W203 Lab 3: Reducing Crime

Hannah Morgan, Christina Papadimitriou, Sharad Varadarajan

1. Introduction

Our team was hired to provide research for a political campaign in North Carolina. Specifically, the objective of this study is to understand the determinants of crime and provide policy recommendations to local government.

The data source used in this study contains crime statistics for a selection of 90 counties in North Carolina in 1987, and was first used by Cornwell and Trumbull[1]. This study uses a cross-section of the original multi-year panel.

An Exploratory Data Analysis (EDA) and Ordinary Least Squares (OLS) Regression will be used to examine the data and explore correlational and potentially causal relationships between crime rate (outcome variable) and a set of explanatory variables from the following categories:

- certainty of punishment
- severity of punishment
- location
- population density and demographics of the counties
- taxes and weekly wages

Additionally, there will be some discussion of variables that could be essential in informing policy recommendations, but are omitted from the dataset.

2. Initial EDA

2.1 Dataset Description

The initial dataset contains 97 observations of 25 variables. The summary of variable descriptions and types, as well as a summary of the initial data set, are shown in Table 1 and Table 2. All of the data provided (besides the county identifier) is numerical in nature, and can be considered a potential model input. The next section takes a deeper look at the data quality and provides our steps taken to prepare the data set for our model building process.

```
crimes = read.csv("crime_v2.csv")
```

```
variable = c("county", "year", "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("county identifier", "1987", "crimes committed per person",
                "'probability' of arrest",
                "'probability' of conviction", "'probability' of prison sentence",
                "avg. sentence, days","police per capita", "people per sq. mile",
                "tax revenue per capita", "=1 if in western N.C.","=1 if in central N.C.",
                "=1 if in SMSA","perc. minority, 1980", "weekly wage, construction",
                "wkly wge, trns, util, commun", "wkly wge, whlesle, retail trade",
                "wkly wge, fin, ins, real est", "wkly wge, service industry",
                "wkly wge, manufacturing", "wkly wge, fed employees",
                "wkly wge, state employees", "wkly wge, local gov emps",
                "offense mix: face-to-face/other", "percent young male")
```

[1] C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data," Review of Economics and Statistics 76, 360-366.

```
type = c("categorical", "discrete", "continuous", "continuous", "continuous", "continuous",
              "continuous", "continuous", "continuous", "continuous", "binary",
              "binary", "binary", "continuous", "continuous",
              "continuous", "continuous", "continuous", "continuous", "continuous",
              "continuous", "continuous", "continuous", "continuous", "continuous")


table1 = data.frame(variable, description, type)

kable(table1, format = "latex", booktabs = T,
      caption = "Variable Descriptions and Types") %>%
  kable_styling(latex_options = c("striped", "hover", "hold_position", "condensed")) %>%
  column_spec(1, bold = T) %>%
  row_spec(0, bold = T)
```

Table 1: Variable Descriptions and Types

variable	description	type
county	county identifier	categorical
year	1987	discrete
crmrte	crimes committed per person	continuous
prbarr	'probability' of arrest	continuous
prbconv	'probability' of conviction	continuous
prbpris	'probability' of prison sentence	continuous
avgsen	avg. sentence, days	continuous
polpc	police per capita	continuous
density	people per sq. mile	continuous
taxpc	tax revenue per capita	continuous
west	=1 if in western N.C.	binary
central	=1 if in central N.C.	binary
urban	=1 if in SMSA	binary
pctmin80	perc. minority, 1980	continuous
wcon	weekly wage, construction	continuous
wtuc	wkly wge, trns, util, commun	continuous
wtrd	wkly wge, whlesle, retail trade	continuous
wfir	wkly wge, fin, ins, real est	continuous
wser	wkly wge, service industry	continuous
wmfg	wkly wge, manufacturing	continuous
wfed	wkly wge, fed employees	continuous
wsta	wkly wge, state employees	continuous
wloc	wkly wge, local gov emps	continuous
mix	offense mix: face-to-face/other	continuous
pctymle	percent young male	continuous

```
kable(t(summary(crimes)), format = "latex", booktabs = T,
      caption = "Data Summary") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down"))
```

Table 2: Data Summary

county	Min. : 1.0	1st Qu.: 52.0	Median :105.0	Mean :101.6	3rd Qu.:152.0	Max. :197.0	NA's :6
year	Min. :87	1st Qu.:87	Median :87	Mean :87	3rd Qu.:87	Max. :87	NA's :6
crmrte	Min. :0.005533	1st Qu.:0.020927	Median :0.029986	Mean :0.033400	3rd Qu.:0.039642	Max. :0.098966	NA's :6
prbarr	Min. :0.09277	1st Qu.:0.20568	Median :0.27095	Mean :0.29492	3rd Qu.:0.34438	Max. :1.09091	NA's :6
prbconv	: 5	0.588859022: 2	‘ : 1	0.068376102: 1	0.140350997: 1	0.154451996: 1	(Other) :86
prbpris	Min. :0.1500	1st Qu.:0.3648	Median :0.4234	Mean :0.4108	3rd Qu.:0.4568	Max. :0.6000	NA's :6
avgsen	Min. : 5.380	1st Qu.: 7.340	Median : 9.100	Mean : 9.647	3rd Qu.:11.420	Max. :20.700	NA's :6
polpc	Min. :0.000746	1st Qu.:0.001231	Median :0.001485	Mean :0.001702	3rd Qu.:0.001877	Max. :0.009054	NA's :6
density	Min. :0.00002	1st Qu.:0.54741	Median :0.96226	Mean :1.42884	3rd Qu.:1.56824	Max. :8.82765	NA's :6
taxpc	Min. : 25.69	1st Qu.: 30.66	Median : 34.87	Mean : 38.06	3rd Qu.: 40.95	Max. :119.76	NA's :6
west	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.2527	3rd Qu.:0.5000	Max. :1.0000	NA's :6
central	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.3736	3rd Qu.:1.0000	Max. :1.0000	NA's :6
urban	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.08791	3rd Qu.:0.00000	Max. :1.00000	NA's :6
pctmin80	Min. : 1.284	1st Qu.: 9.845	Median :24.312	Mean :25.495	3rd Qu.:38.142	Max. :64.348	NA's :6
wcon	Min. :193.6	1st Qu.:250.8	Median :281.4	Mean :285.4	3rd Qu.:314.8	Max. :436.8	NA's :6
wtuc	Min. :187.6	1st Qu.:374.6	Median :406.5	Mean :411.7	3rd Qu.:443.4	Max. :613.2	NA's :6
wtrd	Min. :154.2	1st Qu.:190.9	Median :203.0	Mean :211.6	3rd Qu.:225.1	Max. :354.7	NA's :6
wfir	Min. :170.9	1st Qu.:286.5	Median :317.3	Mean :322.1	3rd Qu.:345.4	Max. :509.5	NA's :6
wser	Min. : 133.0	1st Qu.: 229.7	Median : 253.2	Mean : 275.6	3rd Qu.: 280.5	Max. :2177.1	NA's :6
wmfg	Min. :157.4	1st Qu.:288.9	Median :320.2	Mean :335.6	3rd Qu.:359.6	Max. :646.9	NA's :6
wfed	Min. :326.1	1st Qu.:400.2	Median :449.8	Mean :442.9	3rd Qu.:478.0	Max. :598.0	NA's :6
wsta	Min. :258.3	1st Qu.:329.3	Median :357.7	Mean :357.5	3rd Qu.:382.6	Max. :499.6	NA's :6
wloc	Min. :239.2	1st Qu.:297.3	Median :308.1	Mean :312.7	3rd Qu.:329.2	Max. :388.1	NA's :6
mix	Min. :0.01961	1st Qu.:0.08074	Median :0.10186	Mean :0.12884	3rd Qu.:0.15175	Max. :0.46512	NA's :6
pctymle	Min. :0.06216	1st Qu.:0.07443	Median :0.07771	Mean :0.08396	3rd Qu.:0.08350	Max. :0.24871	NA's :6

2.2 Data Quality and Anomaly Detection

```
crimes['county'] = as.factor(crimes['county']) # fixing data type
crimes['prbconv'] = as.numeric(as.character(crimes['prbconv'])) # fixing data type
crimes = crimes[1:91,] # removing NAs
crimes = crimes[c(1:87,89:91),] # removing duplicate row
crimes = crimes[, c(1,3:25)] # removing column year
crimes$pctmin80 = crimes$pctmin80/100 # scaling variable for consistency
```

The Data Summary table in the previous section served as a starting point for the data cleansing process. The steps taken are outlined below:

1. There were 6 “NA” values reported for each variable; after visual inspection, the team noticed that this was due to 6 empty rows being included as observations. Since no data was included in these 6 rows, the team removed these empty observations. The dataset now includes a total of 91 observations.
2. The structure of the dataset revealed the county identifier was being stored as a vector of integers. Since this variable inherently represents a unique numeric string, rather than a number, the research team converted the data type from integer to factor.
3. The updated structure of the dataset showed that there were 90 unique counties in our dataset, though there are a total of 91 observations. It was discovered that there were two rows for county # 193, with the same data in each of the two rows. The team removed the duplicate row, which left us with a new total of 90 observations in the dataset.
4. The probability of conviction (*prbconv*) was revealed to be a factor variable. Since this variable stores a vector of probability values, the research team converted prbconv to a vector of numerical values
5. There are values greater than 1 for two of our probability variables:
   (a) One observation in the dataset had a Probability of Arrest (*prbarr*) greater than 1; county # 115 has a probability of arrest of 1.09091. Being that this column is proxied by ratio of arrests to offenses, we

initially considered whether or not there can be more arrests than there are offenses. Absent a more clear definition of "offenses", the team decided that there is not enough evidence to categorize this data point as erroneous. We take as an example a group of criminals robbing a bank and getting arrested; in this scenario, multiple arrests stem from one group offense. Therefore, we left the data point as is.

(b) 10 observations of *prbconv* are greater than 1. While initially concerning, the team recognized that it is possible for one arrest to lead to multiple convictions. Therefore, the research team did not remove these values because it is possible to have a prbconv greater than 1.

6. The research team noticed that county # 71 is the only observation in the dataset that has a value of 1 for both West and Central. While the team notes this as a possible anomaly, it is possible for county # 71 to be located on the border of Western and Central North Carolina. Therefore, the team left the observation as is.
7. The year column in the dataset contained the same value (87) for all rows. Since there is no variability in this data column, the team excluded it from analysis and simply noted that all of the observations refer to values from 1987. 24 meaningful variables now remain in the dataset.
8. The research team noticed that the percent minority (*pctmin80*) variable is expressed in a different scale (0-100) than the percent male (*pctymle*) variable (0-1). Therefore, the team decided to divide *pctmin80* by 100 in order to maintain consistent units across the two variables.

The result of the data changes is a final data set with 90 observations and 24 variables, and will be the basis for the subsequent model building process.

3. Model Building Process

3.1 Outcome Variable

The dependent (outcome) variable in this study is the crimes committed per person - i.e. crime rate (*crmrte*). The histogram of this variable (Figure 1), shows that its distribution has a strong positive skew, thus we consider taking the log transformation of this variable. Figure 2 is the histogram of the log of *crmrte*, and it shows that the log variable has a relatively normal distribution. Additionally, the log transformation gives a more intuitive sense to the regression results, since we will be looking at a percent change in crime rate in response to a (unit or percent) change in the explanatory variables and covariates. It is important to note that in the dataset *crmrte* is never less than or equal to zero, which means the log transformation will not result in a loss of data. Finally, the log transformation of the dependent variable is beneficial in ensuring that the residuals of the model are closer to having a normal distribution (i.e. Classical Linear Model Assumption 6 - Normality of errors). Therefore, the research team decided to use the log of crime rate (*log(crmrte)*) as the outcome variable for all the models that will be developed in this study.

```
par(mfrow=c(1,2))
hist(crimes$crmrte, breaks = 10, main = "Figure 1. Histogram of Crime Rate",
    cex.main=0.8, cex.lab=0.7, col = "gainsboro",
     yaxt = "n", xaxt = "n", xlab = "crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)

hist(log(crimes$crmrte), breaks = 10, main = "Figure 2. Histogram of Log of Crime Rate",
    cex.main=0.8, cex.lab=0.7, col = "gainsboro",
     yaxt = "n", xaxt = "n", xlab = "log crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
```

Figure 1. Histogram of Crime Rate

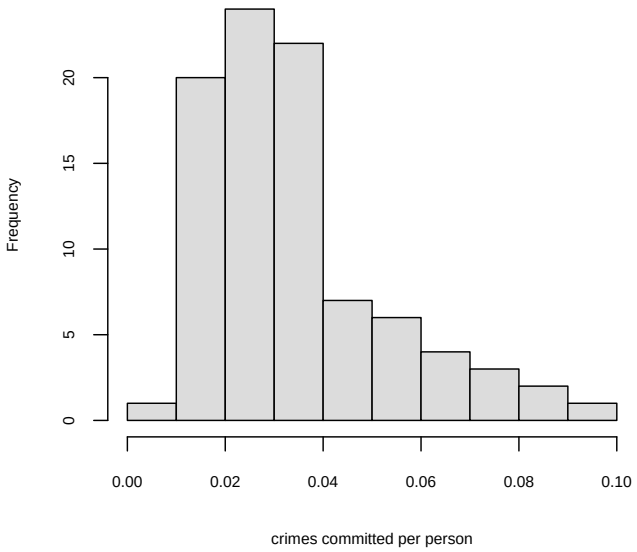
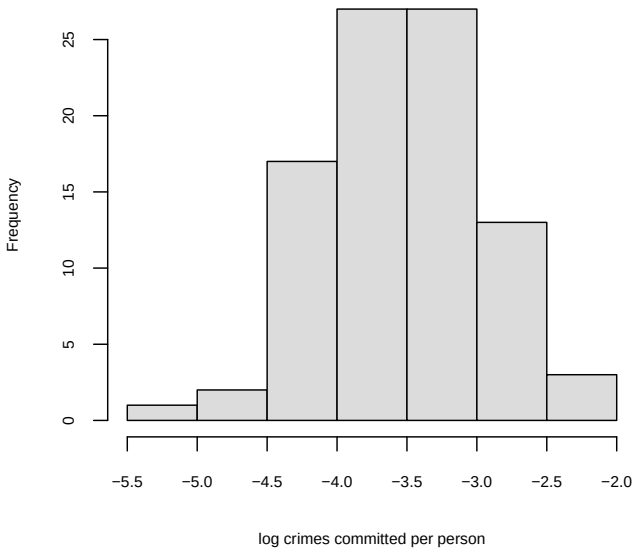


Figure 2. Histogram of Log of Crime Rate

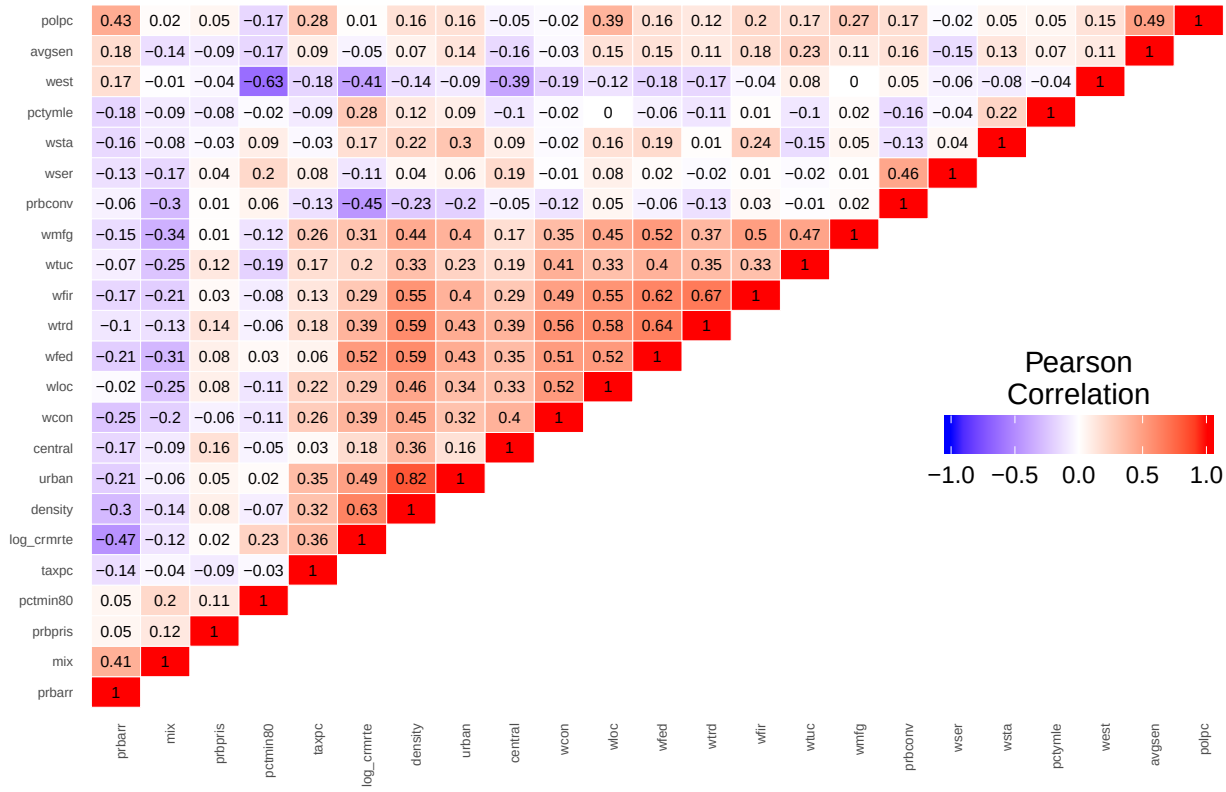


3.2 Correlations

```
crimes_t = crimes
crimes_t$crmrte = log(crimes_t$crmrte)
colnames(crimes_t)[2] <- "log_crmrte"
corr_matrix = cor(crimes_t[,sapply(crimes_t, function(x) !is.factor(x))], use = "pairwise.complete.obs")
get_upper_tri<-function(cormat){
  cormat[lower.tri(cormat)] <- NA
  return(cormat)
}
reorder_cormat <- function(cormat){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}
corr_matrix_3 = reorder_cormat(corr_matrix)
upper_tri <- get_upper_tri(corr_matrix_3)
melted_price_pen_mat = melt(upper_tri, na.rm = TRUE)
ggplot(data = melted_price_pen_mat, aes(x=Var1, y= Var2, fill= round(value,2))) +
  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_minimal() +  theme(axis.text.x = element_text(angle = 90, hjust = 1, size=5),
                           axis.text.y = element_text(size = 5))  +
  geom_text(aes(Var1, Var2, label = round(value,2)), color = "black", size = 2) +
  ggtitle('Figure 3: Correlation Heatmap')  +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
```

```
    legend.justification = c(1, 0),
    legend.position = c(1, 0.3),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```

Figure 3: Correlation Heatmap



After deciding on what the outcome variable would be, the next step is to efficiently quantify the bivariate correlations that exist between the log-transformed crime rate variable and the other variables. To do this, we generated the correlation heat map (Figure 3) above in order to pinpoint variables of interest for the regression models.

The strongest positive correlations for our log-transformed crime rate variable include:

- Log(*crmrte*) vs. *density* (0.63)
- Log(*crmrte*) vs. *wfed* (0.52)
- Log(*crmrte*) vs. *urban* (0.49)

The strongest negative correlations for our log-transformed crime rate variable include:

- Log(*crmrte*) vs. *prbarr* (-0.47)
- Log(*crmrte*) vs. *prbconv* (-0.45)
- Log(*crmrte*) vs. *west* (-0.41)

Section 3.3 takes a closer look at these 6 relationships, with the goal of determining the explanatory variables for the first model.

3.3 Analysis of Key Variables

Probability of Arrest (prbarr)

The variable *prbarr* has a slightly positively skewed distribution (see Figure 4), with a noticeable outlier for county 115, which has a probability of arrest of 1.09. It has a relatively strong negative correlation of (-0.47) with the outcome variable. This is shown in the scatterplot below (Figure 5). Along with this noticeably negative relationship between the two variables, the research team also considered the importance that 'certainty of punishment' (do criminals expect to get caught and face punishment?) may have in influencing the mindset of a criminal. Therefore, *prbarr* was chosen as one of the explanatory variables.

```
cor(crimes$prbarr, log(crimes$crmrte))
```

```
## [1] -0.4727669
```

```
par(mfrow=c(1,2))
hist(crimes$prbarr, breaks = 10, main = "Figure 4. Histogram of Probability of Arrest",
    cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "probability of arrest")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
plot(crimes$prbarr, log(crimes$crmrte), main = "Figure 5. Log of Crime Rate vs. Prob. of Arrest",
     cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "probability of arrest",
     ylab = "log crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
model = lm(log(crmrte) ~ prbarr, data = crimes)
abline(model)
```

Figure 4. Histogram of Probability of Arrest

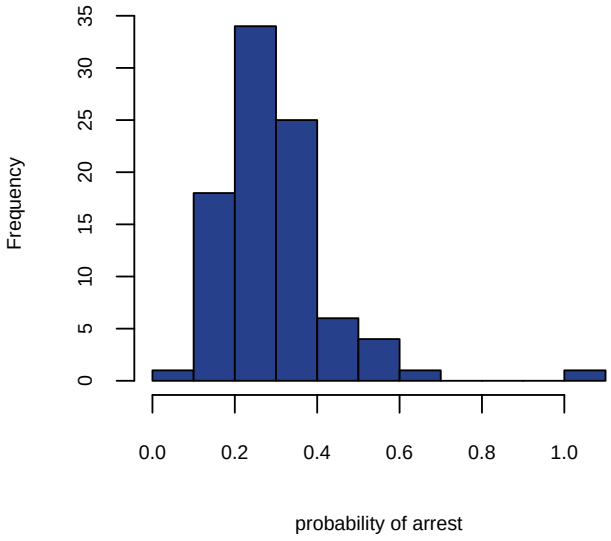
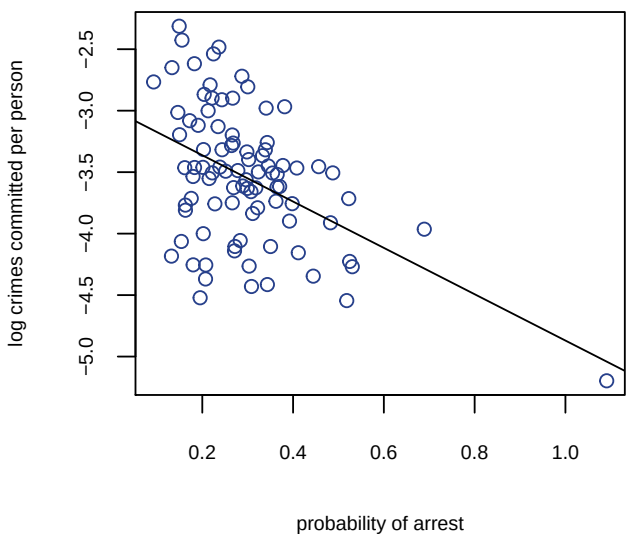


Figure 5. Log of Crime Rate vs. Prob. of Arrest



Probability of conviction (prbconv)

Probability of conviction has a more positively skewed distribution compared to probability of arrest (see Figure 6); this is influenced by 10 counties in the final dataset with *prbconv* values greater than 1. Because of the stronger skew, a log transform on this variable was considered; however, that actually reduced the strength of the correlation (-0.37 compared to -0.45 without the transform). Therefore, this variable remains untransformed. The strength of relationship and nature of this variable also highlights the importance that 'certainty of punishment' may have in influencing both the mindset of a criminal, and policy-making decisions to reduce crime rate. Therefore, the team will use *prbconv* as one of the explanatory variables.

```
cor(crimes$prbconv, log(crimes$crmrte))
```

```
## [1] -0.4468136
```

```r
par(mfrow=c(1,2))
hist(crimes$prbconv, breaks = 10, main = "Figure 6. Histogram of Probability of Conviction",
    cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "probability of conviction")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
plot(crimes$prbconv, log(crimes$crmrte), main = "Figure 7. Log of Crime Rate vs. Prob. of Conviction",
     cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "probability of conviction",
     ylab = "log crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
model = lm(log(crmrte) ~ prbconv, data = crimes)
abline(model)
```

Figure 6. Histogram of Probability of Conviction

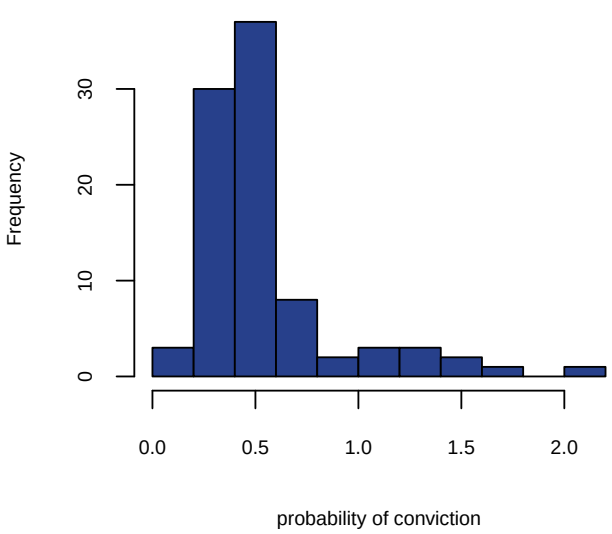
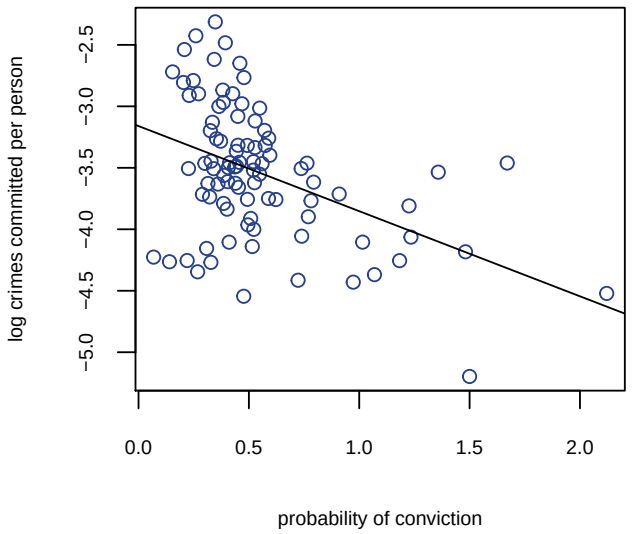


Figure 7. Log of Crime Rate vs. Prob. of Conviction



Binary location variables (urban, west)

Next, the research team explored the two binary location variables that had relatively strong correlations with our outcome variable: whether the county is more urban (*urban*) and whether the county is in the western part of the state (*west*). Our first boxplot in Figure 8 illustrates the distribution of the outcome variable categorized by urban vs. rural counties in North Carolina.

Looking at Figure 8, we observe that the smallest crime rate value for an urban county in the dataset is still greater than crime rates for over 75% of rural counties. However, it is important to note that there are only 8 observations for urban counties (compared to the 82 observations for rural counties). In addition to the small number of urban counties, the team believes the *density* variable provides a more precise measurement for how urban a county is (by measuring people per sq. mile). Therefore, *urban* was not chosen as one of the key explanatory variables.

The boxplot in Figure 9 shows the distribution of the outcome variable, partitioned by West vs Non-West. This boxplot shows that the median crime rate in non-western counties is greater than the median crime rate in western counties. About 25% of observations in the data set are considered western counties ($west = 1$). Because of the non-negligible number of observations that are classified as western, the strong negative correlation (-0.41), and the impact that location based policies could have on crime rate, the team chose to include this variable in the model as an explanatory variable.

```r
par(mfrow=c(1,2))
boxplot(log(crimes$crmrte) ~ crimes$urban, col="gainsboro",
        cex.main=0.7, cex.lab=0.7, yaxt = "n",
```

```
        main="Figure 8. Boxplot of Log Crime Rate by Urban/Rural", staplewex=1,
        ylab="log crimes committed per person",
        names=c("Rural", "Urban"))
axis(2, cex.axis = 0.7)
boxplot(log(crimes$crmrte) ~ crimes$west, col="gainsboro",
        cex.main=0.7, cex.lab=0.7, yaxt = "n",
        main="Figure 9. Boxplot of Log of Crime Rate by West/non-West", staplewex=1,
        ylab="log crimes committed per person",
        names=c("non-West", "West"))
axis(2, cex.axis = 0.7)
```

Figure 8. Boxplot of Log Crime Rate by Urban/Rural

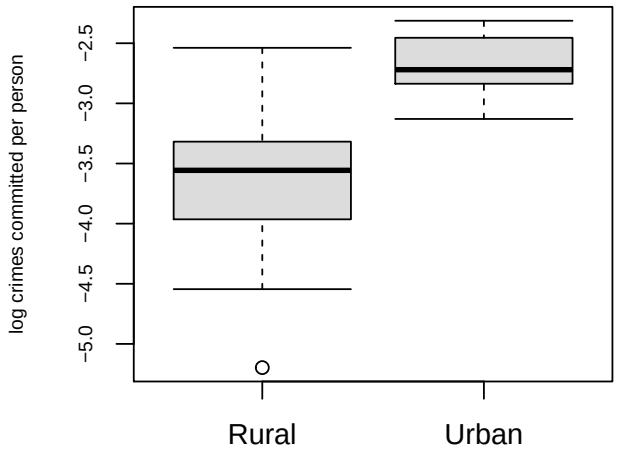
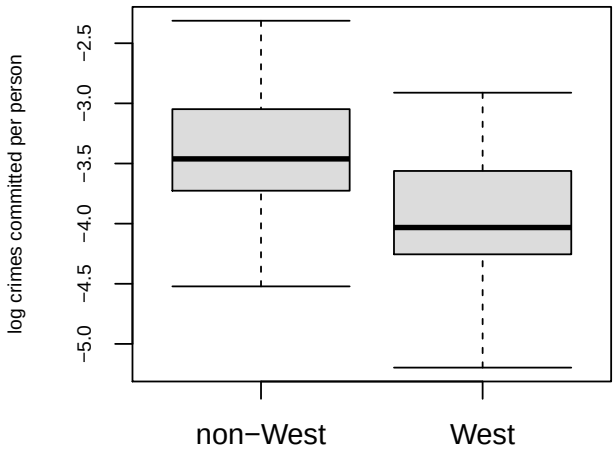


Figure 9. Boxplot of Log of Crime Rate by West/non−West




```
nrow(crimes[(crimes$urban ==1),])
```

```
## [1] 8
```

```
nrow(crimes[(crimes$west ==1),])
```

```
## [1] 22
```

People per sq. mile (density)

Next the team evaluated *density*, which had the highest correlation with the outcome variable (0.63). Looking at Figure 10 below, the distribution for density shows a very positive skew. Therefore, the team decided to explore whether or not a log transformation for density would be beneficial. After taking the log of density, Figure 11 shows that there is no longer a positive skew, but the transformation introduced a rather extreme outlier due to a very small population density for county # 173. The correlation with the outcome variable also decreased from 0.63 to 0.49. While discouraging, the team wanted to explore just how influential that outlier was, and how much it was impacting the correlation between the outcome variable and log of density.

From the residuals vs. leverage plot in Figure 16, the outlier for county # 173 has a Cook's distance considerably greater than 1. When comparing the outcome variable to the log of density, and excluding county # 173's observation, the correlation increases from 0.49 to 0.68, which is higher than the correlation we were seeing before transforming the density variable. This can be better visualized in Figure 15. Unfortunately, just because county # 173's value is an outlier does not mean the data point is erroneous. Therefore, the team does not have enough evidence to exclude this data point from its analysis and will not be applying a transformation to the density variable.

Even though the team will not be applying a transformation, the strong correlation between the outcome variable and the non-transformed *density* is enough to justify its use as an explanatory variable.

```
crimes2 = subset(crimes, log(density) > -10)
cor(crimes$density, log(crimes$crmrte))
```

```
## [1] 0.6330234
```

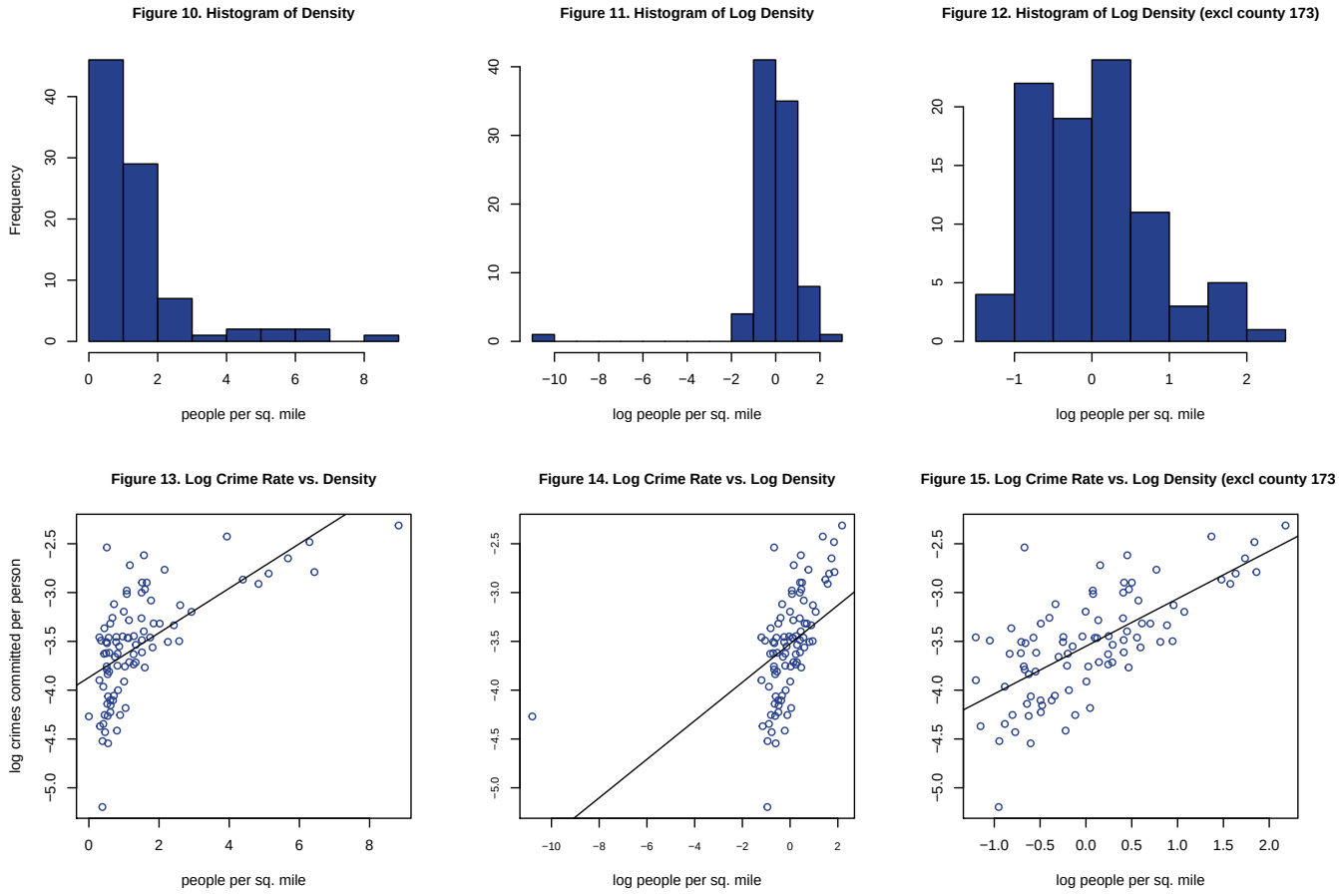
```
cor(log(crimes$density), log(crimes$crmrte))
```

```
## [1] 0.4936425
```

```
cor(log(crimes2$density), log(crimes2$crmrte))
```

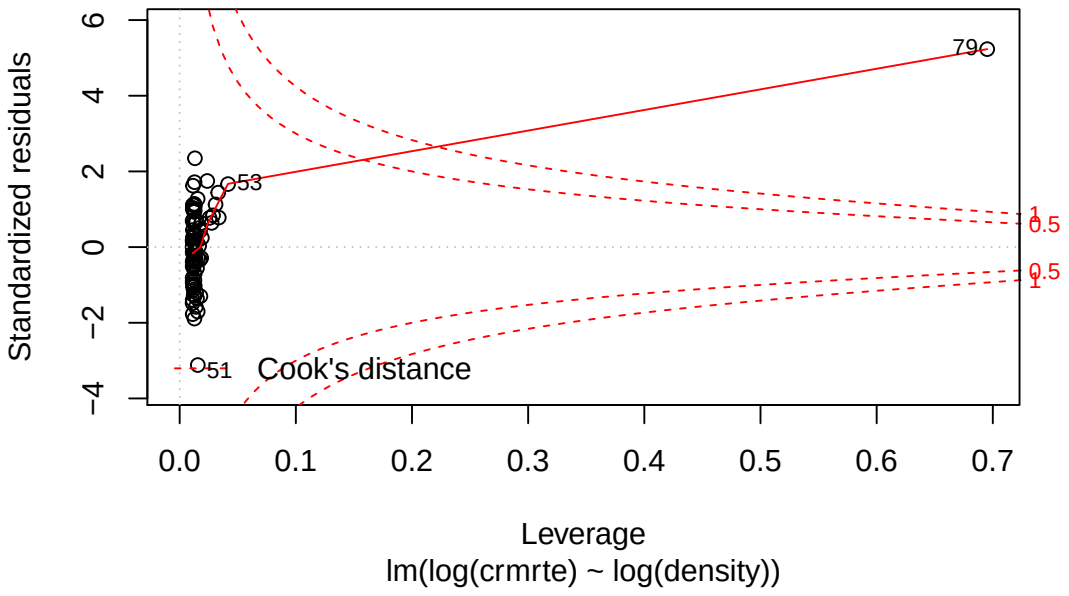
```
## [1] 0.6843266
```

```
par(mfrow=c(2,3))
hist(crimes$density, breaks = 10, main = "Figure 10. Histogram of Density",
    cex.main=1, cex.lab=1, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "people per sq. mile")
axis(2, cex.axis = 1)
axis(1, cex.axis = 1)
hist(log(crimes$density), breaks = 10, main = "Figure 11. Histogram of Log Density",
    cex.main=1, cex.lab=1, col = "royalblue4", ylab = " ",
     yaxt = "n", xaxt = "n", xlab = "log people per sq. mile")
axis(2, cex.axis = 1)
axis(1, cex.axis = 1)
hist(log(crimes2$density), breaks = 10,
     main = "Figure 12. Histogram of Log Density (excl county 173)",
    cex.main=1, cex.lab=1, col = "royalblue4", ylab = " ",
     yaxt = "n", xaxt = "n", xlab = "log people per sq. mile")
axis(2, cex.axis = 1)
axis(1, cex.axis = 1)
plot(crimes$density, log(crimes$crmrte), main = "Figure 13. Log Crime Rate vs. Density",
     cex.main=1, cex.lab=1, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "people per sq. mile",
     ylab = "log crimes committed per person")
axis(2, cex.axis = 1)
axis(1, cex.axis = 1)
model_d1 = lm(log(crmrte) ~ (density), data = crimes)
abline(model_d1)
plot(log(crimes$density), log(crimes$crmrte), main = "Figure 14. Log Crime Rate vs. Log Density",
     cex.main=1, cex.lab=1, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "log people per sq. mile",
     ylab = " ")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
model_d2 = lm(log(crmrte) ~ log(density), data = crimes)
abline(model_d2)
plot(log(crimes2$density), log(crimes2$crmrte),
     main = "Figure 15. Log Crime Rate vs. Log Density (excl county 173)",
     cex.main=1, cex.lab=1, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "log people per sq. mile",
     ylab = " ")
axis(2, cex.axis = 1)
axis(1, cex.axis = 1)
model_d3 = lm(log(crmrte) ~ log(density), data = crimes2)
abline(model_d3)
```




```
plot(model_d2, which=5, caption= "n", main="Figure 16. Residuals vs Leverage")
```

Figure 16. Residuals vs Leverage



Weekly Federal Wages (wfed)

Figure 17 shows that the distribution of weekly federal wages (*wfed*) is approximately normal. Figure 18 illustrates the relationship between weekly federal wages and the outcome variable. The scatterplot indicates that there is a positive linear relationship between the two variables, which is supported by a correlation value of 0.52. Even

though weekly federal wages has an approximately normal distribution and a high correlation with the outcome variable, the team decided not to use it as one of the explanatory variables. The reasoning behind this decision is that there is no intuitive explanation of why decreasing wages of federal employees would decrease crime rate, so this correlation could be the result of a confounding relationship. One potential explanation is that federal wages tend to be based upon cost of living, which would be higher in population dense areas. The correlation between weekly federal wages and density is 0.59 and, as noted earlier, density is also highly correlated with the outcome variable.

```
cor(crimes$wfed, log(crimes$crmrte))
```

```
## [1] 0.5233058
```

```
par(mfrow=c(1,2))

hist(crimes$wfed, breaks = 10, main = "Figure 17. Histogram of Weekly Federal Wages",
    cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "wkly wge, fed employees")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)

plot(crimes$wfed, log(crimes$crmrte), main = "Figure 18. Log of Crime Rate vs. Weekly Federal Wages",
     cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "wkly wge, fed employees",
     ylab = "log crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)

model = lm(log(crmrte) ~ wfed, data = crimes)
abline(model)
```

Figure 17. Histogram of Weekly Federal Wages

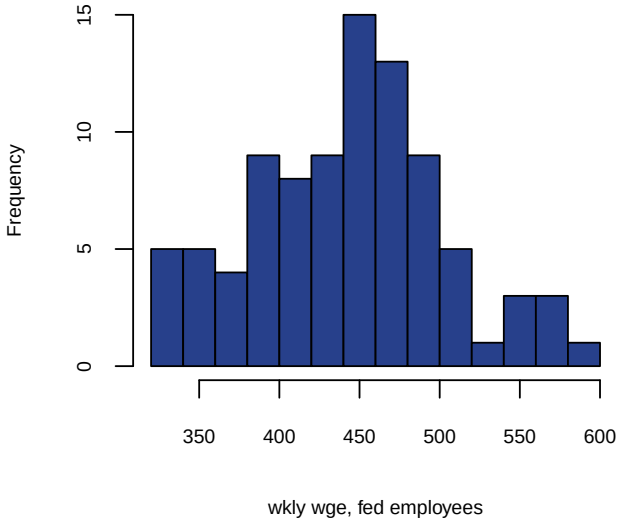
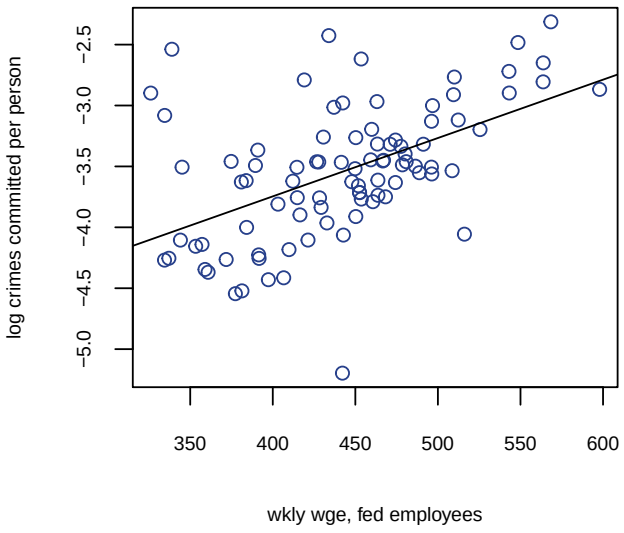


Figure 18. Log of Crime Rate vs. Weekly Federal Wages



3.4 Model Specifications

Model 1: Key Explanatory Variables

In the exploratory analysis (section 3.3), the team identified the key explanatory variables that have strong positive and negative correlations with the outcome variable (log of *crmrte*). The first model of this study will include those explanatory variables that have high correlations with the outcome variable and could be seen as the most important

determinants of crime rate.

First, *prbarr* and *prbconv* will be included in model 1 because increasing the certainty of punishment, proxied by the probabilities of arrest and conviction, is believed to potentially decrease crime rate. Additionally, these two variables can potentially be influenced by political action. Second, areas with higher population density are believed to have higher crime rates, thus the team included *density* in model 1 as well. Lastly, model 1 includes the *west* variable, since it has a relatively strong negative correlation (-0.41) with the outcome variable, and if *west* is proven to affect crime rate, location-based policies could be suggested to decrease crime.

Below is the regression equation for model 1:

$$log(CrimeRate) = \beta_0 + \beta_1 probArrest + \beta_2 probConviction + \beta_3 density + \beta_4 west + u$$

The parameters in the above equation were estimated using OLS regression, and the results will be further explained in section 3.6.

Model 2: More refined and robust than model 1

For the second model, the team set out to improve upon their initial model. The goal was to determine covariates that would improve the accuracy of the model without introducing substantial bias or fully sacrificing parsimony. The two variables that were chosen to be added to the model were police per capita (*polpc*) and the percent minority (*pctmin80*). Our reasons for choosing these variables will be discussed below.

Police per Capita (polpc)

The correlation between *polpc* and the outcome variable is not strong (0.01), so it is not an obvious choice to be considered as a covariate. However, intuitively it makes sense that more or less law enforcement would have an impact on crime rate, so the team chose to further evaluate this variable. Figure 19 shows that the data is noticeably right skewed and Figure 20 illustrates the lack of a linear relationship between the outcome variable and *polpc*; this indicates that a log transform may be necessary. Figure 21 shows the result of that log transform on *polpc*, which depicts a more bell-shaped curve. Also, from Figure 22, the team notices that the log transform increases the correlation with the outcome variable to 0.28, which is much more significant. However, the sign of the impact is contrary to what the team anticipated; the positive relationship between police per capita and crime rate observed is opposite to expected intuition, which is that increasing police per capita would decrease crime rate over time. At first the team was concerned whether there was an omitted variable present, related both to police per capita and the outcome variable, that was influencing this correlation. However, since this is not a time series dataset the team believes that the over-time negative relationship between these two variables is not captured. Therefore, based on the improved model fit of adding this variable into the model (R-squared increase of ~10%), the team included the log of *polpc* in model 2, since it helps explain a lot of the variation in the outcome variable.

```
par(mfrow=c(2,2))
hist(crimes$polpc, breaks = 10, main = "Figure 19. Histogram of Police per Capita",
    cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "police per capita")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
plot(crimes$polpc, log(crimes$crmrte), main = "Figure 20. Log of Crime Rate vs. Police per Capita",
     cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "police per capita",
     ylab = "log crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
model = lm(log(crmrte) ~ polpc, data = crimes)
abline(model)

hist(log(crimes$polpc), breaks = 10, main = "Figure 21. Histogram of Log of Police per Capita",
    cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "log of police per capita")
axis(2, cex.axis = 0.7)
```

```
axis(1, cex.axis = 0.7)
plot(log(crimes$polpc), log(crimes$crmrte), main = "Figure 22. Log of Crime Rate vs. Log of Police per Cap
     cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "log of police per capita",
     ylab = "log crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
model = lm(log(crmrte) ~ log(polpc), data = crimes)
abline(model)
```

Figure 19. Histogram of Police per Capita

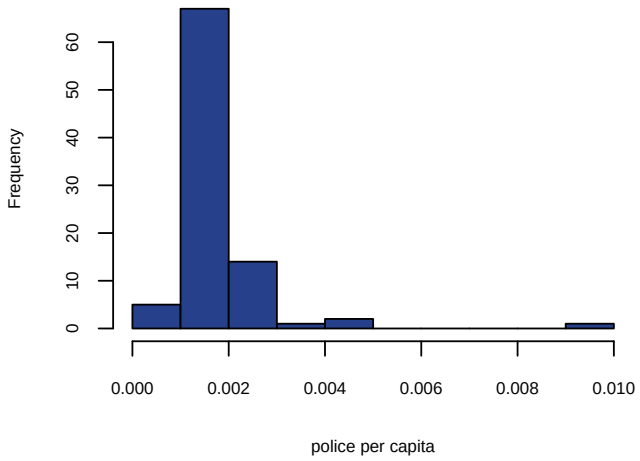


Figure 20. Log of Crime Rate vs. Police per Capita

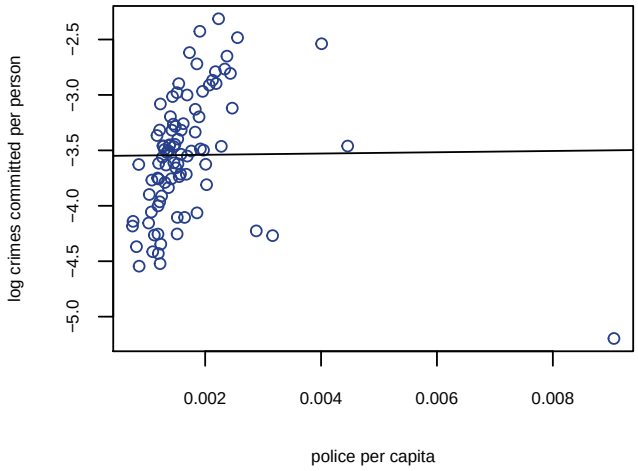


Figure 21. Histogram of Log of Police per Capita

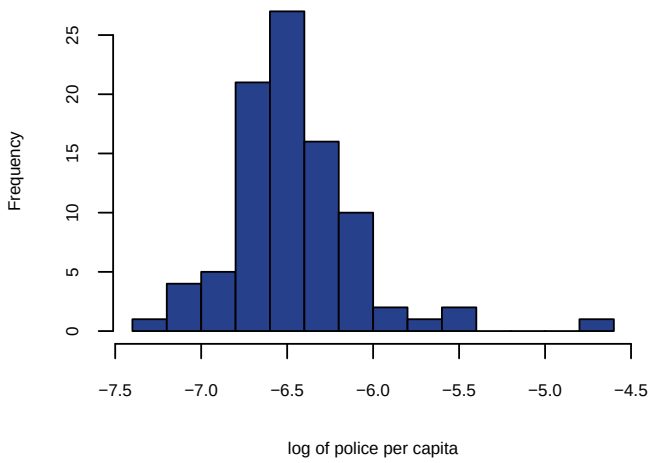
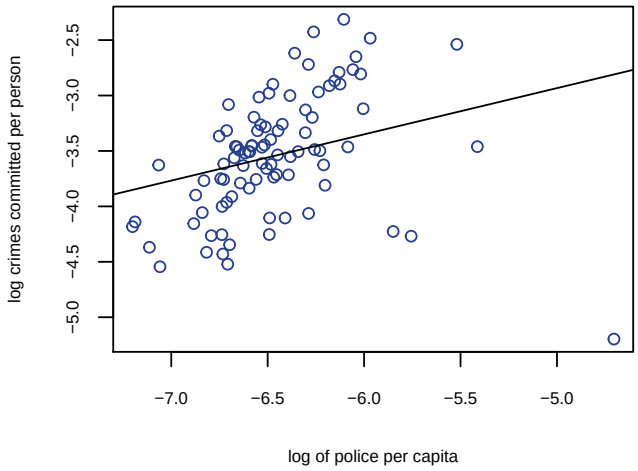


Figure 22. Log of Crime Rate vs. Log of Police per Capita



Percent Minority 1980 (pctmin80).

While not one of the strongest correlations, *pctmin80* does have a correlation of 0.23 with the outcome variable, which is not negligible. Also, this correlation is the second strongest in the entire dataset for *pctmin80*, which points to the possibility of less multicollinearity between pctmin80 and the other variables in our regression equation. Figure 23 shows the histogram of the variable, which exhibits a slight left skew but nothing too strong. Figure 24 further demonstrates the relationship between *pctmin80* and the outcome variable. As a result, this variable was not transformed before it's use in the second model. In addition to these findings, the research team was motivated to introduce this variable as a covariate due to nationwide trends in racial profiling.

```r
cor(crimes$pctmin80, log(crimes$crmrte))
```

```
## [1] 0.2329182
```

```r
par(mfrow=c(1,2))
hist(crimes$pctmin80, breaks = 10, main = "Figure 23. Histogram of Perc. minority, 1980",
    cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "percent minority")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
plot(crimes$pctmin80, crimes$crmrte, main = "Figure 24. Crime Rate vs. Perc. minority, 1980",
     cex.main=0.7, cex.lab=0.7, col = "royalblue4",
     yaxt = "n", xaxt = "n", xlab = "percent minority",
     ylab = "crimes committed per person")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
model = lm(crmrte ~ pctmin80, data = crimes)
abline(model)
```

Figure 23. Histogram of Perc. minority, 1980

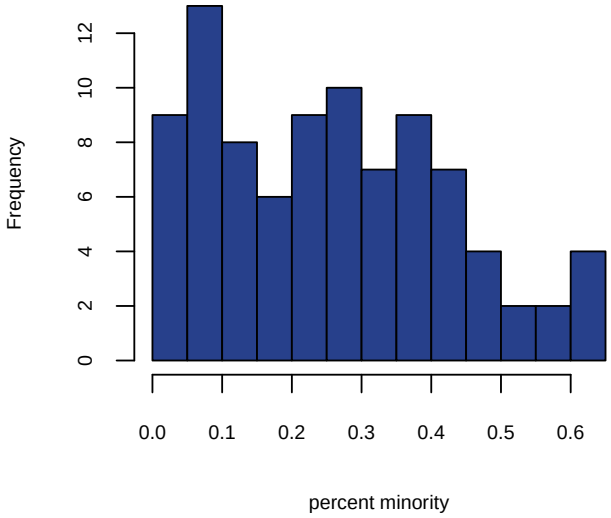
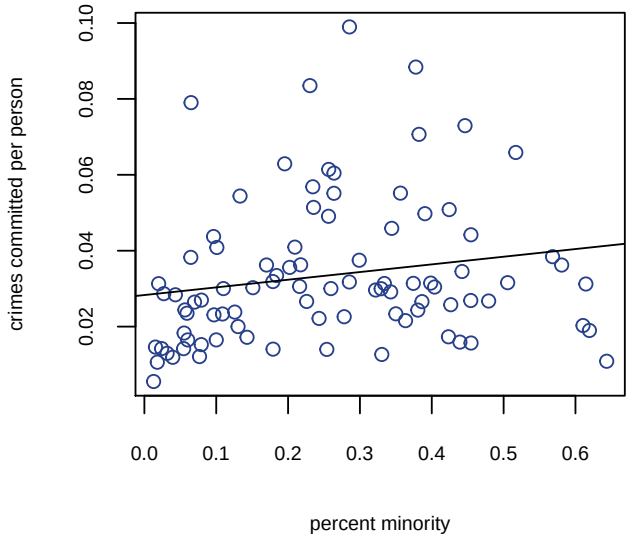


Figure 24. Crime Rate vs. Perc. minority, 1980



While there were higher correlations between other potential covariates and the outcome variable, namely the wage variables, we believe the two variables above introduce less bias to the model. The team chose not to include any of the wage variables in the second model because, in addition to losing parsimony due to the many wage variables, we viewed these weekly wages as potential outcome variables of *density*.

The team noted when running the regression for model 2, which includes both explanatory variables and important covariates, that the *west* variable was no longer statistically significant. Therefore, *west* was removed from model 2. A detailed write-up on statistical significance and removal of *west* from model 2 is available later in the report.

Below is the regression equation for model 2:

$$log(CrimeRate) = \beta_0 + \beta_1 probArrest + \beta_2 probConviction + \beta_3 density + \beta_4 log(policepc) + \beta_5 pctminority + u$$

The parameters in the above equation were estimated using OLS regression, and the results will be further explained in section 3.6.

Model 3: Including almost all variables

For the final model, the team chose to include all of the variables that made it to the final data set, minus the county identifier. The only log-transformed independent variable in this model is *polpc*. This was done in order to see what the highest possible R-squared value could be in relation to the outcome variable, regardless of parsimony. The final R-squared value for the model is 0.863.

The parameters of this model were estimated using OLS regression, and the results will be further explained in section 3.6.

```
model1 = lm(log(crmrte) ~  prbarr + prbconv + density + west, data = crimes)
model2 = lm(log(crmrte) ~  prbarr + prbconv + density + log(polpc) + pctmin80, data = crimes)
model3 = lm(log(crmrte) ~ . -county -polpc + log(polpc), data = crimes)
```

3.5 Assessment of Classical Linear Model (CLM) Assumptions

```
par(mfrow=c(2,2))
plot(model2,which=1, caption ="", main="Figure 25. Residuals vs Fitted")
plot(model2,which=2, caption ="", main="Figure 26. Normal Q-Q")
plot(model2,which=3, caption ="", main="Figure 27. Scale-Location")
plot(model2,which=5, caption ="", main="Figure 28. Residuals vs Leverage")
```

Figure 25. Residuals vs Fitted

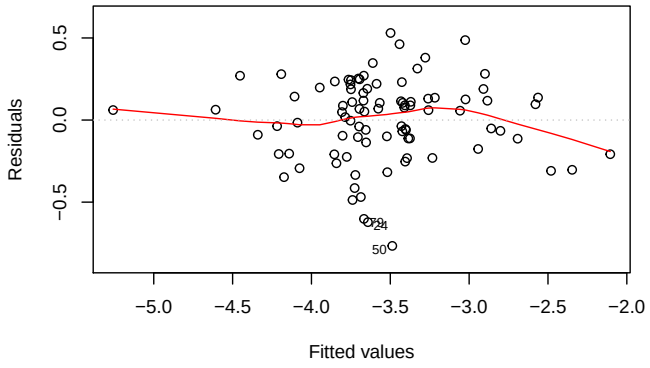


Figure 26. Normal Q−Q

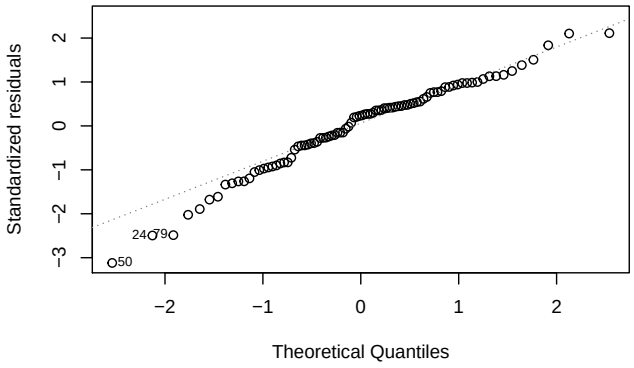


Figure 27. Scale−Location

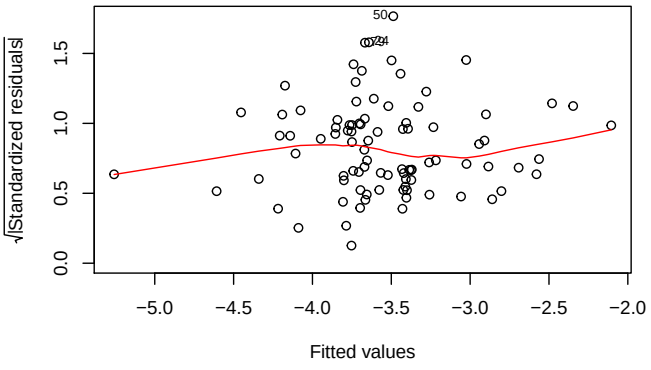
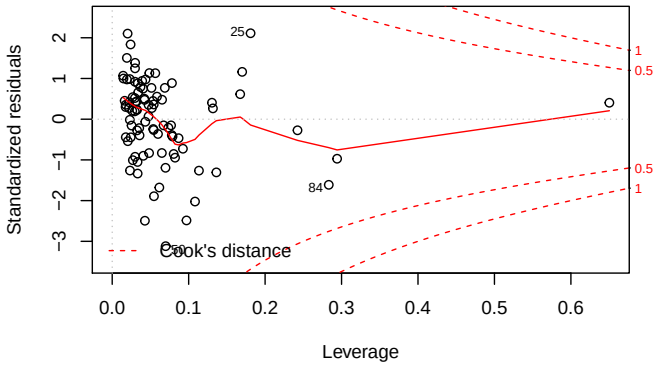


Figure 28. Residuals vs Leverage



While the team was very pleased with the three model specifications, it is crucial to ensure that the models adhere to the Classical Linear Model assumptions; therefore the team proceeded to evaluate the model specifications against the 6 CLM assumptions. The 6 CLM assumptions will be assessed in detail for the most important model specification (model 2), but the team will also highlight major surprises noticed when assessing models 1 and 3. The

team will discuss whether the CLM assumptions are violated and respond to any violations. The diagnostic plots for model 2 (Figures 25 - 28) will be used along with other diagnostic tools to assess the 6 CLM Assumptions.

CLM 1 - A Linear Model

The first CLM assumption states that the population model is linear in parameters (betas). All three models examined in this study are specified such that the dependent variable is a linear function of the independent variables. Therefore, CLM 1 is valid by design for the three models and no further response is required.

CLM 2 - Random Sampling

The second CLM assumption states that the data is a random sample drawn from the population. The dataset used in this study contains 90 observations and each represents a county in North Carolina. Since North Carolina consists of 100 counties, the team noted that the dataset does not contain all the counties. The team did not identify any patterns in the data that would indicate a violation of the random sampling assumption; as long as the scope of observations and results from the model building are limited to application in North Carolina, there is no clustering effect for only using data from North Carolina. With this being said, the team does not have enough information about the data collection process, nor any information on crime rate metrics concerning the 10 excluded counties, to determine if the assumption of random sampling is violated in the selection of only 90 counties worth of data. Therefore, the team assumes that the random sampling assumption holds.

CLM 3 - No Perfect Multicollinearity

The No Perfect Multicollinearity assumption states that none of the independent variables can be constant, and that there can be no exact linear relationship among the independent variables. When evaluating the models in R, the team would have encountered an error if perfect multicollinearity existed in any of models, and therefore would not have been able to run the regressions. Therefore, we can be sure that there is no perfect multicollinearity in any of the model specifications. While this assumption only prohibits perfect multicollinearity, the presence of imperfect multicollinearity can still adversely affect the regression results; imperfect multicollinearity lowers precision and increases the standard error of the estimators. Therefore the team decided to check the variance inflation factor (VIF) of each of estimators.

The VIF results for model 2 can be seen below:

```
vif(model2)
```

```
##     prbarr    prbconv    density log(polpc)   pctmin80
##   1.301423   1.096809   1.454374   1.326523   1.034103
```

All of the variance inflation factors are considerably less than 4, which are not nearly high enough to suggest that corrections need to be made to model 2. The team observed similar results for model 1, with VIF values ranging between 1 and 1.2. For model 3, the team noticed there were two coefficients with VIF values greater than 4: density and urban. When deciding on the variables to include in models 1 and 2, the team chose not to include density and urban together in the same model, primarily due to the high correlation between the variables. The VIF values calculated by the team reinforce this earlier decision.

CLM 4 - Zero-Conditional Mean

The zero-conditional mean assumption states that the error u has an expected value of zero, given any values of the independent variables: $E(u|x_1, x_2, ..., x_k) = 0$. To check whether this assumption is violated, the team examines the plot of residuals against fitted values and pays close attention to the predicted conditional mean spline curve across fitted values. Figure 25 indicates little evidence that the zero-conditional mean assumption is violated. The red spline curve is fairly flat with the exception of a downturn on the right side of the plot, which is likely due to the smaller number of observations in that area.

Additionally, the team will test for the less strong condition of exogeneity. The covariances of the independent variables with the residuals are very close to zero indicating that the exogeneity assumption most likely holds.

```
cov(model2$residuals,crimes$prbarr, use = "pairwise.complete.obs")
```

```
## [1] 2.722235e-19
```

```
cov(model2$residuals,crimes$prbconv, use = "pairwise.complete.obs")
```

```
## [1] 1.259319e-18
```

```
cov(model2$residuals,crimes$density, use = "pairwise.complete.obs")
```

```
## [1] -3.47017e-18
```

```
cov(model2$residuals,log(crimes$polpc), use = "pairwise.complete.obs")
```

```
## [1] -1.108011e-17
```

```
cov(model2$residuals,crimes$pctmin80, use = "pairwise.complete.obs")
```

```
## [1] 2.301341e-18
```

The conclusion is that there is no major evidence that the zero-conditional mean assumption is violated in model 2. However, even if there was, given the large sample size (n>30), the team is confident that due to OLS asymptotics the coefficients are at least consistent. Therefore, no further action is required.

When assessing the zero-conditional mean assumption for model 1, the team observed a less flat shape of the red spline curve in the residuals versus fitted values plot when compared to model 2. However, when testing for the exogeneity assumption the covariances of the independent variables with the residuals are very close to zero, indicating that the exogeneity assumption holds and that the coefficients are at least consistent. Model 3 yields a relatively flat line around zero in the residuals versus fitted values plot, which is a strong indication that the zero-conditional mean assumption holds.

CLM 5 - Homoskedasticity

To check whether the homoskedasticity assumption holds (i.e. whether the variance of the residuals is constant) for model 2, the team examined the residuals versus fitted values plot (Figure 25) and the scale-location plot (Figure 27). In both plots, the observations don't fall within a band of constant thickness. The range of the residuals widens in the middle of the plots, which might be an indication of heteroskedasticity.

In addition to the diagnostic plots, the team performed the Breusch-Pagan statistical test to confirm the presence of heteroskedasticity. The null hypothesis is that there is homoskedasticity. The test yields a p-value larger than 0.05, which is not significant, and therefore, the team fails to reject the null hypothesis that there is homoskedasticity. However, even though the null hypothesis is rejected the team cannot conclude that the homoskedasticity assumption holds, especially because of the contradicting evidence that was observed in the diagnostic plots. Therefore, the team decides to follow a conservative approach to use heteroskedasticity-robust standard errors.

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 9.6696, df = 5, p-value = 0.08516
```

Additionally, when looking at the diagnostic plots for model 1 and 3 there is some evidence of heteroskedasticity. This further supports the team's decision to use heteroskedasticity-robust standard errors for all three models throughout this study.

CLM 6 - Normality of Residuals

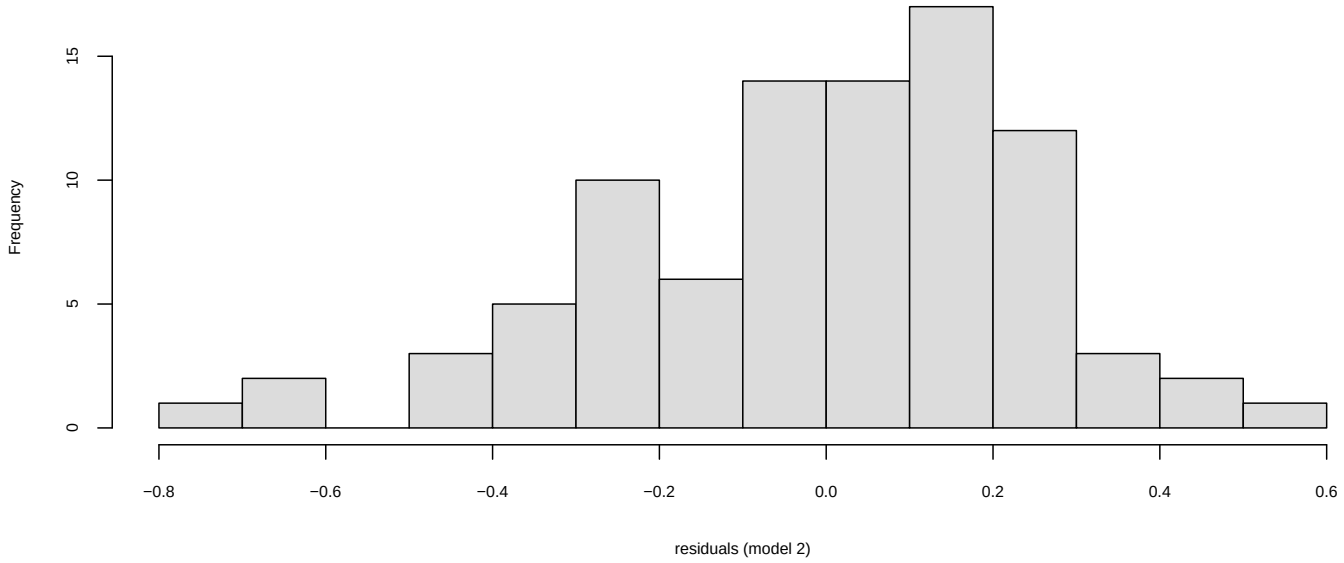
The last assumption the team examined was the normality of the error terms, which assumes that each error is drawn from a normal distribution with a mean of 0. The team took three approaches to assess this assumption:

1. Plotting a histogram the residuals
2. Plotting a q-q plot of the residuals
3. Running the Shapiro-Wilk normality test

Starting with a histogram of the residuals (Figure 29), the team observed a rather bell-shaped curve with a slight left skew for model 2.

```
hist(model2$residuals, breaks = 10, main = "Figure 29. Histogram of Residuals",
    cex.main=0.8, cex.lab=0.7, col = "gainsboro",
     yaxt = "n", xaxt = "n", xlab = "residuals (model 2)")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
```

Figure 29. Histogram of Residuals



The team proceeded to use a second technique to assess the normality of residuals for model 2 by examining the Normal Q-Q plot (Figure 26) and seeing how that compared to the histogram above. From this plot, there is not a considerable deviation from normality; as expected, the data points slightly deviate from the diagonal on the left side of the plot.

Lastly, the team chose to see how the diagnostic plots for model 2 compared to the results from a statistical normality test, namely the Shapiro-Wilk test. For this test, the null hypothesis is that that the errors are indeed normal. This test resulted in a p-value of 0.07, meaning we cannot reject the null hypothesis in favor of the alternative hypothesis that the residuals are non-normal.

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97422, p-value = 0.07034
```

For model 1, the team observed results similar to those were observed for model 2: A rather bell-shaped histogram,

with no considerable deviation from normality in the q-q plot, and a p-value greater than .05 for the Shapiro-Wilk test.

For model 3, the team observed some deviation from normality and observed a p-value less than .05 from the Shapiro-Wilk test. Therefore, we can reject the null hypothesis that the errors are drawn from a normal distribution. Even though we can reject the null hypothesis for model 3, the team understands that this normality test is sensitive to sample size and the results we calculated do not mean that the deviations from normality are large enough for us to be concerned.

Based on the results from the three techniques used above, and since the sample size is rather large (n> 30), the team will rely on the asymptotic properties of OLS, including the central limit theorem (CLT), which implies that OLS coefficients have a normal sampling distribution. Therefore, since OLS estimators are normally distributed for large sample sizes, we do not need the normality assumption to establish the shape of the sampling distribution.

Influential Points

In addition to examining the validity of the CLM assumptions, the team also checked to see if there are any observations with high influence. It is important to note that data points with high leverage have the most potential to affect the OLS coefficients; however, high-leverage data points do not always have high influence on the coefficients. Influence is the amount by which each data point actually changes the regression and is measured by Cook's distance.

The team examined the Residuals vs. Leverage plot (Figure 28) for model 2, to check for high-influence observations. Even though there is a data point with high leverage (x-axis), there is no data point with a large Cook's distance (>1). Thus, the team concluded that there are no observations with undue influence on the model fit.

Model 1 has one observation with Cook's distance ~0.5, which is a little concerning, but no observation with Cook's distance greater than 1. Additionally, model 3 has two high-influence points, one with Cook's distance greater than 0.5 and one with Cook's distance greater than 1. These results further strengthen the team's decision to consider model 2 as the best model in this analysis.

3.6 Regression Table

```
model1 = lm(log(crmrte) ~  prbarr + prbconv + density + west, data = crimes)
model2 = lm(log(crmrte) ~  prbarr + prbconv + density + log(polpc) + pctmin80, data = crimes)
model3 = lm(log(crmrte) ~ . -county -polpc + log(polpc), data = crimes)


model1$AIC <- AIC(model1)
model2$AIC <- AIC(model2)
model3$AIC <- AIC(model3)

se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))


stargazer(model1, model2, model3, type = "latex", intercept.bottom = FALSE,
          se = list(se.model1, se.model2, se.model3), header = F,
           title = "Linear Models Predicting Log of Crime Rate", #float = FALSE,
           omit.stat = c("f","ser"), #float = FALSE, #omit.table.layout = "n",
           star.cutoffs = c(0.05, 0.01, 0.001), column.sep.width = "1pt",
          font.size = "scriptsize") # Omit more output related to errors
```

Table 3 shows the coefficients for the three model specifications that were described above. The three models are linear in the betas and the outcome variable is the log of crime rate across all models. The statistical significance and practical significance (effect size) of the results will be discussed below.

Table 3: Linear Models Predicting Log of Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
Constant	-2.993^{***}	0.200	-0.610
	(0.198)	(0.815)	(1.799)
prbarr	-1.262^{**}	-1.981^{***}	-1.665^{***}
	(0.395)	(0.228)	(0.230)
prbconv	-0.552^{***}	-0.660^{***}	-0.580^{***}
	(0.132)	(0.108)	(0.159)
prbpris			-0.032
			(0.460)
avgsen			-0.012
			(0.017)
density	0.151^{***}	0.109^{***}	0.112^{*}
	(0.026)	(0.028)	(0.050)
taxpc			0.002
			(0.007)
west	-0.363^{***}		-0.168
	(0.073)		(0.124)
central			-0.118
			(0.086)
urban			-0.131
			(0.221)
log(polpc)		0.503^{***}	0.455^{**}
		(0.110)	(0.174)
pctmin80		1.151^{***}	0.907^{**}
		(0.160)	(0.288)
wcon			0.0004
			(0.001)
wtuc			0.0002
			(0.001)
wtrd			0.001
			(0.002)
wfir			-0.001
			(0.001)
wser			-0.0002
			(0.002)
wmfg			-0.0003
			(0.001)
wfed			0.002
			(0.001)
wsta			-0.001
			(0.001)
wloc			0.001
			(0.002)
mix			-0.501
			(0.622)
pctymle			2.173
			(1.583)
Observations	90	90	90
R^2	0.692	0.796	0.863
Adjusted R^2	0.677	0.784	0.818
Akaike Inf. Crit.	52.486	17.116	15.394
Note:			*p<0.05; **p<0.01; ***p<0.001

Statistical Significance

In this section, the focus is a closer look at the heteroskedasticity-robust standard errors and the p-values, in order to assess the statistical significance of the estimated coefficients across the three models. These statistics can be found in Table 3. The p-values for the coefficients of model 1 indicate high statistical significance. Specifically, *prbarr* has p-value < 0.01 and *prbconv*, *density* and *west* have p-values < 0.001.

Next, the coefficients of model 2 are all highly statistically significant with p-values < 0.001. As mentioned earlier, the team made the decision to drop the *west* variable from model 2, as its coefficient went from being highly statistically significant in model 1 to not statistically significant in model2. Additionally, the team ran a linear hypothesis, with the null hypothesis being that the coefficient for *west* $= 0$, to test whether or not *west* could be excluded from model 2. The test returned a p-value of .054, meaning the null hypothesis could not be rejected (see R output below). The differences observed between model 1 and model 2 show that this variable is highly sensitive to model definition and strengthens the decision to exclude it from the more refined model (model 2). Also, exclusion of *west* from the model had a minimal impact on the R-squared of model 2.

```
model2a = lm(log(crmrte) ~  prbarr + prbconv + density + west + log(polpc) + pctmin80, data = crimes)
linearHypothesis(model2a, c("west = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## west = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + density + west + log(polpc) +
##     pctmin80
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1     84
## 2     83  1 3.8207 0.05399 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, all the coefficients of the variables from model 2 remain statistically significant in model 3. All remaining covariates are not statistically significant in model 3. This shows that the results from model 2 are very robust (i.e. not sensitive to model choices).

Practical Significance

Given that the objective of this study is to examine the relationship between the different explanatory variables and covariates with the outcome variable (log of crime rate), the team examines the regression coefficients (betas) as a measure of effect size (i.e. practical significance).

First the team took a look at the coefficients for model 1, which includes the four explanatory variables. The coefficient for *prbarr* (probability of arrest) is -1.26, suggesting that for a 0.01 increase in probability of arrest, there is a 1.26% decrease in crime rate when controlling for the other explanatory variables. The coefficient for *prbconv* (probability of conviction) is -0.55, meaning that when probability of conviction goes up by 0.01 the crime rate decreases by 0.55% when controlling for the other explanatory variables. The standard deviation of probability of arrest and probability of conviction is 0.14 and 0.35, respectively. Therefore, the effect size appears to be practically significant, considering that the 0.01 increase in these two variables is much smaller than their standard deviations.

The coefficient of *density* in model 1 is 0.15 which means that for a unit increase in people per sq. mile, the crime rate will increase by 15% when controlling for the other explanatory variables. In other words, if density decreases by 0.1 then the crime rate decreases by 1.5% when controlling for the other explanatory variables. Density takes values from almost 0 to 8.83 in the dataset, with a mean of 1.44 and standard deviation of 1.52. Therefore, the research

team found this result to be practically significant, since a relatively small decrease in density can decrease the crime rate by 1.51%. Finally, the coefficient for *west* is -0.363, suggesting that when moving from non-West to West and controlling for probability of arrest, probability of conviction and density, the crime rate decreases by 36.3%. Even though this result appears to be very practically significant, the team is hesitant because logarithmic interpretations only hold for differentially small values (i.e. smaller than 10% or 20%). This result further strengthens the team's decision to not include *west* in model 2.

The coefficients in Table 3 under (2) pertain to the second regression equation (model 2), which includes three explanatory variables (prbarr, prbconv and density) and two covariates (log(polpc) and pctmin80). For *prbarr* and *prbconv*, the magnitude of the slope coefficients increased from -1.26 to -1.98 and -0.55 to -0.66 respectively. This indicates that a 0.01 unit increase in probability of arrest, while controlling for the other variables in the model, will decrease crime rate by a higher percentage than the first model indicated. Likewise, the same can be said for probability of conviction. The effect size for these two coefficients once again appears to be practically significant due to the increase in magnitude. For *density*, the magnitude of the slope coefficient ends up decreasing from 0.15 to 0.11, suggesting that a 0.1 unit increase in people per square mile will result in 1.1% increase in crime rate, if controlled for the other variables in the model. Though the coefficient for *density* decreased by about 25% , we still believe this finding to be practically significant, as a 1.1% increase in crime rate is not negligible.

Regarding the first covariate, the research team observes a slope coefficient of 0.503 for the *log(polpc)*, suggesting that a percent increase in police per capita increases crime rate by 0.503%. The team finds this result to be very practically significant, especially considering the positive value for this coefficient. Lastly, the coefficient value for *pctmin80* is 1.51, suggesting that a .01 unit increase in percent minority (1% increase) leads to a 1.51% increase in crime rate, when controlling for the other variables in the model. The mean value for percent minority in dataset is over 25%, giving us the impression that this increase in crime rate is not negligible. Therefore the team finds this coefficient for percent minority to be practically significant.

For model 3, the team notices that the coefficients for the explanatory variables and covariates from model 2 do not change much, indicating that the coefficient estimations in model 2 were rather robust. Additionally there are noticeably small coefficient values for the wage variables in the 3rd model indicating a small effect size. The team also notes that the probability of prison sentence and average sentence have a much smaller effect on log of crime rate when compared to the effect that probability of arrest and probability of conviction have. With regards to *Central*, the coefficient value is over 30% smaller than that of *West*, which was expected given the correlations between these variables and the log of crime rate; however the coefficient value for *Central* of -0.118 is still practically significant, as it suggests a 11.8% decrease in crime rate when moving from non-Central to Central, when controlling for other variables in the model.

The slope coefficient of *mix* (ratio of face-to-face crimes/other) is -0.501, indicating that for a 0.01 increase in *mix*, crime rate decreases by 0.501%. This result appears to be somewhat practically significant. Lastly, the coefficient of percent male (*pctymle*) is 2.173, suggesting that a .01 unit increase in percent male (1% increase) results in a 2.173% increase in crime rate, which is a considerable increase. Even though some of the coefficients in model 3 (discussed above) appear to be somewhat practically significant, the team is confident in the decision not to include these variables in model 2 (best model), since none of the coefficients appear to be statistically significant.

Model Assessment

Table 3 also shows the R-squared, adjusted R-squared, and AIC values across the three models. Adjusted R-squared will be used to evaluate the models since it accounts for the fact that r-squared will always rise when a new variable is added.

The adjusted R-squared for model 1 is 0.677, which means that the model explains 67.7% of the variation in the output variable (log of crime rate) with the 4 independent variables (prob. of arrest, prob. of conviction, density and west). The R-squared for model 2, increases to 0.784, meaning that the model explains 78.4% of the variation in log crime rate with three explanatory variables (prob. of arrest, prob. of conviction and density) and the addition of the 2 covariates (log of police per capita and percent minority). In model 3, where the team added almost all the independent variables into the regression equation, the adjusted R-squared increased to 0.818. It is notable that the addition of the two extra variables from model 1 to model 2 had a more significant effect on adjusted R-squared (from 0.677 to 0.784), than the addition of the 16 remaining variables from model 2 to model 3 (from 0.784 to 0.818).

Additionally, the team uses the Akaike Information Criterion (AIC) for an assessment of model fit that penalizes extra variables. Interestingly, even though model 1 has fewer variables than model 2, its AIC score (52.486) is almost 3 times higher than the AIC score of model 2 (17.116). This indicates that model 2 has a better fit (i.e. lower AIC score) than model 1, even after adjusting for parsimony.

Lastly, the team calculated model significance as a whole for all three models. Each model was found to be jointly significant (p-value $< .001$), meaning each of the three models has considerable predictive power on the whole (see R output below).

```
linearHypothesis(model1, c("prbarr = 0", "prbconv = 0", "density = 0", "west = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## density = 0
## west = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + density + west
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df     F    Pr(>F)
## 1     89
## 2     85  4 43.11 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model2, c("prbarr = 0", "prbconv = 0", "density = 0", "log(polpc) = 0", "pctmin80 = 0"),
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## density = 0
## log(polpc) = 0
## pctmin80 = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ prbarr + prbconv + density + log(polpc) + pctmin80
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     89
## 2     84  5 79.594 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model3, c("prbarr = 0", "prbconv = 0", "density = 0", "log(polpc) = 0", "pctmin80 = 0",
                           "west = 0", "prbpris = 0", "avgsen = 0", "taxpc = 0", "central = 0",
                           "urban = 0", "wcon = 0", "wtuc = 0", "wtrd = 0", "wfir = 0", "wser = 0",
                           "wmfg = 0", "wfed = 0", "wsta = 0", "wloc = 0", "mix = 0",
                           "pctymle = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
```

```
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## density = 0
## log(polpc) = 0
## pctmin80 = 0
## west = 0
## prbpris = 0
## avgsen = 0
## taxpc = 0
## central = 0
## urban = 0
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wfed = 0
## wsta = 0
## wloc = 0
## mix = 0
## pctymle = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ (county + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + west + central + urban + pctmin80 +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc +
##     mix + pctymle) - county - polpc + log(polpc)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     89
## 2     67 22 34.216 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Omitted Variables

There are several variables that are likely present in the true population model to predict crime rate, but are not available in the dataset used in this study. While this list below may not be exhaustive, the team believes that these omitted variables might have a significant impact on the outcome variable and at least one of our independent variables. The omitted variables will be summarized below, and the direction of the omitted variable bias will be hypothesized, along with the size of the bias and whether it is towards or away from zero.

```
om_var = c("unemployment rate", "education level",
           "percent of population below the poverty line",
           "percent of population in low income housing",
           "police force capability", "recidivism")
ind = c("percent minority", "density", "density", "density",
           "probability of arrest", "police per capita")
beta1 = c("positive", "positive", "positive", "positive",
           "negative", "positive")
```

```
beta2 = c("positive", "negative", "positive",
          "positive", "negative", "positive")
delta = c("positive", "positive", "positive",
          "positive", "positive", "positive")
bias = c("positive", "negative", "positive",
          "positive", "negative", "positive")
bias_size = c("++", "-", "++", "++", "- - -", "+++")
zero = c("away","towards","away","away","away","away")

table1 = data.frame(om_var, ind, beta1, beta2, delta, bias, bias_size,zero)
colnames(table1) <- c("Omitted Variable (OV)","Independent Variable (ID) affected","Sign of ID (beta1)", "

kable(table1, format = "latex", booktabs = T,
      caption = "Omitted Variable Bias") %>%
  kable_styling(font_size = 8, latex_options = c("striped", "hover", "HOLD_position", "condensed")) %>%
  column_spec(1, bold = T, width = "10em") %>%
  column_spec(2, width = "9em") %>%
  column_spec(3, width = "4em") %>%
  column_spec(4, width = "4em") %>%
  column_spec(5, width = "6em") %>%
  column_spec(6, width = "6em") %>%
  column_spec(7, width = "4em") %>%
  column_spec(8, width = "5em") %>%
  row_spec(0, bold = T) %>%
  footnote(general = "size of bias: +/- = small; ++/- - = medium; +++/- - - = large")
```

Table 4: Omitted Variable Bias

Omitted Variable (OV)	Independent Variable (ID) affected	Sign of ID (beta1)	Sign of OV (beta2)	Corr. b/w ID and OV (delta)	Sign of Bias (beta2*delta)	Size of Bias	Towards or Away from Zero
unemployment rate	percent minority	positive	positive	positive	positive	++	away
education level	density	positive	negative	positive	negative	-	towards
percent of population below the poverty line	density	positive	positive	positive	positive	++	away
percent of population in low income housing	density	positive	positive	positive	positive	++	away
police force capability	probability of arrest	negative	negative	positive	negative	- - -	away
recidivism	police per capita	positive	positive	positive	positive	+++	away

Note:
size of bias: +/- = small; ++/- - = medium; +++/- - - = large

The table above (Table 4) summarizes the results of the omitted variable bias analysis. These results will be discussed below:

Unemployment rate. The unemployment rate in a county is likely to have an impact on the outcome variable (crime rate). Specifically, the team anticipates a positive relationship between unemployment and crime rate (i.e. increasing unemployment rate leads to increasing crime rate) ($\beta_2 > 0$). Additionally, it is expected for unemployment rate to be positively correlated with one of the independent variables (percent minority) ($\delta > 0$). Thus, the omitted variable bias caused by unemployment rate on the estimated coefficient for percent minority ($\beta_2 * \delta$) is expected to be positive. Also, the team hypothesizes that the unemployment rate would have a large effect on crime rate and will also have a medium size correlation with percent minority. Therefore, the size of the bias is expected to be medium to large. Finally, since the estimated coefficient for percent minority is positive ($\beta_1 > 0$), the omitted variable bias is away from zero.

Education level. The team believes that education level is likely to affect crime rate. The higher the level of education the lower the crime rate ($\beta_2 < 0$). Also, the team expects education level to be positively correlated with *density* (i.e. rural areas with less density are likely to have less educated population than urban areas) ($\delta > 0$). This results in a negative omitted variable bias towards zero (since the estimated coefficient for density, β_2, has a positive sign). The team expects the size of this omitted variable bias ($\beta_2 * \delta$) to be relatively small. Even though education level is likely to have a somewhat large effect on crime rate (i.e. large β_2), the correlation between density and education level is expected to be relatively small (i.e small δ).

Percent of population below the poverty line. Along the same lines as unemployment rate, it is expected that higher rates of poverty would have an impact on the outcome variable; as the poverty rate increases within a county, the team would anticipate the crime rate to increase as well ($\beta_2 > 0$). Similarly, the team expects poverty rate to have a positive correlation with the explanatory variable, density, as population-dense areas tend to have a higher poverty line due to high cost of living ($\delta > 0$). Thus, since density has a positive coefficient in each of the regression models, this results in a positive omitted variable bias away from zero for *density* ($\beta_2 * \delta$). Additionally, the team believes that the poverty rate will have a similar effect as unemployment rate does, in that it will have a large effect on crime rate and will have a medium size correlation with density. Therefore, the size of the bias is expected to be medium to large.

Percent of population in low income housing. The team anticipates the percent of population in low income housing having an impact on crime rate. Specifically, it is expected that a higher rate or residents living in low income housing would lead to a higher crime rate, as has been demonstrated in other areas of the country ($\beta_2 > 0$). Additionally, this omitted variable would also likely have a positive correlation with density, as there is more low income housing (such as Section 8 housing) in population-dense areas ($\delta > 0$). Therefore, leaving this variable out of the model results in a positive bias away from zero ($\beta_2 * \delta$) , since density has a positive coefficient in each of the regression models. As for effect size, the team groups this variable with the previously mentioned socio-economic variables and anticipates a medium to large effect.

Police force capability. The capability of the police force in a county is likely to affect crime rate. The more skilled the police force, the lower the crime rate ($\beta_2 < 0$). The team believes that police force capability is also highly positively correlated with probability of arrest (*prbarr*) ($\delta > 0$). In other words, the more skilled the police force the higher the probability of arrest (i.e. higher ratio of arrests/offences). Therefore, omitting *police force capability* from the model causes a negative bias away from zero, since the coefficient of *prbarr* (β_1) is negative. The team expects this bias to be somewhat large.

Recidivism. Recidivism (i.e. the tendency of a convicted criminal to reoffend) is expected to have a positive effect on crime rate. Areas with high percentages of recidivism are likely to have higher crime rates ($\beta_2 > 0$). Additionally, the team believes that recidivism and the independent variable police per capita (*polpc*) are positively correlated, since areas with higher percentages of repeat offenders are likely to have more police per capita ($\delta > 0$). Therefore, omitting recidivism from the model is likely to cause a positive omitted variable bias on the estimated coefficient of police per capita, and since *polpc* has a positive coefficient ($\beta_1 > 0$) the bias will be away from zero. As discussed in section 3 of this report, the positive coefficient for *polpc* appears to be against the intuitive direction (i.e. negative: more police per capita leads to less crime). The size of this bias is expected to be significant and potentially the driver of the observed positive *polpc* coefficient.

Below the team summarizes the key takeaways from the omitted variable bias analysis:

- The estimated effect of percent minority on crime rate is somewhat overstated due to the positive bias of *unemployment rate* on percent minority.
- *Education level* is an omitted variable that is likely to cause negative bias towards zero on density (i.e. in that case the effect size of density on crime rate is understated). On the other hand, *Percent of population below the poverty line* and *Percent of population in low income housing* are both likely to have a positive bias away from zero on density (i.e. in that case the effect size of density on crime rate is overstated). Considering the opposite direction of the biases mentioned above, some of the bias on the density coefficient is likely going to be canceled out. However, the team believes that the combined positive bias from two omitted variables will have a bigger effect than the negative bias and, thus, the *density* coefficient will be slightly overstated.
- The estimated coefficient of probability of arrest (*prbarr*) is likely overstated due to a negative omitted variable bias (away from zero) caused by omitting *police force ability* from the model.
- The estimated coefficient of police per capita (*polpc*) is positively biased likely due to the omitted variable

recidivism, which could explain the positive sign of the *polpc* coefficient (as opposed to a negative expected impact).

As for potential proxies, the team believes that *taxpc* could be an imperfect proxy for the percent of population living below the poverty line omitted variable. This is because tax revenue per capita gives insight into the earnings of the residents of each county. Therefore, the team would expect a strong negative correlation between *taxpc* and poverty rates in each county. One reason this proxy would be viewed as "imperfect" is because *taxpc* could be very sensitive to outliers; there could be a few significantly wealthy residents in a county that end up skewing the value.

5. Conclusion

The proposed model for percent change in crime rate is model 2, which includes the variables *prbarr*, *prbconv*, *density*, log(*polpc*), and *pctmin80*. All of these variables have a practically significant impact on the outcome variable, log(*crmrte*). Two of the explanatory variables, *prbarr* and *prbconv*, had negative coefficients in the regression model 2. The variables log(*polpc*), *density*, and *pctmin80* all had positive coefficients. These variables not only have independent and joint statistical significance, but the coefficients also show practical significance on the outcome variable as well.

There are several observations that can be made from model 2 to help shape policies and give political recommendations in the state of North Carolina.

- The probability of arrest and conviction had a negative impact on the outcome variable. This indicates that police forces that are more effective help to reduce the crime rate. Coupled with the positive impact that a percent increase in police per capita had on the outcome variable, one can conclude that more police officers is not nearly as effective in reducing crime as good police officers. Therefore, the team suggests increasing spending on police education programs and police officer retention programs, to give more experience and to keep experienced officers.
- The positive impact that density and percent minority have on the outcome variable suggests that the beginning of the suggested programs should start in high density areas, especially high density areas with a higher percent minority population. This focus helps to more effectively execute the new programs to have the highest impact and be the best use of tax dollars.
- Some policies the team would recommend NOT focusing on include increasing mandatory sentencing because of the low impact of probability of prison sentence on the outcome variable. The certainty of punishment appears to have a higher impact than severity of punishment in reducing the crime rate. Also, the team would not recommend simply increasing the number of police officers in a high crime rate area because of the positive coefficient that variable has in the model - police effectiveness appears to have a better impact on crime rate than number of police officers.

With the data and best model built in this analysis, no recommendations can be made at this time for policies regarding income and tax level and location in the state. Additionally, the team identified some key omitted variables, including unemployment rate, education level, police force capability etc., that would help better estimate the true population model. Because these effects, some thought of the omitted variables must be taken into account when constructing the recommended policies. Finally, with the age of the cross sectional data provided, the team would recommend repeating the analysis with more recent data before enacting policies to check if the conclusions stand the test of time.