Modeling Crime in North Carolina

Jonathan Hansen, Graham Miotke, Michael Visconti

Introduction

2020 and 2021 have proven to be very unique years in terms of civil unrest around the nation. Starting with the murder of George Floyd in 2020, rioting and social justice movements began to pop up globally and nationally [1]. In conjunction with these riots, there was a substantial increase in crime, such as homicide around the country [2]. This led to an interesting question: What variables have a substantial impact on crime rates in North Carolina? Furthermore, do crime rates vary from the North Carolina crime rate at smaller levels, such as between county-to-county?

The first important consideration is whether knowledge of a potential punishment has a substantial effect on whether individuals will choose not to commit a crime. This is known as a deterrent effect. One example of this is seen in debates over the death penalty. If a person knows that by killing another human being, they will also be killed through capital punishment, then they may decide not to commit the crime. It stands to reason that if a person has knowledge that committing a crime in a certain place gives them a greater chance of a definite punishment, the overall number of crimes committed in that place should be lower [3]. As a result, potential deterrent effects can be analyzed to see if, in North Carolina, they have an impact on the crime rate. There are three main measures of punishment after committing a crime: arrest, conviction, and prison sentencing.

One of the main talking points over the last two years has been whether governments should defund the police [1]. Therefore, investigating the influence of police densities on crime rates would also be of interest. Mainly, would having denser law enforcement groups in one place deter people from committing more crimes there? Or is the addition of more cops not a crime deterrent? A good way to model police presence would be to look at a county's police density, relative to the overall North Carolina police density.

Finally, most major crime reports come from larger cities. Highly populated places, such as Detroit and Baltimore are generally regarded as the most dangerous places in the country [4]. Is this due to the fact that crime rate is affected by more densely populated areas, or are crimes from these places mainly talked about because of the name value these cities hold?

In order to answer all of these questions, a dataset with information on probability of receiving punishment, presence of police and population density is required. Additionally, since modeling crime on a county level was of interest, the dataset would have to include observations at a county level. The Crime dataset in the Ecdat R package included this information and was used for analysis in this experiment [5].

When we set out to begin modeling the data, we created two models. The first model pooled all the counties in North Carolina together, and for the second, we modeled the data hierarchically, with each county partially-pooled together, to assess how crime changes across countries and fit crime models for each individual county. After fitting the models, the predictive power would be compared to assess which model was a better fit for the data.

Once the models were fit, we found that in both, the only thing which truly deterred people from committing a crime was a higher probability of getting arrested. Probability of receiving a prison sentence, areas with more police per capita or higher population densities were not found to be deterrents.

Data Description

In order to model crime rate off these variables, our group leveraged the R package, Ecdat for crime data. The Crime data set from Ecdat package includes variables covering crime, economics, and demographics for 90 counties in North Carolina. The data spans seven years, from 1981 to 1987, leading to a total of 630 observations. Notably, this dataset contains county-specific information on the probability of arrest, conviction, and prison sentence, as well as the police per capita and population density. To avoid overfitting the model, the rest of the variables were excluded. Some excluded variables include the percentage of the population classified as minority (pctmin), weekly wages for several industries, one of which is federal employees (wfed), and the percentage of the population classified as young males (pctymle).

After consideration, our group selected a subset of the variables which seemed to best answer the research question. The first being the probability of arrest (prbarr), which was calculated as the proportion of the number of arrests to the number of offences. Next, the probability of conviction (prbconv), which was calculated as the proportion of the number of convictions to the number of arrests. We also considered the probability of a prison sentence (prbpris), which was calculated as the proportion of prison sentences to convictions. In order to determine whether a county should increase or decrease the number of police officers on duty, the variable police per capita (polpc) was selected. The variable police per capita is meant to measure a county's ability to detect crimes [4]. The final variable selected for modeling crime rates was the county's population density (density), measured as people (100s) per square mile.

One problem found in the data was observations having "probabilities" greater than 1. In total, there were five observations of the probability of arrest and 71 observations of the probability of conviction with probabilities greater than 1. The five probability of arrest observations all occurred in two counties, county 115 and county 185. As a result, these two counties were removed from the data included in our model. Consequently, the total number of observations was reduced from 630 to 616. Unfortunately, there was no county-specific pattern with probability of conviction observations being greater than 1. Therefore, the probability of conviction variable was removed from the model entirely.

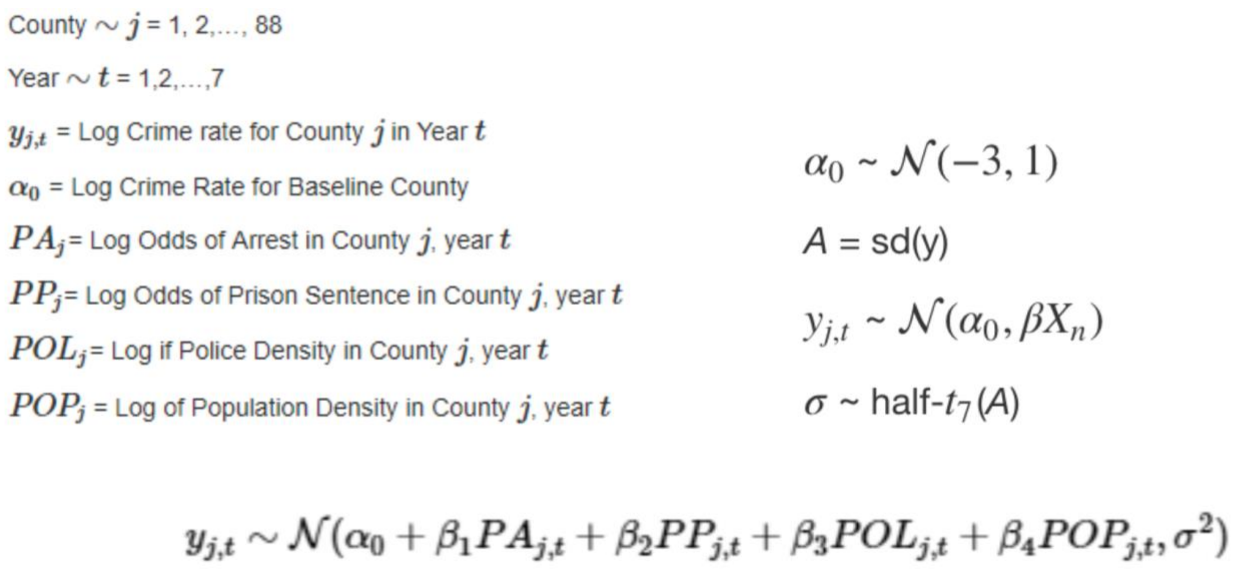
Another pre-processing step we took was to convert all four predictors to the log scale. The natural log odds were taken for the variables probability of arrest and the probability of a prison sentence. For the variable police per capita, the natural log was taken on the proportion of the density of police officers for a county to the average density of police officers for a county in North Carolina. The same process for the police per capita variable was repeated for the population density variable. Finally, the natural log was taken on the response variable, crime rate, which is defined as the crimes committed per person. All the variables were converted to the log scale, so the resulting coefficients on the explanatory variables could be interpreted as elasticities. For example, a 100% increase in a coefficient would result in a 100%*coefficient change in the log crime rate.

Model

Following the selection and transformation of our variables to the log scale, linear regression was utilized to model crime rates. For the model intercepts, the pooled model would use a "baseline" log crime rate for all 88 counties, and the hierarchical model would use a random intercept for county $j = 1, 2,..., 88$. In the pooled model, the intercept would represent a "baseline" county, where the probability of arrest and prison sentence were equally likely at 50%, and the police and population densities were equal to 1. Essentially, if these densities were equal to 1, the county would have the exact density as the average county in North Carolina. County 147, in 1983, was the closest county to this description and was chosen as our baseline. The log crime rate for county 147, in 1983, was -3.0. The rationale in including both a pooled and hierarchical model, was to see if the hierarchical model, with different intercepts, would be able to better predict the crime rate from county-to-county, in year t = 1, 2,...,7. *A priori*, it was unclear which of the models would predict better, so both models were fit to see which predicted better.

Prior Selection

To start, the first six years of data (1981- 1986) for each county were grouped together. The last year, 1987, was excluded in order to test the model's predictions for a county's crime rate in 1987, against the observed crime rate. Next, the mean across the six years for all four explanatory variables, was calculated and plotted against the log crime rate for all 88 counties. From these plots, if the explanatory variable increased and the crime rate decreased, the explanatory variable would be said to have a deterrent effect. However, at the same time, if the crime rate increased, the explanatory variable would be said to not have a deterrent effect. The log odds of arrest seemed to be the only potential deterrent. As a result, for setting priors on the explanatory variables' coefficients, the coefficient for the log odds of arrest would be negative, and the others would be positive. Log odds of arrest and log odds of prison sentence exhibited more variability county-to-county, so the prior variances were set larger. Coefficients were given priors of the increase or decrease in log crime rate following a 100% increase in the explanatory variable. Then, for the pooled model, the prior for the intercept was chosen to be the log crime rate (-3.0) for county 147, in 1983, as mentioned previously. For the hierarchical model, the random intercept was randomly drawn from a Normal(-3, 1) distribution. In order to model the uncertainty from county-to-county, the random draw of the intercept would have additional variation, sigma, following a half-t distribution with a shape parameter of 7 degrees of freedom and a scale equal to the standard deviation of the log crime rate. Then each observation $y_{j,t}$ would be independent and drawn from a Normal($\boldsymbol{\alpha}_j$ + X*$\boldsymbol{\beta}$) distribution. For the pooled model, each observation $y_{j,t}$ would be independent and drawn from a Normal($\boldsymbol{\alpha}_0$ + X*$\boldsymbol{\beta}$) distribution. All the chosen priors, for both the pooled and hierarchical model, are included in the figures below.



County $\sim$ j = 1, 2,…, 88

Year $\sim$ t = 1,2,…,7

$y_{j,t}$ = Log Crime rate for County j in Year t

α_0 = Log Crime Rate for Baseline County

PA_j= Log Odds of Arrest in County j, year t

PP_j= Log Odds of Prison Sentence in County j, year t

POL_j= Log if Police Density in County j, year t

POP_j = Log of Population Density in County j, year t

$\alpha_0 \sim \mathcal{N}(-3, 1)$

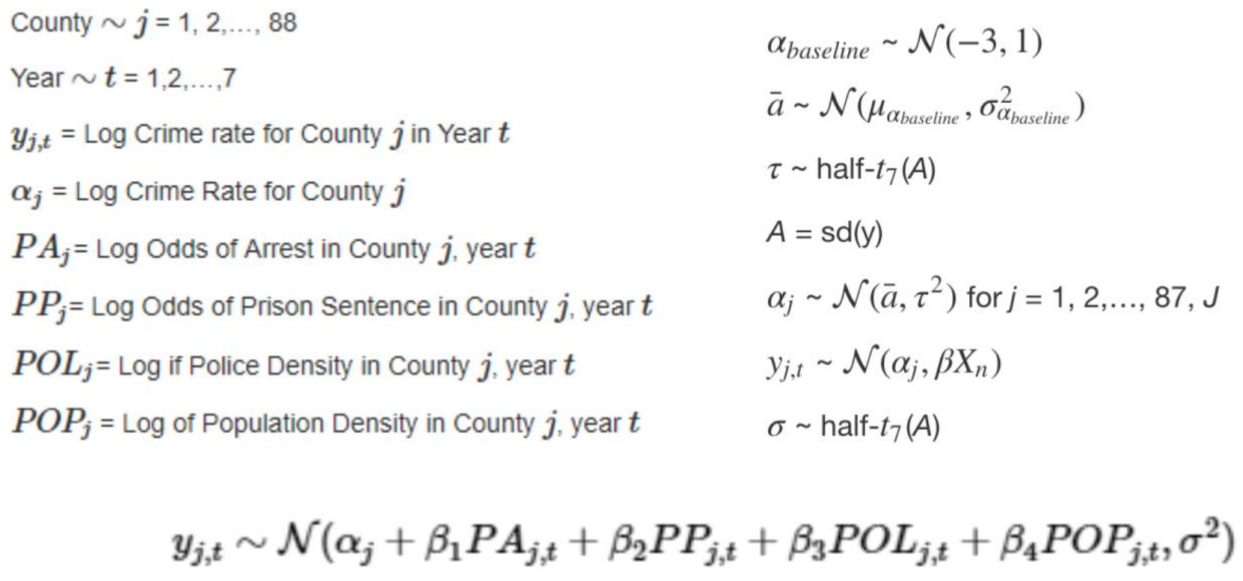
$A = \mathrm{sd}(y)$

$y_{j,t} \sim \mathcal{N}(\alpha_0, \beta X_n)$

$\sigma \sim \text{half-}t_7(A)$

$$y_{j,t} \sim \mathcal{N}(\alpha_0 + \beta_1 PA_{j,t} + \beta_2 PP_{j,t} + \beta_3 POL_{j,t} + \beta_4 POP_{j,t}, \sigma^2)$$

Figure 1. Priors for the Pooled Intercept Model



County $\sim j$ = 1, 2,…, 88

Year $\sim t$ = 1,2,…,7

$y_{j,t}$ = Log Crime rate for County j in Year t

α_j = Log Crime Rate for County j

PA_j= Log Odds of Arrest in County j, year t

PP_j= Log Odds of Prison Sentence in County j, year t

POL_j= Log if Police Density in County j, year t

POP_j = Log of Population Density in County j, year t

$\alpha_{baseline} \sim \mathcal{N}(-3, 1)$

$\bar{a} \sim \mathcal{N}(\mu_{\alpha_{baseline}}, \sigma^2_{\alpha_{baseline}})$

$\tau \sim$ half-$t_7(A)$

A = sd(y)

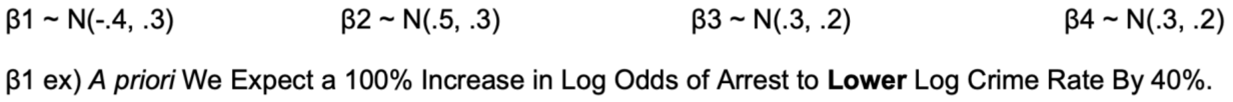
$\alpha_j \sim \mathcal{N}(\bar{a}, \tau^2)$ for j = 1, 2,…, 87, J

$y_{j,t} \sim \mathcal{N}(\alpha_j, \beta X_n)$

$\sigma \sim$ half-$t_7(A)$

$$y_{j,t} \sim \mathcal{N}(\alpha_j + \beta_1 PA_{j,t} + \beta_2 PP_{j,t} + \beta_3 POL_{j,t} + \beta_4 POP_{j,t}, \sigma^2)$$

Figure 2. Priors for the Hierarchical Model



β1 ~ N(-.4, .3) β2 ~ N(.5, .3) β3 ~ N(.3, .2) β4 ~ N(.3, .2)

β1 ex) *A priori* We Expect a 100% Increase in Log Odds of Arrest to **Lower** Log Crime Rate By 40%.

Figure 3. Specified Prior Coefficients for Explanatory Variables

Prior predictive checks were carried out by drawing parameters from the specified prior distributions in order to see if the specified priors were reasonable choices for modeling. Simulated x values of {-2.5, 2.46,..., 1} were used in these predictions. Following the checks, it was verified that the simulated data closely matched the observed data, and the priors were reasonable to use in the model.

Results

When looking at the results of the two different models, the hierarchical model was a better overall fit. To test the models, both predicted the log crime rate, within two standard deviations, for all 88 counties in 1987. Consequently, the hierarchical model captured 83/88 or 94.32% of the out-of-sample predictions in 95% prediction intervals, with an R-Squared value of 0.7373. The pooled model captured 80/88 or 90.09% of out-of-sample predictions in 95% prediction intervals, with an R-Squared value of 0.5214. By capturing an additional three counties, the hierarchical model predicted log crime rates in 1987 better than the pooled model. Specifying a county-specific log crime rate intercept proved out to be beneficial in predicting log crime rates in that county. Furthermore, with the hierarchical model having the higher R-Squared value of the two, the model is superior in explaining log crime rate variation with these variables.

Below are the results of the two model's out-of-sample predictions. A green triangle represents a correct prediction for county *j* and a red triangle represents an incorrect prediction for county *j*.

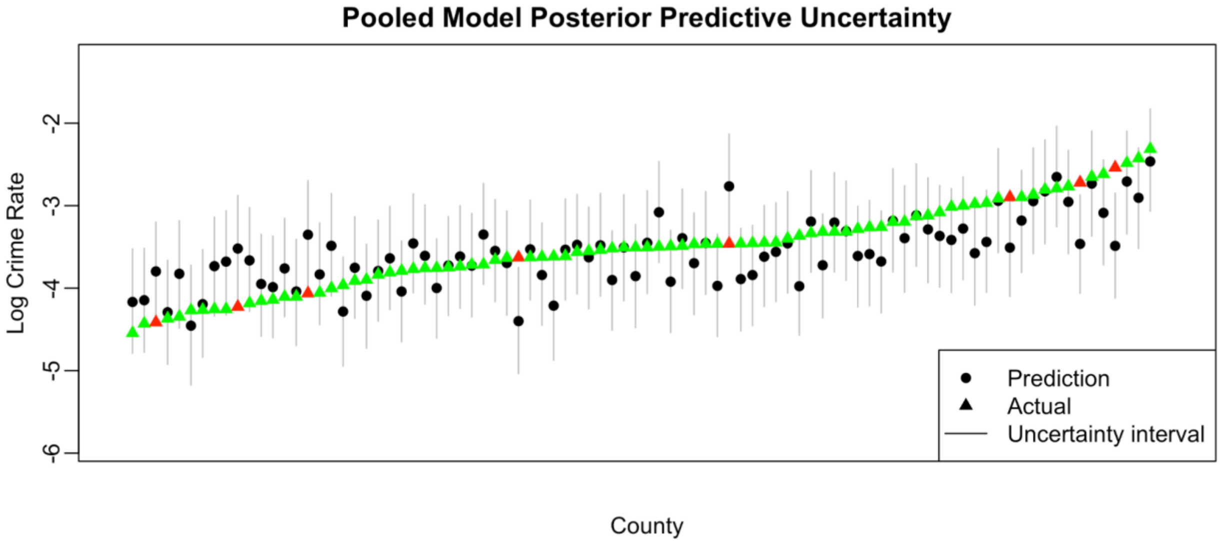


Figure 4. Log Crime Rate Predictions vs. Observed in 1987 for the Pooled Model

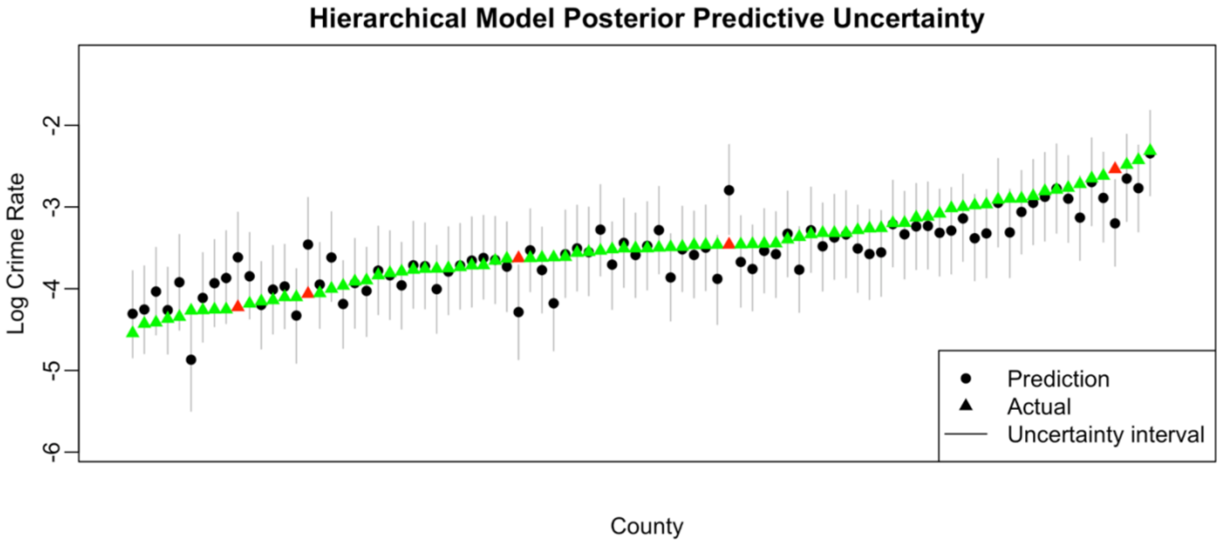


Figure 5. Log Crime Rate 1987 Predictions vs. Observed in 1987 for the Hierarchical Model

The posterior pooled model followed:

$$y_{j,t} \sim \mathcal{N}(-3.41 - 0.17 * PA_{j,t} + 0.35 * PP_{j,t} + 0.43 * POL_{j,t} + 0.33 * POP_{j,t}, 0.31^2)$$

While the posterior for the hierarchical model followed:

$$y_{j,t} \sim \mathcal{N}(\alpha_j - 0.17 * PA_{j,t} + 0.35 * PP_{j,t} + 0.44 * POL_{j,t} + 0.33 * POP_{j,t}, \sigma^2)$$

The four coefficients, in both models, are the means of their posterior distributions included below.

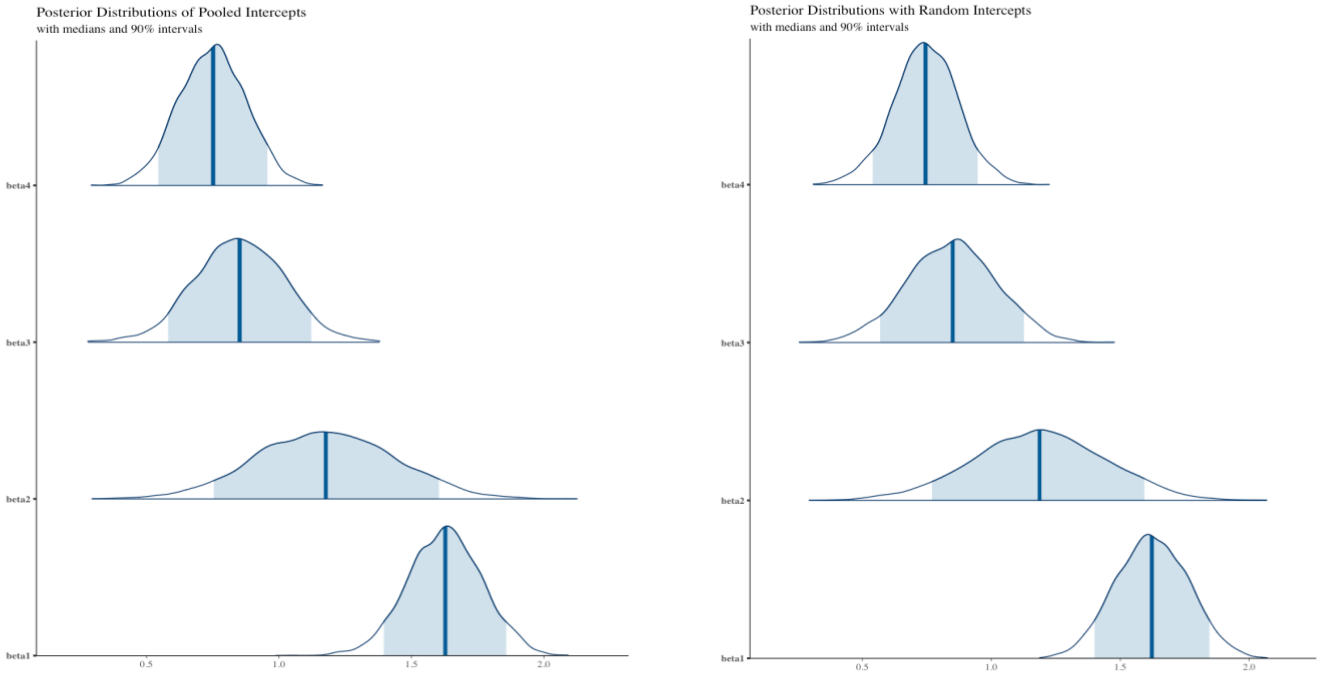


Figure 6. Coefficient Posterior Distributions for the Pooled Model (left) and Hierarchical Model (right)

Lastly, shown in the plot below, are all 88 county’s intercepts for their log crime rate if all other coefficients in the model are set to zero.

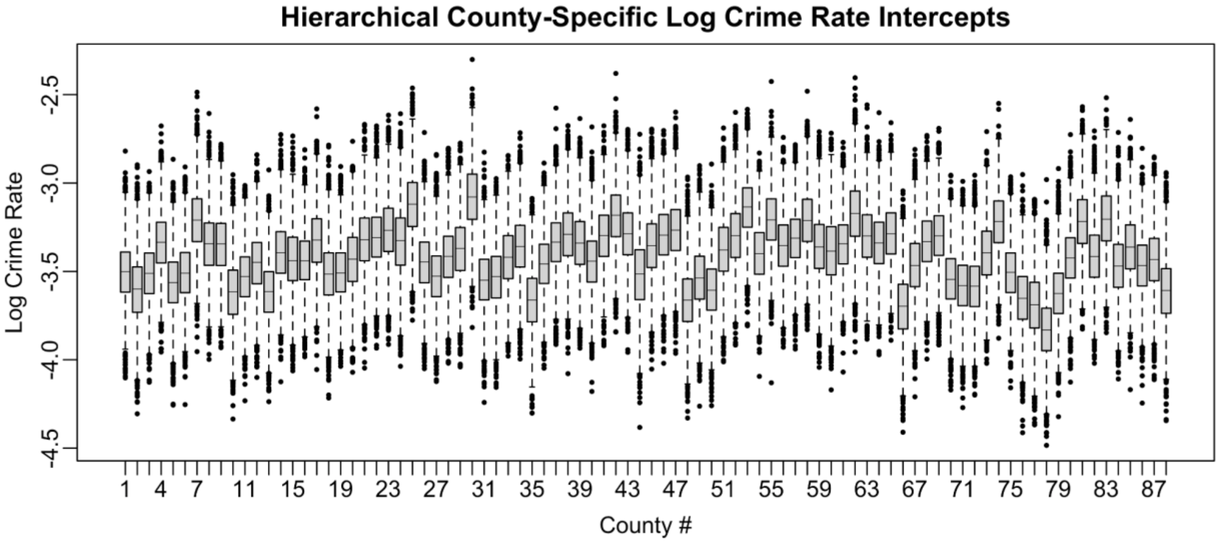


Figure 7. County-Specific Intercepts for Log Crime Rate in Hierarchical Model

Conclusion

Looking at our results, it was found that in North Carolina, during 1981-1987, the only deterrent on the crime rate was the probability of arrest. If a county's log odds of arrest went up, the log crime rate went down. For example, with a 100% increase in log odds or arrest, the log crime rate is expected to decrease 17%. Effectively, as one was more likely to get arrested in an area, that area's crime rate would decrease. Additionally, as the probability of a prison sentence increased or the density of the police or population increased, the crime rate would be expected to increase. Intuitively it makes sense that as there are more people in an area, there would be more crime. However, it's not intuitive that as there are more police in an area, there is higher crime. Going back to the calls to defund the police, if the log density of the police force in an area increased 100%, the log crime rate would increase by 44%. However, this could be interpreted two ways; mainly that an area has more police due to trying to cut down previous high crime rates, or a place has higher crime due to there being more police. Given the increasing calls to defund the police in recent years, our model would say that there may be merit to these calls, as the densest areas in terms of police, had the highest crime.

A major limitation in our analysis was the absent metadata linking each county number to the real county in North Carolina. As a result, we couldn't attach our quantitative findings about county j to qualitative data about county j. There may be several qualitative factors about prominent counties in North Carolina that could've been utilized in our research if the data included more than a county number. For example, imagine that Wake County during the 1970s saw a massive increase in crime. In response, the county may have increased its number of police, and the change in crime rate may have lagged. Without qualitative data, there is no way to determine if denser police forces have always been dense, in a county, or if they are only dense as of recent. In addition, as we set out to research the actuality of defunding the police, our ability to answer questions happening today is limited by the data being from the 1980s. As a result, what impacted the crime rate over 30 years ago may be vastly different from what impacts the crime rate today. Connecting our model to the present faces the same time limitations for all predictors. Lastly, while we wanted to investigate the probability of conviction on crime rate, given the erroneous data observations, we were unable to include this in our analysis. This is unfortunate, as we believed an increase in probability of conviction would be a deterrent factor in committing crimes. Since our final model included only one deterrent, having a balance between deterrents and non-deterrents may have produced a better fit model.

References

1. https://www.brookings.edu/blog/fixgov/2020/06/19/what-does-defund-the-police-mean-and-does-it-have-merit/
2. https://www.pewresearch.org/fact-tank/2021/10/27/what-we-know-about-the-increase-in-u-s-murders-in-2020/
3. https://nij.ojp.gov/topics/articles/five-things-about-deterrence
4. https://rdrr.io/cran/Ecdat/man/Crime.html
5. https://worldpopulationreview.com/us-city-rankings/most-dangerous-cities-in-the-us