. - Short term policy impact are policies that could be implemented quickly, within the months preceeding and immedietly after an election. The support and lobbying for judges with perspectives that would support policies relating to custodial terms and their length,for example. - Medium term policies are those that take some planning and financing, but which can be implemented and results demonstrated during an electoral term. This may include recruiting and training police officers, or upgrades of crime investigation equipment. - Long term policies would be those needing sustained effort or funding, and results will be outside of the duration of a single term. These may be developing incentives and strategies to reduce population density

```r
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	0.02	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	0.18	Long Term
wcon	weekly wage, construction	Negative	0.39	Medium Term
wtuc	wkly wge, trns, util, commun	Negative	0.24	Medium Term
wtrd	wkly wge, whlesle, retail trade	Negative	0.43	Medium Term
wfir	wkly wge, fin, ins, real est	Negative	0.34	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	0.36	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 the 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 terms 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 section will focus on the residuals of all three models.

Dependent variable

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

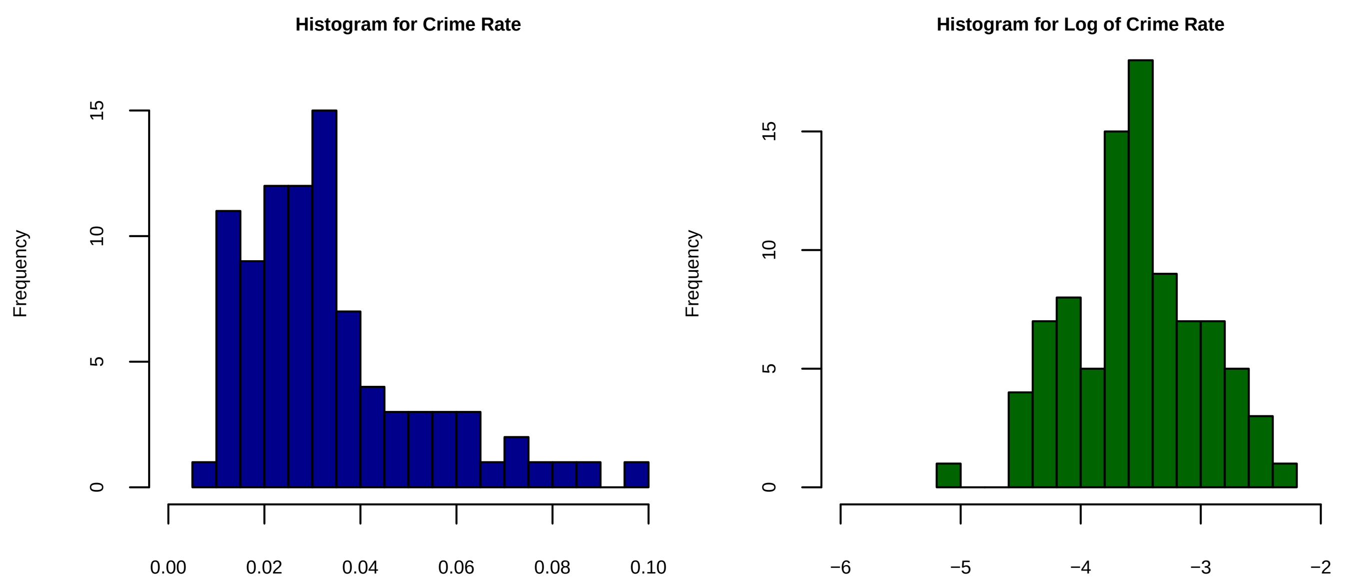
After careful consideration, and in order for us to understand the impact of our main causal effects (probability of arrest and probability of conviction) on crime rate, we decided to transform our dependent variable by taking a natural log.

Since this variable is a ratio (crimes per person), hypothetically it can vary between 0 and 1 (though it's highly unlikely that one could find a county with such a high crime rate). This makes it not very suitable for OLS because this method can predict values outside the 0 to 1 range. The natural log will help us only with part of the problem (avoiding negative values in the prediction of actual crime rate). There's one caveat: in our dataset, the crime rate variable is never equal to zero. Hence, transformation is straightforward. 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 the following: 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 `y` big or small?).

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

```
par(mfrow=c(1,2), mar=c(2,4,1,0))
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 have 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 on 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 would 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):

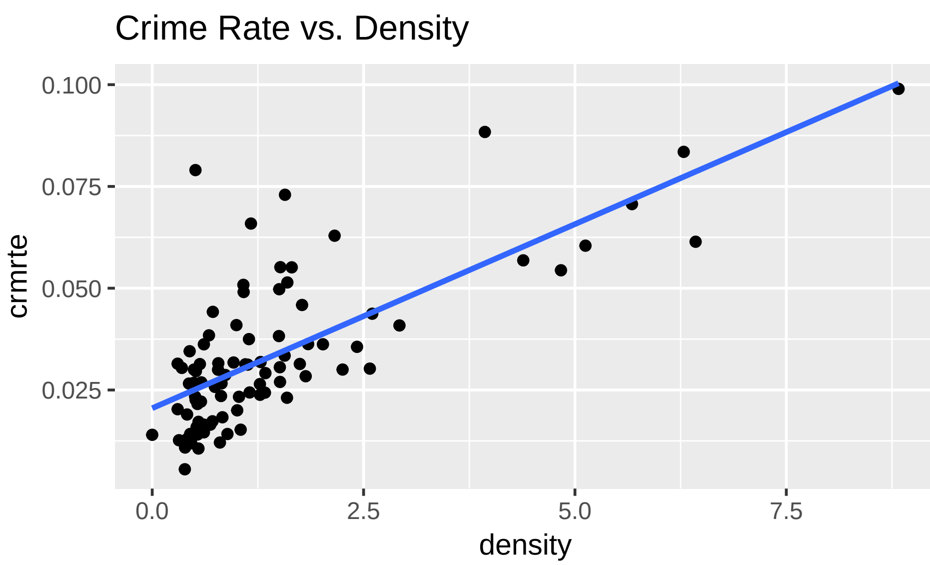
```
crime$region <- ifelse(crime$west == 1, "west",
                       ifelse(crime$central == 1, "central", "other"))
aggregate(crmrte ~ region, data = crime, mean)
```

```
##    region     crmrte
## 1 central 0.03699627
## 2   other 0.03739491
## 3    west 0.02209975
```

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 1.074319
```

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

Another important note about `density` is that it appears to be on a different scale than *people per square mile*. Internet research revealed that current average density in NC is 187.6 *people per square mile*, while in the provided dataset the average value is 1.43 (for 1987). Comparison of county-level density between provided dataset and current numbers provided stronger confirmation that *density* in 1987 *crime* dataset is *population per 100 sq.miles.* This clarification will be important for model coefficient interpretation.

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
```

```
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)
model1a <- lm(model1.formula, data = crime[crime$exclude == 0,])
```

```
interpret1 <- c("","For each person per 100 square miles increase in density,
                crime rate increases by 13.7% when everything else stays the same",
                "Crime rate in the West is 40.1% 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 1.68%",
                "For approximately each percentage increase in probability of conviction,
                crime rate decreases by 0.68%")

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)	-2.7826	
density	0.1374	For each person per 100 square miles increase in density, crime rate increases by 13.7% when everything else stays the same
west	-0.4013	Crime rate in the West is 40.1% lower than in Central and Other regions on average (and controlling for all other included variables)
prbarr_imp100	-0.0168	For approximately each percentage increase in probability of arrest, crime rate decreases by 1.68%
prbconv100	-0.0068	For approximately each percentage increase in probability of conviction, crime rate decreases by 0.68%

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 more impact on crime rate than a 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 a crime (being arrested is easier to relate to than being convicted).

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

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

```
## [1] 0.6722181
```

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

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

```
##
## t test of coefficients:
##
##                  Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)    -2.7825679  0.2020825 -13.7695 < 2.2e-16 ***
## density         0.1374399  0.0270796   5.0754 2.247e-06 ***
## west           -0.4012806  0.0768288  -5.2231 1.234e-06 ***
## prbarr_imp100  -0.0168021  0.0037186  -4.5184 1.998e-05 ***
## prbconv100     -0.0068451  0.0014364  -4.7655 7.697e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One last check is to compare model results for the full dataset and for the one excluding count #71 (where $west = central = 1$):

```r
compare1 <- data.frame(cbind(round(coeftest(model1, vcov = vcovHC, level = 0.05)[,1],4),
                             round(coeftest(model1a, vcov = vcovHC, level = 0.05)[,1],4),
                             round(coeftest(model1, vcov = vcovHC, level = 0.05)[,3],1),
                             round(coeftest(model1a, vcov = vcovHC, level = 0.05)[,3],1)))
colnames(compare1) <- c("Est. w/ #71", "Est. w/o #71", "t.val w/ #71", "t.val w/o #71")
compare1
```

```
##                  Est. w/ #71 Est. w/o #71 t.val w/ #71 t.val w/o #71
## (Intercept)          -2.7826      -2.7776        -13.8         -13.8
## density               0.1374       0.1339          5.1           4.9
## west                 -0.4013      -0.4120         -5.2          -5.1
## prbarr_imp100        -0.0168      -0.0168         -4.5          -4.5
## prbconv100           -0.0068      -0.0068         -4.8          -4.8
```

As we can see, there is a slight change in the coefficient for *west*. However, this change is less than 3%. Hence, we can conclude that this county doesn't change our results and we will keep it in.

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 a 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 crime per capita require more police presence. Therefore, this variable is necessary for control purposes and it improves the fit of the model.

The percent of 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 do counties with more minorities have more gangs? 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:

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

Correlation is strongest between the logs of both variables:

```r
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)	0.5074
log(crmrte), polpc	0.4312
crmrte, polpc	0.4764

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	0.19
Central	0.80
Other	0.59

Now, to the actual model:

```
model2.ind_vars <- c("density", "west", "prbarr_imp100", "prbconv100",
                     "polpc_imp.ln", "pctmin80", "I(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 100 square miles increase in density,
  crime rate increases by 8.9%",
  "(Before: -0.40): This coefficient cannot be interpreted by itself
  as it is now part of interaction with police per capita.
  See explanation below",
  "(Before: -0.0168): 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: -0.0068): 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 0.64% 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)	1.2232	
density	0.0893	(Before: 0.14): The effect of density has decreased as we are controlling for more factors. For each person per 100 square miles increase in density, crime rate increases by 8.9%
west	-4.1247	(Before: -0.40): 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: -0.0168): 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: -0.0068): 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	0.6358	This coefficient indicates that a 1% increase in police per capita is associated with a 0.64% 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
I(west * polpc_imp.ln)	-0.6101	See explanation below

The interaction term (`I(west * polpc_imp.ln)`) is harder to interpret. It applies only to the West region. That means that the `polpc_imp.ln` coefficient of 0.6358 applies only to Central and Other regions. In the West the coefficient for `polpc_imp.ln` is actually 0.0257 (0.6358 - 0.6101) or per each percent change in police per capita in the West, there is only a 0.0257% change in crime rate. This value is relatively close to zero and implies no practically significant relationship between the 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$coefficients["I(west * polpc_imp.ln)"]
coef_west <- model2$coefficients["west"]
coef_west + coef_west_polpc*mean_polpc
```

```
##       west
## -0.1800399
```

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 80.2%, which is 12.9 percentage points higher than our first model:

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

```
## [1] 0.8016033
```

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)             1.2231517  0.8082919  1.5133  0.134060
## density                 0.0893445  0.0248503  3.5953  0.000552 ***
## west                   -4.1247092  1.2955741 -3.1837  0.002056 **
## prbarr_imp100          -0.0201501  0.0032813 -6.1409 2.798e-08 ***
## prbconv100             -0.0074011  0.0011862 -6.2393 1.829e-08 ***
## polpc_imp.ln            0.6357665  0.1209539  5.2563 1.144e-06 ***
## pctmin80                0.0098773  0.0021418  4.6117 1.454e-05 ***
## I(west * polpc_imp.ln) -0.6101401  0.2017213 -3.0247  0.003323 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, the p-value for the F-statistic (derived below) is well below the 1% critical value. This indicates that we can reject the null hypothesis that the additional variables added to model 2 versus model 1 (police per capita, percent minority and the interaction between the Western region and police per capita) jointly have no effect on crime rate.

```
waldtest(model2, model1, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: log(crmrte) ~ density + west + prbarr_imp100 + prbconv100 + polpc_imp.ln +
##     pctmin80 + I(west * polpc_imp.ln)
## Model 2: log(crmrte) ~ density + west + prbarr_imp100 + prbconv100
##   Res.Df Df     F    Pr(>F)
## 1     82
## 2     85 -3 15.24 5.735e-08 ***
## ---
## 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", "I(west * 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] 0.8302507
```

Despite adding a lot more variables, the R^2 of the all-inclusive regression model went up only to 83.0% (from 80.1% in model 2). The Wald test also reveals that if we use 1% as a critical value, we fail to reject the null hypothesis that variables added to model 3 jointly have no effect on crime rate.

```
waldtest(model3, model2, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: log(crmrte) ~ prbarr_imp100 + prbconv100 + prbpris + avgsen +
##     polpc_imp.ln + I(west * 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
## Model 2: log(crmrte) ~ density + west + prbarr_imp100 + prbconv100 + polpc_imp.ln +
##     pctmin80 + I(west * polpc_imp.ln)
##   Res.Df  Df      F  Pr(>F)
## 1     66
## 2     82 -16 2.1983 0.01339 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```r
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***         0.106
##                               (0.027)         (0.025)        (0.062)
## taxpc                                                         0.001
##                                                              (0.006)
## west                         -0.401***       -4.125**        -2.940
##                               (0.077)         (1.296)        (1.828)
## central                                                      -0.120
##                                                              (0.085)
## urban                                                        -0.138
##                                                              (0.292)
## prbarr_imp100                -0.017***       -0.020***      -0.018***
##                               (0.004)         (0.003)        (0.003)
## prbconv100                   -0.007***       -0.007***      -0.007***
##                               (0.001)         (0.001)        (0.001)
## prbpris                                                      -0.300
##                                                              (0.385)
## avgsen                                                       -0.021
##                                                              (0.015)
## polpc_imp.ln                                  0.636***       0.614**
##                                               (0.121)        (0.208)
## pctmin80                                      0.010***       0.009**
##                                               (0.002)        (0.003)
## wcon.ln                                                       0.262
##                                                              (0.225)
## wtuc.ln                                                       0.129
##                                                              (0.282)
## wtrd.ln                                                       0.257
##                                                              (0.300)
## wfir.ln                                                      -0.239
##                                                              (0.311)
## wser_imp.ln                                                  -0.492
##                                                              (0.300)
## wmfg.ln                                                      -0.036
##                                                              (0.161)
## wfed.ln                                                       0.778
##                                                              (0.439)
## wsta.ln                                                      -0.266
##                                                              (0.325)
## wloc.ln                                                       0.040
##                                                              (0.629)
## mix                                                          -0.425
##                                                              (0.513)
## pctymle                                                      2.666**
##                                                              (1.010)
```

```
## I(west * polpc_imp.ln)                        -0.610**         -0.428
##                                                (0.202)          (0.282)
## Constant                     -2.783***         1.223           -1.579
##                               (0.202)          (0.808)          (4.232)
## -----------------------------------------------------------------------
## Observations                     90              90               90
## R2                             0.687           0.817            0.874
## Adjusted R2                    0.672           0.802            0.830
## Residual Std. Error    0.314 (df = 85) 0.244 (df = 82) 0.226 (df = 66)
## =======================================================================
## 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` decreases from model 1 to model 2 as we introduce more variables that likely interact with it (for example, `west`). However, its standard error remains small enough for `density` effect on crime rate to be significant. In model 3, however, its standard error increases most likely due to strong interaction with added variables
- `polpc` maintains its statistically significant coefficient with low standard error in model #3
- `pctmin80` is also robust as its coefficient stays statistically significant in model #3
- All other variables are not statistically significant in model #3, except `pctymle`. However, as we tested this variable in model 2, its standard error was too high. Hence, this variable is not robust enough to keep it in our main model

CLM Assumption Analysis

In this section we discuss the classic linear model (CLM) assumptions of our models. Since model #2 has the highest number of significant coefficients, we will include a full explication of the assumptions for that model, referencing the other models for comparison or when surprising deviations are apparent.

MLR.1: Linearity in Parameters

For all three models, we assume linearity in parameters by default. We have made natural log transformations to the crime rate, police per capita, and wage variables to capture nonlinear relationships between those variables within a linear framework. This does not violate MLR.1.

MLR.2: Random Sampling

Our analysis of the random sampling assumption applies to all three models. Since the observations in the data include 90% (90/100) of all North Carolinian counties, it would be difficult to argue that the observations are not representative of the population. That said, there is always a possibility that the data set could include all counties except those with high density, high crime, low crime, or some other exceptional joint quality. Based on the variances of observations in the various variables, however, it appears that a wide variety of counties are represented, and are in fact representative of the population. For example, the natural log of `crmrate` is approximately normally distributed, and the untransformed `crmrte` variable ranges from close to zero (about 6 crimes per 1000 people) up to almost one hundred per one thousand people. Many other variables have a more or less reasonable variation and distrubution.

```r
cat(paste(
  paste("Minimum crime rate:", round(min(crime$crmrte) * 1000), "per 1000 people", sep = " "),
  paste("Maximum crime rate:", round(max(crime$crmrte) * 1000), "per 1000 people", sep = " "),
  sep = "\n"
))
```

```
## Minimum crime rate: 6 per 1000 people
## Maximum crime rate: 99 per 1000 people
```

One thing that seems to indicate a slight bias in the sampling is the region. It is apparent that there are fewer counties represented in the west than in other regions:

```r
cat(paste(
  paste("West:", nrow(subset(crime, west == 1))),
  paste("Central:", nrow(subset(crime, central == 1))),
```

```r
  paste("Other:", nrow(subset(crime, central == 0 & west == 0))),
  sep = "\n"
))
```

```
## West: 22
## Central: 34
## Other: 35
```

This is actually a fair representation of North Carolina, however, since the western region of the state is narrower than the central and eastern regions, meaning there are actually fewer counties in the western region.

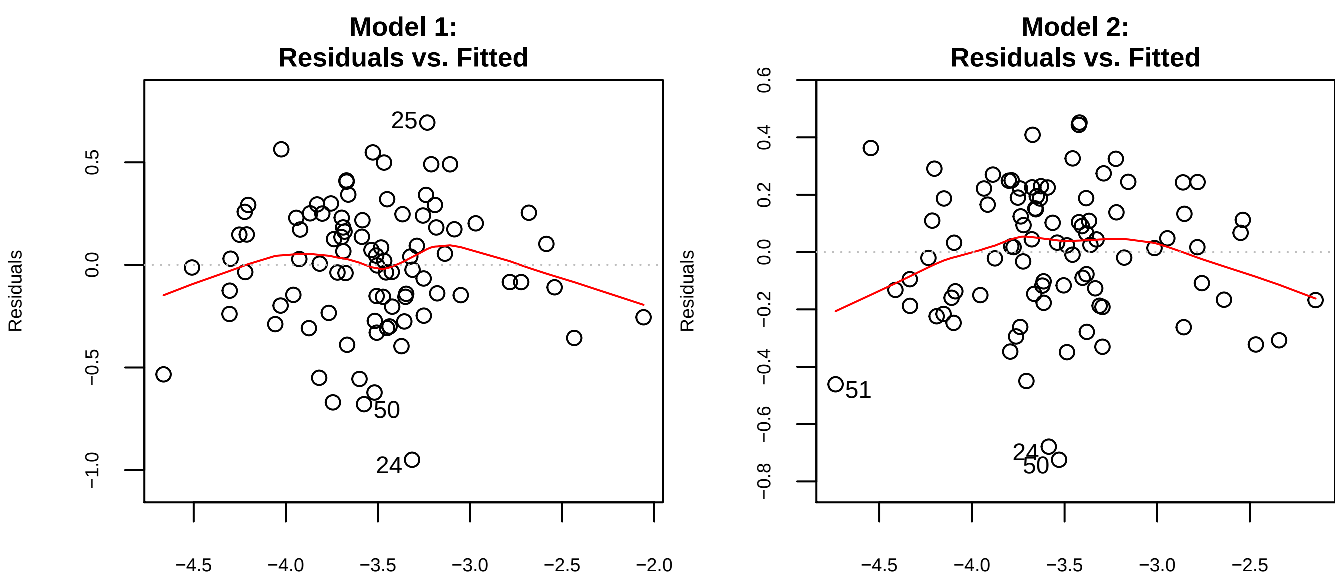
MLR.3: No Perfect Collinearity

If all three regions ("West", "Central", and "Other") were used in the model, then there would be perfect multicollinearity, but R would remove one of them anyone and assumption would no longer be validated. Our models do not include variables that would violate this assumption.

MLR.4: Zero Conditional Mean

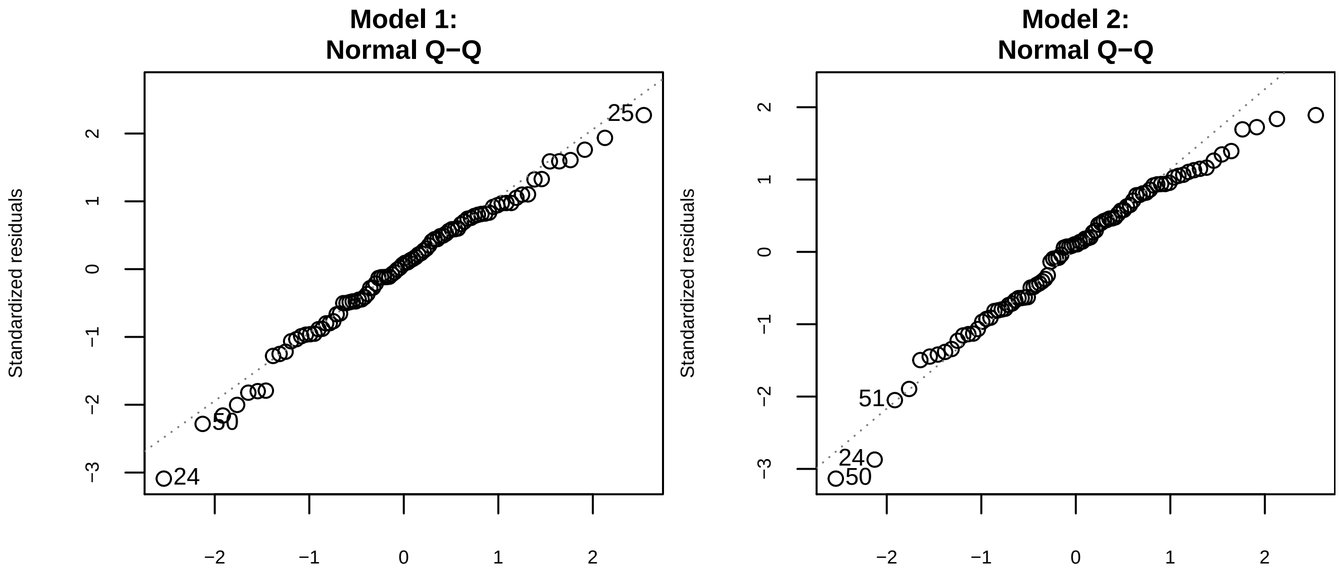
The assumption of zero condition mean of errors seems to be violated in models #1 and #2, as indicated by a seagull-silhouette-shaped residuals vs. fitted lines show here:

```r
par(mfrow=c(1,2), mar=c(2,4,2,0))
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 the Normal Q-Q plot below. The most significant deviation from normality, however, is observed for counties with high crime rates in model #2. This pattern is not as evident in model #1. Hence, more investigation is needed to determine which variables in model #2 contribute to this deviation.

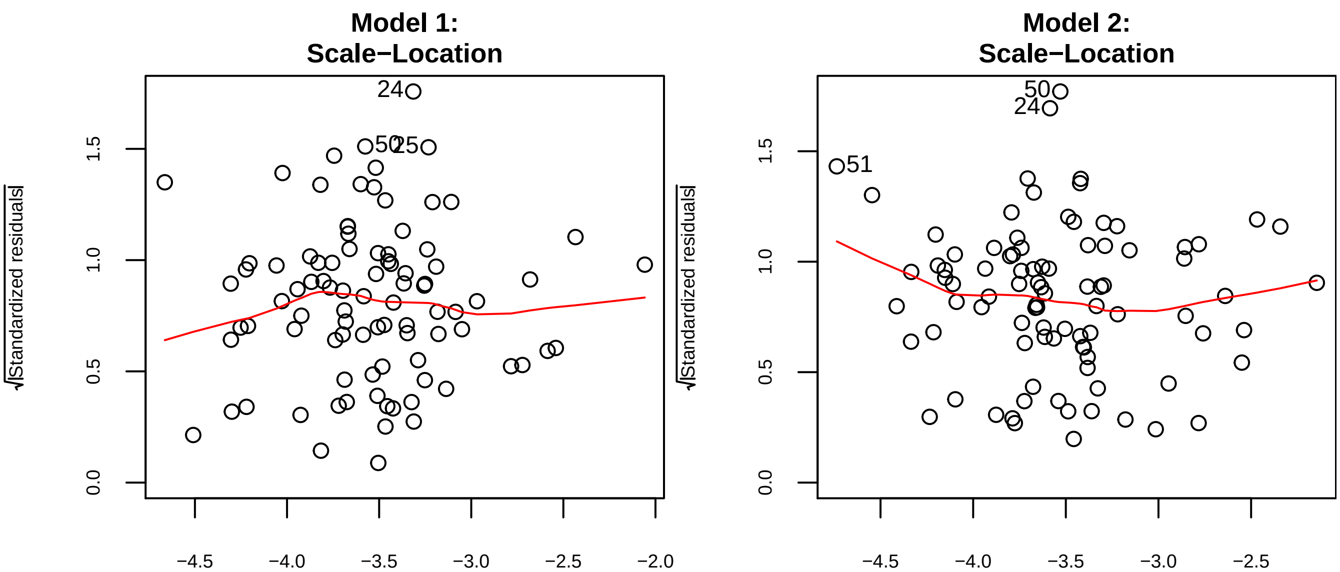
```r
par(mfrow=c(1,2), mar=c(2,4,2,0))
plot(model1, which = 2, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 1:\nNormal Q-Q", caption = "")
plot(model2, which = 2, cex.main = 0.8, cex.axis = 0.6, cex.lab = 0.6,
     main = "Model 2:\nNormal Q-Q", caption = "")
```



MLR.5: Homoskedasticity

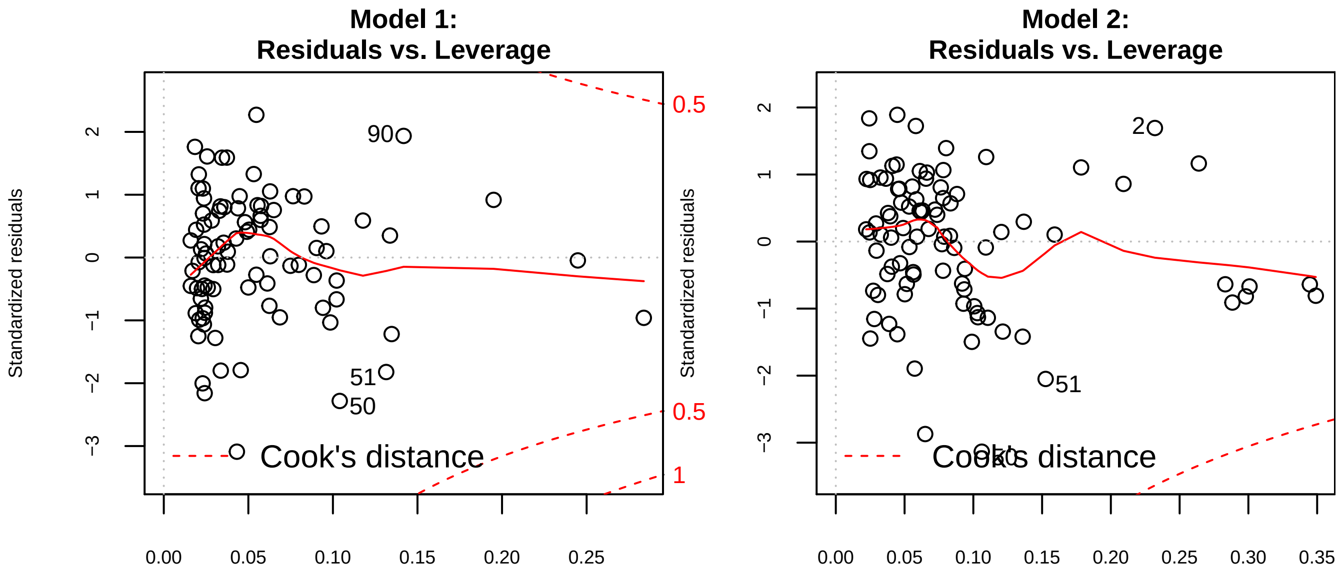
In the Scale-Location plot, again observation #51 is causing some heteroskedasticity in residuals for model #2. In model #1, the standard deviation of residuals, on the other hand, looks constant.

```
par(mfrow=c(1,2), mar=c(2,4,2,0))
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 stand out as outliers, even #51 in the Cook's distance charts below.

```
par(mfrow=c(1,2), mar=c(2,4,2,0))
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 = "")
```

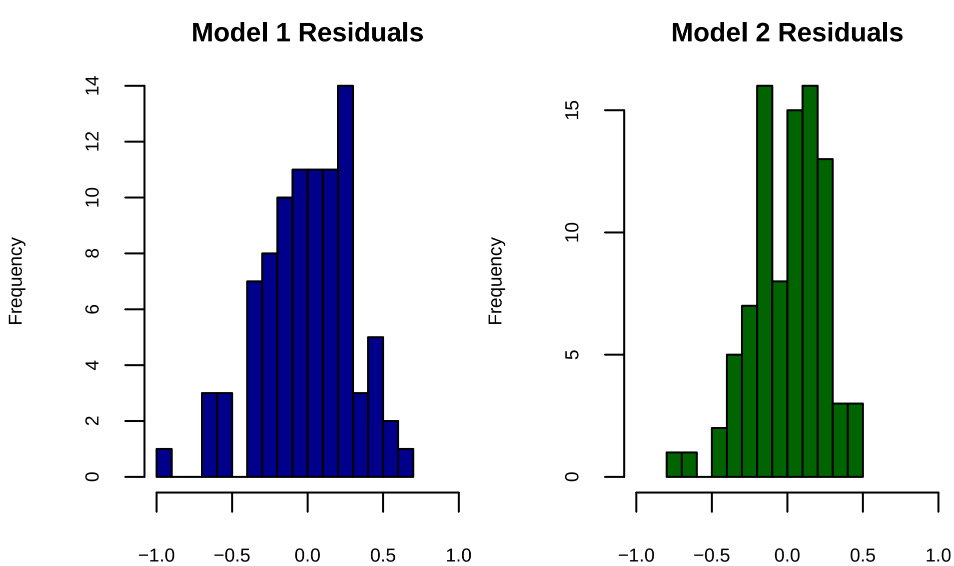


MLR.6: Normality of Errors

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 iterations.

Our dataset contains more than 30 observations (90 to be precise), so we can apply CLT to its residuals and assume they are normal. We still investigate the histograms for model #1 and #2 below. Neither of them fit into the normal distribution particularly well, but they don't have particularly strong skews either.

```
par(mfrow=c(1,2), mar=c(2,4,2,0))
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 shows that our models have some issues with fitting certain values of crime rate (on the high end). We can also see some heteroskedasticity in residuals caused by a few observations.

Omitted Variable Analysis

```r
omv_items <- c(
  "Drug and alcohol abuse levels","", "Unreported crime", "", "Recidivism", "",
  "Unemployment levels", "", "Education levels", "", "Strength of community", "",
  "Income inequality"
)

omv_desc <- c(
  "The presence of drug and alcohol problems in a community is a significant
  contributing factor to crime rates in many areas.", "",
  "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 social isolation or reprisals in smaller communities.
  The presence of gangs, undocumented immigrants, those poor and uninsured, or
  local judicial services being overwhelmed and unavailable may be a cause in
  some more urban areas", "",
  "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 rate", "",
  "The employment levels in a county are likely to have an impact on crime rate", "",
  "The level of education in a county could be an indicator of some crimes", "",
  "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
  burglary and theft. The strength of community can also cause crimes to be unreported and
  dealt with through informal means", "",
  "Inequality of wealth can be a driver of crime. Regardless of average wage value, if the
  difference in wages is generally high, i.e. if there are disparities in the distribution
  amongst the population, crimes will tend to increase"
)

omv_inf <- c(
  "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 more densely populated
  counties and bias the coefficient of 'density' in our model.  However, we have not
  recommended policy decisions base on density, therefore no impact on conclusions", "",
  "May impact rural areas and pockets of urban areas with impacts that would lower those
  coefficients.  Additionally, the greater the police presence, the more likely crimes are
  to be reported which may be artificially increasing crimes recorded in heavily policed
  areas.  While our model does include the policing per capita, no policy decisions
  are based upon it","",
  "It is unclear where this bias may have the strongest effect, but is unlikely to be in
  the most affluent areas with the higher wages and taxes per capita.  If so, this could
  be artificially raising the coefficient of such variable. As these variables are not
  included in the model, this will not impact policy recommendations but may be a factor
  to consider in further study","",
  "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 should be considered where policies include
  this factor", "",
  "There may be covariance between lower levels of education and lower wages, along
  with unemployment. It is likely to have in impact on tax per capita or the lower wages
  number, neither of have been included in the model","",
  "This is another complex factor, but would be expected to explain part of the coefficient
  for population density meaning density, negatively impact models, making the density
```

```
  coefficient less positive","",
  "This is likely to impact higher density areas, as the population is greater and the
  probability of disparities existing is higher. Therefore, this is another factor that
  would detrimentally impact the density coefficient in models"

)

omv_proxy <- c(
  "None available in this dataset, but arrests for drug-specific crimes are likely to
  be available, and deaths caused by drugs are captured by the CDC", "",
  "While police presence may be an indicator of likelihood that offenses are reported,
  it would not be a strong proxy. Assessing numbers of groups less likely to report crime
  may help (undocumented migrants, uninsured property, counties with high case backlog)","",
  "Use of one of the economic variables may act as proxy, but there will be a lot of other
  variable impact in there also.","",
  "A combination of young male and minority may be used as a proxy, however this may
  miss pockets of unemployment in other demographics. Additionally, young males may
  be university students in certain counties","",
  "Use of tax per capita is possible, but may also reflect the underlying wage arbitrage in
  a county.  Its also not clear what tax this relates to and may be unreliable as
  an income predictor","",
  "A good proxy is not clearly available in this dataset, and may be hard to identify in
  general.  Data on church attendance may be useful, as could membership or attendants of
  other civic societies", "",
  "Having made efforts to use the wages in different sectors to understand disparities, this
  was not deemed practical, given the limitations in the understanding of this data as
  outlined in the EDA.  Some more detailed tax information within the county population may
  useful to generate some understanding into this factor"
)
omv_table <- data.frame("Omitted Variable" = omv_items,
                    "Description" = omv_desc,
                    "Inference" = omv_inf,
                    "Proxy Availability" = omv_proxy)

kable(omv_table, booktabs = TRUE) %>%
  kable_styling(font_size = 6,
              full_width = TRUE) %>%
  row_spec(0, bold = TRUE) %>%
  column_spec(1, width = "8em") %>%
  column_spec(2, width = "20em") %>%
  column_spec(3, width = "20em") %>%
  column_spec(4, width = "20em")
```

Omitted.Variable	Description	Inference	Proxy.Availability
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 more densely populated counties and bias the coefficient of 'density' in our model. However, we have not recommended policy decisions base on density, therefore no impact on conclusions	None available in this dataset, but arrests for drug-specific crimes are likely to be available, and deaths caused by drugs are captured by the CDC
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 social isolation or reprisals in smaller communities. The presence of gangs, undocumented immigrants, those poor and uninsured, or local judicial services being overwhelmed and unavailable may be a cause in some more urban areas	May impact rural areas and pockets of urban areas with impacts that would lower those coefficients. Additionally, the greater the police presence, the more likely crimes are to be reported which may be artificially increasing crimes recorded in heavily policed areas. While our model does include the policing per capita, no policy decisions are based upon it	While police presence may be an indicator of likelihood that offenses are reported, it would not be a strong proxy. Assessing numbers of groups less likely to report crime may help (undocumented migrants, uninsured property, counties with high case backlog)
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 rate	It is unclear where this bias may have the strongest effect, but is unlikely to be in the most affluent areas with the higher wages and taxes per capita. If so, this could be artificially raising the coefficient of such variable. As these variables are not included in the model, this will not impact policy recommendations but may be a factor to consider in further study	Use of one of the economic variables may act as proxy, but there will be a lot of other variable impact in there also.
Unemployment levels	The employment levels in a county are likely to have an impact on crime rate	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 should be considered where policies include this factor	A combination of young male and minority may be used as a proxy, however this may miss pockets of unemployment in other demographics. Additionally, young males may be university students in certain counties
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. It is likely to have in impact on tax per capita or the lower wages number, neither of have been included in the model	Use of tax per capita is possible, but may also reflect the underlying wage arbitrage in a county. Its also not clear what tax this relates to and may be unreliable as an income predictor
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 burglary 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 population density meaning density, negatively impact models, making the density coefficient less positive	A good proxy is not clearly available in this dataset, and may be hard to identify in general. Data on church attendance may be useful, as could membership or attendants of other civic societies
Income inequality	Inequality of wealth can be a driver of crime. Regardless of average wage value, if the difference in wages is generally high, i.e. 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 probability of disparities existing is higher. Therefore, this is another factor that would detrimentally impact the density coefficient in models	Having made efforts to use the wages in different sectors to understand disparities, this was not deemed practical, given the limitations in the understanding of this data as outlined in the EDA. Some more detailed tax information within the county population may useful to generate some understanding into this factor

Conclusion

The models that have been generated suggest that there are some opportunities for policies in the criminal justice system that could make a substantial impact to the levels of crime in North Carolina. There is a strong relationship between crime rates and the probability of arrest for each crime, and the probability of conviction given arrest. Care should be taken when converting these findings to practical strategies and policies, however.

For example, it is clear that the number of arrests per crime impacts crime rates. This may be due to the arrest being close in time to the act of the crime, it has a strong deterrent. However, while developing policies, arrests must continue to be targeted at those who the police have belief committed the crimes. Creating metrics to demonstrate increased arrests may simply incentivize any arrest, increasing the number of wrongful arrests and leading to civil liberty groups filing complaints.

The probability of conviction is a strong measure of police efficiency. If the police are arresting the correct suspects, then they will be found guilty and convicted. A higher level of certainty of the Justice system working to a would-be perpetrator clearly reduces the propensity for carrying out a crime. Again, care must be taken in implementing policies to improve conviction rates so as not to impede the criminal justice system in any way. Pressurizing juries to convict or making the system more difficult for defendants, would increase the conviction rate, but may lead to more wrongful convictions, reduce the confidence in the judiciary and may not have the intended impact on crime rates

Policies that will help police quickly identify the correct suspects for arrest should help both variables. Equipment for improved forensic analysis and training for this can be provided. Approaches for crime analysis, interviewing witnesses and community surveillance may also be employed. Additionally, reviewing the practices and policies of those counties that do have lower crime rates due to the better rates of arrest and conviction may uncover some best practices to incorporate into policies

Historically, increasing police numbers are a quick policy decision to reduce crime. While the model presented here might suggest it may actually increase crime, more work should be done to understand this. As mentioned before, the data is panel in nature, and a time series would be more useful in demonstrating the impact of increased policing

It is also pointed out that areas where there was a higher minority presence in 1980 also appear to exhibit higher rates of crime. There is likely to be bivariate relationships that are impacting this, including poverty and unemployment levels which artificially increase the impact of this coefficient. There could also be offsets as there may be a lower propensity to report crime in minority communities.

One of the more striking findings is that the probability of a custodial sentence does not appear to be a strong influencer of crime rate. This may be a result of the limited breakdown of crimes committed. It is unlikely that a custodial sentence would be given for traffic offenses, vandalism and many petty crimes which form the majority of criminality, and therefore prison will not form a disincentive to commit them. The crimes that are likely to result in a prison sentence would have to be more closely analyzed to understand if the probability of committing that crime and going to prison impact the number of such crimes committed. There are policy opportunities in this space. Since judges are elected in North Carolina, campaigns can be formulated to attempt to elect judges that are hard on crime and have higher precedent of custodial sentencing. There is a crime mix indicator provided in the dataset, however it does not clearly distinguish between those offenses punishable by prison and those that are not, and the results of using it provided limited insights. Some more detailed data on crime will be needed to demonstrate how to incorporate this into policies.

It is also worth mentioning that, due to the difference in the crime levels in the West of the state, there may be more specific polities that would apply there. Retirees, second home owners and bay/ocean communities may attract a different type of crime than the mix in the rest of the state.

The number of omitted variables that impact the population density coefficient suggest that there is room for policy development in this area also, but in order to target specific contributors of crime some of those variables must be captured. Providing data on drug abuse, unemployment, poverty and recidivism will help to understand more about the bivariate relationships impacting the coefficient for population density.

Just a note: The Sixth Amendment to the United States Constitution and Article I, Section 24, of the North Carolina Constitution provide a person charged with a crime the right to a trial by a jury. All criminal trials in the superior court are by a jury; the North Carolina Constitution does not allow a judge to decide a criminal defendant's guilt or innocence at trial in superior court