From Probability to Policy - Reducing Crime in North Carolina

Aditi Das, Jake Miller, John Pette, Krysten Thompson

November 30, 2018

1. Introduction and Research Question

A North Carolina gubernatorial candidate's campaign team hired our research firm to identify the determinants of crime and make policy recommendations that would aid the campaign. The candidate's team provided us with a set of crime data from 1987. This introduced an immediate challenge, as the dataset was 30 years old and may no longer be representative of North Carolina crime. However, factors influencing crime should be relatively consistent and we were still able to build a useful model from this dataset. We started with an exploratory data analysis so that we could build models that provided insights into factors related to crime.

We selected crime rate as our dependent variable, and, following our initial examination of the data, chose probability of arrest, population density, and high youth concentration as our initial set of factors demonstrating relationships with crime rates. This led to our research question. We sought to:

Examine the relationship between probability of arrest, population density, and high youth concentration and the dependent variable, crime rate, in order to determine whether related policy changes could be effective in lowering crime rates.

Our approach was to begin with these variables and refine the initial model through additional linear regressions. This would help us identify any variables in the existing data that explained the variation in the model and allow us to pinpoint how much of that explanation is attributed to omitted variables.

Our preliminary assessment of factors that could affect crime was based on our existing knowledge coupled with FBI reports. Many factors we identified as contributing to crime rates in general did not exist in the dataset and we did not find suitable proxy variables. We proceeded with our analysis expecting to see omitted variable effects. However, with further variable analysis and model assessment, we felt we had enough data in place to provide directional guidance and formulate policy recommendations.

2. Exploratory Data Analysis

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

Back Up Dataset

Prior to any data transformations, we made a back up copy of the raw data.

```
crime_raw <- crime
```

Check for Data Completeness

An initial scan of the data set revealed that it contained 6 rows entirely composed of NA data. These provided no information, so we opted to delete them.

```
crime = crime[!is.na(crime$county),]
```

Check for Duplicate Records

We examined the numerical county indicators in order to determine whether it contained any duplicate records.

```
length(unique(crime$county))
nrow(crime)
paste("The number of unique counties is",length(unique(crime$county)) ,
      "while the number of records in the dataset is", nrow(crime))
```

```
## [1] 90
## [1] 91
## [1] "The number of unique counties is 90 while the number of records in the dataset is 91"
```

We found a single duplicate county record. All of the data was consistent between the two records sharing a county ID. We determined that this meant there was a true duplicate and that one of the records could be removed.

```
crime = crime[!duplicated(crime),]
paste("Now there are total of ",nrow(crime)," rows and ",ncol(crime)," columns.")
```

```
## [1] "Now there are total of  90  rows and  25  columns."
```

Note: There are 100 counties in North Carolina, and we had data for only 90 of them. Missing counties could contain data that would skew the analysis, e.g., if a missing county was highly populated with a high crime rate, or highly populated with a low crime rate. We concluded that analyzing the data on a regional level made the most sense.

We did note that the county index numbers appeared to correspond to the Federal Information Processing Standard (FIPS) codes for counties, which allowed us to identify which counties were missing from the data.[1]

Assign Each County to a Region and Check for Data Consistency

North Carolina consists of three main geographic regions: the Atlantic coastal plain, occupying the eastern portion of the state, the central Piedmont region, and the Mountain region in the west. Our dataset only contained regional indicator variables for the West and Central regions. We made the assumption that counties with no assigned region could be considered East, and cross-reference the data with an external reference to validate that assumption.[2] We assigned 35 counties with no prior regional designation to the east region.

```
#Ideally, there should be 1 either in central or west...we should check for bad data.
overlapCounty = sum(crime$central==1 & crime$west==1)
totCentral = sum(crime$central==1 & crime$west==0)
totWest = sum(crime$west==1 & crime$central==0)
crime$east = ifelse((crime$central==0 & crime$west==0), 1,0)
```

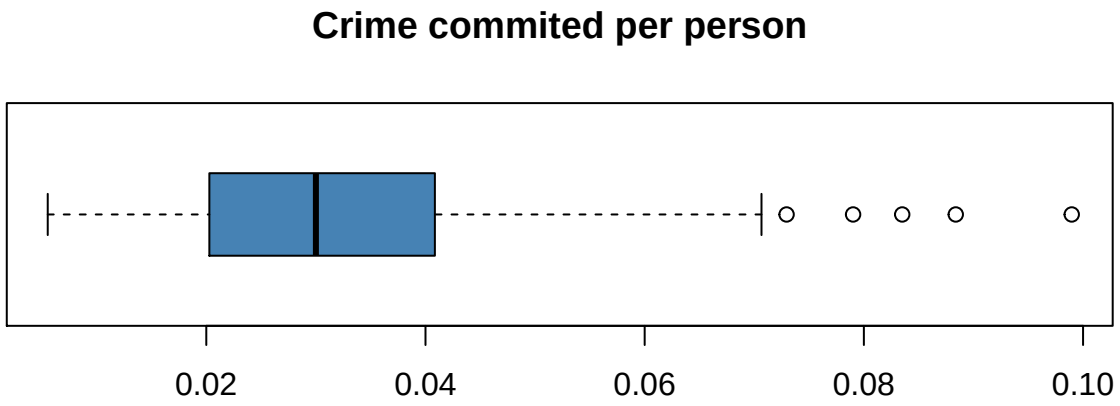
[1] https://www.ncpedia.org/geography/counties
[2] https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina

Note: We noticed that one county was marked as both central and west. Per ncpedia.org, Madison County (FIPS 115) is part of the Central Piedmont region. Hence, we assigned it to the central region and removed the west indicator.

```
a= crime %>% filter(central==1 & west==1) %>% mutate(west=0)
```

Crime Rate Analysis

```
crmrtePlot = boxplot(crime$crmrte, horizontal=TRUE,
                     main = "Crime commited per person" , col=c("steelblue"))
```




```
crmrteOtlr = crime[crime$crmrte >= crmrtePlot$stats[5],]
```

Note: There were six outliers of high crime rate regions. Of those, four were designated as urban areas. We decided not to modify these data points and instead took a closer look at how crime was spread across various regions of North Carolina.

Crime Rate by Region

Each region has several characteristics distinct from the others:

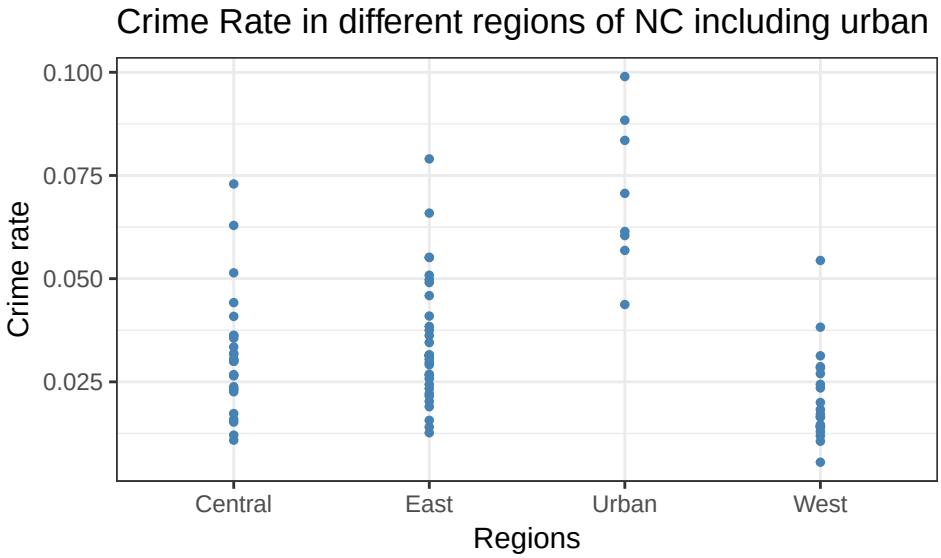
- **Western North Carolina** has one urban county and is mostly characterized by the rugged, mountainous terrain of the Blue Ridge Mountains. Sparsely populated, it encompasses many national parks, national forests, and state parks. Tourism as one of its primary industries: the Blue Ridge Parkway and Great Smoky Mountains National Park were Nos. 1 and 3 on the National Parks Services' most visited destinations in 2017.[3] NC is also the nation's second-leading producer of Christmas trees.[4] Western NC is the most racially homogeneous of the regions: The 15 counties with the lowest `pctmin80` are in the western region; 20 of the region's 22 counties (in this dataset) were in the bottom 30 for `pctmin80`.
- **Central North Carolina ("the Piedmont")** includes its largest metro areas. Charlotte (Mecklenburg County) is one of the nation's leading banking centers. The Piedmont Triad (Greensboro, High Point, and Winston-Salem; Guilford and Forsyth Counties) were hubs of furniture, textile, and tobacco industries. The Research Triangle encompasses Raleigh, Durham, and Chapel Hill and is home to Research Triangle Park, UNC-Chapel Hill, Duke University, and N.C. State University. Outside of the metro areas, the Piedmont includes mill towns centered around furniture and textiles (although these have declined considerably since 1987), as well as poultry, cattle, and other farming.
- **Eastern North Carolina ("Coastal Plain")** is historically the poorest and most agricultural of the three regions. NC is one of the nation's top pork[5] and tobacco producers, with both industries

[3] https://www.nps.gov/orgs/1207/02-28-2018-visitation-certified.htm
[4] https://www.agmrc.org/commodities-products/forestry/christmas-tree-profile
[5] https://www.pork.org/facts/stats/structure-and-productivity/state-rankings-by-hogs-and-pigs-inventory/

concentrated in the Coastal Plain. Other notable crops include soybeans, peanuts, cotton, sweet potatoes, and corn.[6] Tourism, centered around the Outer Banks and other beaches, is another notable industry. Eastern North Carolina's two metro areas include Wilmington (New Hanover County), a deep-water port near the coast, and Fayetteville (Buncombe), which is adjacent to Fort Bragg, the nation's largest army base. Onslow County on the coast includes Camp Lejeune, a large U.S. Marine Corps installation.

We performed an exploratory analysis of crime rates that included regional variables to be aware of and control for any potential regional differences. Each region is distinct enough in industry and demography that we explored certain key metrics by region. Beyond the regional designations, the dataset also included an "urban" variable. We determined that this was not a region unto itself. Rather, it was a separate category. Urban areas could appear in any of the geographic regions. However, as urban areas typically exhibit unique crime characteristics, we thought it would be of value to examine these areas as a group alongside the regions.

```
#Create regions
crime$regWurban = ifelse(crime$urban==1,"Urban",ifelse(crime$west==1,"West",
          ifelse(crime$east==1,"East",ifelse(crime$central==1,"Central","Other"))))

ggplot(crime, aes(x=regWurban, y=crmrte), ) +
  geom_point(size=1,na.rm = TRUE,col="steelblue") +
  labs(title = "Crime Rate in different regions of NC including urban",
       x = "Regions", y = "Crime rate")+
  theme_bw()
```



Note: As we expected, urban areas had substantially higher crime rates than any of the geographic regions. Traditional logic supports the higher instance of crime in urban areas. Among regions, the counties in the west appeared to be the safest in terms of crime.

Crime Rate and Offense Mix

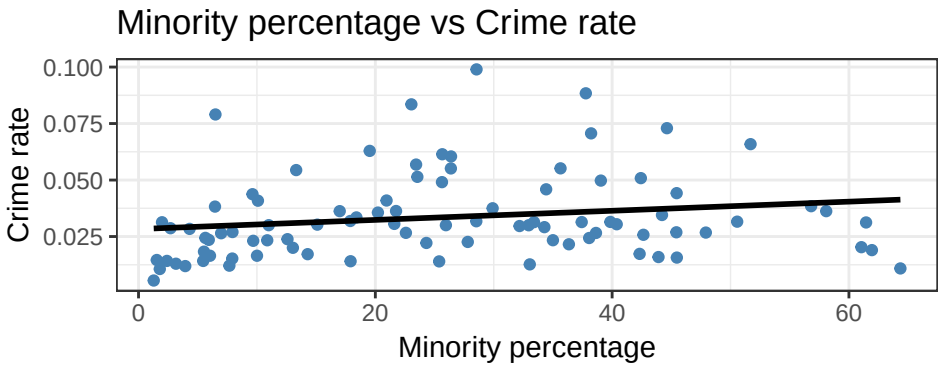
The variable `mix` describes and categorizes crime. We determined it would not make sense to use this as an independent variable in any model in which crime rate was the dependent variable. Exploring the factors in face-to-face crime is a legitimate question, but we chose to focus on the broader crime rate.

[6] https://www.nass.usda.gov/Statistics_by_State/North_Carolina/Publications/County_Estimates/index.php

Minority Percentage

We noticed a slight positive correlation between minority population percentage and crime rate:

```
ggplot(crime, aes(pctmin80, crmrte)) +
  geom_point(size=1.5,na.rm = TRUE,col="steelblue") +
  geom_smooth(method = "lm", color="black",se = FALSE) +
  labs(title = "Minority percentage vs Crime rate",
       x = "Minority percentage", y = "Crime rate")+
  theme_bw()
```



Minority percentage may be an indicator of poverty. North Carolina was a slaveholding and Jim Crow state, and even the demise of *de jure* segregation under Jim Crow was followed by *de facto* segregation, white flight, and urban renewal that affected black businesses in such cities as Greensboro and Durham. As such, many minority communities in North Carolina are historically economically depressed and `pctmin80` may be an indicator of poverty. The state does have poor white communities, though, so it is far from an ideal predictor.

Probability of Arrest, Conviction, Prison Sentence, Average Sentence

Looking at the 'probability' variables, we made the following observations:

- The Probability of conviction field was classed a factor instead of a number; this required conversion to be usable.
- Probability of arrest and probability of conviction were defined as ratios of arrests to offenses and convictions to arrests, respectively. Both fields contained data exceeding 1, which appeared counterintuitive and required additional study.

```
crime$prbconv <- as.numeric(paste(crime$prbconv))
arrGr1 = sum(crime$prbarr>1)
convGr1 = sum(crime$prbconv>1)
```

There was a single record with the ratio of arrests to offenses greater than 1. This anomaly could reflect an error in data gathering, but it is conceivable that the definition of probability of arrest was not confined to a single arrest per crime, and that multiple arrests could be made for a single offense. In this scenario, it is possible that the ratio could exceed 1, so we opted not to adjust this outlier.

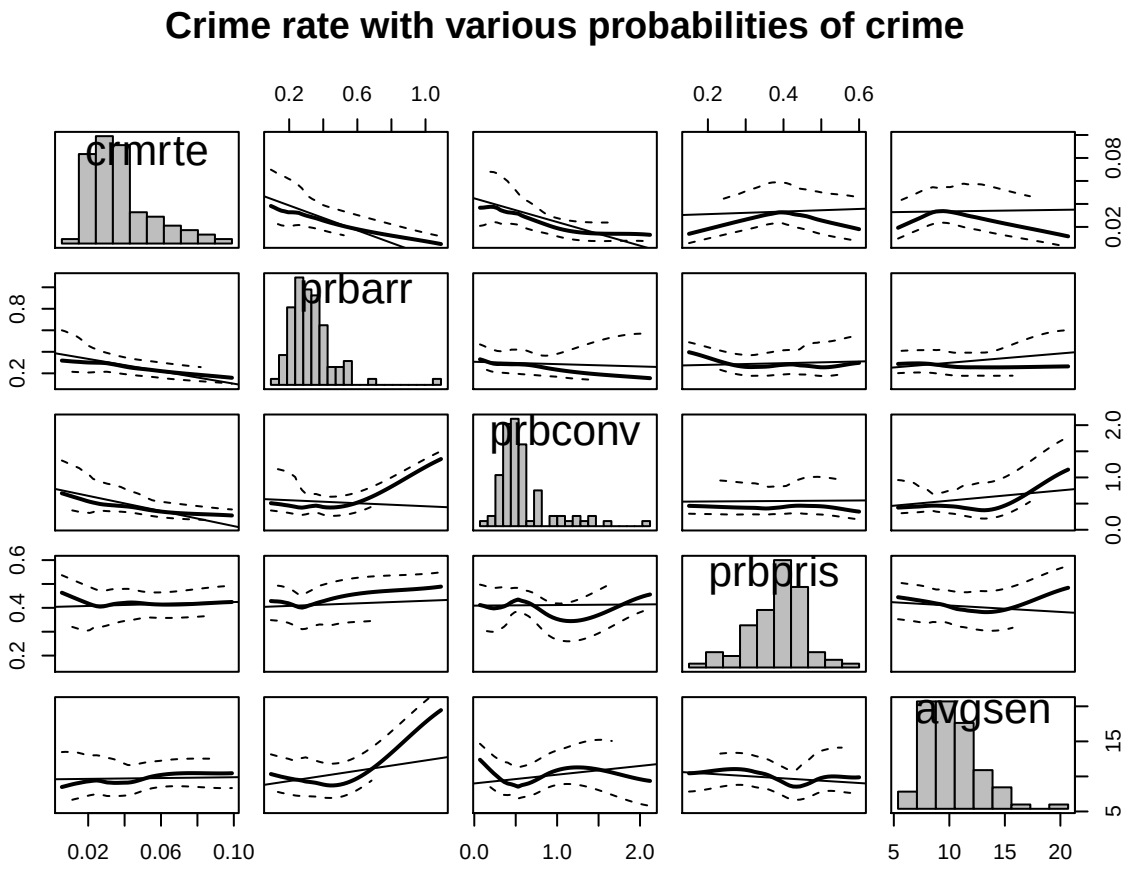
With 10 records, probability of conviction represented a much more substantial problem. Multiple convictions are permitted for a single offense. Convictions may also have occurred in 1987 related to arrests from prior periods, so while unlikely, it is possibile that there could be ratios greater than 1. Therefore, we kept the values unchanged at this stage and made certain to look for high leverage data points as we moved forward.

```
summary(crime[,c("prbarr","prbconv","prbpris","avgsen")])
```

```
##      prbarr           prbconv           prbpris          avgsen      
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.1500   Min.   : 5.380  
##  1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.:0.3642   1st Qu.: 7.375  
##  Median :0.27146   Median :0.45170   Median :0.4222   Median : 9.110  
##  Mean   :0.29524   Mean   :0.55086   Mean   :0.4106   Mean   : 9.689  
##  3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:0.4576   3rd Qu.:11.465  
##  Max.   :1.09091   Max.   :2.12121   Max.   :0.6000   Max.   :20.700
```

In this case, we wanted to observe whether higher rate of arrest, rate of conviction, or length of prison term had any relationship with crime rates.

```
scatterplotMatrix( ~ crmrte + prbarr + prbconv + prbpris + avgsen,
                   data = crime, diagonal=c("histogram"),
  main = "Crime rate with various probabilities of crime" ,
  use=c("complete.obs"), col=c('black','black'), regLine = list(method=lm, col='red'))
```



We did observe lower crime rates associated with increasing probabilities of arrests and convictions. However, the correlation between crime committed (crmrte) and average sentence days (avgsen) was quite low. We then thought it would be interesting to see how probabilities of arrest and conviction were distributed across regions.

Density

The density field was defined as people per square mile. However, the data did not seem to match logically with this definition. As per U.S. Census Bureau data, in 1990, the average number of residents per square mile was 136.4 in North Carolina. The mean in our dataset was 1.42. We made an assumption that our density field was off by a factor of 100. We transformed this field by multiplying density by 100, which produced more logical residents-per-square-mile estimates.

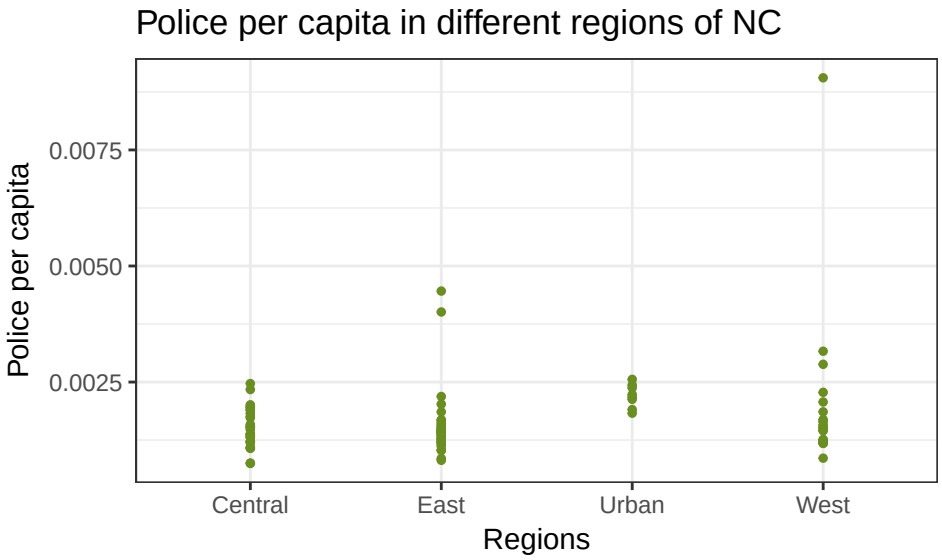
However, even after performing the transformation, there was an abnormally low minimum value of 0.002 residents per square mile for Swain county. As of the 2000 Census, Swain County (FIPS 173) had a population of 11 residents per square mile. Even though we did not modify the data at this stage, we noted the outlier so that we could see later how this high leverage data point influenced our modeling.

```
crime$density100= 100 * crime$density
summary(crime$density100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.002  54.718  97.924 143.567 156.926 882.765
```

Police Per Capita by Region

```
ggplot(crime, aes(x=regWurban, y=polpc), ) +
  geom_point(size=1,na.rm = TRUE,col="olivedrab") +
  labs(title = "Police per capita in different regions of NC",
       x = "Regions", y = "Police per capita")+
  theme_bw()
```



Although the west region had the lowest crime rates, one of its counties had an abnormally high police per capita rate. Further, even though crime is high in urban areas, the police presence per capita did not appear to be overly substantial in these areas.

We hypothesized that we should be able to understand better if we examined crime vs. police per capita across the various regions. The outlier value in the west region had a considerable effect on this comparison. It was extreme - nearly double the next highest rate in the state. Although there is a chance that this was

correct, it is highly unlikely. A comparison with 2015 police employee data[7] revealed that Washington, DC, had the highest U.S. police per capita rate of 0.0065, which is substantially lower than this 0.009 outlier. Hence, we determined that it would be better to replace the outlier with the mean of police per capita for the west region.

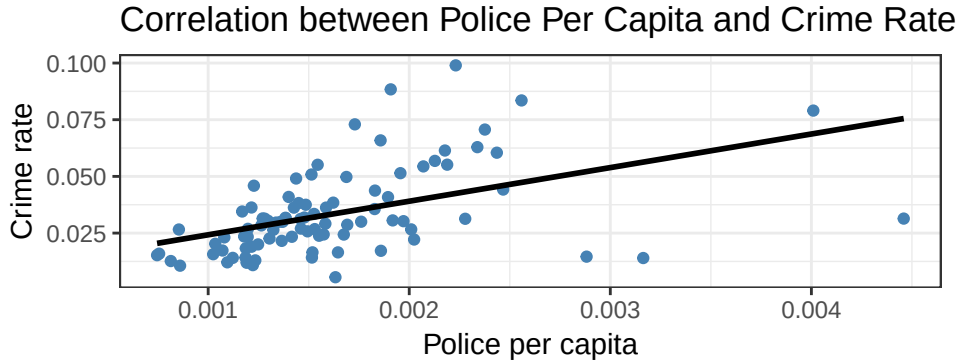
```
crime$polpc =ifelse(crime$polpc == max(crime$polpc),
                    mean(crime[crime$west == 1 & crime$polpc < 0.009,]$polpc),crime$polpc)

summary(crime$polpc)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012378 0.0014897 0.0016255 0.0018587 0.0044592
```

Crime Rate vs Police Per Capita

Initially, we instinctively believed police per capita was a variable that would have a strong relationship with crime rate and planned on including it in our models. However, our exploratory data analysis showed a positive correlation between police per capita and crime rate. Further discussion led us to conclude that police per capita could be thought of as "which came first, the chicken or the egg". Was higher police per capita effective in reducing crime, or was higher police per capita a response to high crime rates or the number of what could be called "severe crimes" (e.g., murder)?

Because of the potentials for reverse causality, we decided not to incorporate the police per capita variable in Model 1 or Model 2.



Tax Per Capita

Note: We assumed that this field represented state income tax. Examining the tax per capita data led us to believe that the scale of this field was off by a factor of 100. Multiplying the data by 100 brought it in line with historical per capita state tax figures.[8] The maximum tax per capita value of $11,976 looked high compared to those in the rest of the counties. We did not modify this point, but noted it for further review.

```
crime$taxpc100= 100 * crime$taxpc
summary(crime$taxpc100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2569    3073    3492    3816    4101   11976
```

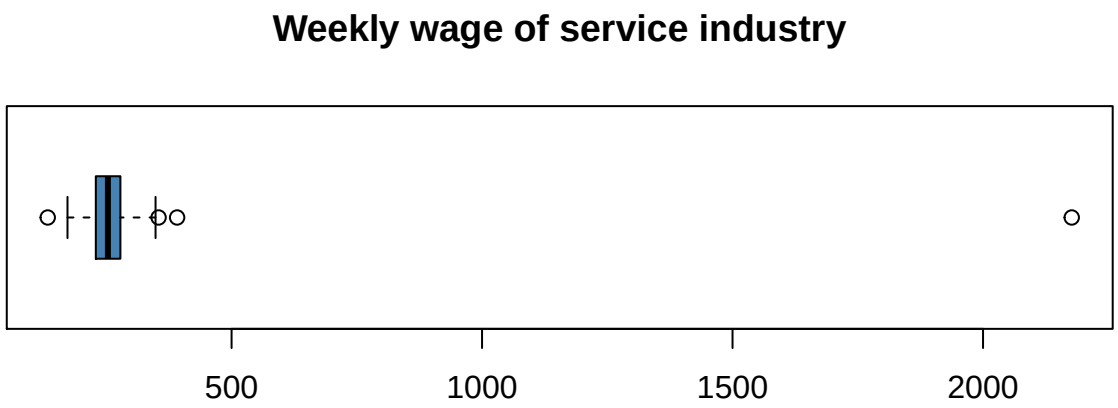
[7] http://www.governing.com/gov-data/safety-justice/police-officers-per-capita-rates-employment-for-city-departments.html

[8] https://taxfoundation.org/state-and-local-tax-burdens-historic-data/

Weekly Wage of Service Industry and Average Wage

Looking at the summary data, we also identified an anomaly in the maximum weekly wage of the service industry ($2,177, vs. the mean of $275). This would be astronomically high for the 1987 time period and the industry, and is likely the result of error. We analyzed how many such instances were present with a box plot.

```
boxplot(crime$wser, main = "Weekly wage of service industry" , col=c("steelblue"),horizontal = TRUE)
```



There was only one county with an abnormally high wage. The value is clearly not representative of the North Carolina population and could have resulted from a data entry error. We decided to replace the lone anomaly with the mean of the rest of that column.

```
crime$wser = ifelse((crime$wser>2100), mean(crime[crime$wser<2100,]$wser),crime$wser)
summary(crime$wser)
```

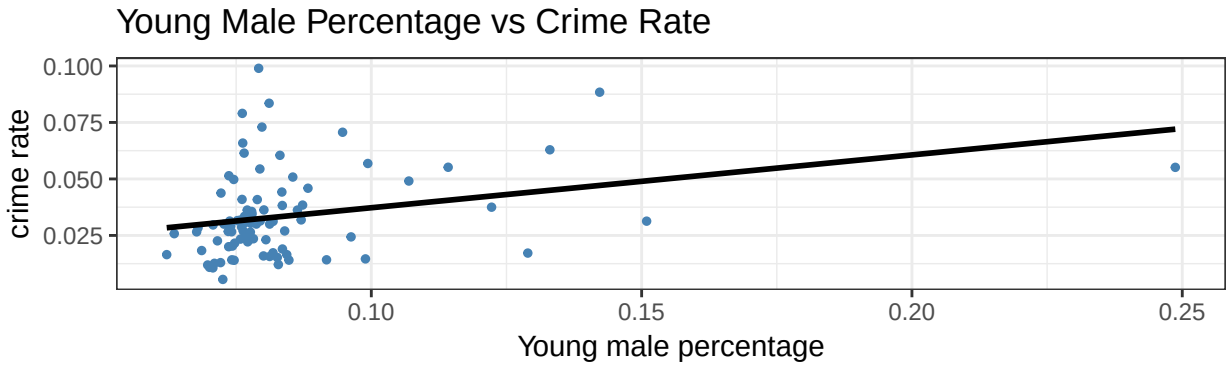
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   254.0   275.9   391.3
```

Crime Rate and Young Male Population

Per the FBI's assessment of variables affecting crime, crime is higher in areas with larger concentrations of youth. This does not specifically call out young males, but research overwhelmingly demonstrates that males commit crimes more often than women.[9] We therefore determined that we wanted to observe the relationship of young male population with crime.

```
ggplot(crime, aes(x=pctymle, y=crmrte), ) +
  geom_point(size=1,na.rm = TRUE,col="steelblue") +
   geom_smooth(method = "lm", color="black",se = FALSE) +
    labs(title = "Young Male Percentage vs Crime Rate",
       x = "Young male percentage", y = "crime rate")+
  theme_bw()
```

[9] https://books.google.com/books?id=CJm4AIc4sZEC&pg=PA88#v=onepage&q&f=false



We did observe that there was one data point for which the young male percentage was much higher than those of the rest of the counties. The census data of 2000 reported a median age of 24 in Onslow County (FIPS 133), home to Camp Lejuene. It is a very young county, explaining the spike in the young male percentage. We assumed this was also true in 1987 and opted not to modify this data point.

EDA Observations

After looking at the data closely, we had several observations:

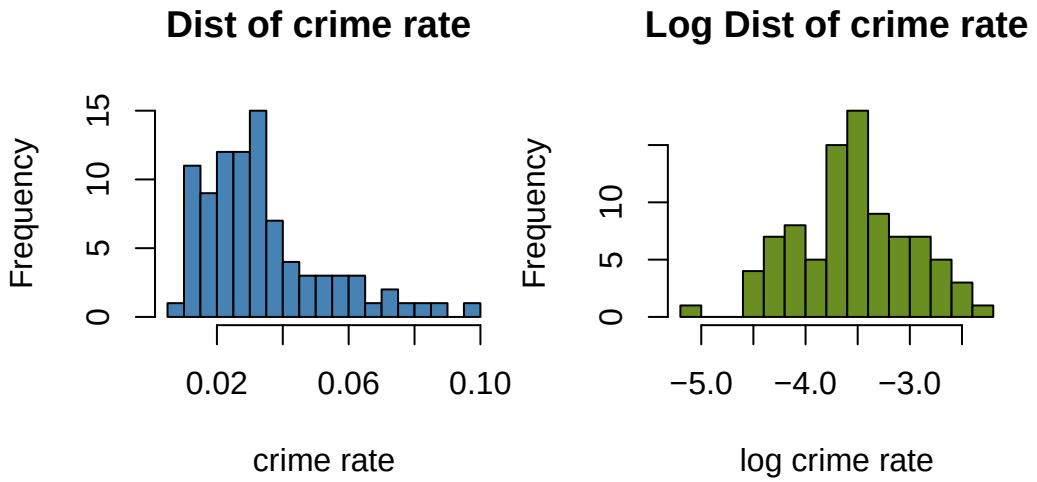
- Crime had a positive correlation with density. This was expected.
- Crime had a negative correlation with probability of arrest and probability of conviction. This was expected as well, as more police intervention should theoretically correlate with less crime.
- Crime had a positive correlation with police per capita. This was quite unexpected and may be a case of reverse causality. We will discuss this in more detail in a later section.
- Crime rates were higher in urban areas than in the rest of the regions, an expected result.

3. Model Specification

Determination of Dependent Variable

In building our model, our first step was to select a dependent variable to represent crime; from the available data, our clear best option was **crime rate**. We first examined the distribution of this variable.

```
hist(crime$crmrte,main="Dist of crime rate ",xlab = "crime rate",breaks = 20,
     col="steelblue")
hist(log(crime$crmrte),main="Log Dist of crime rate",
     xlab = "log crime rate",breaks = 20,col="olivedrab")
```



The histogram of the crime rate data was positively skewed. We also looked at employing a logarithmic transformation to the crime rate data, and the resulting distribution looked mostly normal. Even though there were no regression assumptions that required the dependent variable to be normal, there were many advantages of having a logarithmic dependent variable:

- It allows non-linear or very general relationships between variables.
- In a model with heteroskedasticity, a logarithmic transformation suppresses variation.
- The logarithm suppresses skewness, leading to more normal errors.
- If the dependent variable contains outliers, a logarithmic transformation can reduce the influence of those observations.
- It would allow us represent the crime rate in percentage terms.

With these factors in mind, we decided to make **log(crmrte)** our **predicted variable**.

Determination of Control Variables

Our initial, general research question then followed. We sought to:

Examine the relationship between identified independent variables and the dependent variable, crime rate, in order to determine whether a policy of changing the independent variables would be effective in lowering crime rates.

We next considered this problem intuitively. What factors did we instinctively assume would result in high crime rates? Classically, we would consider major factors to include:

- Urban areas/population density
- Unemployment
- Economic conditions/poverty
- Police presence
- Social services
- High concentration of youth in a population

These are consistent with the FBI's assessment of variables that affect crime.[10] However, the FBI also cites additional factors, including:

- Stability of the population with respect to residents' mobility
- Modes of transportation and highway system
- Cultural factors
- Climate
- Family conditions

[10] https://ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime

- Criminal justice policies
- Citizens' attitudes towards crime

Comparing these established factors against the available data suggested that we might see substantial omitted variable effects, but we concluded there should be enough data to proceed with our analysis.

General Model Assumptions

CLM1 and CLM2 are common across all the models. They are summarized below:

CLM1. Linear in Parameters: Here, the parameters are the coefficients on the independent variables (often marked as β). We made sure that all our models had linear coefficients.

Note: Although the coefficients must be linear, the dependent and independent variables may not be, which allowed us to model nonlinear relationships.

CLM2. Random Sampling: This dataset was first used in a study by Cornwell and Trumball, researchers from the University of Georgia and West Virginia University. Apart from that, we had no knowledge of how the data was collected and would need more details to determine whether we had a true random sample.

Model 1

Based on our EDA analysis and research, we identified the following as our predictor variables:

- **Density** - High population density is classically associated with high crime. It also brings forth factors like urbanization and changes in the area's economic conditions. The density variable in this dataset had very high maximum and minimum values. We determined that taking the logarithm of density would help reduce the residual that would be generated exclusively due to these high leverage points. Moreover, it would help us represent the data as a percentage.
- **Probability of Arrest** - Since probability is between 0 and 1, we were able to multiply by 100 to represent it as a percentage and discuss the percent variation of crime rate with each percentage point variation in arrest rate.
- **central** - This is a dummy variable to help us understand the characteristics of central region, with respect to the eastern region.
- **west** - This is a dummy variable to help us understand the characteristics of western region with respect to eastern region.
- **Young male pct between ages 15-24** - As with probability of arrest, this variable was easily represented as a percentage by multiplying by 100. This made it easy to report the practical outcome of any relationship with crime rates. Our experience told us that population density and youth concentration could have a relationship with crime rates. We did not have a variable in our dataset to represent youth concentration, but we did have percent of young males in the population. We deemed this an adequate proxy for overall youth, but any gender component would also have been built into these figures. Encompassing both of these factors, we determined that it would be a strong candidate for a useful predictor.

```
crime$prbarr100 = crime$prbarr*100
crime$pctymle100 = crime$pctymle*100
```

We chose to include probability of arrest in our analysis. Defined as the ratio of arrests to offenses, we determined that this would serve as an adequate proxy for police effectiveness. At the same time, we want to note that probability of arrest can be impacted by various factors, such as: how well police officers feel they are being compensated (pay could be commensurate with motivation) and bribery, which can result in police "looking the other way". Additionally, the probability of arrest was drawn from the FBI's Uniform Crime Report, which gathers data through "a nationwide, cooperative statistical effort of nearly 18,000 city, university and college, county, state, tribal, and federal law enforcement agencies voluntarily reporting data

on crimes brought to their attention" (note the word "voluntarily" - would police departments with statistics that reflected poor performance report their data?).

Note: We also wanted to analyze the urban indicator variable, but omitted it for two reasons: first, it was highly correlated with density, and could have resulted in multicollinearity problems in our model. Second, urban is not mutually exclusive with any of the regional indicator variables, so we would have had overlap in our data, had we used it. Therefore, we opted to focus exclusively on the geographic categories.

Selection of our initial independent variables resulted in our refined research question. We sought to:

Examine the relationship between probability of arrest, population density, and high youth concentration and the dependent variable, crime rate, in order to determine whether related policy changes could be effective in lowering crime rates.

The equation for the first iteration of our model was:

$$log(crmrte) = \beta_0 + \beta_1 prbarr100 + \beta_2 log(density100) + \beta_3 pctymle100 + \beta_4 central + \beta_5 * west + u$$

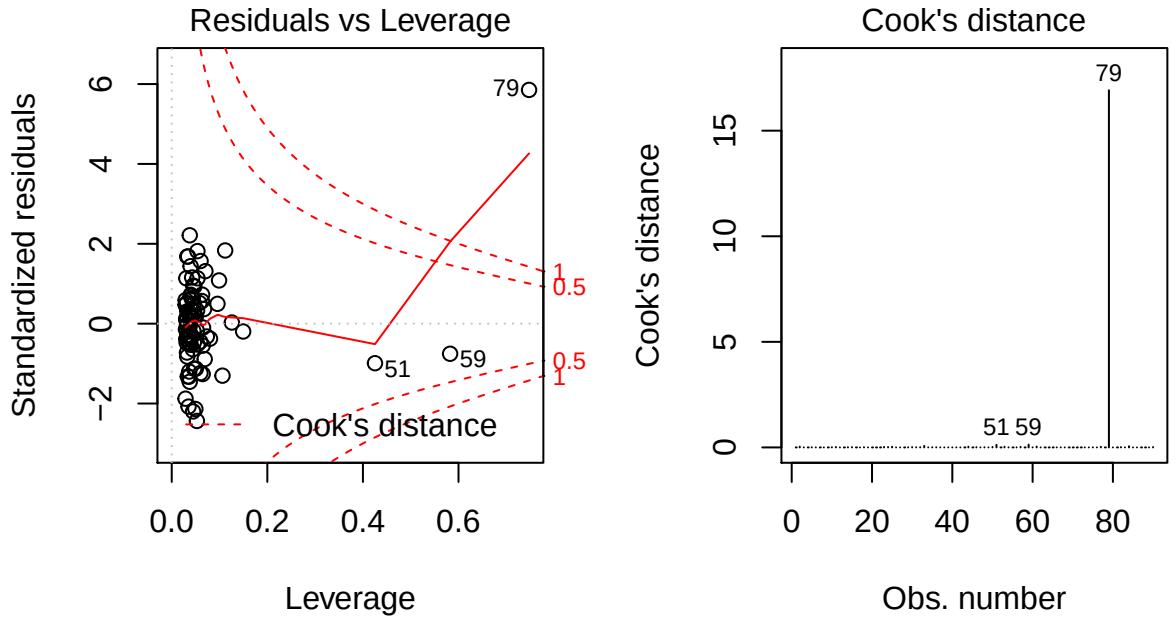
```
# Run model 1
model1= lm(log(crmrte)~prbarr100+log(density100)+pctymle100+central + west,data=crime)
```

Observations from Model 1: Highlights, Model Coefficients, and Goodness of Fit Analysis

Troubleshooting Outliers, High Leverage, and Influence

We analyzed the Residual vs. Leverage diagnostic plot to identify influential cases, extreme values that might influence the regression results when included or excluded from the analysis.

```
par(mfrow=c(1,2));plot(model1,which = 5);plot(model1,which = 4)
```



There were a few high leverage points (#51, #59, and #79) with extreme predictor values, but only #79 proved highly influential, with a Cook's distance greater than 1. We observed that the standardized residual

of the same point (#79) was greater than 3, a definite outlier. In order to address the issue, we looked at point #79 and found it to be the previously identified population density record with an abnormally low value of 0.002 residents per square mile. At this point, we were more confident that the outlier was an error. Therefore, we opted to replace the value with the mean of the density and repeat the regression.

```
crime$density100 = ifelse((crime$density100<0.01),
                          mean(crime[crime$density100>0.01,]$density100),
                          crime$density100)
model1= lm(log(crmrte)~prbarr100+log(density100)+pctymle100+central + west,data=crime)
```

After normalizing the density outlier, we saw no observations of high influence. However, there was still one outlier observation, #25, with standardized residual greater than 3. Looking at the data, we had not identified anything suspicious about that record, so we decided to monitor it for later models.

Goodness of Fit - Model 1

The below table shows measures of fit for Model 1.

	Goodness of Fit for Model 1		
	Multiple R-squared	Adjusted R-squared	AIC
Model 1	0.67	0.65	59.88

With an adjusted R^2 of 0.65, 65% of the variation in crime rate can be explained by probability of arrest, density, and the young male population.

Explanation of Coefficients - Model 1

Model1.Coefficients		Interpretation
(Intercept)	**-5.3408**	
prbarr100	**-0.0084**	For approximately each percentage point increase in probability of arrest, crime rate is associated with a drop of 0.84 %, holding other variables constant.
log(density100)	**0.4622**	For approximately every percent increase in density, crime rate is associated with an increase of 0.46 %, holding other variables constant.
pctymle100	**0.0162**	For approximately every percentage point increase in the young male population, crime rate is associated with an increase of 1.62 %, holding other variables constant.
central	**-0.2874**	Crime rate in the central region is 28.74 % lower than the eastern region.
west	**-0.5217**	Crime rate in the western region is 52.17 % lower than the eastern region.

CLM3. No Perfect Multicollinearity

For a given predictor (x_i), multicollinearity can be assessed by computing a score called the variance inflation factor (or VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 4 indicates a problematic amount of collinearity.

```
kable(t(vif(model1))) %>%
kable_styling(bootstrap_options =  c("striped", "bordered"),
              full_width = F, position = "center")%>%
  add_header_above(c("VIF evaluation for Model1" = 5))
```

VIF evaluation for Model1				
prbarr100	log(density100)	pctymle100	central	west
1.178096	1.387008	1.113744	1.459218	1.21978

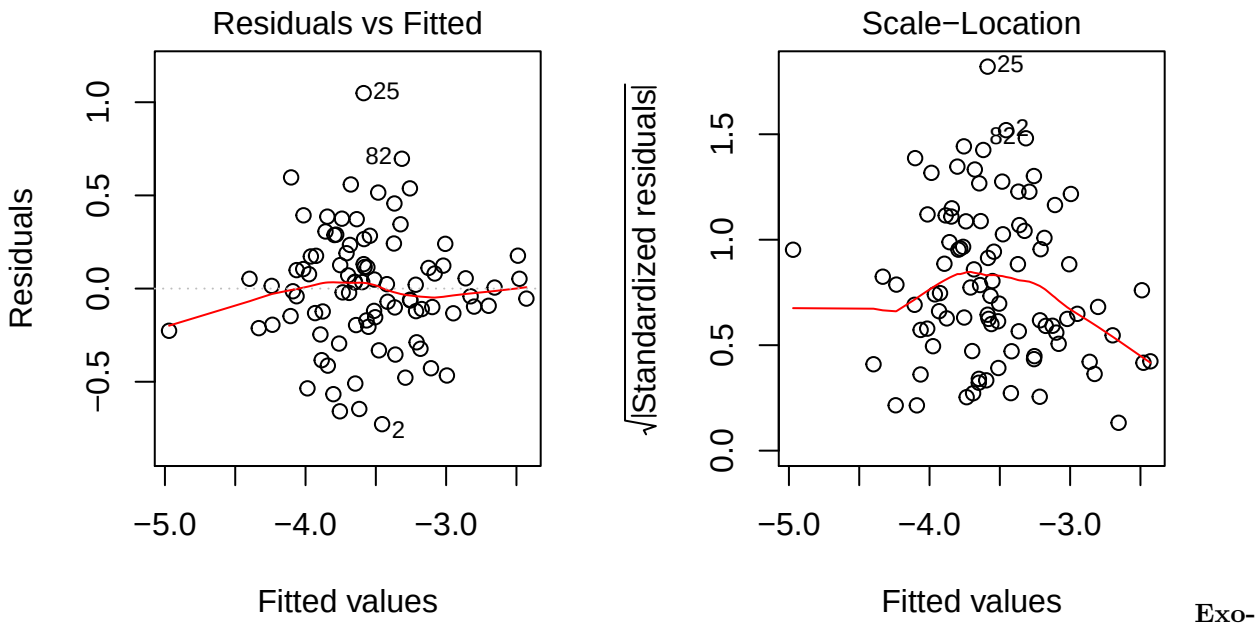
In Model 1, none of the predictor variables showed any problematic amount of collinearity. The VIF values were well below 4 for all predictors.

CLM4-CLM5. Evaluating Zero-Conditional Mean and Homoskedasticity

Zero Conditional Mean: Looking at the Residual vs. Fitted diagnostic plot, we observed some deviations from zero, demonstrating violations of the zero-conditional mean assumption. The more extreme dip on the left of the plot red line dipped towards the left appeared to be mainly due to a lack of data points, but that point notwithstanding, there were other deviations in the center of the plot. This meant that our coefficients would be biased. This may have happened because of:

- Omitted variables
- Functional misspecifications
- Measurement errors in independent variables

```
par(mfrow=c(1,2));plot(model1,which = 1);plot(model1,which = 3)
```



Exogeneity: We had a large sample size (90), so ordinary least squares (OLS) asymptotics should allow us to proceed, despite the violation of zero-conditional mean, if we could meet the less stringent exogeneity assumption, $E(Xu) = 0$ (or that X and u are uncorrelated).

If we were trying to show causality, the zero-conditional mean assumption violation would have been more of a problem, but for our associative model, we just wanted to track the best fit line in the population. In that scenario, our estimates would be consistent, and we would essentially meet exogeneity by definition.

We checked the correlation between the independent variables and the residuals:

```
meanErr1 =round(mean(model1$residuals),digits=2)
covArr1 =round(cov(crime$prbarr100,model1$residuals),digits=2)
```

```
covden1 =round(cov(log(crime$density100) ,model1$residuals),digits=2)
covmle1 =round(cov(crime$pctymle100,model1$residuals),digits=2)
```

The mean of error E(u) = 0
The covariance between prob of arrest and error Cov(prbarr100,u) = 0
The covariance between density and error Cov(log(density),u) = 0
The covariance between young male pc and error Cov(pctymle100,u) = 0

None of the control variables were correlated with the error term, so we could safely infer that **Model 1 was consistent**.

Homoskedasticity: Homoskedasticity says variance of residuals should be constant, with a mean of zero and variance σ^2. For homoskedasticity, we analyzed the Residual vs. Fitted and Scale-Location diagnostic plots.

1. Looking again at the Residuals vs. Fitted plot, it appeared that the band of data points did not have a uniform thickness, indicating heteroskedasticity.
2. The Scale-Location plot did not have a consistent horizontal band of points from left to right, again indicating heteroskedasticity.

3. Breush-pagan test: The null hypothesis is that the model is homoskedastic.

```
#Breush-pagan test
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 9.6533, df = 5, p-value = 0.08568
```

The Breush-pagan test gave us a p-value of 0.08568, which was not sufficient to reject the null hypothesis of homoskedacity in the model. Despite this evidence, we opted to proceed using heteroskedasticisty-robust standard errors, as that is good practice.

CLM6. Normality of Errors

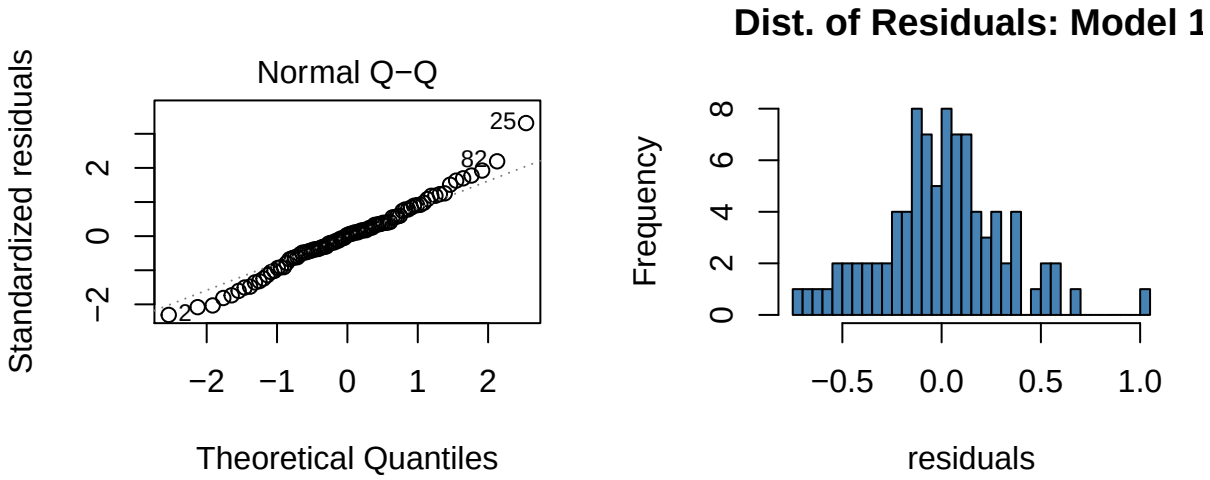
The normality of errors assumption says that the variance of residuals is constant, with a mean of zero and variance of σ^2.

To check normality of errors, we examined:

1. The Quantile-Quantile (Q-Q) plot, part of R's standard diagnostics
2. A histogram of the errors
3. Shapiro-Wilk test (where the null hypothesis is that the errors are normal)

The Q-Q plot of residuals and the histogram of errors can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

```
par(mfrow=c(1,2));plot(model1, which = 2)
#plot a histogram of the residual
hist(model1$residuals, breaks = 30,main="Dist. of Residuals: Model 1",xlab = "residuals",
     col="steelblue")
```



Looking at the above plots, we observed that the errors were not exactly normally distributed and there were outliers pulling the data away from the diagonal. However, since we had a large dataset (>30), we were able to rely on asymptotic properties of OLS, which state that our estimators will have normal sampling distributions for large sample sizes.

```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.98663, p-value = 0.49
```

The Shapiro-Wilk normality test was not sufficient to reject the null hypothesis of normality.

Analyzing Statistical and Practical Significance of Model 1

Statistical significance refers to the unlikelihood that the result is obtained by chance, i.e., the probability that a relationship between two variables exists. We looked at the t-tests to determine whether or not to reject the null hypothesis (which says that the parameters are equal to 0) at a 0.05 level of significance. We used heteroskedasticity-robust standard errors to determine p-values.

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

```
##
## t test of coefficients:
##
##                   Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)     -5.3408213  0.3198298 -16.6990 < 2.2e-16 ***
## prbarr100       -0.0083781  0.0030203  -2.7739  0.006823 **
## log(density100)  0.4622285  0.0531658   8.6941 2.450e-13 ***
## pctymle100       0.0162045  0.0110567   1.4656  0.146497
## central         -0.2873736  0.0978643  -2.9364  0.004282 **
## west            -0.5216919  0.0775565  -6.7266 1.984e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the p-value of each estimate, we observed that probability of arrest, density, and regions (west and central) are statistically significant. However, even though young male population did not have a statistically significant effect on the crime rate, we know from our experience and various research conducted over years [11] that the young male population is susceptible to crime. All of the coefficients in Model 1 were of sufficient size that we deemed all to be of practical significance.

Model 2

Model 2 built on Model 1 by adding tax per capita (taxpc) and probability of conviction (prbconv). We investigated whether the probability of arrest (prbarr100) and prbconv were correlated, but found they were not. We noted that this finding should be further explored, as it could indicate effects of external factors, such as the number of arrests reported were artificially inflated, jails/prisons were overpopulated, or people were being arrested for crimes that did not warrant arrest.

Rationale for including Tax Per Capita and Probability of Conviction:

- **Tax Per Capita** - Taxes are used to fund public services such as police, district attorneys, judges, jails, and all those who work in them. Tax per person is relevant because it means a town or county likely has more financial resources to allocate to safety and protection.

- **Probability of Conviction** - Intuitively, the probability of conviction likely has a strong impact on crime rate given that convictions result in some level of "punishment" in the form of jail time, fines, and/or public service.

The equation for Model 2 was:

$$log(crmrte) = \beta_0 + \beta_1 prbarr100 + \beta_2 log(density100) + \beta_3 pctymle100 + \beta_4 central + \beta_5 west + \beta_6 taxpc + \beta_7 prbconv100 + u$$

```
crime$prbconv100 = crime$prbconv*100
```

```
model2= lm(log(crmrte)~prbarr100 + log(density100) + pctymle100 + central +
             west + taxpc100 + prbconv100 , data=crime)
```

Model 2 Observations and Analysis

Goodness of Fit

First, let's look at goodness of fit and compare it to Model 1. AIC estimates the relative quality of a model. A lower number is better. Note the AIC for Model 2 is 30.95 compared with Model 1 at 59.88, almost a 50% improvement, based on AIC values. In addition, the adjusted R^2 value for Model 2 increases to 0.75 (from Model 1's 0.65). Because adjusted R^2 measures the proportion of variation in the dependent variable (crimerte), a high adjusted R^2 is desired and accomplished with Model 2.

	Goodness of Fit for Model 2		
	Multiple R-squared	Adjusted R-squared	AIC
Model 2	0.77	0.75	30.95

[11] http://www.nber.org/chapters/c6806.pdf

Explanation of Coeficients - Model 2

Model.2.Coefficients		Interpretation
(Intercept)	**-4.8419**	
prbarr100	**-0.0102**	A percentage point increase in probability of arrest is associated with a crime rate drop of 1.02 %, holding other variables constant.
log(density100)	**0.3719**	A 1% increase in population density is associated with a crime rate increase of 0.37 %, holding other variables constant.
pctymle100	**0.0133**	A percentage point increase in the population of young males is associated with a crime rate increase of 1.33 %, holding other variables constant.
central	**-0.2407**	Crime rate in the central region is 24.07 % lower than the eastern region.
west	**-0.4585**	Crime rate in the western region is 45.85 % lower than the eastern region.
taxpc100	**0.0001**	A $1 increase in the taxes paid per person is associated with an increase in the crime rate of 0.01 %, holding other variables constant.
prbconv100	**-0.0046**	A percentage point increase in the conviction rate is associated with a 0.46 % decrease in crime rate, holding other variables constant.

CLM3. Quantifying Multicollinearity

When variables in a model are highly correlated but not perfectly collinear, linear regression works, but estimated values are much less precise.

In the case of Model 2, the VIF scores for each independent variable indicated little to no multicollinearity.

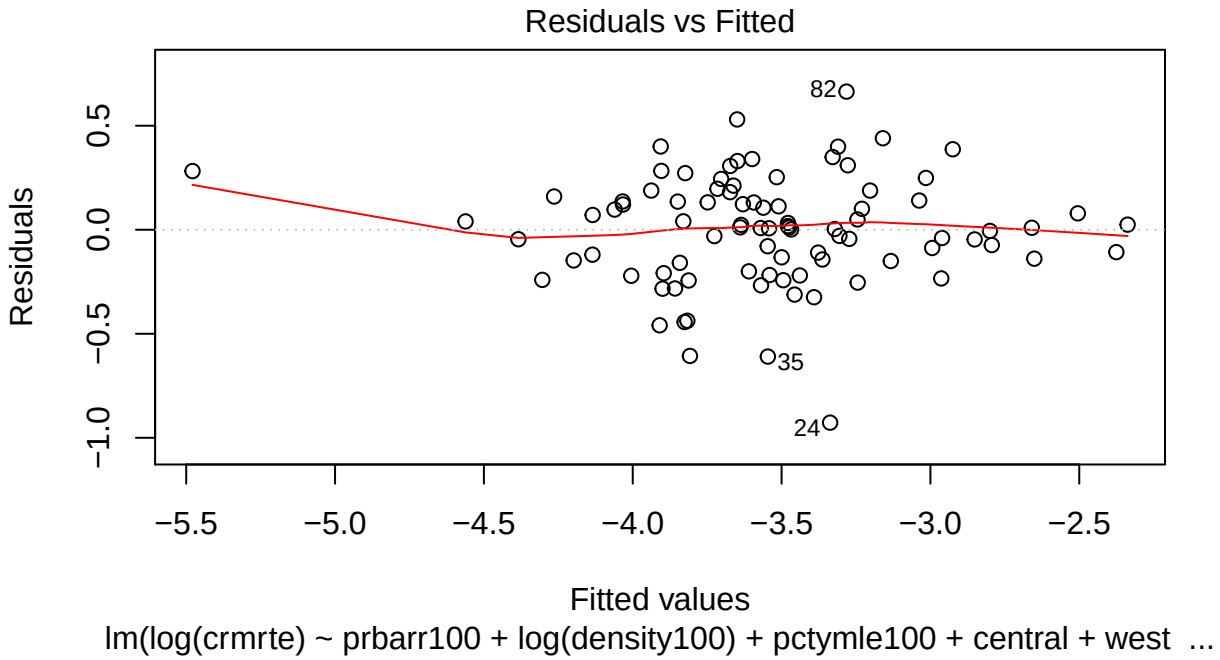
```
kable(t(vif(model2))) %>%
kable_styling(bootstrap_options =  c("striped", "bordered"),
              full_width = F, position = "center")%>%
  add_header_above(c("VIF evaluation for Model 2" = 7))
```

VIF evaluation for Model 2						
prbarr100	log(density100)	pctymle100	central	west	taxpc100	prbconv100
1.228996	1.549802	1.174992	1.498187	1.278611	1.135137	1.140064

CLM4. Zero Conditional Mean

In the Residuals vs. Fitted plot below, the red spline curve should be straight and centered at zero on the y-axis to meet the zero conditional mean assumption. For the most part, the red line is straight at zero on the y-axis and we concluded this assumption was met. The slight upward bend on the left appeared to be the result of a lack of data.

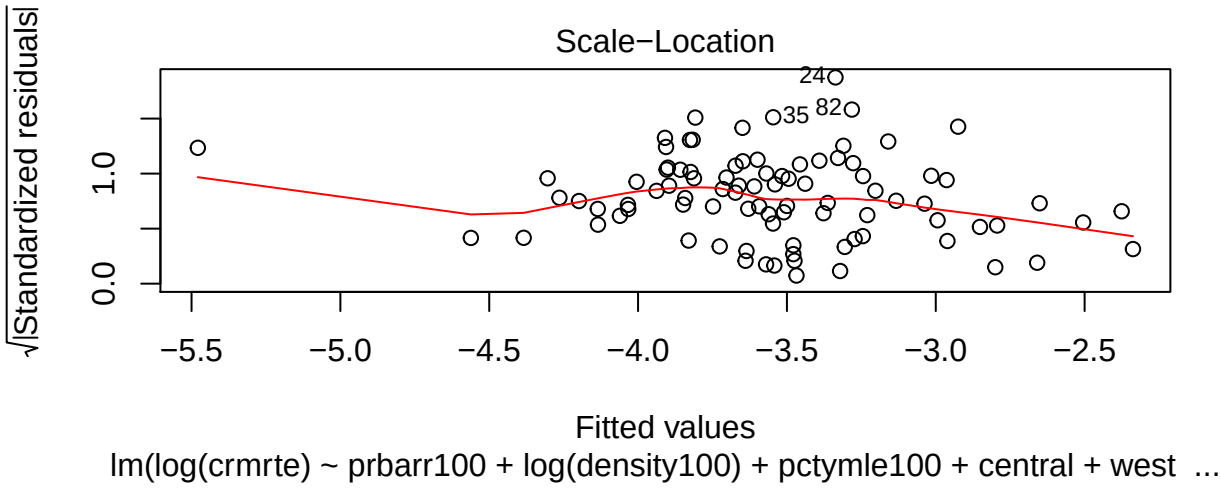
```
plot(model2,which = 1)
```



CLM5. Assessing Heteroskedasticity Levels

We assessed levels of heteroskedasticity by evaluating the Scale-Location plot and conducting a Breusch-Pagan Test. As seen in the Scale-Location plot below, there is a narrow-wide-narrow shape to the data in combination with a "wavy" red line. Both of these indicate the presence of heteroskedasticity.

```
plot(model2,which = 3)
```



In the Breush-pagan output (below), the p-value equals 0.20 which was not sufficient to reject the null hypothesis of homoskedasticity in the model. As with Model 1, we opted to proceed using heteroskedasticisty-robust standard errors.

```
# Confirm heteroskedasticity with Breusch-Pagan test
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
```

```
##
## data:  model2
## BP = 9.768, df = 7, p-value = 0.2021
```

CLM6. Normality of Errors

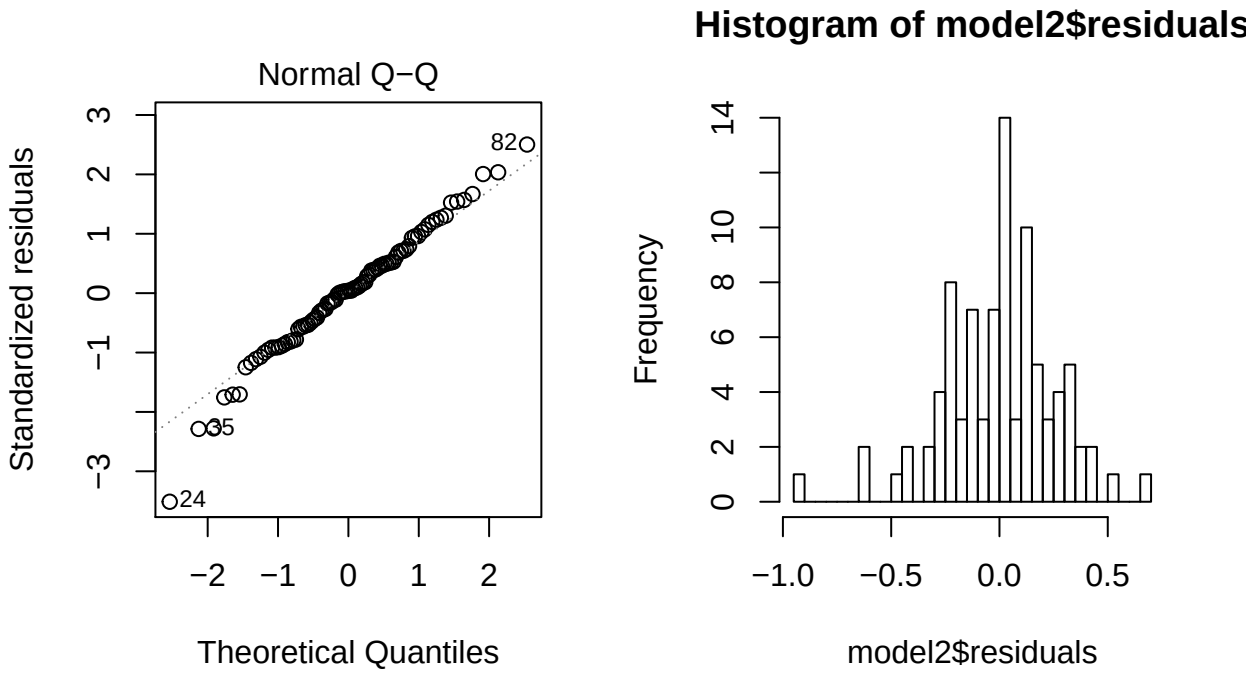
To understand the normality of the residuals, the below shows the Q-Q Plot and a histogram of residuals.

Ideally, the Q-Q plot's data points would perfectly align, and for the most part, we can see that effect below. Due to outliers, data points 24, 35, and 82 contributed toward a deviation from the straight line.

This outcome can also be seen in the adjacent histogram of Model 2 residuals. A residual histogram should reflect a normal distribution. In this case, the outliers resulted in a left-skewness and too much variability in the center values.

In-depth analysis of the above three data points, in addition to considering the possibility of omitted variable bias, should be conducted to coerce a more normal distribution of residuals.

```
par(mfrow=c(1,2));plot(model2, which = 2);hist(model2$residuals, breaks = 40)
```



A further gauge of model strength is measuring the covariance of the dependent variable against each of the independent variables. The calculation should result in "0" for each comparison as it does for Model 2 below.

```
## [1] "Model 2 Residual Mean : 0"
## [1] "Probability of Arrest :  0"
## [1] "Population Density:  0"
## [1] "Percent Young Males:  0"
## [1] "Central Region:  0"
## [1] "West Region:  0"
## [1] "Tax Per Capita:  0"
## [1] "Probability of Conviction:  0"
```

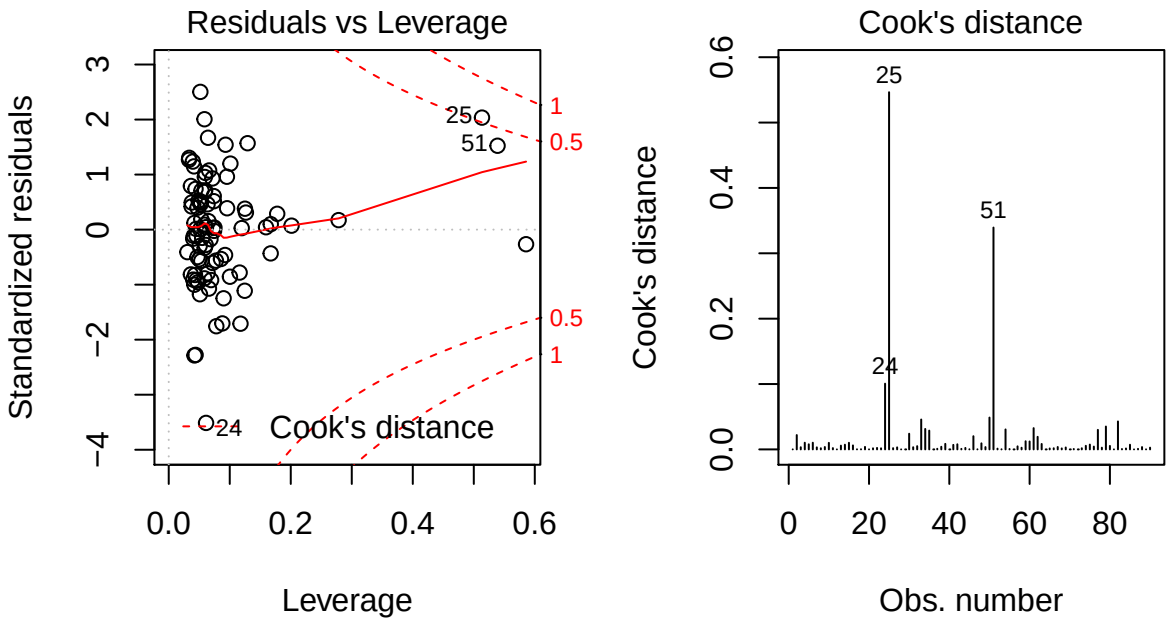
Identifying Outliers, Leverage, and Influence

As we saw in our evaluations of Model 2, outliers influenced some of our diagnostics. The below two charts provide insight into which data points resulted in leverage and influence.

In the Residuals vs. Leverage diagram, data points #25 and #51 fall far to the right on the x-axis, indicating they had sizable leverage. Of even greater note is #25, which fell slightly outside the Cook's distance of .5, indicating substantial influence and the cause of the regression line moving up on the right-hand side.

Point #25 is Dare County, a small county on the Atlantic. It is predominantly white and wealthy, and its main draw is tourism. Its composition could be considered an anomaly when evaluating crime rate, though we did not remove it from this analysis.

```
par(mfrow=c(1,2));plot(model2,which = 5) ;plot(model2,which = 4)
```



Analyzing the Statistical and Practical Significance of Model 2

Statistical significance measures how unlikely it is that something happens by chance. As with Model 1, we looked at the t-tests to determine whether or not to reject the null hypothesis and used heteroskedasticity-robust standard errors to determine p-values.

Based on the p-values below, we can conclude the following variables have statistically significant relationships with crime rate: probability of arrest, density, central, west, and probability of conviction.

Interestingly, tax per capita did not reflect statistical significance. It also showed a slight positive relationship with crime rates in our model, which was counterintuitive, based on our assumptions that more public funding would correlate with lower crime. This may be another effect of population density, as urban centers tend to have higher tax rates (sometimes including city taxes). In order to make a clearer assessment of this variable, we would need to normalize tax rates across the state and re-examine the effects. For those variables with a p-value < 0.05, the coefficients were substantial enough to warrant practical significance.

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

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

```
##                    Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)     -4.8419e+00  3.6377e-01 -13.3104 < 2.2e-16 ***
## prbarr100       -1.0175e-02  3.8813e-03  -2.6216  0.010429 *
## log(density100)  3.7193e-01  5.7207e-02   6.5014 5.833e-09 ***
## pctymle100       1.3345e-02  8.5195e-03   1.5664  0.121102
## central         -2.4066e-01  7.4230e-02  -3.2421  0.001716 **
## west            -4.5850e-01  7.4939e-02  -6.1183 3.084e-08 ***
## taxpc100         5.6376e-05  4.9215e-05   1.1455  0.255334
## prbconv100      -4.5515e-03  1.0771e-03  -4.2256 6.143e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3

Our third model added more coviarates. Before running the model, we saw several pitfalls to this approach:

- Added variables could have correlation with existing variables or each other. We would see if this is the case when examining the model for CLM assumption 3.
- Added variables increase the degrees of freedom in error calculations, reducing precision in all variables.

For variables such as `crmrte` or `density100`, the transformations we applied for the models above were applied here for consistency. Once again, `log(crmrte)` is our outcome variable. We added the following variables:

- `avgsen` and `prbpris` Probability of prison and average sentence are plausible deterrents to crime.
- `pctmin80`. As noted earlier, this variable could be an indicator of poverty.
- *Various wage measures.* We kept each wage variable separate. We considered creating a composite wage metric (e.g., average), but the industries across the diversity of North Carolina counties vary mightily, and many key industries are not represented in this set. Without knowing how many people work in those industries (thus enabling a weighted average), a simple averaging of the wage categories would bear little resemblance to each county's actual average wage or any real-life metric. Further, data on industry-specific weekly wages spanned approximately six categories of industry, but those six categories did not encompass the following top industries in North Carolina: Textiles, Aerospace & Defense, Energy, and Furniture.

```
# Run model 3
crime$prbpris100 = 100*crime$prbpris
model3 <- lm(log(crmrte)~prbarr100+log(density100)+pctymle100+pctmin80+central + west +
    taxpc100 + avgsen + prbconv100 + prbpris100 + wcon + wsta + wfed + wloc + wser + wfir +
    wtrd + wtuc + wmfg, data=crime)

paste("R squared value of model3 : ", round(summary(model3)$r.squared,digits = 2))
paste("Adjusted R squared value of model 3: ",
      round(summary(model3)$adj.r.squared,digits = 2))
paste("AIC of model 3 :" ,round(AIC(model3),2))
```

```
## [1] "R squared value of model3 :  0.86"
## [1] "Adjusted R squared value of model 3:  0.82"
## [1] "AIC of model 3 : 12.61"
```

As expected, the regular and adjusted R^2 increased with the addition of the extra variables. The AIC shrank as well, but there were other issues with this model:

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

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

```
##
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     -5.7550e+00  6.7903e-01 -8.4753 2.434e-12 ***
## prbarr100       -1.2910e-02  5.0515e-03 -2.5557  0.012772 *
## log(density100)  3.0456e-01  7.6099e-02  4.0022  0.000154 ***
## pctymle100       3.4722e-02  1.0369e-02  3.3487  0.001310 **
## pctmin80         8.8659e-03  3.3479e-03  2.6482  0.009993 **
## central         -1.3175e-01  7.8450e-02 -1.6794  0.097537 .
## west            -1.0786e-01  1.2314e-01 -0.8759  0.384077
## taxpc100         8.6520e-05  7.2033e-05  1.2011  0.233754
## avgsen           3.7984e-03  1.3864e-02  0.2740  0.784910
## prbconv100      -5.1317e-03  1.2428e-03 -4.1291 9.920e-05 ***
## prbpris100       1.9028e-03  4.0520e-03  0.4696  0.640101
## wcon             1.3526e-04  9.6465e-04  0.1402  0.888894
## wsta            -1.2606e-03  7.0172e-04 -1.7964  0.076740 .
## wfed             2.5354e-03  7.7216e-04  3.2835  0.001602 **
## wloc             1.9584e-03  1.9623e-03  0.9980  0.321711
## wser            -2.1843e-03  1.1806e-03 -1.8502  0.068506 .
## wfir            -8.1943e-04  1.0301e-03 -0.7955  0.429038
## wtrd            -3.7554e-04  1.7665e-03 -0.2126  0.832262
## wtuc             3.1746e-04  6.2397e-04  0.5088  0.612509
## wmfg             5.8805e-05  4.3581e-04  0.1349  0.893052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fewer than half the variables were statistically significant. Moreover, many variables were not *practically* significant, either. The wage measures were consistently small, as was `avgsen`. Of the new variables added, only `wfed` was statistically significant. All others, even combined, were not:

```
linearHypothesis(model3, c("avgsen = 0", "prbpris100 = 0", "wcon = 0", "wsta = 0", "wloc = 0",
                           "wser = 0", "wfir = 0", "wtrd = 0", "wtuc = 0", "wmfg = 0"),
                           vcov = vcovHC)[4]
```

```
##   Pr(>F)
## 1
## 2  0.326
```

Model 3 Assumptions

Despite the model's shortcomings, it is worth briefly considering the CLM assumptions:

CLM3. No Perfect Multicollinearity

```
vif(model3)
```

```
##       prbarr100 log(density100)      pctymle100        pctmin80
##        1.586542        2.993415        1.415751        2.684619
##         central            west        taxpc100          avgsen
##        2.073516        3.155513        1.453810        1.573661
##      prbconv100      prbpris100            wcon            wsta
##        1.485782        1.187407        2.185729        1.509293
##            wfed            wloc            wser            wfir
##        3.188810        2.302690        2.587199        2.714289
##            wtrd            wtuc            wmfg
##        3.031354        1.760963        1.881098
```

The VIFs for Model 3 were consistently higher than 1 and 2, but not so high as to constitute a violation.

CLM4-CLM5. Evaluating Endogeneity vs Exogeneity and Homoskedasticity

Model 3's diagnostic plots (not shown) violated the zero conditional mean assumption and demonstrated clear heteroskedasticity. Model 3 did meet the lesser standard of exogeneity:

```
c <- 0
for (var in list(crime$prbarr100, log(crime$density100), crime$pctymle100, crime$pctmin80,
     crime$central, crime$west, crime$taxpc, crime$avgsen,
     crime$prbconv100, crime$prbpris100, crime$wcon, crime$wsta, crime$wfed,
     crime$wloc, crime$wser, crime$wfir, crime$wtrd, crime$wtuc, crime$wmfg)){
  if (abs(cov(model3$residuals,var)) > 1e-10){
    c <- c+ 1
  }
}
paste(c, "variables have (absolute value of) covariance with the residuals greater than 1x10^-10")
paste("The mean of the residuals is", round(mean(model3$residuals),2))
```

```
## [1] "0 variables have (absolute value of) covariance with the residuals greater than 1x10^-10"
## [1] "The mean of the residuals is 0"
```

CLM6. Normality of Errors

Model 3's Normal Q-Q plot showed violations of normality among errors at both ends, especially at the lower end. As with previous models, our sample size allowed us to tolerate deviations from this assumption.

4. Model Comparison

The following table compares all three of our models. For space considerations, we have omitted five wage variables from Model 3 whose coefficients were less than 0.001.

```
not_prac <- c()
for(j in 1:20){
  if (abs(model3$coefficients[j]) < 1e-3){# if a |coefficient| is less than 0.001
    if(names(model3$coefficients[j]) != "taxpc100"){ #keep taxpc100 as it is in earlier models
      not_prac <- append(not_prac, names(model3$coefficients[j]))
    }
  }
}

# robust standard errors
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))

stargazer(model1, model2, model3, type = "latex",
          no.space = TRUE,
          font.size = "small",
          column.labels = c("Model 1", "Model 2", "Model 3"),
          report = "vcs*", # report SEs
          title = "Linear Models Predicting Crime Rate",
          omit = not_prac, # omit the variables found earlier as not prac significant
```

```
          keep.stat = c("rsq", "adj.rsq", "n"),
          add.lines=list(c("AIC", round(AIC(model1),2), round(AIC(model2),2), round(AIC(model3),2)),
                         c("BIC", round(BIC(model1),2), round(BIC(model2),2), round(BIC(model3),2))),
          se = list(se.model1, se.model2,se.model3), # Add robust SEs
          star.cutoffs = c(0.05, 0.01, 0.001)) # stringent star cutoffs
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 30, 2018 - 14:19:28

Table 1: Linear Models Predicting Crime Rate

	Dependent variable:		
	log(crmrte)		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
prbarr100	−0.008	−0.010	−0.013
	(0.003)**	(0.004)**	(0.005)*
log(density100)	0.462	0.372	0.305
	(0.053)***	(0.057)***	(0.076)***
pctymle100	0.016	0.013	0.035
	(0.011)	(0.009)	(0.010)***
pctmin80			0.009
			(0.003)**
central	−0.287	−0.241	−0.132
	(0.098)**	(0.074)**	(0.078)
west	−0.522	−0.459	−0.108
	(0.078)***	(0.075)***	(0.123)
taxpc100		0.0001	0.0001
		(0.00005)	(0.0001)
avgsen			0.004
			(0.014)
prbconv100		−0.005	−0.005
		(0.001)***	(0.001)***
prbpris100			0.002
			(0.004)
wsta			−0.001
			(0.001)
wfed			0.003
			(0.001)**
wloc			0.002
			(0.002)
wser			−0.002
			(0.001)
Constant	−5.341	−4.842	−5.755
	(0.320)***	(0.364)***	(0.679)***
AIC	59.88	30.95	12.61
BIC	77.38	53.45	65.11
Observations	90	90	90
R^2	0.673	0.773	0.858
Adjusted R^2	0.653	0.754	0.820
Note:		*p<0.05; **p<0.01; ***p<0.001	

Parsimony

As expected, R^2 increased as more variables are added to the models. The adjusted R^2, which penalizes models with more variables, noted a sizeable difference between 1 and 2, but a more modest gain between 2 and 3. In terms of parsimony, Model 3 scored best in AIC while Model 2 scored best in BIC. This is due to the difference in the two calculations; BIC assesses harsher penalties than AIC for extra variables as the number of observations increases – as in the case here, where n = 90 (the multiplier of k, number of parameters, is 2 for AIC and $log(90)$ or ~ 4.5 for BIC).

Significance

In statistical significance, four of five Model 1 coefficients ranked as significant, compared to five of seven for Model 2, and six of 19 for Model 3. Model 3 was also by far the weakest in practical significance.

Variance of Coefficient Estimates

All three models have produced unbiased linear estimates for each parameter. It is also worth noting which model had the smallest variances. Consider the following table, which compares variances for common parameters across models:

```
kable(data.frame(c(round(diag(vcovHC(model1))[2:6],6),"",""),
                 round(diag(vcovHC(model2))[c(2:6,7,8)],6),
                 round(diag(vcovHC(model3))[c(2:4,6:8,10)],6)),
      format = "latex",
      col.names = c("Model 1", "Model 2", "Model 3"))
```

	Model 1	Model 2	Model 3
prbarr100	9e-06	0.000015	0.000026
log(density100)	0.002827	0.003273	0.005791
pctymle100	0.000122	0.000073	0.000108
central	0.009577	0.005510	0.006154
west	0.006015	0.005616	0.015162
taxpc100		0.000000	0.000000
prbconv100		0.000001	0.000002

Model 1 "wins" for `prbarr100` and `log(density100)`, while Model 2 wins for `pctymle100`, `central` and `west`, while also beating Model 3 (albeit by slim margins) for `prbconv100`. We noted at the beginning of Model 3 that adding variables decreases the precision of coefficient estimates, and none of its coefficient variances were lowest when compared to Models 1 and 2.

Conclusion

When compared side-by-side, each model has relative strengths and weaknesses. Model 1 has strong significance in its coefficients, but poor overall model metrics. Model 3 has a strong adjusted R^2, but the quantity of coefficients reduced the accuracy of its coefficients. Model 2 is strong on both fronts and, all factors considered, our strongest model.

5. Omitted Variables

We identified many potential omitted variables in our data and, consequently, our models. The table below summarizes the most important variables we identified as missing in this analysis (in our model-building process, we identified many more, but we wanted to highlight the most critical variables here). We also estimated the sizes of biases and the directions in which they would vary with crime rate (with respect to zero) and the other independent variables.

Omitted Variable	Bias with y	Bias with x's
Unemployment	Positive/large. Lack of employment may push more individuals to such crimes as petty theft.	Positive/large
Education level/ School quality	Negative/large. We believe education level positively affects income, while school quality keeps younger people engaged.	Negative/moderate
Poverty level	Positive/large. The pressure of being poor for long periods of time could lead individuals to commit crimes in search of money.	Positive/large
Income/cost of living ratio	Negative/large. A paycheck that fully covers one's needs can help avoid crime-inducing circumstances.	Negative/small
Availability/quality of social services	Negative/moderate. Strong social safety nets can help individuals weather periods of unemployment or potentially crime-inducing life circumstances.	Could be positive or negative
Cultural factors	Could be positive or negative, with varying size.	
Religious attendance	Negative/large. Belonging to a church or similar civic institution confers levels of support to individuals and strengthens their bonds with their neighbors.	Could be positive or negative, with varying size.
Climate	Positive (with temperature)/moderate. Warm weather months typically show an increase in crime. The same may be true for warm weather counties.	None
Family support	Negative/large	Negative/large
Police corruption	Positive/moderate	Could be positive or negative, with varying size
Criminal justice policies	Could be positive or negative, with varying size	Could be positive or negative, with varying size
Citizens' attitudes towards crime	Could be positive or negative, with varying size	Could be positive or negative, with varying size

Further, as noted in Model 2, we also suspect that the tax per capita variable contained omitted variable bias.

We could not accurately estimate the collective effect of the omitted variables above on the outcome variable. We determined that there were more variables that were likely to bias in the positive direction (away from zero) than negative with the dependent variable. However, most of these would also vary in the positive direction with many of the independent variables. Our best guess is that there was an overall moderate positive bias from the identified omitted variables.

We identified no good potential proxies for the omitted variables in the dataset. The income elements of poverty level and income/cost-of-living ratio are captured to some degree in the wage data, but the expense portions are not. Police corruption and criminal justice policies are buried along with several other omitted variables within probability of arrest and probability of conviction, but we did not believe that these represented the principal drivers in either variable.

Some of our data fields were proxies for multiple omitted variables. We discussed the inherent issues with

probability of arrest earlier in our analysis. Probability of conviction contained similar problems. Although conviction data was sourced from the North Carolina Department of Correction, factors such as bribery, corruption, overpopulation in jails, and funding impact conviction rates and represent omitted variable bias.

6. Policy Suggestions and Conclusions

When creating our models, we also considered what factors could be influenced by the governor. It is one thing to understand the determinants of crime, but for this project, we were specifically asked to identify determinants of crime and make related policy recommendations. We could not make any firm causal claims based on the information available, but we observed substantial enough relationships to reach associative conclusions and formulate recommendations for the campaign.

The variable in our models that carried the greatest weight was population density, which had a positive relationship with crime rates. Population density is not something the candidate can control; it would not be practical to say, "I will relocate people from urban areas into rural ones to reduce crime." However, it does help guide us in determining where to focus our recommendations and allocate resources geographically.

The young male percentage of the population is similarly something that policy cannot affect reasonably. When including percent young males in the models, though, we did so with the understanding that further analysis may indicate an issue with the quality of schools, lack of attendance, economic conditions, and unemployment, all of which could be included in a governor's platform.

Policy changes are most effective in other areas, particularly those that deal with the uses of government funding. Based on our best model, we recommend that the candidate stress policy changes that would affect two of these areas: probability of arrest and probability of conviction. Our model suggested a negative relationship between crime rates and probability of arrest, which we were using as a proxy variable for police effectiveness. Probability of conviction also demonstrated a negative relationship with crime rates, indicating a connection with the certainty of punishment.

Policy Recommendations

We recommend that the candidate focus on the following policy changes:

- Increased resources for law enforcement, in the form of additional staff and training/development for existing staff. The intent would be to increase the probability of arrest and the quality of the work going into each arrest, which would improve the probability of conviction.
- Increased resources for the judiciary, including incentives to recruit quality prosecutors. Success in this area would lead to higher probability of conviction.
- Renewed focus on community relations, in order to increase the public's quantity and quality of interactions with the police. This could increase their willingness to provide quality information to police and, if necessary, testify. We hypothesize that community involvement would increase both probability of arrest and conviction.
- Increase availability of victim counseling services, so that, after reporting a crime, the public would not be intimidated by the judicial process. We expect this would increase probability of conviction.
- Due to the positive correlation between population density and crime rates in our model, concentrate these efforts first in densely populated areas.
- Next, focus on the counties in the eastern region, which exhibited higher average crime rates than the central and western regions.

Note: Prior to implementation, community relations and victim counseling would require follow-up studies to ensure that we had a firm understanding of the current extent and effectiveness of these programs.

As discussed, we considered the effects of per capita tax rates to be inconclusive, potentially containing some omitted variable bias, in addition to being statistically insignificant. We would not recommend the candidate base any tax policy assertions on the slight positive relationship, such as "lowering taxes reduces crime." Understanding tax effects further would require additional data and research.

The Omitted Variables section above is an overview of data we considered critical for conducting more in-depth analyses and providing finely tuned policy recommendations. Without understanding unemployment, household income/poverty levels, types of crimes, and corruption levels, we kept our recommendations broad.