UNITED STATES MILITARY ACADEMY

FINAL PROJECT REPORT

MA478: GENERALIZED LINEAR MODELS

SECTION H28

COL CLARK, NICHOLAS

BY

CDT LUCAS VILLANTI '24, CO G4

WEST POINT, NEW YORK

06 MAY 2024

LGV I CERTIFY THAT I HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT I USED TO COMPLETE THIS ASSIGNMENT AND THAT I ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

_____ I CERTIFY THAT I DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTANCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURE: Lucas Garrett Villanti

Modeling the Influence of Socioeconomic and Temporal Factors on Burglary Rates in Chicago

CDT Lucas Villanti

May 6, 2024

Abstract

Chicago's nickname "Chiraq," derived from its high crime rates, underscores the city's struggle with burglaries influenced by socioeconomic disparities. This study examines burglary patterns across Chicago neighborhoods, focusing on the impact of socioeconomic factors, seasonal changes, and spatial-temporal variability over a seven-year period. Utilizing Chicago crime data, we applied three statistical models to analyze the relationships between crime rates and various predictors: a basic Poisson regression model, a mixed-effects model incorporating random effects, and a Bayesian Hierarchical Poisson Model referred to as the Chiraq model. The Poisson model, while foundational, exhibited signs of underfitting and overdispersion. The mixed-effects model showed improved fit and provided insights into the random variations across different neighborhoods and years. However, it was the Chiraq model, with its advanced Bayesian hierarchical structure, that demonstrated a superior balance of complexity and fit, evidenced by its lower Deviance Information Criterion (DIC) and more favorable marginal log-likelihood compared to the other models. This model effectively captured the nuanced effects of unemployment, wealth, and temporal trends on burglary rates, suggesting that higher unemployment and wealth levels are significantly associated with increased burglary rates. Despite some limitations in predicting future rates and potential overdispersion issues, the Chiraq model's comprehensive approach offers valuable insights for targeted crime prevention strategies, emphasizing the need for continued refinement and ethical considerations in future research.

Key words: Chicago, burglary patterns, socioeconomic factors, Bayesian hierarchical model, Mixed Effects Model

1 Introduction

Chicago, often referred to in hip-hop as ”Chiraq,” earned this nickname due to its history of crime and violence. Although controversial, the term highlights the city's ongoing struggle with these issues, particularly burglaries. The incidence of burglaries in Chicago has historically fluctuated, with factors like poverty, inequality, and limited access to education and job opportunities significantly contributing to crime's persistence. However, the city's challenges vary by neighborhood. Research shows a clear spatial distribution of burglaries, emphasizing the role of local socioeconomic factors. Areas with higher poverty and unemployment rates typically experience more burglaries, in contrast to more affluent neighborhoods. This disparity underscores the dynamic nature of crime patterns, requiring a detailed understanding of their causes. In this report statistical modeling is a crucial approach to dissecting the complex factors influencing burglary rates, thereby informing targeted crime prevention and community safety measures Monroe [2015].

2 Literature Review

Literature on burglary patterns in Chicago has examined various spatial and temporal dimensions to understand crime occurrences and inform policing strategies. Luo [2017] conducted research spanning 2006 to 2016, focusing on census blocks and police beat areas, integrating hourly, daily, and monthly temporal dimensions. Their analysis identified burglary hot spots and observed shifts over time and space, aiding in targeted policing efforts.

Similarly, Bernasco et al. [2017] investigated street robbery location choices in Chicago, considering temporal variations in location attributes. Their study debunked assumptions about the opportunistic nature of street robbery, highlighting consistent preferences for proximity to cash economies and transit hubs, with only high schools showing significant influence during school sessions.

Comparatively, our study also delves into burglary patterns in Chicago, focusing on location attributes. Like Luo [2017], we consider temporal dimensions, although at a yearly and monthly level. Additionally, while Bernasco et al. [2017] examined cash economies and school presence, we explore the wealth of location and seasonal effects akin to school presence but different. Despite our study's limitations, it aligns with prior research by investigating similar covariates to identify burglary trends. However, our temporal scope differs from both Luo [2017] and Bernasco et al. [2017], as we focus on yearly and monthly trends rather than hourly, daily, and monthly ones.

3 Methodology

Our problem statement for this project is given the noticable disparties in burglary rates across different Chicago neighborhoods, what model can accurately capture the impact of seasonal changes, population, and wealth index on burglary counts, and account for spatial and temporal variability. From here we established two research questions to answer:

1. What are the key socioeconomic and demographic factors that influence burglary rates in Chicago neighborhoods?
2. How do seasonal patterns, district socio-economic profiles, and wealth indices impact the spatial and temporal variability of robbery incidence in Chicago over a seven-year period?

Then our hypothesis from this questions are:

1. **H1:** Higher levels of poverty and unemployment in a Chicago neighborhood are positively associated with higher burglary rates.
2. **H2:** The application of a mixed-effects model, accounting for both fixed (socioeconomic status, unemployment, etc.) and random (seasonal variations) effects, will provide a more accurate prediction of burglary rates across Chicago neighborhoods than models considering only fixed effects.

3.1 Generalized Linear Models

We came up with three models to answer this problem statement. The first one is using a basic Poisson regression model, second being a mixed effects model accounting for location and month as a random effect, and then our last model which we call our Chiraq model which is a Bayesian Hierarchical Poisson Model.

3.1.1 Poisson Regression Model

To evaluate the impact of various socio-economic factors on crime rates, we applied a Poisson regression model. This approach is appropriate for modeling count data, where the response variable represents the number of crime events. The model is expressed as follows:

$$\begin{aligned}\log(\lambda_i) = \beta_0 + \beta_1 \cdot \text{unemp}_i + \beta_2 \cdot \text{ym}_i + \beta_3 \cdot \text{wealth}_i \\ + \beta_4 \cdot \text{year}_i + \beta_5 \cdot \text{month\_num}_i\end{aligned} \tag{1}$$

where λ_i denotes the expected count of crimes in the i-th neighborhood which is ID. The coefficients β_1 through β_5 measure the effect of each variable on the log of the expected crime count.

3.1.2 Mixed Effects Model

In our study, we extend the basic Poisson regression model to a mixed effects model to better handle the hierarchical structure of our data. This model incorporates random effects to account for variations within clusters (in this case, individual IDs(Location) and years), which cannot be explained by the observed variables alone. The model is specified as follows:

$$\begin{aligned}\log(\lambda_{ij}) &= \beta_0 + \beta_1 \cdot \text{unemp}_{ij} + \beta_2 \cdot \text{ym}_{ij} + \beta_3 \cdot \text{wealth}_{ij} \\ &\quad + \beta_4 \cdot \text{month\_num}_{ij} + u_i + v_j \\ Y_{ij} &\sim \text{Poisson}(\lambda_{ij})\end{aligned}$$

where λ_{ij} denotes the expected count of crimes for each observation i in year j. The model includes random intercepts u_i and v_j for each ID and year, respectively. The random effects are assumed to be normally distributed, capturing the unobserved heterogeneity across IDs and years.

3.1.3 Chiraq Model

To analyze the complex data structure and account for potential dependencies across hierarchical levels, we utilize a Bayesian hierarchical model with the Integrated Nested Laplace Approximations (INLA) method. The model is formulated to include both fixed and random effects, capturing variations at different hierarchical levels:

$$\begin{aligned}
n_{it} &= \beta_0 + \beta_1 \cdot \text{unemployment}_i + \beta_2 \cdot \text{wealth}_i + \beta_3 \cdot \text{year}_i \\
&\quad + u_i + v_{t,\text{year}} + w_{t,\text{month}} \\
Y_{it} &\sim \text{Poisson}(\lambda_{it}) \\
n_{it} &= \log(\lambda_{it})
\end{aligned}$$

λ_{it} is the expected count of crime rates in the i-th district at time t, represented by Y_{it}. The model intercept is β_0, with β_1, β_2, and β_3 serving as the coefficients for the unemployment rate, wealth index, and year, respectively. u_i indicates the iid random effect for district i, while $v_{t,\text{year}}$ is the random effect for year t, modeled as a random walk of order 1 (RW1). The random effect for month t, $w_{t,\text{month}}$, is also modeled with RW1 and assumes a cyclic structure.

4 Data Description/ Preprocessing

Our dataset originates from the Chicago crime records available on the nick3703 GitHub repository. It encompasses count data for burglaries across 552 distinct locations in Chicago, with observations recorded monthly from January 2010 to December 2015. Additional data attributes include wealth indices, unemployment rates, and population.

To facilitate analysis, we consolidated multiple CSV files sourced from the GitHub repository into a single comprehensive dataset. This integration was followed by the standardization of date formats to ensure consistency across records. We subsequently restructured the dataset into a long format, which allowed for more efficient handling of temporal variables such as months. Lastly, we introduced a 'crime rate' variable, calculated as the number of crimes per month per location, to serve as a primary response variable.

5 Data Exploration Analysis

To gain more understanding of crime patterns over time and their correlation with other variables, we have presented a series of visual analyses depicted in Figures 1 through 3. Figure 1 displays a time series plot of mean burglary rates in Chicago, capturing data from January 2010 through December 2015. This plot is particularly informative as it illustrates the general trend of burglary rates across the period, emphasized by a fitted trend line which highlights a decrease over time. Highlighting the possibility of considering year into our models as not just a random effect but as a fixed effect. In Figure 2, a density plot showcases the distribution of total crime rates, showing a right skewed plot and that some areas have higher crime than others. Finally, Figure 3 explores the relationship between total crime rates and population size, with each point colored to represent different unemployment rates. This showcases that total crime in an area is related to socioeconomic factors such as unemployment in the area. In that there is a trend for crime to happen in wealthier areas.

6 Model Summary Results

6.1 Poisson Regression

Table 1: Poisson Regression Model Coefficients

Predictor	Estimate	Std. Error	z-value	P-value
Intercept	0.1979	0.0131	15.078	$< 2 \times 10^{-16***}$
Unemployment Rate (unemp)	0.0241	0.0047	5.156	$2.52 \times 10^{-7***}$
Young Male Population (ym)	0.0023	0.0001	21.579	$< 2 \times 10^{-16***}$
Wealth Index (wealth)	0.1319	0.0047	28.050	$< 2 \times 10^{-16***}$
Year (year)	-0.1651	0.0027	-60.164	$< 2 \times 10^{-16***}$
Month Number (month_num)	0.0328	0.0013	24.655	$< 2 \times 10^{-16***}$

When deciding variables to pick for this model, we chose unemployment, young male population, wealth index, year, and month as our covariates for looking at burglary rates. For our first model, we wanted to include all the variables and then include year and month as variables to encompass the trend of summers having an increase in crime and the decreasing trend in overall burglaries over this 5-year period.

The Poisson regression model provides a detailed look at the factors influencing crime rates. The intercept indicates a baseline crime rate when other predictors are held constant. Specifically, a one-unit increase in unemployment is linked to a 2.4% rise in the expected count of crimes, suggesting that higher unemployment may exacerbate crime rates. Similarly, an increase in the proportion of young males slightly raises crime rates, reflecting this demographic's potentially higher involvement in criminal activities. Moreover, wealthier areas experience higher crime rates, possibly due to more attractive targets for criminal activities. The model also captures a significant yearly decrease in crime rates, indicating effective crime reduction strategies or societal changes over time. Additionally, the month number effect reveals seasonal trends, with crime rates increasing as the year progresses, highlighting the influence of time-based variables on crime dynamics. This comprehensive analysis underscores the complex interplay of socio-economic and temporal factors in shaping crime patterns.

6.2 Mixed Effects Model

Table 2: Mixed Effects Poisson Regression Model Coefficients

Predictor	Estimate	Std. Error	z-value	P-value
Intercept	-0.3444	0.1250	-2.755	0.00587 **
Unemployment Rate (unemp)	0.0099	0.0221	0.448	0.65398
Young Male Population (ym)	0.0025	0.0006	4.613	0.00000397 ***
Wealth Index (wealth)	0.1385	0.0241	5.742	0.00000000938 ***
Month Number (month_num)	0.0328	0.0013	24.723	$< 2 \times 10^{-16}$ ***

Table 3: Selected Model Statistics for the Mixed Effects Model

Parameter	Value
AIC	112083.0
Log Likelihood	-56034.5
Deviance	112069.0
Number of Observations	39744

6.2.1 Random Effects

The model includes random effects to account for unobserved variability across different IDs and years. These effects are summarized as follows:

- **ID:** Variance = 0.24857, Standard Deviation = 0.4986, indicating significant variability in crime rates among different IDs, reflecting individual differences or unique characteristics of locations.
- **Year:** Variance = 0.08548, Standard Deviation = 0.2924, suggesting variability over years, which could capture changes in crime rates due to policy shifts, economic cycles, or other temporal factors.

6.2.2 Interpretation of Model Results

The mixed effects model extends our analysis by incorporating random effects for IDs and years to account for inherent variability that cannot be captured by fixed effects alone. This model allows us to understand how individual and temporal variations influence the observed crime rates, besides the fixed effects of unemployment, young male population, wealth, and month.

The coefficients of the mixed effects model provide insights into factors influencing crime rates while accounting for random variations across different IDs and years. The negative intercept suggests a baseline decrease in crime rates when other variables are zero. A slight, non-significant increase in crime is associated with unemployment, as indicated by its coefficient and large p-value, suggesting weak or no evidence of its impact. The presence of young males in the population and wealth are both significantly associated with increases in crime rates, highlighting socioeconomic impacts. The significant positive coefficient for month suggests a seasonal pattern in crime occurrence, with higher rates later in the year. These results demonstrate the complex dynamics of crime influenced by both socio-economic conditions and inherent variabilities across time and individuals.

6.3 Chiraq Model

Table 4: Fixed Effects for the Chiraq Model

Effect	Mean	SD	2.5% Quantile	Median	97.5% Quantile
Intercept	0.349	0.135	0.075	0.350	0.622
Unemployment Rate (unemp)	0.022	0.022	-0.022	0.022	0.066
Wealth Index (wealth)	0.186	0.022	0.142	0.186	0.229
Year (year)	-0.142	0.053	-0.250	-0.143	-0.034

Table 5: Overall Statistics for the Chiraq Model

Statistic	Value
DIC (Deviance Information Criterion)	110227.27
DIC (Saturated)	51698.67
Effective Number of Parameters	539.36
Marginal Log-Likelihood	-55760.47

6.3.1 Random Effects and Model Hyperparameters

The model includes several random effects with corresponding hyperparameters to adjust the precision of these effects, reflecting their variability:

- **Year Factor**: Modeled as a random walk of the first order (RW1), with a precision mean of 113.54 and standard deviation of 84.623.

- **Month Factor**: Also modeled with RW1 but cyclic, indicating seasonal effects, with a precision mean of 47.15 and standard deviation of 18.980.

- **ID**: Independent and identically distributed effects for each ID, with a precision mean of 3.86 and standard deviation of 0.257.

6.3.2 Interpretation of Model Results

The Bayesian hierarchical model employed incorporates both fixed and random effects, using the INLA method to efficiently estimate model parameters. This model structure allows us to capture and quantify the inherent variability across IDs, months, and years, while assessing the impact of unemployment, wealth, and time on crime rates.

The model indicates that both unemployment and wealth have a measurable effect on crime rates, with wealth showing a more consistent influence across the estimates. The year effect suggests a decrease in crime rates over time, potentially reflecting successful interventions or socio-economic improvements. The precision parameters for the random effects underscore the significant variability across years and the cyclic nature of crimes across months, enhancing our understanding of how temporal factors uniquely influence crime dynamics.

7 Model Selection

7.1 Poisson Model

The Poisson Model displayed distinct curved bands in the residuals plot (Figure 4), suggesting underfitting and potential overdispersion. The "Fitted vs. Actual" plot (Figure 5) showed poor alignment of data points along the expected 45-degree line, indicating significant deviations from ideal predictions, particularly at higher fitted values.

7.2 Mixed Effects Model

Similarly, the Mixed Effects Model showed patterns of underfitting and overdispersion in the residuals plot (Figure 6). However, the "Fitted vs. Actual" plot (Figure 7) demonstrated a relatively better alignment along the 45-degree line compared to the Poisson model, suggesting improved accuracy in predicting actual values, although not entirely optimal.

7.3 Chiraq Model

Lastly, the Chiraq Model also exhibited curved bands in the residuals plot (Figure 8), consistent with the previous models, pointing to underfitting issues. The "Fitted vs. Actual" plot (Figure 9) revealed a slight improvement in alignment compared to the Poisson Model but still displayed discrepancies at higher fitted values.

7.4 Final Selection

While all three models showed signs of underfitting and potential overdispersion, the Mixed Effects Model appears to perform slightly better in terms of alignment in the ”Fitted vs. Actual” plot (Figure 7). However, when looking at the Model Statistics, The Chiraq model has a lower DIC than the AIC of the Mixed Effects model, suggesting a better overall fit when considering the balance between model complexity and goodness of fit. Also, the marginal log-likelihood of the Chiraq model is less negative than the Mixed Effects, showing that the Chiraq model might be more probable in terms of explaining the variations. Although the plots suggested that the Mixed Effects model might provide better alignment at certain ranges, the metrics favor the Chiraq model.

8 Discussion and Conclusions

8.1 Conclusions

The mixed effects and Bayesian hierarchical models have provided detailed insight into the factors influencing burglary rates, accounting for fixed and random influences across different IDs, months, and years. We found that unemployment and wealth significantly affect crime rates, with wealth consistently showing a strong correlation. The temporal analysis revealed decreased crime rates over the years and confirmed the seasonal patterns in crime occurrence. In the end, we chose the Chiraq model for its complexity and ability to incorporate fixed and random effects along with individual variation and temporal trends. It could be better due to its tendency to underestimate higher crime rates, but it is a more comprehensive and informative approach to the problem compared to the other models.

8.2 Model Limitations

The models used have highlighted significant trends and effects but have inherent limitations. The Poisson assumption that the mean and variance of crime rates are equal can lead to overdispersion issues, where the variance exceeds the mean, potentially leading to underestimating the standard errors. Furthermore, using random walks for years and months assumes smooth temporal effects, which might not effectively capture abrupt changes or nonlinear trends. Additionally, the current models have limited predictive power for future burglary rates, a significant drawback for practical applications in crime prevention strategies.

8.3 Future Work

Future research should consider several avenues to build on the current analysis and address these limitations. Exploring other model families, such as Negative Binomial models, can handle overdispersion better than Poisson models. Incorporating additional variables such as school vacation periods might provide further insights into temporal crime patterns, especially about shifts in population density and youth activity. Enhancing the model's predictive capabilities by integrating more dynamic variables and using machine learning techniques could lead to better forecasting of crime rates. These steps will help refine the models and increase their utility in real-world applications, contributing to more effective crime prevention and policy-making.

8.4 Ethical Considerations

When looking at sensitive data such as crimes, it is important to consider ethical factors before sending the results to the public. The most important thing to consider is the interpretation of the results. The goal of conducting research like this is to make scientific efforts to help promote public safety and address the root cause of the crime without reinforcing or communicating biases or stereotypes. Our selected model, the Chiraq Model, has its limitations and shows the need for further refinement. As research continues, it is essential to prioritize ethical research conduct and strive for positive and informative solutions rather than negative ones.

9 References

References

Wim Bernasco, Stijn Ruiter, and Richard Block. Do street robbery location choices vary over time of day or day of week? a test in chicago. *Journal of research in crime and delinquency*, 54(2):244–275, 2017.

Jun Luo. Multi-spatiotemporal patterns of residential burglary crimes in chicago: 2006-2016. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:193, 2017.

Sydney Monroe. Chi-raq and the myth of chicago gang wars. *The New York Times Magazine*, Dec 2015. URL `https://www.nytimes.com/2015/12/07/magazine/chi-raq-and-the-myth-of-chicago-gang-wars.html`.

OpenAI. Chatgpt, 2024. Retrieved May 06, 2024, from OpenAI. Assistance to the Author. CHAT GPT helped me first with the format of my paper laying out the foundation for the sections I created. I also plugged in my models from Rstudio to created equations and summary tables that I used in my report. In addition, I used it to help me with my model selection process. I did not really know how to interept the fitted vs residuals plots so it gave me gudiance on that. Finally, I used chatgpt to write my abstract. I prompted it with my entire paper and it give me a less than 250 word abstract. The link to my conversations are in the Appendix code.

10 Figures

10.1 Time Series of Mean Burglary Rates

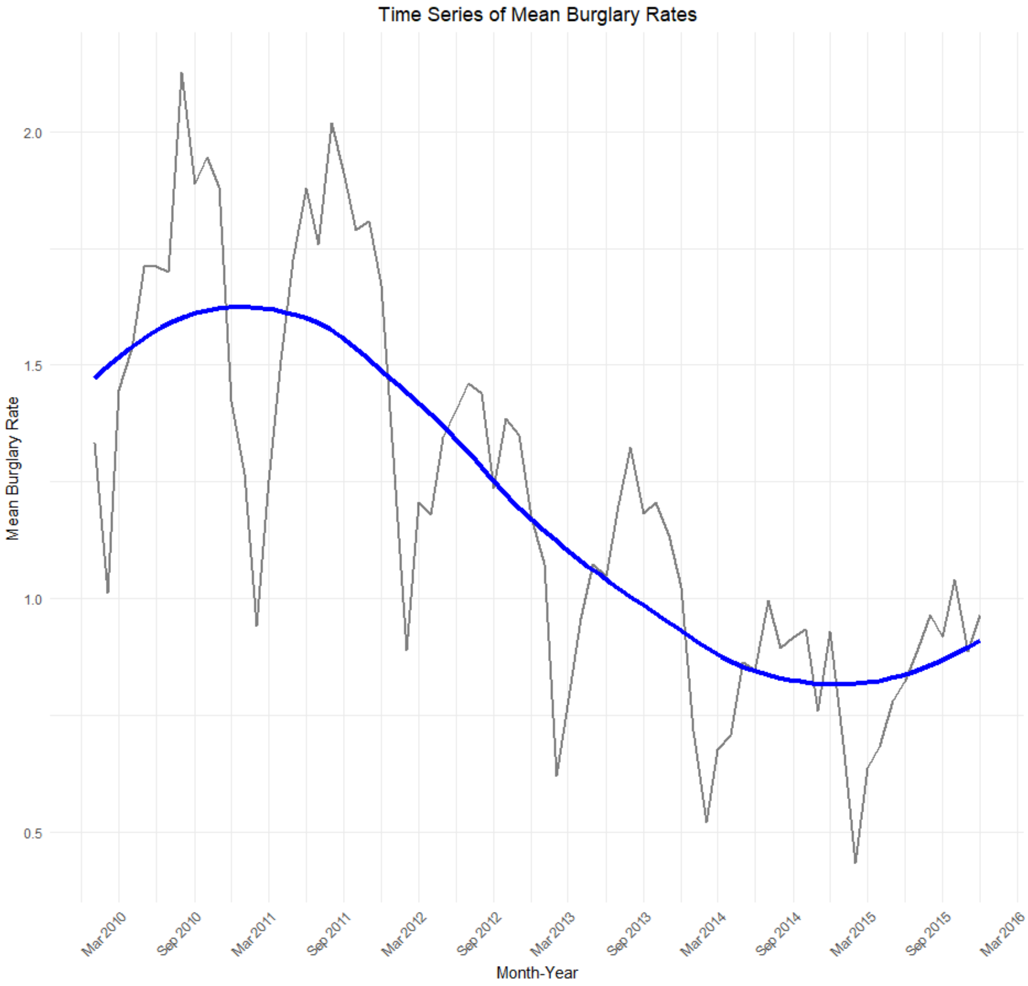

Figure 1: Time Series plot showing the mean burglary rates in Chicago from January 2010 to December 2015. A trend line indicates the overall movement across the period.

10.2 Density Plot of Total Crime Rates

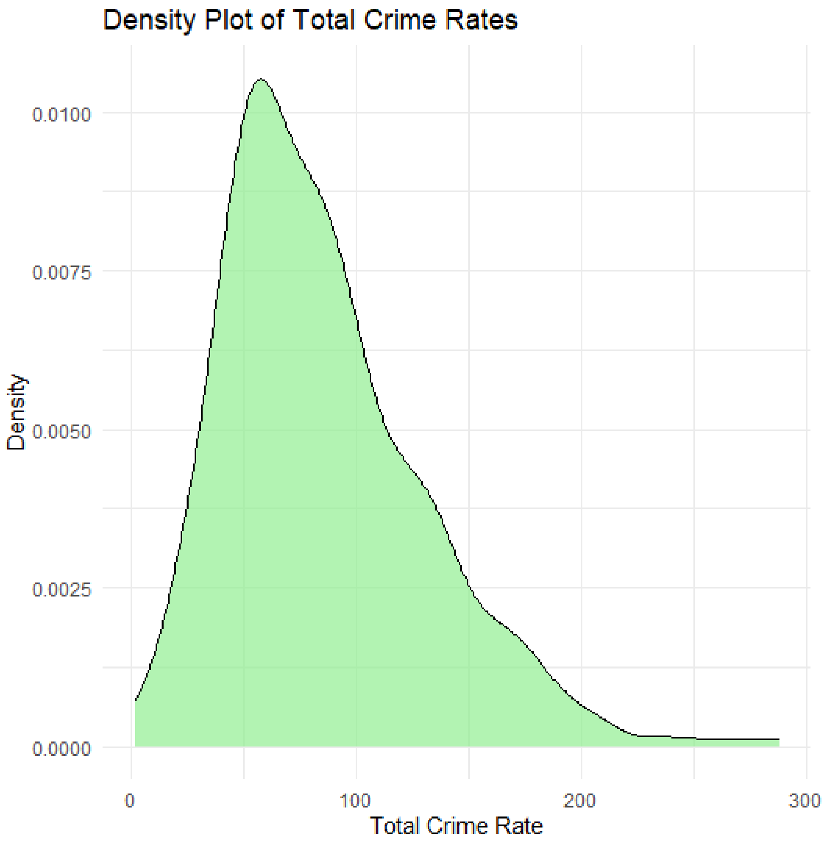

Figure 2: Density plot of total crime rates highlighting the distribution and spread of crime data points over the study period.

10.3 Scatter Plot of Total Crime vs. Population

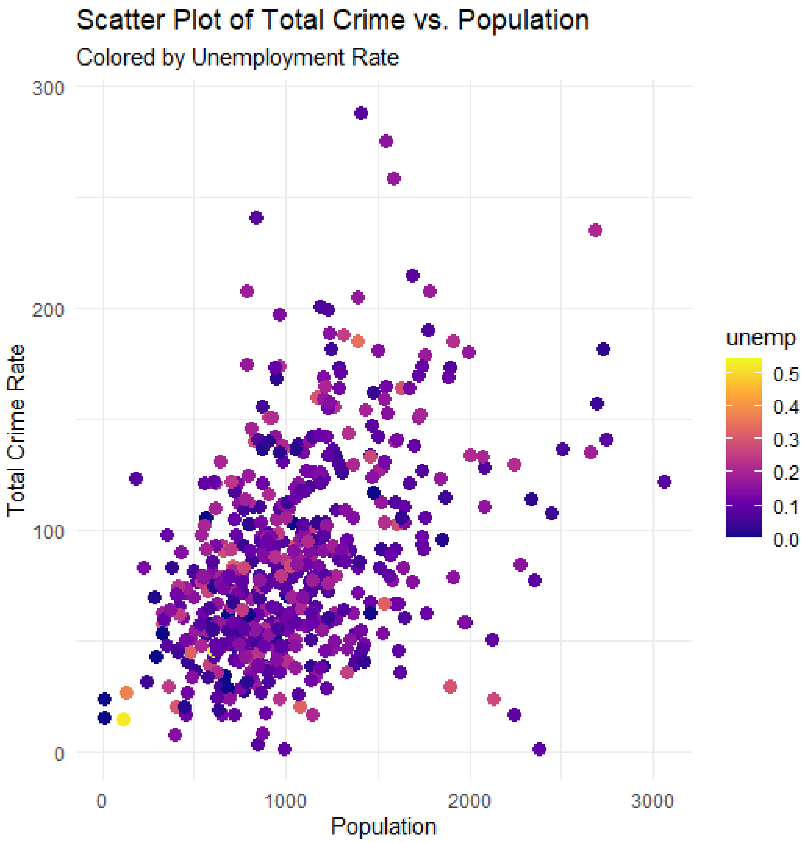

Figure 3: Scatter plot of total crime rates versus population size, colored by unemployment rate, showing the relationship between population density and crime occurrence.

10.4 Poisson Model

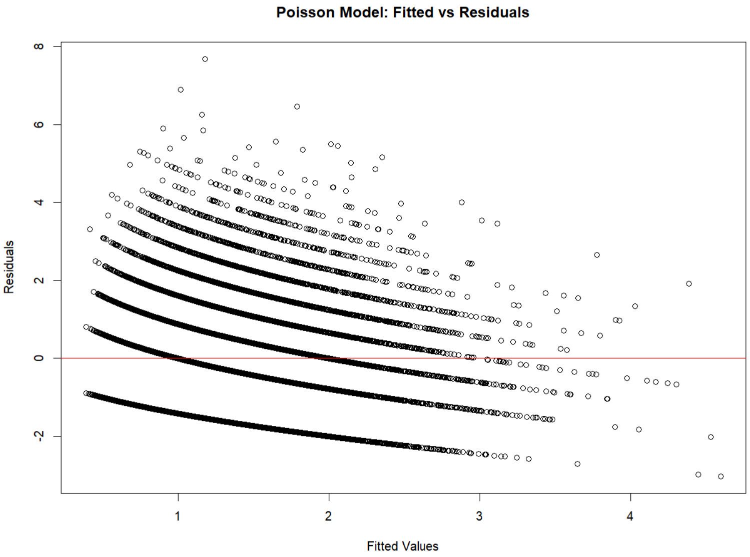

Figure 4: Poisson Model: Fitted vs Residuals

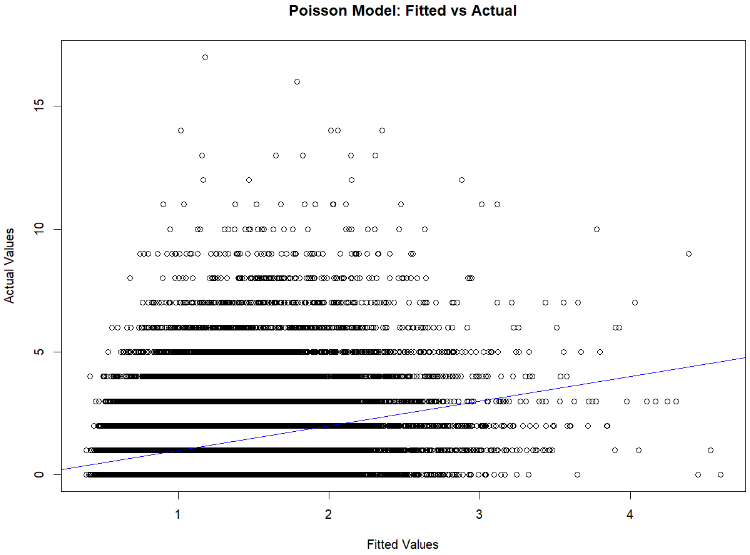

Figure 5: Poisson Model: Fitted vs Actual

10.5 Mixed Effects Model

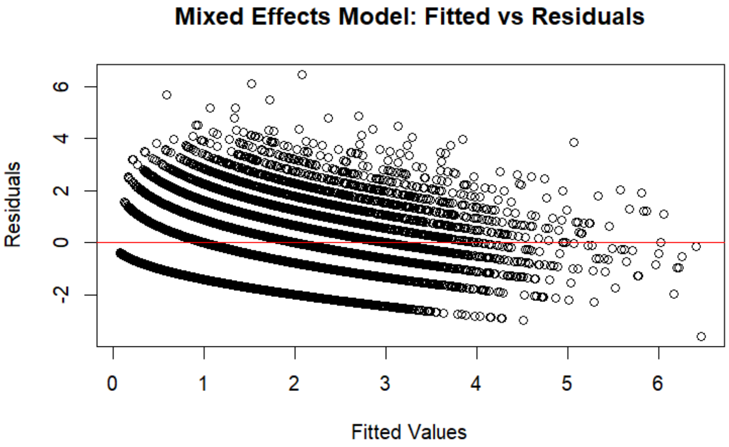

Figure 6: Mixed Effects Model: Fitted vs Residuals

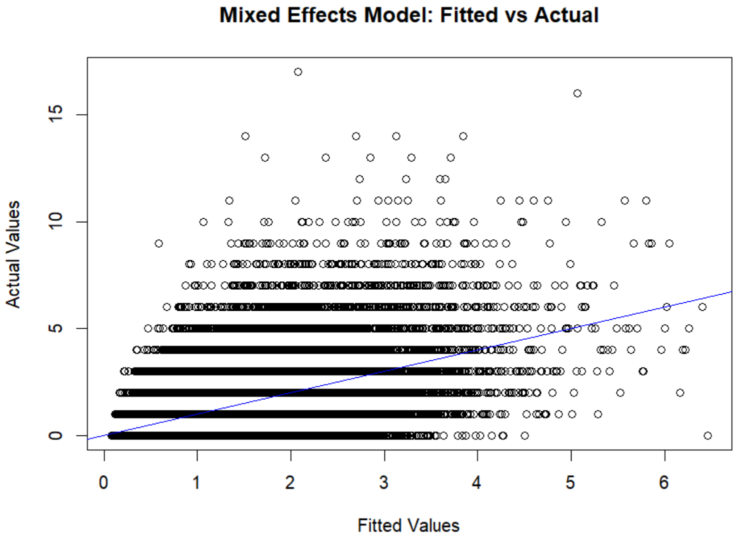

Figure 7: Mixed Effects Model: Fitted vs Actual

10.6 Chiraq Model

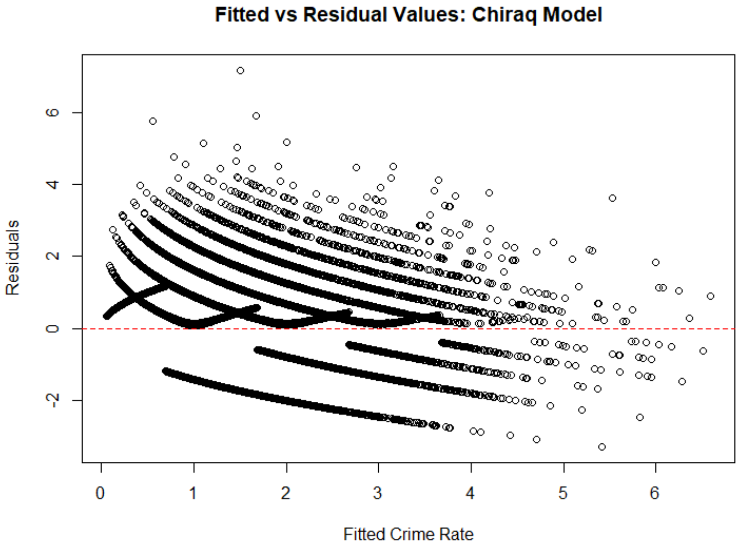

Figure 8: Chiraq Model: Fitted vs Residual Values

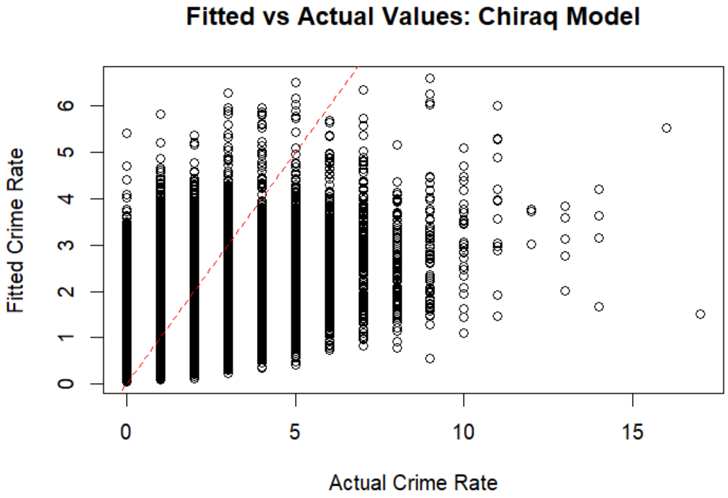

Figure 9: Chiraq Model: Fitted vs Actual Values

11 R Code

For further details on our project, please visit the **Project Repository** on GitHub:

Lucas and Garrett's Project Repository

For a detailed log of discussions and prompts used during the project's development, refer to the **ChatGPT Interaction Log** OpenAI [2024]:

CDT Lucas Villanti CHATGPT Log