Lab 3 Final Report for W203 Section 5

Nathan Peper, Ernesto Oropeza, Stacy Irwin

16 April 2020

1. Introduction

1.1 Background and Motivation

Crime rates have always been a foundationally important metric to people, but continue to become increasingly complex and debated in today's society. Not only are the debates difficult to compare due to the varying nature of crime rates across the country, but the debates and crime rates also have several differences and influencing factors even at the state and local levels. However, with the evolution of modern media broadcasting and social media platforms combined with the universal desire for safety, information that promotes fear is constantly amplified. This creates a disproportionate perception of crime for people compared to what the actual crime rates suggest. This report attempts to provide some clarity and insight to the crime rates across North Carolina through the analysis of a subset of data from 1987, in order to allow for the development of potential policy recommendations that might positively impact society and the criminal justice system.

One of the highest responsibilities of any politician is to develop laws or policies that keep the region stable while advancing it forward towards its goals. While the boundaries of these regions are generally fixed, the people within the boundaries are affected by these laws and policies and their resulting effects on the society in the region. Within a small region, the people are very different and have their own personal wants and desires. However, if we take a psychologist's view of the population using a well-known theory, such as Maslow's hierarchy of needs, we can move the discussion from a bottom's up approach of worrying about each individual in a region to top's down approach of how to support the universal needs of society and enable intrinsic motivation to help advance society towards its goals.

At the bottom of the Maslow's hierarchy are people's physiological needs, which tend to include food, water, sleep, health, and shelter. Once these most basic needs are met, people tend to be concerned with various areas of safety and let these needs guide personal behavior. Most people typically have different ideas about what makes them feel generally safe, but a common need usually includes a safe environment, or low crime.

On the surface, crime seems to be a very simple and easy to understand concept for the following analysis. However, when we begin to analyze a specific crime and consider the full timeline, process, and motivators that led to that particular crime, it quickly shows its vast complexities. For example, how were these crimes defined? Were they morally charged? Did they change over time or by region? What was the process to establish or change the definition of the crimes and who needed to approve? Once we understand that simply defining crime is complex, how did we measure "crime," or how should we have? This is especially important for any data-driven analysis or decision making. For the crime data in discussion from North Carolina, the data is drawn from the Federal Bureau of Investigation's (FBI) Uniform Crime Reports and the prison and probation files of the NC Department of Correction. How could the recording party's data be influenced by internal or external forces?

Even more complex is now determining the specific cause of a crime. While there are entire industries of research in areas such as criminology due to how complex of a topic this is, many people have a strong view that some simple factor is responsible for a crime. However, these are typically very primitive models that blame crime on broken households, explicit music lyrics, video games, and demographic stereotypes. Most

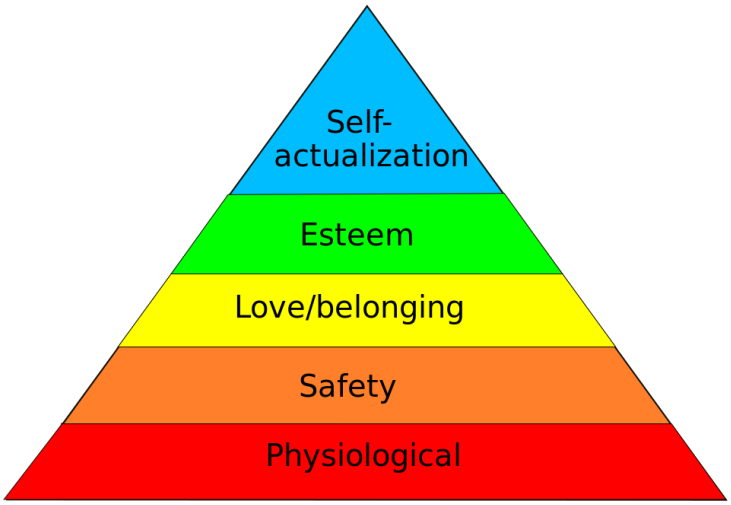


Figure 1: Hierarchy of Needs Image

people tend to avoid taking a stance or casting blame on something that they can't relate to or simply don't understand. While this analysis is not about the specific details of any individual crime, we want to highlight the complexities of determining causation for a "crime" and only hope to develop high-level indicators from the given data set that correlate with the crime rate.

Lastly, as we progress through the analysis of the data set, it is important to consider that our goal for the analysis is providing insight into a given area's crime and generating policy suggestions. However, these policy suggestions will subsequently create their own effects on the population that we must consider. A powerful and recent example of this can be seen in New York City's Operation Impact, where the NYPD increased police presence in high-crime areas in an effort to reduce crime. However, the measurable unintended effects of the program were a significant reduction in test scores for African American boys, age 9 to 15. Examples like this highlight the importance of considering how the analysis is performed, what inferences are drawn from the analysis, and what recommendations result.

1.2 Research Focus

For a political campaign that is looking to keep its region stable and advance society within its given boundaries, we recommend using the provided data set to answer the following research question:

What aspects of the criminal justice system are most effective at reducing crime rates?

While it is likely that both the criminal justice system and economic factors have an impact on local crime rates, we believe that the criminal justice system is an area that policy makers can directly affect change while the economy has more external influences that are out of our control. Additionally, we'll use the provided data to show that the correlation and model fit for the criminal justice indicators are stronger than the economic indicators to support our assumption.

Related to the criminal justice system, some of the sub-questions that we'll focus on throughout the analysis include:

- Are the number of police important with respect to reducing crime rates?
- Which steps in the criminal justice system process are most important to reducing the crime rate?

In order to answer these questions, we'll build three distinct models. The first model will focus on the criminal justice system variables available in the dataset as a baseline for our continued analysis. This model is important because it will enable us to understand the potential influence these variables may have on

crime rates as they are the variables we also have the greatest ability to affect in the near term from a political perspective. The second model will continue by incorporating appropriate demographic data from the exploratory data analysis. While we are not able to influence these variables, they are important for understanding the environment that we are creating policy recommendations and incorporating into our models. The third and final model, will build on our proposed model by methodically including all of the applicable independent variables from the dataset. This will allow us to initially see the significance of the economic variables and then assess the robustness of the proposed model once all variables are included with our model.

Due to the complexity of modern-day society, policy makers must use every tool possible to ensure that they are making educated and fact-based decisions. This data-driven approach to assessing crime rates looks to support recommendations for maintaining stability and advancing the local region.

1.3 Crime Data

We analyzed data from a 1994 study by C. Cornwall and W. Trumball. The data set contains crime and economic data from ninety counties in North Carolina (NC).

1.4 R Packages

We used several R packages for data analysis and visualization:

```
# Packages for data manipulation
library(purrr)

# Packages for linear regression
library(lmtest)
library(sandwich)

# Packages for visualization
library(car)
library(stargazer)
```

1.5 Initial Data Checks

The data set consists of a single, comma separated value (CSV) file. There are three problems with the data that are readily apparent when viewing the final dozen rows (see table 1):

1. There is a stray apostrophe in the *prbconv* column that is causing R to convert this variable to a factor.
2. There is a duplicate row for county 193.
3. The final six rows of the CSV file contain no data.

```
crime_raw <- read.csv("crime_v2.csv")
tail(crime_raw[, 1:8], 12)
```

```
##    county year    crmrte   prbarr     prbconv  prbpris avgsen      polpc
## 86    189   87 0.0313130 0.161381 0.300577998 0.288462  12.27 0.00227837
## 87    191   87 0.0458895 0.172257 0.449999988 0.421053   9.59 0.00122733
## 88    193   87 0.0235277 0.266055 0.588859022 0.423423   5.86 0.00117887
## 89    193   87 0.0235277 0.266055 0.588859022 0.423423   5.86 0.00117887
## 90    195   87 0.0313973 0.201397 1.670519948 0.470588  13.02 0.00445923
## 91    197   87 0.0141928 0.207595 1.182929993 0.360825  12.23 0.00118573
## 92     NA   NA        NA       NA                   NA     NA         NA
## 93     NA   NA        NA       NA                   NA     NA         NA
## 94     NA   NA        NA       NA                   NA     NA         NA
## 95     NA   NA        NA       NA                   NA     NA         NA
```

```
## 96     NA   NA         NA       NA                    NA     NA          NA
## 97     NA   NA         NA       NA             `      NA     NA          NA
```

We can omit the blank rows and errant apostrophe by reading only the first 91 rows from the CSV file, and we can eliminate the duplicate row with R's `unique()` function.

```
crime <- unique(read.csv("crime_v2.csv", nrow=91))
tail(crime[, 1:8])
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 85    187   87 0.0345231 0.332669 0.443114 0.432432   6.98 0.00116911
## 86    189   87 0.0313130 0.161381 0.300578 0.288462  12.27 0.00227837
## 87    191   87 0.0458895 0.172257 0.450000 0.421053   9.59 0.00122733
## 88    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
## 90    195   87 0.0313973 0.201397 1.670520 0.470588  13.02 0.00445923
## 91    197   87 0.0141928 0.207595 1.182930 0.360825  12.23 0.00118573
```

We verified that the remaining data was free of duplicate or empty rows. A result of *FALSE* indicates there is no duplication or missing values.

```
any(duplicated(crime))
```

```
## [1] FALSE
```

```
anyNA(crime)
```

```
## [1] FALSE
```

```
paste("Rows in original data set:", nrow(crime_raw), "Rows in corrected data set:", nrow(crime))
```

```
## [1] "Rows in original data set: 97 Rows in corrected data set: 90"
```

2. Justice System Model

2.1 Justice System Variables

The purpose of our research is to identify the aspects of the criminal justice system that have the greatest impact on crime rates. Our data set includes the following variables related to the criminal justice system:

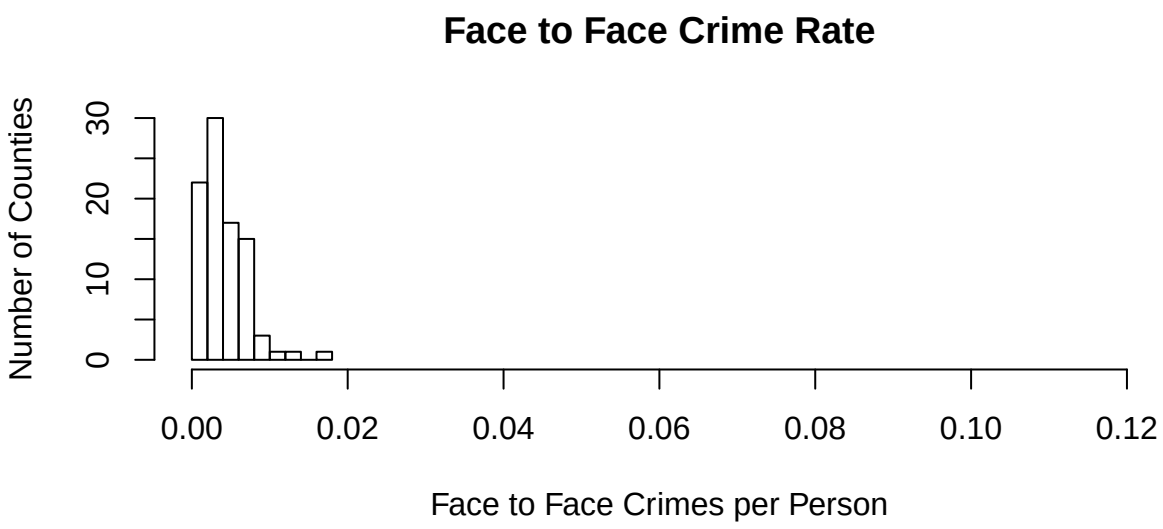
- Percentage of crimes that are face-to-face: *mix*
- Probability of arrest: *prbarr*
- Probability of conviction: *prbconv*
- Number of policemen per capita: *polpc*
- Probability of going to prison: *prbpris*
- Average length of sentences: *avgsen*

2.1.1 Mix Variable

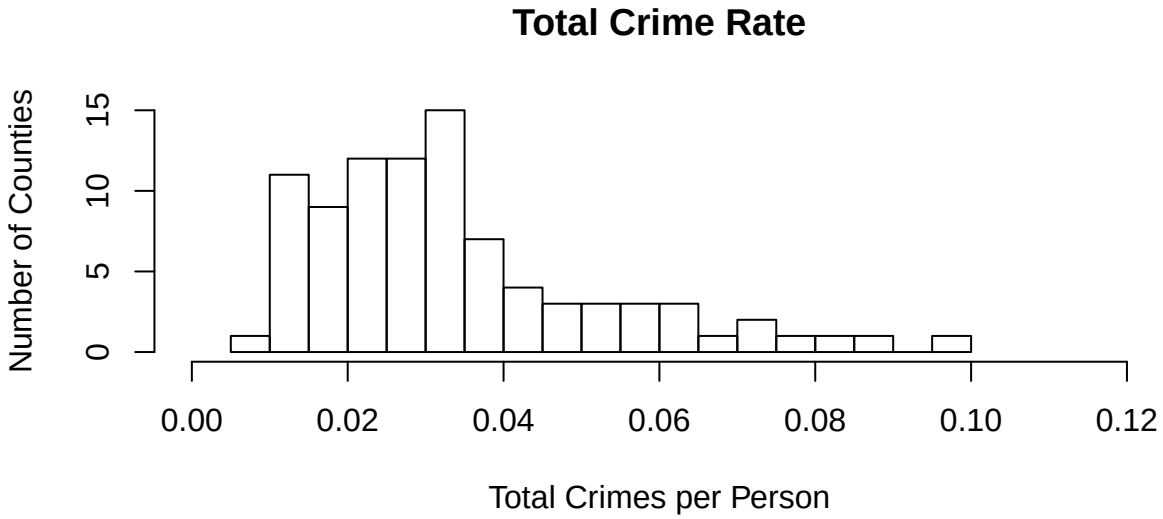
The variable *mix* can be multiplied by the crime rate to calculate separate crime rates for face-to-face and other crimes. The *mix* variable is a modification to the dependent variable, *crmrte*, therefore we will not use *mix* as an independent variable.

The following two plots show the distribution of crime rates for face-to-face and other crimes.

```
hist(crime$mix * crime$crmrte, xlim = c(0, 0.12), main = "Face to Face Crime Rate",
     xlab = "Face to Face Crimes per Person", ylab = "Number of Counties")
```




```
hist(crime$crmrte, xlim = c(0, 0.12), main = "Total Crime Rate",
     xlab = "Total Crimes per Person", ylab = "Number of Counties",
    breaks = 25)
```



Face-to-face crime rates are approximately 20% of the total crime rate. There is little correlation between the percentage of crimes that are face-to-face and the total crime rate (see below).

```
cor(crime$mix, crime$crmrte)
```

```
## [1] -0.1320004
```

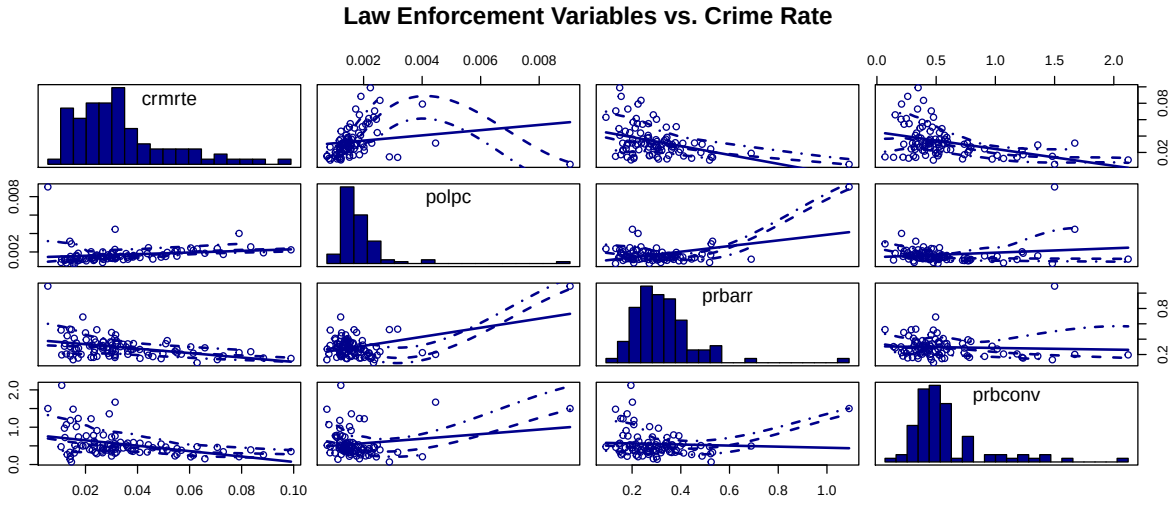
We expect that voters in North Carolina would be most anxious about face-to-face crimes. However, focusing solely on face-to-face crimes would exclude approximately 80% of crimes from our analysis, potentially increasing the impact of random effects on our model's results. Furthermore, we have no means to identify which arrests, convictions, or prison sentences are for face-to-face or other crimes. Our model will focus on the total crime rate. This decision is further justified by the low correlation between total crime rate and the

mix of crimes, which indicates there is not a strong tendency for counties with high crime rates to have a different percentage of face-to-face crimes.

2.1.2 Exploration of Justice System Variables

The matrix of scatter plots below compares law enforcement variables to crime rate. We grouped these variables together because they are related to the likelihood that an offender will be apprehended, stand trial, and be convicted. We include the police per capita, *polpc* variable, with this group because police play a role in arresting offenders and providing evidence for trials, but less of a role in sentencing.

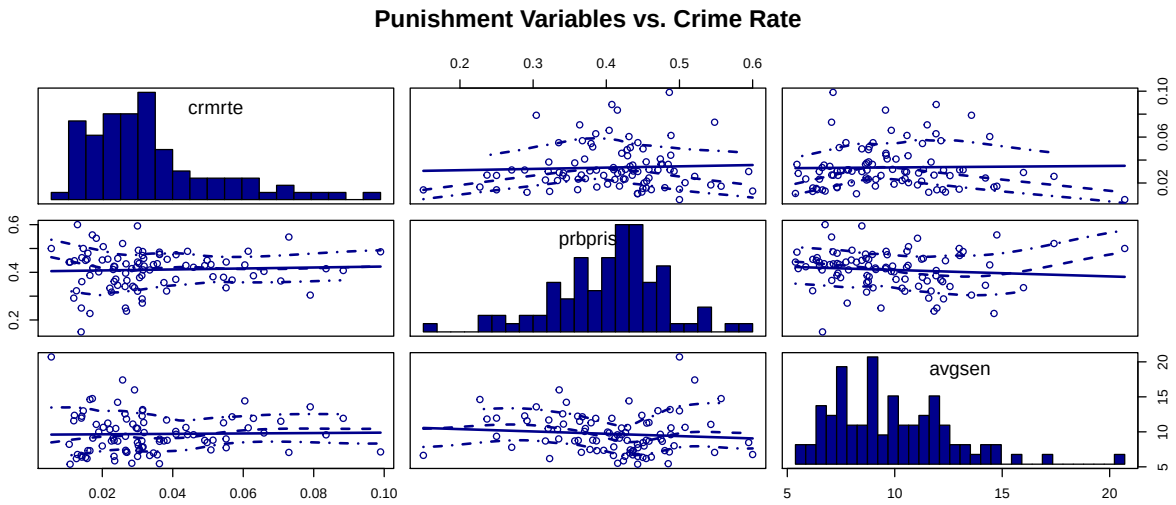
```
scatterplotMatrix(~ crmrte + polpc + prbarr + prbconv, data = crime, col = "dark blue",
                  diagonal=list(method ="histogram", breaks=25),
                  main = "Law Enforcement Variables vs. Crime Rate", cex.labels = 1.4)
```



We can see that the crime rate is higher when the *prbarr* and *prbconv* is low. Surprisingly, the crime rate is also higher where the police per capita is higher. We doubt that counties can effectively reduce their crime rates by eliminating police officers. Instead, we suspect that *polpc* is a result of crime, not a driver. Residents of counties with higher crime rates may be more willing to support larger police forces.

The following matrix of scatter plots compares variables related to sentencing with crime rates.

```
scatterplotMatrix(~ crmrte + prbpris + avgsen, data = crime, col = "dark blue",
                  diagonal=list(method ="histogram", breaks=25),
                  main = "Punishment Variables vs. Crime Rate", cex.labels = 1.4)
```

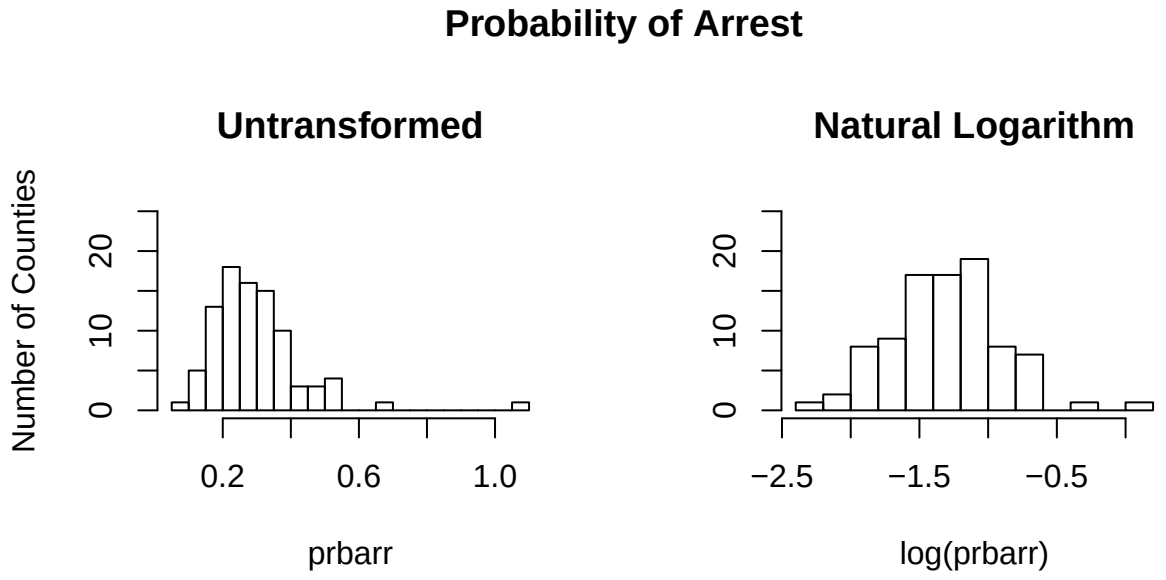


There are no indications of postive or negative correlations between crime rate and the probability of receiving a prison sentence or the average length of a prison sentence.

2.2 Arrest Model

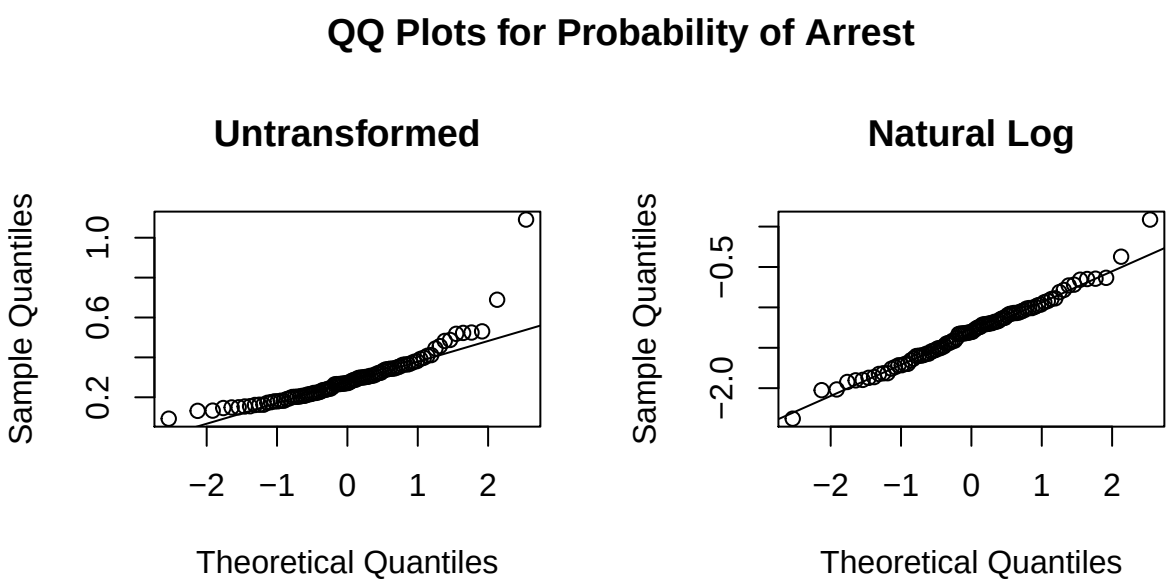
We will start with a simple model that regresses crime rate on the probability of being arrested. A histogram of the *prbarr* variable shows a slight rightward skew, which is slightly improved by taking the natural logarithm.

```
par(mfrow = c(1, 2), oma = c(0, 0, 1.2, 0))
hist(crime$prbarr, main = "Untransformed", xlab = "prbarr",
     ylab = "Number of Counties", ylim = c(0, 25), breaks = 15)
hist(log(crime$prbarr), main = "Natural Logarithm", xlab = "log(prbarr)",
     ylab = NULL, ylim = c(0, 25), breaks = 15)
title(main = "Probability of Arrest", outer = TRUE)
```



Taking the natural logarithm improves the normality of the *prbarr* variable.

```r
par(mfrow = c(1, 2), oma = c(0, 0, 1.2, 0))
qqnorm(crime$prbarr, main = "Untransformed"); qqline(crime$prbarr)
qqnorm(log(crime$prbarr), main = "Natural Log"); qqline(log(crime$prbarr))
title("QQ Plots for Probability of Arrest", outer = TRUE)
```

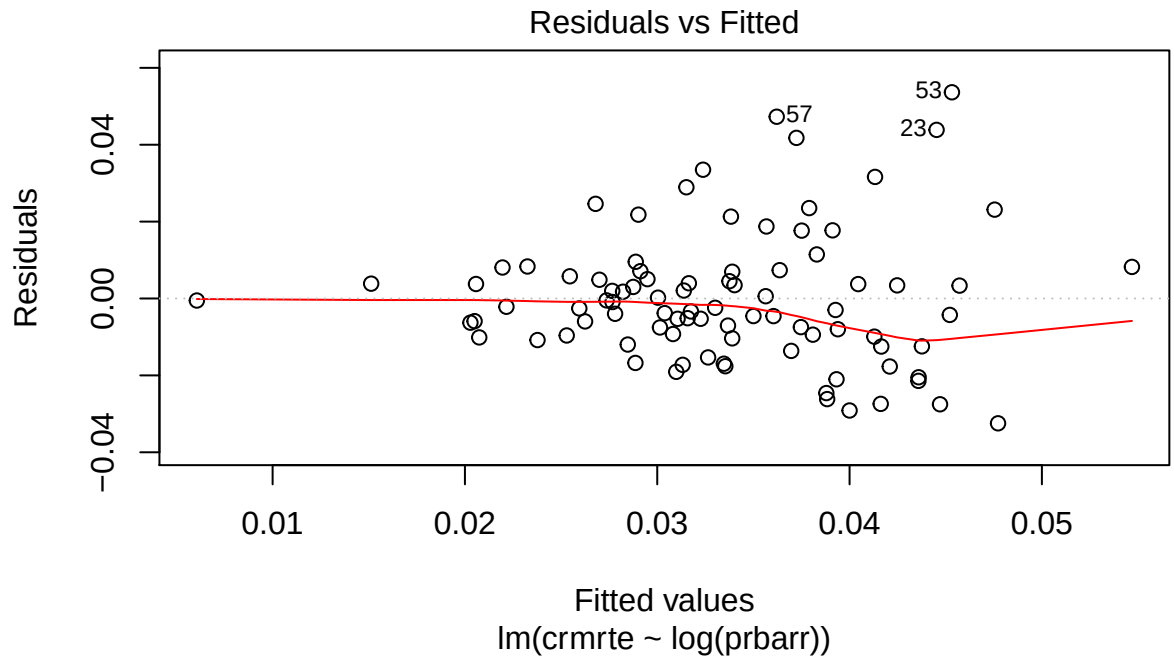


We will use the log transform for the *prbarr* variable because moves the variable closer to a normal distribution. Taking the log transform will also simplify evaluation of practical signicance. The units of the probability variables are not intuitive and taking the logarithm will enable us to express changes in percentage terms.

```r
lemod1 <- lm(crmrte ~ log(prbarr), data = crime)
```

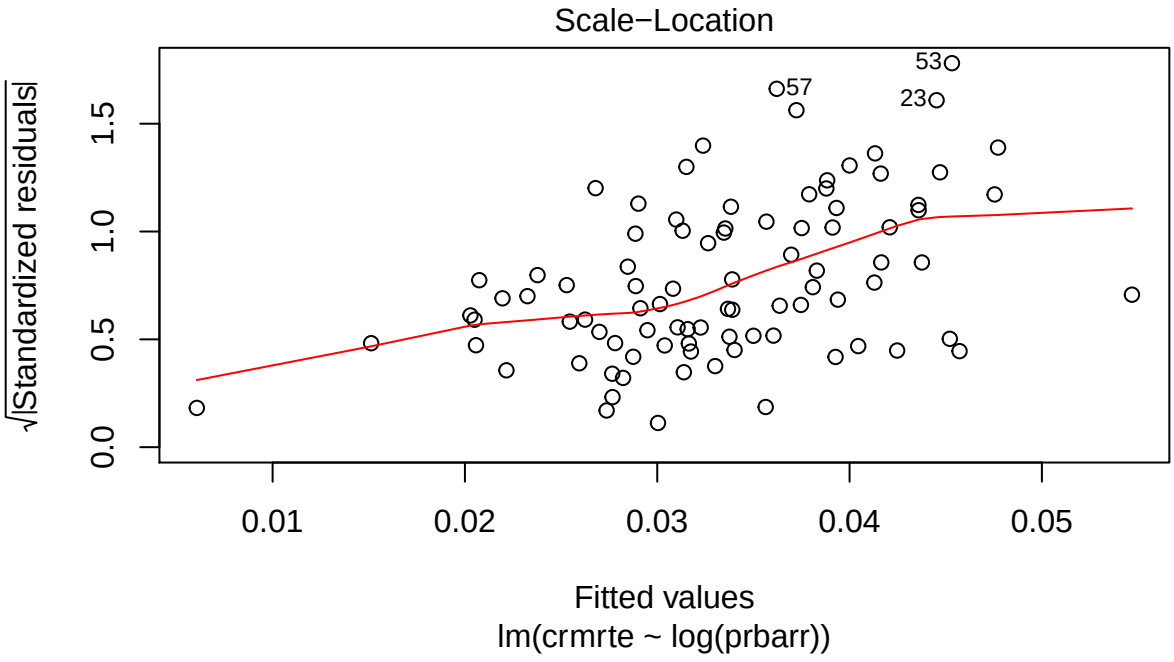
We will use heteroskedastic-robust standard errors because both the residuals and standardized residuals plots (see below) indicate heteroskedasticity.

```r
plot(lemod1, 1)
```



The residuals between 0.02 and 0.03 are closer to zero than the residuals between 0.03 and 0.05, indicating heteroskedasticity.

```
plot(lemod1, which = 3)
```



The linear model indicates a negative correlation between the probability of arrest and crime rates. The standardized residuals appear to slope upwards, also indicating heteroskedasticity.

The model results indicate a negative relationship between crime rates and arrests and the result is statistically significant. The t-test for the arrest coefficient is significant to a confidence level of 99%.

```
lemod1_skd <- coeftest(lemod1, vcov = vcovHC)
print(lemod1_skd)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.0077773  0.0050940  1.5268    0.1304
## log(prbarr) -0.0197279  0.0046590 -4.2344 5.616e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("R-Squared:", summary(lemod1)$r.squared)
```

```
## R-Squared: 0.1770441
```

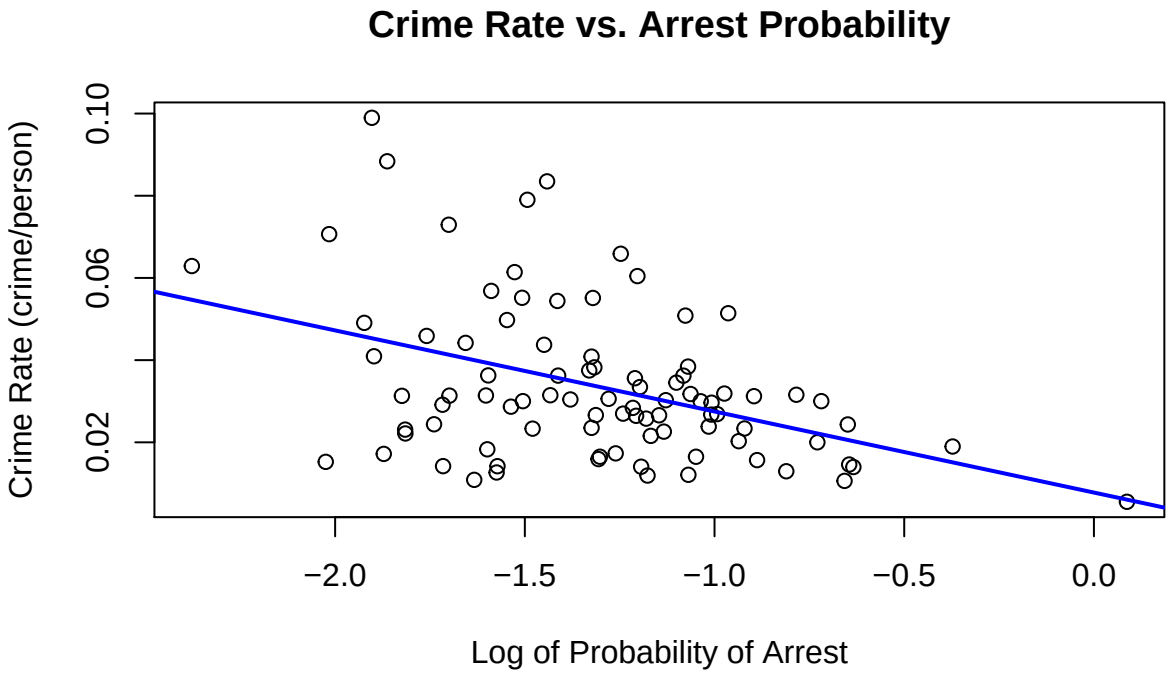
In terms of practical significance, a 10% increase in the probability of arrest corresponds to a reduction of 0.002 crimes per person. For counties with crime rates near the median of 0.03 crimes per person, a 10% increase in arrest probability corresponds to a 6.5% decrease in expected crime rate.

```
0.1 * coef(lemod1)[2]/median(crime$crmrte)
```

```
## log(prbarr)
## -0.06575544
```

The following plot shows the model for crime rate regressed on the probability of arrest.

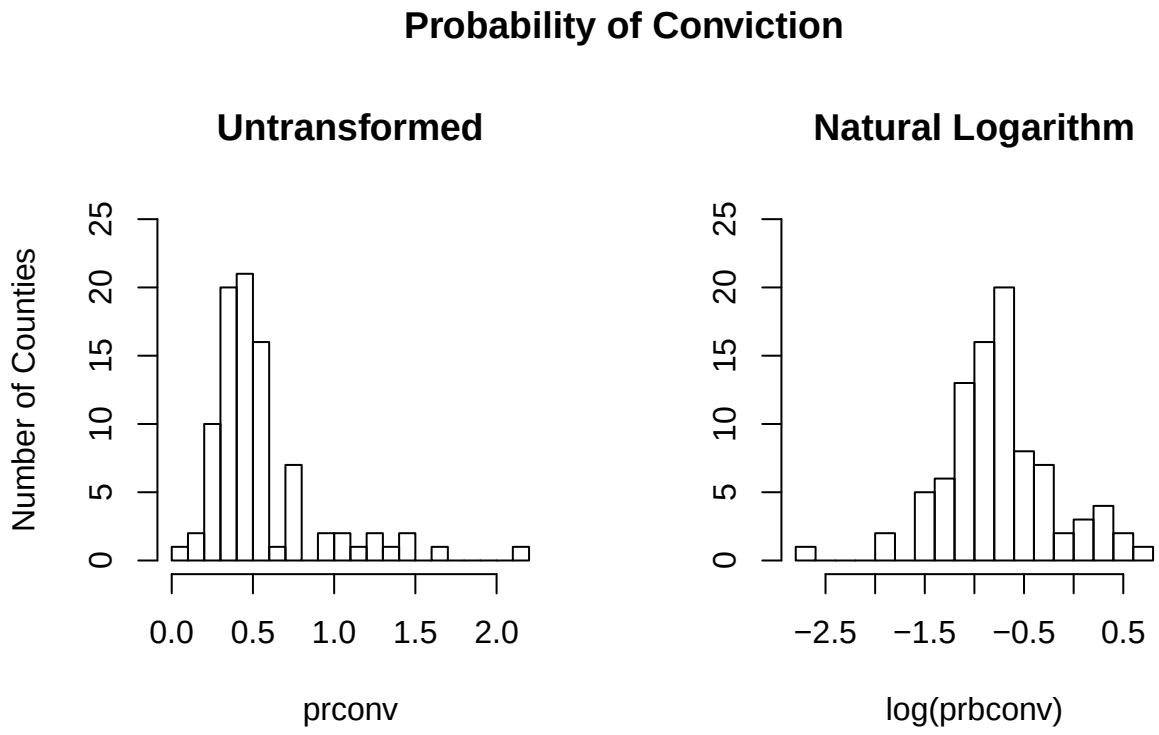
```
plot(log(crime$prbarr),crime$crmrte, main = "Crime Rate vs. Arrest Probability",
     xlab = 'Log of Probability of Arrest', ylab = 'Crime Rate (crime/person)')
abline(lemod1_skd[1:2, 1], col = "blue" , lwd = 2)
```



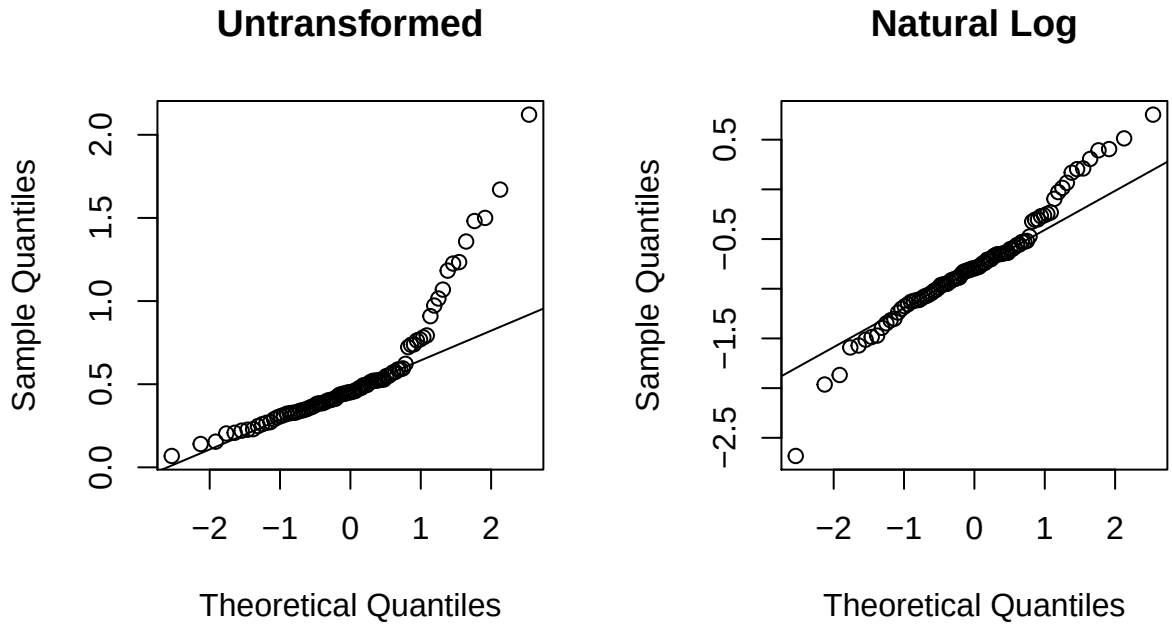
2.3 Arrest and Convictions Model

Similar to the probability of arrest, the probability of conviction has a rightward skew and it appears more normal after taking the natural logarithm. These affects are apparent in the histograms and QQ plots of *prbconv*.

```
par(mfrow = c(1, 2), oma = c(0, 0, 1.2, 0))
hist(crime$prbconv, main = "Untransformed", xlab = "prconv",
     ylab = "Number of Counties", ylim = c(0, 25), breaks = 15)
hist(log(crime$prbconv), main = "Natural Logarithm",
     xlab = "log(prbconv)", ylab = NULL, ylim = c(0, 25), breaks = 15)
title(main = "Probability of Conviction", outer = TRUE)
```




```r
par(mfrow = c(1, 2), oma = c(0, 0, 1.2, 0))
qqnorm(crime$prbconv, main = "Untransformed"); qqline(crime$prbconv)
qqnorm(log(crime$prbconv), main = "Natural Log"); qqline(log(crime$prbconv))
title("QQ Plots for Probability of Conviction", outer = TRUE)
```



In this model we will add the logarithm of *prbconv* as an independent variable. Transforming the variable with a natural logarithm will help to reduce the impact of outliers and the rightward skew. The log transform will make comparisons between arrest and conviction coefficients more intuitive. We continue to use heteroskedasticity-robust standar errors for the model.

```
lemod_arr_conv = lm(crmrte ~ log(prbarr) + log(prbconv), data = crime)
tcoe3 <- coeftest(lemod_arr_conv, vcov = vcovHC)
print(tcoe3)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  -0.0099487  0.0088923 -1.1188 0.2663067
## log(prbarr)  -0.0241640  0.0056086 -4.3084 4.309e-05 ***
## log(prbconv) -0.0158032  0.0045895 -3.4433 0.0008859 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("R-Squared:", summary(lemod_arr_conv)$r.squared)
```

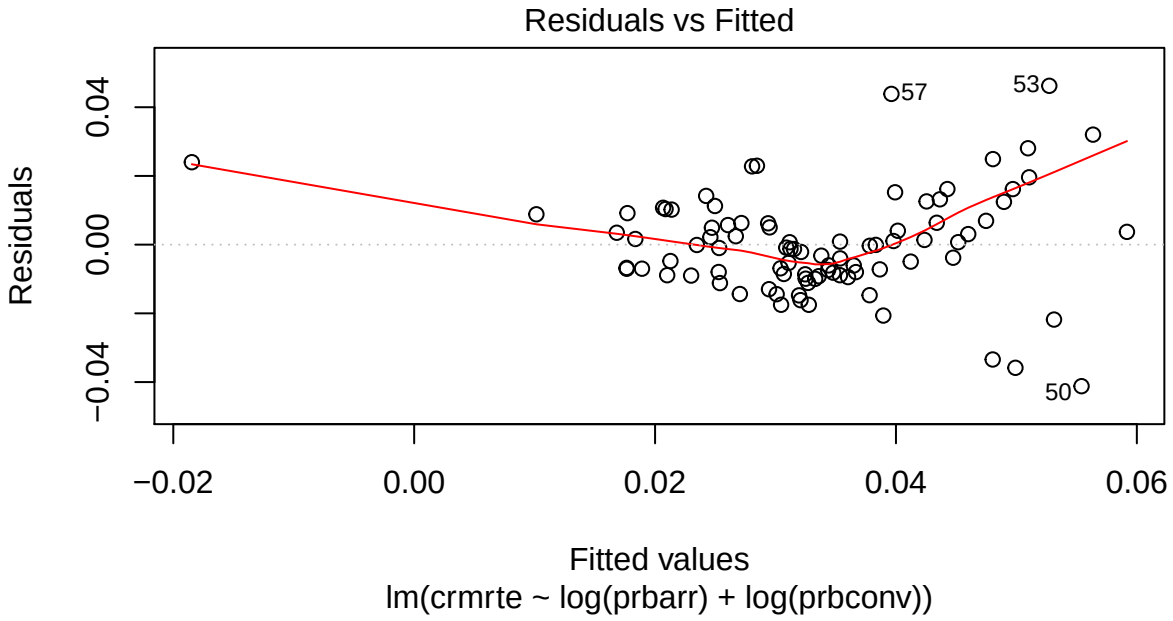
```
## R-Squared: 0.3867098
```

The coefficients for both *prbarr* and *prbconv* are statistically significant. The practical significance of *prbarr* has increased slightly with a 10% increase in *prbarr* corresponding to a reduction of 0.0024 crimes per person, compared to a reduction of 0.02 crimes per person for the model that omitted *prbconv*. The probability of conviction has slightly less practical significance, with a 10% increase resulting in a reduction of 0.0016 crimes per person. The R-squared value has doubled from the prior model.

Unfortunately this model has indications of a zero-conditional-mean violation, suggesting there is an omitted variable that has a relationship with both the probability of conviction and the crime rate. The residuals plot shows a strong u-shape to the residuals. Furthermore the model is predicting a negative crime rate for one county. In the next section we will evaluate whether adding additional justice system parameters will mitigate the zero-condictional mean violation.

```
par(oma = c(0, 0, 1, 0))
plot(lemod_arr_conv, which = 1)
title(main = "Arrests and Convictions Model", outer = TRUE)
```

Arrests and Convictions Model



The county with the negative fitted value is Madison County in far western North Carolina (NC). The county is identified in the data set by its federal information processing standard (FIPS) code of 115. Madison County is the only county for which a negative crime rate is predicted.

```
head(sort(predict(lemod_arr_conv, crime)))
```

```
##          51          58           8          84           5          52
## -0.01845891  0.01014628  0.01682007  0.01763684  0.01764828  0.01770610
```

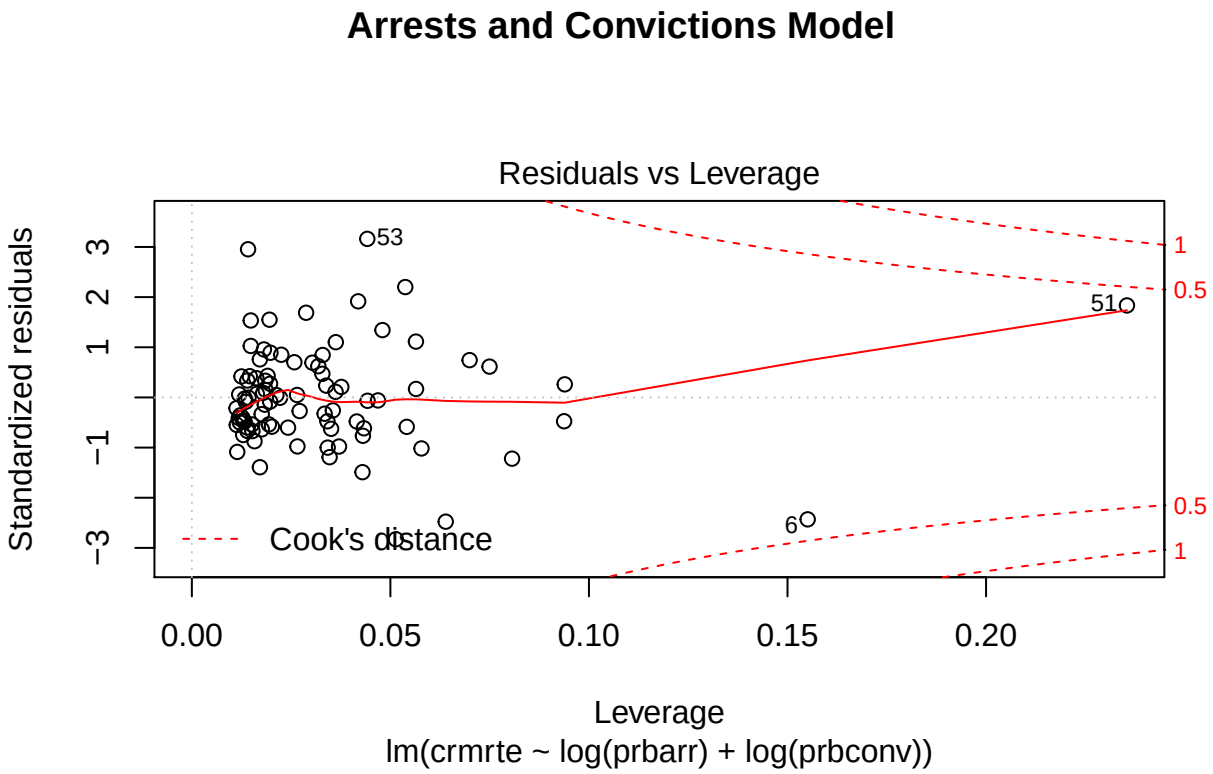
Madison County has the highest probability of arrest of all counties in NC, and one of the highest probabilities of conviction. These factors are causing the negative predicted crime rate.

```
head(crime[order(-crime$prbarr, -crime$prbconv),
           c("county", "crmrte", "prbarr", "prbconv")])
```

```
##    county    crmrte   prbarr   prbconv
## 51    115 0.0055332 1.090910 1.5000000
## 58    131 0.0189848 0.689024 0.4955750
## 79    173 0.0139937 0.530435 0.3278690
## 6      11 0.0146067 0.524664 0.0683761
## 62    139 0.0243470 0.522696 0.2894740
## 5       9 0.0106232 0.518219 0.4765630
```

Madison County is in far western NC and has a farily small population (23,000 in 2017). The county is 97.5% white and is a dry county, meaning the sale of alcohol is not allowed unless specifically allowed by individual towns. There are also indications that until the early 1900s, Madison county was one of several sundown counties in NC (J.W. Lowen, *Sundown Towns, A Hidden Dimension of American Racism*, 2005). In spite of these attributes, Madison County is not exerting a concerning amount of leverage on this model, as indicated by its Cook's distance of less than 0.5 (see point 51 on leverage chart below). We will leave Madison County in the data set.

```
par(oma = c(0, 0, 1, 0))
plot(lemod_arr_conv, which = 5)
title(main = "Arrests and Convictions Model", outer = TRUE)
```



The variance inflation factor indicates there is little multicolinearity between *prbarr* and *prbconv*.

```
vif(lemod_arr_conv)
```

```
##  log(prbarr) log(prbconv)
##     1.042696     1.042696
```

2.4 Model with All Criminal Justice Parameters

The model in section 2.3 focused on probabilities for conviction and arrest because our exploratory data analysis suggested a relationship between these variables and crime rate. We observed indications of a zero-conditional mean violation and a possible omitted variable after we added probability of conviction to the model. In this section we will add other justice system variables to the model to evaluate whether these variables mitigate the zero-conditional mean violation.

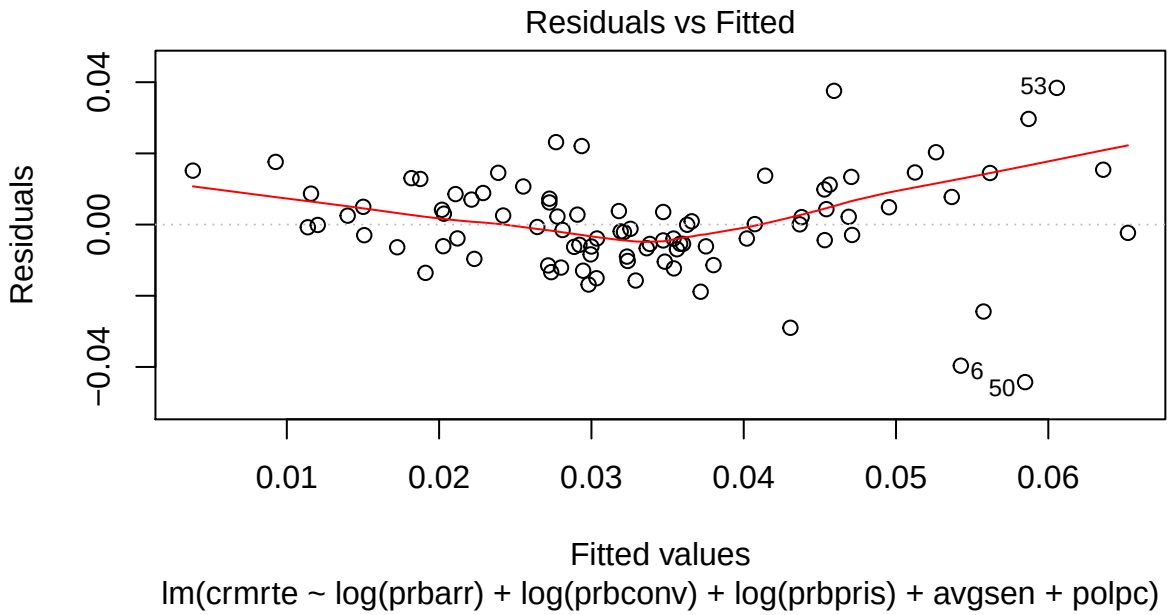
We will take the natural logarithm of the probability of receiving a prison sentence so that all probabilities are handled consistently.

```
lemod_all = lm(crmrte ~ log(prbarr) + log(prbconv) + log(prbpris) +
               avgsen + polpc, data = crime)
```

Adding the additional parameters eliminated the negative fitted crime rate, but the indications of a zero-conditional-mean violation is even more readily apparent, due to the more even distribution of data along the x axis.

```
par(oma = c(0, 0, 1.2, 0))
plot(lemod_all, which = 1)
title(main = "All Justice System Variables", outer = TRUE)
```

All Justice System Variables



A review of the heteroskedastic standard errors for this model shows that none of the additional parameters are statistically significant.

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

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -0.01592050  0.01112571 -1.4310  0.156150
## log(prbarr)  -0.02731265  0.00608581 -4.4879 2.269e-05 ***
## log(prbconv) -0.01652389  0.00521018 -3.1715  0.002118 **
## log(prbpris)  0.00495274  0.00875542  0.5657  0.573122
## avgsen       -0.00054092  0.00066269 -0.8162  0.416667
## polpc         6.48675899  4.25880127  1.5231  0.131479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In fact, if we test whether all three added variables show joint significance (below) we obtain a large p-value (0.4112) and low F statistic (0.9694). Therefore, we can not reject the null hypothesis that *log(prbpris)*, *avgsen* and *polpc* are irrelevant to crime rate.

```
linearHypothesis(lemod_all, c("log(prbpris) = 0", "avgsen = 0", "polpc = 0"),
                 vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(prbpris) = 0
## avgsen = 0
```

```
## polpc = 0
##
## Model 1: restricted model
## Model 2: crmrte ~ log(prbarr) + log(prbconv) + log(prbpris) + avgsen +
##     polpc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     87
## 2     84  3 0.9694 0.4112
```

We will evaluate whether controlling for demographic variables mitigates the zero conditional mean violation in the next section.

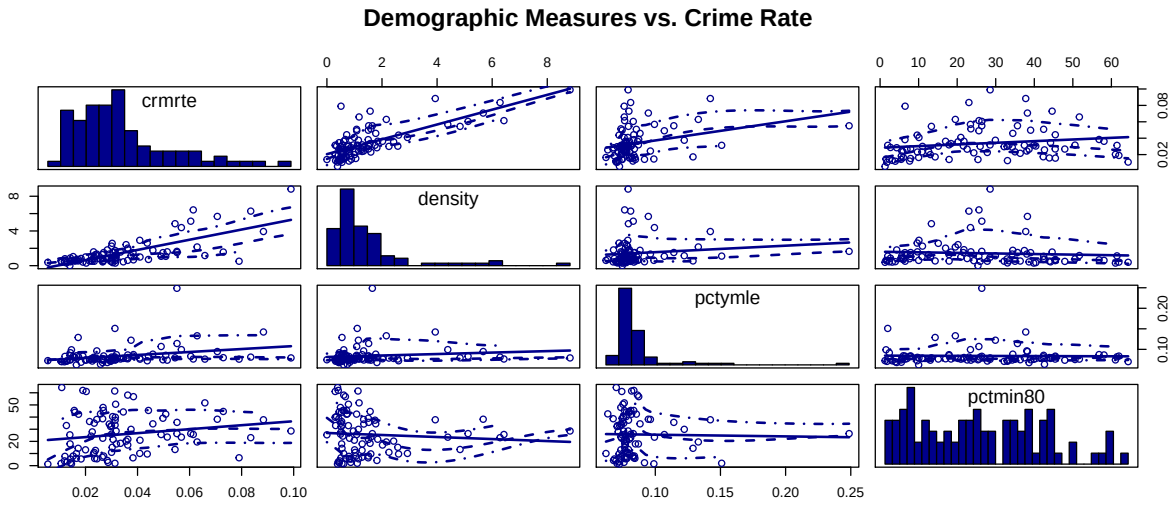
3. Demographic Model

The fitted values vs. residuals plots from the basic model suggested violations of the zero conditional mean assumption, possibly due to omitted variables that are affecting the crime rate (crmrte). There are several variables that represent each county's demographic information that may have a relationship to crime rate. While it is impractical or impossible for government leaders to influence these demographic variables, including them in crime models may improve our ability to identify other variables that have a relation to crime.

There are three metric demographic variables in the data set:

- density: the number of people per square mile
- pctymle: the proportion of the population between the ages of 15 and 24
- pctmin80: the proportion of the county's population that identified as a minority in the 1980 census.

```
scatterplotMatrix(~ crmrte + density + pctymle + pctmin80, data = crime,
                  col = "darkblue", diagonal=list(method ="histogram",
                                                  breaks=25),
                  main = "Demographic Measures vs. Crime Rate",
                  cex.labels = 1.4)
```



The plots indicate that population density has at least a moderate positive correlation with crime rates, which is confirmed by calculating the correlation between crime rate and the other variables:

```
cor(crime[c("crmrte", "density", "pctymle", "pctmin80")])[1, ]
```

```
##    crmrte   density   pctymle  pctmin80
## 1.0000000 0.7283706 0.2903397 0.1816506
```

The correlation matrix suggests weak positive correlations between crime rate and the proportion of young males or minorities in a county.

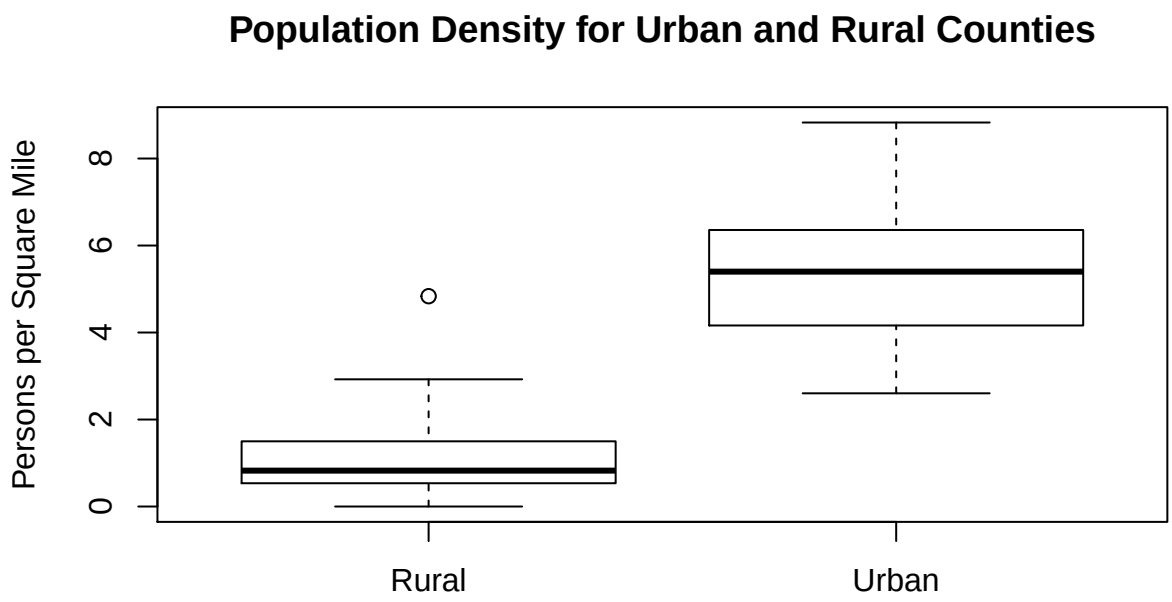
In addition to the metric variables, there are three demographic indicator variables in the data set:

- west: 1 if the county is in western NC, 0 otherwise
- central: 1 if the county is in central NC, 0 otherwise
- urban: 1 if the county is located in a standard metropolitan statistical area (SMSA), 0 otherwise.

3.1 Evaluation of Population Density and Urban Status

It is plausible that urban counties would have higher population densities.

```
boxplot(density ~ urban, data = crime, names = c("Rural", "Urban"),
        main="Population Density for Urban and Rural Counties",
        xlab = NULL, ylab = "Persons per Square Mile")
```



Only two rural (non-SMSA) counties have population densities that overlap with urban counties. While the two variables are not strictly colinear because density is metric and urban is logical, they do appear to be measuring the same feature. Including both variables in a model could result in larger standard errors. We prefer to use the density variable because it provides more information than the urban Bernoulli variable.

The R summary of the density variable reveals a potentially spurious density value for one county. The minimum density is more than 10,000 times smaller than the first quartile value.

```
summary(crime$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## 0.00002 0.54718 0.97925 1.43567 1.56926 8.82765
```

In fact, the next smallest density is still 15,000 times larger than the minimum value.

```
pop_density <- crime[order(crime$density), "density"]
head(pop_density)
```

```
## [1] 0.0000203422 0.3005714420 0.3009985690 0.3167155390 0.3503981830
## [6] 0.3858093020
```

```
pop_density[1] * 1562
```

```
## [1] 0.03177452
```

The largest county in NC in terms of land area is Dare county, with an area of 1,562 square miles. The minimum population density corresponds to a population of 0.03 people. The population density of 0.00002 is likely spurious and we will eliminate this point from the data set.

```
crimef <- crime[crime$density > 0.01, ]
paste("Rows removed: ", nrow(crime) - nrow(crimef))
```

```
## [1] "Rows removed:  1"
```

```
paste("New minimum density:", min(crimef$density))
```

```
## [1] "New minimum density: 0.300571442"
```

3.2 Evaluation of County Location

To assist with graphing, we will convert the location variables to a factor. We assume that counties that are neither in the west or central portion of NC are in eastern NC.

```
# Create a "loc" column of type factor
cloc <- function(west, central){
    if(west) loc_lbl = "west"
    else if(central) loc_lbl = "central"
    else loc_lbl = "east"
    return(loc_lbl)
}
crimef$loc <- pmap_chr(list(crimef$west, crimef$central), cloc)
crimef$loc <- as.factor(crimef$loc)
```

We will also add an *east* indicator column to the dataframe, which will be used in later models.

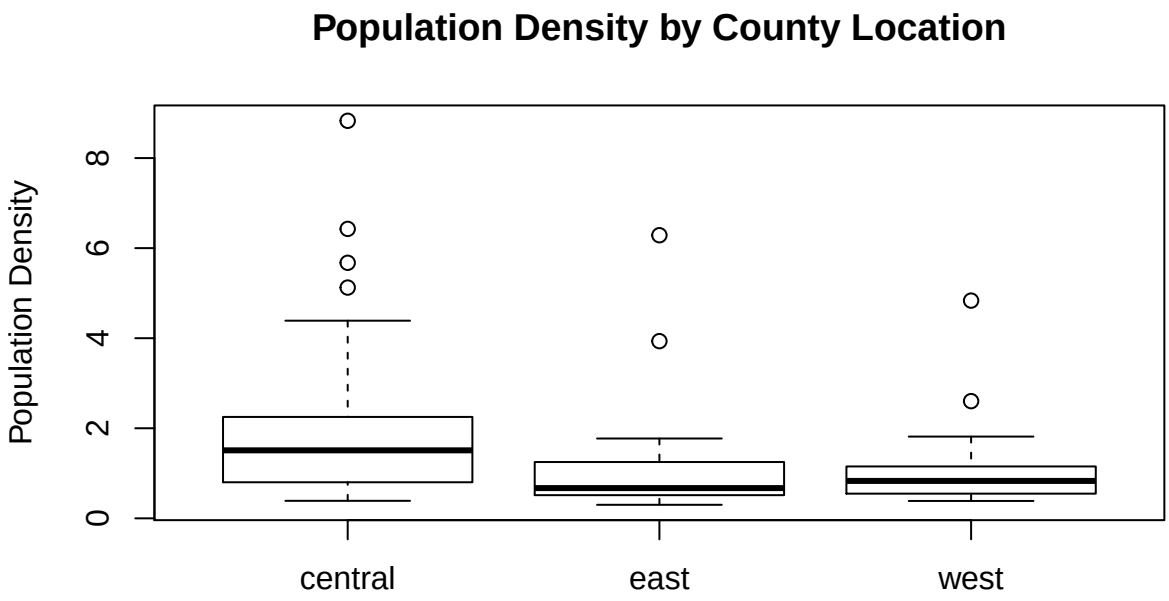
```
east <- function(west, central){
    if(!west & !central) label = 1
    else label = 0
    return(label)
}
crimef$east <- pmap(list(crimef$west, crimef$central), east)
crimef$east = as.integer(crimef$east)
```

A check of crime rate by county location indicates that western NC counties have noticeably lower crime rates.

```
boxplot(crmrte ~ loc, data = crimef,
        main="Crime Rate by County Location",
        xlab = NULL, ylab = "Crimes per Person")
```

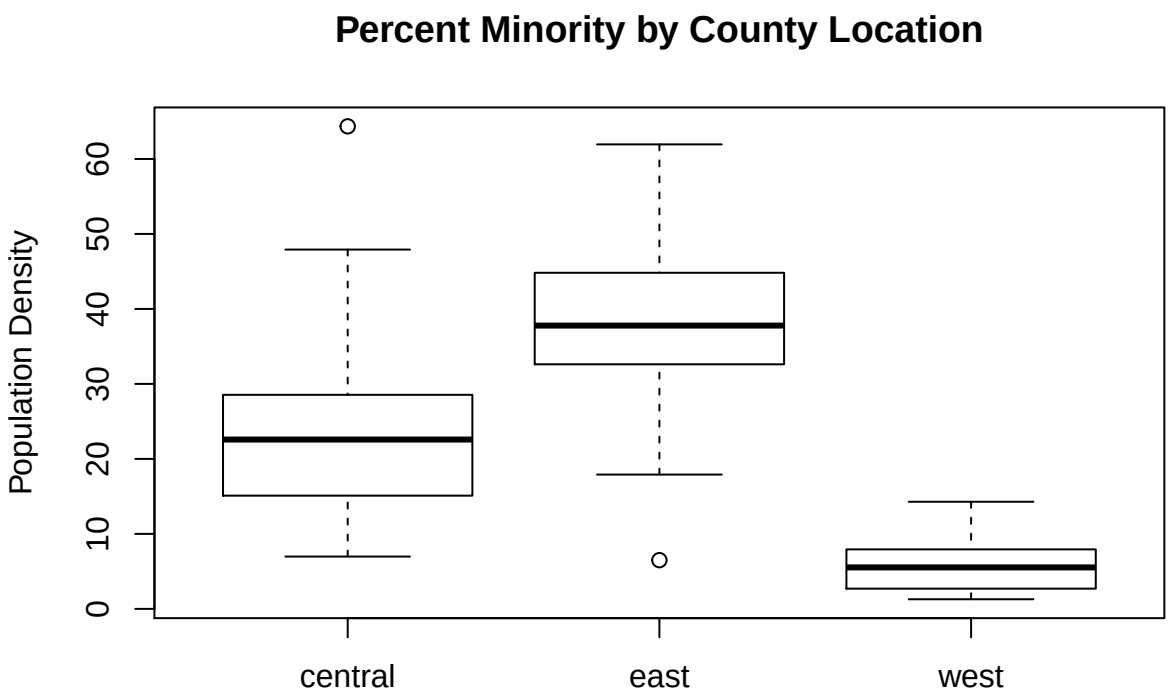
Comparison of population densities indicates that density is not responsible for the difference in crime rates. Eastern and western counties have a similar range of population densities but the median crime rate is higher in eastern counties.

```
boxplot(density ~ loc, data = crimef,
        main="Population Density by County Location",
        xlab = NULL, ylab = "Population Density")
```



The percent of minorities is noticeably lower in western counties.

```
boxplot(pctmin80 ~ loc, data = crimef,
        main="Percent Minority by County Location",
        xlab = NULL, ylab = "Population Density")
```



The effect of county location on crime rate suggests that models can be improved by controlling for location. There is a strong relationship between the percentage of minorities and location, which can be seen by conducting a two-sample t-test.

```
t.test(crimef$pctmin80 ~ crimef$west)
```

```
##
##  Welch Two Sample t-test
##
## data:  crimef$pctmin80 by crimef$west
## t = 12.912, df = 86.275, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  21.81308 29.75135
## sample estimates:
## mean in group 0 mean in group 1
##       31.799908        6.017693
```

Including both the percent minorities and the county location in a model could result in smaller coefficients due to the relationship between minorities and county location. We chose to include an indicator variable *west* in our model. The culture of western NC has differed from eastern NC since colonial times and we suspect that the percentage of minorities is just one aspect of the cultural differences.

There is one county that is flagged as both a central and a western county.

```
crimef[crimef$west + crimef$central > 1,
       c("county", "west", "central", "crmrte")]
```

```
##    county west central    crmrte
```

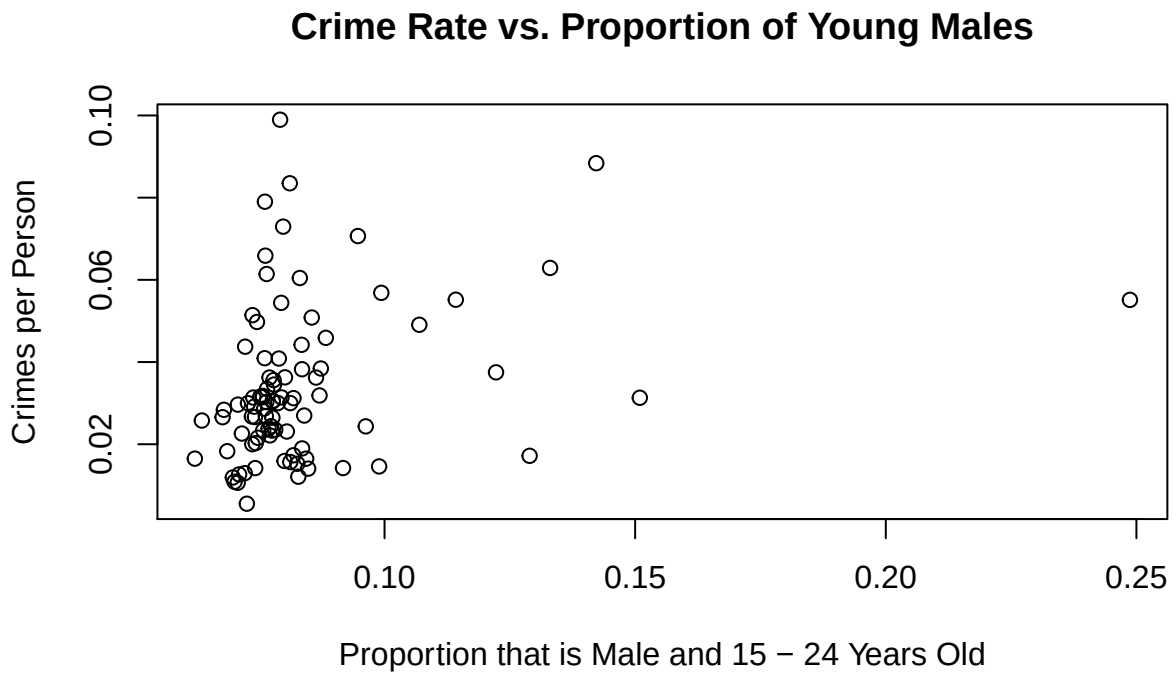
```
## 33     71    1       1 0.0544061
```

County 71 is Gaston County, which is in the southwestern part of the Charlotte metropolitan area. Gaston county is located towards the western part of North Carolina and is considered to be a western county per our analysis.

3.3 Percentage of Young Males

The data set includes a variable that represents the percentage of the population that is male and between the ages of 15 and 24.

```
plot(crmrte ~ pctymle, data = crimef,
     main = "Crime Rate vs. Proportion of Young Males",
     xlab = "Proportion that is Male and 15 - 24 Years Old",
     ylab = "Crimes per Person")
```



There is one county, Onslow County, that appears to be an outlier, with close to 25% of its population being young and male. We were able to identify the county by the Federal Information Processing Standards (FIPS) code, which is used in the data set to identify each county. Onslow county is the site of Camp Lejuene, an active Marine Corps base with over 40,000 active-duty personnel, which explains the high proportion of young males.

```
head(crimef[order(-crimef$pctymle), c("county", "pctymle")], 3)
```

```
##    county   pctymle
## 59    133 0.2487116
## 86    189 0.1509264
## 23     51 0.1422378
```

We will leave Onslow county in the data set. Unlike the county that was eliminated due to an impossibly low population density, we believe the reported proportion of young males for Onslow County is correct. In addition, the percentage of young males does not appear to have as strong of a relationship to crime rates as

population density. We will check models to ensure the Onslow County data point is not overly influencing the results.

3.4 Crime Model with Demographics

Earlier we evaluated a model that only included factors related to criminal justice. Now we will construct a model that includes the significant demographic information as previously discussed in this analysis

For our next model we will regress crime rate on density, the *west* indicator variable, the percentage of young males in the population, and the natural logarithms of the probabilities for arrest and conviction. The logarithms for arrest and conviction probabilities flatten the data in the residuals plots and they make the results of the regression easier to understand. Because of the logarithms, a fixed percentage change in a probability will have the same impact on crime rate, regardless of the magnitude of the initial probability.

```
md_cd <- lm(crmrte ~ log(prbarr) + log(prbconv) + density +
            west + pctymle, data = crimef)
```

3.4.1 Evaluation of CLM Assumptions

The classical linear model (CLM) consists of six multiple linear regression model (MLR) assumptions. While only the first five assumptions (MLR 1 - 5) are needed for multiple least squares regression to provide the best linear unbiased estimator (BLUE), MLR 6 is needed to justify use of a t distribution to evaluate significance of coefficients.

For this model we will conduct a detailed review of the CLM assumptions.

3.4.1.1 MRL 1: Model is Linear in Parameters MLR 1 requires that the population model be represented by an equation of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

We are unable to determine if crime rates in NC follow this model. Nevertheless we believe the use of such a linear model is justified due to its flexibility and its demonstrated utility in a wide range of problem domains. Furthermore, our intent is not to precisely predict crime rates in individual counties, but to assist in identification of policies that have the potential to reduce crime rates. Identifying a linear model that best fits the population will help us to identify promising factors that can be targeted with policy initiatives.

3.4.1.2 MLR 2: Random Sampling MLR 2 requires that our data is randomly selected. Our data set contains 90 of the 100 counties in North Carolina. The ten missing counties and their FIPS codes are listed below:

- 029, Camden County
- 031, Carteret County
- 043, Clay County
- 073, Gates County
- 075, Graham County
- 095, Hyde County
- 103, Jones County
- 121, Mitchell County
- 177, Tyrrell County
- 199, Yancey County

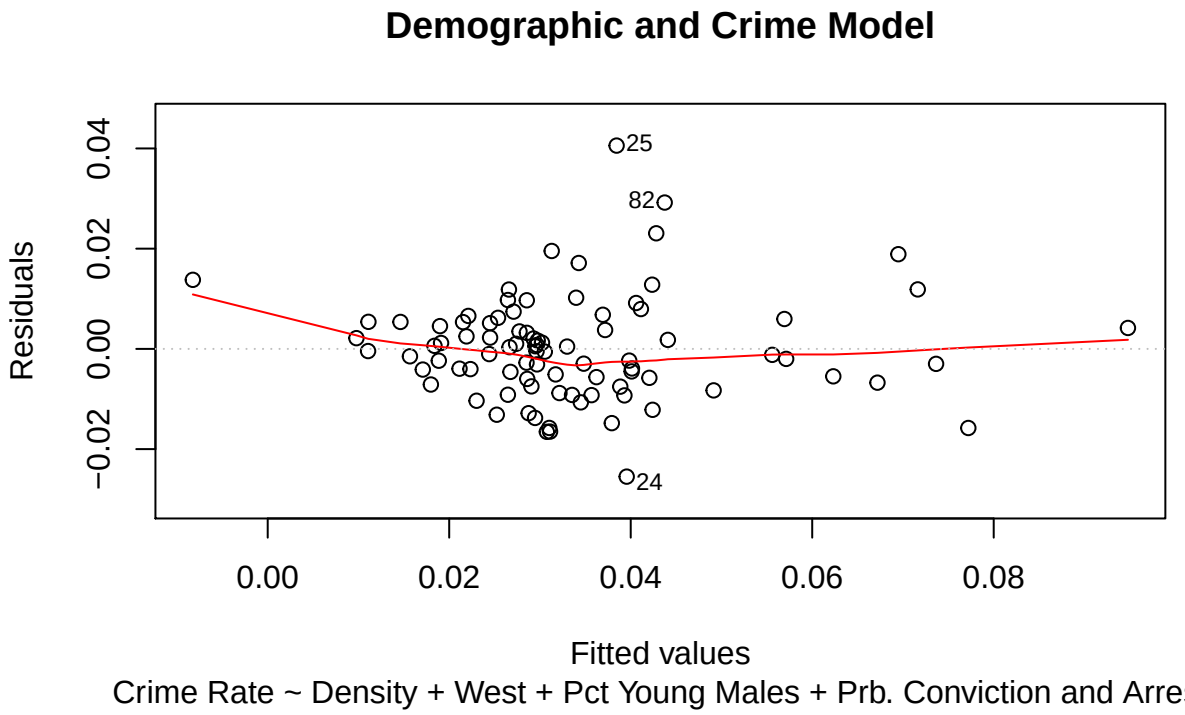
This list includes the six least populated counties in NC, as well as the the eighth least populated county per the 2017 census estimate. The remaining three counties range in population from approximately 10,000 people to 17,000 people. This indicates that the counties were not randomly selected and that counties with the lowest populations are underrepresented in the sample. We do not expect this cluster of omitted

counties to significantly degrade the results of our analysis. This group of counties is likely to have a wider random variation in metrics due to their low population. Also, government policies that are effective in more populated counties will benefit a larger number of people. Nevertheless, researchers who are interested in the behavior of crime rates in sparsely populated counties may choose to consider a different data set.

3.4.1.3 MLR 3: No Perfect Colinearity We have purposefully constructed our model to avoid using highly correlated independent variables in the same model. For example, we decided to use *west* and *density* and exclude *pctmin80* and *urban* in our model. It is highly unlikely that such unrelated variables as county location, probability of conviction, or percentage of young males would exhibit perfect Colinearity

3.4.1.4 MLR 4: Zero Conditional Mean To evaluate MLR 4 we will first view the residuals vs. fitted values plot for the demographic model.

```
capt = paste("Crime Rate ~ Density + West + Pct Young Males",
"+ Prb. Conviction and Arrest")
plot(md_cd, which = 1, main = "Demographic and Crime Model",
     caption = NULL, sub.caption =capt)
```

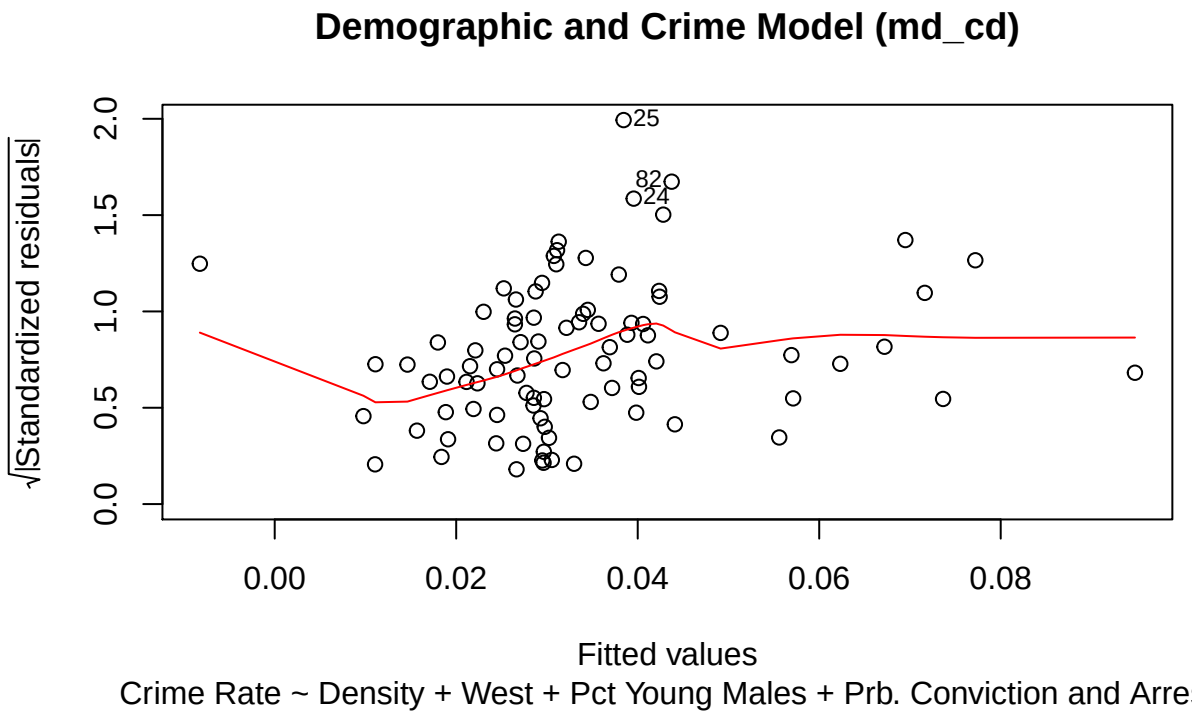


The magnitude of the residuals is similar to the magnitude of the fitted values, indicating that there can be significant errors in the prediction of crime rates for individual counties. Also, even after adding demographic variables, there is still one county for which the model predicts a negative crime rate (data point at far left in plot above). This data point corresponds to county 115 which was discussed in section 2. The remaining data points on the residuals plot are centered around 0 and there is no readily apparent upward or downward slope. This indicates that we are not violating MLR 4.

3.4.1.5 MLR 5: Homoskedasticity MLE 5 requires that the variance of the residuals is a constant and is not conditional on our independent variables: $\text{Var}(u|x_1, ...x_k) = \sigma^2$.

```
capt = paste("Crime Rate ~ Density + West + Pct Young Males",
"+ Prb. Conviction and Arrest")
```

```
plot(md_cd, which = 3, main = "Demographic and Crime Model (md_cd)",
     caption = NULL,
     sub.caption = capt)
```



The standardized residuals plot above suggests that variance of residuals is smaller at very low crime rates. We will further investigate the possibility of hetero heteroskedasticity with a Breusch-Pagan test.

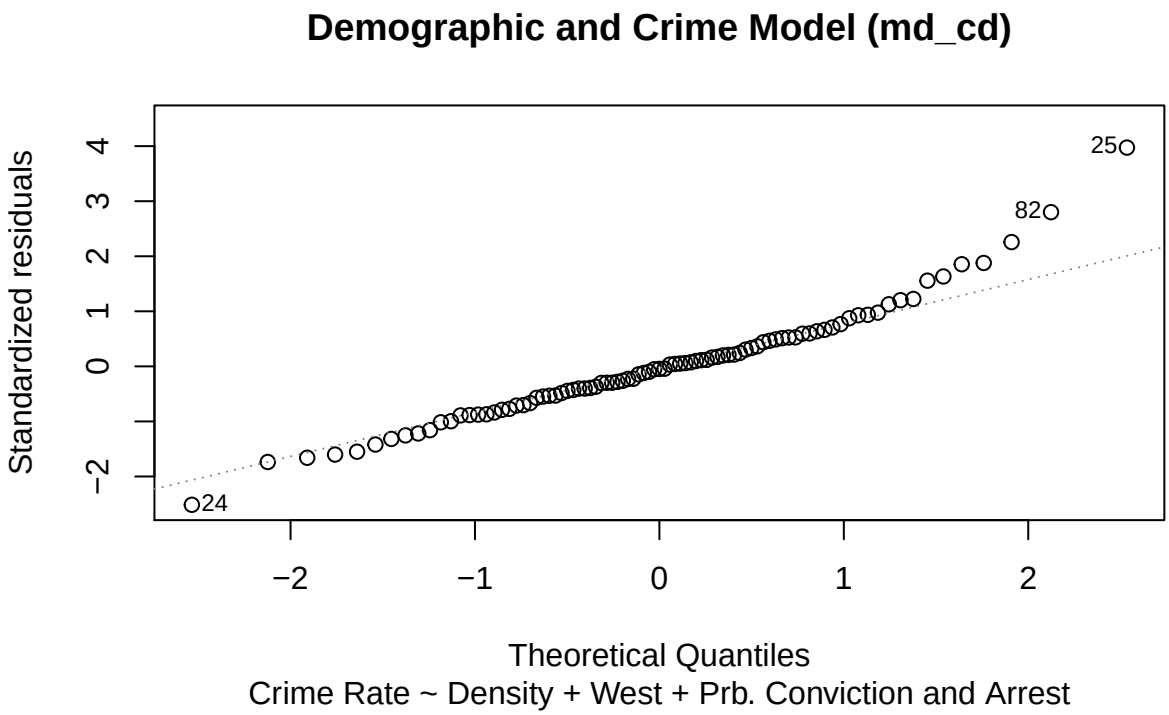
```
bptest(md_cd)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  md_cd
## BP = 16.457, df = 5, p-value = 0.005654
```

For the Breusch-Pagan test, H_0 is that the model is homoskedastistic. The Breusch-Pagan test calculated a p-value of 0.003, which results in rejecting the null hypothesis at a confidence level of over 99%. We will use heteroskedasticity-robust standard errors to compensate for this violation of MLR-5.

```
plot(md_cd, which = 2, main = "Demographic and Crime Model (md_cd)",
     caption = NULL,
     sub.caption = "Crime Rate ~ Density + West + Prb. Conviction and Arrest")
```

3.4.1.6 MLR 6: Normality of Population Errors



The plot above indicates some departure from normality at the tails of our distribution. With a sample size of 89 we conclude that per the central limit theorem (CLT), the distribution of the linear regression coefficients will be sufficiently normal, allowing the use of the t distribution to evaluate significance.

3.4.1.7 Summary of MLR Assumptions MLR 1, MLR 3, and MLR 4 are met. MLR 5, homoskedasticity, is not met, but use of heteroskedastistic robust standard errors will compensate. MLR 6 is not met, but our sample size is sufficiently large for the coefficient distributions to be normal per the CLT. Finally, MLR 2 is violated to the extent that sparsely populated counties are underrepresented in our sample. We expect that the data set is sufficient to evaluate crime rate trends for moderately to densely populated counties, but the results may not extend to counties with small populations.

3.4.2 Demographic Model Significance

```
md_cd_skd <- coeftest(md_cd, cov = covHC)
md_cd_skd
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -0.00214890  0.00545111 -0.3942 0.6944343
## log(prbarr)  -0.01100768  0.00329795 -3.3377 0.0012660 **
## log(prbconv) -0.00888646  0.00226272 -3.9273 0.0001768 ***
## density       0.00675083  0.00084007  8.0360 5.516e-12 ***
## west         -0.01056894  0.00268102 -3.9421 0.0001678 ***
## pctymle       0.08857477  0.05139813  1.7233 0.0885565 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will refer to the median crime rate in NC during our discussion of practical significance.

```
cat("Median Crime Rate:", median(crimef$crmrte), " Crimes per Person")
```

```
## Median Crime Rate: 0.0300184  Crimes per Person
```

All variables used in the demographic model, except for the percentage of young males, are statistically significant to at least 99% confidence.

- Both the probabilities of arrest and conviction have negative coefficients, indicating a rise in arrests or convictions would reduce the expected crime rate. A 10% rise in either the probability of arrest or conviction results in an increase in the crime rate of approximately 0.001, or a 3% reduction for a county with the median crime rate.
- For every unit increase in population density, the expected value of crime rate will increase by 0.007, or approximately 20% of the median crime rate.
- Counties located in the western part of NC have an expected crime rate that is 0.01 crimes per person lower than non-western counties, all other factors being equal. 0.01 crimes per person is 33% of the median crime rate.
- The percentage of young males in the county is statistically significant to only 90% confidence. A 1% increase in the percentage of the population that is young and male will result in an increase in crime rate of approximately 0.0009 crimes per person, which would be a 3% increase in crime for a county with the median crime rate.

The demographic model accounts for 70% of the variation observed in the data set.

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

```
## [1] 0.7003108
```

4. The Expanded Model

The next model will include unused economic, demographic, and criminal justice variables to evaluate whether any of these variables sufficiently improve the model without adding unnecessary complexity.

4.1 Financial Variables

The previous model did not include financial variables, which might might provide insight into the economic health of a county. We will add these variable into the next model and show that there is no joint significance. We will compare an expanded model with financial variables to a restricted model using an F test to show if the general economic health, as indicated by the average wages and tax revenue per capita, of a county has significance in predicting the crimerate.

This first set of financial variables that we'll be adding to our proposed model include *taxpc*, *wcon*, *wtuc*, *wtrd*, *wfir*, *wser*, *wmfg*, *wfed*, *wsta*, and *wloc*.

For this test, our null and alternative hypotheses are stated as:

$$H_0 : \beta_{taxpc} = 0, \beta_{wcon} = 0, \beta_{wtuc} = 0, \beta_{wtrd} = 0, \beta_{wfir} = 0, \beta_{wser} = 0, \beta_{wmfg} = 0, \beta_{wfed} = 0, \beta_{wsta} = 0, \beta_{wloc} = 0.$$

$$H_1 : H_0 \text{ is not true.}$$

First we can start by building our model and looking at the effects of adding the financial variables individually.

```
md_cd_fin <- lm(crmrte ~ log(prbarr) + log(prbconv) + density + west +
                taxpc + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
                wsta + wloc, data = crimef)
coeftest(md_cd_fin, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  -2.3452e-02  2.8395e-02 -0.8259 0.411503
## log(prbarr)  -1.2937e-02  4.9447e-03 -2.6163 0.010771 *
## log(prbconv) -9.4579e-03  3.9539e-03 -2.3921 0.019295 *
## density       5.4209e-03  2.1168e-03  2.5608 0.012480 *
## west         -9.1270e-03  2.7839e-03 -3.2786 0.001591 **
## taxpc         2.6404e-04  3.2694e-04  0.8076 0.421906
## wcon         -1.9025e-05  4.0819e-05 -0.4661 0.642530
## wtuc         -4.7396e-07  2.3931e-05 -0.0198 0.984252
## wtrd         -3.7753e-05  9.5895e-05 -0.3937 0.694943
## wfir         -2.4192e-05  3.4637e-05 -0.6984 0.487085
## wser         -6.2905e-06  7.9397e-05 -0.0792 0.937064
## wmfg          3.3531e-06  1.9992e-05  0.1677 0.867258
## wfed          6.0898e-05  3.9138e-05  1.5560 0.123982
## wsta         -1.6126e-05  3.4980e-05 -0.4610 0.646137
## wloc          5.7796e-05  7.8460e-05  0.7366 0.463672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
r2_ur <- summary(md_cd_fin)$r.squared
r2_ur
```

```
## [1] 0.7400244
```

We can see in the above table that the unrestricted model coefficients have decreased in significance for all of the independent variables except *west*, that none of the financial variables achieve strong significance on an independent basis, but the r-squared value has increased. Next we will continue with the F test to determine if the financial variables may provide joint significance to our model.

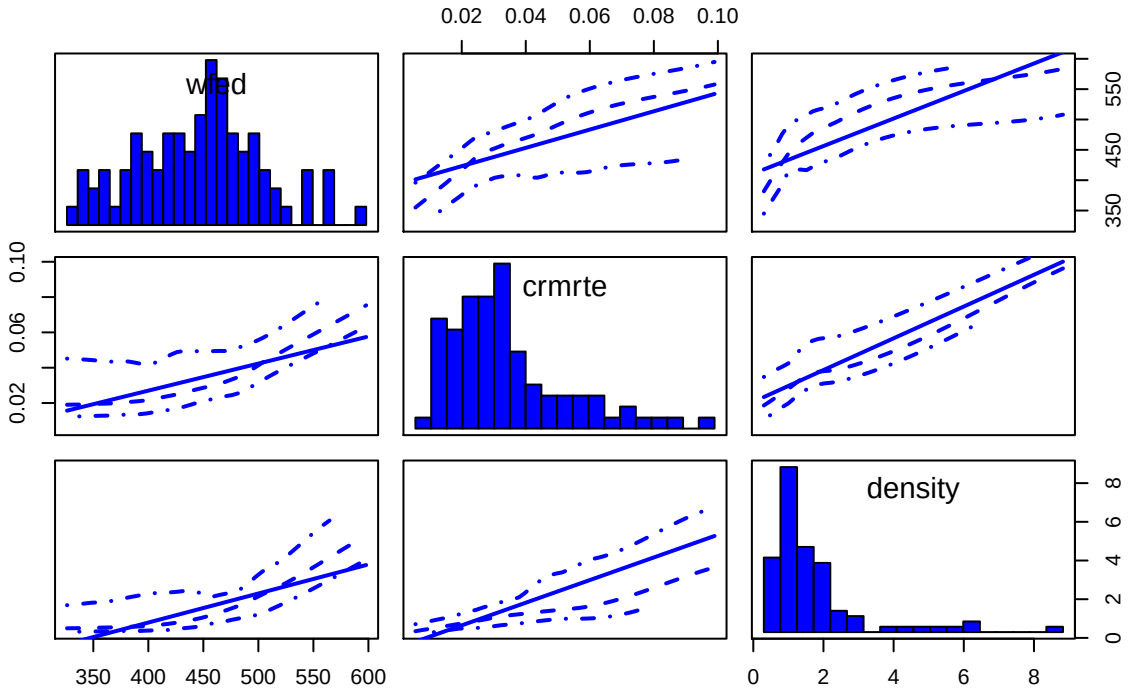
```
linearHypothesis(md_cd_fin, c("taxpc= 0", "wcon= 0", "wtuc= 0", "wtrd= 0",
                              "wfir= 0", "wser= 0", "wmfg= 0", "wfed= 0",
                              "wsta= 0", "wloc= 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## taxpc = 0
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wfed = 0
## wsta = 0
## wloc = 0
##
## Model 1: restricted model
## Model 2: crmrte ~ log(prbarr) + log(prbconv) + density + west + taxpc +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc
##
## Note: Coefficient covariance matrix supplied.
##
```

```
##   Res.Df Df      F Pr(>F)
## 1     84
## 2     74 10 0.9361 0.5058
```

From this comparison we can see that the F statistic calculated from our proposed and with financial variable models is less than our critical value. This means that our financial variables are not considered jointly significant and we cannot reject our null hypothesis. If we continue to dissect the financial variables, we will see the there is a positive correlation between federal wages and crime rate, but we assume that is due to the additional correlation with population density. This would include higher cost of living areas where wages are typically higher and federal employee compensation structure with housing adjustments. The scatter plots below indicate a positive relationship between federal wages and population density.

```
options(repr.plot.width=12, repr.plot.height=12)
scatterplotMatrix(~ wfed + crmrte + density, data = crimef, col = "blue",
                  diagonal=list(method ="histogram", breaks=25),
                  main = "Federal Wages vs. Crime Rate and Density",
                  plot.points = FALSE,
                  regLine = TRUE,
                  smooth = TRUE,
                  cex.labels = 1.4)
```



Although we initially believed there would be an obvious correlation between crime and viable alternatives, such as wages and tax revenue that could potentially be used for community support programs, it was not apparent in the provided cross-sectional data.

4.2 Adding All Criminal Justice System Variables

The next model adds all financial variables and unused criminal just variables (*prbpris*, *avgsen*, and *polpc*) to the demographic model.

```
md_cd_fin_pol <- lm(crmrte ~ log(prbarr) + log(prbconv) + density + west
                    + taxpc + wcon + wtuc + wtrd + wfir + wser + wmfg +
                    wfed + wsta + wloc + log(prbpris) + avgsen + polpc,
                    data = crimef)
```

As we can see from the t-tests, none of the additional variables are statistically significant and there is only a small increase in r-squared.

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

```
##
## t test of coefficients:
##
##                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  -1.5666e-02  2.8773e-02 -0.5445 0.587812
## log(prbarr)  -1.5160e-02  5.1786e-03 -2.9274 0.004589 **
## log(prbconv) -9.8351e-03  4.4937e-03 -2.1887 0.031913 *
## density       5.1729e-03  2.0383e-03  2.5378 0.013350 *
## west         -9.9640e-03  2.9198e-03 -3.4126 0.001066 **
## taxpc         1.9242e-04  2.9087e-04  0.6616 0.510398
## wcon         -1.0928e-06  3.4076e-05 -0.0321 0.974508
## wtuc          5.4792e-07  2.4536e-05  0.0223 0.982246
## wtrd         -1.7967e-05  9.4124e-05 -0.1909 0.849157
## wfir         -2.5697e-05  3.3889e-05 -0.7583 0.450797
## wser         -6.5616e-06  7.7942e-05 -0.0842 0.933145
## wmfg         -1.9143e-06  1.9814e-05 -0.0966 0.923307
## wfed          5.4448e-05  3.9620e-05  1.3743 0.173687
## wsta         -9.3638e-06  3.7035e-05 -0.2528 0.801124
## wloc          4.9234e-06  8.4437e-05  0.0583 0.953666
## log(prbpris) -7.5133e-04  7.0141e-03 -0.1071 0.914999
## avgsen       -5.6040e-04  5.9350e-04 -0.9442 0.348252
## polpc         3.9294e+00  2.7089e+00  1.4505 0.151311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## [1] 0.7592055
```

4.3 Adding All Demographic Variables

Finally, we'll include the remaining independent variables that are not already being used from the demographic data, such as *central*, *east*, *urban*, and *pctmin80*.

```
md_cd_fin_pol_dem <- lm(crmrte ~  log(prbarr) + log(prbconv) + density +
                        west + pctymle + taxpc + wcon + wtuc + wtrd +
                        wfir + wser + wmfg + wfed + wsta + wloc +
                        log(prbpris) + avgsen + polpc + central + urban +
                        pctmin80, data = crimef)
coeftest(md_cd_fin_pol_dem, vcov = vcovHC)
```

```
##
## t test of coefficients:
```

```
##
##                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  -5.1047e-02  2.8031e-02 -1.8211 0.073057 .
## log(prbarr)  -1.6263e-02  5.7159e-03 -2.8452 0.005882 **
## log(prbconv) -9.1610e-03  4.1044e-03 -2.2320 0.028965 *
## density       5.2769e-03  1.6835e-03  3.1345 0.002555 **
## west         -1.3874e-03  5.1257e-03 -0.2707 0.787477
## pctymle       1.0764e-01  5.6689e-02  1.8989 0.061891 .
## taxpc         2.4160e-04  3.3546e-04  0.7202 0.473911
## wcon          1.8121e-05  3.3135e-05  0.5469 0.586270
## wtuc          5.6926e-06  2.2419e-05  0.2539 0.800334
## wtrd          2.3517e-05  8.8681e-05  0.2652 0.791683
## wfir         -3.9490e-05  3.8677e-05 -1.0210 0.310921
## wser         -1.0031e-05  7.9549e-05 -0.1261 0.900035
## wmfg         -5.3052e-06  1.9379e-05 -0.2738 0.785110
## wfed          5.1741e-05  3.3981e-05  1.5226 0.132554
## wsta         -1.8083e-05  4.1301e-05 -0.4378 0.662917
## wloc          2.9913e-05  8.4327e-05  0.3547 0.723912
## log(prbpris) -2.2085e-03  7.0068e-03 -0.3152 0.753597
## avgsen       -6.4970e-04  5.1720e-04 -1.2562 0.213414
## polpc         4.0274e+00  2.3393e+00  1.7216 0.089760 .
## central      -3.2894e-03  4.0231e-03 -0.8176 0.416462
## urban        -7.8921e-04  8.2602e-03 -0.0955 0.924168
## pctmin80      3.3903e-04  1.4005e-04  2.4208 0.018203 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## [1] 0.8355125
```

In the this maximal model, *west* has lost statistical significance and the percentage of minorities, *pctmin80* has gained significance. We expect this is due to multicollinearity between *west* and *pctmin80* that was indicated in section 3.2.

The number of police per capita continues to be statistically significant to 90% confidence, with a positive coefficient. We do not believe hiring more police would result in more crime. We suspect that communities with higher crime rates are willing to support larger police forces.

A comparison of residuals plots for our demographic model from section 3.4 and the maximal models from this section show little improvement over the demographic model.

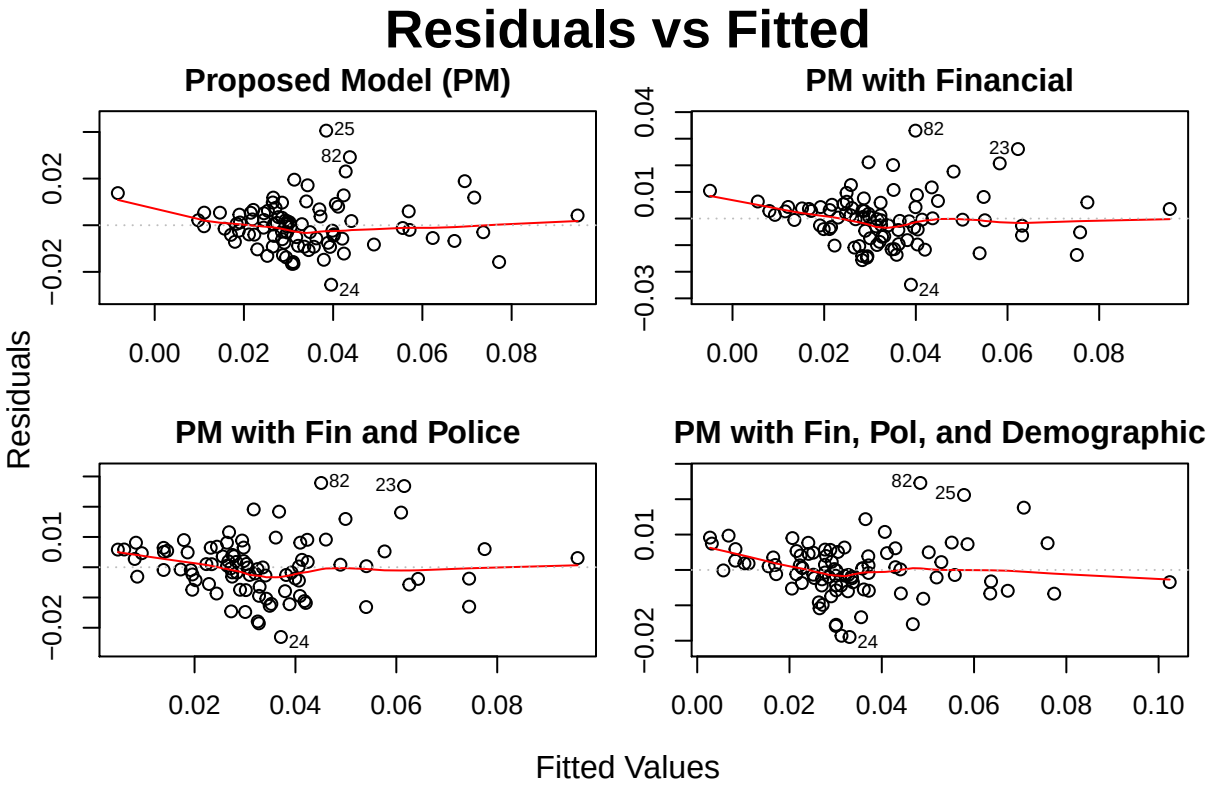
```
par(oma=c(2,2,3,0), mar=c(3,2,2,1), mfrow=c(2,2))

plot(md_cd, which=1, main="Proposed Model (PM)", caption="",
     sub.caption="")
plot(md_cd_fin, which=1, main="PM with Financial", caption="",
     sub.caption="")
plot(md_cd_fin_pol,  which=1, main="PM with Fin and Police",
     caption="", sub.caption="")
plot(md_cd_fin_pol_dem, which=1, caption="", sub.caption="",
     main="PM with Fin, Pol, and Demographic")

mtext(text="Fitted Values",side=1,line=0,outer=TRUE)
mtext(text="Residuals",side=2,line=0,outer=TRUE)

title(main="Residuals vs Fitted", line=0, outer=TRUE,
```

```
      cex.main=2)
```



The proposed model appears to be centered around zero with a small negative deviation and no apparent slope. With financials added to the model, the line appears generally unchanged but the line looks to deviate more than the original model. When the criminal justice system variables are included, the deviation increase even more. Finally, with the remaining demographic variables added to the model, the deviations continue to increase and there appears to be a negative slope. The bottom two models do eliminate the negative fitted value, but we concluded this advantage does not justify the additional model complexity.

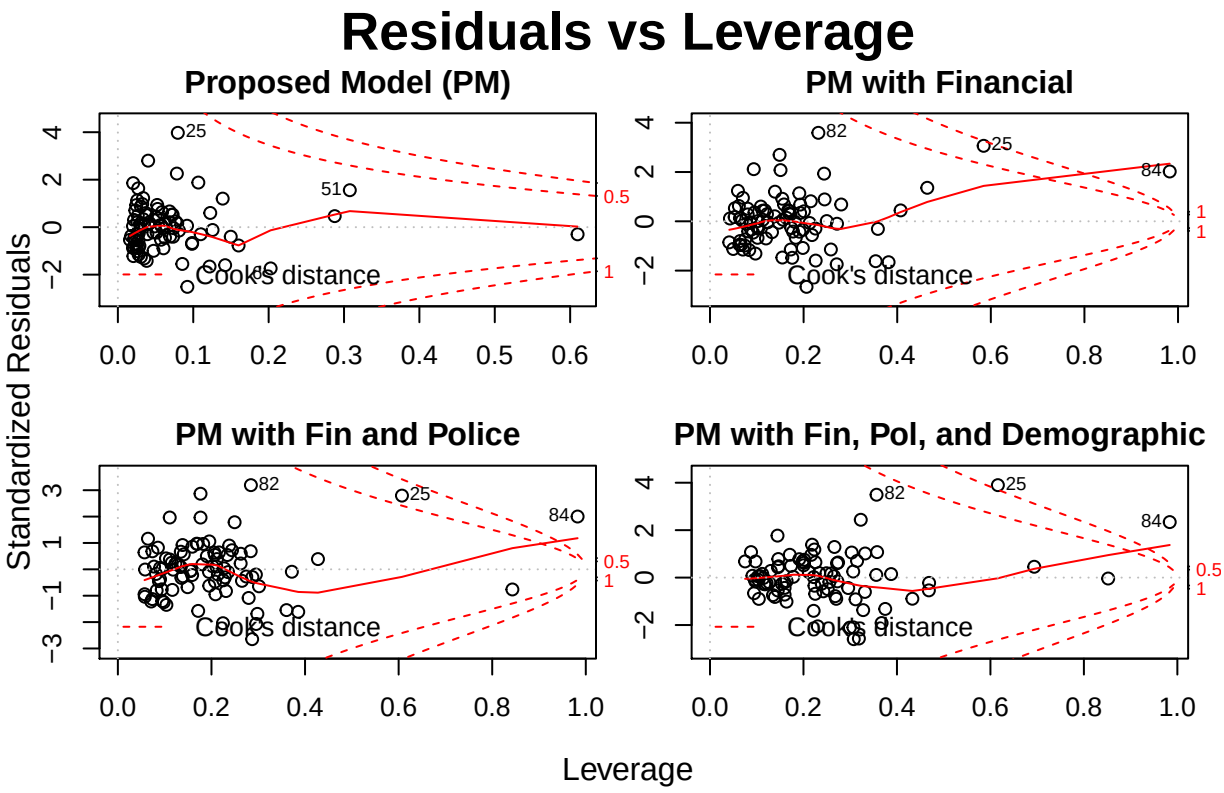
Leverage plots for the four models (see below) indicate that the proposed model performs the best with respect to minimizing the number of data points with high Cook's distance.

```
par(oma=c(2,2,3,0), mar=c(3,2,2,1), mfrow=c(2,2))

plot(md_cd, which=5, main="Proposed Model (PM)", caption="",
     sub.caption="")
plot(md_cd_fin, which=5, main="PM with Financial", caption="",
     sub.caption="")
plot(md_cd_fin_pol,  which=5, main="PM with Fin and Police",
     caption="", sub.caption="")
plot(md_cd_fin_pol_dem, which=5, caption="", sub.caption="",
    main="PM with Fin, Pol, and Demographic")

mtext(text="Leverage",side=1,line=0,outer=TRUE)
mtext(text="Standardized Residuals",side=2,line=0,outer=TRUE)

title(main="Residuals vs Leverage", line=0, outer=TRUE, cex.main=2)
```



5. Regression Table for Summarized Model Comparison

This section will compare the base (arrest and conviction) model from section 2, the proposed model (demographic) from sectin 3, and the expanded model with all criminal, financial, and demographic variables from section 4.

```
robust_se.md_cd <- coeftest(md_cd, vcov = vcovHC)
robust_se.md_cd_fin <- coeftest(md_cd_fin, vcov = vcovHC)
robust_se.md_cd_fin_pol <- coeftest(md_cd_fin_pol, vcov = vcovHC)
robust_se.md_cd_fin_pol_dem <- coeftest(md_cd_fin_pol_dem,
                                        vcov = vcovHC)
stargazer(lemod_arr_conv, md_cd, md_cd_fin_pol_dem,
          se = list(tcoe3[, "Std. Error"], robust_se.md_cd[, "Std. Error"],
                    robust_se.md_cd_fin_pol_dem[, "Std. Error"]),
          add.lines = list(c("AIC", toString(AIC(lemod_arr_conv)),
                             toString(AIC(md_cd)), toString(AIC(md_cd_fin_pol_dem)))),
          omit.stat = c("f", "adj.rsq", "ser"),
          column.labels = c("Base Model", "Proposed Model", "Expanded Model"),
          star.cutoffs = c(0.05, 0.01, 0.001),
          type=stargazer_type)
```

```
##
## ===============================================================
##                                 Dependent variable:
##              -----------------------------------------------
##                                        crmrte
```

```
##                  Base Model      Proposed Model    Expanded Model
##                      (1)              (2)              (3)
## ------------------------------------------------------------------
## log(prbarr)      -0.024***         -0.011*          -0.016**
##                   (0.006)          (0.005)           (0.006)
##
## log(prbconv)     -0.016***         -0.009*          -0.009*
##                   (0.005)          (0.004)           (0.004)
##
## density                           0.007***          0.005**
##                                   (0.001)           (0.002)
##
## west                              -0.011***         -0.001
##                                   (0.002)           (0.005)
##
## pctymle                            0.089             0.108
##                                   (0.047)           (0.057)
##
## taxpc                                               0.0002
##                                                    (0.0003)
##
## wcon                                                0.00002
##                                                    (0.00003)
##
## wtuc                                                0.00001
##                                                    (0.00002)
##
## wtrd                                                0.00002
##                                                    (0.0001)
##
## wfir                                               -0.00004
##                                                    (0.00004)
##
## wser                                               -0.00001
##                                                    (0.0001)
##
## wmfg                                               -0.00001
##                                                    (0.00002)
##
## wfed                                                0.0001
##                                                    (0.00003)
##
## wsta                                               -0.00002
##                                                    (0.00004)
##
## wloc                                                0.00003
##                                                    (0.0001)
##
## log(prbpris)                                        -0.002
##                                                     (0.007)
##
## avgsen                                              -0.001
##                                                     (0.001)
##
```

```
## polpc                                                  4.027
##                                                       (2.339)
##
## central                                               -0.003
##                                                       (0.004)
##
## urban                                                 -0.001
##                                                       (0.008)
##
## pctmin80                                              0.0003*
##                                                      (0.0001)
##
## Constant          -0.010           -0.002            -0.051
##                   (0.009)          (0.006)           (0.028)
##
## ------------------------------------------------------------------
## AIC          -496.069980761364 -548.293352186705 -569.685460940159
## Observations        90                89                89
## R2                 0.387             0.700             0.836
## ==================================================================
## Note:                                *p<0.05; **p<0.01; ***p<0.001
```

A few coefficients are consistent across multiple models. The coefficients for probability of arrest and conviction are significant and have similar practical significance in all three models. Population density is statistically significant in both the proposed and expanded model. As noted earlier, the *west* indicator variable loses significance in the expanded model, most likely due to collinearity with the percentage of minorities. Our decision to use the *west* variable in our model is based on our suspicion that western counties have cultural differences that go beyond differences in the percentage of minorities when compared to other counties in NC. However the expanded model suggest that replacing *west* with the percentage of minorities may also result in a suitable model.

The expanded model has the best Akike information criterion (AIC) score, which evaluates the model fit while incorporating parsimony. However the improved AIC score comes at the cost of adding sixteen variables. In spite of the expanded model's improved AIC score, we prefer the simpler proposed model for identifying suitable policy initiatives.

6. Omitted Variables

The direction of omitted variable bias can estimated with the equation $\text{Bias}(\beta_i) = \beta_{ov}\delta_{ov-i}$ where β_i is the coefficient of the variable that is included in the model, β_{ov} is the coefficient of the omitted variable, and $\delta_{ov,i}$ is the covariance of the included and omitted variables, x_ia nd x_{ov}.

1. Local unemployment rate

We expect that the local unemployment rate and crime rates would have a positive correlation, resulting in a positive value of $\beta_{unemploy}$. We also expect that unemployment would have a small negative correlation with the probability of conviction. Higher unemployment would result in less tax revenue and possibly cutbacks on the resources of local governments. This results in a small negative bias on the coefficient of probability of conviction, resulting in this coefficient becoming more negative.

2. High School Graduation Rate

We predict that high school graduation rates will be negatively correlated with crime rates, due to criminal activity having a greater opportunity costs for persons with higher levels of education. We also expect that high school graduation rates will be positively correlated with probabilities of arrest and conviction, due to higher graduation rates being correlated with greater economic activity and availability of resources for local

law enforcement agencies. * $\beta_{gradrate} < 0$ * $\delta_{grad-prb} > 0$ * The bias to the coefficients for *log(prbarr)* and *log(prbconv)* is negative, away from zero. We suspect the correlation between graduation rate and the arrest and conviction probabilities is low, resulting in a small effect.

3. County Economic Activity

We expect that a county's economic activity, determined through a variable similar to GDP, is correlated with crime rate.

$$crmrte = \beta_0 + \beta_1 * prbarr + \beta_2 * economy + u$$
$$economy = \alpha_0 + \alpha_1 * prbarr + u$$

If $\beta_2 < 0$ and $\alpha_1 > 0$ then $OMVB = \beta_2\alpha_1 < 0$ and if $\beta_1 < 0$ then the OLS coefficient on *prbarr* will be scaled away from zero (more negative) gaining statistical significance.

4. Measure of Social Services

We expect that a county's strength of social services, is correlated with crime rate.

$$crmrte = \beta_0 + \beta_1 * prbarr + \beta_2 * socialservices + u$$
$$socialservices = \alpha_0 + \alpha_1 * prbarr + u$$

Similar to high school graduation rates, we expect that availability of social services is positively correlated with probabilities of arrest and conviction because it may be a proxy for availability of resources. We expect the positive correlation to be small, therefore we expect the overall bias to be small, but negative and away from zero.

5. Percent of Population with Criminal Record

We expect that a county's percentage of the population with a history of criminal activity is correlated with the crime rate.

$$crmrte = \beta_0 + \beta_1 * prbarr + \beta_2 * pctcrimrec + u$$
$$pctcrimrec = \alpha_0 + \alpha_1 * prbarr + u$$

If $\beta_2 > 0$ and $\alpha_1 > 0$ then $OMVB = \beta_2\alpha_1 > 0$ and if $\beta_1 < 0$ then the OLS coefficient on *prbarr* will be scaled toward zero (less negative) losing statistical significance.

Omitted Var	Omitted Coef. β_{ov}	Included Var	Cov $\delta_{ov,i}$	Bias Dir.	Bias Size
Local Unemployment rate	Positive	Prob. Conviction	Negative	Negative	Small
HS Grad. Rate	Negative	Prob. Conviction	Positive	Negative	Small
Economic Activity	Negative	Prob. Conviction	Negative	Negative	Small
Social Services	Negative	Prob. Conviction	Positive	Negative	Small
Pct. Criminal Record	Positive	Prob. Conviction	Positive	Positive	Small

7. Conclusion

Our optimal linear model suggests that local governments can most effectively reduce crime by increasing the likelihood that an offender will be arrested, and by increasing the likelihood that once arrested, an offender will be convicted. We did not conduct analysis to determine whether higher probabilities for arrest and conviction cause lower crime rates or if the variables are merely correlated. Nevertheless, we believe the relationship demonstrated by this analysis is sufficient for consideration with respect to policy proposals for reducing crime.

One could argue that our results are obvious, that the general public is aware that arresting and convicting offenders will lower crime rates. But consider what this analysis did not find. We found little relationship between crime rates, average sentences, and the probability that an offender would be sent to prison. While long prison terms will likely increase both the prison population and our tax burden, our analysis suggests that they will not lower crime rates.

Furthermore, we did not find a useful relationship between wages and crime rates. While policy proposals to increase wages may be worthwhile, our analysis on the provided cross-sectional data provides no basis for claiming that increasing wages will reduce crime. Instead we found that higher wages were correlated with higher crime rates. The general public is unlikely to support programs that aim to reduce crime by lowering wages. We do not believe that higher wages cause crime. It is more likely that the positive correlation between wages and crime is occurring because many categories of wages are positively correlated with population density, which is positively correlated with crime rates. Local, state, and federal wages had the highest correlation with population density, possibly due to cost of living adjustments, and possibly because higher-level managers with higher salaries are likely to work in urban areas.

Another valid critique of this analysis is that it provides no guidance on how to increase arrests or convictions. There is a moderate positive correlation between police per capita and the probability of arrest, suggesting that law enforcement agencies must have adequate resources to be effective at reducing crime. With respect to convictions, there is little correlation between the probability of conviction and anything else in the data set. Surprisingly, there is nearly zero correlation between the probability of conviction and the probabilities for arrest or receiving a prison sentence. Therefore, in order to identify practices that promote arrests and convictions, we recommend conducting a comparative study of law enforcement practices in counties with both high and low probabilities of arrest and conviction.

We are especially interested in the probability of conviction. Law enforcement agencies have the ability to arrest individuals without support or concurrence from other organizations. But law enforcement agencies cannot convict an offender. A conviction represents effective cooperation between law enforcement, other parts of government such as prosecutors and judges, and for jury trials, the community. Consequently, we consider the probability of conviction to be a better measure of the professionalism of law enforcement agencies than the probability of arrest. It is also possible that the probability of conviction is a proxy measurement for the overall effectiveness of local governments and for the relationship between the law enforcement agencies and their local communities. If this is correct, focusing on the overall health of local governments may have greater impact on the probability of conviction than focusing on evidence or judicial proceedings.

In summary, we recommend that policy proposals for reducing crime focus on adequately resourcing law enforcement agencies, promoting professional practices within those agencies, and promoting product relationships between law enforcement agencies, other branches of local government, and their communities. Our analysis does not indicate that higher wages or longer prison sentences will reduce crime.