Lab 3 Final Version

TK Truong, Dili Wang

12/09/2018

Introduction

In this report, we will be studying the relationship between crime rate and other variables for 90 counties in the state of North Carolina in 1987. The purpose of our exploratory data analysis and regression analysis will be to ascertain which of the variables in the data provided, if any, are convincing determinants of crime. Research has shown a relationship between high crime rates in dense, urban areas due to several factors such as higher populations, income disparity, and differences in social or environmental norms (*Krivo & Peterson, 1996*). Our study will focus on the effect of county density and wealth on predicting crime rate. In addition to measuring the effect that these variables have on the crime rate, we will also discuss which of the variables are confounding and which variables omitted in the data could further influence this model. We hope that our results will help further inform campaign policies that serve to lower crime rates in this state.

Exploratory Data Analysis

```
library(ggrepel)
library(dplyr)
library(car)
library(reshape2)
library(ggplot2)
library(cowplot)
library(stargazer)
library(lmtest)
library(sandwich)
```

```
crime = read.csv("crime_v2.csv", header = TRUE)
```

A. Cleaning the Dataset

Initially, we have a data set of 97 observations and 25 columns. In examining a summary of the crime data set, we found that there are 6 rows where there are NA's in every column. We will remove these as a part of our cleaning process. We also discovered a duplicate row of county 193, which will be removed as well.

The summary also revealed that the prbconv variable was read as a factor instead of a numeric variable as we expected, thus we transformed that variable into a numeric one.

```
# removing NA rows
ncrime <- crime[!is.na(crime$county), ]
# removing any duplicate rows
ncrime <- ncrime[!duplicated(ncrime), ]
# transform prbconv to numeric
ncrime$prbconv <- as.numeric(as.character(ncrime$prbconv))
```

After this cleaning process, the data set contains 90 observations.

B. Examining the data

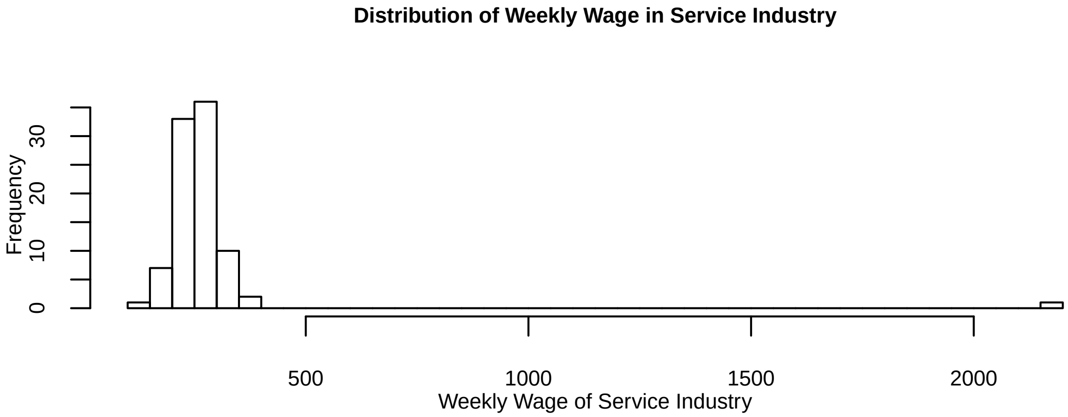
To examine all of the metric variables more closely, we used the following code to plot a histogram of the distribution of the raw numeric data for each metric variable. We excluded the variables county, year, west, central and urban from this step, since they are categorical and not metric variables. The histogram outputs have been suppressed here, but please refer to the .rmd file for more details.

```
for (i in c(seq(3, 10, 1), seq(14, 25, 1))){
  col_name = names(ncrime)[i]
  hist(ncrime[[col_name]], breaks = 30, xlab = col_name, main = col_name)
}
```

From this initial histogram analysis, we noticed outliers for several of the variables - probability of arrest (prbarr), police per capita (polpc), tax per capita (taxpc), percent young male (pctmyle) - which may influence our later analysis. None of the outliers for these variables were unusual enough for us to consider them as erroneous data points.

However, we did notice an extreme outlier for variable weekly wage of service industry (wser) in its distribution histogram, displayed here:

```
hist(ncrime$wser, breaks = 30,
     xlab = '', ylab = "",
     main = "Distribution of Weekly Wage in Service Industry",
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Weekly Wage of Service Industry", side=1, cex=0.7, padj = 4)
mtext("Frequency", side=2, cex=0.7, padj = -4)
```



In the histogram of the wser variable we see a huge outlier for county 185 with a value of 2177.0681. Given that the mean without this outlier is 253.9701, we strongly suspected this data point is the result of the decimal being in the wrong place (possibly a mistake during data entry) and that the proper value should be approximately 217.71 instead. We corrected this data point by replacing the value of 2177.0681 with 217.71 before further analysis.

```
mean(ncrime$wser)
```

```
## [1] 275.3379
```

```
mean(ncrime$wser[ncrime$wser < ncrime$wser[84]])
```

```
## [1] 253.9701
```

```
# resetting the wser value for county 185
ncrime$wser[84] <- ncrime$wser[84] * 0.1
```

Other noticeable features from our histogram analysis include several variables that show right skew - crime rate (crmrte), probability of arrest (prbarr), probability of conviction (prbconv), average prison sentence (avgsen), tax per capita (taxpc), and offense mix (mix). For variables that do not have skew, the distribution for most is symmetric and approximately normal, with the exception of pctmin80 which appears to be uniformly distributed. We did not notice any other strange data points, such as negative numbers or unreasonable numbers. Percent variables percent minority (pctmin80) and percent young male (pctymle) also reasonably fall within the range of 0 to 1.

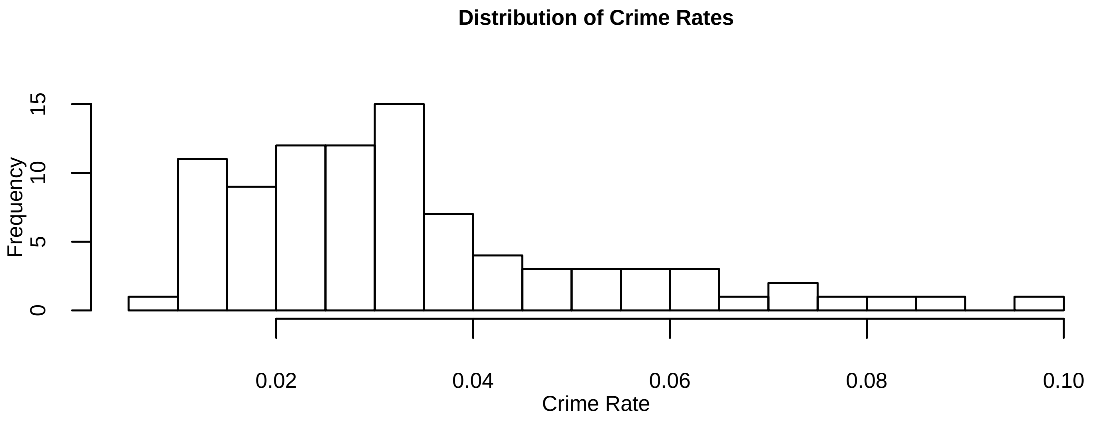
We did notice that variables probability of arrest (prbarr) and probability of conviction (prbconv) have values greater than 1, but did not find this to be alarming, given the definition of these variables. Probability of arrest is defined as the ratio of arrests to offenses, which would mean that there must be some crimes which resulted in arrests but was not categorized as offenses, resulting in higher arrests counts than offense counts and leading prbarr > 1. Probability of conviction, the ratio of convictions to arrests, can also be greater than 1 in a county for a couple of reasons. Perpetrators can be convicted of crimes in court through a citation (notice to appear in court), where an arrest. The other way that convictions could be greater than arrests is if there are perpetrators who are arrested once but convicted of multiple crimes. On the other hand, probability of prison sentence (prbpris) should only contain values between 0 and 1, as the data shows, since this is the percent of convictions that result in a prison sentence. Assuming every conviction either results in 1 or 0 prison sentences, there would be no way for prbpris to be > 1.

C. Univariate Analysis of Dependent Variable

Since the objective of this study is to find the determinants of crime, our dependent outcome variable for all versions of our model building process will be the crime rate (crmrte). Before, proceeding, we would like to examine some of the characteristics of the crime rate variable.

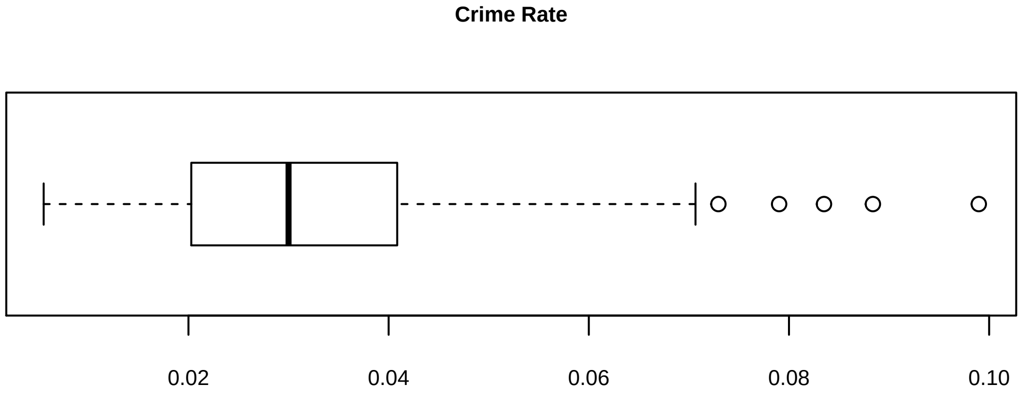
First, we examine the distribution of the crmrte variable more closely via a histogram:

```
hist(ncrime$crmrte, breaks = 30,
     xlab = '', ylab = '',
     main = "Distribution of Crime Rates",
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Crime Rate", side=1, cex=0.7, padj = 4)
mtext("Frequency", side=2, cex=0.7, padj = -4)
```



Here, we see that the data is skewed right but does not appear to be any other alarming features or discontinuities in the data. There are also appears to be 5 outliers to the right, with values greater than Q3+1.5(IQR), according to the boxplot below:

```
boxplot(ncrime$crmrte, main = "Crime Rate",
        horizontal = TRUE, cex.main = 0.7, cex.axis = 0.7)
```



However, in examining both the histogram and boxplot, we would not characterize any of these outliers as extreme. Thus, we do not currently have sufficient reason to believe that these 5 data points require additional cleaning (either removal or correction). We will examine later if any outliers exert considerable influence in our model. Since this variable represents a rate, it would not be appropriate to perform a log transformation. We would like to note that in some versions of our analysis, not included in this final report, we found that all versions including log transformation of crmrte also did not result in better fitting models. Therefore, we will perform the remainder of the analysis here without transformation of the crmrte variable.

D. Bivariate Regression Considerations:

Based on our background knowledge and research, we suspect that density and some measure of wealth are strong predictors of crime rate. However, we will conduct bivariate analyses to examine other potential relationships with crime rate. To help us determine which variables will enhance the accuracy of the 3 different versions of our multilinear regression model, we built the following table which displays the correlation coefficient between our dependent crime rate (crmrte) variable and every metric variable (excluding the 3 categorical variables). The left column shows the variable name and the right column shows the strength of linear correlation between crime rate and that variable. The rows are then ordered in this table by the absolute value of the correlation coefficient:

```
ncrime_subset <- select(ncrime, crmrte, prbarr:pctymle)
rows <- ncol(ncrime_subset)-1
correlations <- data.frame(variable=character(length=rows),
                 correlation=numeric(length=rows),
                 stringsAsFactors=F)
for (i in 1:rows) {
  temp1 <- colnames(ncrime_subset[i+1])
  temp2 <- cor(ncrime_subset[,1], ncrime_subset[,i+1])

  correlations[i,1] <- temp1
  correlations[i,2] <- temp2
}
correlations <- subset(correlations, variable != 'west'
                & variable != 'central' & variable != 'urban')
#order by strength of correlation
correlations[order(abs(correlations$correlation), decreasing = TRUE), ]
```

```
##    variable correlation
## 6   density  0.72837061
## 18     wfed  0.48991633
## 7     taxpc  0.44871511
## 14     wtrd  0.42722262
## 1    prbarr -0.39528302
## 12     wcon  0.39296155
## 2   prbconv -0.38596559
## 20     wloc  0.35982934
## 16     wser  0.35439832
## 17     wmfg  0.35256117
## 15     wfir  0.33602609
## 22  pctymle  0.29033966
## 13     wtuc  0.23599574
## 19     wsta  0.19984675
## 11 pctmin80  0.18165059
## 5     polpc  0.16728163
## 21      mix -0.13200035
## 3   prbpris  0.04799540
## 4    avgsen  0.01979653
```

The purpose of this correlation table is to aid with our strategy in building models 2 and 3. As we refine the accuracy of our model from model 1 to model 2, we will give priority to examine variables that already exhibit stronger bivariate correlation with crmrte.

A couple characteristics we can already see from our correlations table are the following:

- Comparatively, prbpris, avgsen, polpc, pctmin80 and mix variables have some of the lowest bivariate correlations with the crime rate variable.
- The correlation coefficient that is most surprising is that of the police per capita (polpc) variable, which could be important to our objective of informing policy. In addition, the correlation between crmrte and polpc is positive, which is slightly puzzling. One could argue that more law enforcement in a county should then lead to a reduction in crime rate, but this is not the case. Given that the correlation is also not strong, this variable is a covariate we will examine more closely in model 2.

- Other surprising positive correlations are crmrte with taxpc and all of the wage variables. Both tax revenue per capita and weekly wages can arguably serve as a measurement of the overall wealth of a county, and one would reasonably assume that the initial assumption is that crime is less likely to occur in wealthier areas, but this is not shown in the bivariate correlation coefficients.

- The correlation between crime rate and density is, by far, the strongest of all of the other variables with coefficient of 0.7284. This further supports our initial speculation that population density could be a possible determinant of crime.

E. Effects of Urban and Location

Since the 3 categorical variables - urban, west, central - only contain values 0 or 1, we decided that it would not be proper to test if any of these variables have a strong correlation with our dependent crmrte variable through correlation coefficient calculation. Instead, we perform the following exploratory data analysis to determine which of the categorical variables may play a significant role in our regression model building later on.

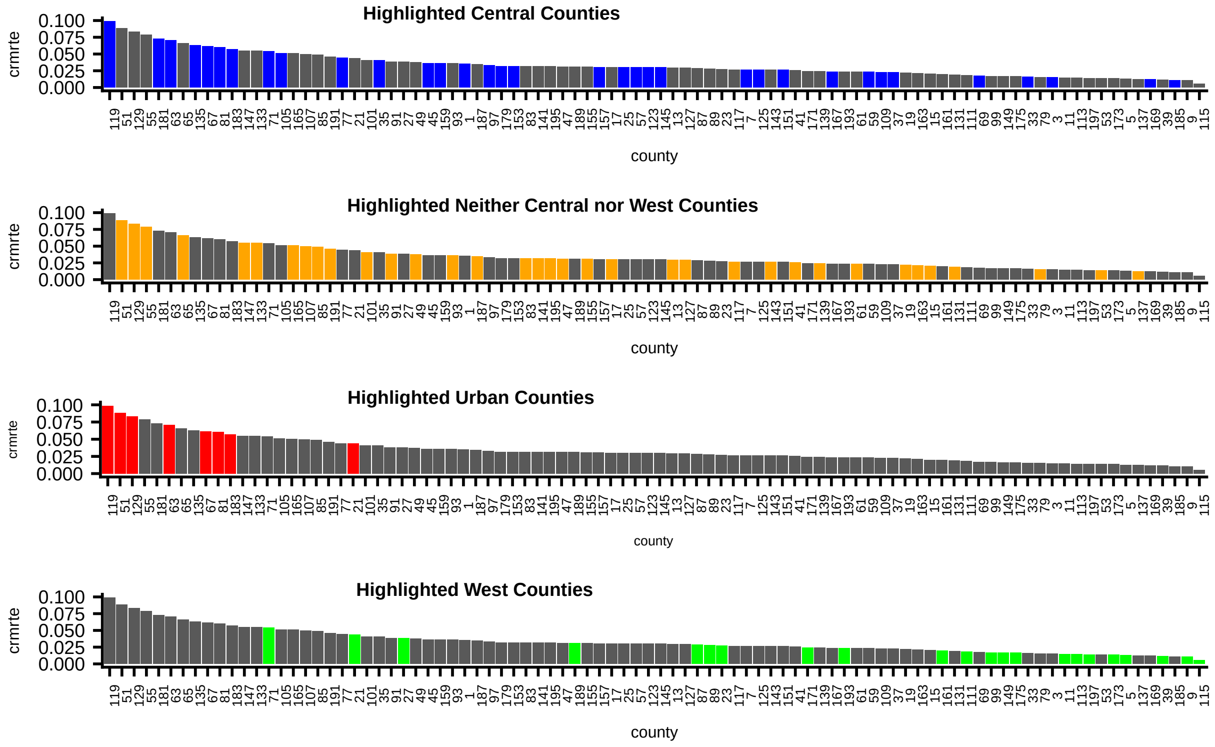
First, we examined the categorical variables of west and central, which refers to the location of the county in North Carolina. There are 22 west counties and 34 central counties, meaning there are some that are neither. As a result, we added an additional column (notwestcen) to denote this. We note that one county (71) was labeled as both west and central, but we do not have enough information to conclude whether this is an error.

```r
ncrime$notwestcen <- 0
ncrime$notwestcen[ncrime$west == 0 & ncrime$central == 0] <- 1
```

Next, we explored whether or not the 4 categorical variables - urban, central, west, and neither central nor west (notwestcen) has any noticeable effect on the dependent variable of crime rate. Using the following code, we plot 4 bar charts of crime rate per county, where the bars are ordered from highest to lowest crime rate. However, we use a different color to highlight counties where each indicator variable = 1:

```r
# Central
ncrime$county <- factor(ncrime$county, levels = ncrime$county[order(-ncrime$crmrte)])
central_bar = ncrime %>%
  mutate(highlight_flag = ifelse(central == 1, T, F)) %>%
  ggplot(aes(x = county, y = crmrte)) +
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge",
  show.legend = FALSE) + theme(text=element_text(size=6),
  axis.text.x=element_text(size=5, angle=90), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'blue'))
# Not Central and Not West
```

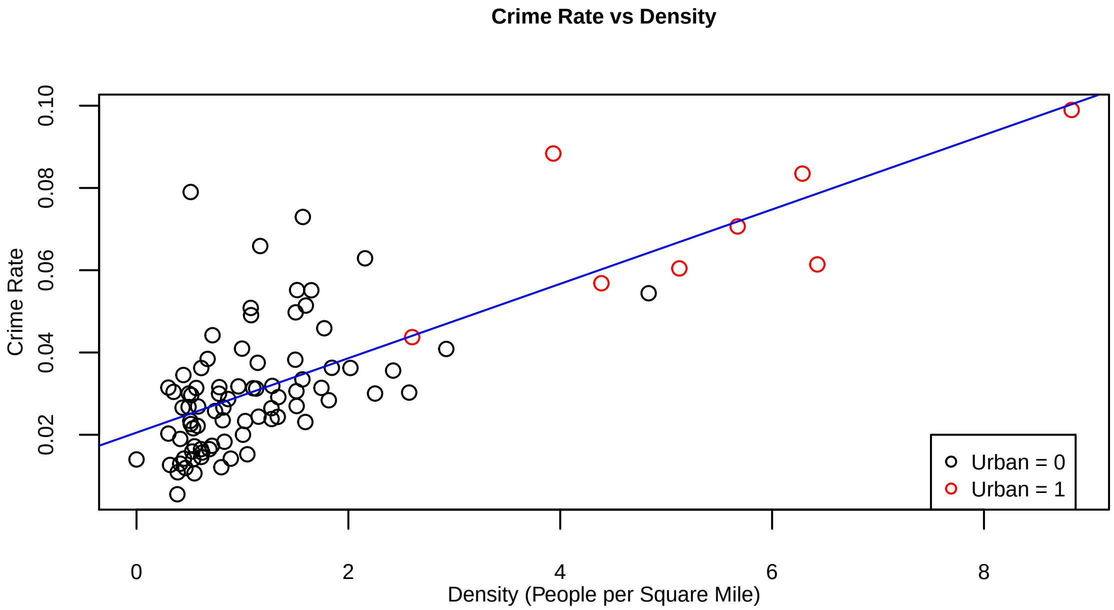
```r
ncrime$county <- factor(ncrime$county, levels = ncrime$county[order(-ncrime$crmrte)])
notwestcen_bar = ncrime %>%
  mutate(highlight_flag = ifelse(notwestcen == 1, T, F)) %>%
  ggplot(aes(x = county, y = crmrte)) +
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge",
  show.legend = FALSE) + theme(text=element_text(size=6),
  axis.text.x=element_text(size=5, angle=90), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'orange'))
# Urban
ncrime$county <- factor(ncrime$county, levels = ncrime$county[order(-ncrime$crmrte)])
urban_bar = ncrime %>%
  mutate(highlight_flag = ifelse(urban == 1, T, F)) %>%
  ggplot(aes(x = county, y = crmrte)) +
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge",
  show.legend = FALSE) + theme(text=element_text(size=5),
  axis.text.x=element_text(size=5, angle=90), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'red'))
# West
ncrime$county <- factor(ncrime$county, levels = ncrime$county[order(-ncrime$crmrte)])
west_bar = ncrime %>%
  mutate(highlight_flag = ifelse(west == 1, T, F)) %>%
  ggplot(aes(x = county, y = crmrte)) +
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge",
  show.legend = FALSE) + theme(text=element_text(size=6),
  axis.text.x=element_text(size=5, angle=90), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'green'))
plot_grid(central_bar, notwestcen_bar, urban_bar, west_bar, scale = 1,
          labels = c("Highlighted Central Counties", "Highlighted Neither Central nor West Counties",
          "Highlighted Urban Counties", "Highlighted West Counties"), hjust = c(-1.5,-0.9, -1.5, -1.6),
          label_size = 7, nrow = 4, ncol = 1)
```



From these bar charts, we see that effects of central and notwestcen variable on crime rate are actually quite random, since we see central counties and neither west nor central counties with all values of crime rate, both low and high. Comparatively, the effect of the urban categorization is not random on the crime rate variable, as we see 7 of 8 urban counties have some of the highest crime rates with numbers in the 4th quartile. Unlike the urban indicator, the west categorization contains a greater range of crime rates, but it is still difficult to characterize its effects as random since we see that there are no west counties with very high crime rates. As we further refine model 2, we will study the effects of the 2 non-random indicator variables, west and urban, in our regression analysis.

By looking at the relationship between the urban and density variables, we can also start to formulate a plausible argument for why urban counties may have higher crime rates than non-urban counties, To do so, we created this crime rate vs density scatterplot where the urban counties are highlighted in red.

```r
plot(ncrime$density, ncrime$crmrte, col = ncrime$urban+1,
     main = "Crime Rate vs Density", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
legend(7.5, 0.02, legend=c("Urban = 0", "Urban = 1"),
       col=c("black", "red"), cex=0.7, pch = 1:1)
mtext("Density (People per Square Mile)", side=1, cex=0.7, padj = 4)
mtext("Crime Rate", side=2, cex=0.7, padj = -4.5)
abline(lm(crmrte ~ density, data = ncrime), col = 'blue')
```



From this plot, we can see, perhaps unsurprisingly, that 7 of the urban counties have the highest population densities and the highest crime rates, which further supports our claim that density should be an explanatory variable. One speculation which we considered is that crime may be more likely to occur in urban counties where the population density is high, because higher populations within close proximity allow for more opportunities for perpetrators to commit crimes. For instance, perpetrators may be less likely to commit crimes in rural areas simply because crime locations are either too far away or there are less people to victimize. Also, we wanted to point out that while it is not possible for us to measure a correlation coefficient between a metric like density and an indicator variable like urban, the two variables clearly share a strong relationship, where urban counties can be definitively characterized as highly dense. Given this relationship, it may be confounding to include both variables in the same model. We will test for the inclusion of urban indicator variable in Model 2, but only the density variable will be included as an explanatory variable in model 1.

Model Building

Model 1

For the initial model, there are 2 major sets of variables that we would characterize as directly explanatory for the crime rate:

- As we discussed in the EDA, high population density could provide "opportunity" in the sense of increased proximity, a greater amount of potential victims, and the convenience for multiple crimes to be committed by the same perpetrator. Given the strong correlation between population density and crime rate, density will be the first explanatory variable we study.

- Based on EDA and research, the second explanatory variable we wish to study is the effect of a county's overall "wealth" on the crime rate. "Wealth" of a county can be measured by a wide array of variables, such as average/median income, quality of amenities for residents, quality of schools, etc.. Our initial speculation is that crime rate would be lower in wealthier counties and higher in poorer areas. While we do not have a direct manifestation of a "wealth" variable in our data set, we can find quite a few proxies.
  - The first proxy for a "wealth" variable would be the taxpc variable, which has the 3rd highest correlation strength with our dependent crime rate variable. Generally, tax revenue is a percentage of income and thus, taxpc can be considered a good proxy of county wealth where we would expect counties with high tax revenue to consist of wealthier inhabitants. However, we also want to point out that income may not be the only driver for tax rates. For example, taxes can be influenced by several other factors outside of income, including number of children, cost of property and mortgage, cost of amenities (parks, recreation, school programs), or the current political climate.
  - The second proxy for a "wealth" variable would be the wage variables (wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc). As we see from our correlation analysis during EDA, wfed, wtrd, and wcon also have very strong correlations

with dependent crime rate variable. The issue with inclusion of wage variables is that none of the wage variables alone are a good representation of the overall wealth of a county. We have no information the distribution of county residents in each of the 9 categories of wage, and therefore we cannot give proper weights to each of the 9 wage variables. Ultimately, we believe that taxpc instead of wage variables is a better representation of the "wealth" variable, but we will study the effect of inclusion of wage variables in further iterations of model 1.

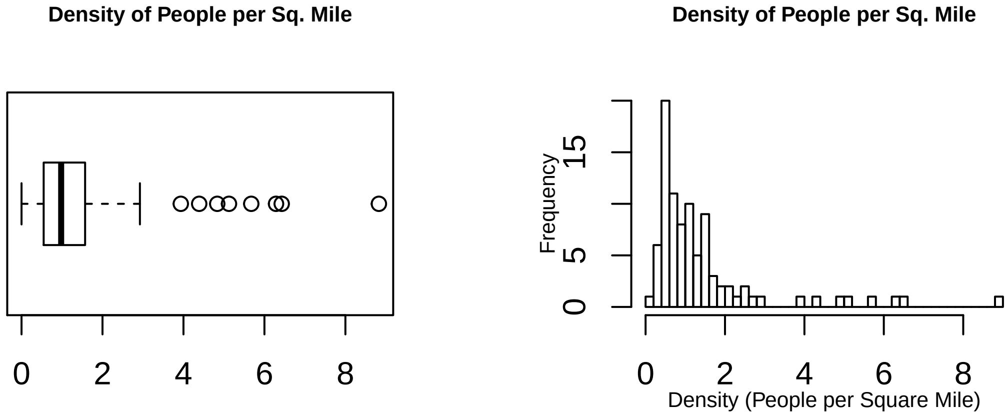
For our initial version of model 1, we seek to find the coefficients in the following multivariate model:

$$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + u$$

Density

In the following box plot and histogram, we examined the density variable more closely:

```r
par(mfrow=c(1,2))
boxplot(ncrime$density,
        main = "Density of People per Sq. Mile",
        horizontal = TRUE, cex.main = 0.7)
hist(ncrime$density, breaks = 40, xlab='', ylab='',
     main = "Density of People per Sq. Mile",
     cex.main = 0.7, cex.lab = 0.7)
mtext("Density (People per Square Mile)", side=1, cex=0.7, padj = 4)
mtext("Frequency", side=2, cex=0.7, padj = -4.5)
```

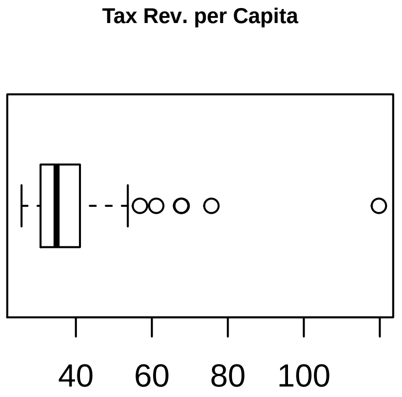


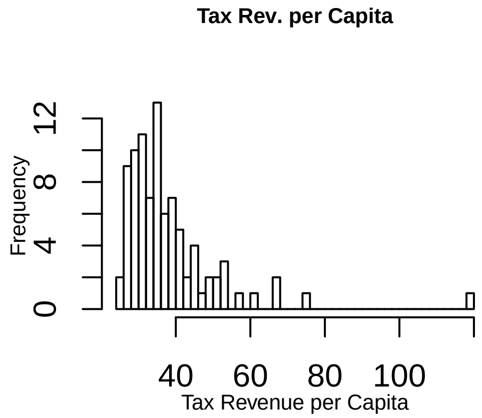
We see from these plots that the density variable is asymmetric with right skew and has several outliers greater than Q3+1.5(IQR). While some of most of the outliers are not very extreme, there is one to the far right with a value of 8.83 people per square mile for the county 119. As we run diagnostics on our model later on we will also assess if the outliers for density contribute to erroneous or alarming influences in our model. While the larger values contribute to both the skew and appearance of gaps in the distribution, we do not believe a log transformation is proper for this variable, given that it is a ratio of the measure of people to area.

Tax Revenue per Capita

We also closely examined the taxpc (tax revenue per capita) variable in the following boxplot and histogram:

```r
par(mfrow=c(1,2))
boxplot(ncrime$taxpc,
        main = "Tax Rev. per Capita",
        horizontal = TRUE, cex.main = 0.7)
hist(ncrime$taxpc, breaks = 40,
     xlab='', ylab='', main = "Tax Rev. per Capita",
     cex.main = 0.7, cex.lab = 0.7)
mtext("Tax Revenue per Capita", side=1, cex=0.7, padj = 4)
mtext("Frequency", side=2, cex=0.7, padj = -4.5)
```





There is clear right skew in the taxpc variable with a handful of outliers past Q3+1.5(IQR). Most of the points past the right whisker are not extreme outliers, but there is one data point that is particularly extreme for county 55 with a value of 119.76145. Given that we do not have reason to believe that this is erroneous, we will not remove or correct it at this point, and we will study its influence later in our diagnostic plot of Cook's Distance.

We also decided to perform no further log transformation on the taxpc variable given that it is a ratio of total tax revenue to number of people. We did want to note here that in previous versions of our analysis, we did calculate the correlation coefficient between log(crmrte) and log(taxpc) and found that the result, 0.3398, was not better than our earlier reported correlation coefficient of 0.4487 between untransformed crmrte and taxpc variables, which further ruled out the need for log transformation.

Now that we have thoroughly examined our explanatory variables, we build the first version of model 1, an OLS regression model with taxpc and density as independent variables and crmrte as the dependent variable:

```
model1_1 = lm(crmrte ~ density + taxpc, data = ncrime)
model1_1$coefficients
```

```
##  (Intercept)      density        taxpc
## 0.0086976849 0.0080869319 0.0003459569
```

From this output, we see that the estimated density coefficient is one order of magnitude greater than that of taxpc. The estimated coefficient for density is 0.0087 which means that for each density increase of 1 person per square area, we would expect crime rate to increase by 0.008 crimes per person. Rephrased more practically, for each increase of 100 people per square mile, we would expect crime rate to increase by ~0.87 crimes per person. The estimated coefficient for taxpc has less practical significance with a value of 0.0003. This means that a tax revenue per capita increase of $1 correlates with an expected increase in crime rate by 0.0003 crimes per person, or equivalently, an increase in $1,000 of tax revenue per capita correlates with an increases the crime rate by ~0.35 crimes per person.

We would argue that estimated slope of density variable on crime rate is of practical significance. While the density from county to county does not vary as much as 100 people per square mile, the density between an city within a county and a suburban/rural area in the same or different county could differ by as much as 100 people per square mile. If, for an example, a family considered moving from a suburban area to an urban city where the population density did differ by this amount or more, they may think twice if they knew that the crime rate could vary by ~0.87 crimes per person, or equivalently, an increase of 8 to 9 crimes per 10 people.

On the other hand, we would argue that a change in taxpc that corresponds to a practically significant change in crime rate, would be much harder to achieve. In order for predicted crime rate to vary as much as 3-4 crimes per 10 people, tax revenue per capita must vary by as much as $1,000. We conducted a quick search to discover that tax revenue per capita (*Naverson, 2018)*) between the states of Louisiana and New York differ by only as much as $2000 in 2015. If we adjust for lack of inflation in 1980, we would predict that it is difficult for taxpc to differ as much as 4 figures from state to state, and even less difficult to vary by this amount within the same state.

Rather puzzling from this conclusion is that crime rate increases with taxpc, our proxy for "wealth", which contradicts our earlier speculations that the wealthier the county, the higher the taxpc and the lower the crime rate. There could be several explanations for the positive slope of taxpc, but before we discuss this further we will need to first assess whether or not we wish to include any of the wage variables, as inclusion of other variables into our model will change the coefficient for taxpc.

We use the following to assess the general fit of our model:

```
summary(model1_1)$r.squared
```

```
## [1] 0.582296
```

```
summary(model1_1)$adj.r.squared
```

```
## [1] 0.5726936
```

```
AIC(model1_1)
```

```
## [1] -530.6359
```

Our initial model 1 has an AIC value of -530.6359 with R^2 of 0.5823 and adjusted R^2 of 0.5727. This means that 57.3% of the variation in our dependent crime rate variable can be explained by our regression model. The AIC calculated here will serve as a baseline to help us determine if further model additions result in better or worse fit. Next, we will test whether or not the addition of wage variables to this model will be appropriate.

Wage Variables

In the field of criminology, one theory to explain the reasoning behind criminal behavior is rational choice theory. This assumes that an individual will weigh the benefits and costs of committing a crime and act accordingly. While this is a big assumption and ignores other social factors that contribute to crime, we recognize the importance of legal opportunities to earn a living. If criminals are driven by financial incentives, higher availabiltiy of jobs and decent wages could be related to a lower crime rate.

In the data, we are given 9 wage variables, describing the weekly wage in different industries. One of the major limitations of this data is that we do not have the percentage allocation of county population in each of these industries. We must also assume that it is possible that some residents of a county may have occupations that are not represented in these wage variables that could potentially have major effects on the average wage and total income. Therefore, it is impossible for us to calculate directly from these 9 variables the total income of a particular county.

Another caveat we considered is the inclusion of taxpc and any wage variable as independent variables in the same linear regression model. It is arguable that because wage is related to income and taxpc is a percentage of income, taxpc is possibly an outcome of a wage variable and could confound the effects of the wage variable on the dependent crmrte variable. However, we also considered the fact that taxpc is not autocorrelated with wage (please see table below) and taxpc can depend on other factors that are not specifically related to wage, such as property tax of the county, the quality of schools in a county, and average number of children of residents. For the latter reason, we decided to assess the inclusion of wage variables by testing their inclusion into our initial MLR model that already included taxpc and density as independent variables.

To further study the wage variables and decide if they should be included in the model at all, we constructed the following table showing each wage variable's mean, standard deviation, and correlation coefficient between each variable and the taxpc variable. The rows are then displayed in order of descending median.

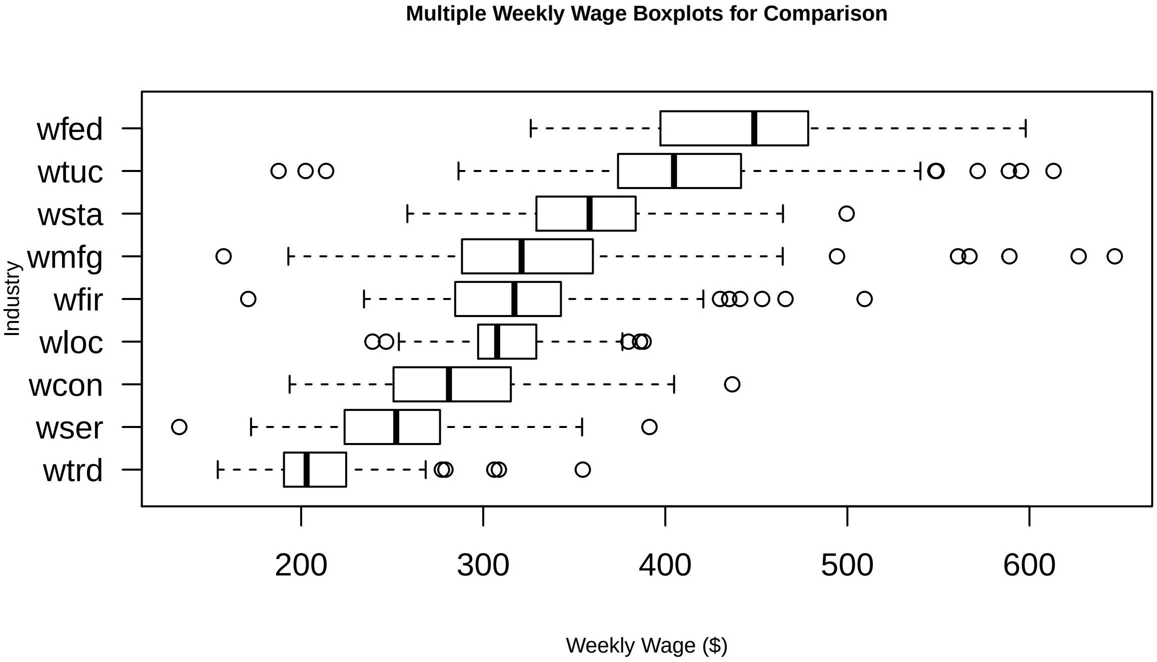
```
wages <- data.frame(variable=character(length=9),
                    median=numeric(length=9),
                    sd=numeric(length=9),
                    correlation_with_taxpc=numeric(length=9),
                    stringsAsFactors=F)

count = 1

for (i in seq(15, 23, 1)){
  col_name = names(ncrime)[i]

  wages[count, 1] = col_name
  wages[count, 2] = median(ncrime[[col_name]])
  wages[count, 3] = sd(ncrime[[col_name]])
  wages[count, 4] = cor(ncrime$taxpc, ncrime[[col_name]], use = 'complete.obs')

  count = count + 1
}

wages[order(wages$median, decreasing = TRUE), ]
```

```
##   variable   median       sd correlation_with_taxpc
## 7     wfed 448.8550 59.95125              0.0620723
## 2     wtuc 404.7800 77.35523              0.1712900
## 8     wsta 358.4000 43.29420             -0.0349883
## 6     wmfg 321.0500 88.23064              0.2586084
## 4     wfir 317.1257 53.99862              0.1309436
## 9     wloc 307.6500 28.13213              0.2199012
## 1     wcon 281.1624 47.75272              0.2639568
## 5     wser 252.2183 43.99123              0.2563944
## 3     wtrd 202.9879 33.87036              0.1839214
```

As we see from this table, the wage variables are actually weakly correlated with the taxpc variable. The wsta variable actually has negative, but close to 0, correlation with the taxpc variable. This further supports our claim that the inclusion of any of the wage variables will likely not detract from the accuracy of our model if any of them are included as independent variables along with taxpc.

We also created this boxplot comparison chart for the 9 wage variables that provides an aggregated view of the data:

```
boxplot(ncrime[c(17,19,15,23,18,20,22,16,21)],
        main = "Multiple Weekly Wage Boxplots for Comparison",
        xlab = "Weekly Wage ($)", ylab = "Industry",
        horizontal = TRUE, las = 1,
        cex.main = 0.7, cex.lab = 0.7)
```



Given that we do not have proper weights for each of the wage variables in terms of the percentage of each county's residents allocated to each of the 9 industries, we ruled out the possibility of creating a single wage variable proxy. For instance, if we took the average of all 9 wage variables, we will be creating an artificial variable that is no longer representative of any of the wage variables.

Since choosing one wage variable is not representative of all 9, one possibility we considered is the inclusion of a subset of the wage variables that both span the range of all wage variables and potentially increases the accuracy of our linear model. In looking at the correlations table in our EDA, where we calculated the correlation of each variable with crmrte, we found that wfed (weekly wage of federal employees), wtrd (weekly wage of wholesale and retail trade employees), and wcon (weekly wage of construction industry employees) have some of the strongest correlations with the dependent crime rate variable with correlation coefficients of 0.4899, 0.4272, and 0.393 respectively. If we take a look at the bar chart above, wfed also has the highest median wage, wtrd has the lowest median wage and wcon is ranked 7th in median wage. Together, the wage variables capture almost the entire range of all wage variables, so we felt that they are quite representative of all nine wage variables. Therefore we decided to build a second MLR model, called model1_2, which adds these 3 wage variables to the our previous model that already includes taxpc and density as independent variables:

```
model1_2 = lm(crmrte ~ taxpc + density + wfed + wcon + wtrd, data = ncrime)
```

The purpose of creating this second model, model1_2, is to test if these 3 wage variables have joint significance in our linear model. Essentially, we wish to know whether or not these 3 wage variables collectively are associated with a change in our dependent variable crmrte. To test the joint significance of the 3 wage variables, we perform the Wald test to assess if their inclusion produces a significant F-statistic. A significant F-statistic allows us to reject the null hypothesis and provides evidence for the alternative hypothesis which is that the 3 wage variables are jointly significant. Alternatively, an insignificant F-statistic provide evidence for the exclusion the 3 wage variables from our linear model.

```
waldtest(model1_1, model1_2)
```

```
## Wald test
##
## Model 1: crmrte ~ density + taxpc
## Model 2: crmrte ~ taxpc + density + wfed + wcon + wtrd
##   Res.Df Df      F Pr(>F)
## 1     87
## 2     84  3 1.2311 0.3036
```

From this output, we see that the p-value associated with our F-statistic of 1.2311 is 0.3036. This p-value is not less than 0.05 and is thus not significant enough to reject the null hypothesis. As a result, we have sufficient reason to exclude the 3 wage variables (wfed, wcon, wtrd) from our original linear model consisting of density and taxpc.

For good measure, we wanted to also make sure that we are able to exclude all 9 wage variables from our original linear model. Below, we created a 3rd version of a linear model which includes both density and taxpc as independent variables and all 9 wage variables. We then ran the Wald test again:

```r
model1_3 = lm(crmrte ~ density + taxpc + wcon + wtuc + wtrd + wfir + wser +
                wmfg + wfed + wsta + wloc, data=ncrime)
waldtest(model1_1, model1_3)
```

```
## Wald test
##
## Model 1: crmrte ~ density + taxpc
## Model 2: crmrte ~ density + taxpc + wcon + wtuc + wtrd + wfir + wser +
##     wmfg + wfed + wsta + wloc
##   Res.Df Df      F Pr(>F)
## 1     87
## 2     78  9 1.2195 0.2955
```

From the above output, we found that the F-statistic for comparing the restricted model to the unrestricted model with all 9 wage variables is 1.2195 with a p-value of 0.2955. Again, this is not significant enough to reject the null hypothesis. We conclude that we have sufficient reason to exclude all 9 wage variables from our linear model.

To further support our decision to exclude the wage variables, we also created the following table that compares the goodness of fit between the 3 models we have built thus far:

Model Version	**Model Equation**	R^2	Adjusted R^2	AIC
model1_1	$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + u$	0.582296	0.5726936	-530.6358574
model1_2	$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + \beta_3 \cdot wfed + \beta_4 \cdot wcon + \beta_5 \cdot wtrd + u$	0.5998879	0.5760717	-528.508397
model1_3	$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc +$ all 9 wage var $+ u$	0.6338204	0.5821797	-524.4842709

Our criteria for a better fitting model is one that has a greater adjusted R^2 value and a lower AIC. We expect the unadjusted R^2 value to increase from model1_1 to model1_2 and from model1_2 to model1_3 purely from the inclusion of more variables. For this reason, we also calculated the adjusted R^2 values, which only increase if the inclusion of new terms improves the model more than would be expected by chance. From this table, we see that the model including 3 wage variables and the model including all 9 wage variables result in higher AIC values and a negligible increase in adjusted R^2. This means that both model1_2, which includes 3 wage variables, and model1_3, which includes 9 wage variables, do not produce a better fitting model than model1_1 which has no wage variables and achieves the best level of parsimony.

Since it is not representative to include only one wage variable, we found that both the inclusion of a representative subset of wage variables and the inclusion of all wage variables yield no joint significance and worse fitting linear models. In conclusion, we have sufficient evidence from our Wald Tests, AIC and adjustedR^2 calculations that the wage variables should be excluded from our model of explanatory variables.

As a result, we have decided that the final version for model 1 consisting of only explanatory variables, should be model1_1:

$$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + u$$

A regression model with crime rate as the dependent variable and density and taxpc as the only independent variables. Next, we test the CLM assumptions with model1_1.

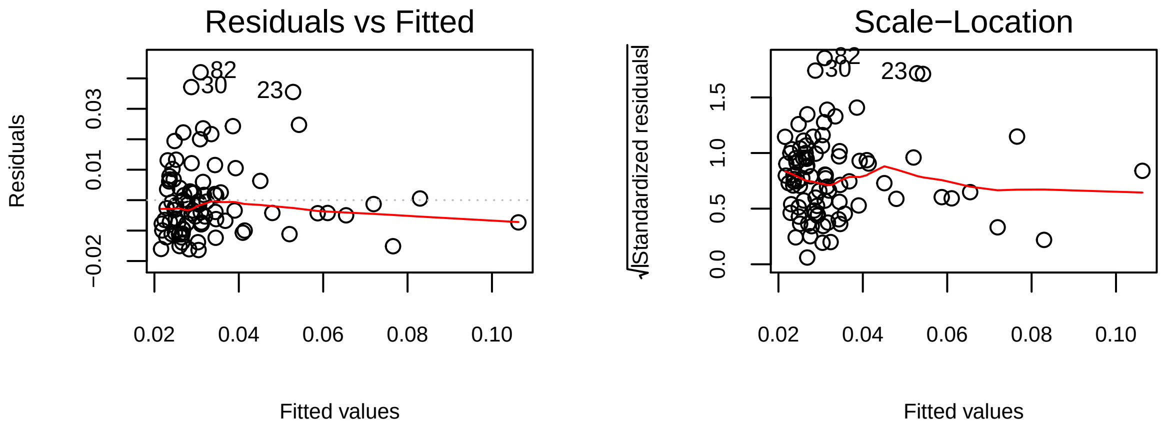
CLM Assumptions Testing on Model 1, Version 1

1. **MLR 1 Linearity in Parameters**: The first assumption is our regression model is linear in parameters, which means that the dependent variables are written as a linear combination of the independent variables. We have met this assumption by creating our model1_1 above which consists strictly of linear terms. There is nothing further to test, since we have not constrained the error term yet.
2. **MLR 2 Random Sampling**: Although we did not collect the data ourselves, we will assume that random sampling methods were used to obtain the sample of North Carolina counties. Our data set consists of 90 counties in 1980, which is very close to the present day count of 100 counties in North Carolina. Since we aim to predict crime rates for North Carolina counties (and not the entire United States, for example), our data is a large enough subset of counties to sufficiently satisfy random sampling. We know that there is no clustering because our counties are chosen from different locations and not all counties are urban in the EDA.

3. **MLR 3 Multicollinearity**: The correlation between taxpc and density is 0.3201, which is sufficient to satisfy the assumption of no perfect collinearity, since $R \neq 1$ or $R \neq -1$.

We will be using the following plots in our discussion of MLR 4 and 5.

```
par(mfrow=c(1,2))
plot(model1_1, which=1, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
plot(model1_1, which=3, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
```



4. **MLR 4 Zero Conditional Mean**: To test for zero conditional mean, which states that $E(u|x_i) = 0$ for all x_i, we plotted the [Residuals vs Fitted] values plot of our model. To check for zero conditional mean, we are examining the red line in the [Residuals vs Fitted] plot, which represents the average of the residuals. If zero conditional mean is met, we should see a relatively straight horizontal line from left to right that lies closely to the residuals = 0 axis line. As we can see in [Residuals vs Fitted] plot above, the red line is relatively horizontal and quite close to the residuals = 0 axis line. On the right side, there is a slight drop to -0.005 due to the right most point, but this could be an artifact of having much fewer data points on the right than on the left. Since there are so few points on the right side used for calculating the average of residuals, we would expect some slight deviations in the residuals average from 0 to occur. Ultimately, given that the deviations from residuals = 0 are not severe, we believe that zero conditional means has been met.

5. **MLR 5 Homoskedasticity**: Homoskedasticity is defined by the residuals having the same variance with respect to all values of the independent variables, and this assumption is met if $var(u|x_1, x_2, ...x_k) = 0$. Graphically we test for homoskedasticity by examining both the [Residuals vs Fitted] Plot and the [Scale Location] Plot above for our model1_1:

- In the [Residuals vs Fitted] plot, we expected to see a band of points of equal thickness of from left to right. In our [Residuals vs Fitted] plot above, We see that the width of the thickness of points in this plot is not the same from left to right, but we also need to consider that we have far more points on the left side than on the right side. While a first glance of [Residuals vs Fitted] shows possible heteroskedasticity, further diagnostics will be needed.
- In the [Scale Location], homoskedasticity is represented by a horizontal band of points from left to right. Compared to the [Residuals vs Fitted] plot, our band of points in the [Scale Location] plot is a bit more even in thickness from left to right, although the left side is clearly wider in thickness than the right. To see if the band of points is flat from left to right, we can examine the whether or not the red line in this plot is horizontal from left to right. In our plot, the red line itself is pretty close to a straight line, with the exception of a sight uptick around the fitted values = 0.45 point.

From looking at the two plots, it is still difficult to conclude the homoskedasticity is met. The [Residuals vs Fitted] plot shows some heteroskedasticity, but the red line is in the [Scale Location] plot shows that the band of points is mostly flat, which aligns with homoskedasticity.

To completely ascertain whether or not homoskedasticity is met, we also run the Breusch-Pagan test:

```
bptest(model1_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1_1
## BP = 0.10109, df = 2, p-value = 0.9507
```

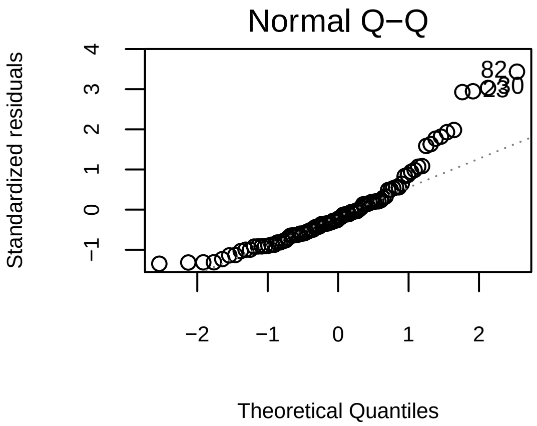
The results of the Breusch-Pagan test shows a p-value of 0.9507, which is not significant enough to reject the null hypothesis of homoskedasticity. We did note that when using the Breusch-Pagan test, we must pay attention to the sample size. For a large sample consisting of 200 points or more, any amount of heteroskedasticity will appear as significant, whereas for small data sets, the test will rarely be significant. Since our data set is 90 points, we questioned whether or not heteroskedasticity will be detected by the bptest. However, our p-value for this bptest is very high, almost close to 1. If our p-value was lower, say <0.2, we may have considered heteroskedasticity does exist even if the test did not return sufficient significant results to reject the null of homoskedasticity. In addition, we know that heteroskedasticity is detectable in our sample size of 90 points by bptest we ran for later iterations of the model below. Considering the large p-value from our bptest and the relatively flat band in the [Scale

Location] Plot, we believe that the homoskedasticity assumption is met which means we can use standard errors to assess the significance of our model coefficients.

Since we have met CLM Assumptions MLR 1-5, we know that the coefficients in our OLS regression model, model1_1, is the best in class of linear unbiased estimators.

6. **MLR 6 Normality of Errors**:

```
par(mfrow=c(1,2))
plot(model1_1, which=2, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
```



The 6th CLM Assumption is that the errors terms in our model are normally distributed $(u_i\ N(0,\sigma^2))$, and the distribution is independent of all values of our independent variables x_i. To assess if our model meets assumption 6, we can take a look at the [Normal Q-Q] plot shown above. The dashed diagonal line represents the trajectory of normally distributed residuals while the points represent the actual residuals of our model. Deviations from the dashed diagonal line, which we see for some of the points on the right, means that our errors do deviate from normality and that this assumption may be violated.

We also ran the Shapiro-Wilk test confirms that the normality assumption is violated with a p-value < 0.001:

```
shapiro.test(model1_1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1_1$residuals
## W = 0.89061, p-value = 1.576e-06
```

In response to the MLR 6 violation, we can use the fact that our data set of 90 points is sufficient large (>30) to rely on the asymptotic properties of the OLS regression. From the async, we know that there is a version of the Central Limit Theorem which states the the OLS estimators are normally distributed for large sample sizes. In this case, we will need to rely on OLS regression asymptotic properties to satisfy the 6th assumption.

Model 1 Conclusions

Now that we have tested assumptions 1-6, we can conclude that OLS coefficients for linear model1_1 are approximately normally distributed. We can then use their standard errors (since MLR 5 of homoskedasticity is satisfied) to assess whether or not the coefficients are significant. From the following output, we see that the coefficient β_1 for density variable is highly significant with p-value $<< 0.001$. The The coefficient β_2 for the taxpc variable is still significant with p-value < 0.01:

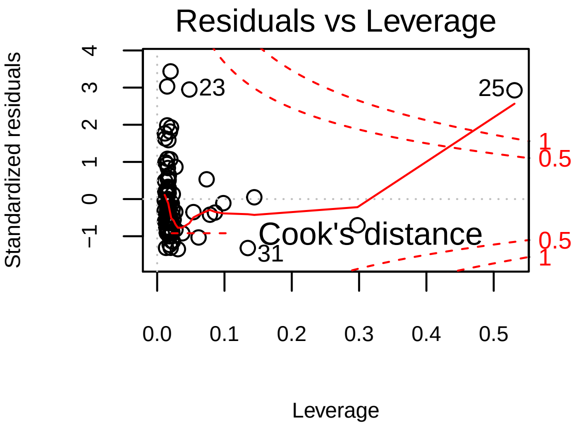
```
summary(model1_1)
```

```
##
## Call:
## lm(formula = crmrte ~ density + taxpc, data = ncrime)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.016390 -0.007844 -0.002918  0.003907  0.042006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0086977  0.0040251   2.161  0.03345 *
## density     0.0080869  0.0009079   8.908 6.94e-14 ***
## taxpc       0.0003460  0.0001054   3.284  0.00148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.01235 on 87 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5727
## F-statistic: 60.64 on 2 and 87 DF,  p-value: < 2.2e-16
```

Now that we have tested all 6 CLM assumptions and measured the goodness of fit of our model, we want to comment further on the fact that β_2, the coefficient of the taxpc variable, is positive in our model 1. As stated earlier, this contradicts our original hypothesis that crime rate and tax per capita should be negatively correlated. Even now that we have a multivariate linear model where we have controlled for density, the relationship between taxpc and crmrte is still positively correlated. One thing to take into consideration is that there is an outlier in the taxpc variable as seen in this diagnostic plot:

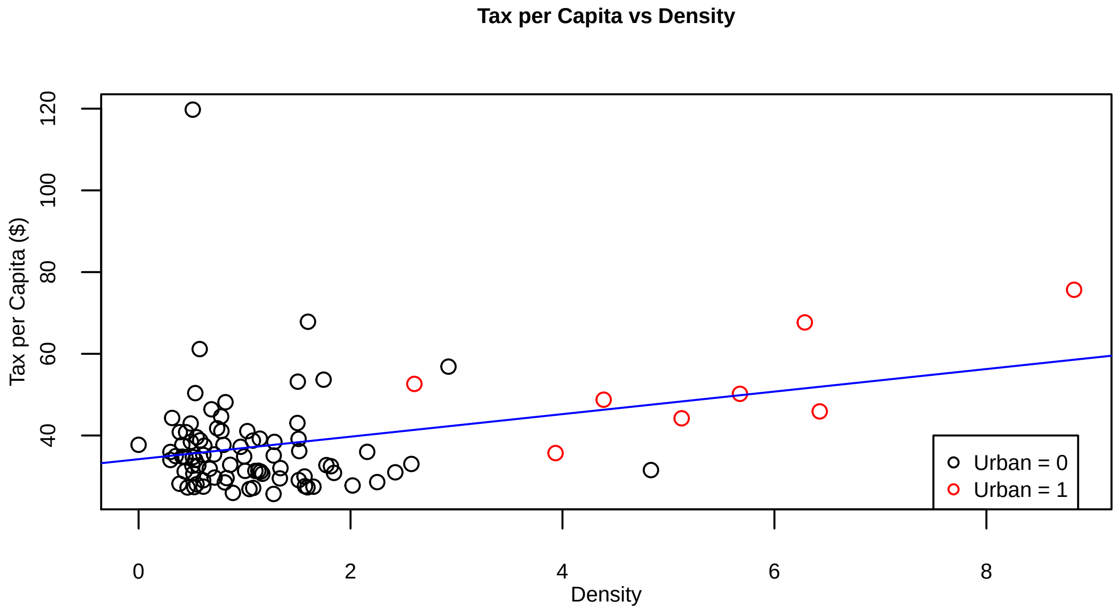
```
par(mfrow=c(1,2))
plot(model1_1, which=5, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
```



From the [Residuals vs Leverage] plot, we see that point number 25 corresponds to county 55 which we noted earlier in the EDA to be an outlier for the taxpc variable. Given that this point has Cook's Distance greater than 1, we know that this point possesses great influence in our regression model. As stated in our original univariate analysis of taxpc, we did not have reason to suspect that the taxpc value of 119.76 was considered erroneous and thus we do not have sufficient reason to remove it from our model. Instead, we wanted to point out that this influential outlier may affect the accuracy of estimate for β_2 and could mean that our estimate deviates from the true value of β_2. However, a Cook's Distance calculation is insufficient for us to conclude whether or not the true value of β_2 is actually positive or negative.

In addition, we can consider an alternative interpretation for a positive β_2 value. One thing to note is that there is a positive correlation between taxpc and density:

```
plot(ncrime$density, ncrime$taxpc,
     main = "Tax per Capita vs Density", xlab = '', ylab = '', col = ncrime$urban+1,
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
legend(7.5, 40, legend=c("Urban = 0", "Urban = 1"),
       col=c("black", "red"), cex=0.7, pch = 1:1)
mtext("Density", side=1, cex=0.7, padj = 4)
mtext("Tax per Capita ($)", side=2, cex=0.7, padj = -4.5)
abline(lm(taxpc ~ density, data = ncrime), col = 'blue')
```



In denser counties where tax per capita is high, there may also be greater diversity in the resident incomes due to the large number of residents. We know that several of the counties with highest densities are also urban, and we know that in general, urban areas contain greater diversity of jobs and therefore wages and incomes. As a result, crime rates may increase even in areas of high tax per capita due to the the possibility that income disparity is greater in denser areas, leading to potential perpetrators committing crimes under the influence of income disparity. We will explore this further in the omitted variables discussion.

In conclusion, our two explanatory variables, density and tax per capita, explain 57.3% of the variation in crime rate and both coefficients in our OLS model are significant and positive. Based on our final version of model 1, crime rate increases by 0.87 crimes per person for a density increment of 100 people per square mile, while controlling for tax per capita. In addition, crime rate increases by 0.35 crimes per person for a tax per capita increment of $1,000 when controlling for density. We ultimately decided to exclude both a subset of the wage variables and all of the wage variables from our linear model as they did not exhibit joint significance.

With regards to informing campaign policy, our regression analysis of model 1 allows us to draw a few conclusions. While density has greater practical significance in affecting crime rate, it is difficult to create policies that change population density of a county. There is also no guarantee that population density decrease will necessarily decrease crime rate as we do not have sufficient evidence here to claim a causal relationship, only that crime rate and density are highly correlated. On the other hand, campaign policies can be formulated to not only change tax per capita but also change the allocation of taxes to resources within each county. We see that taxpc has a positive correlation with both density and crime rate, which means that denser counties have both higher tax per capita and higher crime rates. Higher tax revenue per capita may mean higher potential for these counties to allocate resources towards crime reduction. While we do not have information on tax revenue allocation, we would recommend that campaign policies consider prioritizing contributions of tax revenue towards crime reduction efforts for denser counties where crime rate is high.

Model 2

To further increase the accuracy of our model, we considered introducing criminal justice related covariates, such as probability of arrest (prbarr), probability of conviction (prbconv), probability of prison (prbpris), police per capita (polpc), and average prison sentence (avgsen). We suspect that they cannot be explanatory or causal variables given that it is unclear how well they can affect the behavior of perpetrators even if we apply a rational choice theory framework. Such statistics may not be well known by the public and perpetrators may rely more on anecdotal knowledge of police presence and how likely they are to get caught, convicted, and sentenced. In constructing Model 2, we will specifically focus on the effects of adding prbarr and prbconv since they had higher correlations with crmrte.

We will also include polpc although bivariate analysis did not show a strong correlation with crmrte. Research has been inconclusive in terms of the impact of larger police forces (Lim, Lee, & Cuvelier, 2010). However, we see its importance in terms of potential policy recommendations regarding resource allocation.

Lastly, we explore the effects of the location indicator variables since we suspect there may be regional differences in crime rate.

A table summarizing the results of specifications for Model 2 is located at the end of this section for reference.

Model 2_1: Including prbarr

First, we will introduce prbarr and examine how it affects our model.

```
model2_1 = lm(crmrte ~ taxpc + density + prbarr, data = ncrime)
```

This version of model 2 shows no violations of the CLM assumptions aside from a lack of normality in the errors as we have seen with previous model specifications (please see .rmd file for full set of test on CLM assumptions). With a sample size of 90, we are not overly concerned with this violation given that we can rely on asymptotic properties.

The introduction of prbarr yields an adjusted R^2=0.5989 and AIC=-535.3614, which is an improvement from our results from model 1. As we expect, the coefficient for prbarr is negative, which means it contributes to a reduction in crmrte. All of the coefficients are statistically significant with $p < 0.05$ (see table below). In particular, prbarr has practical significance since a one unit change in prbarr results in an approximately reduction of 2.5 crimes per 100 people.

Model 2_2: Including prbarr and prbconv

Next, we will introduce prbconv into our model specification.

```
model2_2 = lm(crmrte ~ taxpc + density + prbarr + prbconv, data = ncrime)
```

Similar to the previous version, there are no violations of the CLM assumptions aside from the lack of normality of the errors (please see .rmd file for full set of test on CLM assumptions).

The introduction of prbconv yields an adjusted R^2=0.6543 and AIC=-547.7948, which is an improvement from our previous model with just prbarr included. The coefficient for prbconv, like prbarr, is also negative as we would expect. All of the coefficients are statistically significant with $p < 0.05$ (see table below). We would argue that the coefficient for prbconv is also practically significant since the order of magnitude, and thus units, are comparable to the coefficient for prbarr.

Model 2_3: Including prbarr, prbconv, and polpc

Finally, we utilize the previous model and include polpc as well.

```
model2_3 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc, data = ncrime)
```

Like the previous two models, there is a violation of normality of errors (please see .rmd file for full set of test on CLM assumptions). However, this model also has a violation of homoskedasticity, so in response, we will use heteroskedastic-robust standard errors to examine statistical significance.

The introduction of polpc yields an adjusted R^2=0.696 and AIC=-558.4386, which is a greater improvement than the previous iterations. The coefficient for polpc (β_5=5.2678) is surprising, not only in that it is positive, but also that it is relatively larger than the coefficients for the other variables. The coefficient for taxpc has lost statistical significance in this model, but all of the other coefficients are statistically significant with $p < 0.05$. We would also argue that the estimated coefficient of the police per capita variable has high practical significance, since it is 2 orders of magnitude greater than that of both prbarr and prbconv. We see that an increase in polpc by 1 is associated with an estimated increase of ~5 crimes per person per our regression model.

Please see the following table for a summary of measures of fit for the discussed specifications. We also included the fit measurements from the final version of model 1 for comparison.

Model Version	**Model Equation**	Adjusted R^2	AIC
Final Model 1	$crmrte = \beta_0+\beta_1\cdot density+\beta_2\cdot taxpc+u$	0.5726936	-530.6358574
model2_1	$Model1 + prbarr$	0.5988508	-535.3614346
model2_2	$Model1 + prbarr + prbconv$	0.6542697	-547.7947705
model2_3	$Model1 + prbarr + prbconv + polpc$	0.6960063	-558.4385577

Examining Indicator Variables

We also wanted to consider potential regional differences by using the indicator variables in the data set (urban, west, central, notwestcen) as covariates. Earlier in the EDA, we noticed that the effect of central and notwestcen on crime rate is rather random and suspected that they could be excluded from our models. Here, we further test our theory by regressing crmrte on each respective indicator variable. Only regressions with urban and west (which we previously categorized as non-random indicators) yielded statistically significant results as seen in the table below. As a result, we have decided to exclude central and notwestcen from further iterations of our regression model.

We included further details on calculating the statistical significance of the coefficient of each of the indicator variables on our dependent crime rate variable in the .rmd. Here is a summary of our findings:

Model Equation	t Statistic	p-value
crmrte ~ urban	7.318	1.12e-10
crmrte ~ west	-3.454	0.000851
crmrte ~ central	1.578	0.118

Model Equation	t Statistic	p-value
crmrte ~ notwestcen	1.569	0.12

```
urban_I = lm(crmrte ~ urban, data = ncrime)
urban_I$coefficients
```

```
## (Intercept)       urban
##  0.02990170  0.04059258
```

In this model, the intercept ($\beta_0 = 0.0299$) is the average crime rate for non-urban counties. The coefficient for urban (β_1=0.0406) represents the difference between average crime rate for urban and non-urban counties. This is in line with what we would expect in that urban counties tend to have higher crime rates than non-urban counties.

```
west_I = lm(crmrte ~ west, data = ncrime)
west_I$coefficients
```

```
## (Intercept)        west
##  0.03720145 -0.01510170
```

In this model, the intercept ($\beta_0 = 0.0372$) is the average crime rate for non-western counties. The coefficient for west (β_1=-0.0151) represents the difference between average crime rate for western and non-western counties. This confirms what we saw in our previous exploratory data analysis in that western counties tend to have a lower crime rate compared to non-western counties.

Based on these results, we will include specifications with urban and west respectively.

Including Urban Indicator Variable

```
model2_4 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + urban, data = ncrime)
```

```
summary(model2_4)$adj.r.squared
```

```
## [1] 0.6927108
```

```
AIC(model2_4)
```

```
## [1] -556.546
```

This model shows evidence of heteroskedasticity, so we will be using heteroskedastic-robust standard errors to examine statistical significance. Compared to that of our original model, the introduction of urban to the model specification does not show an improvement in adjusted R^2 or AIC. In this model, the coefficients for urban and taxpc are not statistically significant. We also noted in the EDA that urban and density share a strong relationship, which may reduce precision if they are included in the same model. For these reasons, we will not incorporate urban into our model going forward.

Including West Indicator Variable

```
model2_5 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + west, data = ncrime)
```

```
summary(model2_5)$adj.r.squared
```

```
## [1] 0.7495835
```

```
AIC(model2_5)
```

```
## [1] -574.9657
```

According to the Breusch-Pagan test, this model nears statistical significance ($p = 0.054$) for rejecting the null hypothesis of homoskedasticity. Thus, to be conservative, we will use heteroskedastic-robust standard errors to examine statistical significance of the coefficients.

This model with the west indicator variable shows evidence of better fit with an adjusted R^2=0.7496 and AIC = -574.9657, which is an improvement from our model2_3. Even with the conservative heteroskedastic-robust standard errors, all of the coefficients are statistically significant, excluding that of taxpc, with $p < 0.001$.

The coefficient for west is negative as we expected from our regression of crmrte on west. While many of the coefficients have not change much in size, we noticed that the coefficient for polpc increased by over 1, which may be practically significant.

Summary of Model 2

The table below summarizes our efforts with model specification as described in the above sections.

```
se.model2_1 = coeftest(model2_1)[ , "Std. Error"]
se.model2_2 = coeftest(model2_2)[ , "Std. Error"]
se.model2_3 = sqrt(diag(vcovHC(model2_3)))
se.model2_4 = sqrt(diag(vcovHC(model2_4)))
```

```r
se.model2_5 = sqrt(diag(vcovHC(model2_5)))

stargazer(model2_1, model2_2, model2_3, model2_4, model2_5,
          type = "latex", title = "Versions of Linear Model 2",
          omit.stat = c("f","n"), se = list(se.model2_1,
          se.model2_2, se.model2_3, se.model2_4, se.model2_5),
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC",
          round(AIC(model2_1),1), round(AIC(model2_2),1), round(AIC(model2_3),1),
          round(AIC(model2_4),1), round(AIC(model2_5),1))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Mon, Dec 10, 2018 - 19:07:01

Table 4: Versions of Linear Model 2

	Dependent variable:				
	crmrte				
	(1)	(2)	(3)	(4)	(5)
taxpc	0.0003^{**}	0.0003^{**}	0.0002	0.0002	0.0001
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
density	0.007^{***}	0.007^{***}	0.006^{***}	0.005^{**}	0.005^{***}
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
prbarr	-0.025^{**}	-0.030^{**}	-0.052^{***}	-0.052^{***}	-0.051^{***}
	(0.010)	(0.009)	(0.013)	(0.014)	(0.012)
prbconv		-0.013^{***}	-0.018^{***}	-0.018^{***}	-0.018^{***}
		(0.003)	(0.004)	(0.004)	(0.004)
polpc			5.268^{**}	5.287^{**}	6.356^{***}
			(1.715)	(1.755)	(1.664)
urban				0.002	
				(0.010)	
west					-0.011^{***}
					(0.002)
Constant	0.017^{***}	0.028^{***}	0.034^{***}	0.035^{***}	0.039^{***}
	(0.005)	(0.006)	(0.008)	(0.009)	(0.008)
AIC	-535.4	-547.8	-558.4	-556.5	-575
R^2	0.612	0.670	0.713	0.713	0.766
Adjusted R^2	0.599	0.654	0.696	0.693	0.750
Residual Std. Error	0.012 (df = 86)	0.011 (df = 85)	0.010 (df = 84)	0.010 (df = 83)	0.009 (df = 83)
Note:					*p<0.05; **p<0.01; ***p<0.001

Our final version of model 2, a more accurate multilinear model, has the following form

$$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + \beta_3 \cdot prbarr + \beta_4 \cdot prbconv + \beta_5 \cdot polpc + \beta_6 \cdot west + u$$

with the following coefficients:

```r
model1_1$coefficients
```

```
##  (Intercept)      density        taxpc
## 0.0086976849 0.0080869319 0.0003459569
```

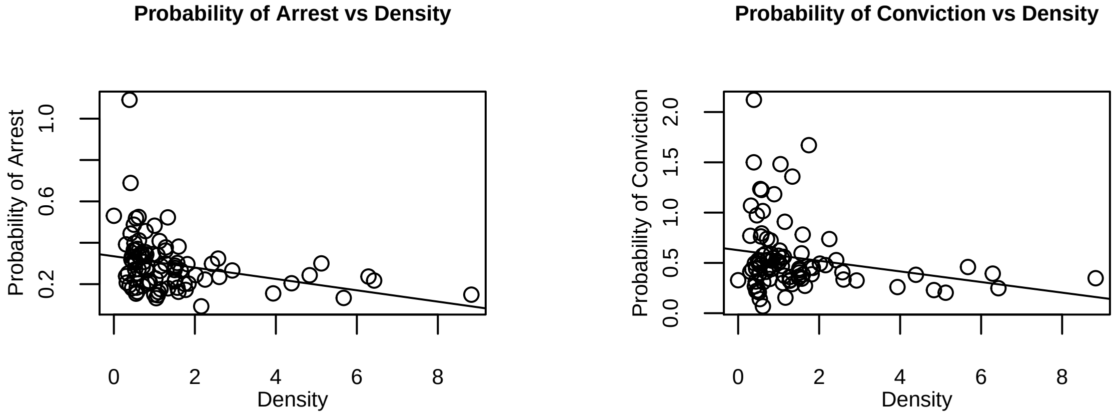
```
model2_5$coefficients
```

```
##   (Intercept)         taxpc       density        prbarr       prbconv 
##  0.0386964186  0.0001135914  0.0052769297 -0.0514567715 -0.0184690215 
##         polpc          west 
##  6.3563657705 -0.0105866949
```

This model has an adjusted R^2 of 0.7496, meaning 75% of the variation in crmrte is explained by our model after adjusting for the number of predictors. Of all of the versions we attempted in the creation of model 2, this model (model2_5 or version 5) has both the best fit measured by the lowest AIC and clearly the highest adjusted R^2. Compared to our model 1, the coefficients of taxpc and density have decreased in size with the introduction of these criminal justice related covariates and the west location indicator variable. As mentioned taxpc loses statistical significance in this iteration of the model as well. This suggests that these covariates have more of an impact that we previously expected in theorizing our models.

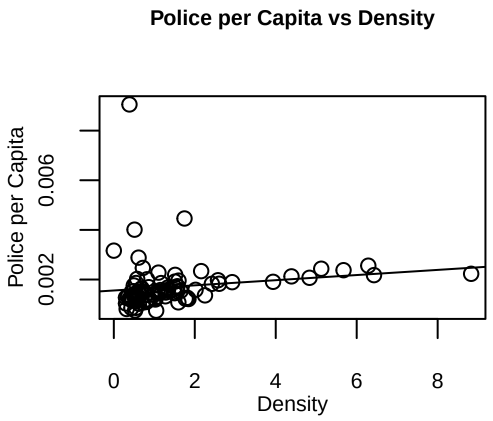
Considering the practical significance of prbarr, prbconv, and polpc, we examined their respective bivariate relationships to density in order to inform policy recommendations.

```
par(mfrow=c(1,2))
plot(ncrime$density, ncrime$prbarr,
     main = "Probability of Arrest vs Density", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Density", side=1, cex=0.7, padj = 4)
mtext("Probability of Arrest", side=2, cex=0.7, padj = -4.5)
abline(lm(prbarr ~ density, data = ncrime))

plot(ncrime$density, ncrime$prbconv,
     main = "Probability of Conviction vs Density", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Density", side=1, cex=0.7, padj = 4)
mtext("Probability of Conviction", side=2, cex=0.7, padj = -4.5)
abline(lm(prbconv ~ density, data = ncrime))
```


Probability of Arrest vs Density

Probability of Conviction vs Density

```
par(mfrow=c(1,2))
plot(ncrime$density, ncrime$polpc,
     main = "Police per Capita vs Density", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Density", side=1, cex=0.7, padj = 4)
mtext("Police per Capita", side=2, cex=0.7, padj = -4.5)
abline(lm(polpc ~ density, data = ncrime))
```



The correlations with density for prbarr (R=-0.3027) and prbconv (R=-0.2267) are medium in effect size, but they are surprisingly negative. In the [Probability of Arrest vs Density] and [Probability of Conviction vs Density] plots above, we actually see that the prbarr and prbconv values for denser urban counties are actually comparable to less dense counties since the slopes in the regression lines are quite shallow This may not be a desirable outcome since crmrte rates are higher in denser urban counties, so the goal should be to achieve arrest rates and conviction rates that scale positively with density, in order to reduce crime in these counties.

Given that crime rates are negatively correlated with prbarr and prbconv, even in our multivariate model, we know that harsher practices in law enforcement, meaning higher probability of arrest and conviction, could be correlated to reduction in crime rates. In order to achieve higher arrest and conviction rates, resources such as personnel (police per capita) and tax revenue would need to be allocated to denser counties. However, we find that the correlation between polpc and density R=0.1591) to be small. While there is an overall positive correlation, it may be that the increase in police per capita per increase in density is too insufficient to increase arrest and conviction rates accordingly. In other words, there there may be not enough police per capita in denser counties to influence crime reduction and therefore policies can be developed to shift resources toward increasing the police per capita in denser counties.

Model 3

In our construction of model 3, we first decided to add the covariates of pctymle (percent young male between the ages of 15 and 24) and pctmin80 (percent minority). In our EDA, we discovered that both of these variables had higher bivariate correlation with crime rate than the variable polpc (police per capita) which we included in model 2. Both variables were previously excluded because we felt that they could not be categorized as explanatory variables nor could they be categorized as variables that could influence policy in a meaningful way.

We constructed 2 more models by respectively adding pctymle and pctmin80 to our best version of model 2.

```
model3_1 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + west + pctymle, data = ncrime)
model3_2 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + west + pctmin80, data = ncrime)
```

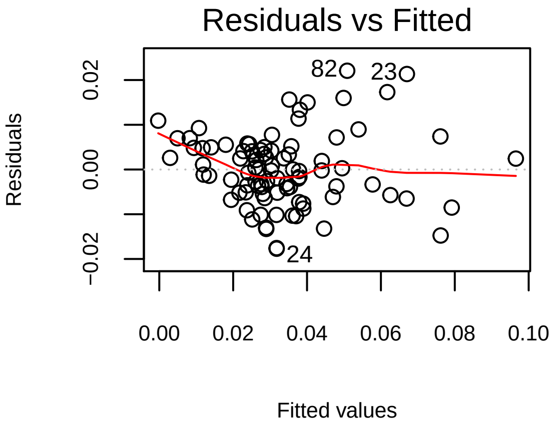
We then constructed the following table to compare measures of fit between the best version of Model 1, best version of Model 2 and the 2 models we constructed here:

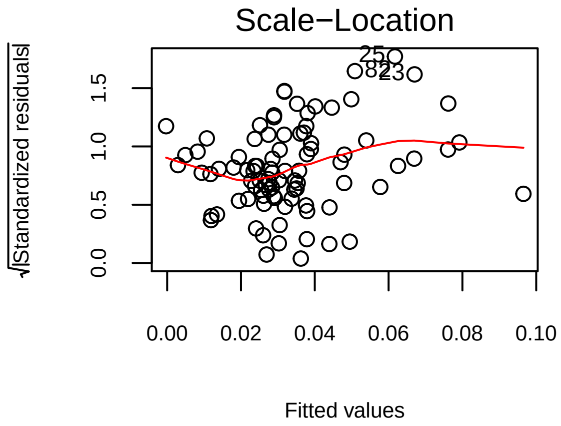
Model Version	**Model Equation**	Adjusted R^2	AIC
Final Model 1	$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + u$	0.5726936	-530.6358574
Final Model 2	$Model1 + \beta_3 \cdot prbarr + \beta_4 \cdot prbconv + \beta_5 \cdot polpc + \beta_6 \cdot west$	0.7495835	-574.96573
model3_1	$Model2 + \beta_7 \cdot pctmyle$	0.7589113	-577.4731157
model3_2	$Model2 + \beta_7 \cdot pcctmin80$	0.7982818	-593.5195341

We discovered that the inclusion of pctymle variable as a 7th independent variable alone does not show significant increase in adjusted R^2 from model 2 and a rather negligible improvement in fit as seen by a small change in AIC. However, when we include pctmin80 as a 7th independent variable, we find that not only do we achieve a better adjusted R^2 but also a jump in AIC from -574.9657 in model 2 to -593.5195. The change in adjusted adjusted R^2 due to pctmin80 is small compared to the difference between model 1 and model 2, but we found this sudden decrease in AIC to be rather interesting.

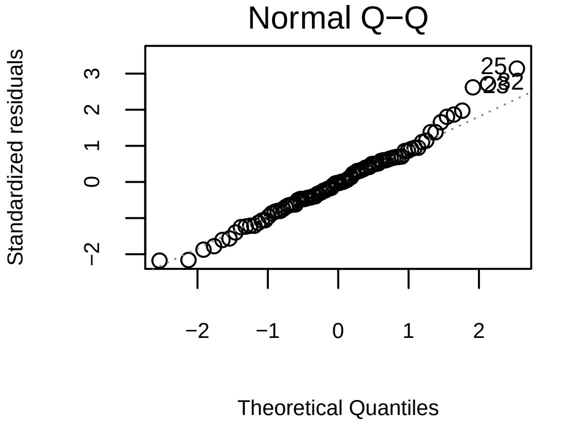
Next, we decided to test if the coefficient for pctmin80 in our model is significant and if this new version of the model has significant coefficients for the other independent variables we have included. To do so, we first performed a test on the 6 CLM Assumptions of model3_2, our best fitting model thus far with the inclusion of pctmin80.

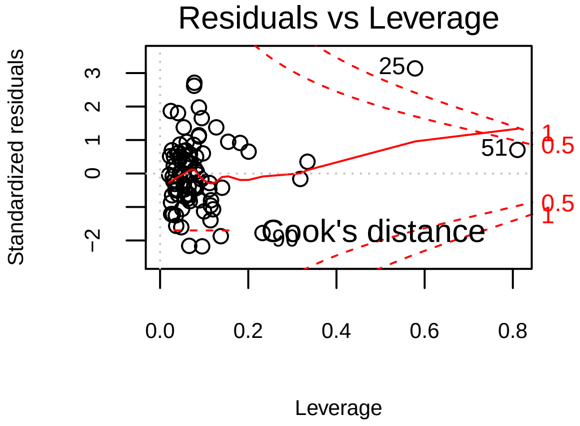
```r
par(mfrow=c(1,2))
plot(model3_2, which=1, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
plot(model3_2, which=3, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
```





```r
par(mfrow=c(1,2))
plot(model3_2, which=2, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
plot(model3_2, which=5, cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
```





```r
bptest(model3_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3_2
## BP = 17.32, df = 7, p-value = 0.01545
```

```r
shapiro.test(model3_2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3_2$residuals
## W = 0.98635, p-value = 0.4722
```

Based our diagnostic plots as well as the results from the Breush-Pagan and Shapiro-Wilk tests, we see no violation of zero-conditional mean or normality of errors, but there is some evidence of heteroskedasticity since there is a significant result (p-value<0.05) in the bptest. Therefore, we will examine coefficients' significance in this model using heteroskedastic-robust errors:

```r
coeftest(model3_2, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  2.9276e-02  6.7733e-03  4.3223 4.310e-05 ***
## taxpc        1.3413e-04  2.3842e-04  0.5626 0.5752463
## density      5.3694e-03  1.4093e-03  3.8101 0.0002674 ***
## prbarr      -6.0961e-02  1.2148e-02 -5.0183 2.978e-06 ***
## prbconv     -2.0466e-02  3.7151e-03 -5.5087 4.059e-07 ***
## polpc        7.3343e+00  1.9429e+00  3.7750 0.0003016 ***
## west        -1.8352e-03  2.4869e-03 -0.7380 0.4626513
## pctmin80     3.3446e-04  7.7203e-05  4.3323 4.153e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

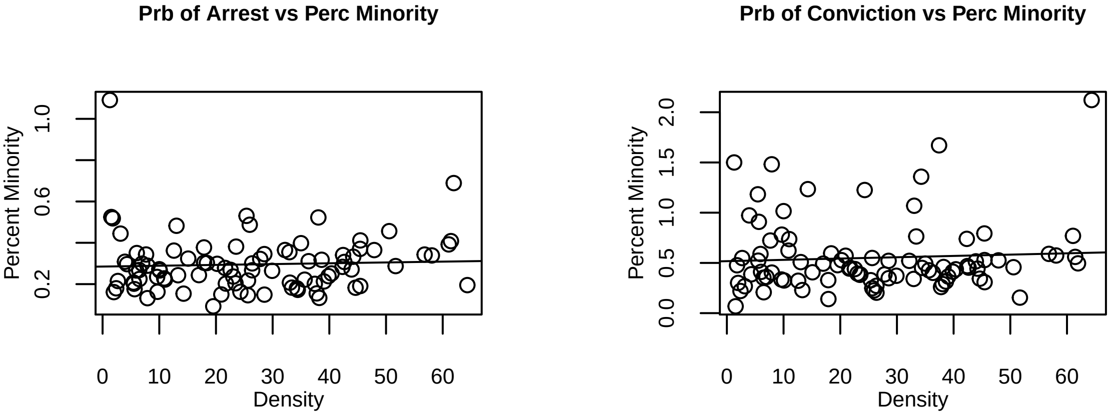
In this result, we see that the coefficient for the pctmin80 covariate is highly significant. The loss of significance of taxpc's coefficient already occurred prior to the inclusion of pctmin80. However, we see that the inclusion of pctmin80 did lead to the loss of significance of the coefficient for the west variable.

The eventual result of the loss of significance taxpc's coefficient is perhaps not surprising. Since model 1, we have noticed that data point 25, which corresponds the extreme taxpc outlier, has consistently appeared in our [Residuals vs Leverage] diagnostic plots as having Cook's Distance greater than 1. Since model 1, we had suspected that taxpc's outlier could introduce error, and therefore greater variance, to the estimate of β_2, which has the potential to result in an insignificant estimate.

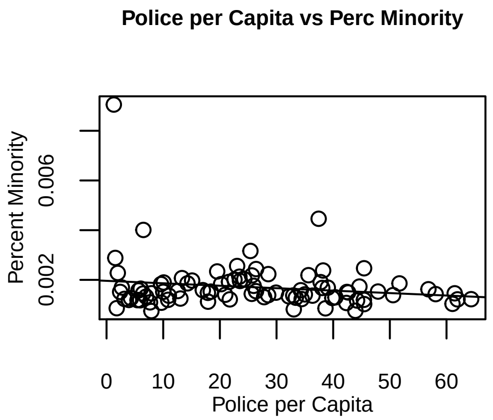
While the inclusion of pctmin80 shows a significant β_7, we do not feel that the estimated coefficient is of practical significance. An increase in minority of 10% is only correlated with and increase of 0.003 crimes per person. This coefficient is one order of magnitude less than that of other coefficients in our linear model and pretty comparable to taxpc, which was already arguably not practically significant.

As shown in the plots below, pctmin80 has weak relationships with the criminal justice related covariates that we included. We note that the large shift in AIC is interesting, but we are unsure as to what underlying mechanism may be responsible for it. We also recognize that in southern US during the 1980s, racial bias and prejudice may a greater factor when considering crime rate and minorities. Based on our background knowledge, it is reasonable to believe that minorities at that time may have been differentially punished by the criminal justice system as well as apprehended for crimes they may not have committed.

```r
par(mfrow=c(1,2))
plot(ncrime$pctmin80, ncrime$prbarr,
     main = "Prb of Arrest vs Perc Minority", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Density", side=1, cex=0.7, padj = 4)
mtext("Percent Minority", side=2, cex=0.7, padj = -4.5)
abline(lm(prbarr ~ pctmin80, data = ncrime))

plot(ncrime$pctmin80, ncrime$prbconv,
     main = "Prb of Conviction vs Perc Minority", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Density", side=1, cex=0.7, padj = 4)
mtext("Percent Minority", side=2, cex=0.7, padj = -4.5)
abline(lm(prbconv ~ pctmin80, data = ncrime))
```



```r
par(mfrow=c(1,2))
plot(ncrime$pctmin80, ncrime$polpc,
     main = "Police per Capita vs Perc Minority", xlab = '', ylab = '',
     cex.main = 0.7, cex.lab = 0.7, cex.axis = 0.7)
mtext("Police per Capita", side=1, cex=0.7, padj = 4)
mtext("Percent Minority", side=2, cex=0.7, padj = -4.5)
abline(lm(polpc ~ pctmin80, data = ncrime))
```



To test the robustness of this iteration of our model, we included other variables to the model and examined changes in measures of fit.

```
model3_3 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + west + pctmin80
              + mix, data = ncrime)
model3_4 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + west + pctmin80
              + mix + prbpris, data = ncrime)
model3_5 = lm(crmrte ~ taxpc + density + prbarr + prbconv + polpc + west + pctmin80
              + mix + prbpris + avgsen, data = ncrime)
```

Model Version	**Model Equation**	Adjusted R^2	AIC
model3_2	$Model2 + \beta_7 \cdot pctmin80$	0.7982818	-593.5195341
model3_3	$Model2 + \beta_7 \cdot pctmin80 + \beta_8 \cdot mix$	0.8040107	-595.2168541
model3_4	$Model2 + \beta_7 \cdot pctmin80 + \beta_8 \cdot mix + \beta_9 \cdot prbpris$	0.8015873	-593.2288683
model3_5	$Model2 + \beta_7 \cdot pctmin80 + \beta_8 \cdot mix + \beta_9 \cdot prbpris + \beta_10 \cdot avgsen$	0.7999654	-591.6282675

We can then conclude that model3_2 is robust against the inclusion of all other variables since we do not see any meaningful changes in adjusted R^2 or AIC in these model specifications.

Our final version of model 3 is the following

$$crmrte = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot taxpc + \beta_3 \cdot prbarr + \beta_4 \cdot prbconv + \beta_5 \cdot polpc + \beta_6 \cdot west + \beta_7 \cdot pctmin80 + u$$

Omitted Variables Discussion

We considered several potential omitted variables in our model specifications, and how they may cause bias in the coefficients of our explanatory variables, density and taxpc.

Our base model is:

$$y = \beta_0 + \beta_1 density + \beta_2 taxpc + u$$

In the following sections, we will discuss how specific variables affect our base model estimates. We discuss each omitted variable as if it would take the position of a third independent variable, represented by x, as show in the equation below.

$$y = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 x + u$$

Next, we derive the omitted variable bias terms treating x as the omitted variable. To determine the omitted variable bias, we first regress x on the other independent variables and substitute it back into the regression model.

$$x = \delta_0 + \delta_1 density + \delta_2 taxpc + \nu$$

This results in the following form for a true model which includes the omitted variable.

$$\begin{aligned} y &= \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3\delta_0 + \delta_1 density + \delta_2 taxpc + \nu + u \\ &= \beta_0 + \beta_3\delta_0 + (\beta_1 + \beta_3\delta_1) density + (\beta_2 + \beta_3\delta_2) taxpc + \beta_3\nu + u \end{aligned}$$

Thus, the omitted variable bias for density is $\beta_3\delta_1$ and for taxpc is $\beta_3\delta_2$. β_3 represents the coefficient of each of the omitted variables discussed below, if the variable had been included in a regression model with crime rate.

Median Income of County

In developing our base model, we wanted to include a measure of county wealth or prosperity. Given the data set and the results of our analysis, we chose to proxy this with tax per capita (taxpc). However, median income of county could be a more effective measure of this given that taxpc could be related to other factors outside of income alone. Although we were provided with weekly wages for 9 industries, this did not provide a way to examine the expected income of residents.

If we expect median income and crime rate are negatively correlated, or $\beta_3 < 0$, and median income and density to be positively correlated $\delta_1 > 0$, then the omitted variable bias due to median income would be $\beta_3\delta_1 < 0$. Given that β_1, the estimated coefficient for density in our model, is positive, we expect the omitted variable of median income causes a downward bias in this OLS coefficient toward zero.

With regards to taxpc, we also expect a positive correlation with median income, so $\delta_2 > 0$, and the omitted variable bias is $\beta_3\delta_2 < 0$. β_2, the estimated coefficient for taxpc, is also positive and we also expect the omitted variable causes a downward bias in this coefficient toward zero.

Income Inequality

Recognizing that although the wealth of a county may be important for predicting crime, we must also consider economic inequality within the county as well. Many major urban cities have great wealth, but we would be remiss not to acknowledge the differences between high-income and lower-income residents and how this may contribute to societal or environmental pressures that foster crime. Since there are many ways to operationalize income inequality, we are unsure what would be the most appropriate statistic would be. One potential measure could be the Gini Index, which examines the dispersion of income (Desilver, 2015).

We predict that an increase in the dispersion of income will lead to an increase in crime rate, so $\beta_3 > 0$. Given that income dispersion is greater in areas of higher density, we expect $\delta_1 > 0$, which means the omitted variable bias for β_1 is $\beta_3\delta_1 > 0$. This causes a upward bias away from zero in our estimated coefficient for density.

For taxpc, we predict positive correlation with income dispersion leading to $\delta_2 > 0$ under the assumption that wealthier counties have have higher diversity of incomes. We do recognize, however, if the data were more granular, say collected for cities and towns of a certain area where income is more homogeneous, then it would be possible for δ_2 to be negative. In the case of data collected for counties, $\beta_3\delta_2 > 0$ causes a upward bias away from zero in the coefficient β_2 for taxpc. Initially, the estimate for taxpc surprised us because it was positive, but the omitted variable bias from income inequality could explain why we see this positive correlation between taxpc and crime rate, instead of expected negative relationship, in our estimated models.

Unemployment Rate

Similarly, unemployment rate by county may be a key omitted variable because it shows another aspect of the counties that may be masked by wages or even median income. While the data set included weekly wages for 9 industries, we could not discern how many people were employed by each. Counties with high unemployment rates have less legal opportunities for its residents to earning a living. By rational choice theory, they may turn to crime assuming that perpetrators of crime are motivated by financial incentives.

Following the logic of crimes being possibly motivated by financial incentives, we expect $\beta_3 > 0$, where an increase in unemployment rate would correlate with an increase in crime rate. $\delta_1 < 0$ because we expect areas of higher density (such as cities for instance) to have greater opportunities for employment and therefore lower unemployment rates. This leads to the omitted variable bias for β_1, the coefficient of density, to be $\beta_3\delta_1 < 0$. Given that our estimated β_1 is positive, we expect the omitted variable of unemployment rate causes a downward bias toward zero in the OLS coefficient for density.

We would naturally expect that taxpc is also lower in counties where unemployment is high, thus $\delta_2 < 0$, and the omitted variable bias for β_2 is $\beta_3\delta_2 < 0$, which causes a downward bias in taxpc's coefficient toward zero.

Average Years of Education of Residents

Along the same lines, average years of education for county residents may also be an omitted variable that could have strong correlations with crime rate. In fact, it may also be related to median income and unemployment rate as well. Those with more education tend to have higher incomes, may be less likely to be unemployed, and less likely to engage in social circles that engage in unlawful activities.

Therefore, we predict that $\beta_3 < 0$, as we would expect crime rate to reduce due to an increase in average years of education. We expect that average years of education is likely positively correlated with density, as denser regions may have more schools and more opportunities for education, so $\delta_1 > 0$. This means that the omitted variable bias on β_1 is $\beta_3\delta_1 < 0$ due to the average years of education variable.

We would expect that taxpc and average years of education to also be positively correlated, since better educated residents would lead to an overall wealthier county. In addition, some tax revenue may be allocated towards local schools, so we would also expect positive correlation in the association of an increase in taxpc with an increase in more educational opportunities and higher average years of education. Therefore, $\delta_2 > 0$, and the omitted variable bias is $\beta_3\delta_2 < 0$.

Since β_1 and β_2 are both positive and both omitted variable biases are negative, we expect that the omitted variable bias for average years of education to cause a downward bias in our OLS coefficients toward zero.

Percentage of Juvenile Delinquents

We also considered the effect of perpetrators that were involved with crime before adulthood. However, it is possible that this omitted variable would be auto-correlated with average years of education since it is likely that juvenile delinquents go on to achieve less years of education. With the given data set, percentage of young males (ages 15-24) is a partial proxy for this variable since juvenile delinquents tend to be male, but percentage of juvenile delinquents is more precise in that these minors have committed some kind of offense whereas that may not be the case with young males in general.

We expect the percentage of juvenile delinquents to be positively correlated with crime rate, or $\beta_3 > 0$, because juvenile delinquency may be more likely for minors raised in areas where crimes are pervasive. For density, we expect $\delta_1 > 0$, purely from the fact that juvenile delinquency is more likely in areas with more people and more diversity, so $\beta_3\delta_1 > 0$ corresponding to an upward bias away from zero in our estimated coefficient for density.

If we can assume that crimes by minors could also financially motivated, then we predict a negative relationship with taxpc, resulting in $\delta_2 < 0$ and $\beta_3\delta_1 < 0$. This causes a downward bias in our estimated coefficient for taxpc towards zero

Percentage of Recidivism

Percentage of recidivism or repeat offenders is also an important omitted variable to consider. Although it may not be directly causal, the amount of recidivism may highlight the effectiveness of the criminal justice system and law enforcement in a county. Higher recidivism may also be related to higher crime rates and therefore $\beta_3 > 0$. This also provides an opportunity for policy reform in this area, particularly in the areas of criminal justice policies. Unfortunately, there was no proxy for this in the given data set.

We predict density to be positively correlated with percentage of recidivism, $\delta_1 > 0$, which aligns with our earlier argument that crime is more likely to occur in denser areas due to higher presence of opportunity and influence for both first-time and repeat offenders. Therefore, the omitted variable bias on β_1 due to percentage of recidivism is $\beta_3\delta_1 > 0$, which would cause an upward bias away from zero in our estimated coefficient for density.

On the other hand, we predict that the percentage of recidivism is lower for wealthier counties and that $\delta_2 < 0$ resulting in a bias of $\beta_3\delta_1 < 0$. This causes a downward bias in our estimated β_2 for taxpc coefficient towards zero.

Summary of Omitted Variable Bias

The following table summarizes our discussion above on omitted variable bias.

Omitted Variable (x)	corr. with crmrte(β_3)	corr. with density(δ_1)	OMVB on β_1 ($\beta_3\delta_1$)	corr. with taxpc (δ_2)	OMVB on β_2 ($\beta_3\delta_2$)
Median Income	<0 (-)	>0 (+)	(-) towards 0	>0 (+)	(-) towards 0
Dispersion of Income	>0 (+)	>0 (+)	(+) away from 0	>0 (+)	(+) away from 0
Unemployment Rate	>0 (+)	<0 (-)	(-) towards 0	<0 (-)	(-) towards 0
Avg Years of Educ	<0 (-)	>0 (+)	(-) towards 0	>0 (+)	(-) towards 0
% Juvenile Delinquents	>0 (+)	>0 (+)	(+) away from 0	<0 (-)	(-) towards 0
% Recidivism	>0 (+)	>0 (+)	(+) away from 0	<0 (-)	(-) towards 0

One of the more interesting conclusions from this table is that 5 of the 6 omitted variables we identified drive the estimate of β_2, the coefficient on tax revenue per capita, downwards towards 0. On the other hand the omitted variable biases of only half of these variables drive the density coefficient β_1 towards 0. While we do not have information on the significance of these omitted variables in our regression model, the difference of their aggregate result on the estimated coefficients for density and taxpc did catch our attention. The fact that so many omitted variables result in driving β_2 downward towards 0 leads us to believe that it is highly likely that the estimate for β_2 in our model is lower than its true value. Given that the true β_2 would be more greater than the estimated β_2, this would mean that the relationship between crime rate and tax revenue per capita, is more positive than OLS regression predicts. This result further contradicts our initial hypothesis that crime rate and tax revenue per capita would share a negative correlation.

Conclusion

Please see the table below for the final versions of our model specifications:

```
se.model1_1 = coeftest(model1_1)[ , "Std. Error"]
se.model2_5 = sqrt(diag(vcovHC(model2_5)))
se.model3_2 = sqrt(diag(vcovHC(model3_2)))

stargazer(model1_1, model2_5, model3_2,
          type = "latex", title = "3 Final Versions of Model Specifications",
          omit.stat = c("f","n"), se = list(se.model1_1,
          se.model2_5, se.model3_2),
          star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC",
          round(AIC(model1_1),1), round(AIC(model2_5),1), round(AIC(model3_2),1))))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Mon, Dec 10, 2018 - 19:07:05

Table 8: 3 Final Versions of Model Specifications

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
density	0.008***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)
prbarr		−0.051***	−0.061***
		(0.012)	(0.012)
prbconv		−0.018***	−0.020***
		(0.004)	(0.004)
polpc		6.356***	7.334***
		(1.664)	(1.943)
west		−0.011***	−0.002
		(0.002)	(0.002)
pctmin80			0.0003***
			(0.0001)
taxpc	0.0003**	0.0001	0.0001
	(0.0001)	(0.0002)	(0.0002)
Constant	0.009*	0.039***	0.029***
	(0.004)	(0.008)	(0.007)
AIC	-530.6	-575	-593.5
R^2	0.582	0.766	0.814
Adjusted R^2	0.573	0.750	0.798
Residual Std. Error	0.012 (df = 87)	0.009 (df = 83)	0.008 (df = 82)
Note:			*p<0.05; **p<0.01; ***p<0.001

Our initial OLS regression model consisting of explanatory variables, density and tax revenue per capita, showed statistical significant estimated coefficients for both independent variables. We found that the effect of density on crime rate to be of greater practical significance than that of tax revenue per capita. In our analysis, we ruled out the inclusion of wage variables given their lack of joint significance in the regression model. Our base model resulted in AIC of -530 and an adjusted R^2 of 0.57, meaning that 57% of the variation in crime rate could be explained by the regression model of 2 explanatory variables. The correlation between tax revenue per capita and crime rate was positive, contrary to our original theory that wealthier counties with higher taxpc would have lower crime rate. However, we additionally noticed a positive relationship between the independent variables, taxpc and density, leading us to recommend that campaign policies consider prioritizing the increased tax revenue towards crime reduction in denser counties where both taxpc and crime rate are high.

The addition of criminal justice related covariates, prbarr, prbconv, and polpc, as well as the west location indicator variable was shown to increase the fit of our model. This final version of model 2 yielded an adjusted R^2 of 0.75 meaning 75% of the variation in crmrte is explained by the model after adjusting for the addition of more variables. At the same time, we also saw an improvement in fit, signified by a reduction in AIC from model 1 to this iteration of model 2, of -531 to -575. The results of this regression were unexpected in that we saw both statistical and practical significance for the coefficients for these covariates while the practical significance of our explanatory variables, density and taxpc, decreased. At the same time, the coefficient for taxpc lost statistical significance in the final model 2. However, in discovering the importance of prbarr, prbconv, and polpc, we shaped our policy recommendations to address potential gaps in these areas.

To demonstrate the robustness of our model, we included variables in the data set that we had previously not considered. However, with the addition of pctmin80, we saw a substantial improvement in fit delineated by a decrease in AIC, from -574.97 to -593.52. There was also an increase in the adjusted R^2 from 0.75 to 0.80. While we are unsure what about the addition of pctmin80 might be responsible for this increase in goodness-of-fit, its impact could not be ignored. Including additional covariates to this version of the model did not show such an effect on model fit. Therefore, this led us to produce a final version of model 3, which includes the covariate pctmin80. In this model, only the coefficients for taxpc and west lose statistical significance. We see slight decreases in the coefficients for prbarr and prbconv as well as a one unit increase in the coefficient for polpc. These criminal justice related variables remain both statistically and practically significant. Although the coefficient for poctmin80 is statistically significant, it is arguably not practically significant since it is an order of magnitude smaller than the other coefficients by comparison.

As we constructed 2nd and 3rd models of better fit, we discovered that the coefficient of the taxpc variable, which initially had statistical significance in the base model, lost significance in later model iterations. In addition, a common theme in our [Residuals vs Leverage] diagnostic plots for each model version show that the taxpc outlier for county 55 (point 25) exhibits great influence (with Cook's Distance > 1), which we suspected would be problematic for our estimates of β_2, the coefficient for taxpc, as this could introduce noticeable errors leading to greater variance eventual loss of significance. Furthermore, we discovered that 5 of the 6 omitted variables we identify may be driving β_2 towards 0 which would mean that the true value of β_2, and therefore the slope between crime rate and tax revenue per capita, is more positive than estimated by our OLS regression model. All things considered, we believe that it may still be possible that our original theory, which is that crime rate decreases as the overall wealth of counties increases, could still be true, but that taxpc may be a poor or insufficient proxy for county wealth. The loss of statistical significance of taxpc's estimated coefficient in a better fitting model shows that this variable may not be as strong of a determinant of crime as we had initially expected. Perhaps this result speaks to the limitations of the data, such as unknowns in tax revenue sources (income, property, sales, others) or tax revenue allocations that would be needed to help us more precisely measure county wealth.

Throughout our exploratory data analysis and model specifications, we also considered other factors that may affect the accuracy of our work. In particular, we considered that crime may be under-reported in rural areas as compared to urban areas, which means the crime rate given in the data set may not fully capture actual crimes per person that occurred. We also believe due to limitations of the data set, the lack of granularity in the data may have prevented us from building a more accurate model. Many of the data points were aggregated over an entire county, but we recognize that one county can consist of smaller different regions (city, suburbs, rural) where the variation in our independent variables, such as density, would be greater from region to region. However, that variation in independent variables is reduced when comparing measurements for different counties due to the aggregation or averaging of data over entire counties.

In order to reduce crime, based on our analysis, we recommend crafting policies that address proper resource allocations, especially for denser counties where the crime rate tends to be higher. We also found that denser counties tended to have higher tax per capita. Thus, given the statistical and practical significance of probability of arrest, probability of conviction, and police per capita, policies should focus on a redistribution of tax dollars to adequately staff the police force. If counties do not have the necessary funding, policies should aim to increase tax per capita in order to properly fund law enforcement efforts. Specifically in dense counties, we expect that an increase in police per capita may also lead to a proportional increase in the probability of arrest and probability of conviction, which may aid in combating the higher crime rate that we observed.

References

- DeSilver, D. (2015, September 22). The many ways to measure economic inequality. Retrieved from http://www.pewresearch.org/fact-tank/2015/09/22/the-many-ways-to-measure-economic-inequality/
- Krivo, L. J., & Peterson, R. D. (1996). Extremely disadvantaged neighborhoods and urban crime. Social forces, 75(2), 619-648.
- Lim, Hyeyoung & Lee, H & Cuvelier, Steven. (2010). The impact of police levels on crime rates: A systematic analysis of methods and statistics in existing studies. Asian Pacific Journal of Police and Criminal Justice. 8. 49-82.

- Sampson, R. J., & Wilson, W. J. (1995). Toward a theory of race, crime, and urban inequality. Race, crime, and justice: A reader, 1995, 37-56.
- Skogan, W. G. (1976). Efficiency and effectiveness in big-city police departments. Public administration review, 278-286.
- Naverson, D.B. (2018). https://www.businessreport.com/article/tax-foundation-louisiana-ranks-third-lowest- per-capita-income-taxes