Lab 3: Reducing Crime

Stage 3: Final Report. W203, Spring 2018

Cameron Kennedy and Ehsan Yousefzadeh

1. Introduction

This analysis explores a 1987 dataset on crime and several demographics potentially related to crime from counties in North Carolina. It seeks to understand predictive, and where possible, causal relationships that drive crime rates. Motivating this analysis is a political campaign that has requested research on the determinants of crime for the purpose of making policy recommendations for local governments.

Although the request for this research came from a political campaign, the research team is independent of any political organization or belief system and did not approach this analysis seeking to support or reject any specific claims. The findings in this report are objective, based solely on the data provided, and are given to the political campaign to use as they see fit.

2. Initial Hypotheses and Exploratory Data Analysis

Initial Hypotheses

Having only reviewed the variable names and their descriptions, the research team formulated initial hypotheses prior to exploring the data. These hypotheses all address the primary research question, which is understanding the drivers of crime rates, and are as follows:

- When individuals contemplate committing crimes, it seems reasonable the likelihood and severity of punishment will factor into their decision. Therefore, we hypothesize that higher numbers of police, higher probability of arrest, higher probability of conviction, and higher average sentences will all drive down crime rates.
- High density areas are often associated with inner-city areas that can be more neglected and impoverished, which, in turn, can increase crime. Therefore, we hypothesize that areas with higher population density and those designated as 'urban' will have higher crime rates. We also expect these two variables to be strongly correlated.
- Another factor people may weigh when considering criminal activity is the extent to which their basic needs (e.g., food, shelter) are met. People with low income may have a greater incentive to commit crimes motivated by desperation to meet these basic needs. Therefore, poverty may contribute to crime, and while poverty is not directly included in the data set, we hypothesize that wage rates and tax revenue may serve as (inversely related) proxy variables to poverty, and that as they increase, crime will decrease.
- Finally, common knowledge tells us that a disproportionate amount of crime is committed by young males (relative to the whole population). Therefore, we hypothesize that counties with higher percentages of young male citizens will have higher crime rates.

The analysis in this report tests these hypotheses, but as we perform our exploratory data analysis, we are open to discovering additional findings that may cause us to revise our hypothesis and modeling choices.

Dataset Overview

The data set contains a total of 91 observations and 25 variables. Each observation represents the data for a specific county in North Carolina. Table 1 refers to the full data summary with the variable name, description, range of values, and the type of variable. We found the data set to be relatively clean with no NAs in the main data. At the end of the data set we noticed a stray mark a few rows after the main data ended. To fix this error, we deleted rows where the county was NA (there was no data in these rows other than the stray mark). Next, we had to convert the prbconv data set to a numerical value. Finally, we decided to remove the year variable since all the observations are in the year 1987.

The code for the initial library loading and data wrangling is listed here:

```
#Clear existing data
rm(list = ls())

#Load Libraries
library(data.table)
library(ggplot2)
library(reshape2)  #Used for ggplot multiple histograms
library(tidyr)
library(stargazer)
library(knitr)
library(kableExtra)
library(lmtest)
library(sandwich)
library(car)

#Open and wrangle data
dt <- fread('crime_v2.csv', na.strings=c("NA","N/A","null",""))  #Read in data
dt <- dt[!is.na(county)]  #Stray mark at end of data; fixing by deleting rows where county = NA
dt$prbconv <- as.numeric(dt$prbconv)  #Convert prbconv to numeric
dt$county <- as.factor(dt$county)  #Convert county a factor
dt[, year:=NULL] #Delete year, since it's all the same value
dt <- dt[!duplicated(dt)]  #Remove duplicate row (county 193)
```

Univariate Analysis

The univariate analysis reviews each variable in the dataset individually To begin, the table below lists each of our variables, along with its description (as provided), range of values, and data type:

```
Variables = c("crmrte", "prbarr", "prbconv", "prbpris", "avgsen", "polpc", "density", "taxpc",
              "west","central","urban","pctmin80","wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg",
              "wfed", "wsta", "wloc", "mix", "pctymle")
Description = c("crimes committed per person", "probability of arrest", "probability of conviction",
                "probability of prison sentence", "average sentence in days", "police per capita",
                "people per sq mile", "tax revenue per capita", "=1 if western NC", "=1 if central NC",
                "=1 if urban", "Percent minority in 1980", "Weekly wage, construction",
                "Weekly wage, trns, util, commun", "Weekly wage wholesale, retail",
                "Weekly wage, finance, ins, real estate", "Weekly wage, service industry",
                "Weekly wage, manufacturing", "Weekly wage, federal employees",
                "Weekly wage, state employees", "Weekly wage, local government employees",
                "Offense mix: face-to-face, other", "Percent young male")
Range = c("0.005533 to 0.098966", "0.09277 to 1.09091", "0.06838 to 2.12121", "0.1500 to 0.6000",
          "5.380 to 20.700","0.0007459 to 0.0090543","0.00002 to 8.82765","25.69 to 119.76","0 or 1",
```

```
          "0 or 1","0 or 1","1.284 to 64.348","193.6 to 436.8","187.6 to 613.2","154.2 to 354.7",
          "170.9 to 509.5", "133.0 to 2177.1", "157.4 to 646.9", "326.1 to 598.0", "258.3 to 499.6",
          "239.2 to 388.1", "0.01961 to 0.46512", "0.06216 to 0.24871")
Type = c("continuous","continuous","continuous","continuous","continuous","continuous","continuous",
         "continuous","discrete","discrete","discrete","continuous","continuous","continuous",
         "continuous","continuous","continuous","continuous","continuous","continuous","continuous",
         "continuous","continuous")
table = data.frame(Variables, Description, Range, Type)
kable(table, format = "latex", booktabs = T,
      caption = "Variable Descriptions, Range and Type") %>%
  kable_styling(latex_options = c("striped","hold_position"))
```

Table 1: Variable Descriptions, Range and Type

Variables	Description	Range	Type
crmrte	crimes committed per person	0.005533 to 0.098966	continuous
prbarr	probability of arrest	0.09277 to 1.09091	continuous
prbconv	probability of conviction	0.06838 to 2.12121	continuous
prbpris	probability of prison sentence	0.1500 to 0.6000	continuous
avgsen	average sentence in days	5.380 to 20.700	continuous
polpc	police per capita	0.0007459 to 0.0090543	continuous
density	people per sq mile	0.00002 to 8.82765	continuous
taxpc	tax revenue per capita	25.69 to 119.76	continuous
west	=1 if western NC	0 or 1	discrete
central	=1 if central NC	0 or 1	discrete
urban	=1 if urban	0 or 1	discrete
pctmin80	Percent minority in 1980	1.284 to 64.348	continuous
wcon	Weekly wage, construction	193.6 to 436.8	continuous
wtuc	Weekly wage, trns, util, commun	187.6 to 613.2	continuous
wtrd	Weekly wage wholesale, retail	154.2 to 354.7	continuous
wfir	Weekly wage, finance, ins, real estate	170.9 to 509.5	continuous
wser	Weekly wage, service industry	133.0 to 2177.1	continuous
wmfg	Weekly wage, manufacturing	157.4 to 646.9	continuous
wfed	Weekly wage, federal employees	326.1 to 598.0	continuous
wsta	Weekly wage, state employees	258.3 to 499.6	continuous
wloc	Weekly wage, local government employees	239.2 to 388.1	continuous
mix	Offense mix: face-to-face, other	0.01961 to 0.46512	continuous
pctymle	Percent young male	0.06216 to 0.24871	continuous

The table above requires no specific analysis, but is used for reference throughout the report.

Next, we provide a table of each variable and its summary statistics:

```
disp_summary_table <- function(table_obj) {
  kable(summary(table_obj), format = "latex", booktabs = T, caption = "Data Summary") %>%
  kable_styling(latex_options =c("striped", "hold_position", "scale_down"))}
disp_summary_table(dt[,2:9])
```

```
disp_summary_table(dt[,10:17])
```

```
disp_summary_table(dt[,18:24])
```

Table 2: Data Summary

crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc
Min. :0.005533	Min. :0.09277	Min. :0.06838	Min. :0.1500	Min. : 5.380	Min. :0.0007459	Min. :0.00002	Min. : 25.69
1st Qu.:0.020604	1st Qu.:0.20495	1st Qu.:0.34422	1st Qu.:0.3642	1st Qu.: 7.375	1st Qu.:0.0012378	1st Qu.:0.54718	1st Qu.: 30.73
Median :0.030002	Median :0.27146	Median :0.45170	Median :0.4222	Median : 9.110	Median :0.0014897	Median :0.97925	Median : 34.92
Mean :0.033510	Mean :0.29524	Mean :0.55086	Mean :0.4106	Mean : 9.689	Mean :0.0017080	Mean :1.43567	Mean : 38.16
3rd Qu.:0.040249	3rd Qu.:0.34487	3rd Qu.:0.58513	3rd Qu.:0.4576	3rd Qu.:11.465	3rd Qu.:0.0018856	3rd Qu.:1.56926	3rd Qu.: 41.01
Max. :0.098966	Max. :1.09091	Max. :2.12121	Max. :0.6000	Max. :20.700	Max. :0.0090543	Max. :8.82765	Max. :119.76

Table 3: Data Summary

west	central	urban	pctmin80	wcon	wtuc	wtrd	wfir
Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. : 1.284	Min. :193.6	Min. :187.6	Min. :154.2	Min. :170.9
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:10.024	1st Qu.:250.8	1st Qu.:374.3	1st Qu.:190.7	1st Qu.:285.6
Median :0.0000	Median :0.0000	Median :0.00000	Median :24.852	Median :281.2	Median :404.8	Median :203.0	Median :317.1
Mean :0.2444	Mean :0.3778	Mean :0.08889	Mean :25.713	Mean :285.4	Mean :410.9	Mean :210.9	Mean :321.6
3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:0.00000	3rd Qu.:38.183	3rd Qu.:315.0	3rd Qu.:440.7	3rd Qu.:224.3	3rd Qu.:342.6
Max. :1.0000	Max. :1.0000	Max. :1.00000	Max. :64.348	Max. :436.8	Max. :613.2	Max. :354.7	Max. :509.5

Table 4: Data Summary

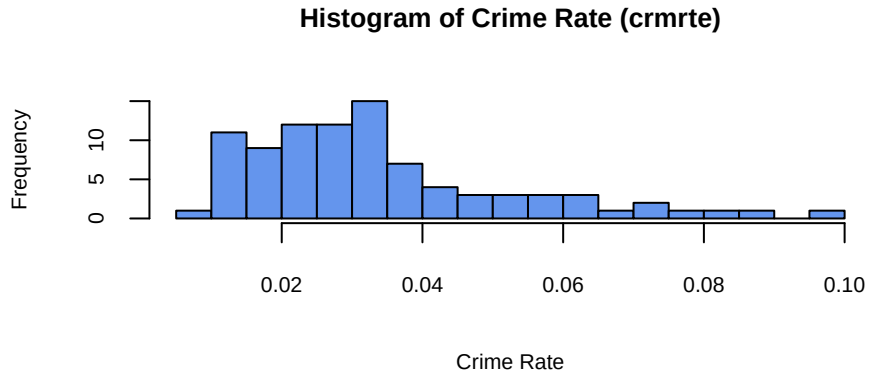
wser	wmfg	wfed	wsta	wloc	mix	pctymle
Min. : 133.0	Min. :157.4	Min. :326.1	Min. :258.3	Min. :239.2	Min. :0.01961	Min. :0.06216
1st Qu.: 229.3	1st Qu.:288.6	1st Qu.:398.8	1st Qu.:329.3	1st Qu.:297.2	1st Qu.:0.08060	1st Qu.:0.07437
Median : 253.1	Median :321.1	Median :448.9	Median :358.4	Median :307.6	Median :0.10095	Median :0.07770
Mean : 275.3	Mean :336.0	Mean :442.6	Mean :357.7	Mean :312.3	Mean :0.12905	Mean :0.08403
3rd Qu.: 277.6	3rd Qu.:359.9	3rd Qu.:478.3	3rd Qu.:383.2	3rd Qu.:328.8	3rd Qu.:0.15206	3rd Qu.:0.08352
Max. :2177.1	Max. :646.9	Max. :598.0	Max. :499.6	Max. :388.1	Max. :0.46512	Max. :0.24871

From the table above, rather than commenting on each variable here, we use the data throughout the report, largely in the remainder of the univariate analysis section.

Dependent Variable (Crime Rate)

The dependent variable in our analysis is the crime rate (crmrte) which is a continuous variable with a range of 0.0055 to 0.0989 crimes committed per person. The distribution of the crime rate variable appears to be positively skewed, with a mean of 0.033 and a median of 0.030. No clear outliers are shown in the crime rate. However, there appear to be a handful of higher crime rate values on the right tail.

```
plot_hist <- function(var_obj, title_name, xlab_name) {
  hist(var_obj, col = "cornflowerblue", breaks = 20,
       main = title_name, cex.main=0.8, cex.lab=0.7,
       yaxt = "n", xaxt = "n", xlab = xlab_name)
  axis(2, cex.axis = 0.7)
  axis(1, cex.axis = 0.7)
}
plot_hist(dt$crmrte, "Histogram of Crime Rate (crmrte)", "Crime Rate")
```

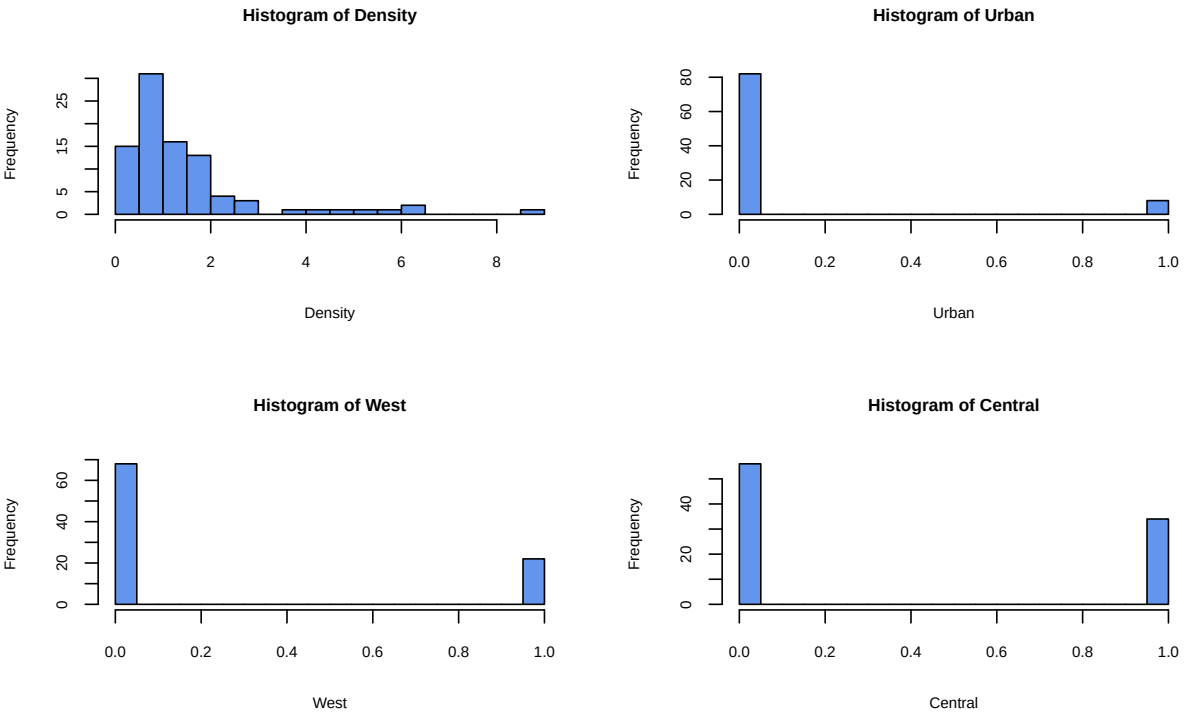


Input Variables

The independent variables have been grouped into several categories: geographic variables, judicial process and policing variables, wage variables, demographic variables, and government variables.

Geographic Variables

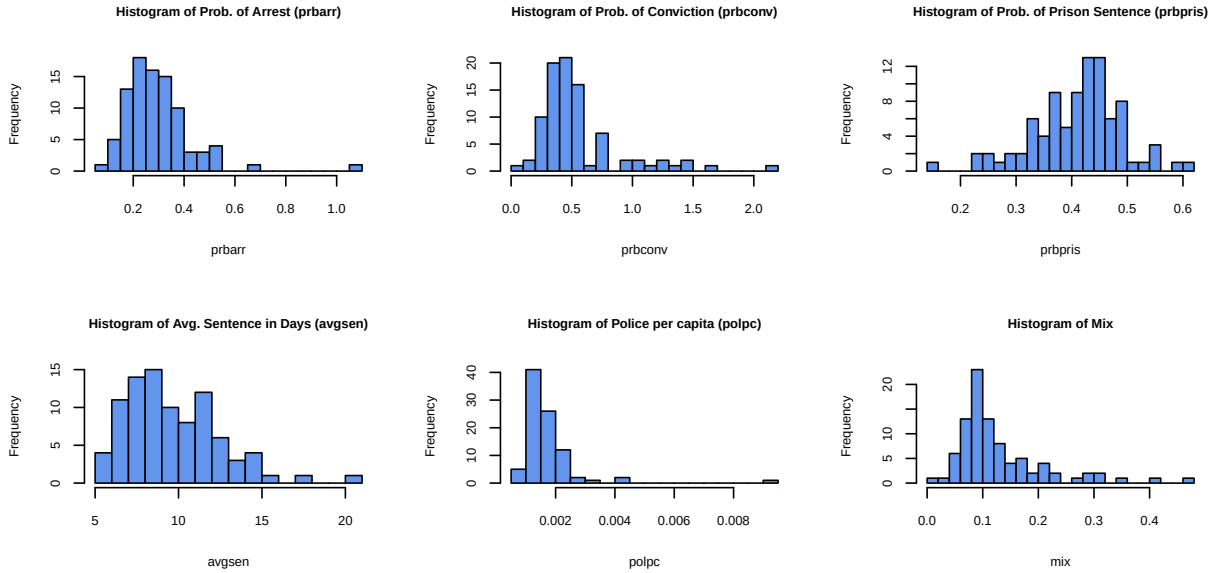
```
par(mfrow=c(2,2))
plot_hist(dt$density, "Histogram of Density", "Density")
plot_hist(dt$urban, "Histogram of Urban", "Urban")
plot_hist(dt$west, "Histogram of West", "West")
plot_hist(dt$central, "Histogram of Central", "Central")
```



The geographic variables are density, urban, west, and central. Density is a continuous variable that is in terms of people per square mile. It has a range of 0.0002 and 8.8277. The density variable is not normally distributed with the majority of values clustered towards the lower bound. The density variable appears to contain an outlier in the upper bound with a value of 8.83, but the team does not consider this value implausible enough warrant deletion. Urban, west, and central are all discrete variables with binary outcomes with "1" meaning positive and "0" meaning negative. The urban variable has only a small amount of positive values with a mean of 0.088. The west variable has more positive values compared to urban, with a mean of 0.25. The central variable has an even greater amount positive values with a mean of 0.37.

Judicial Processes and Policing Variables

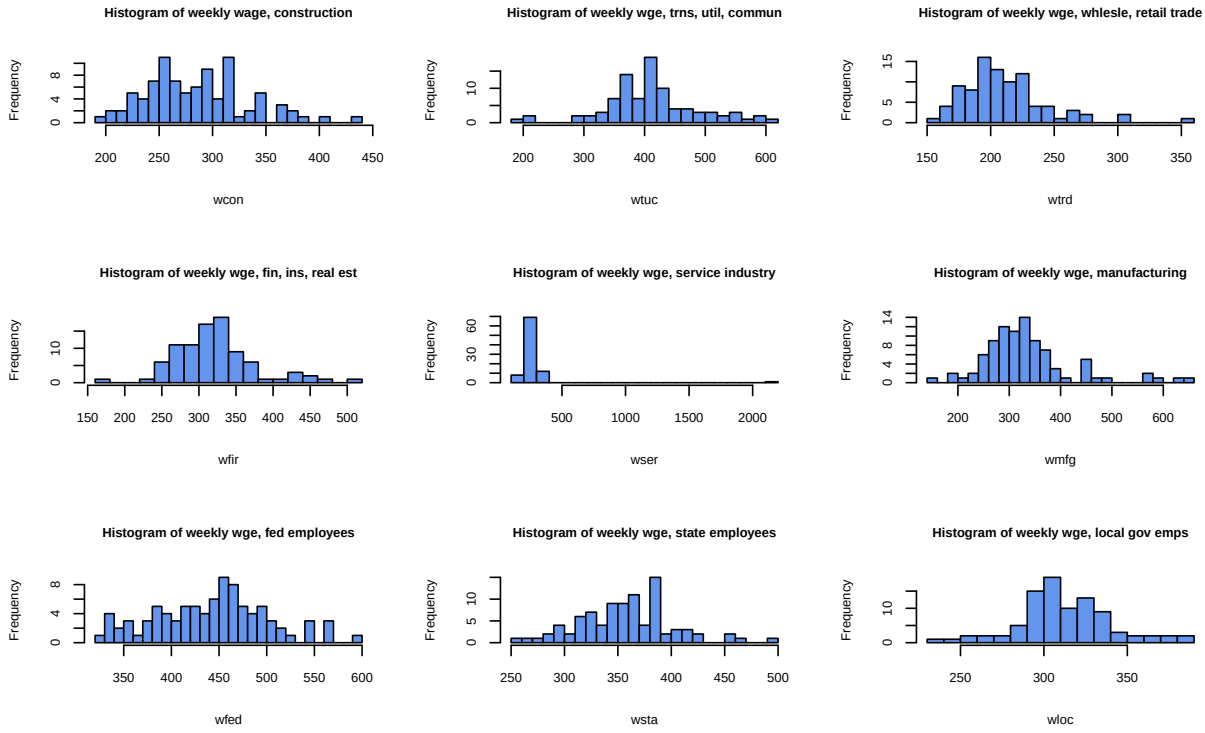
```
par(mfrow=c(2,3))
plot_hist(dt$prbarr, "Histogram of Prob. of Arrest (prbarr)", "prbarr")
plot_hist(dt$prbconv, "Histogram of Prob. of Conviction (prbconv)", "prbconv")
plot_hist(dt$prbpris, "Histogram of Prob. of Prison Sentence (prbpris)", "prbpris")
plot_hist(dt$avgsen, "Histogram of Avg. Sentence in Days (avgsen)", "avgsen")
plot_hist(dt$polpc, "Histogram of Police per capita (polpc)", "polpc")
plot_hist(dt$mix, "Histogram of Mix", "mix")
```



The judicial processes and policing variables are probability of arrest, probability of conviction, probability of sentence, average sentence (days), police per capita, and offense mix. The probability of arrest variable is continuous and the distribution appears to be positively skewed with a mean of 0.29 and a median of 0.27. The probability of arrest variable has an outlier value of 1.09 which does not make sense since probability cannot exceed a value of 1. The probability of conviction is a continuous variable and the distribution does not appear to be normally distributed with the majority of values clustered towards the lower bound. The probability of conviction variable has several values above 1 which makes us put into the question the validity of this variable as probability should have an upper limit of 1. The probability of prison sentence is a continuous variable with a distribution that appears to be normally distributed with a mean of 0.41 and a median of 0.42. The probability of prison sentence has a range of 0.15 to 0.60. The police per capita variable is a continuous variable that does not appear to be normally distributed as the majority of values are towards the lower bound. The police per capita variable appears to have an outlier value 0.009 on the upper limit. This can be considered an outlier as the mean is 0.0017 and the 3rd quartile is 0.0018. Thus, a value of 0.009 appears to be extraordinary. The mix of offense variable is continuous and does not appear to be normally distributed as the majority of values are clustered towards the lower bounds.

Wage Variables

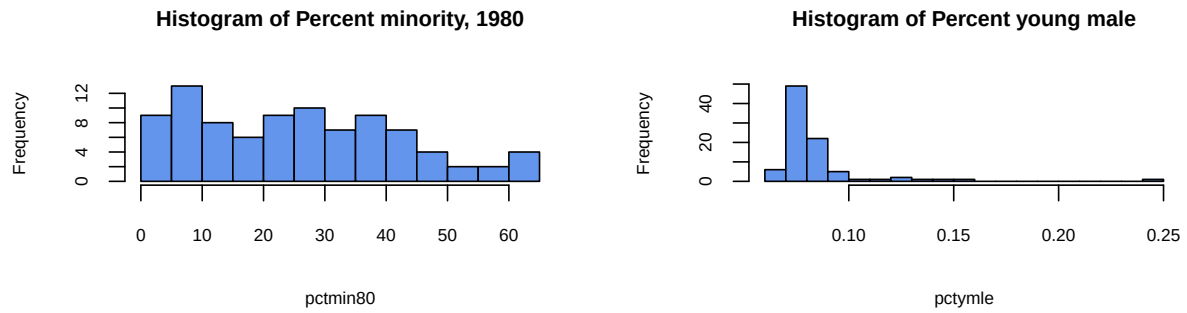
```
par(mfrow=c(3,3))
plot_hist(dt$wcon, "Histogram of weekly wage, construction", "wcon")
plot_hist(dt$wtuc, "Histogram of weekly wge, trns, util, commun", "wtuc")
plot_hist(dt$wtrd, "Histogram of weekly wge, whlesle, retail trade", "wtrd")
plot_hist(dt$wfir, "Histogram of weekly wge, fin, ins, real est", "wfir")
plot_hist(dt$wser, "Histogram of weekly wge, service industry", "wser")
plot_hist(dt$wmfg, "Histogram of weekly wge, manufacturing", "wmfg")
plot_hist(dt$wfed, "Histogram of weekly wge, fed employees", "wfed")
plot_hist(dt$wsta, "Histogram of weekly wge, state employees", "wsta")
plot_hist(dt$wloc, "Histogram of weekly wge, local gov emps", "wloc")
```



The wage variables are weekly wages for construction, trns/util/commun, fin/ins/real estate, service industry, manufacturing, federal employees, state employees, and local government employees. All these variables are continuous variables. The wage construction (wcon) variable appears to be normally distributed with a mean of 285 and a median of 281. The weekly wage of trns, util, and commun (wtuc) also appears to be normal with a mean of 412 and median of 406. The wage of fin, ins, and real estate (wfir) also appears to be normally distributed with a mean of 322 and a median of 317. The weekly wage of the service (wser) does not appear to be normally distributed. The wser variable also appears to contain an outlier with a value of 2177 which is far greater than the mean of 276. The wage of manufacturing (wmfg) appears to be somewhat normally distributed with a mean of 335 and median of 320. The wmfg also appears to contain several possible outlier values with a maximum value of 647. The wage of federal employees (wfed) appears to be normally distributed with a mean of 443 and a median of 450. The wage of state employees (wsta) also appears to be normally distributed with both a mean and median of 358. The wage of local government employees (wloc) appears to be normally distributed with a mean of 313 and a median of 308.

Demographic Variables

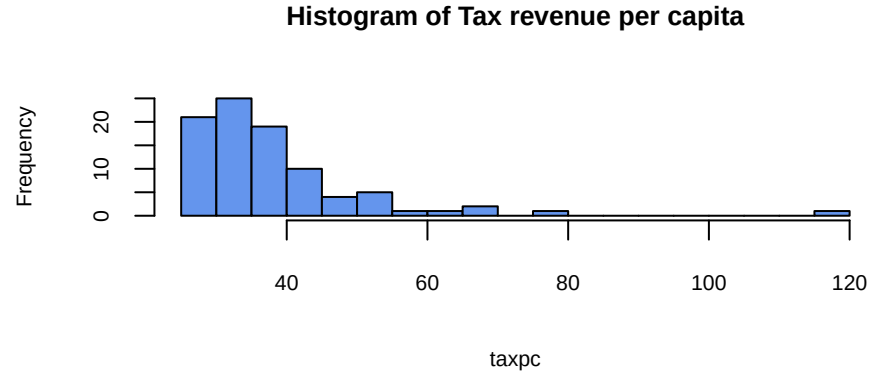
```
par(mfrow=c(1,2))
plot_hist(dt$pctmin80, "Histogram of Percent minority, 1980", "pctmin80")
plot_hist(dt$pctymle, "Histogram of Percent young male", "pctymle")
```



The demographic variables are percent minority in 1980 (pctmin80) and the percent young male (pctymle). Percent minority variable does not appear to be normally distributed as the majority of the values are clustered towards the lower bound. The percent young male also does not appear to be normally distributed and there appears to be an outlier in the upper bound. The percent young male variable has a mean of 0.084 and the outlier value is 0.249.

Government Variable

```
plot_hist(dt$taxpc, "Histogram of Tax revenue per capita", "taxpc")
```



The final category of variables is the government variable which only contains the tax revenue per capita variable (taxpc). The tax per capita variable does not appear to be normally distributed, it has a mean of 38 and a median of 34.9. There appears to be an outlier value of 120 on the upper bound.

Data Transformation

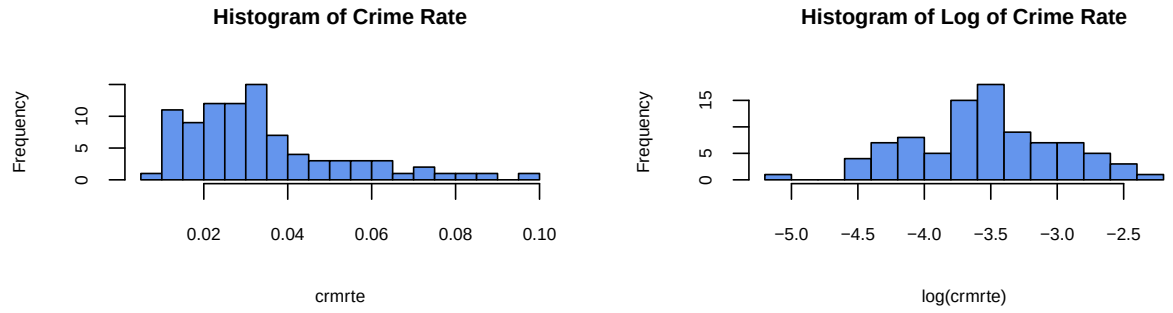
Natural nonlinearities in the data can obscure an ordinary least squares model, therefore, the team considered several data transformations.

Dependent Variable Transformation

The team first considered transforming crime rate (crmrte), the dependent variable, which has a noticeable right skew. By itself, the skew wasn't strong enough to present a compelling reason to transform this variable. However, upon building the linear regression models and testing them against standard assumptions, the team was able to meet some previously violated assumptions by taking a log transformation of the crime rate variable. Furthermore, the non-transformed units of crime are crimes committed per person, and since it ranges from 0.0055 to 0.0990, the team considers these original values less interpretable than the log-transformed values of crime rate, which, in a linear model, will provide increases or decreases to crime rate on a percentage basis (assuming coefficients are $< 20\%$). Therefore, all the linear regression models use the log of crime rate. In the exploratory data analysis, the non-transformed version of crime rate is used; however, the team also created scatter plots of all potential input variables vs. the log of crime rate which can be found in the Appendix.

The plot below shows side by side histograms of the crime rate variable before and after its log transformation:

```
par(mfrow=c(1,2))
plot_hist(dt$crmrte, "Histogram of Crime Rate", "crmrte")
plot_hist(log(dt$crmrte), "Histogram of Log of Crime Rate", "log(crmrte)")
```



We see the log transformation makes the crime rate variable considerably more normally distributed.

Independent Variable Transformation.

Given the skewness in their distributions, the team gave serious consideration to transforming the following independent variables: prbconv, avgsen, polpc, density, taxpc, mix, and pctymle, all of which had a noticeable right skew. Despite their skew, the team ultimately found that transforming these variables did not help improve any modeling assumptions, and they also either did nothing to improve interpretability, or in the cases of variables measured in percentage terms (e.g., prbconv) we felt they actually hurt interpretability because a percent increase of a variable that's already measured in percentages could be confusing. Therefore, because the team found no value in transforming these variables, it left them as is.

The team also considered adding a variable for 'east' that would signify counties in the eastern portion of the state, calculating it by giving it a value of 1 where central or west were both equal to 0, and 0 otherwise. However, the team realized that adding this variable would make it a perfect linear combination of the central and west variables, and thus it would violate the no perfect multicollinearity assumption, so it did not add it.

Finally, the team also contemplated adding a variable that combined (multiplied) the probability of arrest, probability of conviction, and probability of prison sentence to form a new variable that would gauge the overall probability that a crime committed would result in a perpetrator being imprisoned. The team ultimately elected not to add this variable because of the possibility that one arrest can have multiple convictions (discovered upon noticing prbconv had several counties with values > 1, and there was no information available to clarify this finding; see Data Quality and Anomalies section for more detail), making

the multiplication of these variables potentially inaccurate. Even if this new proposed variable would have shown a strong association with crime, it could have been difficult to interpret given the possibility of multiple convictions per arrest. For these reasons, the team did not add this variable.

Data Quality and Anomalies

This section discusses issues found in the data and how the team addressed them. Each issue is noted in a separate bullet point below.

- Data quality issues addressed prior to the univariate analysis:
  - The year variable contained the same value, 87 (representing the year 1987), for the entire data set. Because it's a constant, it has no variation, and therefore cannot contribute to the model. Because of this the team removed this data point from the dataset for analysis.
  - County 193 was listed twice in the dataset (the entire row was an exact duplicate), therefore the team removed this duplicate row.
- Data quality issues addressed after the univariate analysis:
  - The service industry wages variable (wser) has one data point that is significantly outside of the range of the other values in this variable. With a value of 2,177 this one data point is over 9 standard deviations higher (one standard deviation of this variable is 206) than the next highest value in the data field (391). It is highly unlikely that one county would have a wage whose average is that much higher for the same industry sector. Furthermore, if we make the assumption these values are denominated in US dollars, then $2,177 per week as an average translates to over $113,000 annually, which also seems highly implausible for service industry workers in 1987. Because its value is so vastly higher than the rest of the data both within this variable and across all wage variables, and because it doesn't seem plausible that an average service industry worker would earn this much money, the team considers the likelihood of this data point being erroneous as much greater than it being legitimate, and therefore decided removed it from the analysis.
  - The probability of arrest variable (prbarr) contained one data point of 1.09 (the remaining values in this variable ranged from 0.09 to 0.69). A value greater than one indicates that there are more arrests than there are crimes committed, and it seems implausible to consider a situation where police officers are arresting 9% more people than there are total crimes committed. Given this extreme unlikelihood, and because the research team could not think of a plausible explanation for how the number of arrests could be greater than the number of crimes committed, the team considered this point suspicious enough to be erroneous and removed it from the analysis.
  - The probability of conviction variable (prbconv) contains several values greater than 1, indicating that there are several counties with more convictions than arrests. One possible explanation for this might be that a single arrest could lead to multiple convictions (e.g., a criminal arrested once but convicted of both breaking and entering, and of theft). Because no further explanation is provided and there is no way to further research this point, the team elected to leave this variable as is. However, the reader should be aware of this phenomenon in the data, and the research team recommends a deeper explanation of this variable if further research is performed.
  - A quick internet search revealed North Carolina contains 100 counties, the most recent of which was established in 1911, therefore this dataset (with its 1987 data for 90 counties) is missing 10 counties. These missing counties may or may not be an issue, but the reader should be aware that if the missing counties are biased in any way (i.e., not representative of the provided data set), then this analysis, while valid for the 90 counties contained in the data set, may not be valid for all of North Carolina.
  - The percent minority variable is from 1980, whereas the rest of the data is from 1987, meaning the sampling date of percent minority variable lags all other variables by 7 years. Although the research team does not believe this finding warrants any action (such as deleting), it did think

it was worth mentioning to make the reader aware of it. A subsequent study could reduce the impact (or perception of impact) of any shortcomings this finding may present, either by verifying that minority percentage in counties changes very little in 7 years, or by taking the data from the 1990 census and performing a linear interpolation to arrive at more accurate figures for the minority percentages in 1987.

A few other variables had data points that might have been outliers, but none of them had strong enough evidence to warrant removing them.

The remaining analysis in this report omits the one likely erroneous data point from the wser variable and the one likely erroneous data point from the prbarr variable, both as justified above.

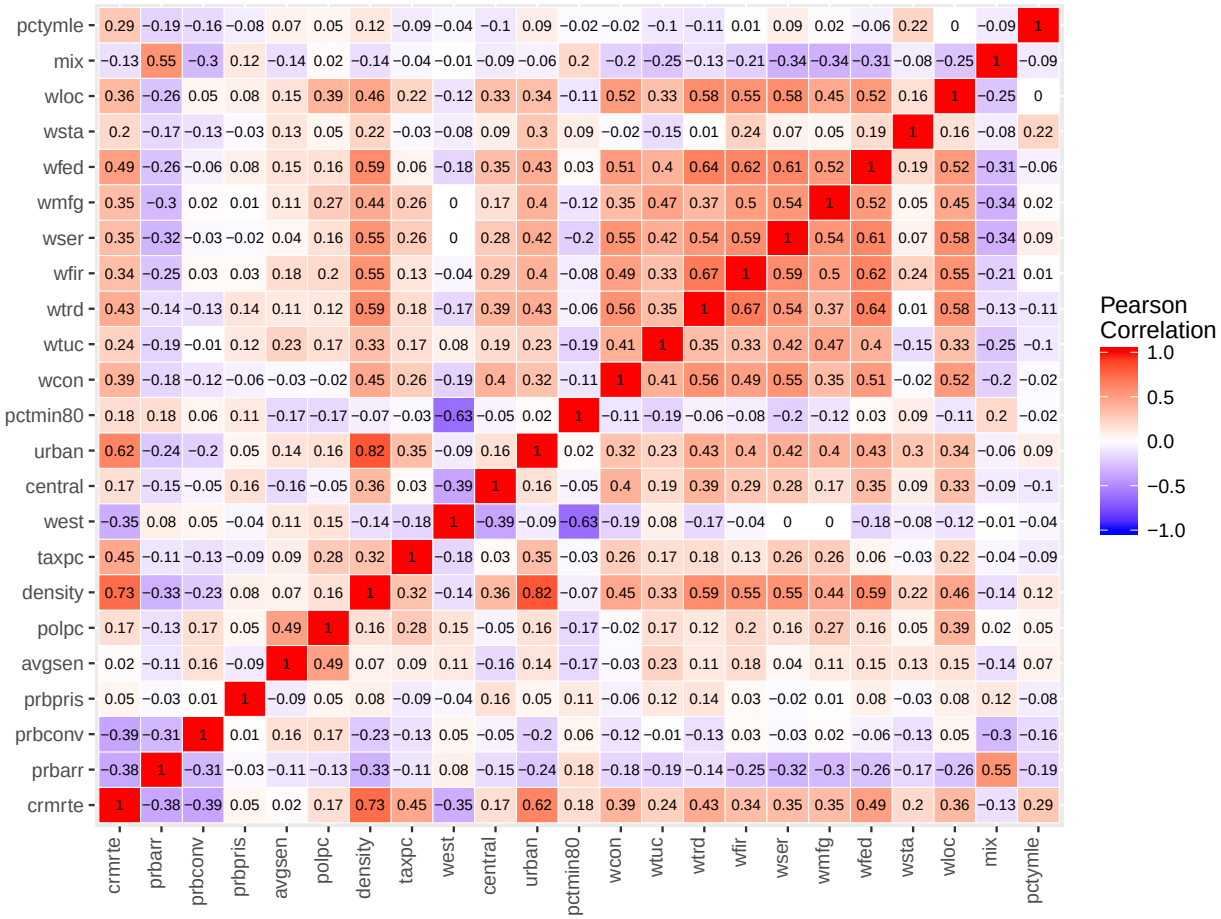
Analysis of More Than One Variable

As part of the exploratory data analysis, the team generated a heatmap to quickly identify bivariate relationships between variables:

```
#Removing erroneous et al data
dt[dt[, wser > 2000], wser:=NA] #Removing outlier data point
dt[dt[, prbarr > 1], prbarr:=NA] #Removing outlier data point

#Now let's make a heat map ...
cor_dt <- melt(round(cor(dt[, .SD, .SDcols = sapply(dt, is.numeric)], use = "pairwise.complete.obs"),2))

ggplot(data = cor_dt, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
    midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0, hjust = 1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank()
        ) +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 2.5)
```



Helping calibrate the analysis of the heatmap, while this report is focused on the data and not the underlying psychology of crime, the authors do note that because crime is a human behavior, and because behaviors often have weaker correlations than other types of data, lower correlation coefficients that may not be noteworthy in other domains may be meaningful for this analysis.

Several items stand out from analyzing this heatmap. First, notice the dependent variable, Crime Rate (crmrte) on the bottommost row and leftmost column of the heatmap. Looking at crime rate's correlations to all other variables, we make the following observations:

- Its highest correlation is with density (people per square mile) at 0.73, a variable which, by itself, explains over half the variance of crime ($0.73\hat{}2 \sim 0.53$, or 53% of the variance).
- Its second highest correlation is with the urban variable at 0.62, also a large value. (We discuss the clear relationship between density and urban in the next section.)
- Beyond these first two, 11 variables have an absolute correlation coefficient with crime in the 0.25 - 0.60 range:
  - 7 of these are wage (out of 9 total wage variables): wcon, wtrd, wfir, wser, wmfg, wfed, and wloc, all of which vary positively with crime rate. And even the other two, wsta and wtuc, have correlation coefficients of just below 0.25.
  - Of the other 4 in this range, prbarr and west vary negatively with crime rate, while taxpc and pctymle vary positively.

Moving beyond relationships between crime rate and the remaining variables, the team noted several strong relationships or groups of relationships between independent variables as seen in the heatmap, the most

notable of which include:

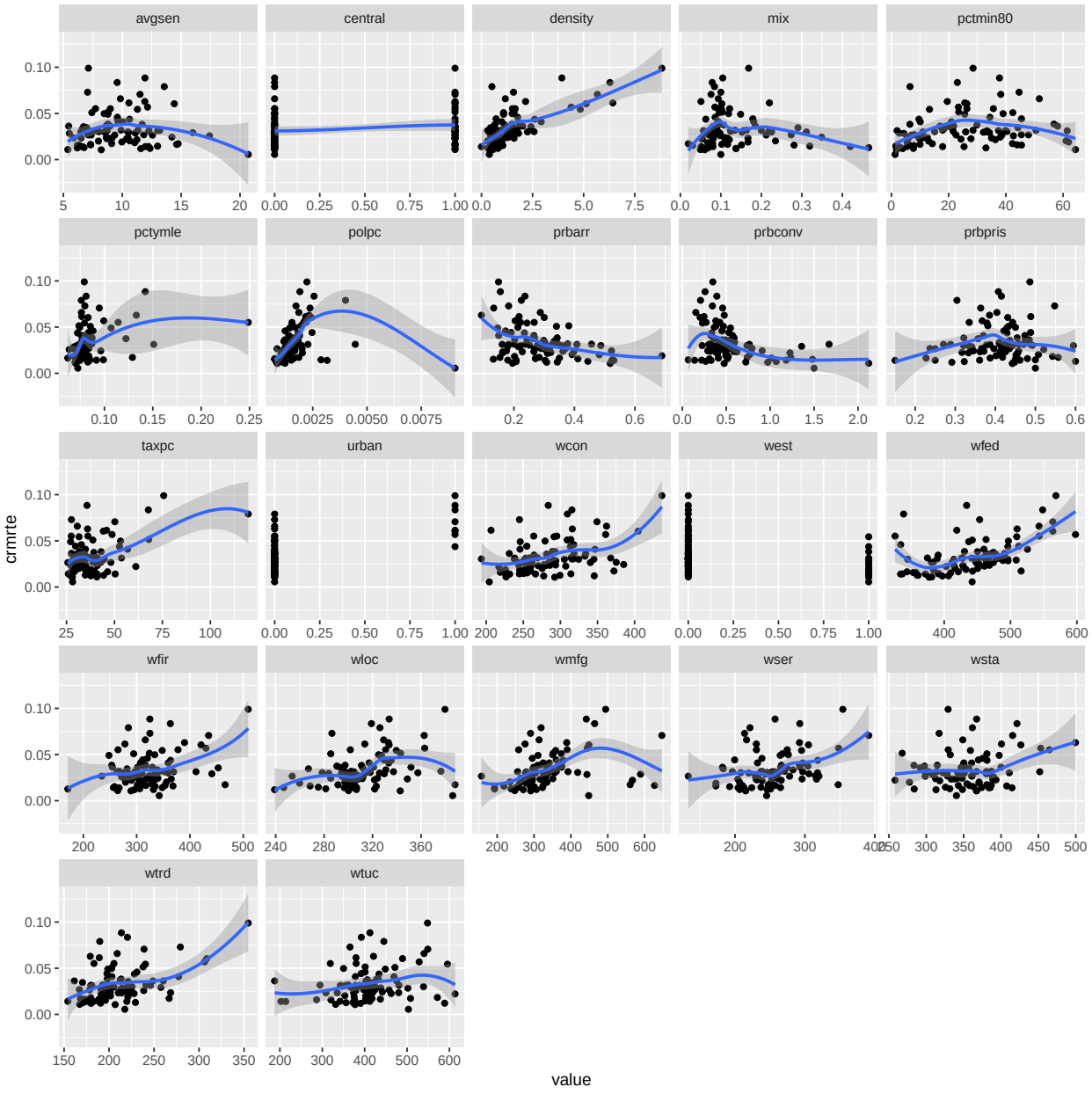
- Density variable with urban variable. Here we see a strong relationship of 0.82 between these two variables, meaning that one is highly explanatory of the other.
- Many of the wage variables are highly correlated with each other, as can be seen by the large, darker red area near the upper right of the heat map. The wage variable that doesn't seem to follow this pattern is wsta (state employees). Possible reasons state employee wages may not trend with other wage groups is that state wages may be fixed by role across the state and therefore may be shielded from market effects that could drive wages up or down, and/or state employees may be concentrated in the state capital, with other counties having few employees which could contribute to a higher variation among the remaining counties due to the small n size of state employees in those counties.
- Density with most wage variables. Here we see positive correlations between density and most of the wage variables, ranging between 0.22 and 0.59.
- Percentage of minority citizens with west variable. Here we note that counties in the part of the state designated as west are likely to have lower proportions of minority citizens, with a -0.64 correlation between the two variables.

Taking a deeper look at the relationship between crime rate and the other variables, the team created several scatter plots:

```
dt_hist <- melt(dt[, .SD, .SDcols = sapply(dt, is.numeric)])

scdata_dt <- gather(dt, -county, -crmrte, key = "var", value = "value")

ggplot(data=scdata_dt, aes(x = value, y = crmrte)) +
  geom_point() +
  stat_smooth() +
  facet_wrap(~ var, scales = "free_x")
```



As an overview, in each of the plots above, the x-axis represents the values of each of the variables listed at the top of each scatterplot, and the y-axis represents crime rate and is on a fixed scale across all plots. The blue line and darker grey bands running through each plot give a visual representation of how each variable varies with crime.

Visual analysis of these scatter plots reveals the following insight:

- Confirming our finding from the heatmap, density appears to have a tight upward relationship with crime. And it seems to have one of the most obvious linear relationships among all the variables.
- Pctymle appears to have a strong positive relationship with crime for pctymle values under 0.1, but afterwards tapers off. With only a few data points greater than 0.1, it could be that there's so little data that it's too hard to form a strong relationship in this area, or it could mean that pctymle's effects on crime do actually plateau above 0.1.
- Polpc appears to increase with crime, but practically, although it demonstrates an associative relationship,

it's hard to hypothesize that more police per person are causing more crime. If anything, the more plausible effect is in the opposite direction: more crime motivates communities to staff more police. Because of this more likely relationship, and because police rate is not the variable in question, the research team is not further considering this variable as part of its potential list of independent variables on which to model crime.

- Prbarra and prbconv appear to have downward relationships with crime, consistent with our initial hypothesis.
- Tax revenue per capita (taxpc) shows an upward trend with crime, though there is notable variation, and we speculate this variable may have some overlap with density. This relationship contradicts our initial hypothesis.
- Most of the wage variables (wloc, wsta, wfed, wmfg, wser, wfir, wtrd, wtuc, and wcon) show a relatively tight relationship with crime, especially in their denser regions (though not as much in their outlying points). And in these more dense regions, they also appear to have an upward sloping relationship with crime; that is, as wages increase, so does crime. These positive relationships between wages and crime contradict our initial hypothesis.
- For the binary coded variables (urban, west, and central), urban areas tend to have higher crime, areas flagged as west tend to have lower crime, and central areas tend not to show a notable difference on crime. Higher crime in urban areas matches our initial hypothesis.
- Percent minority as of the 1980 census (pctmin80) is interesting, in that crime appears to be highest in the middle of its range. This supports a hypothesis that communities with higher percentage of minorities may face unfair socioeconomic disadvantages, which may lead to higher crime rates. Because minority percentage could be a predictor of crime, we will consider it at least one of the linear regression models.
- The remaining variables, avgsen, mix, prbpris, did not seem to have a noteworthy relationship with crime.

Any scatter plots not specifically mentioned are deemed to not contribute significantly enough to the modeling or conclusions to be worthy of commentary.

3. Model Building Process

The model selection process builds off our initial hypotheses and exploratory data analysis in the previous sections. The research team built three models in total. To summarize, each model contained the following variables:

Model 1. Key explanatory variables only: density + prbarr + prbconv + pctymle

Model 2. Key explanatory variables + enhancing covariates: Model 1 + west + pctmin80

Model 3. Comprehensive model: All variables except prbpris, avgsen, polpc, mix, and urban.

Model 1

Model 1 Variable Selection

For the first model, our goal was to first identify a few key variables that we believe most likely explain our crime rate dependent variable. Our initial hypotheses considered crime relative to the following variables: police per capita, probability of arrest, probability of conviction, probability of prison sentence, average sentence length, population density, urban status (binary), wage variables, tax revenue per person, and percent of young males in the population. After exploring our data, we removed several of these variables for the following reasons:

- Police per capita: Removed because it increases as crime rates increase, and we don't believe that more police cause more crime. Rather, as stated earlier, a more plausible explanation is that as crime increases, communities respond by adding more police, making polpc appear as more of a dependent variable in this study, and thus making it invalid for our analysis.
- Probability of prison sentence and average sentence length: Removed because the exploratory data analysis showed almost zero relationship to crime.
- Urban: Removed because of its very strong relationship to density, which makes it seem duplicative. We chose density over urban because we felt density has a greater likelihood of predicting crime because it is a continuous variable as opposed to a binary one. Supporting this theory, our heatmap showed density had a 0.73 correlation with crime rate, whereas urban was less at 0.62.
- Tax revenue and wages: Removed because our hypothesis was that crime decreases as these increase, but our exploratory data analysis showed they actually increase as crime rates increase.

We may reintroduce some of these in subsequent models.

This leaves us with the following four variables for Model 1: Probably of arrest, probability of conviction, density, and percent of young males in the population, which had correlations of 0.38, 0.39, 0.73, and 0.29, respectively. Visual inspection of the scatterplots of these variables vs. crime all suggest relationships in agreement with our hypotheses, further supporting the choice of these data fields.

As a reminder, all models are built with the dependent variable as the natural log of crime rate to help meet linear regression assumptions, as discussed in the Data Transformation section of the report. Therefore, regression coefficients will represent how changes in the independent variables correspond to percent changes in crime rate, as opposed to absolute changes in crime rate.

```
#Model 1: Explanatory variables only
dt <- dt[!is.na(prbarr)]  #Allows for testing covariance
model1 <- lm(log(crmrte) ~ density + prbarr + prbconv + pctymle , data = dt)
```

Model 1 Assumptions

The classical linear model assumptions will be discussed in detail for Model 1 (log(crmrte) ~ density+ prbarr + prbconv + pctymle). Specifically, we are testing for the following 6 CLMs:

1. Linear population model
2. Random Sampling
3. No perfect multicollinearity
4. Zero-conditional mean
5. Homoskedasticity
6. Normality of Errors

Assumption 1: Linear Population Model: This assumption is automatically met because we created a linear model.

Assumption 2: Random Sampling: Random sampling is tested by evaluating the data collection process. The data set we have is a collection of data from 90 counties in North Carolina, and the state has a total of 100 counties. Although 90 samples out of 100 is high, it begs the question, "why weren't all 100 counties selected, and therefore what bias might the unselected counties have on the results?" Because we do not have the answer to this question, it requires additional exploration on the data sampling process.

This assumption further depends on what we're considering as the full population. If North Carolina is our population, then this data is more likely to be representative than if this study was being considered for understanding the drivers of crime for the whole country. In this latter case, North Carolina may not be a representative sample, and we'd likely want to sample a more representative distribution of the US population.

Given the considerations stated above, because we cannot say with absolute confidence that the data is indeed randomly sampled, we caution the reader that this assumption may be violated.

Assumption 3: No Perfect Multicollinearity:

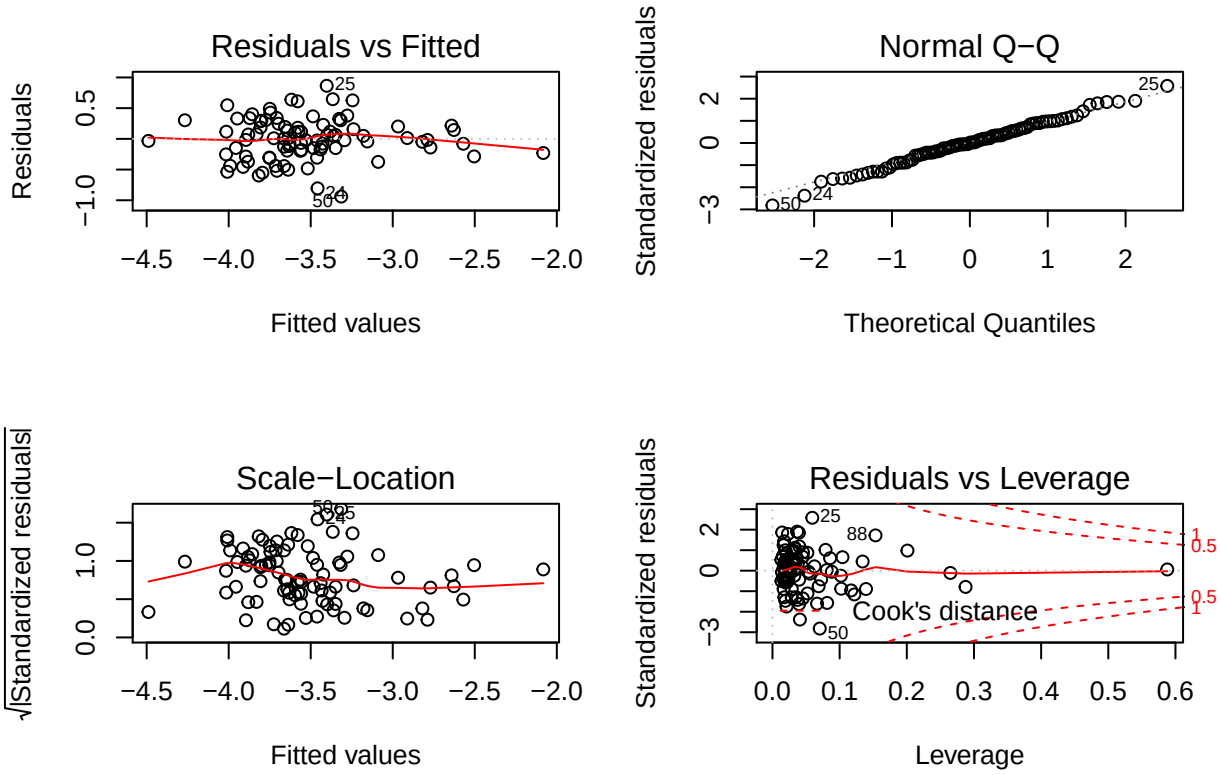
```
vif(model1)
```

```
##  density   prbarr  prbconv  pctymle
## 1.278900 1.427889 1.324947 1.094245
```

Visual inspection of the scatterplot data of our variables against crime in the Analysis of More Than One Variable section above shows no perfectly linear relationships and ample scatter of all variables, suggesting this assumption is not violated. Furthermore, we confirm our data does not have high but imperfect collinearity by calculating the variance inflation factor, which explains how much the standard error of our coefficients is inflated due to other variables in our model. The VIF values for our model range from 1.09-1.43 (shown above), which is well below the rule-of-thumb level of 4. Thus, our model does not violate this assumption.

Assumption 4: Zero-Conditional Mean:

```
par(mfrow=c(2,2))
plot(model1)
```



This assumption is measured by viewing the Residuals vs Fitted plot above to see that the expected value of the errors is zero. From the plot, we can see that the mean for the values is a relatively flat line and is close to zero. Thus, this assumption has not been violated. Also, from the Residuals vs Leverage plot, we see that no single observation has a large Cook's distance, thus indicating that no single observation has undue influence on the model.

```
#Test for covariance with residuals
cov(model1$residuals, dt$prbarr)
```

```
## [1] -1.579793e-18
```

```
cov(model1$residuals, dt$prbconv)
```

```
## [1] -3.349476e-18
```

```
cov(model1$residuals, dt$density)
```

```
## [1] -2.857427e-18
```

```
cov(model1$residuals, dt$pctymle)
```

```
## [1] -2.077348e-19
```

Finally, we calculate the covariance of the residuals against the model's variables to see they are all nearly zero, further suggesting no violation of this assumption.

Assumption 5: Homoskedasticity: This assumption is viewed from the Residuals vs Fitted plot by viewing a constant band of thickness among the values. As the values increase from left to right, which shows a decrease in thickness on the right side. The Scale-Location plot also shows a line that's not horizontal, giving further evidence that we have likely violated the homoskedasticity assumption. Because of this violation, we use more robust standard errors that account for heteroskedasticity.

Assumption 6: Normality of Errors: This assumption is measured from the Normal Q-Q plot above. We can see that our standardized residuals fit nicely along the dotted line, thus indicating that our errors are normally distributed and that our model does not violate this assumption.

Model 1 Results

This section discusses the results and performance of Model 1.

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.206180   0.299543 -10.7036 < 2.2e-16 ***
## density      0.153784   0.028267   5.4404 5.146e-07 ***
## prbarr      -1.512038   0.500877  -3.0188 0.0033600 **
## prbconv     -0.573547   0.163761  -3.5023 0.0007415 ***
## pctymle      2.414283   1.076667   2.2424 0.0275721 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All four variables in our model show statistical significance, with the probability of arrest at a high level ($P < 0.01$), and both population density and probability of conviction at very high levels ($P < 0.001$). Therefore, we can reject the notion (reject the null hypothesis) that the relationships between these variables and the log of crime rate are simply due to chance.

From a practical significance standpoint, we make the following observations:

- As density increases by 1 person per square mile, crime increases by about 15%. Given that the density variable ranges from approximately 0 to 9 people per square mile across all counties in the sample, this seems like a very significant driver of crime.
- As probability of arrest increases by 0.01 (not the same as a 1% increase), we see crime decreases by ~1.5%. Looking at its interquartile range, moving from the 1st quartile of 0.20 to the 3rd quartile of 0.34 (a 0.14 difference), the model suggests a 21% reduction in crime across this range (holding all

other variables constant); therefore, we see this variable as having considerable opportunity to make a practically significant reduction in crime.

- As probability of conviction increases by 0.01, (not the same as a 1% increase), we see crime decreases by ~0.5%. Although this variable has some outlier values at the high end of its range (as high as 2.1), we still consider it practically significant because looking at its interquartile range of 0.34 to 0.59 (a more conservative approach), this range is still large enough to presents an opportunity to reduce crime. Holding all other variables constant, increasing convictions from the first quartile to the 3rd quartile would increase the rate by 0.25, which this model predicts would reduce crime by 12.5%.
- As the proportion of a county's young male population increases by 0.01, we see crime increase by ~2.4%. For most counties, this variable has a low range, noting that its lowest value is 0.06, and the 3rd quartile is 0.08. However, even across this small range, the model predicts a 4.8% reduction in crime (holding all other variables constant), which we consider practically significant.

Model 2 (with some comparisons to Model 1)

Model 2 Variable Selection

For Model 2, we added two more variables: The 'west' (binary) variable and the percent of minority population in a county.

The west variable had a -0.35 correlation with crime rate, and the Model 1 did not have a variable that described location, so we thought it safe to add as we anticipate little interaction with other variables.

Percent minority had a smaller, 0.18 correlation with crime, but it stood out as a variable that was generally unique to the data set (it showed minimal interaction with other variables). While it adds a second demographic variable to the model, it had almost no correlation (-0.02) to the other demographic variable, percent young male, suggesting a low chance that these two variables would have unwanted interactions in the model.

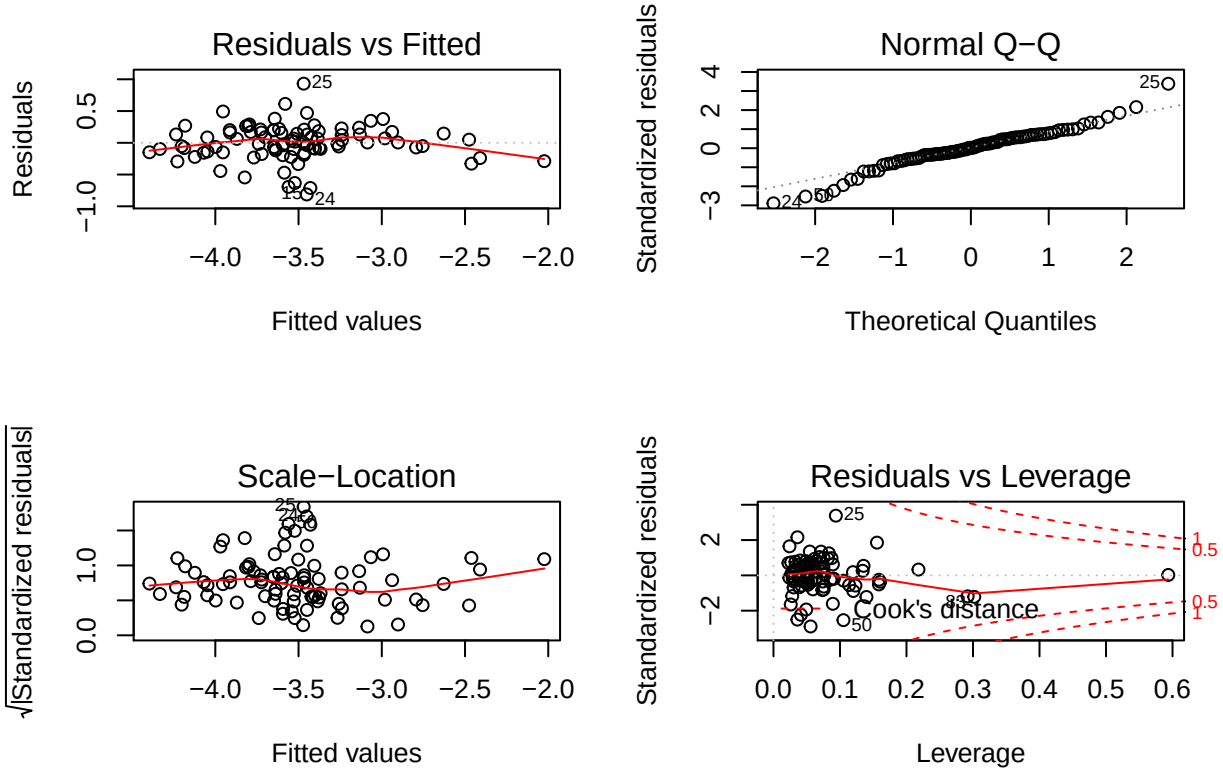
We chose to leave other variables out of this model as we wanted to keep it relatively simple and reduce the potential for variable interaction, also knowing that we will introduce several more variables in Model 3.

```
#Model 2: Explanatory variables + enhancing covariates
model2 <- lm(log(crmrte) ~ density +  prbarr + prbconv + pctymle + west + pctmin80, data = dt)
```

Model 2 Assumptions

For Model 2, the same tests were performed to assess CLM assumptions, as shown below:

```
par(mfrow=c(2,2))
plot(model2)
```




```
vif(model2)
```

```
##  density   prbarr  prbconv  pctymle     west pctmin80
## 1.296036 1.590400 1.399574 1.097771 1.824251 1.925495
```

```
#Test for covariance with residuals
cov(model2$residuals, dt$prbarr)
```

```
## [1] -1.202185e-18
```

```
cov(model2$residuals, dt$prbconv)
```

```
## [1] 7.416928e-19
```

```
cov(model2$residuals, dt$density)
```

```
## [1] -1.29144e-17
```

```
cov(model2$residuals, dt$pctymle)
```

```
## [1] 1.113096e-19
```

```
cov(model2$residuals, dt$west)
```

```
## [1] 2.487197e-19
```

```
cov(model2$residuals, dt$pctmin80)
```

```
## [1] -1.066399e-16
```

For Model 2, we observe that it potentially violates the random sampling assumption discussed in Model 1.

Homoskedasticity is violated because the band of fitted vs observed values starts narrow on the left side, widens towards the middle, then narrows again on the right side. Because of this violation, we use more robust standard errors that account for heteroskedasticity.

We also notice a potential violation of normality of errors, as the values on the Q-Q Norm plot start to stray from the dotted line on both ends of the axis. Because the dataset is sufficiently large (n = 90, larger than the 30 rule-of-thumb threshold), we rely on the asymptotic properties of OLS regression and are not overly concerned with this violation.

Model 2 Results and Comparison to Prior Model

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

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.1970202  0.2692620 -11.8733 < 2.2e-16 ***
## density      0.1445510  0.0292735   4.9380 4.094e-06 ***
## prbarr      -1.8620845  0.3658354  -5.0900 2.237e-06 ***
## prbconv     -0.6697316  0.1515342  -4.4197 3.002e-05 ***
## pctymle      2.0425928  0.6864257   2.9757  0.003840 **
## west        -0.1533491  0.1077964  -1.4226  0.158651
## pctmin80     0.0086136  0.0028395   3.0335  0.003237 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at our new variables, our first observation of coefficients in Model 2 is that counties designated as 'west' do not have a statistically significant relationship with crime rate (P=0.16), meaning that although counties in the west show lower crime rates, there is insufficient evidence to reject the notion that this isn't simply due to random chance.

Next, we see percent minority population does have a statistically significant relationship with crime (P=0.003). But from a practical standpoint, its coefficient of 0.008 tells us that for every 0.01 increase in percent minority (i.e., for every 1 percentage point increase of minority population) there is a 0.008% increase in crime. This value is so small we consider it to be meaningless, and therefore assert that we do not see practical significance with this variable.

Finally, we notice some small changes to the coefficients of the four variables included in Model 1. We see density remained about the same, moving from 0.15 in Model 1 to 0.14 in Model 2. Probably of arrest's coefficient increased from -1.5 in Model 1 to -1.9 in Model 2, and probably of conviction went from 0.57 to 0.67 in Model 2, both of which tell us that that their magnitude increased, suggesting that our two added variables (west and percent minority) in combination were reducing their effects when they were omitted variables in Model 1. Lastly, we observe that percent young male dropped from 2.4 in Model 1 to 2.0 in Model 2, suggesting that the two variables omitted in Model 1 but added in Model 2 weakened its effect.

Model 3 (with some comparisons to Model 1 and Model 2)

Model 3 Variable Selection

The third model is the culmination of all variables but excluding variables that we feel were either not representative to crime rate or added too much additional bias and covariance. We included most of the variables available, both to see their effects, as well as to see their impact on the significance of variables in Models 1 and 2.

We included all the tax and wage variables in this model, because we do see strong correlations between these variables and crime (correlation coefficients vs. crime ranging from 0.20 to 0.49), and although our exploratory analysis indicated they moved (relative to crime rate) in the opposite direction than we initially hypothesized, we subsequently hypothesized that wealth might actually attract impoverished people. That is, when the average taxes and wages of a county increase, some of this additional wealth might be utilized on resources for impoverished people (e.g., homeless shelters). Therefore, these variables may still be valid proxies for poverty, but in the opposite direction than was initially hypothesized. Of note, this study isn't designed to test or refute this new hypothesis, so the fact that it seems plausible is sufficient rationale for including these variables in our third model.

We also added the 'central' variable as a demographic variable, as it has a 0.17 correlation coefficient with crime rate, and may help explain additional effects of location.

We initially also included the urban variable in Model 3, but we ultimately removed it because of its very strong relationship with density, and we believe density is a better variable to include because of its stronger relationship with crime and because it's not binary and can therefore give a more nuanced insight into predicting crime rates. As evidence of their strong overlap, the variance inflation factor (VIF) for density when urban was included in the model was 5.3, exceeding the rule-of-thumb threshold of 4, but when urban was removed, density's VIF reduced to 2.5. We consider this a worthwhile tradeoff to make, sacrificing the incremental predictive power that urban provided for the sake of meeting linear modeling assumptions. Finally, removing urban also helps with interpretability, because the coefficients of these linear models make the assumption that all other variables are held constant, yet it seems not possible in most cases to hold density constant while changing a county's urban status. Said differently, it seems unlikely two different counties have roughly the same density with one considered urban and the other not.

We omitted probability of imprisonment (prbpris) and average sentence length (avgsen) from this model because they show nearly zero relationship with crime (with correlation coefficients of 0.05 and 0.02, respectively).

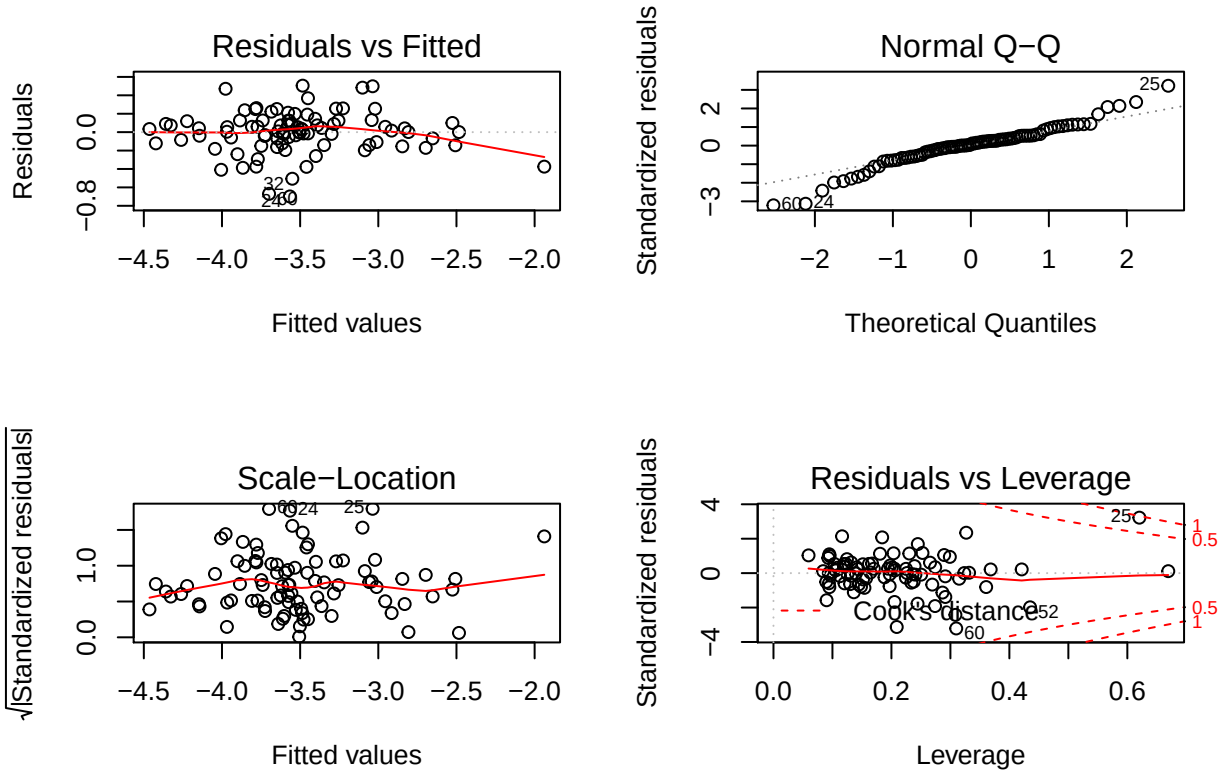
We omitted the mix variable because we felt it difficult to interpret, and because of its low relationship to crime rate (correlation coefficient is -0.13)

Finally, we chose not to include police per capita (polpc) because, as stated earlier, the data show this variable as being an outcome of crime, as opposed to a predictor of crime.

```
#Model 3: Previous model + additional covariates
model3 <- lm(log(crmrte) ~  prbarr+prbconv+density+taxpc+west+central+pctmin80+wcon+wtuc+wtrd+wfir+wser
```

Model 3 Assumptions

```
par(mfrow=c(2,2))
plot(model3)
```




```
vif(model3)
```

```
##   prbarr  prbconv  density    taxpc     west  central pctmin80     wcon
## 1.726331 1.518853 2.475517 1.575732 3.289910 2.137778 2.806754 2.125234
##     wtuc     wtrd     wfir     wser     wmfg     wfed     wsta     wloc
## 1.573826 2.996704 2.757576 2.575882 1.857702 2.828530 1.474007 2.294408
##  pctymle
## 1.345716
```

For Model 3, we observe that it potentially violates the random sampling assumption discussed in Model 1.

Similar to Model 2, homoskedasticity appears violated because the band of fitted vs observed values starts narrow on the left side, widens towards the middle, then narrows again on the right side. Because of this violation, we use more robust standard errors that account for heteroskedasticity.

Also similar to Model 2, we notice a potential violation of normality of errors, as the values on the Q-Q Norm plot start to stray from the dotted line on both ends of the axis. We note that the deviation from the diagonal line is slightly greater than in Model 2, but because the dataset is sufficiently large (n = 90, larger than the 30 rule-of-thumb threshold), we rely on the asymptotic properties of OLS regression and are not overly concerned with this violation.

Finally, the 'west' variable had a somewhat higher VIF of 3.3, but because it is under 4, we are not overly concerned with multicollinearity effects.

Model 3 Results and Comparison to Prior Models

```
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4.5222e+00  7.6847e-01 -5.8847  1.25e-07 ***
## prbarr      -1.7634e+00  4.3091e-01 -4.0922 0.0001128 ***
## prbconv     -6.1606e-01  1.5869e-01 -3.8822 0.0002318 ***
## density      9.4872e-02  4.8811e-02  1.9436 0.0559593 .
## taxpc        7.7752e-03  8.5323e-03  0.9113 0.3652854
## west        -7.4251e-02  1.4489e-01 -0.5125 0.6099319
## central     -8.1850e-02  1.0301e-01 -0.7946 0.4295393
## pctmin80     8.8738e-03  4.0219e-03  2.2064 0.0306411 *
## wcon         3.0850e-04  8.8069e-04  0.3503 0.7271669
## wtuc         5.4251e-05  6.1888e-04  0.0877 0.9303978
## wtrd        -3.8952e-04  1.9708e-03 -0.1976 0.8438989
## wfir        -7.4180e-04  1.3779e-03 -0.5384 0.5920325
## wser        -1.6352e-03  1.3688e-03 -1.1947 0.2362487
## wmfg        -1.2367e-04  4.7176e-04 -0.2622 0.7939750
## wfed         3.2675e-03  9.7674e-04  3.3453 0.0013240 **
## wsta        -1.4682e-03  8.6852e-04 -1.6905 0.0953793 .
## wloc         1.9757e-03  2.0316e-03  0.9725 0.3341740
## pctymle      4.1908e+00  1.1256e+00  3.7232 0.0003943 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our results of Model 3, we notice that most of the variables that showed both statistical and practical significance in Models 1 and 2 also show significance here, namely the probabilities of arrest (P=0.0001) and conviction (P=0.0002), and the percent young male (P=0.0004).

Probability of arrest's coefficient was -1.8, similar to its values of -1.9 in Model 2 and -1.5 in Model 1, suggesting agreement among the models.

Probability of conviction's coefficient of -0.62 in Model 3 was right at the midpoint of a tight range of all three models, with Model 1 at -0.57 and Model 2 at -0.67, also suggesting agreement among the models.

Notably, percent young male's coefficient increased from 2.4 in Model 1 and 2.0 in Model 2 to 4.2 in Model 3. Holding all these other factors constant that were not included in Models 1 and 2, percent young male has a much greater impact on crime rate, at the rate of a 4.2% increase in crime for every 1 percentage point increase in a county's percentage of young male citizens.

Percent minority also retains statistical significance in Model 3, but its coefficient remains practically insignificant just as we saw in Model 2.

Density, interestingly, went from definitively statistically significant in Models 1 and 2, to borderline significant in Model 3 (P=0.056). Intentionally, we did not tweak Model 3 to fish for the slightly better P value that would have moved density into the statistically significant realm of $P < 0.05$, but we do note that this P value is still very low, with only a 5.6% chance that we should not reject the null hypothesis (that density's effect is due to pure chance). Therefore, we still consider density to have a reasonably strong likelihood of being a real driver of crime.

One of our wage variables, wfed, shows statistical significance (P=0.0013), and although its coefficient of 0.0033 might seem small, recall that this variable has a range of 330 - 600. So for every $10 increase in wages (assuming its units are dollars), we see a 3.3% increase in crime rate. Observing this data alone suggests practical significance, yet none of the other wage variables from other sectors have statistical significance, leaving us to ask the question, "why do increased wages of federal workers in particular drive up crime rates?". Absent the ability to answer this question, we consider this result questionable, and therefore do not assert that wages (federal or other sectors) impact crime.

The remaining variables did not show any statistical significance with crime, and therefore we also conclude that they have no practical effect on crime rates.

4. Regression Table

```
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))
stargazer(model1, model2, model3,
  type = "latex",
  title = "Linear Models Predicting NC Crime Rate",
  no.space=TRUE,
  single.row=TRUE,
  omit.stat=c("LL","ser","f"), omit.table.layout="n",star.cutoffs = c(0.05, 0.01, 0.001),
  add.lines=list(c("BIC", round(BIC(model1),1), round(BIC(model2),1),round(BIC(model3),1)),
    c("AIC", round(AIC(model1),1), round(AIC(model2),1),round(AIC(model3),1))),
  se = list(se.model1, se.model2, se.model3))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Apr 17, 2018 - 4:32:44 PM

Table 5: Linear Models Predicting NC Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
density	0.154^{***} (0.028)	0.145^{***} (0.029)	0.095 (0.049)
taxpc			0.008 (0.009)
prbarr	-1.512^{**} (0.501)	-1.862^{***} (0.366)	-1.763^{***} (0.431)
prbconv	-0.574^{***} (0.164)	-0.670^{***} (0.152)	-0.616^{***} (0.159)
pctymle	2.414^{*} (1.077)	2.043^{**} (0.686)	4.191^{***} (1.126)
west		−0.153 (0.108)	−0.074 (0.145)
central			−0.082 (0.103)
pctmin80		0.009^{**} (0.003)	0.009^{*} (0.004)
wcon			0.0003 (0.001)
wtuc			0.0001 (0.001)
wtrd			−0.0004 (0.002)
wfir			−0.001 (0.001)
wser			−0.002 (0.001)
wmfg			−0.0001 (0.0005)
wfed			0.003^{***} (0.001)
wsta			−0.001 (0.001)
wloc			0.002 (0.002)
Constant	-3.206^{***} (0.300)	-3.197^{***} (0.269)	-4.522^{***} (0.768)
BIC	85.2	60.4	71.4
AIC	70.3	40.5	24.3
Observations	89	89	88
R^2	0.583	0.714	0.808
Adjusted R^2	0.563	0.694	0.762

The regression table displays the coefficients for each of our three models. The statistical significance and practical significance of the coefficients in all models, discussed in the individual model sections above, can be seen here for an easy comparison across all the models.

The first model has an adjusted R-squared of 0.563, the second model has 0.694, and the third model has 0.762. These make sense because as we add more variable into our model we always expect the R-squared value to increase. All models explain a large amount of the variation in crime rates (56%, 69%, and 76%, for Models 1, 2, and 3, respectively).

As additional measures of each model we ran Akaike's An Information Criterion (AIC) and Bayesian Information Criterion (BIC), the results for which can be seen at the bottom of the regression table above. Both measures penalize the number of parameters in the model, and the smaller the AIC or BIC value, the better the fit. Against these measures, AIC showed Model 3 to be the best, but BIC, with its larger penalty for additional input variables, favored Model 2.

5. Omitted Variables and Other Modeling Considerations

This section discusses variables that are not present in the data set, and therefore are not contained in the analysis, but that may also serve as either associative or causal factors toward influencing crime rate. Each item below lists a potential omitted variable or concept, the bias it might have on the model from being excluded (both direction and magnitude if possible), any potential variables included in the dataset that might serve as a proxy for the omitted variable.

- One omitted variable to consider is police officer ability. We surmise that as police officer ability increases, crime rates decrease. However, because this variable would be nearly impossible to measure, we make the assumption that police officer ability is evenly distributed throughout the data and therefore does not bias the results.
- A related omitted variable to consider is policing resources. Examples include, more squad cars, sophisticated communication equipment, or a data science team that helps predict and understand the drivers of crime, and we expect the presence of any or all of these resources to drive down crime rates.
- We initially hypothesized that poverty would have a positive relationship with crime (surmising that as poverty increases, so does crime), and considered wage and tax revenue as inverse proxies for poverty. Upon exploring the data, the team found the opposite relationship to be true, and now believes that average income in a county may not be a good substitute for its levels of poverty. Therefore, in this analysis, poverty may still be an omitted variable, such that higher levels of poverty drive up crime rates.
- Related to poverty, external economic / employment factors are another group of potential omitted variables that could affect crime. For example, if any large businesses (e.g., an auto-plant employing thousands of workers) started or closed in a particular county, that could affect crime in that county or nearby counties, the assumption being that as jobs increase, crime decreases. So if a large job boom hit a few counties, that could bias crime rates downward relative to other demographics included in the data, and conversely, if a massive layoff hit, then that could bias crime rates upwards. It is difficult to speculate on the magnitude of such effects, as they would depend on the size of the employment change and the relationship between employment and crime, whether linear or not.
- Illegal drugs (or supply of drugs) may be another omitted variable influencing crime rates. 1987 was a time when drug supply from foreign cartels was impacting the country, and to the extent that increased drug supply impacted some North Carolina counties more than others, and with the presumption that increased drugs also increases crime, then these counties may have had higher rates of crime, biasing the crime rates farther upward than they otherwise would have been (absent a surge in illegal drug prevalence) relative to the variables included in the dataset.

- Type of crime is a potential omitted variable. For example, small crimes like shoplifting may happen more frequently, while severe crimes like homicide may be rare, but the public, and therefore politicians representing the public, likely care about them much more. Because crime rate in this dataset measures all crimes equally, a better measure may be some sort of weighted crime index as the dependent variable. Because this data is absent, it could bias the crime rate in either direction. Specifically, if actual crimes committed tended to have higher levels of severity (homicides, aggravated assaults, etc.) then the current data set may be underreporting a crime relative to a severity-weighted crime dependent variable. And vice versa, if actual crimes committed were on the lighter side (petty theft, jaywalking, etc.), then this current data set might be over-reporting crime relative to a severity-weighted crime dependent variable.
- Granularity of data. While not a specific omitted variable, the data is only supplied at the county level, which means that it could be masking effects of other demographics. For instance, counties with a large urban population in one small area with an otherwise rural population in the rest of the county may generate different findings than if that same territory were split into two counties, one for the urban area and the other for the rural one. With North Carolina county boundaries drawn decades ago, it's plausible that they don't fall along breakpoints that may better help describe crime. This averaging may tend to "water down" results, lower effects of many demographics that would otherwise be stronger if we had data at a more granular level, or if county boundaries were drawn differently.

6. Conclusions

After a careful review and analysis of the data, the team formed the following main conclusions.

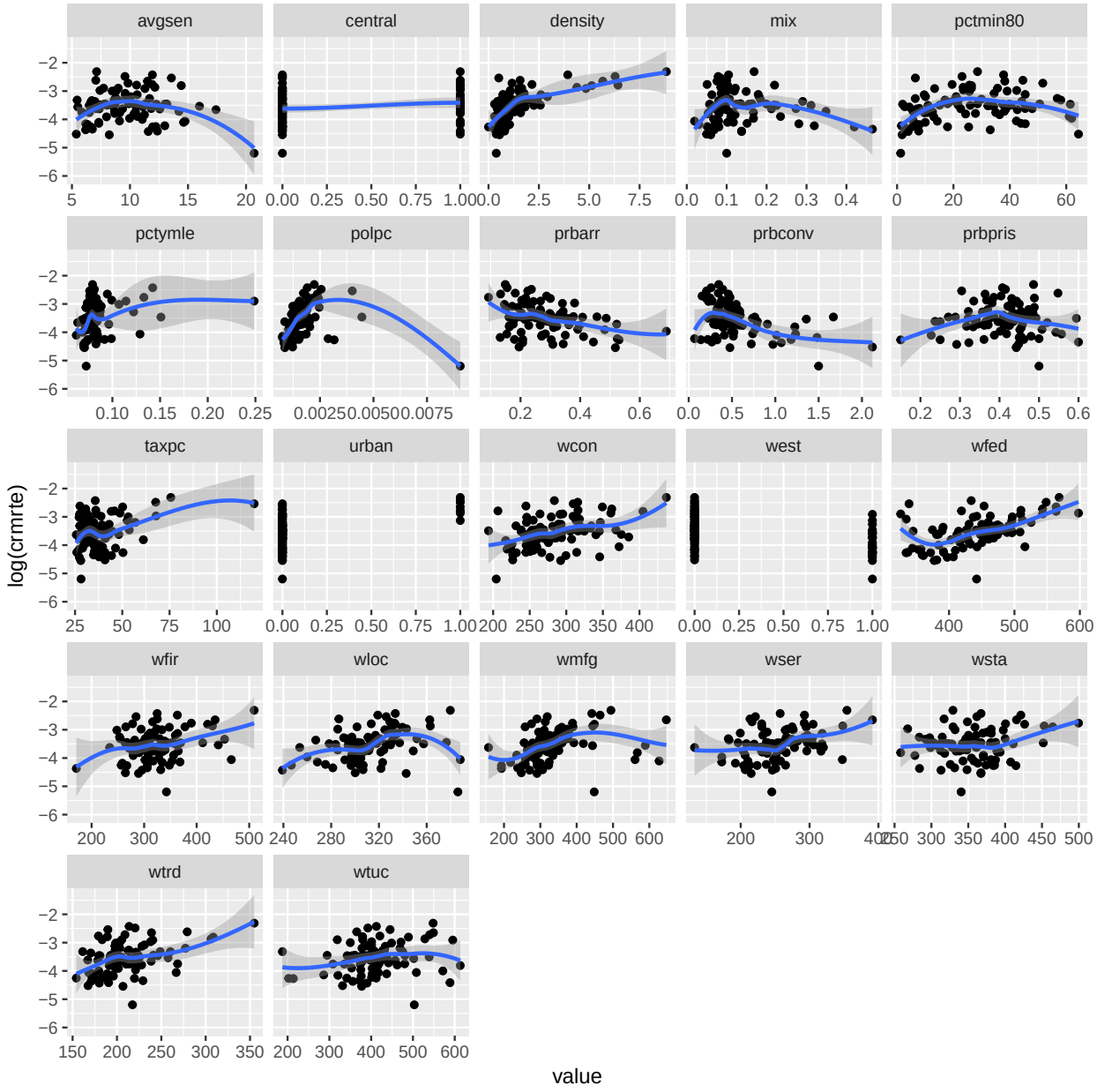
- A big predictor of crime is population density, with more densely populated areas having not just more total crime, but more crime per person. Although it moved just barely out the range of statistical significance in Model 3, its effect's likelihood of being due to random chance is still very small, and thus the team considers it practically significant. For the political campaign, obviously changing population density is infeasible, so the research team recommends targeting areas with high population density and dedicating resources to address crime in these areas.
- Both probability of arrest and probability of conviction are also predictive of crime, and the research team suggests the political campaign focus on methods that increase arrest and conviction rates for crimes, noting that locations with higher arrest and conviction rates see significantly lower crime rates. Both these variables are also robust to all three model specifications.
- The final driver of crime is the percentage of young males in a county, which also shows significance across all three models and is therefore also considered robust. As with density, changing the percentage of this demographic is infeasible, so the research team recommends targeting areas with concentrations of young males and dedicating resources to address crime in these areas.

Appendix

Scatter Plot of Input Variables vs. Log of Crime Rate

This section shows the same scatter plots as in the main document, but with the log of crime rate instead of the raw crime rate data.

```
ggplot(data=scdata_dt, aes(x = value, y = log(crmrte))) + geom_point() + stat_smooth() + facet_wrap(~ va
```



Viewing these results, the team saw little to no difference in the scatter plot shapes and distribution. It is, however, less intuitive to interpret the y-axis (the log of crime rate), which is why it is in the Appendix and not the main body of the document.

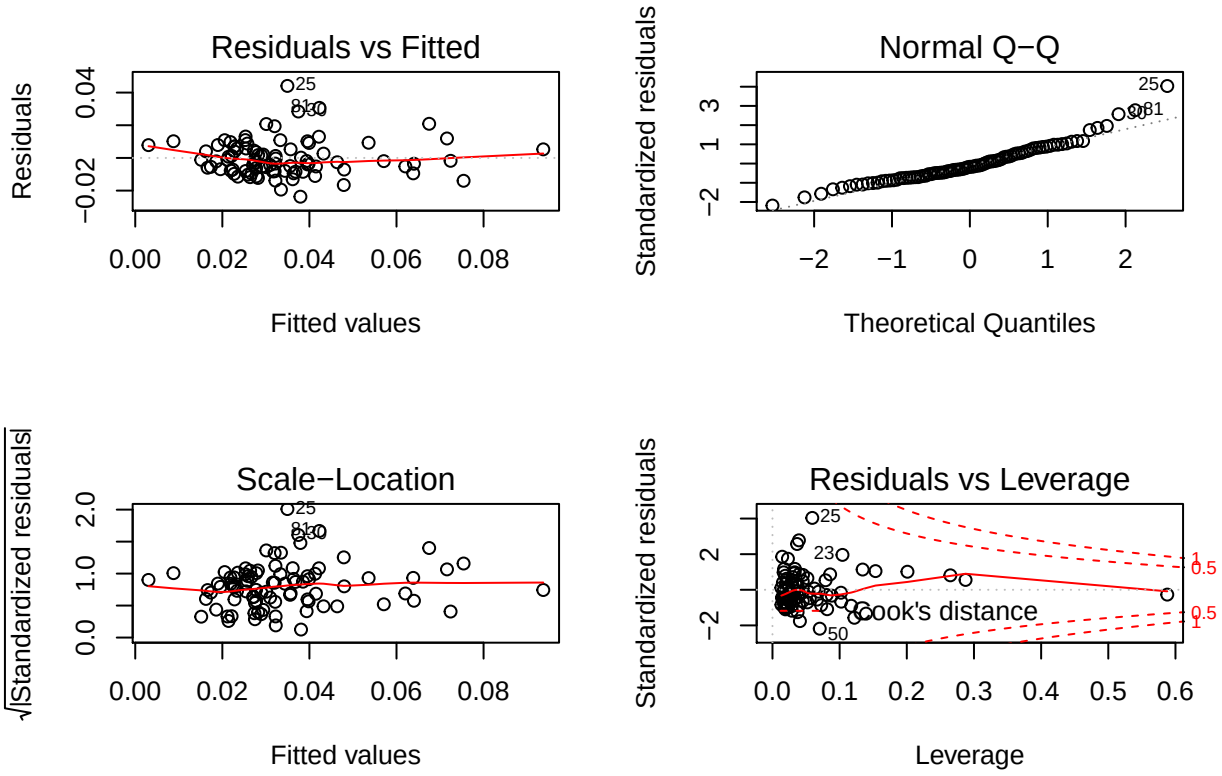
Model 1 Without Log Transformation of Crime Rate

This section shows the test for the classic linear model assumptions of the same variables in Model 1, but without taking the log of crime rate.

```
model1_nl <- lm(crmrte ~ density + prbarr + prbconv + pctymle , data = dt)
vif(model1_nl)
```

```
##  density   prbarr  prbconv  pctymle
## 1.278900 1.427889 1.324947 1.094245
```

```
par(mfrow=c(2,2))
plot(model1_nl)
```




```
cov(model1_nl$residuals, dt$prbarr)
```

```
## [1] -8.818106e-20
```

```
cov(model1_nl$residuals, dt$prbconv)
```

```
## [1] -2.405959e-19
```

```
cov(model1_nl$residuals, dt$density)
```

```
## [1] 4.632356e-19
```

```
cov(model1_nl$residuals, dt$pctymle)
```

```
## [1] 2.012079e-21
```

```
coeftest(model1_nl, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0376078  0.0111185  3.3824 0.001093 **
## density      0.0068907  0.0010491  6.5684 4.01e-09 ***
## prbarr      -0.0461780  0.0164195 -2.8124 0.006121 **
## prbconv     -0.0166265  0.0054239 -3.0654 0.002923 **
```

```
## pctymle      0.1003466  0.0503350  1.9936 0.049444 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared to the assumptions in Model 1 shown in the main document, the biggest difference the team noticed with this version that does not log transform the dependent variable, crime rate, is that it violates Assumption 6: Normality of Errors, as can be seen from the data points deviating from the line on the Q-Q plot. The remaining assumption tests showed similar results to Model 1.