UNITED STATES MILITARY ACADEMY

PROJECT REPORT

MA478: GENERALIZED LINEAR MODELS

SECTION H2-4

COL CLARK, NICHOLAS

By

CADET SAMIN KIM ‘24, CO A2

WEST POINT, NEW YORK

08 MAY 2024

SK______ MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

______ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: KIM.SAMIN.1594266885 Digitally signed by KIM.SAMIN.1594266885
Date: 2024.05.08 00:45:31 -04'00'

Influence of Socioeconomic and Demographic Factors on Chicago Burglary Rate

CDT Samin Kim

May 8, 2024

Abstract

This study investigates the correlations between socioeconomic and demographic factors and burglary rates across 552 census block groups in Chicago, utilizing data from 2010 to 2015. Through the application of generalized linear models, this research identifies significant predictors of burglary occurrences, with a particular focus on the impacts of wealth disparities, unemployment rates, and demographic indicators such as the proportion of young males. Additionally, environmental influences, notably temperature, are examined to understand their effects on seasonal variations in crime rates. The findings indicate that both the young male population and higher temperatures correlate strongly with increased burglary rates, suggesting that these factors significantly influence the fluctuation of crime rates throughout the year. This study not only enhances the understanding of the dynamics influencing burglary rate but also supports policymakers and community leaders in implementing effective strategies to mitigate burglary rates. Further research is suggested to explore additional causal relationships and refine predictive models.

1 Introduction

Chicago, a city with complex socioeconomic dynamics, has struggled with reducing the crime rate. Over the decades, the city has seen fluctuating crime rates, with burglaries being a significant concern for public safety. The landscape of crime in Chicago continues to evolve, showing patterns and challenges that require further research.

Recent analyses, including a 2023 report by CBS Chicago, indicate that while some crimes decreased during the pandemic due to reduced public presence, other crimes, such as motor vehicle thefts and robberies, surged as life returned to normalcy. Notably, burglaries have reached historic lows, yet the nature of these crimes has shifted, with increased break-ins at commercial establishments rather than residences. This shift shows the changing dynamics of urban crime, which seem to be influenced by complex socioeconomic factors, including unemployment, wealth disparity, and even seasonal changes.[1]

Understanding these dynamics is critical not just for academic inquiry but for practical policymaking. This research focuses on how various socioeconomic and demographic factors, such as wealth levels, unemployment rates, and the demographics of neighborhoods, specifically the proportion of young males, contribute to the incidence of burglaries across Chicago's 552 Census block groups. By employing generalized linear models, this research seeks to offer an understanding of the predictors of burglary rates, providing actionable recommendations that could help create effective crime prevention strategies.

2 Literature Review

Identifying and understanding the impacts of multiple factors on the burglary rate is critical for developing effective preventive strategies and policies. Burglary significantly impacts community safety and imposes considerable economic and psychological costs on individuals and societies. This literature review section focuses on previous studies that analyze the relationship between socioeconomic factors, environmental conditions, and burglary rates.

Socioeconomic and Demographic Impact on Crime

Chang's (2011) and Rosenfeld & Messner's (2009) research exemplify the relationship between socioeconomic factors and urban crime, indicating that socioeconomic conditions play a crucial role in shaping the criminal landscape of urban areas. Chang's study shows the impact of economic and spatial factors, such as the accessibility and configuration of urban spaces, on burglary rates.[2] Similarly, Rosenfeld and Messner's analysis highlights how economic trends correlate with declines in crime, suggesting a direct link between broader economic health and crime rates.[3] These findings provide insights to include such predictors in analyzing Chicago's burglary rates.

Environmental Factors and Their Impact on Crime

The influence of environmental factors, particularly temperature, on crime has also been a focus of research. The study by Farrell and Pease (1994) shows seasonal trends in crime rates, with temperature being a significant predictor.[4] This goes with the routine activities theory, which argues that the opportunity for crime increases when people are more likely to be away from their homes, such as during warmer weather. This theory supports the inclusion of temperature as a key variable in the statistical models.

3 Methodology

The primary objective of this study is to identify key factors that correlate with variations in burglary rates across different census block groups in Chicago. By integrating data on socioeconomic, demographic, and environmental factors, the research seeks to:

- Quantify the impact of socioeconomic variables such as unemployment and wealth on the burglary rate.
- Assess the demographic influences focusing on the proportion of young males and the total population.
- Explore the relationship between environmental factors, specifically average monthly temperatures, and burglary rates.

3.1 Hypothesis

Based on the literature review, the study tests the following hypotheses:

1. Higher unemployment rates and greater levels of wealth within a census block are associated with higher burglary rates, reflecting the impact of economic disparities.

2. Warmer temperatures correlate with increased burglary rates, potentially due to more frequent absence from homes and increased pedestrian traffic, providing more opportunities for burglary.

This research aims to clarify how these factors interact and which are the most critical in shaping burglary patterns in Chicago.

3.2 Data Preparation

Each variable was initially stored in separate CSV files and meticulously merged into a main data frame in R. The main data frame was originally stored as wide format and was transformed as long format to conduct analysis through multiple models. The dataset was scrutinized for completeness and accuracy, with no missing data or significant outliers, ensuring the data broadly followed expected trends and distributions without anomalies that could skew the analysis.

For this research, three distinct data frames were prepared to facilitate different aspects of the analysis:

- Original Data: Maintained in its detailed monthly format per census block for exploratory and detailed temporal analyses.
- Aggregated by Month: This data frame aggregates burglary counts across all census blocks for each month, simplifying the data to a single count per month. This aggregation supports the analysis of temporal effects on burglary rates.
- Aggregated by Census Block: This data frame aggregates the monthly burglary counts for each census block over the entire study period, resulting in a single burglary count per block that represents the total incidents over five years. This aggregation supports the analysis of static effects of socioeconomic and demographic variables on burglary rates using multi-linear regression models.

3.3 Data Exploration

The dataset utilized in this study was sourced from a dataset collected by COL Nicholas Clark at the United States Military Academy. It comprises comprehensive data for 552 census block groups in Chicago, encompassing a range of socioeconomic, demographic, and environmental variables relevant to urban crime analysis.

The dataset includes the following key variables for each census block group:

- Burglary Counts: The response variable representing the number of burglaries reported each month from January 2010 to December 2015.
- Population: Total population per census block, providing a base for normalization and per capita analyses.
- Unemployment Rate: Percentage of the labor force that is unemployed and actively seeking employment, serving as a key economic indicator.
- Wealth: An index of wealth levels within each census block intended to capture economic status and disparity.

- Young Male Count: The number of young males in the population of the census block.
- Temperature: Average monthly temperatures, categorized by month to assess potential seasonal effects on burglary rates.

Visualizations

Figure 2 displays the frequency distribution of burglary rate across all census blocks in Chicago. The majority of census blocks exhibit relatively low crime counts, with a peak frequency for blocks reporting around 50-100 crimes over the study period. The distribution of the burglary rate resembles normal distribution with a slight skew, which allows us to apply linear regression model.

The histogram of burglary rates from aggregated by month data frame (Figure 3) shows the aggregated total crimes per month across all years. The data shows a normal distribution with some fluctuation, suggesting seasonal trends or periodic fluctuations in crime rates. This pattern underscores the importance of considering temporal factors in the modeling process, as month-to-month variations could significantly influence the predictive accuracy of the models.

The time series plot (Figure 4 combines the total crime counts per month with temperature data, represented in different colors for varying temperature ranges. There is a visible relationship between higher temperatures and increased crime rates, particularly noticeable during peak summer months. This visualization is crucial for hypothesizing about the potential impact of weather conditions on crime dynamics and supports the decision to include temperature as a predictor in the analysis.

Figure 1 details the frequency of crime counts across the census blocks as captured in the original unaggregated data. The distribution is highly skewed, with the vast majority of census blocks exhibiting very low crime counts (0 to 2 crimes) and a rapid decrease in frequency as crime counts increase. This skewed visualization of the burglary rate proves that the burglary rate has a Poisson distribution in the original data frame.

Collinearity

To ensure that the variables included in the models do not introduce multicollinearity, we conducted a correlation analysis. A correlation heat map was generated to visualize the strength of relationships between multiple variables: population, unemployment, and wealth.

The heat map (Figure 5) shows a very high correlation between population and wealth (0.90), suggesting that these two variables are highly correlated when predicting burglary rates. Given the potential for multicollinearity, careful consideration was required to select only one of two variables. The correlation between unemployment and both population and wealth was negligible (-0.02 and -0.03, respectively), indicating that unemployment could be included in the models without concerns of overlapping influence with the other two variables.

Population was chosen over wealth as a key predictor in the regression models. This decision was made by the lower p-values and higher coefficients associated with population in preliminary model tests, suggesting that the population offers more significant explanatory power for the variations in burglary rates across census blocks.

Variable Selection

A systematic approach was employed using R code to identify the most predictive combination of variables for burglary rates. We generated multiple regression models to evaluate all possible combinations of predictor variables, which included population, unemployment, wealth, and young male count. The selection process was done through the following R script that iterated over every possible combination of predictors.

```r
for (i in 1:length(predictors)) {
  combns <- combn(predictors, i, simplify = FALSE)
  for (comb in combns) {
    formula <- as.formula(paste("total_crimes ~", paste(comb, collapse = "+")))
    model <- glm(formula, data = block_data, family = gaussian(link = "identity"))
    summary_model <- summary(model)
    results[[paste("Model with predictors:", paste(comb, collapse = ", "))]] <-
   list(
      Formula = formula,
      Summary = summary_model
    )
  }
}

for (model_name in names(results)) {
  print(model_name)
  print(results[[model_name]]$Summary)
}
```

The R script constructs and evaluates models, extracting and recording the summary statistics for each, which include AIC scores, coefficients, and standard errors, providing an overview of each model's performance. The primary criterion for model selection was the Akaike Information Criterion (AIC). Models with lower AIC values were considered better as they provided a better fit to the data.

Table 4 is an example of model comparisons, highlighting the AIC scores and coefficients for each model. For example, the model including both population and young male count as predictors showed a promising balance between complexity and explanatory power.

Table 1: Summary of Model Evaluations with AIC and Coefficients

AIC	Population	Unemployment	Wealth	YM_Count
5666.8	0.03376	-	-	0.167674
5667.3	0.03419	29.30912	-	0.160886
5668.3	0.03911	-	-2.65849	0.165555
5668.8	0.03948	39.19565	-2.62859	0.158814
5677.4	0.04196	-	-	-
5678.1	0.0418	38.891726	-3.52844	-
5678.6	0.04892	-	-3.62041	-
5679.3	0.04897	-	12.12432	0.22551
5684.9	-	39.309	12.27342	0.22006
5707.6	-	-	16.388	-
5708.1	-	-	16.263	-

The decision on the final model was based on the lowest AIC score and the practical significance of the predictors. This approach ensured that the selected model was both statistically sound and relevant for policy recommendations and practical applications.

4 Experimentation

4.1 Aggregated Data by Block

Upon examining the histogram of aggregated burglary rates by block, which displayed a roughly normal distribution albeit with some skew, a linear regression model was selected as a suitable initial approach. This choice was predicated on capturing the linear relationship between demographic factors and the rates of burglary, providing a straightforward interpretation of these effects. As a result of the automated process of variable selection, the variables that produced the lowest AIC score was population and young male count. The equation of the linear regression model is the following:

$$\eta_{ij} = \beta_0 + \beta_1 * (\text{Population}_{ij}) + \beta_2 * (\text{Young Male Count}_i) + \epsilon_{ij}$$

$$y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

Where, i = month, j = location, and $y_{ij} = \eta_{ij}$ = number of burglaries in month i and j

4.2 Aggregated Data by Month

The analysis began with a basic linear regression model using temperature as the sole predictor, motivated by the observed seasonal patterns in a separate histogram of monthly burglary rates. To further refine the model and incorporate the temporal variability, a linear regression with mixed effects was introduced, allowing the model to account for month-to-month fluctuations while assessing the overall temperature impact on burglary rates. The equation of the final linear regression model is the following:

$$\eta_{ij} = \beta_0 + \beta_1 * (\text{Temperature}_j) + u_i + \epsilon_j$$

$$y_{ij} \sim N(\mu_j, \sigma^2)$$

Where, i = month, j = location, and $y_{ij} = \eta_{ij}$ = number of burglaries in location j

4.3 Original Data

The choice of count data models was driven by the initial data exploration of the original, unaggregated dataset. The histogram of the burglary counts showed a Poisson distribution with a large number of zeros. Based on the observation, the first model that was created is the Poisson model.

Poisson Model was chosen as the initial approach for count data modeling, reflecting the basic assumption of equal mean and variance in count data. The Poisson model with the lowest AIC was the model with unemployment, population, and young male count as predictor variables. The Poisson model failed the goodness of fit test. As seen from Figure 1, the burglary counts had a large number of zeros. As a response to the excess zeros in the dataset, a zero-inflated Poisson (ZIP) model was developed. The ZIP model produced lower AIC than the Poisson model but still failed to pass the goodness of fit test. The equation for the final Poisson and ZIP model is the following:

$$\eta_{ij} = \beta_0 + \beta_1 * (\text{Young Male}_{ij}) + \beta_2 * (\text{Population}_{ij}) + \beta_3 * (\text{Unemployment}_{ij})$$

$$y_{ij} \sim Poiss(\lambda_{ij}),\ log(\lambda_{ij} = \eta_{ij})$$

Where, i = month, j = location, and $y_{ij} = \eta_{ij}$ = number of burglaries in location i and j

After the goodness of fit test of the ZIP model, the overdispersion in the dataset was identified. The mean and variance ratio of the dataset was 1.6, meaning the variance was 60% larger than the assumption of the Poisson model. After identifying overdispersion, the negative binomial model was introduced, which was better suited to handle the greater variability in the data. The negative binomial model also utilized the same predictor variables as the Poisson and ZIP models, since the combination of young male count, population, and unemployment as predictor variables produced the lowest AIC value. Compared to the Poisson and ZIP model, the negative binomial model had the lowest AIC value, which indicates a better fit to the data.

Lastly, the Quasi-Poisson model was also tested to provide a flexible approach to addressing overdispersion by allowing the variance to differ from the mean, a feature particularly relevant given the data's characteristics. Due to the characteristics of the Quasi-Poisson model, it is impossible to compare AIC values with other models.

5 Result

Aggregated by Block Data

The Table 2 shows the coefficients and the p-values of the variables in the final model. The AIC value of the model was 5666

Table 2: Summary of Linear Regression Model for Block Data

Variable	Coefficient	P-value
Intercept	42.63406	< 2e-16
YM_Count	0.16767	0.000274
Population	0.03376	1e-13

Aggregated by Month Data

The Table 3 shows the coefficients and the p-values of the variables in the linear regression model with temperature and months as mixed effects.

Table 3: Summary of Linear Regression Model with Mixed Effects for Monthly Data

Variable	Coefficient	P-value
Intercept	434.849	<0.0001
Temperature	4.477	0.00278

Original Data

The Table 4 shows coefficients and p-values of variables in each model. Of four models, the negative binomial model had the lowest AIC value, which is 118577. The ZIP model had the second lowest AIC value, which is 121012, and the Poisson model had the highest AIC value of 124526. The Quasi-Poisson and Negative Binomial models showed similar performance in the original dataset. Figure 6 shows the residuals vs. fitted values of the Quasi-Poisson model and Figure 7 shows the residual vs. fitted values of the Negative Binomial model. Both figures show similar shapes and trends, but the difference is that the Quasi-Possion model shows a banding pattern in the plot. Overall, the Negative Binomial model shows a slightly better fit to the data, but the difference is very minimal.

Table 4: Model Summary for Four Models: Original Data

Model	Coefficient	P-value
Poisson Model		
Intercept	-0.3396	$< 2e-16$
YM_Count	1.658e-03	$< 2e-16$
Population	3.553e-04	$< 2e-16$
Unemployment	0.3553	2.22e-08
Negative Binomial Model		
Intercept	-0.3601107	$< 2e-16$
YM_Count	0.0015786	$< 2e-16$
Population	0.0003803	$< 2e-16$
Unemployment	0.3365601	4.52e-05
Zero-Inflated Poisson Model		
Intercept	1.900e-02	0.296
YM_Count	1.066e-03	1.72e-11
Population	2.863e-04	$< 2e-16$
Unemployment	5.637e-01	2.66e-10
Quasi-Poisson Model		
Intercept	-3.396e-01	$< 2e-16$
YM_Count	1.658e-03	$< 2e-16$
Population	3.553e-04	$< 2e-16$
Unemployment	3.553e-01	2.16e-05

6 Conclusions and Discussion

This study's analysis has identified significant correlations between socioeconomic variables and burglary incidence in Chicago. Specifically, the young male population and the overall population are both positively correlated with burglary rates. Additionally, we found out that seasonal variations, particularly temperature, play a substantial role in the fluctuation of burglary rates throughout the year. The higher temperature in the area caused the higher incidence rate of burglary in Chicago.

Recommendation for Policy and Practice

The strong correlation between burglary rates and the young male population points to a potential demographic focus for community intervention programs. Policymakers may implement targeted programs aimed at engaging this demographic with a higher young male population in highly populated areas, particularly during the summer months. Such initiatives could include summer employment programs and educational opportunities designed to provide constructive alternatives to criminal activities.

The correlation between burglary rates and factors like urban development suggests that improvements in these areas could reduce crime. Enhancements in street lighting, the use of surveillance cameras, and the design of public spaces that promote natural surveillance and territorial reinforcement in urban areas can prevent additional burglaries.

Future Research

Further research is needed to explore the other factors behind the correlations observed in this study. Adding spatial data, such as zip codes, as categorical variables may provide insight into the spatial impact on burglary rates. Also, this result tried to handle the overdispersion through the Negative Binomial model but failed to capture the large number of 0s in the data. Constructing additional models from the data may develop more precise insight into the correlation.

Stake Holder Analysis

The study on burglary rates in Chicago draws significant interest from various stakeholders, each with different levels of influence and specific implications based on the study's findings.

Local Government and Policy Makers

Local government officials and policymakers exhibit a high interest in reducing crime rates, improving public safety, and efficiently allocating resources. They have significant influence as they can implement policies and initiate community programs based on the study's analysis. This study can guide policy adjustments and resource allocation, focusing on high-risk demographics and areas. Implementing targeted interventions based on the study could enhance their political credibility and effectiveness in governance.

Law Enforcement Agencies

Law enforcement agencies are interested in developing effective strategies for crime prevention and control. Their influence ranges from moderate to high, as they can adjust operational strategies within the constraints of broader policy frameworks. The study's findings on burglary hotspots and correlations with certain demographic factors can aid in tactical planning, patrolling, and community policing strategies, potentially leading to more efficient crime prevention and lower crime rates.

Local Communities and Residents

Residents, especially those in neighborhoods with high burglary rates, have a very high interest in the outcomes of this study. The study's findings enable residents to pursue more resources or targeted programs in their neighborhoods. Additionally, a better understanding of crime patterns can enhance community vigilance and encourage preventative measures at both individual and neighborhood levels.

Youth Organizations

Youth organizations show a high interest in engaging youth and preventing juvenile delinquency. Their influence is moderate, as their programs directly impact the youth but depend on funding and policy support. Insights from the study could help design more effective programs that engage young males during high-risk periods, potentially preventing them from engaging in criminal activities.

Stake Holder Analysis: Conclusion

This stakeholder analysis shows the various impacts of the study's findings on various groups, highlighting the importance of strategic communication and collaboration among stakeholders to leverage the insights for public benefit effectively. Each stakeholder group can contribute to a holistic approach toward addressing the issues of burglary in Chicago.

Final Conclusion

The findings from this study provide valuable insight into burglary in Chicago for decision-making in crime prevention and urban planning. By understanding the demographic and temporal patterns associated with burglary rates, as well as the spatial distribution of these crimes, policymakers can implement more effective strategies that address the potential root causes of burglaries in the city.

Appendix A

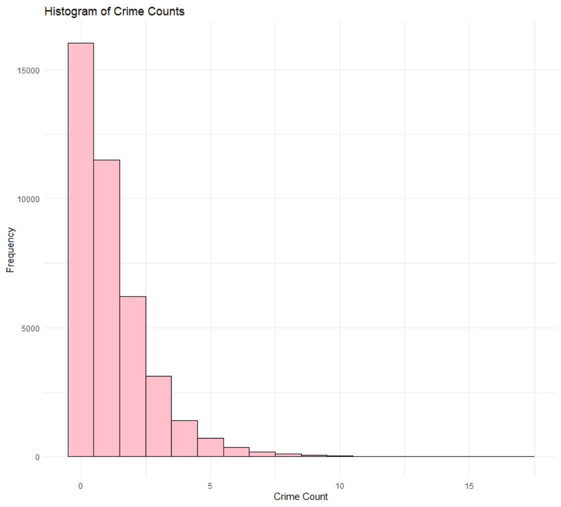

Figure 1: Histogram of Burglary Rate in Chicago (Original Data)

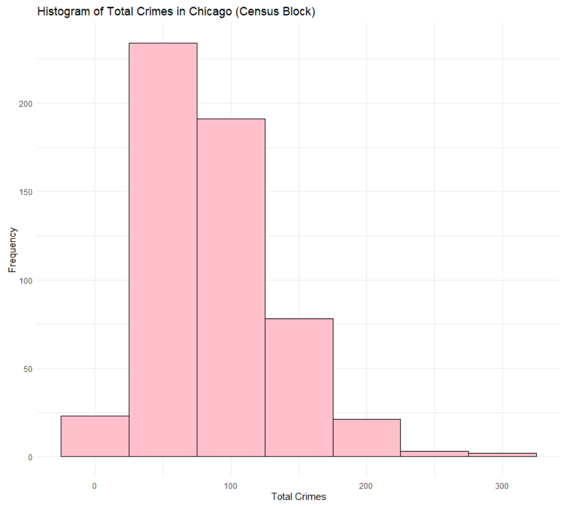

Figure 2: Histogram of Burglary Rate in Chicago (Aggregated by Census Block)

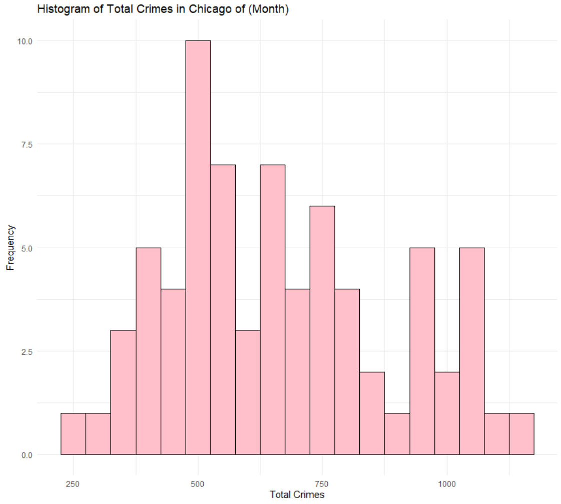

Figure 3: Histogram of Burglary Rate in Chicago (Aggregated by Month)

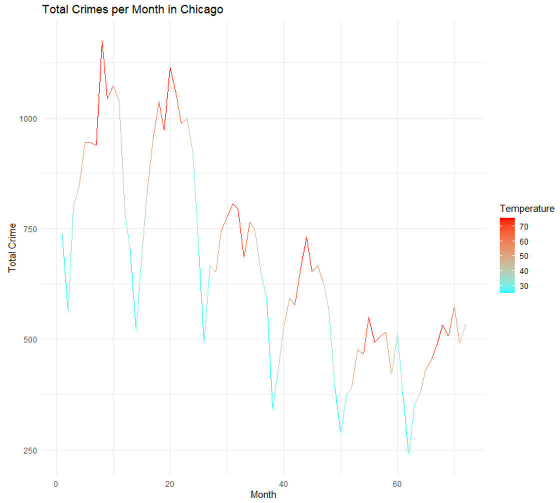

Figure 4: Burglary Rate per Month with Temperature

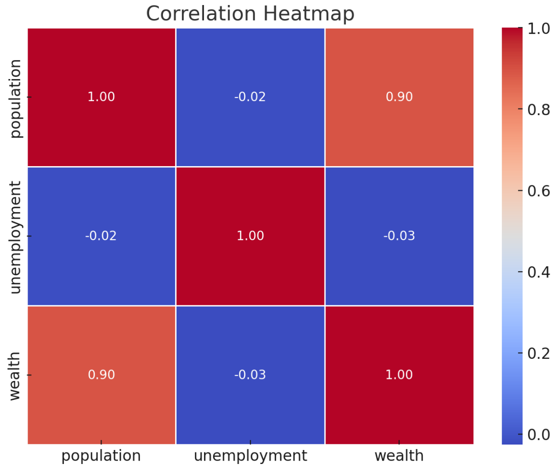

Figure 5: Correlation Heat Map of Variables

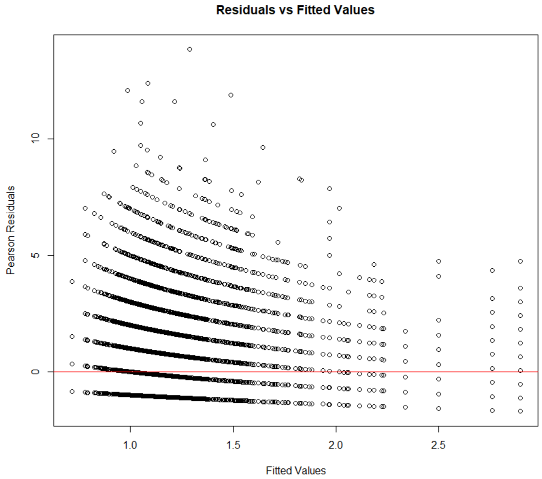

Figure 6: Quasi Poisson Model: Residuals vs. Fitted Values

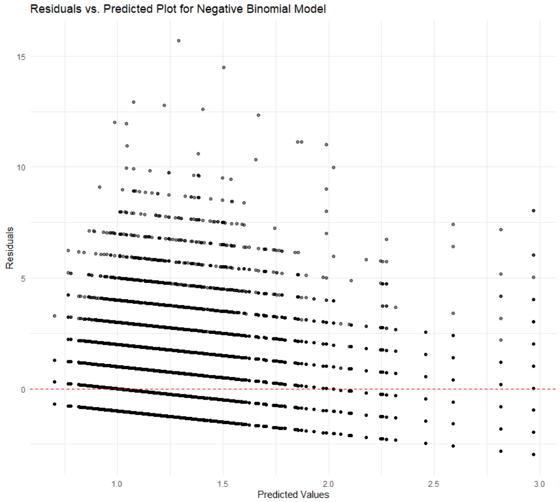

Figure 7: Negative Binomial Model: Residuals vs. Fitted Values

Appendix B

R Code

```r

library(CARBayesST)
library(rstan)
library(Matrix)
library(dplyr)
library(tidyverse)
library(lubridate)
library(knitr)
library(ggplot2)
library(dplyr)
library(glm2)
library(sf)
library(spdep)
library(tidyverse)
library(rstan)
library(Matrix)
library(spdep)
library(INLA)
library(geepack)
library(wesanderson)
library("lme4")
library("spatstat")
# library("maptools")
library("lattice")
library(pscl)
library(boot)
library(pROC)
library(ROCR)

Chi_dat=read.csv("https://raw.githubusercontent.com/nick3703/Chicago-Data/master/crime.csv")
#Population by Census Block Group
pop=read.csv("https://raw.githubusercontent.com/nick3703/Chicago-Data/master/pop.csv")
#Percentage Unemployed by Census Block Group
un.emp=read.csv("https://raw.githubusercontent.com/nick3703/Chicago-Data/master/unemp.csv")
#Centered and Scaled Average Family Income by Census Block Group (2015 Dollars)
wealth.std=read.csv("https://raw.githubusercontent.com/nick3703/Chicago-Data/master/wealth.csv")
#Number of Young Males by Census Block Group (15-20 yr olds)
ym=read.csv("https://raw.githubusercontent.com/nick3703/Chicago-Data/master/ym.csv")

#Here, we aggregate the data by the month and by the census block. This will allow us to perform a simple analysis of the data #with a basic linear model before we attack the more complex models.



long_data <- pivot_longer(Chi_dat, cols = starts_with("count."),
                          names_to = "date",
                          values_to = "count",
                          names_prefix = "count.")

long_data$date <- as.Date(paste0(substr(long_data$date, 1, 4), "-", substr(long_data$date, 5, 6), "-01"))

block_data <- long_data %>%
  group_by(X) %>%
  summarise(total_crimes = sum(count))
monthly_data <- long_data %>%
  group_by(date) %>%
  summarise(total_crimes = sum(count))
monthly_data$date_numeric <-as.numeric(as.factor(monthly_data$date))
monthly_data <- monthly_data %>%
  mutate(month = month(date, label = TRUE))

monthly_data <- subset(monthly_data, select = -c(date))





#Here, we merge the data with all extraneous data presently given to us.

names(pop)[names(pop) == "x"] <- "population"
names(un.emp)[names(un.emp) == "x"] <- "unemployment"
names(wealth.std)[names(wealth.std) == "x"] <- "wealth"
names(ym)[names(ym) == "x"] <- "ym_count"

block_data <- merge(block_data, un.emp, by = "X")
block_data <- merge(block_data, wealth.std, by = "X")
block_data <- merge(block_data, ym, by = "X")
block_data <- merge(pop, block_data, by = "X")
# <- merge(merged_df, Chi_dat, by = "X")

#We are given weather data based on the month's in Chicago so this is also merged with the monthly data.



weather <- data.frame(
  month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"),
  # HIGH = c(31.6, 35.7, 47.0, 59.0, 70.5, 80.4, 84.5, 82.5, 75.5, 62.7, 48.4, 36.6),
  # LOW = c(18.8, 21.8, 31.0, 40.3, 50.6, 60.8, 66.4, 65.1, 57.1, 45.4, 34.1, 24.4),
  AVERAGE = c(25.2, 28.8, 39.0, 49.7, 60.6, 70.6, 75.4, 73.8, 66.3, 54.0, 41.3, 30.5)
  # HEATING_DEGREE_DAYS = c(1234, 1015, 808, 468, 198, 31, 2, 4, 77, 355, 713, 1069),
  # COOLING_DEGREE_DAYS = c(0, 0, 2, 8, 60, 199, 326, 276, 116, 15, 0, 0),
  # PRECIPITATION = c(1.99, 1.97, 2.45, 3.75, 4.49, 4.10, 3.71, 4.25, 3.19, 3.43, 2.42, 2.11),
  # SNOWFALL = c(11.3, 10.7, 5.5, 1.3, 0.01, 0.0, 0.0, 0.0, 0.0, 0.2, 1.8, 7.6)
)

```

```r
weather <- data.frame(
  month = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
  # HIGH = c(31.6, 35.7, 47.0, 59.0, 70.5, 80.4, 84.5, 82.5, 75.5, 62.7, 48.4, 36.6),
  # LOW = c(18.8, 21.8, 31.0, 40.3, 50.6, 60.8, 66.4, 65.1, 57.1, 45.4, 34.1, 24.4),
  AVERAGE = c(25.2, 28.8, 39.0, 49.7, 60.6, 70.6, 75.4, 73.8, 66.3, 54.0, 41.3, 30.5)
  # HEATING_DEGREE_DAYS = c(1234, 1015, 808, 468, 198, 31, 2, 4, 77, 355, 713, 1069),
  # COOLING_DEGREE_DAYS = c(0, 0, 2, 8, 60, 199, 326, 276, 116, 15, 0, 0),
  # PRECIPITATION = c(1.99, 1.97, 2.45, 3.75, 4.49, 4.10, 3.71, 4.25, 3.19, 3.43, 2.42, 2.11),
  # SNOWFALL = c(11.3, 10.7, 5.5, 1.3, 0.01, 0.0, 0.0, 0.0, 0.0, 0.2, 1.8, 7.6)
)


monthly_data <- merge(monthly_data, weather, by = "month")

monthly_data <- monthly_data %>%
  rename(AVERAGE = AVERAGE.x)


#Now we will explore the data to see if there are any relationships between the variables.

ggplot(monthly_data, aes(x=total_crimes)) +
  geom_histogram(binwidth = 50, fill="pink", color="black") +
  labs(title="Histogram of Total Crimes in Chicago of (Month)", x="Total Crimes", y="Frequency") +
  theme_minimal()

ggplot(block_data, aes(x=total_crimes)) +
  geom_histogram(binwidth = 50, fill="pink", color="black") +
  labs(title="Histogram of Total Crimes in Chicago (Census Block)", x="Total Crimes", y="Frequency") +
  theme_minimal()

ggplot(monthly_data, aes(x = date_numeric, y = total_crimes, color = AVERAGE)) +
  geom_line() +
  scale_color_gradient(low = "cyan", high = "red") +
  labs(title = "Total Crimes per Month in Chicago",
       x = "Month",
       y = "Total Crime",
       color = "Temperature") +
  theme_minimal()




q <- block_data[, sapply(block_data, is.numeric)]

cor(q)
q1 <- monthly_data[, sapply(monthly_data, is.numeric)]
cor(q1)

```

```r
#There is strong correlation between the wealth value and population in the block data.


#Simple GLM
# glm1 <- glm(total_crimes ~ population + wealth + ym_count + unemployment, data = block_data, family = gaussian(link = identity))
# summary(glm1)
# glm2 <- glm(total_crimes ~ wealth + ym_count + unemployment, data = block_data, family = gaussian(link = identity))
# summary(glm2)
# glm3 <- glm(total_crimes ~ ym_count + unemployment, data = block_data, family = gaussian(link = identity))
# summary(glm3)
# glm4 <- glm(total_crimes ~ unemployment, data = block_data, family = gaussian(link = identity))
# summary(glm4)


#extract names of all predictor variables

predictors <- names(block_data)[!names(block_data) %in% c("X", "total_crimes")]

results <- list()


for (i in 1:length(predictors)) {
  combns <- combn(predictors, i, simplify = FALSE)
  for (comb in combns) {
    formula <- as.formula(paste("total_crimes ~", paste(comb, collapse = "+")))
    model <- glm(formula, data = block_data, family = gaussian(link = "identity"))
    summary_model <- summary(model)
    results[[paste("Model with predictors:", paste(comb, collapse = ", "))]] <- list(
      Formula = formula,
      Summary = summary_model
    )
  }
}


for (model_name in names(results)) {
  print(model_name)
  print(results[[model_name]]$Summary)
}


#The lowest AIC was found in the model with the predictors: "ym_count, population". This model will be used for the spatial #analysis.

glm5 <- glm(total_crimes ~ ym_count + population, data = block_data, family = gaussian(link = identity))

summary(glm5)
#Goodness of fit test

```

```r
1-pchisq(glm5$deviance, glm5$df.residual)

AIC_linear_block <- AIC(glm5)

glm_m <- glm(total_crimes ~ ym_count + population, data = monthly_data, family = gaussian(link = identity))

summary(glm_m)


#Now Linear model with temperature


#Simple GLM
monthly_data$month <- as.factor(monthly_data$month)
glm1 <- glm(total_crimes ~ AVERAGE, data = monthly_data, family = gaussian(link = identity))
summary(glm1)

1-pchisq(glm1$deviance, glm1$df.residual)


#Creating a new dataframe with the data we need for the spatial analysis.


merged_df <- merge(pop, un.emp, by = "X")
merged_df <- merge(merged_df, wealth.std, by = "X")
merged_df <- merge(merged_df, ym, by = "X")
pdf <- merge(merged_df, long_data, by = "X")

head(pdf)

ggplot(pdf, aes(x = count)) +
  geom_histogram(binwidth = 1, fill = "pink", color = "black") +
  labs(title = "Histogram of Crime Counts", x = "Crime Count", y = "Frequency") +
  theme_minimal()

pois <- glm(count ~ ym_count + population + unemployment, data = pdf, family = poisson(link = log))

pois2 <- glm(count ~ ym_count + population, data = pdf, family = poisson(link = log))
summary(pois)
#Goodness of fit test
1-pchisq(pois$deviance, pois$df.residual)

AIC_pois <- AIC(pois)
BIC_pois <- BIC(pois)

#None of these models work so far regarding the goodness of fit. We will try to use a negative binomial model instead.

nb <- glm.nb(count ~ ym_count + population + unemployment, data = pdf)

summary(nb)
#Goodness of fit test
```

~~~

1-pchisq(nb$deviance, nb$df.residual)

AIC_nb <- AIC(nb)
BIC_nb <- BIC(nb)
```

```{R}
predicted_values <- predict(nb, type = "response")
residuals_values <- resid(nb, type = "response")
plot_data <- data.frame(
  Predicted = predicted_values,
  Residuals = residuals_values
)

ggplot(plot_data, aes(x = Predicted, y = Residuals)) +
  geom_point(alpha = 0.5) +  # Use points with some transparency
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +  # Add a horizontal line at y = 0
  labs(x = "Predicted Values", y = "Residuals") +
  theme_minimal() +  # Use a minimal theme
  ggtitle("Residuals vs. Predicted Plot for Negative Binomial Model")



#There is a mass amount of zeroes present in the data, so we will try to use a zero-inflated poisson model instead.

zip <- zeroinfl(count ~ 1, data = pdf, dist =  "poisson")
summary(zip)

#Goodness of fit test
zip <- zeroinfl(count ~ ym_count + population + unemployment, data = pdf, dist = "poisson")
summary(zip)

AIC_zip <- AIC(zip)
BIC_zip <- BIC(zip)


pois_mm <- glmer(count ~ 1 + (1|date) + population + unemployment, family = poisson, data = pdf)

summary(pois_mm)

AIC(pois_mm)

observed_deviance <- sum(resid(pois_mm, type = "pearson")^2)
df_residual <- df.residual(pois_mm)
observed_deviance / df_residual


#Overdispersion Value of 1.6: This value means that the observed variance is 60\% greater than what would be expected under the Poisson model assumption (where the dispersion would ideally be 1). This indicates that the Poisson model is
~~~

```
not a good fit for the data. We will try to use a negative binomial model instead.

# pdf$date <- as.Date(pdf$date)
pdf$month <- month(pdf$date)


nb_mm <- glmer.nb(count ~ + ym_count + population + unemployment + (1|month), data = pdf)

summary(nb_mm)


AIC_nb_mm <- AIC(nb_mm)


observed_deviance <- sum(resid(nb_mm, type = "pearson")^2)
df_residual <- df.residual(nb_mm)
observed_deviance / df_residual

monthly_data$month <- as.factor(monthly_data$month)

#merge long data with all the other data

merged_df <- merge(pop, un.emp, by = "X")
merged_df <- merge(merged_df, wealth.std, by = "X")
merged_df <- merge(merged_df, ym, by = "X")
long_data <- merge(merged_df, long_data, by = "X")


gaus_mm1 <- lmer(total_crimes ~(1|month), data = monthly_data)

summary(gaus_mm1)

gaus_mm2 <- lmer(total_crimes ~ (1|month) + AVERAGE, data = monthly_data)
summary(gaus_mm2)


AIC(gaus_mm1)
AIC(gaus_mm2)


residuals_data1 <- data.frame(residuals = residuals(gaus_mm1),
                              fitted = fitted(gaus_mm1))

ggplot(residuals_data1, aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Fitted Values", x = "Fitted Values", y = "Residuals")

residuals_data2 <- data.frame(residuals = residuals(gaus_mm2),
                              fitted = fitted(gaus_mm2))

ggplot(residuals_data2, aes(x = fitted, y = residuals)) +
  geom_point() +
```

```r
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Fitted Values GLMM", x = "Fitted Values", y = "
    Residuals")

residuals_data3 <- data.frame(residuals = residuals(glm5),
                             fitted = fitted(glm5))

ggplot(residuals_data3, aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Fitted Values LM", x = "Fitted Values", y = "
    Residuals")


rand_effects <- ranef(gaus_mm2, condVar = TRUE)
plot(rand_effects, main = "Random Effects Distribution")


model_qp <- glm(count ~ population + ym_count + unemployment,
                family = quasipoisson(link = "log"),
                data = pdf)

summary(model_qp)



# model_qp <- glm(count ~ population + ym_count + unemployment + (1|date),
#                  family = quasipoisson(link = "log"),
#                  data = pdf)

summary(model_qp)

log_likelihood <- logLik(model_qp)

# Print the log likelihood
print(log_likelihood)


LL_pois <- logLik(pois)
LL_nb <- logLik(nb)
LL_zip <- logLik(zip)
LL_qp <- logLik(model_qp)

plot(fitted(model_qp), residuals(model_qp, type = "pearson"),
     xlab = "Fitted Values", ylab = "Pearson Residuals",
     main = "Residuals vs Fitted Values")
abline(h = 0, col = "red")


ggplot(data = pdf, aes(x = date, y = count)) +
  geom_point() +
  labs(x = "Month", y = "Count", title = "Observed Counts")

ggplot(data = pdf, aes(x = date, y = predicted_counts)) +
  geom_point() +
  labs(x = "Month", y = "Count", title = "Predicted Counts")
```

```r


ggplot(data = pdf, aes(x = date, y = count)) +
  geom_point() +
  labs(x = "Month", y = "Count", title = "Observed Counts")

ggplot(data = pdf, aes(x = date, y = predicted_counts)) +
  geom_point() +
  labs(x = "Month", y = "Count", title = "Predicted Counts")
```

Works Cited

[1] Elliott Ramos. Here's what's happening with crime in chicago in 2023. CBS News Chicago, 2023. https://www.cbsnews.com/chicago/news/heres-whats-happening-with-crime-in-chicago-in-2023/.

[2] Daniel Chang. Social crime or spatial crime? exploring the effects of social, economical, and spatial factors on burglary rates. *Environment and Behavior*, 43(1):26–52, 2011.

[3] Richard Rosenfeld and Steven F. Messner. Crime and the economy: Insights from a macrolevel examination of the usa and europe. *Criminology & Public Policy*, 8(4):649–686, 2009.

[4] Graham Farrell and Ken Pease. Crim seasonality: domestic disputes and residential burglary in merseyside 1988–90. *British Journal of Criminology*, 34(4):487–498, 1994.