W203 Statistics Lab 3

The Significant Effects (Haerang Lee, Soravit Sophastienphong, Alex Heaton)

December 11, 2019

Introduction

The objective of this lab is to gather meaningful, actionable insights about the determinants of crime in North Carolina to generate data-driven policy suggestions applicable to local government. Using a cross-section of data from a study conducted in 1994 by Cornwell and Trumball, researchers of Georgia and West Virginia University, we would like to explore **how the strictness of law enforcement affects criminal activity in different regions of North Carolina.** We hypothesize that stricter enforcement leads to reduced crime.

Refer to the data dictionary attached to the lab 3 instructions for variable names, descriptions, and more information about the dataset.

Data Cleaning

The North Carolina crime data for 1987 has 27 columns and 90 records, after we remove 6 records with completely missing data and a duplicate record for county 193. We also converted the data type of the field `prbconv`, representing the probability of conviction per arrest, from factor to numeric.

Data Explorations and Transformations

We first explore the data to identify anomalous values and coding features, starting with the variables that are not indicator variables and are unrelated to wages. We plotted histograms for these variables in Figure 1. The histograms of these variables do not show any spikes near the tails, which is common to top-coded variables. A few variables had outliers at the right end of their graphs: the probability of arrest (`prbarr`), the probability of conviction (`prbconv`), police per capita (`polpc`), tax percapita (`taxpc`), percentage of young males (`pctymle`), and population density (`density`).

`pbarr` and `pbconv` include values greater than 1 (as shown in Figure 1), although these represent probabilities which theoretically should be less than or equal to 1. The data dictionary explains that the numerator and denominator for these variables may come from different sources or periods, which may result in the ratios greater than 1. So we will treat these variables as valid estimators of the probabilities they represent without further modifications.

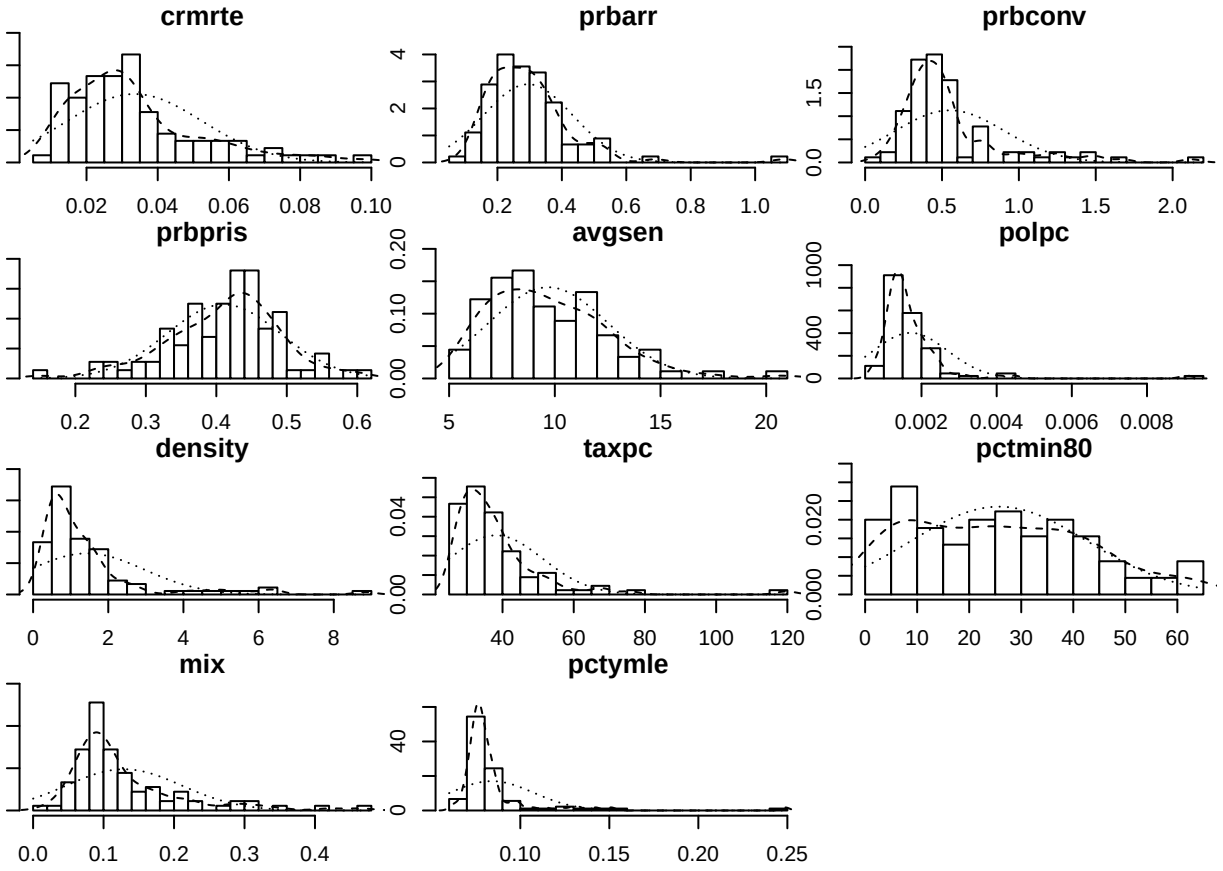


Figure 1: Histograms for variables other than `year`, `county`, wages, and indicators

Descriptive Statistics of `polpc`:

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012378 0.0014897 0.0017080 0.0018856 0.0090543
```

There are five counties with `polpc` lower than 0.001, indicating that the minimum value of 0.0007 is not an outlier. Given that these counties are not urban and do not have particularly high crime rates, it makes sense that there might be a lower number of police per capita.

County 115 has the maximum `polpc`. It also has relatively high `prbarr`, `prbconv`, `prbarr`, and `avgsen` values and the lowest crime rate in the data. While its pattern differs from that of the other counties, it may be an example of a county having a low crime rate as a result of strict law enforcement. Its `prbconv` and `prbarr` values over 1, as demonstrated before, are not signs of errors. Hence, despite notable differences from the rest of the counties, we do not believe that county 115 is erroneous in any way and will keep it in the dataset as-is.

Descriptive Statistics of `taxpc`:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.73   34.92   38.16   41.01  119.76
```

The range for tax revenue per capita is quite broad and exhibits a right skew. Although the data dictionary did not provide the units and we could not find any information in the original data source in R's Ecdat package, it appears to be in thousands of dollars. We examined the counties with large `taxpc` values and identified only two counties with `taxpc` greater than 75, one being the max of 119.76 for county 55. While this is a relatively large value, we do not have a reason to believe that it is erroneous, misleading, or otherwise necessary to remove. If the county is very wealthy, then it can plausibly have a higher tax revenue per capita.

Descriptive Statistics of `pctymle`:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06216 0.07437 0.07770 0.08403 0.08352 0.24871
```

Similarly, `pctymle` has a heavy right skew. There are two counties with `pctymle` greater than 0.15. Again, these outliers do not warrant removing. Young men may be attracted to the counties that are home to job opportunities that fit their goals ane needs, which may lead to a higher concentration of young males in a county that has a concentration of particular industries or employers.

Descriptive Statistics of `density`–Original Value:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.54718 0.97925 1.43567 1.56926 8.82765
```

Density values range from 0 to 8 in the dataset. According to the codebook, this is the number of people per sq. mile. However, that would mean that, at a maximum, a county has up to 8 people per square mile, whereas the density of the entire NC state in 1987 is about 213 people per square mile (calculated from data in the US Census: 10 million people over 4,800 sq. miles). The original `crime` in the Ecdat package of R–which our data was derived from–has similar numbers, which appears unrealistic. We found the 1990 county-level data from the University of Minnesota (Schroeder, Jonathan P, 2016, https://conservancy.umn.edu/handle/11299/181605), which showed the population density per county ranging from 9 to 976 people per sq. mile. Although we cannot get an accurate figure of what the density in our dataset is supposed to be, based on the Census data and the University of Minnesota data, we determined that the existing `density` values represent the number of people per 100 sq miles. We will multiply the density numbers by 100 for it to represent the number of people per sq. mile.

Descriptive Statistics of `density`–Updated Value:

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

The updated median and maximum values are more aligned with other public data sources. However, the minimum value is 0.002 people per square mile, which seems nonsensical. According to the US Census, the smallest county in NC is Clay County at 221 sq mi, which, at the minimum density of our data set, would have only 0.442 people living there in 1987. Even the largest county, Dare County, at 1,562 sq mi would have just 3.1 people living there, which is highly unlikely. Hence, this observation will be considered missing. We do not have a reason to believe that deleting this data point would create a bias for the data, but we want to preserve all other data points from the same county. Therefore, we will impute the mean rather than drop county 173 entirely from the dataset.

Descriptive Statistics of `avgsen`:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.380   7.375   9.110   9.689  11.465  20.700
```

The maximum average sentence is only 20.7 days and the mean is ~9.69 days. Intuitively, we would think that those with long prison sentences of multiple months or years would significantly impact the average, but the data indicates that a more significant number of convicted criminals receive short sentences, decreasing the average.

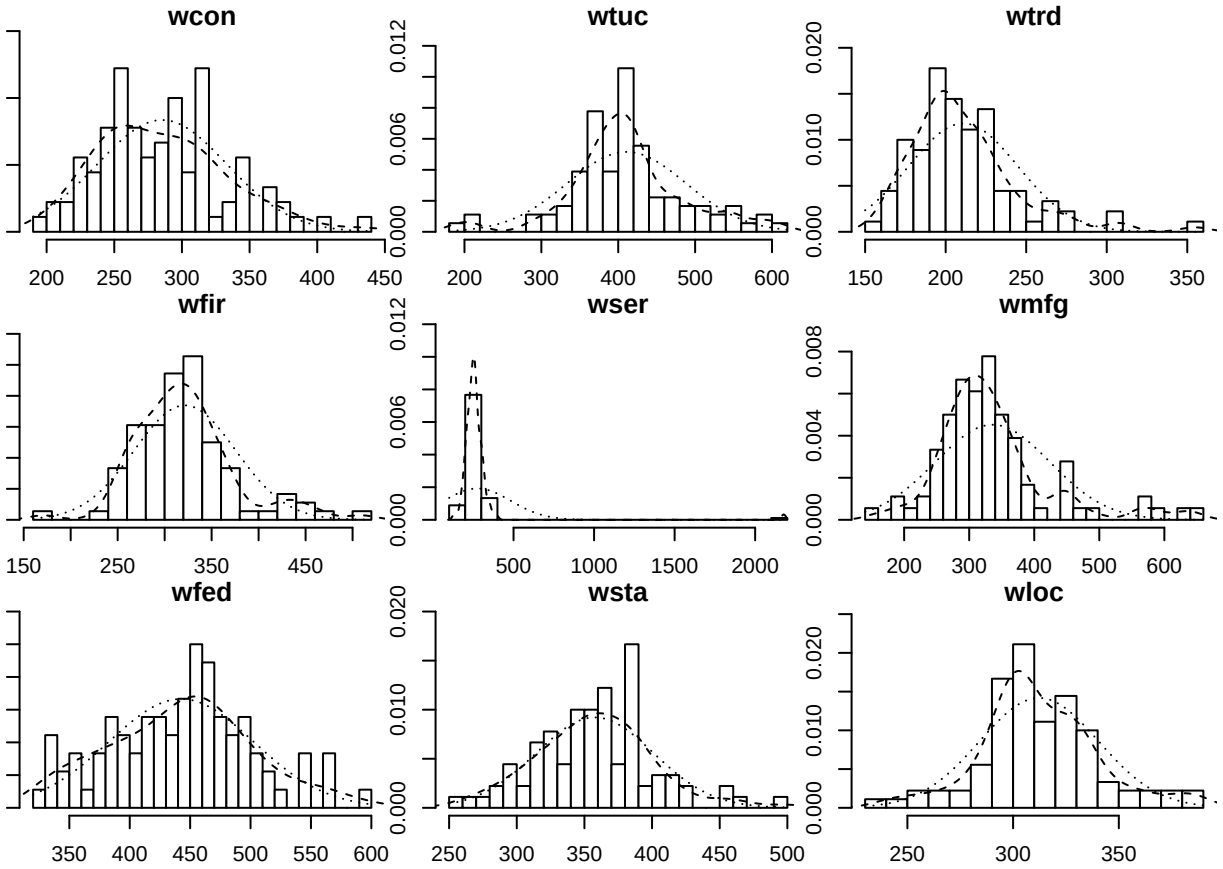


Figure 2: Histograms for wage variables

Last, we analyze a series of wage variables. Figure 2 contains the histograms of nine variables representing weekly wages from various industries. None of the wage histograms shows a spike at the tail end, indicating a low risk of top coding. We identified some variables with a maximum value that look close to a rounded value, such as the maximum value of `wfed` right under 500 and `wsta` right under 600. However, each variable only has one observation of these almost-round values, so we determine that they are not top-coded.

Some wage histograms spike in the middle of the curve. For instance, `wcon`, `wsta`, and `wfed` show very tall and very short histogram bars right next to each other in the center. However, we did not find any signs that these spikes were a result of coding or anomalous values.

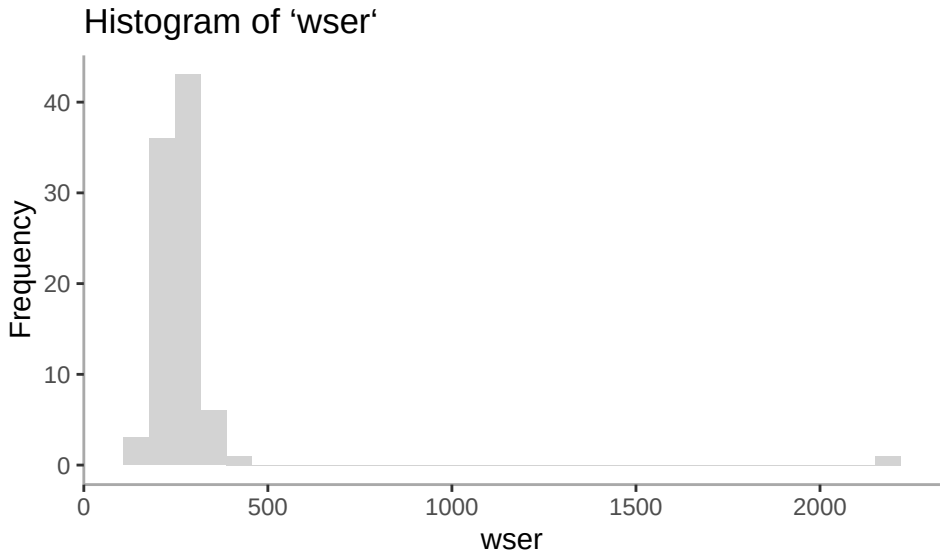


Figure 3: Histogram of Wages for Service Industry

Figure 3 shows that `wser` has an extreme value. We examined this further.

Descriptive Statistics of `wser`

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   275.3   277.6  2177.1
```

For comparison, every county had a `wser` of $391 or below, except county 185, which had a `wser` of $2,177.

Top 3 values of `wser`

```
##     county      wser
## 185    185 2177.0681
## 63      63  391.3081
## 119    119  354.3007
```

The maximum values of all other wage variables across all counties were under $700/week.

Maximum values of all wage variables

```
##      wcon      wtuc      wtrd      wfir      wser      wmfg      wfed      wsta
##  436.7666  613.2261  354.6761  509.4655 2177.0681  646.8500  597.9500  499.5900
##      wloc
##  388.0900
```

Given the context, $2,177 for any wage variable seems unrealistic. Since this maximum value is about ten times the size of the `wser` median, it is a possible transcription error related to placing the decimal point in the wrong place. We will treat this as a missing value and impute the mean.

Choosing an Outcome Variable: Log of Crime Rate

We chose the crimes committed per person (`crmrte`, also called the crime rate) as the outcome variable of our analysis because we would like to provide insight into policy changes to reduce crime. Since the counties vary in area and population size, examining offenses committed per person, rather than the total number of incidents, allows us to make a logical comparison of crime levels among counties.

We use *log*(crmrte), because a log transformation helped us meet the Gauss-Markov assumptions, which are preconditions for making our Ordinary Least Squares estimators BLUE (Best Linear Unbiased Estimators) and for hypothesis testing using statistical methods such as the t-test.

Choosing Explanatory Variables

Before choosing the explanatory variables, we explored the relationship between `crmrte` and all variables in the dataset to gain a better understanding of a potential influence that the variables may have on our research question. An analysis of correlation, shown in Figure 4, is one such exploratory method.

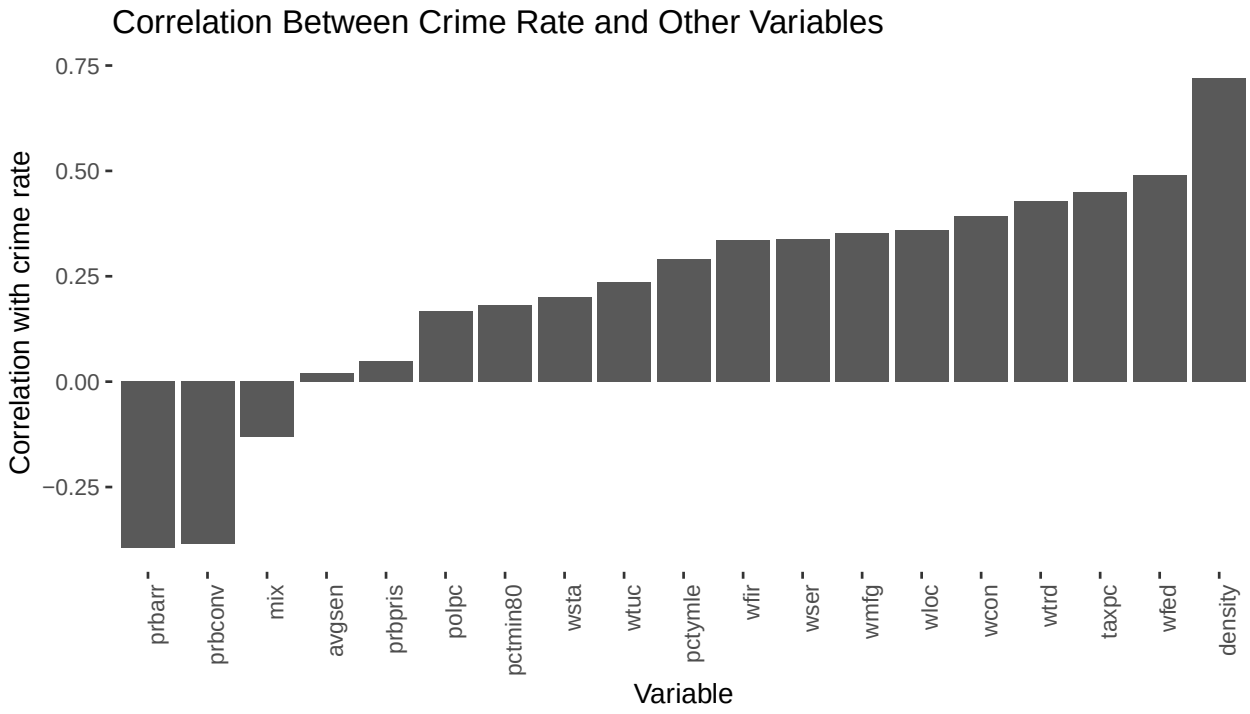


Figure 4: Correlation between crime rate and explanatory variables

The strictness of law, the potentially salient driver of crime reduction, could be operationalized through variables such as the probability of arrest and the probability of conviction (`prbarr`, `prbconv`). The average length of sentences (`avgsen`) could represent this conceptual variable, too, but in the chart above, it has a relatively low correlation with the crime rate. Based on relevance and correlation, the following are the variables that warrant further analysis.

Density

`density` has the strongest positive correlation with `crmrte`. According to the National Bureau of Economic Research (NBER: https://www.nber.org/papers/w5430), dense, urban areas are generally known to have higher crime rates, which may explain the patterns in the data. However, population density cannot be changed by policy, so we will treat this as a relevant covariate, but not an explanatory variable in our model.

Probability of Arrest and Probability of Conviction

`prbarr` and `prbconv` have the strongest negative correlation with `crmrte`, which seems to support the common sense that stricter criminal justice policy would deter citizens from criminal activities. Since `pbrarr` is the probability of arrest given offense and `prbconv` is the probability of conviction given arrest, we decided that multiplying the two would give a better measure of the strictness of criminal justice: the probability of conviction given offense. We will call this new variable `pbarrandconv` and include it as a key variable in our analysis.

$$\text{pbarrandconv} = \text{pbarr} * \text{pbconv} = \frac{\text{Arrests}}{\text{Offenses}} * \frac{\text{Convictions}}{\text{Arrests}} = \frac{\text{Confictions}}{\text{Offenses}}$$

Log of Police Per Capita

Even though police per capita (`polpc`) does not have as high of a correlation with crime rate as other variables, we wanted to include it in our model because it can be controlled via policy. For example, the police can station more officers on standby in specific neighborhoods, which may increase the perceived level of law enforcement and deter criminal activities. Therefore, we include it as an explanatory variable in our model. We use the log of police per capita (log(polpc)) to transform the data to satisfy the Gauss-Markov assumptions for linear regressions.

Linear Regressions

Model 1: Base Model

We wanted to explore how well we could predict crime rates using the probability of conviction per offense (`pbarrandconv`) and the police per capita (`polpc`). Depending on the outcome of this model, we could suggest the creation of policies to modify the perceived strictness of the law, by changing the number of arrests made or the level of police presence. Our OLS model yielded the following results, as detailed in Table 1.

$$\begin{aligned}\log(\text{crmrte}) &= \hat{\beta}_0 + \hat{\beta}_1 \log(\text{polpc}) + \hat{\beta}_2 \log(\text{pbarrandconv}) \\ &= -2.450 + 0.338 \log(\text{polpc}) - 0.531 \log(\text{pbarrandconv})\end{aligned}$$

First and foremost, the adjusted R-squared value of 0.424 indicates that our model explains less than 50% of log(`crmrte`). While this level of goodness of fit does not invalidate or bias our model, we will consider how it may be improved in the future iterations of the model. Second, we identified heteroskedasticity as outlined in the CLM assumption analysis section below. Therefore, we used heteroskedasticity-robust errors in Table 1.

Table 1: Model 1

	log(Crime Rate)
log(Probability of Conviction per Offense)	-0.531^{***}
	(0.177)
log(Police Per Capita)	0.338
	(0.295)
Constant	-2.450
	(2.198)
N	90
R^2	0.437
Adjusted R^2	0.424
Residual Std. Error	0.417 (df = 87)
F Statistic	33.731^{***} (df = 2; 87)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Probability of Conviction given Offense

Our model indicates that a 1% increase in the Probability of Conviction per offense corresponds to an approximately 0.531% decrease in crime rate. This coefficient is statistically significant to the 0.1% level and relatively practically significant. For instance, a 10%-increase (not percentage point) in the number of convictions per offense–from 15.99% to 17.59%–may decrease the crime rate from 3.35% to 3.19%, which is a 4.93% change.

Police per Capita

The coefficient is not statistically significant using heteroskedasticity-robust errors. And its value greater than 0 indicates that for every 1% increase in police per capita, there is a 0.34% *increase* in crime rate. While this is not the result that we were expecting, there is some intuition to this. A higher police per capita could both be the cause of reduced crime and the result of increased crime. For instance, the government may wish to station more law enforcement in counties with high crime rates that are likely to need additional support.

In order to make a more definitive statement about how the police per capita may influence crime rate *ceteris paribus*, we would require historical data on both crime rate and police per capita. Then we may calculate how the *change* in police per capita in each county may be related to the *change* of crime rate in the following period. Unfortunately, we only have information from a single year–1987. We could also argue that these missing variables (historical information on crime and police per capita) could be creating an omitted variable bias.

CLM Assumptions

Model 1: MLR.1 Linear in Parameters

We have built a model that is linear in parameters (coefficients), based how the assumption is defined as found in Wooldridge p.74:

> The model in the population can be written as
>
> $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u,$$
>
> where β_j values are unknown parameters constants) of interest and u is an unobserved random error or disturbance term.

Since the error term `u` in the equation can take on any value, it absorbs any nonlinearity in our model. Hence, this assumption is satisfied.

Model 1: MLR.2 Random Sampling

The observations in the NC crime data are *not* a random sample, but a selection of almost all counties in one state in one year. Naturally, the observations are not a representative sample of the U.S., or even North Carolina, as different factors may affect the crime rate in each year, such as the rise of a particular criminal organization or new regulations in a certain year. Since we only have the data for 1987, we should be cautious not to make hasty causal claims in the absence of the information about other years in these counties.

Additionally, the observations are *not* independent of each other and their attributes are not identically distributed. For instance, the criminals or police from one county could be active in a neighboring county, resulting in dependent crime rates, arrest rates, or other measures among neighboring counties.

Although we cannot go back randomly select the samples again, we can instead try to measure and control for the lack of independence. The data does not provide the exact location of the counties or the distances among them, but the indicator variables `central` and `west` can be used to gauge the proximity of the counties. Putting these into the next variation of the model would help us mitigate the issue of non-random sampling. We checked whether every county fell under one of the three regions: Western, Central, and Eastern.

```
##
##               Not Western Western
##   Central              33       1
##   Not Central          35      21
```

The 35 counties that are neither Central nor Western can be considered Eastern. County 71 is marked as both Western *and* Central, but this data may not necessarily be erroneous, as the county may be located right in the middle of the two regions and considered either Western or Central by the residents. So we will use this data as-is.

A quick analysis shows that Central NC has the most number of urban counties, even though it comprises overall the smallest number of counties. Eastern has the most number of rural counties. To counter the effects of the mentioned omitted biases of location, we can consider adding `west` and `central` to our next model.

Model 1: MLR.3 No Perfect Collinearity

This assumption states that there is no exact linear relationship between any of our independent variables. To test this, we generate a scatterplot with `polpc` and `pbarrandconv` in Figure 5. The data points do not fit perfectly on the regression line and the two variables are not perfectly collinear.

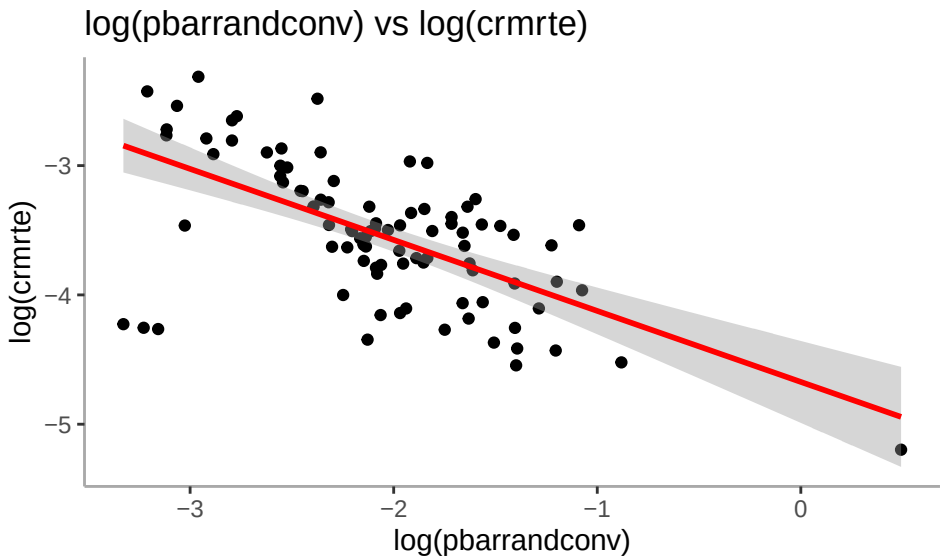


Figure 5: Model 1: Scatter Plot of polpc and pbarrandconv

Second, we use variance inflation factors to detect multicollinearity in our model. VIFs are calculated by running regressions of each predictor on every other predictor and measure how the variance of coefficients is inflated as a result of linearity with other predictors. There is no set rule on what constitutes a concerning VIF, although 10 is a commonly used threshold by researchers. Nonetheless, our values do not raise any concern, as they are near 1 and signify that the variance of the coefficients in our model are barely inflated. Additionally, we have an intuitive sense that police per capita should not be *perfectly* correlated with convictions per offense. Given these reasons, the two predictors are not perfectly correlated, and our assumption is satisfied.

```
## log(pbarrandconv)        log(polpc)
##          1.008134          1.008134
```

Model 1: MLR.4 Zero Conditional Mean

Looking at the plot of residual vs. fitted visuals in Figure 6, the LOWESS (locally weighted scatterplot smoothing) line in red is sloping upward instead of lying flat, so the zero conditional mean assumption is violated. The steady upward trend of $E(u|x)$ indicates that we may have an omitted variable bias, such as the crime rate of the prior years.

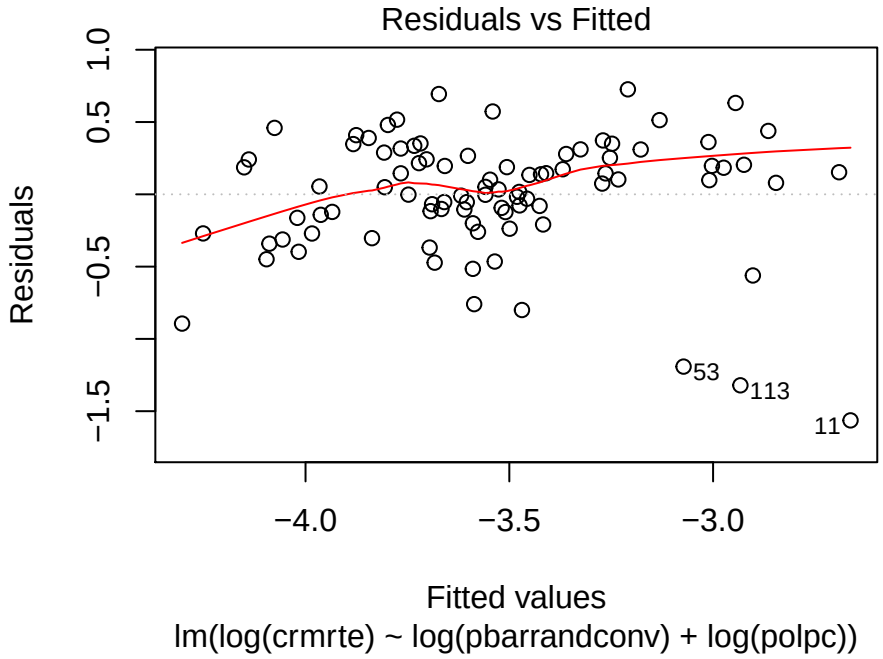


Figure 6: Model 1: Residuals vs fitted values

To examine the impact of omitting the past year's crime rate, assume that the "true" model of crime rate is as follows:

$$\text{crmrte} = \beta_0 + \beta_1 \text{ polpc} + \beta_2 \text{ past crmrte} + \beta_3 \text{ pbarrandconv} + u$$

Our model currently cannot estimate β_2, the coefficient of past crime rates, but we can assume that the crime rates from year to year are positively correlated.

$$\beta_2 > 0$$

The past crime rate can be predicted using the current year `polpc`.

$$\text{past crmrte} = \gamma_0 + \gamma_1 \text{ polpc} + v$$

Ceteris paribus, governments would place more police in counties that have higher crime rates. Therefore, these two variables are positively correlated.

$$\gamma_1 > 0$$

Without the data on the past crime rates, our coefficient for current year `polpc` may be biased by the amount of $\beta_2 \cdot \gamma_0$, as shown below.

$$
\begin{aligned}
\text{crmrte} &= \beta_0 + \beta_1 \text{ polpc} + \beta_2 \text{ pastcrmrte} + \beta_3 \text{ pbarrandconv} + u \\
&= \beta_0 + \beta_1 \text{ polpc} + \beta_2 \ (\gamma_0 + \gamma_1 \text{ polpc} + v) + \beta_3 \text{ pbarrandconv} + u \\
&= \beta_0 + \beta_1 \text{ polpc} + (\beta_2 \ \cdot \gamma_0 + \beta_2 \ \cdot \gamma_1 \text{ polpc} + \beta_2 \ \cdot v) + \beta_3 \text{ pbarrandconv} + u \\
&= \beta_0 + (\beta_1 \ + \beta_2 \ \cdot \gamma_1) \text{ polpc} + \beta_3 \text{ pbarrandconv} + (u + \beta_2 \ \cdot \gamma_0 + \beta_2 \ \cdot v)
\end{aligned}
$$

Since $\gamma_1 \ > 0$ and $\beta_2 \ > 0$, the omitted variable bias of the past crime rate is positive. Depending on the magnitude of β_1 and the bias, the bias may be toward (if $\beta_1 < \beta_2 \ \cdot \gamma_0$) or away from zero (if $\beta_1 < \beta_2 \ \cdot \gamma_0$). Therefore, the coefficient we found for `polpc` would decrease if we add the past crime rate, possibly to a number below 0. Unfortunately, we cannot determine this, and `polpc` is unhelpful in predicting the level of crime for policy recommendation purposes, so in the next iteration, we will exclude this variable.

Model 1: MLR.5 Homoskedasticity

The variance of the error u should be the same given any value of the explanatory variables. Looking at the scale-location plot in Figure 7, the model shows signs of heteroscedasticity. In particular, the red line shows a distinct U shape. Without meeting this assumption, we cannot say that our OLS regression is BLUE–Best Linear Unbiased Estimators.

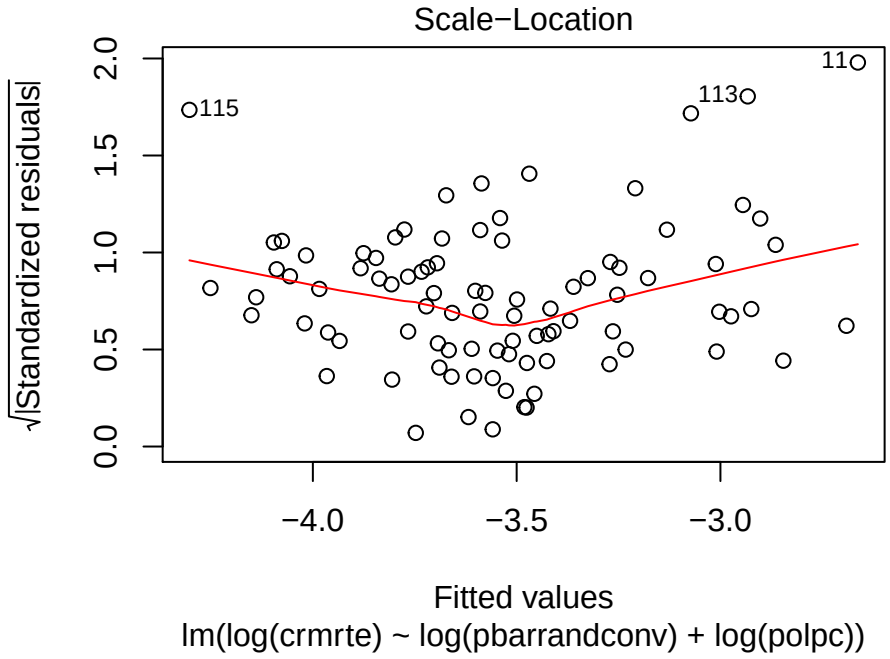


Figure 7: Model 1: Scale-location plot

We can use the studentized Breusch-Pagan test to check the homoskedasticity further. The Breusch-Pagan test has a null hypothesis that there is homoskedasticity.

```
bptest(base_model_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  base_model_1
## BP = 8.8767, df = 2, p-value = 0.01182
```

The p-value is statistically significant at the 5% level. So we reject the hypothesis that we have homoskedasticity and should, therefore, use heteroscedasticity-robust errors to evaluate our model.

Model 1: MLR.6 Normal Distribution of Errors

A normal distribution of errors implies that the OLS coefficients ($\hat{\beta}_j$) will be normally distributed, and therefore we can conduct a hypothesis testing of the coefficients.

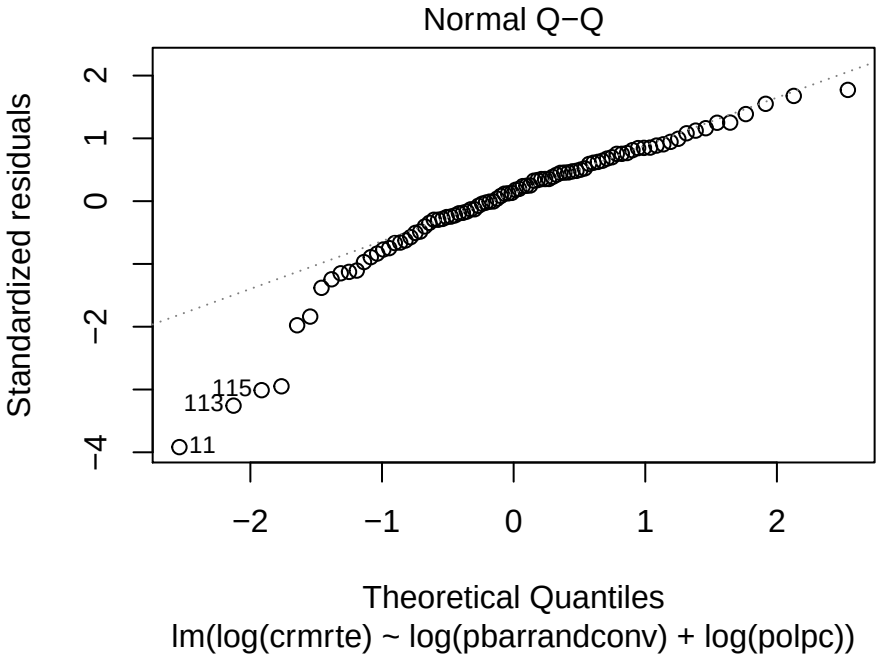


Figure 8: Model 1: QQ plot to check normal distribution of errors

In Figure 7, the Q-Q plot indicates a left-skewed distribution of residuals. Indeed, the histogram of residuals matches this assessment.

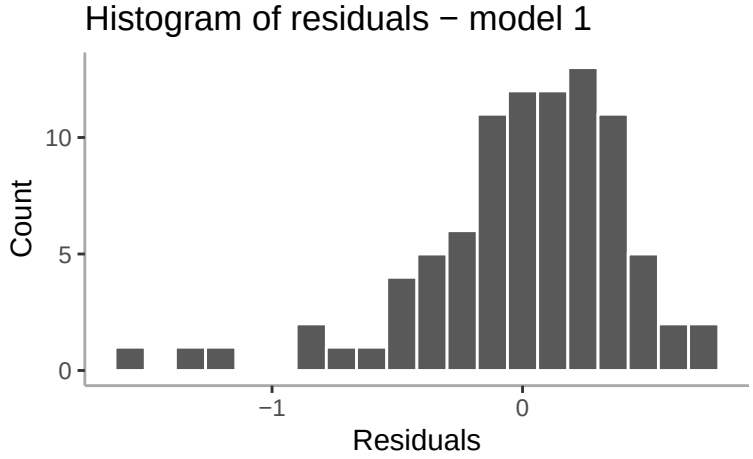


Figure 9: Model 1: Histogram of residuals

The histogram of residuals indicates that while a lot of our predicted values are not too far off from 0, there are a few outliers to the left where we are underpredicting and may impact our interpretation of the p-values.

Model 1 Interpretation

For the probability of conviction given offense, although the coefficient was statistically significant and we assessed its practical significance to be somewhat important, we cannot take this significance at a face-value due to two reasons. First, we evaluated the Gauss-Markov assumptions and found violations of random sampling, zero-conditional mean, and homoskedasticity. Therefore, the t-test that we rely on to test the hypothesis concerning our OLS coefficient may be unreliable.

Second, we will exclude `polpc` in the next model, and consequently change the regression line and the statistical significance of the coefficient of `pbarrandconv`. The reason for this decision is that in order to evaluate `polpc`, we require additional variables that are currently unavailable. The omitted variable biases the coefficient of `polpc` in an unpredictable manner. We determined that the research would yield more insightful results if we replaced `polpc` with other variables.

Model 2: Key explanatory variables and only covariates that may increase accuracy

In this iteration of the model, we replaced `polpc` with a new set of additional variables that could explain the crime rate.

Log of Density

A potential source of omitted variable bias from our base model was density. Areas with higher population density lend themselves to more crime, such as major cities. For example, one pickpocket could snatch a lot more valuables in a busy street than in sparse areas. Since we would like to understand how the percent change in density affects the percent change in the crime rate, we use the log of density in the model.

Average Sentence

The average sentence also reflects the severity of punishment. We already know this to be less correlated with the crime rates, so it is not likely to be a key driver, but helpful to have in the model nonetheless.

Percent Minority

There is a well-known gang called MS-13 that was active in North Carolina in the early 2000s, and they were a dangerous international gang, comprising of El Salvadorans. Although this particular gang's well-known activities in NC are much later than at the time of the data collection, we are curious to see if the percentage of minorities may lead to higher crimes.

Percent Young Males

Gangs tend to recruit young men, so we are curious if a higher number of young men present in an area increases the crime rate.

Log of Tax Revenue per Capita

Tax is a proxy for wealth in each county. Poorer counties may be prone to more crime, in which case, the policies could be developed to target and reduce poverty in such communities. We use the log of tax revenue per capita, so that we can understand the elasticity of crime rate to percent changes in tax revenue.

Ratio of Face-to-Face Crimes to Other Types of Crimes

Face-to-face crime serves as a proxy for violence, which can be associated with more severe crimes. It is interesting to consider whether more serious crimes impact lower or higher crime rates. Perhaps areas with high crime rates tend to have more crimes that are less serious.

West and Central

Given our violation of random sampling in our base model, we add these location variables in order to counteract the lack of independence caused by neighboring counties. Adding in these indicator variables allows us to use a single regression equation that represents separate and distinct groups with possibly different patterns.

Model 2 Generation and Comparison to Model 1

The Adjusted R2 increased from 0.424 to 0.740. Unlike R2, the Adjusted R2 penalizes the addition of new variables, so the increase in Adjusted R2 indicates that the second model is better at predicting the values of the dependent variable, even if we account for the number of additional independent variables. In Table 2 below, we used heteroskedasticity-robust errors.

Table 2: Model 2

	log(Crime Rate)	
	Base Model	Second Model
	(1)	(2)
log(Convictions per Offense)	-0.531^{***} (0.177)	-0.341^{***} (0.094)
log(Police per Capita)	0.338 (0.295)	
Average Sentence		-0.003 (0.013)
Percent Minority		0.008^{**} (0.004)
Percent Young Male		0.578 (1.137)
Face-to-Face / Other		-0.516 (0.582)
log(Tax Revenue per Capita)		0.179 (0.244)
log(Density)		0.390^{***} (0.084)
West		-0.260^{*} (0.153)
Central		-0.171^{*} (0.103)
Constant	-2.450 (2.198)	-6.740^{***} (0.683)
N	90	90
R^2	0.437	0.766
Adjusted R^2	0.424	0.740
Residual Std. Error	0.417 (df = 87)	0.280 (df = 80)
F Statistic	33.731^{***} (df = 2; 87)	29.123^{***} (df = 9; 80)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

The following is the equation for our model.

$$\begin{aligned}\log(\text{crmrte}) = &-6.740 - 0.341\ \log(\text{pbarrandconv}) \\ &+ 0.008\ \text{pctmin80} + 0.578\ \text{pctymle} + 0.179\ \log(\text{taxpc}) \\ &- 0.516\ \text{mix} - 0.003\ \text{avgsen} + 0.390\ \log(\text{density}) \\ &- 0.260\ \text{west} - 0.171\ \text{central}\end{aligned}$$

We will interpret our results after evaluating the CLM assumptions.

Model 2: MLR.1 Linear in Parameters

This assumption is satisfied.

Model 2: MLR.2 Random Sampling

This assumption is not fully satisfied, as mentioned in model 1 analysis, but mitigating factors such as our addition of `west` and `central` allow us to proceed with the OLS regression and estimator interpretations.

Model 2: MLR.3 No Perfect Collinearity

While some of our new covariates are correlated, this assumption only assumes no *perfect* correlation, which is not the case with any of our variables. Judging by intuition, none of our variables measure the same concept, nor can they be derived from each other directly. We also calculated the VIFs, and the highest VIF we have is that of `west`, meaning its variance is 170% inflated by other covariates. We will take note of this, but it is not high to the extent that we should be overly concerned. Hence, our assumption is satisfied.

```
## log(pbarrandconv)           avgsen         pctmin80          pctymle
##          1.571694         1.131494         2.405810         1.279256
##               mix       log(taxpc)   log(density)             west
##          1.136673         1.229475         1.630497         2.739049
##           central
##          1.817052
```

Model 2: MLR.4 Zero Conditional Mean

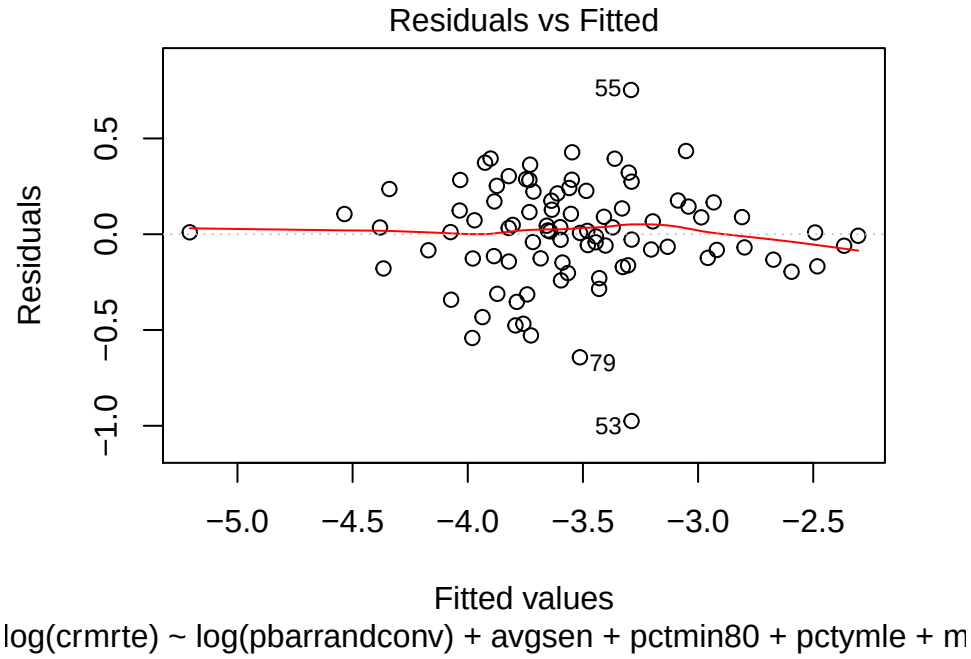


Figure 10: Model 2: Residuals vs. fitted values

The LOWESS (locally weighted scatterplot smoothing) line in red is mostly flat, and much flatter than model 1. We believe that by addressing some of the omitted variables, we have now satisfied the requirements of this assumption.

Model 2: MLR.5 Homoskedasticity

The variance of the error u should be the same given any value of the explanatory variables.

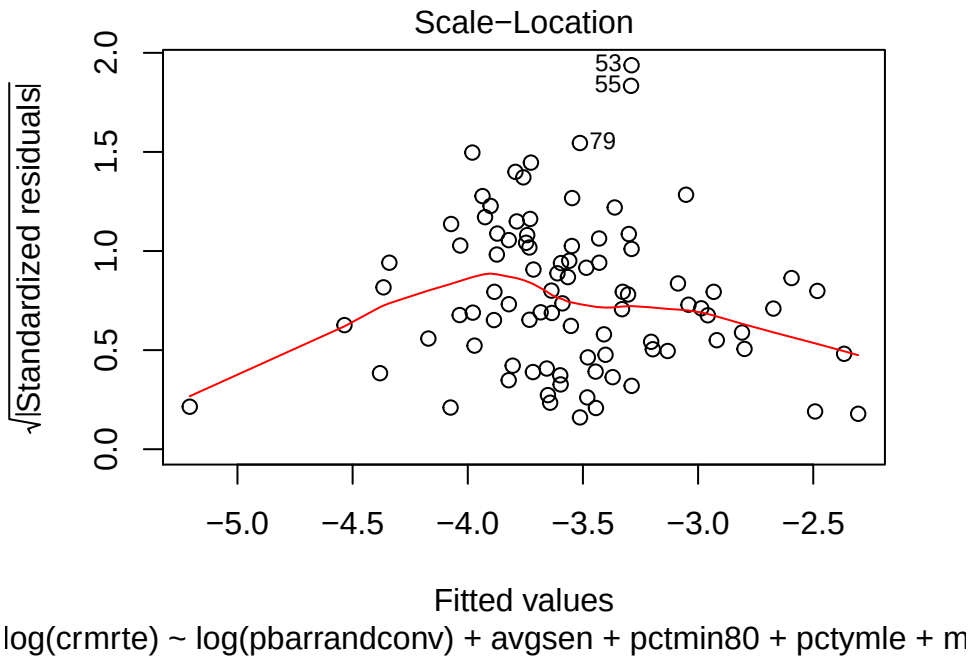


Figure 11: Model 2: Scale-location plot

In the scale-location plot, the red line is still not straight, indicating heteroskedasticity. We can use the studentized Breusch-Pagan test to validate the assumption further.

```
bptest(model_2h)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2h
## BP = 28.229, df = 9, p-value = 0.0008732
```

For this model, the Breusch-Pagan test has a p-value less than 0.01, which is highly statistically significant. Therefore, we reject the null hypothesis that the model has homoskedasticity assume that there is heteroskedasticity. The issues of heteroskedasticity can be mitigated by using heteroskedasticity-robust errors, which we have used.

Model 2: MLR.6 Normal Distribution of Errors

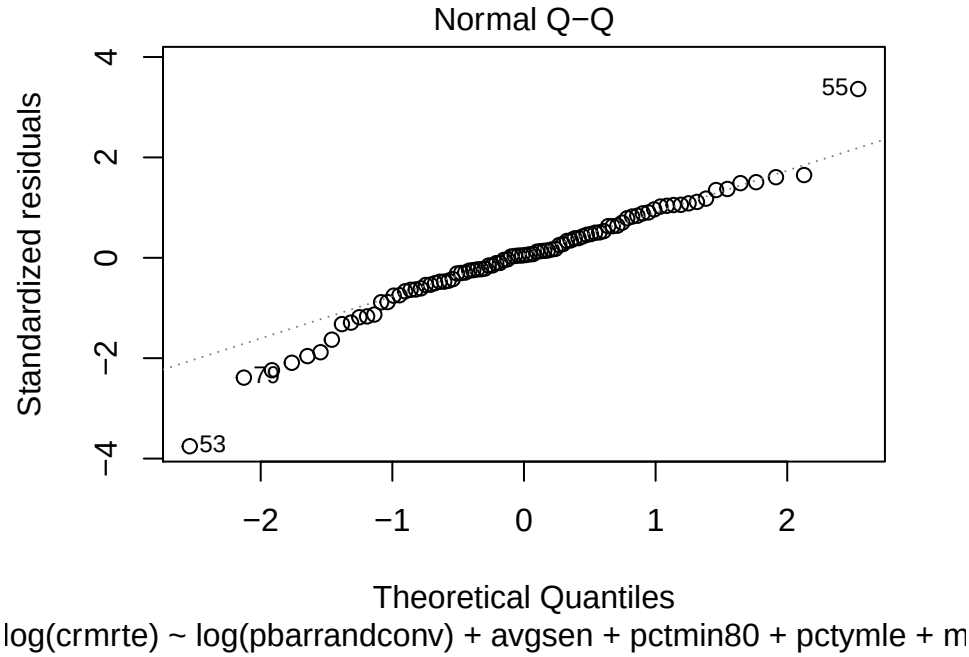


Figure 12: Model 2: QQ plot to check normal distribution of errors

The residuals for the newer model is pretty normally distributed and allows for a relatively accurate interpretation of the t-values.

Model 2 Interpretation

After establishing that the CLM assumptions are satisfied, we interpret the regression coefficients.

Statistical Significance of Each Coefficient

First, we review the results of a t-test on each of the coefficients, using heteroskedasticity-robust errors (vcovHC from the `sandwich` package.) According to Table 2 above, `log(pbarrandconv)`, `pctmin80`, and `log(density)` are statistically significant at the 5% level.

Joint Statistical Significance of Coefficients

However, that does not mean that the remaining variables (`pctymle`, `log(taxpc)`, `mix`, `avgsen`, `west`, and `central`) are jointly irrelevant. The F-test on the remaining variables, which were *individually* statistically insignificant, shows that these variables are *jointly* statistically significant.

```
##
##  F test to compare two variances
##
## data:  model_2h and model2_restricted
## F = 0.53106, num df = 80, denom df = 84, p-value = 0.004746
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3437814 0.8222849
## sample estimates:
## ratio of variances
##          0.5310592
```

The F-test compares the variance of the unrestricted, original model to the restricted model that excludes the group of variables for which we want to test the joint significance. With a p-value of 0.47%, we reject the null hypothesis that the difference in the variances is 0.

Practical Significance

A 1% increase in the probability of conviction given arrest (`pbarrandconv`) decreases the crime rate by approximately 0.341%. To put this in real-world terms, a 10% increase in the probability of conviction given arrest (`pbarrandconv`)–from 15.99 percentage points to 17.59 percentage points–decreases the crime rate from 3.35 percentage points to 3.24 percentage points, a 3.19% change. The severity of the law seems to decrease the crime rate by a trifling amount, especially in light of the large increase in the conviction rate required. Moreover, `avgsen`, another variable that measures the strictness of the law, was not statistically significant and we could not reject the null hypothesis that `avgsen` had no impact on the crime rate.

The percentage of minorities (`pctmin80`) is statistically significant in our model of `crmrte` and has important practical implications. Approximately 0.84% increase in `crmrte` for each 1-percentage point increase in minorities. A 10-percentage point increase in the percentage of minorities would increase a 3.35 percentage points crime rate to 3.64 percentage points, which is a 8.76% increase. Our model predicts that the difference in the crime rates of a White-majority county (10% minorities) and a racial minority-majority county (60% minorities) could be over 52%. In other words, if the White-majority county has a crime rate of 3.35 percentage points, *ceteris paribus*, a similar but racial minority-majority county would have a crime rate of 5.10 percentage points.

The geography variables, `west` and `central`, were not statistically significant at the 5% level, but significant at the 10% level. Despite the weak statistical significance, the two coefficients have salient policy implications. It appears that the counties in the Western NC experience 26.0% less crime than Eastern NC, and the Central NC counties experience 17.1% less than the Eastern counties, even after wealth (through `taxpc`) and the ratio of minorities have been controlled for in the statistical model. The policymakers should consider this fact and try to allocate more crime reduction resources toward the Eastern counties, which experience significantly more crime than the other regions.

Eastern NC, especially, appears to be home to more racial minorities. Figure 13 shows a box and whiskers plot of the percentage of minorities in the three regions. On average, the Western counties have 30.36-percentage points less minorities in `pctmin80` than the Eastern counties. Central NC has 12.40-percentage points less minorities than Eastern. In this context, we can deduce that higher crime rates in the Eastern regions are somewhat linked to the higher percentage of racial minorities.

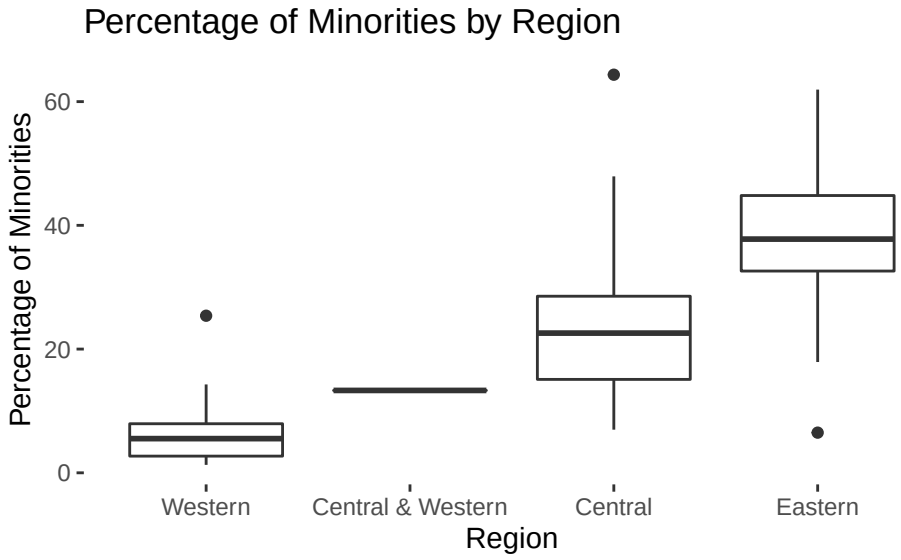


Figure 13: Percentage of Minority per Region of NC

Model 3

In this iteration, we add more variables to see if they add much more insight to the model 2 above. The additional variables are the log of all nine wage variables, log of `prbpris`, and `urban`. We performed log transformations in order to standardize our wage variables such that we can explain changes in crime rate by changes in wages.

Log of Weekly Wage Variables

Wages are one of the most actionable variables from a policy standpoint. For example, the minimum wage could be increased. Since wage also plays a large part in determining the standard of living and access to opportunities, it makes sense to understand how it influences crime. We will include the log of all 9 available wage variables in the model.

Urban

While we know that this is highly correlated with `density`, we would like to capture how other characteristics of large cities affect crime rate.

Probability of Prison Sentence

`pbrpris` is another variable related to the strictness of law enforcement, which will help us understand whether fear of prison affects whether people commit crimes.

Model 3 Generation and Comparison to Previous Models

Table 3: Model 3

	log(Crime Rate)		
	Base Model	Second Model	Third Model
	(1)	(2)	(3)
log(Convictions per Offense)	-0.531^{***} (0.177)	-0.341^{***} (0.094)	-0.356^{***} (0.110)
log(Police per Capita)	0.338 (0.295)		
Average Sentence		-0.003 (0.013)	-0.010 (0.016)
Percent Minority		0.008^{**} (0.004)	0.008^{**} (0.004)
Percent Young Male		0.578 (1.137)	2.366 (1.542)
Ratio of Face-to-Face Crimes to Other Crimes		-0.516 (0.582)	-0.181 (0.674)
log(Tax Revenue per Capita)		0.179 (0.244)	0.277 (0.317)
log(Density)		0.390^{***} (0.084)	0.357^{***} (0.096)
West		-0.260^{*} (0.153)	-0.207 (0.139)
Central		-0.171^{*} (0.103)	-0.197^{**} (0.090)
Urban			-0.148 (0.174)
log(Probability of Prison Sentence)			0.032 (0.176)
log(Weekly Wage: Construction)			0.181 (0.264)
log(Weekly Wage: Trns, Util, Commun)			0.221 (0.347)
log(Weekly Wage: Whlsle, Retail Trade)			-0.076 (0.423)
log(Weekly Wage: Fin, Ins, Real Est)			-0.155 (0.334)
log(Weekly Wage: Service Industry)			-0.581 (0.366)
log(Weekly Wage: Manufacturing)			0.156 (0.172)
log(Weekly Wage: Fed Employees)			0.913^{**} (0.387)
log(Weekly Wage: State Employees)			-0.312 (0.316)
log(Weekly Wage: Local Gov Emps)			0.318 (0.619)
Constant	-2.450 (2.198)	-6.740^{***} (0.683)	-11.340^{***} (3.809)
N	90	90	90
R^2	0.437	0.766	0.821
Adjusted R^2	0.424	0.740	0.769
Residual Std. Error	0.417 (df = 87)	0.280 (df = 80)	0.264 (df = 69)
F Statistic	33.731^{***} (df = 2; 87)	29.123^{***} (df = 9; 80)	15.807^{***} (df = 20; 69)

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Model 3: MLR.1 Linear in Parameters

This assumption is satisfied.

Model 3: MLR.2 Random Sampling

This assumption is not satisfied, but mitigating factors would still make a BLUE OLS model and hypothesis testing possible.

Model 3: MLR.3 No Perfect Collinearity

Again, none of our variables are *perfectly* correlated, although the high number of variables means some are highly correlated, the wage variable especially. However, this is okay, since they are all wages of different jobs, such that none of them are multiples of the other.

```
## log(pbarrandconv)           avgsen         pctmin80          pctymle
##          1.918995         1.490552         2.886066         1.550065
##               mix       log(taxpc)     log(density)             west
##          1.422016         1.789145         4.222223         3.102980
##           central            urban     log(prbpris)        log(wcon)
##          1.999366         2.505713         1.245254         2.055412
##         log(wtuc)        log(wtrd)        log(wfir)        log(wser)
##          1.760155         2.940868         2.571311         2.505564
##         log(wmfg)        log(wfed)        log(wsta)        log(wloc)
##          2.207166         3.131297         1.645957         2.172968
```

Our highest VIF is now `density`, with a value of 4.22. This value raises some concern, since the variance of population density is inflated by over 320%. We still do not have perfect correlation, but this may be a sign that some variables in our model do not necessarily add value. The only thing that may be of concern is whether the tax revenue may be collinear with the wage variables. However, we determined that this is not a cause for concern of multicollinearity, as there are other types of taxes than the income tax (e.g., property tax and sales tax), there are more industries than the 9 in our data, and the tax revenue per capita is more than merely a function of average weekly wages.

Model 3: MLR.4 Zero Conditional Mean

Looking at the plot of residual vs. fitted visuals, the LOWESS (locally weighted scatterplot smoothing) line in red is mostly flat, though less flat than model 2.

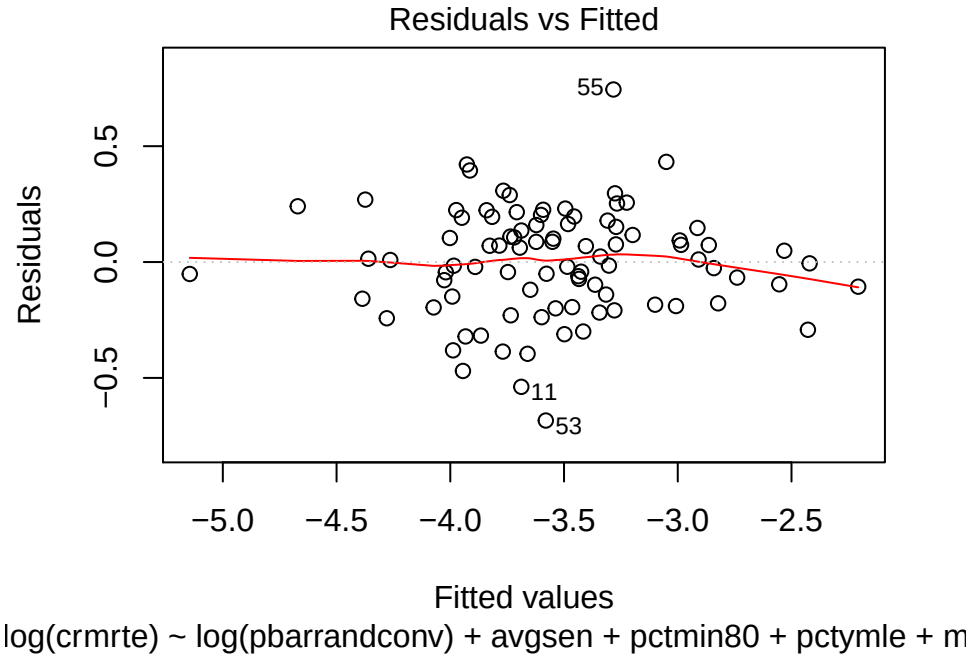


Figure 14: Model 3: Residuals vs. fitted visuals

Looking at the plot of residual vs. fitted visuals, the LOWESS (locally weighted scatterplot smoothing) line in red is still mostly flat.

Model 3: MLR.5 Homoskedasticity

In the scale-location plot, the red line is not entirely flat, which raises concerns about heteroskedasticity.

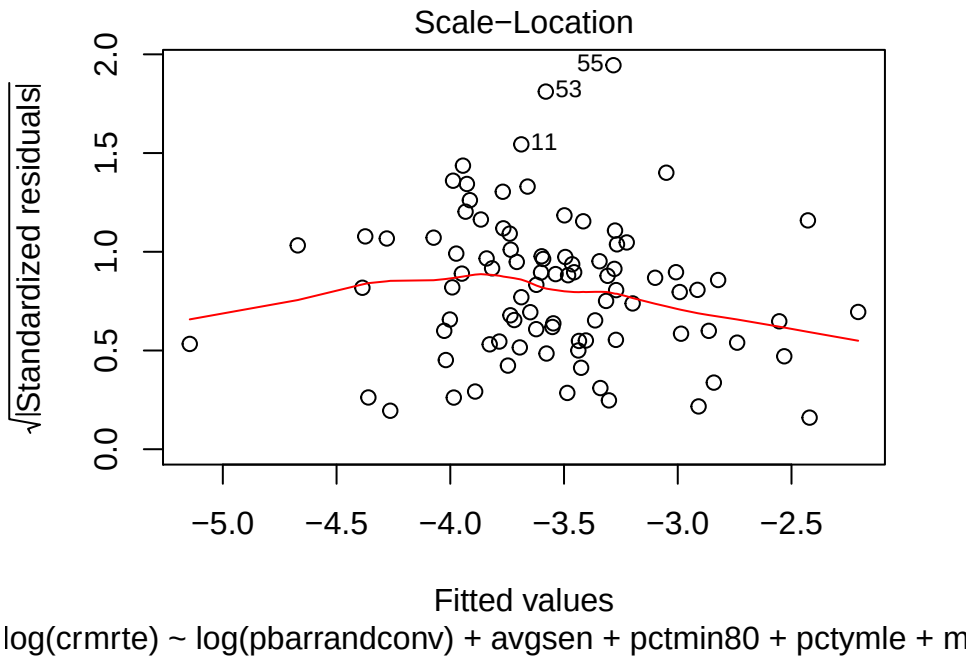


Figure 15: Model 3: Scale-location plot

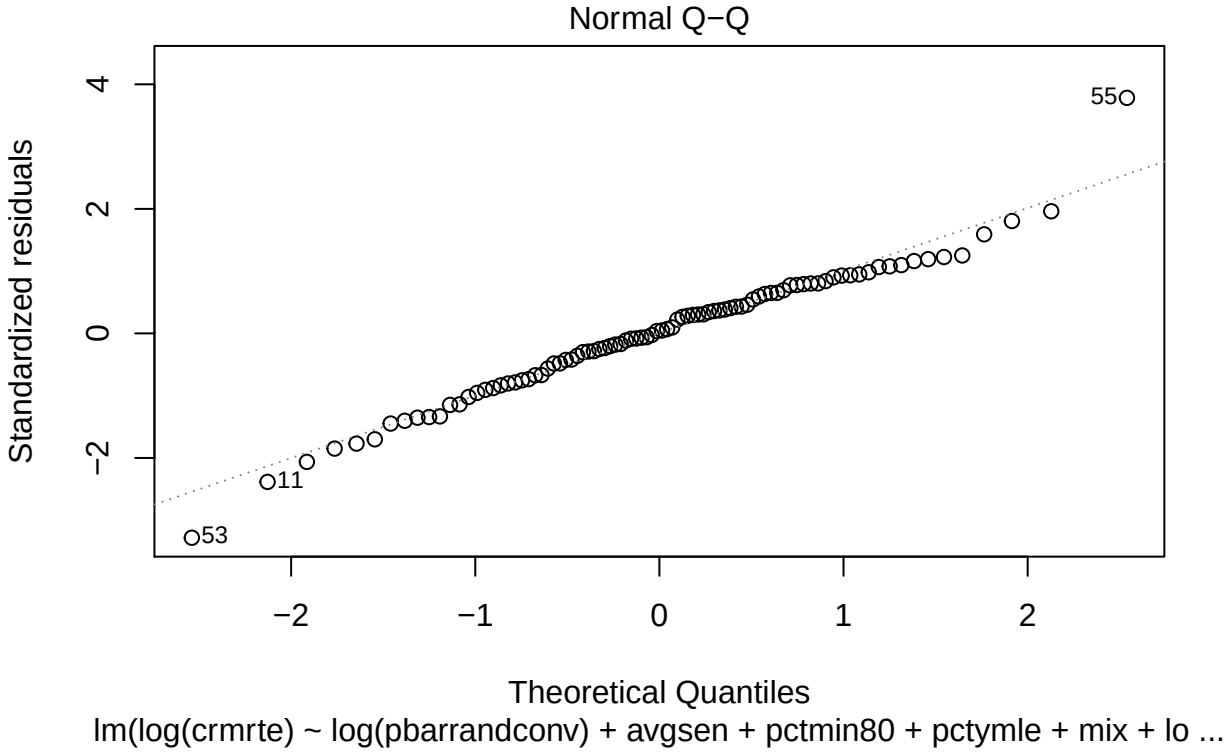
We can use the studentized Breusch-Pagan test to further validate the assumption.

```
bptest(model_3d)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3d
## BP = 45.371, df = 20, p-value = 0.0009826
```

For this model, the Breusch-Pagan test has a p-value less than 0.01, which is highly statistically significant. As a result, we reject the null hypothesis that the model has homoskedasticity and used the heteroskedasticity-robust errors for interpreting the coefficients.

Model 3: MLR.6 Normal Distribution of Errors



Except for a handful of counties on either end of the distribution, the residuals have an approximately normal distribution. This allows us to use the t-distribution to test the hypothesis.

Model 3 Interpretation

Statistical Significance of Each Coefficient

We once again use heteroskedasticity-robust standard errors to run the t-test on each coefficient. According to Table 3 above, some of the variables statistically significant in model 2 are also significant in model 3 at a 5% level or less: `log(pbarrandconv)`, `pctmin80`, and `log(density)`. Unlike model 2, `central` is now statistically significant at the 5% level, while `west` is no longer significant even at 10%. `log(wfed)` is also statistically significant.

We will compare model 2 and 3 to determine whether all the variables added to create model 3 from 2 are jointly significant and the coefficients of the shared variables between the two models did not change much.

Joint Statistical Significance of Coefficients

```
##
##  F test to compare two variances
##
## data:  model_3d and model_2h
## F = 0.88826, num df = 69, denom df = 80, p-value = 0.6159
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5635583 1.4111077
## sample estimates:
## ratio of variances
##          0.8882639
```

All of the variables added to model 2 to create model 3 turn out to be jointly *irrelevant*. We cannot reject the null hypothesis that the variances of model 2 and 3 are equal, because the p-value of the F test to compare the two variances is 63.11%. In other words, model 3 does not provide much more insight than model 2. This result illustrates that model 2 is quite robust.

Practical Significance

The practical significance does not change much from model 2, because all additional variables are jointly insignificant.

Final Model Selection

To predict the crime rate, we recommend using model 2. Even though model 3 has a slightly higher R-squared value, it does not seem to be adding much more insight than what we already knew from model 2. Then, all things being equal, simpler models with fewer variables are better than overfitted models.

Conclusion and Policy Implications

Our results indicate that stricter criminal justice does not have practical implications in reducing crime. The average sentence length did not have a statistically significant relationship with the crime rate. Additionally, we could not comment on the impact of increased police concentration, due to the lack of data on the prior year police per capita or prior year crime rates. And although the ratio of convictions to offense appeared to have a statistically significant correlation coefficient, its practical implications were negligible. In this regard, we recommend that the political campaign focus less on more severe punishment of crime, given that additional convictions may be an inefficient use of tax dollars that has a trivial impact. We also found that average wages across various industries as a group do not predict or influence the crime rate. Therefore, we cannot make any data-based policy recommendations for economic development programs.

On the other hand, we found that the percentage of minorities had a high statistical and practical significance. Although the regional indicator variables were not statistically significant at the 5% level, it is worth noting their implications in the context of the number of racial minorities, because the counties in Eastern North Carolina not only experienced higher rates of crime, but also included higher percentages of minorities.

Another noteworthy predictor of crime was population density. After controlling for a variety of other variables, density continued to be a robust explanatory variable. This may be explained by network effects or well-developed transportation systems that may bring an influx of criminals and targets alike. Although we cannot comment on the root cause with the information available from this study, we would recommend that any policies be first directed toward high-density counties for a maximum impact.

Perhaps another study that examines the socioeconomic statuses, opportunities, and other factors between racial minorities and Whites can provide insights into why a higher percentage of minorities or the Eastern counties are correlated with higher crime rates, which may help devise better social policies. In the meantime, for a political campaign with the goal of getting more votes, a political candidate may consider creating a platform around increased economic opportunities and reduced criminal punishments. Such platforms may appeal to the constituents of the Eastern counties that are racial minorities, as those constituents face severe probability of arrest or higher sentences and subsequent loss of economic opportunities.