Analyzing Determinants of Crime

Rowan Cassius, Michael Steckler, Julian Pelzner

12/3/2019

Introduction

Understanding the determinants of crime is an age-long question that is well studied in economic literature. This report aims to contribute to the discussion through an analysis of crime rates within counties in North Carolina in 1987. The data we analyze contains information about crime rates and various socioeconomic, geographic and demographic factors at the county level for 91 of North Carolina's 100 counties.

This study was conducted for the reelection campaign of the governor of North Carolina. Political consultants have identified crime as a primary voter initiative. Through our analysis we seek to understand how effective increased penalization is at combating crime and how the living conditions within a county affect instances of crime.

In this report, we provide an analysis that illuminates the determinants of crime, and we suggest policy actions that can be implemented at the local government level. The report is structured as follows:

1. First, we check for any holes in the data and clean the dataset accordingly.
2. Next we perform an exploratory analysis of the variables which we reason are the strongest proxies for the concepts we wish to study. This is then followed by fitting a baseline regression model and evaluating all classical linear model assumptions.
3. Then, we fit several more regression models to control for omitted variables and examine the robustness of the baseline results.
4. Lastly, we generate policy suggestions according to our findings.

Data Cleaning

Exploration of the dataset for data validity uncovered the several problems with, after addressing which, the data were vetted and ready for analysis.

- 6 of the original 97 rows have all values missing. We eliminated such rows from the dataset before analysis.
- The `prbconv` variable, representing the ratio of convictions to arrests in a given county, was registered as a factor. In repsonse, we cast probability of conviction as a numeric variable.
- Examination of the distribution summary table uncovered that all of the data appeared within expected bounds with the exception of `prbarr`, a variable representing the probability of arrest in a given county, because the maximum value belonging to this variable is > 1, outside the bounds of a probability. To address this we took the following course of action:

```
# Examining outlier in probability of arrest.
ggplot(crime, aes(x=prbarr)) +
  geom_histogram(bins = 20, fill = "dodgerblue4") +
  xlab("Probability of Arrest") +
  ylab("Count")
```

By looking at the histogram of `prbarr`, the variable has one outlier at 1.09. This is puzzling because it suggests that the number of convictions in the associated county was greater then the number of arrests in 1987. Some plausible explantions for this apparent outlier include the following:

- The number of convictions includes spillover from previous years. That is, the number of convictions includes convictions from 1986 or earlier while the number of arrests only includes arrests from 1987.

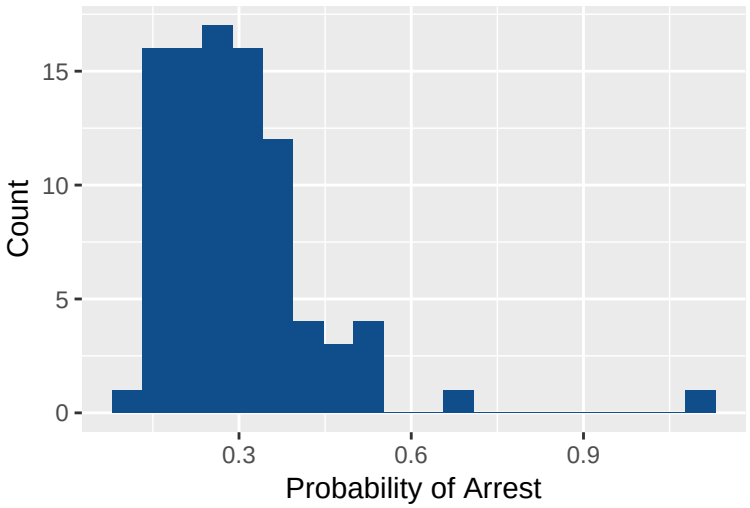

Figure 1: Distribution of Probility of Arrest

- The number of convictions includes spillover from neighoring counties. That is, some criminals were arrested in neighboring counties and tried in the county associated with this outlier, letting the apparent number of convictions exceed the number of arrests.
- The outlier is a mistranscription. The official who reported the figure meant to report 0.09.

While plausible, it is not safe to assume the outlier is a mistranscription, so we stongly speculate that it is a result of spillover either from a different year or a different county or both. Both cases of spillover suggest that the observation containing this outlier did not come from the underlying population we intend to study: counties which each report their own crime and crime-related information exclusively in the year of 1987. On these grounds, we removed the observation containing this outlier from the analysis.

Exploratory Analysis

Variable Selection

In this analysis, we seek to understand the extent to which penalizing criminal activity and living conditions interact with crime, and it is therefore necessary to choose variables representing all of these concepts. Among all the variables in the dataset we select `crmrte`, representing the number of crimes committed per person in a given county, as the dependent variable used to measure crime. We also select `prbarr`, representing the ratio of arrests to total offenses, also known as the probability of arrest, as a proxy for measuring counties' efforts to penalize crime on the street level.

While there are other variables in the data which can argualy proxy counties' efforts to penalize crime, such as the probability of conviction (after arrest), the probability of imprisonment (after conviction), and the average prison sentence length, we choose probability of arrest because it is the most tangible form of penalty among all those available and the also first opportunity for a legal institution to prosecute an offender. Apprehension is a county's immediate response to a criminal's offense and is necessary before the criminal can experince any other ensuing forms of penalty. Additionally, because prison sentence is a metric only based on crimes serious enough to warrant imprisonment, it does not reflect a county's effort to penalize all crimes but only the most serious subset of them. Therefore, we consider the probability of arrest to be the strongest proxy for crime penalization that is most likely to deter a criminal from committing an offense. In addition, a higher probability of arrest could indicate that the police are more effective at detaining people who commit crimes in the county.

Moreover, we hypothesize that an increased likelihood of arrest stifles crime rates, because if criminals are less likely to escape punishment after committing offenses, we argue that this will make them less inclined to

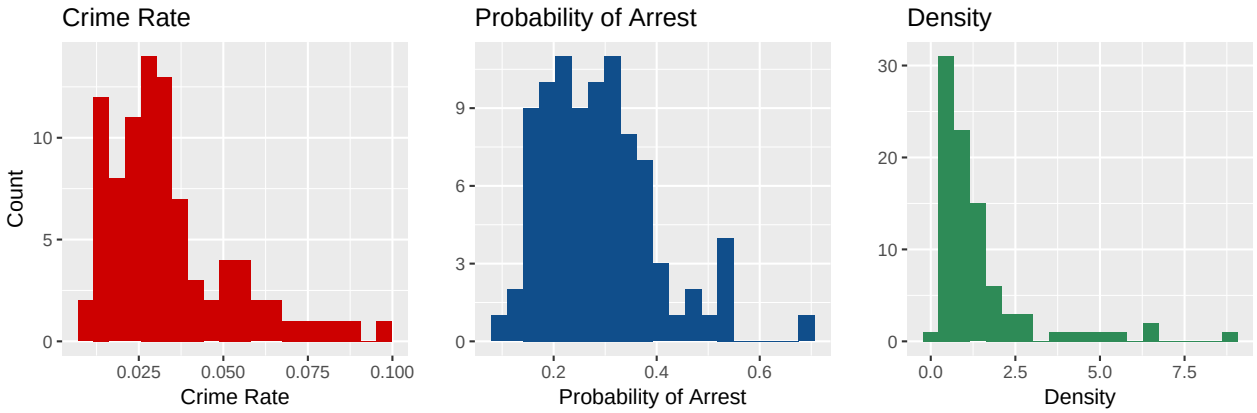

Figure 2: Marginal Distributions of Crime Rate, Probability of Arrest and Density

commit offenses in the very first place.

The variable of interest we identify as most relevant to a county's living conditions is `density`, representing the the number of people per square mile. We hypothesize that higher population density leads to an increase in crime rate for two main reasons:

1. In a county with a high population density, there are more vulnerable denizens per square mile for criminals to perpetrate crimes against than there there are in a county with lower density.
2. Prior research suggests that denser areas tend to make people more irritable due to increased economic competition and lack of living comfortability, thus making people in denser areas more likely to act out through criminal behavior.

In summary, the key explanatory variables in our baseline model will be probability of arrest and density. The next figure shows the marginal distributions of crime rate and both key explanatory variables.

```
# Crime histogram
plot.crime = ggplot(crime, aes(x=crmrte)) +
  geom_histogram(bins = 20, fill = "red3") +
  xlab("Crime Rate") + ylab("Count") + ggtitle("Crime Rate")

# Arrest Probability histogram
plot.arrest = ggplot(crime, aes(x=prbarr)) +
  geom_histogram(bins = 20, fill = "dodgerblue4") +
  xlab("Probability of Arrest") + ylab("") + ggtitle("Probability of Arrest")

# Density histogram
plot.density = ggplot(crime, aes(x=density)) +
  geom_histogram(bins = 20, fill = "seagreen4") +
  xlab("Density") + ylab("") + ggtitle("Density")

grid.arrange(plot.crime, plot.arrest, plot.density, nrow = 1, ncol = 3)
```

Each distribution has some right skew and a couple outliers. This means that some observations may have high leverage over a linear regression. Examination of bivariate scatter plots will alert us to whether any of the leverage is problematic.

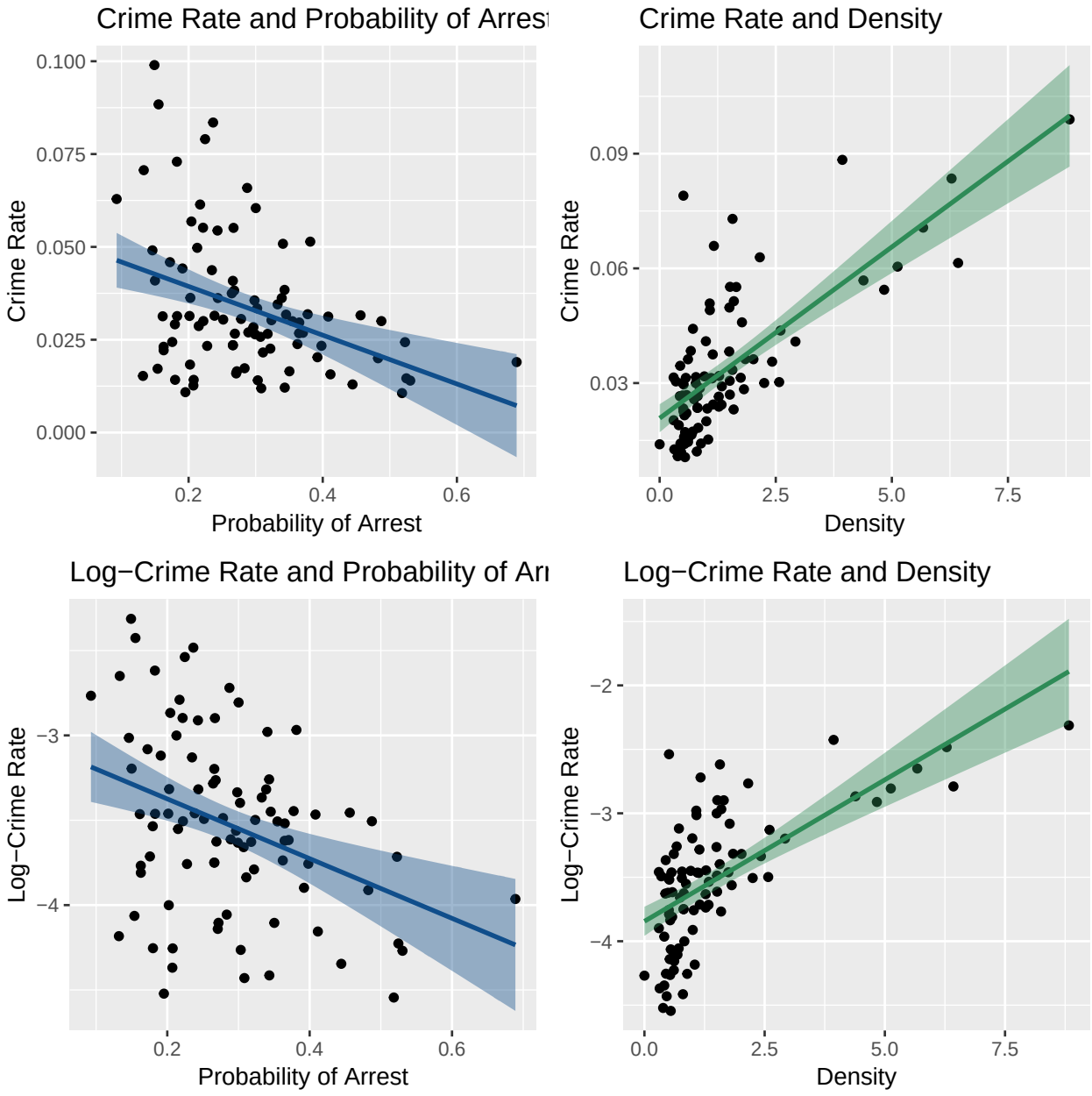

Figure 3: Scatter Plots of Crime against Probability of Arrest and Density

Bivariate Analyses

The set of plots above compares the scatter plots of crime rate with each primary explanatory variable and the scatter plots of log-crime rate with the same variables. It appears that the untransformed specifications have the strongest linear relationship with crime rate, so we will specifiy our baseline model without transforming crime rate. Refraining from transforming the response variable will enable us to directly interpret the model coefficients as the associated changes in crime rate for a single unit increases in each explanatory variable. Additionally, none of the points currently appear to have problematically high influence.

The first scatter plot illustrates that likelihood of arrest has a negative correlation with crime rate, supporting our hypothesis that increasing crackdown on crime is associated with a lower crime rate. The second plot supports our second hypothesis that increases in popualation density are associated with increased crime rates. We will proceed to fit a regression model which regresses crime rate on probability of arrest and density to precisely understand their relationships.

Baseline Model Specification.

$$crimerate = \beta_0 + \beta_1 prbarr + \beta_2 density + \epsilon$$

Assessing CLM Assumptions

1. **Linearity**

   After our bivariate exploratory analyses of crime rate with probability of arrest and crime rate with density, we chose the specification offering relationships that appear linear. However, we can argue that the model is linear nonetheless because we have not yet constrained ϵ.

2. **Random Sampling**

   We can infer that because there are 91 data points, out of 100 possible counties in the state, that the researchers intended to complete a census of the population of counties in the state. However, because data from a few counties was inaccessible, they limited the study to a convenience sample of the remaining counties. Nonetheless, while the counties represened in the dataset don't form a fully random sample, since over 90% of counties in the state represented, we conclude that the results of our analysis will still offer useful insights into the causal factors behind crime rate at the county level. We assume that the inavailability of data in some counties is not connected to the crime rates therein in any important way.

3. **Multicollinearity**

   To test for multicollinearity, we have calculated the inflated variance factors for both regressors in the first model. The variance inflation factors for both probability of arrest and density are 1.12, 1.12 respectively, each of which is close to 1, so there is no problematic multicollinearity and, therefore, no perfect multicollinearity.

4. **Zero Conditional Mean**

   To examine the zero conditional mean assumption, we have plotted the residuals as a function of the fitted values. By inspection of the residuals vs fitted values plot (see Appendix), there is not substantial evidence of violation of the zero conditional mean assumption. The conditional mean remains close to zero but tends downward slightly as the fitted values increase.

5. **Homoscadascity**

   By examination of the residuals vs fitted values plot (see Appendix), there appears to some degree of heteroscedasticity because the magnitude of the residuals tapers off at both the highest and lowest fitted values. To address this, we will use heteroscedasticty robust standard errors when making inferences.

6. **Normality of Errors**

   By inspection of the residuals' normal Q-Q plot and their distribution (see Appendix), the error distribution suffers from a fat right tail and a short left tail. While this means the error distribution deviates from normality, the sample size of 90, which is far greater than 30, enables us to confidently invoke the central limit theorem and conclude that in spite of the errors' non-normality, we can still approximate the sampling distributions of the model's coefficients and make sound inferences about them.

Model Results

The baseline regression analysis (see Table 1) suggests that both probability of arrest and density both have significant effects on crime at the 5% and 0.1% significant levels respectively, with coefficients -0.028 and 0.008 respectively.

The first coefficient suggests that in a county of N residents, a 10% increase in the probaility of arrest is associated with a $0.1 \times 0.028 \times N$ reduction in the number of crimes the county experiences per year. For example in a county of 100,000, a 10% increase in the chance of arrest is expected to reduce the number of crimes the county experiences by 280, which would be a welcome outcome to county residents.

Similary, density's coefficient of 0.008 suggests that in the same county as the one described above, de-densifying the county by 10 people per square mile will reduce reduce the number of crime in the county by 8000! While both initial results are exciting, we are careful not to recommend policies based on these results exclusively because the regressors are likely endogeneous to some degree and their effects on crime may be the artifacts of omitted variables.

Assessing Exogeneity

Exogeneity of density:

Plausible ommitted variables that may have biased the effect of density include the following:

- Whether a county is urban

  Urban localities are known for having higher crime rates than suburban and rural areas and they are also more dense. Therefore the omission of this variable is likley to have biased the effect of density on crime rate upward from zero. Because `urban` is an indicator variable in the dataset for whether a county is urban, we can control for this omitted variable directly by including urban in subsequent regression models.

- Economic inequality

  Economic inequlity is likely to be positively correlated with both crime rate and desnity. Therefore the omission of this variable has most likely inflated the postive effect of density on crime rate. In the absence of a metric for economic disparity such as a gini coefficient, we can mitigate the bias from this omitted variable by including variables descibing other economic conditions, such as the various wage variables and tax per capita variable that are provided in the given dataset.

- Social Cohesion.

  Social cohesion is likely to be negatively correlated with density. In other words, denser areas may experience higher rates of competition, which can cause distrust and social unrest. In addition, social cohesion is likely to be negatively correlated with crime rate. We expect that the more cohesive a society (county) is, the lower its crime rate will be. Thus, the omission of this variable is likely to have biased the effect of density on crime rate upwards away from zero. However, metrics for this concepts do not exist in the given dataset. To that end, future work should strive to collect more data related to social cohesion. Useful metrics could include level of neighbor trust, and rates of riots, protests, and other forms of social unrest.

- Geographic region

There may be varying levels of crime based on the region in North Carolina in which counties are located. It is difficult to postulate the sign of correlation between region and density, or between region and crime rate, and therefore it is difficult to postulate the direction of the bias results from the omission of this variable. Nonetheless, it is a relevant concept which should be controlled for in our analysis. Fortunately, there are west and central dummy variables in the dataset, which proxy geographic factors in different parts of North Carolina.

Exogeneity of probability of arrest:

Plausible ommitted variables that may have biased the effect of probability of arrest include the following:

- Police presence

  We speculate that police presence on the street is likely to be negatively correlated with crime rate, but positively correlated with the probability of arrest. This speculation suggests that omitting police presence has inflated the negative effect of the probability of arrest on crime rate by pushing the coefficient away from zero. We can control for police presence by including the police per capital variable in subsequent models.

- Harshness of police force

  We reason that the harshness of a county's police force is negatively correlated with the crime rate in that county and positively correlated with the probability of arrest in that county. Therefore, omission of this variable from our model would suggest that the effect of probability of arrest on crime rate is biased downwards away from zero. Metrics for this concept do not exist in the given dataset either. Useful metrics could include rate of violent arrests and the degree to which police are armed with weapons. To that end, future work should strive to collect more data measuring the harshness of police forces.

- Proportion of face-to-face crimes

  In this context, we are thinking about variety of offenses in the sense that some counties may have a mix of offenses which contain a higher proportion of personal or face-to-face crimes relative to other counties. We theorize that the proportion of face-to-face crimes is positively correlated with probability of arrest, because these crimes are generally more serious in nature and could provide victims with descriptions of suspects which are o great police when searching for criminals. Additionally, we reason that proportion of face-to-face crimes is positively correlated with crime rate. Therefore, omission of this variable would has biased the effect of probability of arrest on crime rate upwards towards zero. The dataset includes the variable `mix`, which directly represents the proportion of face-to-face offenses in the dataset, and we will use it to control for the proportino of face-to-face offenses.

Future Model Specifications

To combat the biases that may have arisen from omitting the aforementioned variables, we will proceed by fitting new models which control for particular omitted variables. After including the controls, we will evaluate the robustness of the effects observed in the baseline regression. We will proceed to fit the following two regressions:

1. In our first subsequent regression model, we only include the variables which we believe would contribute the most to mitigating omitted variable biases. These variables include the urban indicator variable, `urban`, the police per capita metric variable, `polpc`, as well as the offense mix proportion variable, `mix`.
2. The final iteration of our model includes controls for geographic fixed effects, utilizing the regional indicator variables `west` and `central`. Additionally, this iteration of our model controls for economic conditions by including wage variables and the tax per capita variable, `taxpc`. Furthermore, we decide to include the variable `pctymle`, which describes the percentage of young males under the age of 24 that reside in a given county, because young men are the most fit demographic for committing face-to-face crimes.

Model 2 Specifcation

$$Y = \beta_0 + \beta^T X + \gamma^T Z + \epsilon$$

Model 3 Specifcation

$$Y = \beta_0 + \beta^T X + \gamma^T Z + \theta^T W + \epsilon$$

$$Y = crimerate, X = \begin{bmatrix} prbarr \\ density \end{bmatrix}^T , Z = \begin{bmatrix} urban \\ polpc \\ mix \end{bmatrix}^T$$

and W are the columns of the dataset pertaining to wage, geographic and demogaphic factors.

Regression Summaries

We reference the stargazer table below (Table 1) to display the results of all three regression models. After fitting regressions which control for ommitted variables, our findings are as follows:

Throughout all three specifications, `prbarr` and `density` remain statistically significant and their coefficient estimates remain relatively stable. Therefore, the previous interpretation of the coefficients from our baseline model is still valid and we will generate policy recommendations with regards to those two key explanatory variables. However, characteristics of the key regressors were not assigned randomly to the counties in the dataset, so we can not definitively conclude that these variables were the true causes of any observed differences in crime rates. Nonetheless, the robustness of the coefficients to controlling for geographic, economic and demographic factors provides strong evidence that density and probability of arrest have causal effects on crime. Additionally, the adjusted R^2 value increases for each specification, which suggests that the controls we introduced accounted for some of the variation in our response variable and were important to include.

It is also worth noting that in the thrid specification that the percent young male variable (`pctymle`) has a signigicant positive effect at the 5% level. Besides percent young male, none of the additional regressors associated with the third model are statistically significant. However, the significance of `pctymle` suggests that this variable is relevant to crime rates and we take it into account when considering policy recommendations.

Table 1:

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
prbarr	-0.028^{*}	-0.028^{*}	-0.036^{*}
	(0.012)	(0.013)	(0.016)
density	0.008^{***}	0.006^{***}	0.007^{***}
	(0.001)	(0.002)	(0.002)
polpc		7.832	3.902
		(5.649)	(7.751)
urban		0.005	-0.002
		(0.010)	(0.010)
mix		0.002	0.002
		(0.018)	(0.022)
pctmin80			0.0003
			(0.0002)
pctymle			0.168^{*}
			(0.076)
Geographic fixed effects	No	No	Yes
Wages fixed effects	No	No	Yes
Observations	90	90	90
R^2	0.554	0.613	0.789
Adjusted R^2	0.543	0.594	0.732
Residual Std. Error	0.013 (df = 87)	0.012 (df = 85)	0.010 (df = 70)
Note:			*p<0.05; **p<0.01; ***p<0.001

Recommendations

Based on our findings, we have 2 major recommendations for policies that could be implemented in counties across North Carolina to reduce crime: allowing stop-and-frisk procedures by local police forces, and funding more community centers and programs for the youth.

We recommend that the governor should collaborate with the state and local governments and their police departments to enact and implement a stop-and-frisk policy. This crime penalization policy would effectively provide police forces with the ability to increase the probability of arrest in their jurisdictions. Considering that this dataset provides evidence that a higher probability of arrest is related to a decrease in crime rates, we reason that a policy enabling police to stop and frisk locals will result in a noticeable reduction in crime. Next steps involve consulting state, city, and district attorneys and police chiefs to craft a series of internal policies and bills that will support law enforcement in these efforts. Furthermore, in an effort to preempt any racial bias concerns, we suggest that police forces undergo a series of mandatory implicit bias trainings as well.

We also recommend that the governor fund community building programs throughout the state, especially for the youth. This policy is aimed as ameliorating issues associated with social cohesion and economic opportunity as they pertain to living conditions. Supporting youth organizations and community centers are major components of building positive social networks and reducing crime rates. Given that the dataset provided evidence of a connection between a large propotion of of young male in a given county to higher crime rates, it behooves policy makers to invest in ther success by providing them mentorship and career-building support. The governor should rally support among statewide and citywide legislatures to ensure these laws can be passed and implemented in a timely fashion.

Conclusion

This study explored the effects of density and penilization of criminality on crime. Through a regression analysis, we found that a county's probability of arrest, density, and percentage of youth males are characteristics that are significantly related to it's crime rate. Specifically, a regression analysis controlling for economic, geographic and demogrphic factors showed that while density and percentage of young males are positively correlated with crime rate, higher chance of arrest is expected to decrease crime rate.

According these results, we generated policy recommendations for the re-election campaign of North Carolina's governor. We suggest that probability of arrest can be strengthened by implementing more aggressive policing policies such as a stop-and-frisk procedure. However, counties' densities can not easily be lowered after the fact, so we defer any policy suggestions related to population density. Additionally, we encourage investment in community resources and youth programs so that young people have more guidance are less likely to enage in criminal behavior.

Appendix

- Summary Statistics

Table 2:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
county	91	101.615	58.794	1.000	52.000	152.000	197.000
year	91	87.000	0.000	87.000	87.000	87.000	87.000
crmrte	91	0.033	0.019	0.006	0.021	0.040	0.099
prbarr	91	0.295	0.137	0.093	0.206	0.344	1.091
prbpris	91	0.411	0.080	0.150	0.365	0.457	0.600
avgsen	91	9.647	2.847	5.380	7.340	11.420	20.700
polpc	91	0.002	0.001	0.001	0.001	0.002	0.009
density	91	1.429	1.514	0.00002	0.547	1.568	8.828
taxpc	91	38.055	13.078	25.693	30.662	40.948	119.761
west	91	0.253	0.437	0.000	0.000	0.500	1.000
central	91	0.374	0.486	0.000	0.000	1.000	1.000
urban	91	0.088	0.285	0.000	0.000	0.000	1.000
pctmin80	91	25.495	17.017	1.284	9.845	38.142	64.348
wcon	91	285.358	47.487	193.643	250.782	314.795	436.767
wtuc	91	411.668	77.266	187.617	374.632	443.436	613.226
wtrd	91	211.553	34.216	154.209	190.864	225.126	354.676
wfir	91	322.098	53.890	170.940	286.527	345.354	509.466
wser	91	275.564	206.251	133.043	229.662	280.541	2,177.068
wmfg	91	335.589	87.841	157.410	288.875	359.580	646.850
wfed	91	442.901	59.678	326.100	400.240	478.030	597.950
wsta	91	357.522	43.103	258.330	329.325	382.590	499.590
wloc	91	312.681	28.235	239.170	297.265	329.250	388.090
mix	91	0.129	0.081	0.020	0.081	0.152	0.465
pctymle	91	0.084	0.023	0.062	0.074	0.083	0.249

- CLM Diagnostic Plots

```
ggplot(crime, aes(x=model.1$fitted.values, y=model.1$residuals)) +
  geom_point() +
  geom_smooth() +
  xlab("Fitted Values") +
  ylab("Residuals") +
  ggtitle("Residuals vs Fitted Values")
```

```
residual <- data.frame(model.1$residuals)
p.resid <- ggplot(residual,
                  aes(x=model.1.residuals)) +
  geom_histogram(bins=20) +
  xlab("Residual") +
  ggtitle("Residual Histogram")
q.resid <- ggplot(residual, aes(sample = model.1.residuals)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Residual Q-Q Plot")

grid.arrange(p.resid, q.resid, nrow = 1, ncol = 2)
```

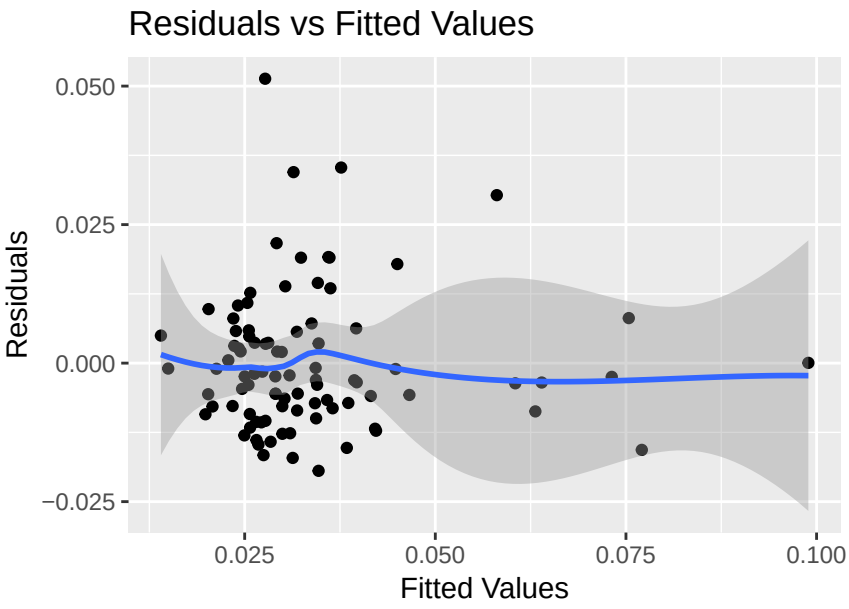

Figure 4: Baseline Models Residuals vs Fitted Values Plot

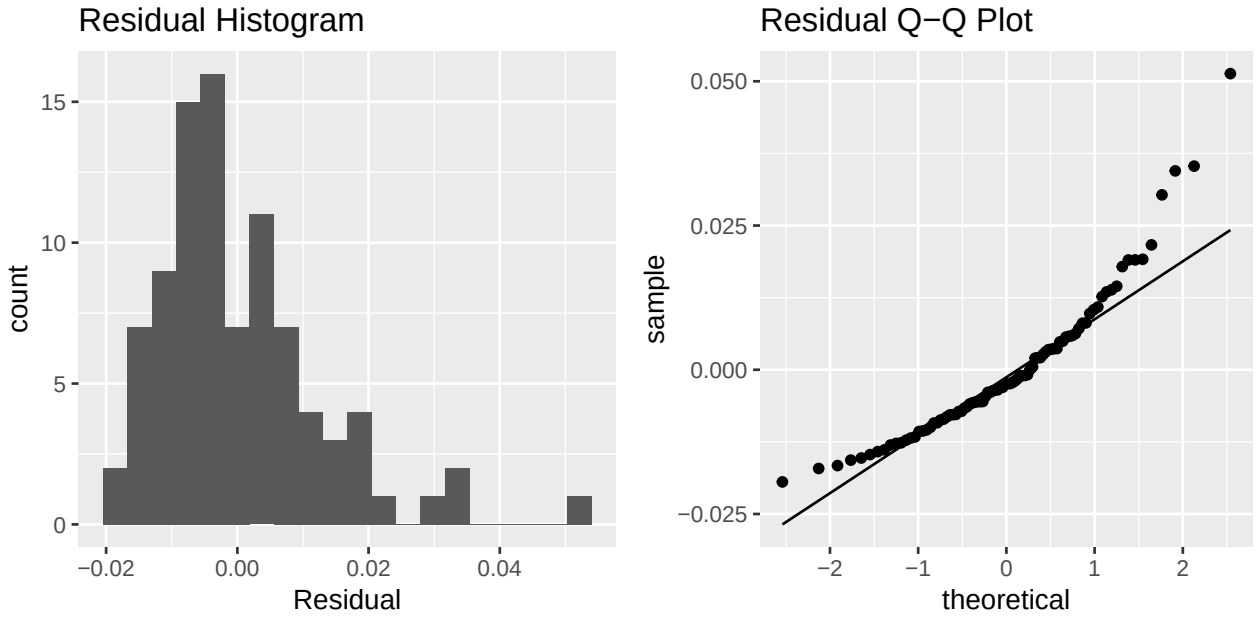

Figure 5: Residuals' distribution and Normal Q-Q Plot