MIDS-Lab3-Draft

Alexandra Iov, Haifeng Lin, Michael Reiter, Keith LoMurray

22 July 2018

Introduction

Reducing crime rates is a major goal of many political campaigns. The goal of this project is to use statistical analysis of a historical dataset on crime rates of counties in North Carolina to understand the determinants of crime and to generate policy suggestions for a political campaign. The assumption is the crime rate is driven by a variety of factors.

There are numerous hypotheses that could be proposed about what factors affect crime rates based on an individual's political views. For example, it could be proposed that the police per capita affect the crime rate. More police leads to less people committing crimes due to the prevalance of the police and drives down the crime rate. Another hypothesis could be that crime rate changes based on the probability of arrest, probability of convinction, and probability of prison sentence. This model could be described as a rational actors model. It is based on the assumption that people that commit crimes weigh the pros and cons before committing the crime. That includes factoring the likelihood of getting caught and being sentenced.

In order to avoid setting an alternative hypothesis after exploring the data, for this report we are working with some basic assumptions to drive our primary hypothesis tests. These include that regions with more people, *density*, regions with lower income, *taxpc*, regions with more young males, *pctymle*, and areas with a lower likelihood of being arrested, *prbarr*, have higher crime rates. This model could be viewed as a sociological explanation of crime, as it assumes crime is affected more by the makeup of the community rather than individual policing decisions.

Our first model will be developed with these variables to measure the impact on crime rates. We will follow up model one by adding additional variables to see if we can increase the fit of the model. The third model will contain all the variables provided in the dataset. This model will be used to compare the fit to model 1 and model 2.

Data Exploration and Cleaning

Our exploratory data analysis starts with a table of summary statistics, followed by histograms for select variables, as well as correlations by variable. The file contained six rows of NA data. The file also contained one duplicate row; county 193 was a complete duplicate on every variable. The NA rows and duplicates were excluded from the dataset, leaving a total of 90 observations for the analysis.

```
crime <- crime[!duplicated(crime),] #removes duplicated record
#Analyzing the datatypes, the probability of conviction is a factor and not numeric

crime <- na.omit(crime)
crime <- data.frame(crime)
nrow(crime)
```

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

```
length(crime) #gives the number of variables, not the number of records
```

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

Upon loading the raw data, prbconv was loaded as a character due to a comma in one of the NA rows in the source csv file. The following code transforms this variable to a numeric data type for proper analysis.

```
crime$prbconv <- as.numeric(paste(crime$prbconv)) #casts prbcon as a numeric variable
```

Probabilities (*prbarr, prbconv, prbpris*): The probability of conviction is a factor, not a numeric variable. Therfore a new column must be created to cast it as numeric.

We found that probability of conviction (1) and probability of arrest have values greater than 1. We know this is not possible as we cannot have more arrests than offenses or more convictions than arrests. There are 10 observations that have a probability of conviction of greater than 1. We do not want to remove these values as they represent more than 11 percent of our remaining observations. Therefore, we have top-codee these variable records to be equal to the maximum possible value, 1.

Since the probability of prison sentence is directly related to the probability of conviction, a variable we know is not perfectly accurate as it has been recorded, we need to be cautious when analyzing it.

```
length(which(crime$prbconv > 1))
```

```
## [1] 10
```

```
crime$prbconv_tc <- ifelse(crime$prbconv>=1, 1, crime$prbconv)
crime$prbarr_tc <- ifelse(crime$prbarr>=1, 1, crime$prbarr)
```

Average Sentence (*avgsen*): The average sentence in days ranges from 5.38 to 20.70, with a median of 9.11 and a mean of 9.69. This indicates that, on average, the types of crimes are relatively minor.

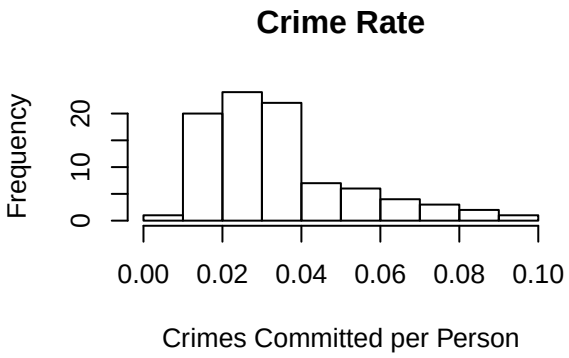
Density (*density*): The unit for this variable's measurement is unclear from the data descriptions, but comparisons to current population demographics indicate that the units are in *hundreds* of people per square mile. Under this assumption, the population density ranges from 0.002 to 883.77, with a median of 97.93 and a mean of 143.57, indicating a strong positive skew.

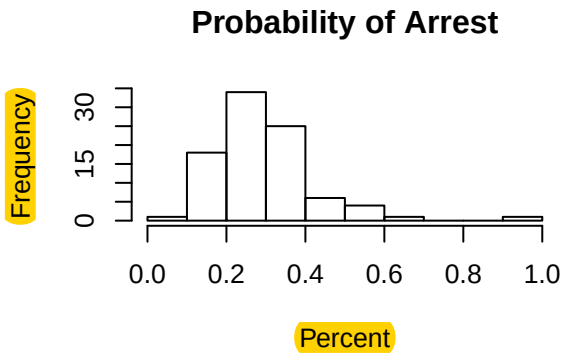
Percent Minority (*pctmin80*): The percentage of minorities across counties ranges from 1.28 to 64.35, with a a median of 24.85 and a mean of 25.71. The distribution shows a positive skew, however, the mean and the median remain relatively close. Overall, the distribution of this variable seems reasonable, and there are no obvious outliers.

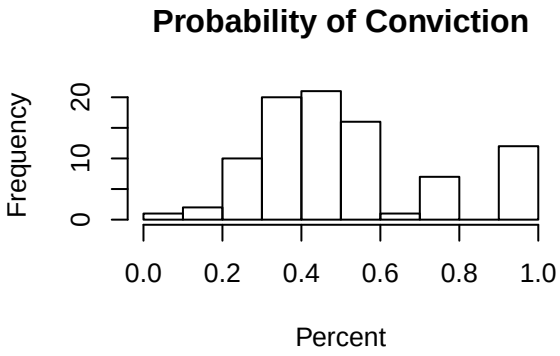
Percent Young Male (*pctymle*): The percentage of young males across counties ranges from 6.22 to 24.87, with a median of 7.77 and a mean of 8.40 indicating a positive skew. Of the total 90 observations, 89 fall within the range 6.22 to 15.09, with one extreme value at nearly 25 percent. While this is considered an outlier, we believe that it is possible due to the clustering of Universities in one area - such as Duke University and the University of North Carolina - thus driving up the youth population. Therefore, we will not remove or transform this outlier.

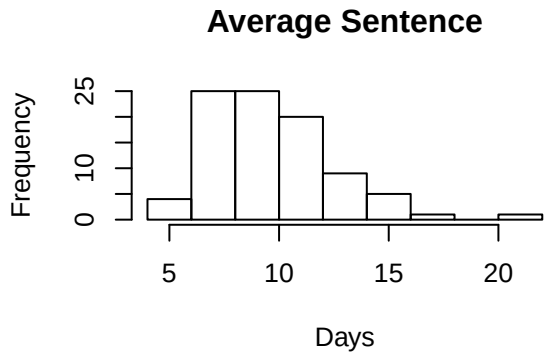
Distributions

```
attach(crime)
layout(matrix(c(1,2,3,4,5,6,7,8,9,10), 2, 2, byrow = TRUE))
hist(crmrte, main = "Crime Rate",
     xlab = "Crimes Committed per Person")
hist(prbarr_tc, main = "Probability of Arrest",
     xlab = "Percent")
hist(prbconv_tc, main = "Probability of Conviction",
     xlab = "Percent")
hist(avgsen, main = "Average Sentence",
     xlab = "Days")
```

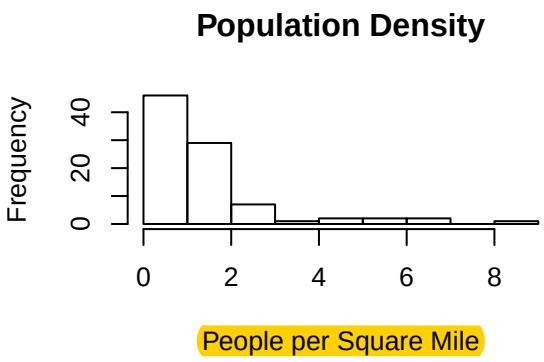


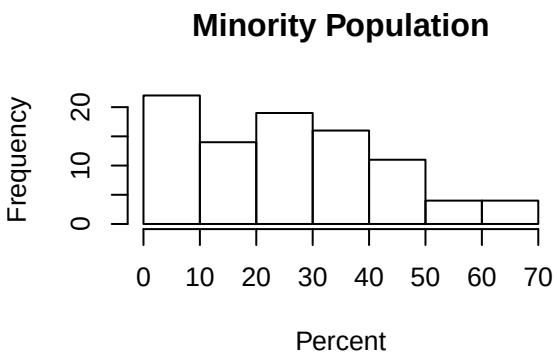


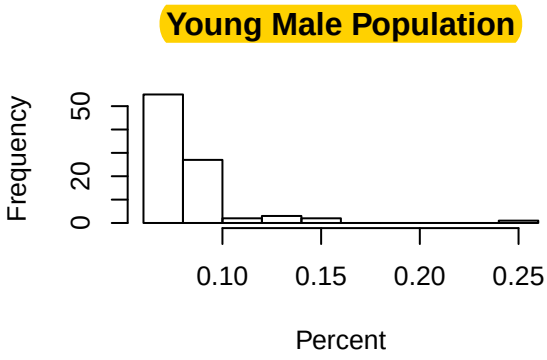


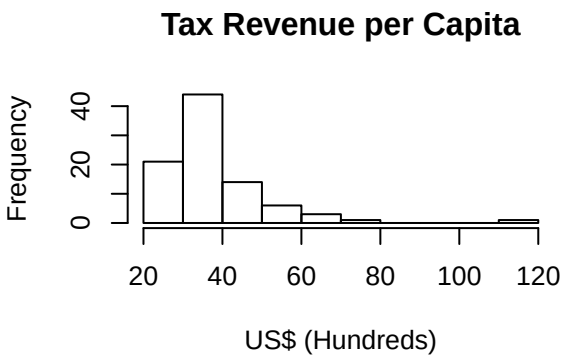



```
hist(density, main = "Population Density",
     xlab = "People per Square Mile")
hist(pctmin80, main = "Minority Population",
     xlab = "Percent")
hist(pctymle, main = "Young Male Population",
     xlab = "Percent")
hist(taxpc, main = "Tax Revenue per Capita",
     xlab = "US$ (Hundreds)")
```

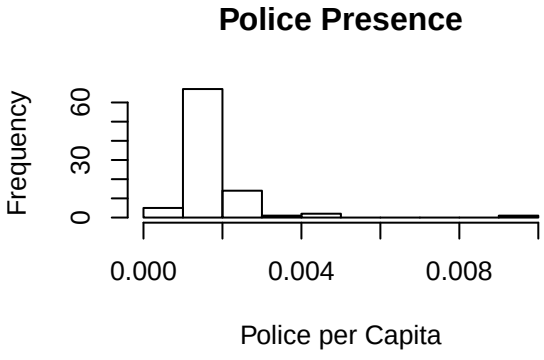


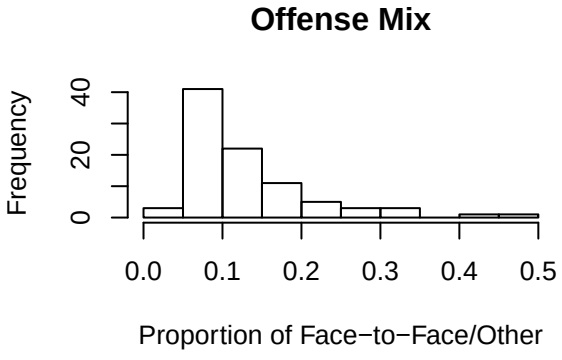







```
hist(polpc, main = "Police Presence",
     xlab = "Police per Capita")
hist(mix, main = "Offense Mix",
     xlab = "Proportion of Face-to-Face/Other")
detach(crime)
#ggplot(mix) + geom_histogram()
```





Key Variable Findings

Dependent Variable (Crime Rate)

Crimes committed per person (*crmrte*) is our dependent variable. Values range from 0.55 percent to 9.90 percent with a median of 3.00 percent and a mean of 3.34 percent. The values are positively skewed, but s[ow] a fairly normal distribution.

Geography

The geographic variables are binary variables including west, central, and urban. In viewing the summary statistics, 24.4 percent of the counties are categorized as west, 37.8 percent are categorized as central, 8.9 percent are categorized as urban, and 36.7 percent are not categorized at all. In comparing across regions, there is one county that is listed as both west and central, one county that is listed as both west and urban, and 5 counties that are listed as both central and urban. It is unclear from the data descriptions whether these geographic divisions should be mutually exclusive; however, we assume that overlap between urban and west/central is plausible and that the one record indicating overlap between west and central will not significantly affect the analysis. Therefore, these records are not altered for the analysis.

Income Levels

Tax Revenue (taxpc): The data on tax revenue ranges from 25.69 to 119.76, with a mean of 34.92 and a mean of 38.16, indicating a positive skew. Of the total of 90 observations, 89 fall within the range of 25.69 and 75.67, with one potential outlier at nearly 120. As we cannot determine a valid reason to remove this outlier, it will remain in the analysis. The units for this variable are unclear from the data descriptions, but based on secondary reasearch our conclusion is that the units are in *hundreds* of US dollars. This variable appears to be the most suitable proxy for income among the nine income-related variables and will serve as a key explanatory variable for model one.

Weekly Wages (*wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc*): These variables represent average weekly wage rates for common job types. However, since we do not know the number of people of each occupation in each county, these fields do not serve as a good proxy for income. If one county with a high crime rate has a high weekly wage for federal employees, for example, then we would be tempted to draw conclusions about the relationship between federal employee wages and crime. However, it could be the case that there is only one federal employee living in that county and that they are in a role with a high wage. We will use tax revenue as a proxy for wealth in each county.

Other

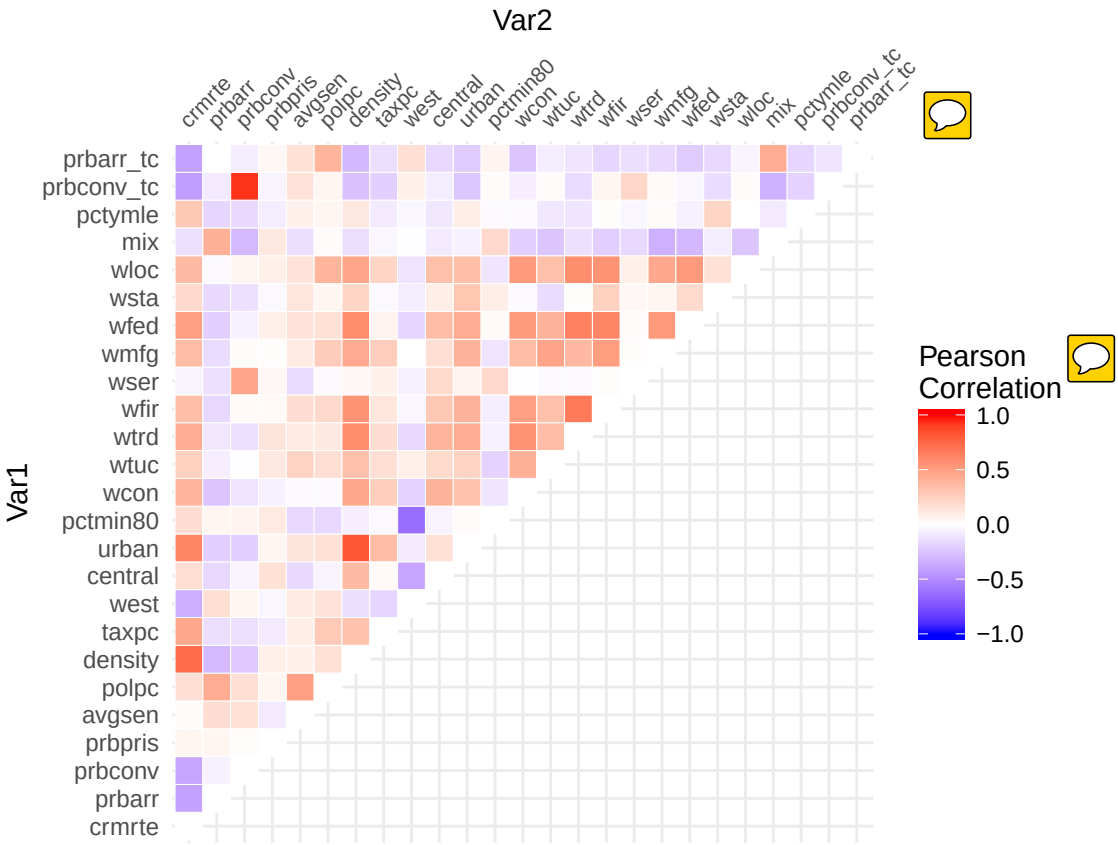
Police per Capita (*polpc*): Police per capita are consistently low, even with the 3rd quartile under 0.2%. One county has a Police per Capita of .9%. This can either be a busy county that requires many police officers or it is a small county with few residents and a relatively high number of police officers.

Offense Mix (*mix*): The unit on this variable is unclear, but is assumed to be the ratio or percent of face-to-face crimes with respect other crimes (vs. total crimes). Values range from 1.96 to 46.51, with a median of 10.10 and a mean of 12.91, indicating a positive skew.

Variable Correlations

```
crime2 <- subset( crime, select = -c(year, county) )

cor_crime <- round(cor(crime2),2)

get_lower_tri<-function(cor_crime){
    cor_crime[upper.tri(cor_crime)] <- NA
    return(cor_crime)
  }

# Get upper triangle of the correlation matrix
get_upper_tri <- function(cor_crime){
```

```
    cor_crime[lower.tri(cor_crime)]<- NA
    return(cor_crime)
  }


lower_tri <- get_lower_tri(cor_crime)
# Melt the correlation matrix
melted_cormat <- melt(lower_tri, na.rm = TRUE)

melted_cormat[melted_cormat$value == 1,]$value <- 0

# Create a ggheatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
  scale_x_discrete(position = "top") +
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 8, hjust = 0))+
 coord_fixed()
# Print the heatmap
print(ggheatmap)
```



Notes On Correlation

- Urban and density appear to be strongly colinear, which is logical.
- Crime rate has the strongest positive correlation with density & urban followed by wfed, taxpc, and wtrd.
- Crime rate the strongest negative correlation with prbarr, prbconv, and west.
- There appears to be colinearity with many of the wage variables. Mix and the wage variables appear to have an inverse relationship.

Model Development

For interpreaability of the models, we multipled the probability and percent variables by 100. Density was also multiplied by 100 to convert it from hunderd people per square mile to person per square mile and to ease intepretability.

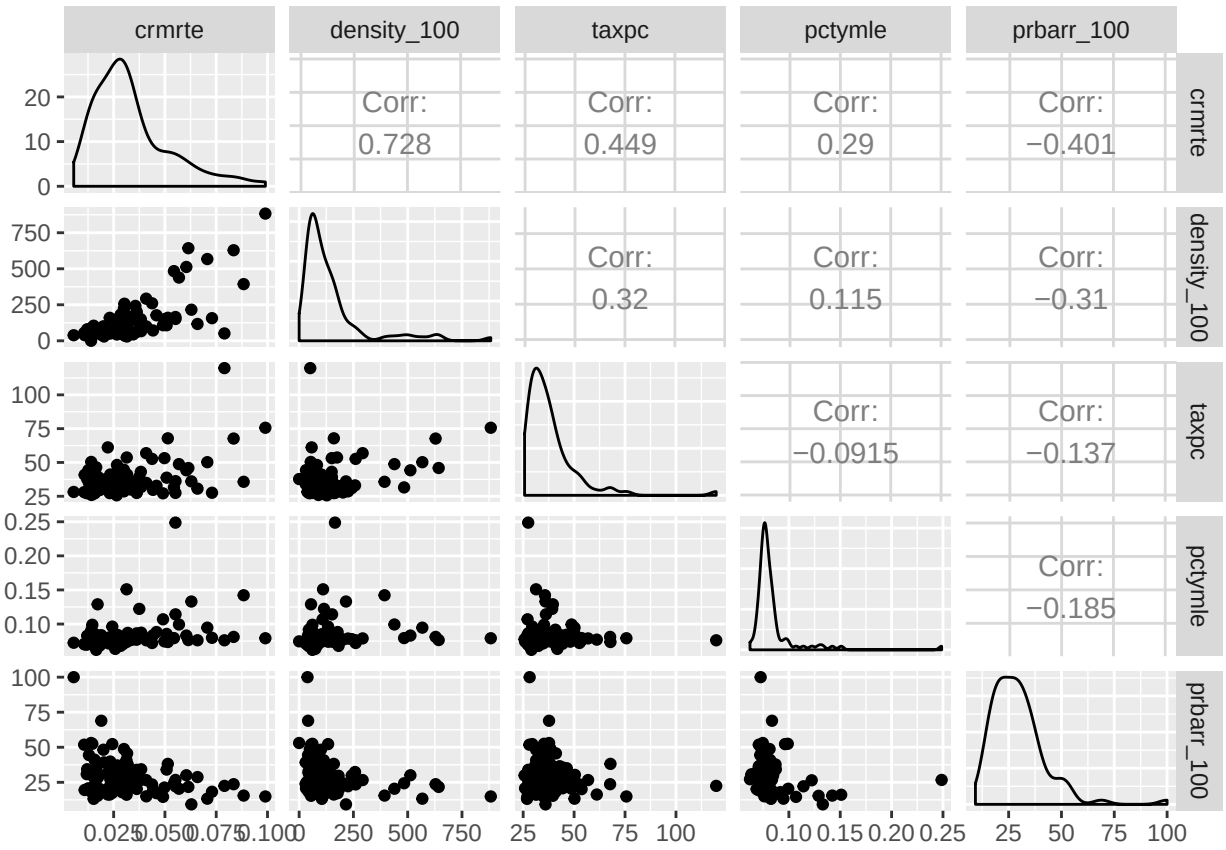
```
#Multiple probability by 100 for inteprebility
crime$density_100 <- crime$density*100
crime$pctymle_100 <- crime$pctymle*100
crime$prbarr_100 <- crime$prbarr_tc*100
crime$prbconv_100 <- crime$prbconv_tc*100
crime$prbpris_100 <- crime$prbpris*100
```

Model 1: Base Model

The base model was developed based on the sociological hypothesis proposed [illegible]ng the introduction. The exploratory data analysis including univaritate analysis plus correlation measures support this model as a strong candidate to explain the facors affecting crime rate.

Some of the variables in model 1 are acting a proxy values for the ideal variable the hypothesis would like to include. Taxpc is the tax revenue per capita. This variable is being used as a proxy for income since an income statistic is not available. There could be reasons beyond income that tax revenue varies. For example, urban areas typically have high tax liabilities than rural areas. In general though it would be expected that higher income areas have more tax revenue per capita.

```
prime_model_df <- subset(crime, select = c(crmrte, density_100, taxpc, pctymle, prbarr_100) )

ggpairs(prime_model_df)
```



The scatterplot matrix highlights some of the correlations seen in the correlation heatmap. For example, density appears to be strongly correlated with crime rate. The independent variables do not appear to be strongly correlated with each other. We would have expected stronger colinearity among the independent variables, if the model selected multiple variables measuring similar items. See model 4 for an example of this issue.

```
## Primary Model
model1 = lm(crmrte ~ density_100 + taxpc + pctymle_100 + prbarr_100, data = crime)
model1$coefficients
```

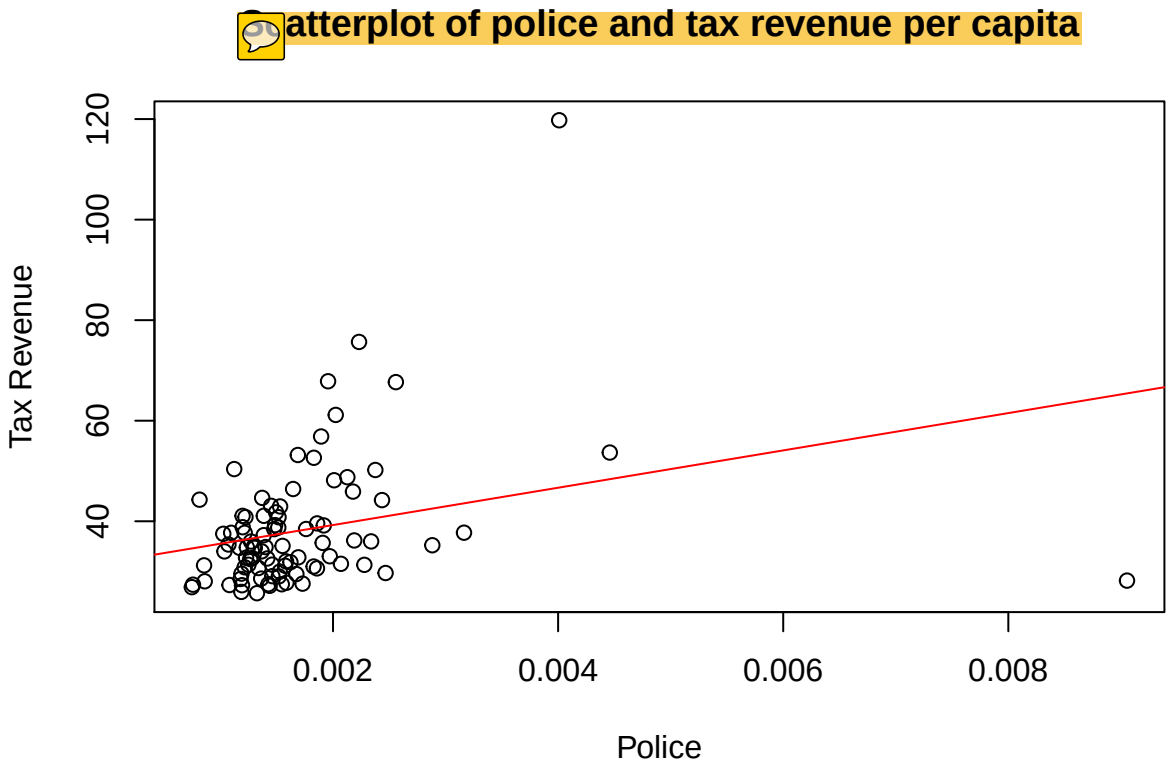
```
##   (Intercept)   density_100         taxpc   pctymle_100    prbarr_100
## -1.773028e-04  7.106879e-05  3.828146e-04  1.786089e-03 -2.084214e-04
```

Comparing the base model to our initial hypothesis we see some consistency and some deviations from the hypothesis. Holding other independent variables constant, increasing density, tax revenue, or percent of population that is young males increases the crime rate. Increasing the probability of arrest corresponds to a reduction in the crime rate. The one unexpected result from this model is the increase in crime rate with an increase in tax revenue. Our hypothesis leveraged tax revenue as a proxy for wealth and assumed that increases in wealth in a community corresponded to a reduction in crime. According to the model this is not the case. This is potentially explained by tax revenue not being a true proxy for the wealth of a community. Potentially tax revenue is a better proxy for urban vs. rural areas with urban areas having higher crime and higher taxes. According to the correlations there does seem to be a positive correlation between urban and tax revenue with a Pearson's correlation coefficient of 0.35. It could also be that counties with high crime rates increase their taxes in order to support crime reduction measures. Police per capita is also positively associated with crime rate according to the correlation heatmap. The scatterplot of police and tax revenue shows a slight positive trend between number of police and tax revenue per capita, which could support a hypothesis that counties with high crime increase taxes and the size of the police force to combat crime.

Additional research would need to monitor other attributes of counties with higher tax rates to see if the rates

are higher based on higher income or counties taking a greater share of income. In order to draw conclusions it would also be important to monitor trends to see if increasing taxes and/or police was effective at reducing crime compared to counties that took other approaches.

```
attach(crime)
plot(polpc, taxpc, main='Scatterplot of police and tax revenue per capita', xlab = 'Police', ylab='Tax
abline(lm(taxpc ~ polpc), col='red')
```




```
detach(crime)
```

Model 2: Enhanced Model

Model 2 took the indepdent variables from model 1 and added additional variables that according to exploratory data analysis had potential to be good predictors of crime rate. The additional variables include police per capita, percent minority. As mentioned previously the police per capita does have a correlation with crime rate. It would also make sense that the number of police directly affects certain aspects of crime such as the probability of arrest. Percent minority is also associated with the crime rate. One should be careful to draw to many conclusions from percent minority. As will be discussed in the section on omitted variables, it is possible with being a minority is associated with other attributes correlated to crime. Percent minority will be included in the enhanced model to represent some of the omitted factors.

```
model2 = lm(crmrte ~ density_100 + taxpc+ pctymle_100 + prbarr_100 + polpc + pctmin80 , data = crime)
model2$coefficients
```

```
##   (Intercept)   density_100         taxpc   pctymle_100    prbarr_100
## -5.140387e-03  6.947972e-05  3.238819e-04  1.627790e-03 -3.290154e-04
##         polpc      pctmin80
##  2.666199e+00  3.019770e-04
```

```
paste("AIC Model2  = ", AIC(model2))
```

```
## [1] "AIC Model2  =  -562.630038903181"
```

```
paste("AR2 Model2  = ", summary(model2)$adj.r.squared)
```

```
## [1] "AR2 Model2  =  0.71279709912108"
```

A general comparison of model performance will be discussed in the section below. However, in model 2 we see similar trends to model 1 including the coefficients of the variables included in model 1 continuing in the same direction. The additional variables also trended in the direction predicted. As police per capita increased, we see an increase in crime rate. The same trend is seen with the percent minority variable. Although, as mentioned, one should be careful when drawing conclusions based on the direction of the coefficients only.

Model 3: All Variable Model

The third model includes all variables that make sense to include. We excluded the variables year and county number, but included all other variables. Model 3 will allow us to compare model performance of the other models.

```
crime1 = crime[ , !(names(crime) %in% c("pctymle", "density", "prbarr", "prbconv", "prbpris", "crmrte_l
```

```
model_all <- lm(crmrte ~ . , data = crime1[,-c(1:2)])
model_all$coefficients
```

```
##   (Intercept)        avgsen         polpc         taxpc          west
##  1.368594e-02 -4.538430e-04  5.386952e+00  1.907781e-04 -1.546193e-03
##       central         urban      pctmin80          wcon          wtuc
## -3.728301e-03  1.791764e-04  3.352009e-04  2.611542e-05  7.746638e-06
##          wtrd          wfir          wser          wmfg          wfed
##  1.868225e-05 -3.586569e-05 -1.136264e-05 -8.507163e-06  3.739193e-05
##          wsta          wloc           mix   density_100   pctymle_100
## -1.911118e-05  1.970759e-05 -1.686701e-02  5.349190e-05  1.061689e-03
##    prbarr_100   prbconv_100   prbpris_100
## -5.025758e-04 -2.260597e-04 -4.451190e-06
```

Model 4: Alternative Model

We chose to highlight one additional model. As mentioned in the introduction, one could hypothesize a rational actor model to crime. In this model, as the probability of arrest, convection, and going to prison increases the crime rate would be expected to go down. To highlight the model performance against the sociological model, we will run the alternative model and compare performance to the other models.

```
#Model focused on crime stats including prob arrest, conviction, prison, and the police per capita
model3 = lm(crmrte ~ prbarr_100 + prbconv_100 + prbpris_100 + polpc, data = crime)
model3$coefficients
```

```
##   (Intercept)    prbarr_100   prbconv_100   prbpris_100         polpc
##  6.330221e-02 -8.910049e-04 -3.971159e-04  6.350306e-05  8.190360e+00
```

CLM Assumptions

In this section, we evaluate the CLM assumptions for all four models.

Linearity in Parameters

From the Model Comparison section, we can see that all those models have linear coefficients and that they are correctly specified.

Random Sampling

Currently, North Carolina is comprised of 100 counties, while the dataset used in this analysis contains 90 county observations. While these 90 observations should provide sufficient representation of the state, we have no knowledge as to why 10 counties are missing from the dataset. If there is inherent bias behind why the counties were excluded, this assumption could be affected.

No Perfect Collinearity

From the heatmap under Variable Correlations section, we do not see any two independent variables that demonstrate a perfect linear relationship. We do notice some independent variables with correlation, such as density and urban, however, their correlations do not violate this assumption since they are not perfectly linear.

Zero Conditional Mean

As discussed in the omitted variables section, we do find omitted variable bias (OVB) in all four models as the dataset has left out some important causal factors. In other words, we expect that there is correlation between the error term (that captures all external factors and omitted variables) and our independent variables, i.e. the mean of the error term conditioned on each independent variable is zero. More details about omitted variables can be found in the following discussion.

Regression Table

```
#Report
stargazer(model1, model2, model_all, model3, type='text',
          report='vc',
          title='Linear models predicting crime rate',
          keep.stat=c("n", "adj.rsq"),
          omit.table.layout = "n",
          add.lines=list(c("AIC", round(AIC(model1),1), round(AIC(model2),1), round(AIC(model3),1), roun
```

```
##
## Linear models predicting crime rate
## ================================================
##                       Dependent variable:
##               --------------------------------
##                             crmrte
##                 (1)      (2)      (3)      (4)
## ------------------------------------------------
## avgsen                            -0.0005
##
## density_100   0.0001   0.0001    0.0001
##
## taxpc         0.0004   0.0003    0.0002
```

```
##
## west                          -0.002
##
## central                       -0.004
##
## urban                         0.0002
##
## pctymle_100    0.002   0.002   0.001
##
## prbarr_100   -0.0002 -0.0003  -0.001  -0.001
##
## prbconv_100                  -0.0002  -0.0004
##
## prbpris_100                  -0.00000 0.0001
##
## polpc                 2.666   5.387    8.190
##
## pctmin80             0.0003   0.0003
##
## wcon                         0.00003
##
## wtuc                         0.00001
##
## wtrd                         0.00002
##
## wfir                         -0.00004
##
## wser                         -0.00001
##
## wmfg                         -0.00001
##
## wfed                         0.00004
##
## wsta                         -0.00002
##
## wloc                         0.00002
##
## mix                           -0.017
##
## Constant     -0.0002 -0.005   0.014    0.063
##
## ---------------------------------------------
## AIC          -544.9  -562.6   -516.3   -580
## Observations   90      90       90       90
## Adjusted R2   0.643   0.713   0.794    0.510
## =============================================
```

A comparison of models highlights some interesting findings. The alternative rational actor model had the lowest Akaike's Information Criterion (AIC) score, meaning controlling for model complexity it performed the best. The model with all variables included performed the worst on AIC, which makes sense given AIC penalizes model complexity.

However, comparing adjusted R squared, which accounts for much of the variation a model explains, the model with all variables performed the best. The enhanced model performed better on adjusted R square and AIC compared to the base model. The alternative model performed the worst on R squared.

Given the overall information about the models, please the lack of intepretability of the all variable model, it would be recommended to focus on the enhanced model to draw conclusions.

Interpretation of Coefficients

The regression coefficients for each variable in our model (detailed in the regression table above) tell us the degree of change in the dependent variable, crime rate, for each unit change in the independent variable, assuming the other variables in the model are fixed.

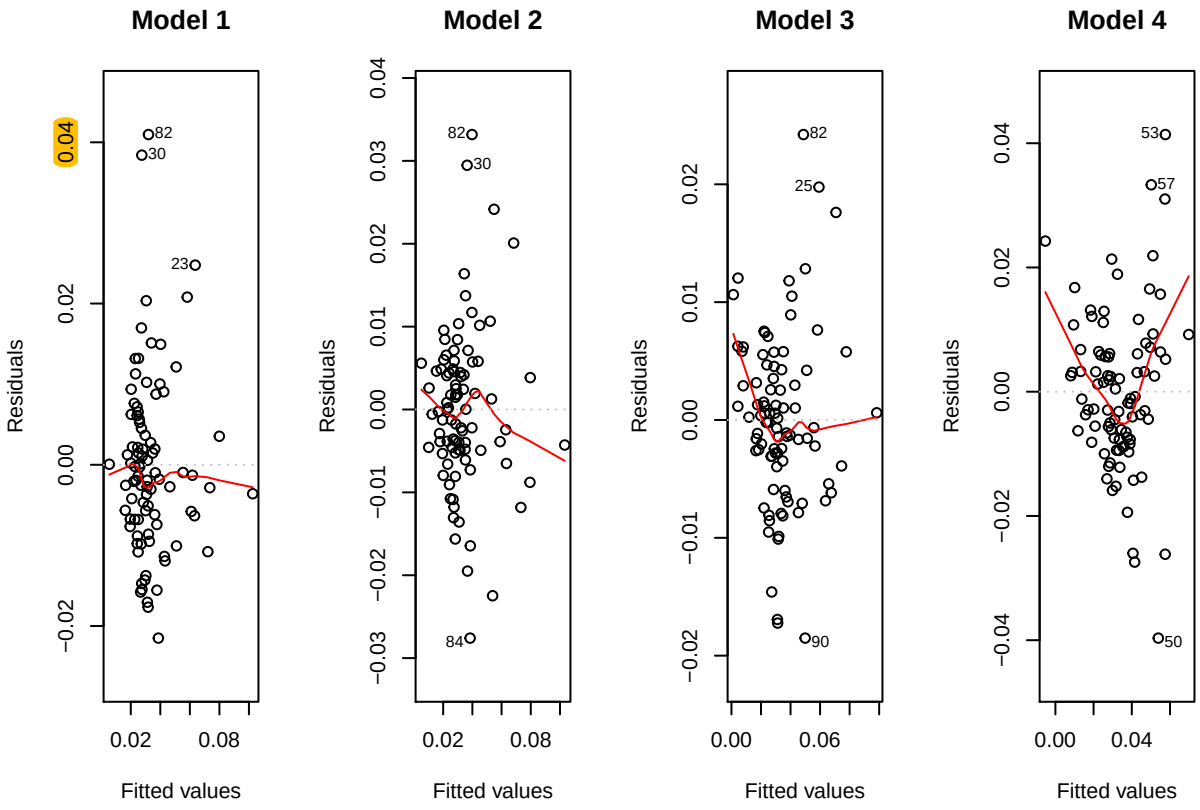
Therefore, for model 1 the coefficients can be described as:

- density_100: each percent change in population density is associated with a change of 0.0001 in crime rate, assuming all other variables are held fixed.
- taxpc: each $100 change in the tax rate per capita is associated with a 0.0004 change in crime rate, assuming all other variables are held fixed.
- pctymle_100: each percent change in the proportion of youth to total population is associated with a change of 0.002 in crime rate, assuming all other variables are held fixed.
- prbarr_100: each percent change in the probability of arrest is associated with a change of -0.0002 in crime rate, assuming all other variables are held fixed.

Models 2, 3, and 4 follow the same logic.

Model Comparison

```
old.par <- par(mfrow=c(1, 4))
plot(model1, which=1, caption="", main = "Model 1")
plot(model2, which=1, caption="", main = "Model 2")
plot(model_all, which=1, caption="", main = "Model 3")
plot(model3, which=1, caption="", main = "Model 4")
```




```
par(old.par)
```

Comparing the fitted values to residuals, models 1 and 2 have a moderate flat trend, while model 3 (including all variables) has a strong peak on the left. The alternative model has significant issues, but by taking the log of the crime rate, the residuals vs fitted flattens somewhat. The AIC is no longer comparable to the other models, but the adjusted R squared remains the same.

```
crime$crmrte_log <- log(crime$crmrte)
model3_log = lm(crmrte_log ~ prbarr_100 + prbconv_100 + prbpris_100 + polpc, data = crime)

model3_log$coefficients
```

```
##   (Intercept)    prbarr_100   prbconv_100   prbpris_100         polpc
## -2.436571e+00 -2.641090e-02 -1.206252e-02  4.902271e-04  1.548945e+02
```

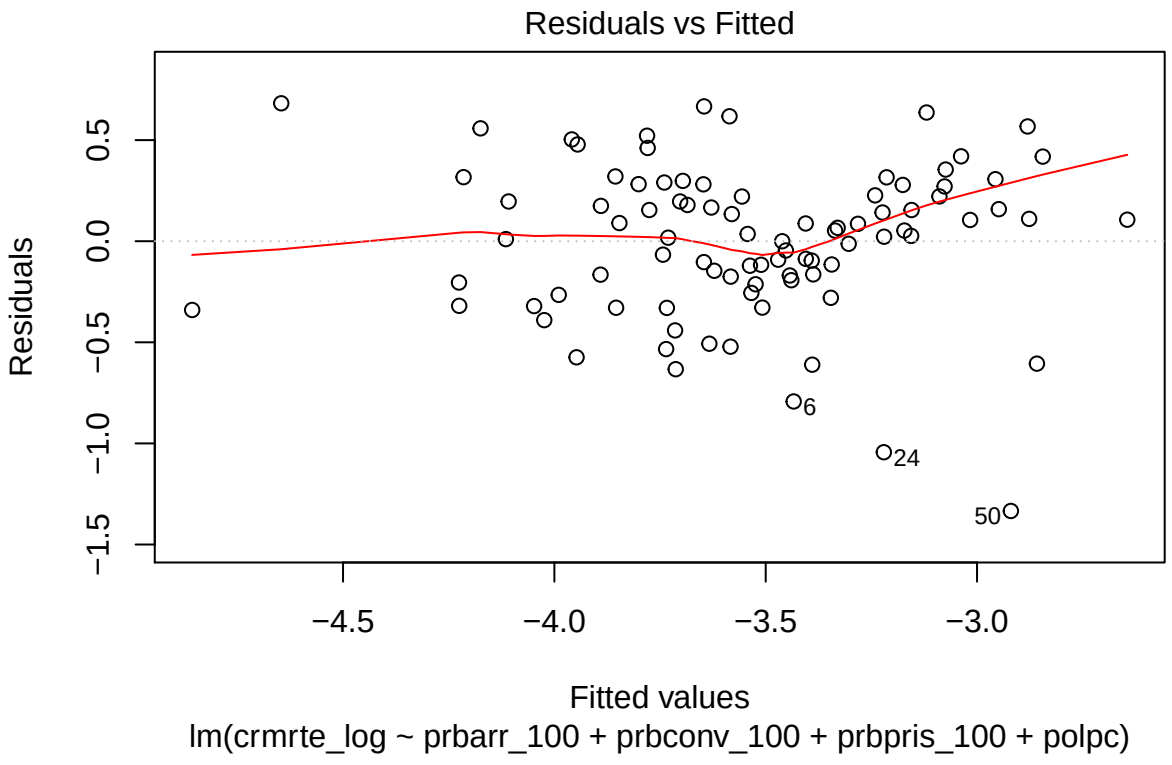
```
paste("AIC Model2  = ", AIC(model3_log))
```

```
## [1] "AIC Model2  =  89.6093081686469"
```

```
paste("AR2 Model2  = ", summary(model3_log)$adj.r.squared)
```

```
## [1] "AR2 Model2  =  0.51234509427935"
```

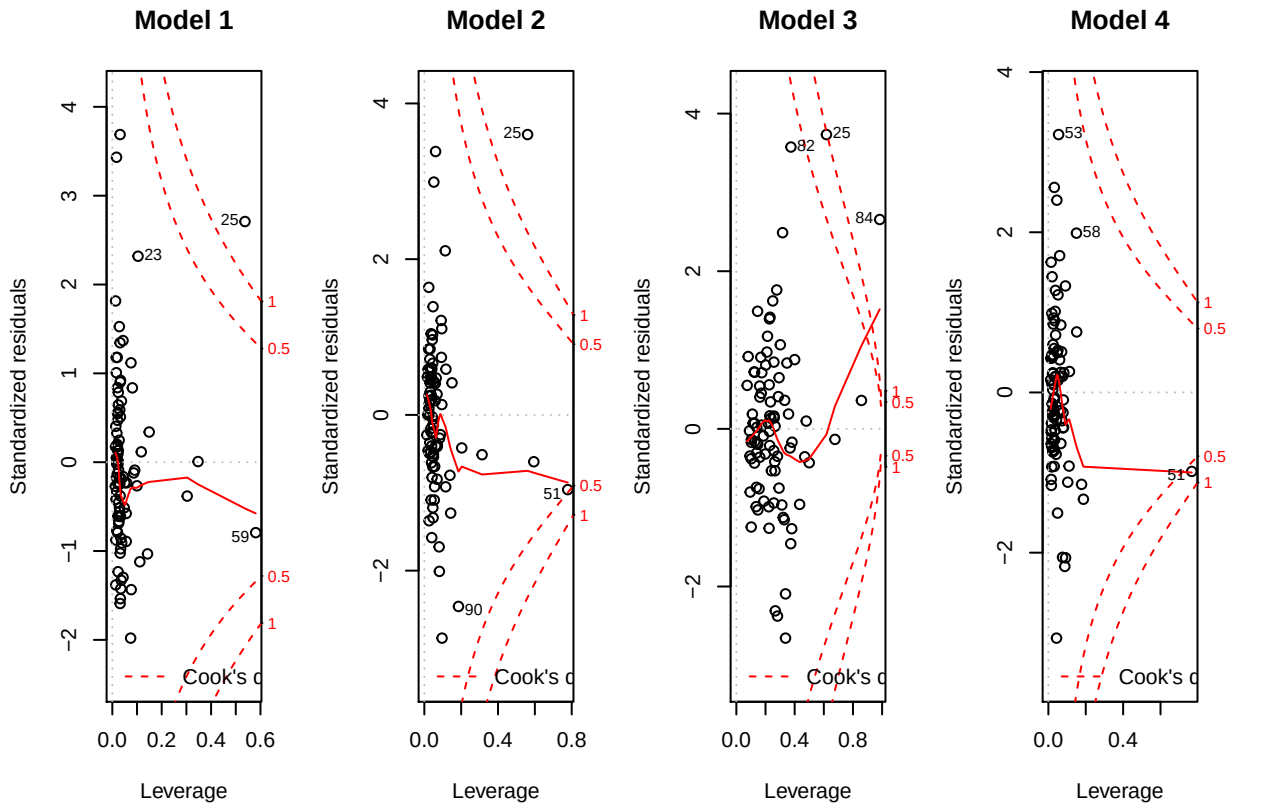
```
plot(model3_log, which=1)
```




```
old.par <- par(mfrow=c(1, 4))
plot(model1, which=5, caption="", main = "Model 1")
plot(model2, which=5, caption="", main = "Model 2")
plot(model_all, which=5, caption="", main = "Model 3")
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
plot(model3, which=5, caption="", main = "Model 4")
```




```
par(old.par)
```

It is also important to check the residuals vs. leverage of each model. There is one strong influence point seen in multiple models. Observation 25 has a tax revenue of 119.76. this is a significant outlier and has high leverage on the model. There is no evidence for excluding this point though since it doesn't appear to be inaccurate. Model 3 also has a leverage point with observation 84.

Omitted Variables

β_2 represents the coefficient between the dependent variable (crime rate) and the omitted variable, α_1 represents the coefficient between the predictor and the omitted variable, and β_1 represents the coefficient between the dependent variable (crime rate) and the predictor in discussion.

Drugs and Alcohol abuse:

Alcohol assumption directly increases the risk of criminal violence. Firstly, alcohol has influence in increase of aggression. For drug abuse, crime could arise from competition among drug suppliers and from drug addicts who need to fund their addition.

Drug and alcohol abuse has a positive correlation with Percent of Young Male because in general young males are more likely to have such problem.

$\beta_2 > 0$ and $\alpha_1 > 0$, so OVB > 0. Since β_1 for Percent of Young Male is > 0, the OLS coefficient on Percent of Young Male will be scaled away from zero (more positive) gaining statistical significance.

Income levels:

Crime rate in low-income areas is usually higher because people are more motivated to offend as a means of overcoming their disadvantage. Overall, the correlation between income level and crime rate is negative,

meaning higher income level leads to lower crime rate. We use Tax Revenue per Capita as the proxy for income level in our model and assume that areas with higher income level have higher tax revenue per capita.

Income level has positive correlation with Tax Revenue per Capita in the dataset.

$\beta_2 < 0$ and $\alpha_1 > 0$, so OMVB < 0. Since β_1 for Tax Revenue per Capita is > 0, the OLS coefficient on Tax Revenue per Capita will be scaled toward zero (less positive) losing statistical significance.

Unemployment rate:

In general, unemployment increases the risk of individuals becoming involved in crime. Unemployment rate has positive correlation with Crime Rate and negative correlation with Tax Revenue per Capita in the dataset.

$\beta_2 > 0$ and $\alpha_1 < 0$, so OMVB < 0. Since β_1 for Tax Revenue per Capita is > 0, the OLS coefficient on Tax Revenue per Capita will be scaled toward zero (less positive) losing statistical significance.

Criminal Opportunity

Offenders commit more crime when there are more opportunities and incentives for committing it. Some factors can create opportunities or incentives for crime. These include lax physical security, lax personal security, attractive commercial or residential targets and easy opportunities for selling or disposing of stolen goods. Criminal opportunity has positive correlation with crime rate. No proxy variable available in the model.

Recidivism

If areas have low recidivism rate, they could have low crime rate as well. Therefore, the recidivism rate has positive correlation with crime rate. No proxy variable available in the model.

Degree of Inadequate Parenting

Factors associated with or indicative of inadequate parenting could significantly increase the risk of juvenile involvement in crime. Most persistent adult offenders generally start offending as adolescents. We assume that higher degree of inadequate parenting, the higher crime rate could be. No proxy variable available in the model.

Conclusion

Reducing crime rates is a major goal of many political campaigns. The goal of this project is to use statistical analysis of a historical dataset on crime rates in North Carolina counties to understand the determinants of crime and to generate policy suggestions for a political campaign.

Before presenting our findings, we would like to preface that, given the limitations of this dataset, any politican should be wary of drawing policy prescriptions to address crime from the data. This data is over 30 years old, represents only a single year of data, and is also missing critical information that could have an influence on crime rates.

As previously mentioned, there are many variables that could have enriched the analysis such as drug and alcohol abuse, education, income levels, and unemployment rate. Criminal punishment and prosecutions could also have important racial overtones, which would not be captured in this data. Our models would also benefit from knowing more about the type of crimes that are occuring. The face-to-face and average prison sentence measures do not tell us enough about the severity of the crimes. Are these organized crimes, corporate crimes, violent crimes, or other crimes?

We purposefully excluded the minority variable as we believe this can be because of other related factors. Over-policing in areas with higher minority populations often leads to higher arrest rates in those communities. People of color are more often arrested (increasing the number of reported "crimes") on small drug charges than white people. We do not want to make any conclusions on the relationship between minority percentage and crime rates because we believe that these could easily be misinterpreted or misused. We also want to

point out that race is easy to mislabel and is not always uniformly reported. Many people fit into multiple races and we have no understanding of how that was handled in the collection of these observations.

The variable we decided to include in our model was taxpc, or Tax Revenue per Capita. We assumed this would be the best proxy for income and we assumed that as taxpc increased, crime rate would decrease. The first assumption, that taxpc was a good proxy for income, can be incorrect because in general, populated cities have higher tax rates and tax revenue does not reflect income but rather the rates themselves. Our model found that our second assumption was incorrect as well. We have multiple theories as to why an increase of tax revenue per capita would lead to an increase in crime rate but we wanted to mention one: we believe that the increased tax revenue could be associated the number of police in a county. As police increase, the number of crimes reported also increase. Crime rate is not a true measure of how many crimes actually occur but rather how many crimes are reported.

With all of this being said, our ultimate goal is to make policy recommendations. Understanding our data limitations is key to understanding next steps. If our rational actor model had yielded impressive results, we would recommend to increase the number of arrests, convictions, and prison sentences. However, this model was a poor predictor of crime rates, so we do not recommend increasing these numbers. While we cannot difinitively say what causes crime, we can recommend that policy focuses on alleviating crime in high density areas with high young male and minority communities, because those are the areas most affected by crime. Further research will need to be done to say what needs to be done in those areas. We recommend to study the omitted variables we highlighted earlier, education rates, and types of crimes closer.