UNITED STATES MILITARY ACADEMY

FINAL PROJECT

(CHICAGO BURGLARY QUEST)

MA478: GENERALIZED LINEAR MODELS

SECTION H2

COL CLARK, NICHOLAS

By

CADET JOSHUA BLACKMON '24, CO C1

WEST POINT, NEW YORK

07 MAY 2024

BLACKMON.JOSHUA.LEE.1381780356 Digitally signed by BLACKMON.JOSHUA.LEE.1381780356 Date: 2024.05.07 11:51:15 -04'00' 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: BLACKMON.JOSHUA.LEE.1381780356 Digitally signed by BLACKMON.JOSHUA.LEE.1381780356 Date: 2024.05.07 11:51:45 -04'00'

Chicago Burglary Quest

CDT J. L. Blackmon

Abstract

This project delves into the complex dynamics of burglary rates in Chicago, aiming to inform potential policing reforms and community initiatives. By analyzing burglary incident data spanning from January 2010 to December 2015, we seek to uncover underlying trends and factors driving burglaries across different neighborhoods. Our research questions center on understanding the relationships between socioeconomic variables: wealth, unemployment, and population demographics, and burglary rates in Chicago. Additionally, we investigate the influence of temperature fluctuations on burglary occurrences. Through a combination of linear regression models, linear mixed effects models, and Poisson model variants, we analyze the associations between various variables and burglary rates. Our findings reveal significance between burglary rates and demographic factors, such as the proportion of young males and overall population density. Furthermore, temperature patterns emerge as reliable predictors of burglaries throughout the year. However, our analysis also highlights limitations, including the challenge of directly comparing models due to differing response variables and the need for further exploration of spatial effects on crime patterns. Policymakers and community leaders should consider these implications when implementing data-driven approaches to address crime. Our research provides valuable insights into burglary patterns in Chicago and offers recommendations for policymakers and community leaders to address socioeconomic disparities, understand environmental influences, and prioritize community engagement to reduce burglary rates and enhance the well-being of Chicago residents.

Keywords — Crime Analysis, Burglary Rates, Chicago Neighborhoods, Socioeconomic Factors, Urban Development, Community Policing, Statistical Modeling

1 Introduction

For the past 60 years, Chicago has faced significant challenges related to crime, with burglary being one of the most prevalent issues affecting the safety and security of its communities [1]. The city's neighborhoods experience varying levels of burglary rates, influenced by factors such as socioeconomic status, urban development, and community policing efforts. For this project, we have been asked to conduct a comprehensive analysis of burglary rates across different parts of Chicago in an effort to inform policing reform in the city. By leveraging detailed data on burglary incidents, we intend to provide actionable insights that can inform policymakers, Chicago PD, and community leaders. This could ultimately contribute to the reduction of burglary rates and better the well-being of Chicago's residents.

With that being said, we have deliberated on our Research Questions:

- What is the relationship between socioeconomic variables and the incidence of burglaries in Chicago?

- What other determinants significantly impact the frequency of burglaries in Chicago?

2 Literature Review

With our task surrounding the accurate research surrounding what affects crime within Chicago, we understand that various socioeconomic factors within each block of Chicago can prove to be essential to research. In Table 1, you can see what data was considered and covered in our research.

Explained Variation	**Unexplained Variation**
Wealth of Area	Officer Presence
Young Male Population	Age of Population
Population per Block	Education Level
Temperature	Reinvestment per Block
Temporal Factors	Race Makeup

Table 1: Factors Contributing to Explained and Unexplained Variation

2.1 Weather's Effect on Crime

Current research yields strong evidence that temperature has a positive effect on most types of property and violent crime. This differs with rainfall or sunshine amount as it is found to be independent from crime [2]. Anecdotally, this makes sense since higher temperatures cause people to spend more time outside the home. Time spent outside the home, in line with routine activity explanations for crime, has been shown to increase the risk of criminal victimization for most types of crime. The results suggest that temperature is one of the main factors to be taken into account when explaining quarter-to-quarter and month-to-month variations in recorded crime [2]. For this reason, we focused on temperature rather than precipitation, UV, etc. to research. This specific research is however notably in the United Kingdom where gun violence is not as prevalent as in Chicago. This will be considered.

2.2 Population Demographics

While studies on spatial mismatch in the low-skilled labor market have documented the poor job opportunity possessed by youth in neighborhoods within Atlanta, Chicago, and Detroit, there exists no evidence on the role job opportunity plays in explaining the dramatic spatial variation in crime within urban areas. Controlling for time and fixed effects, studies suggest that young males' job opportunity plays a key role in the high variation in crime across urban neighborhoods [3].
Extensive research has seen these results and has delved into the impacts of age and unemployment on criminal activity, but there has been limited exploration of how age and unemployment interact to influence rates of crime. By analyzing arrest and unemployment data specific to different age groups in the United States from 1958 to 1995, this study unveils two key findings: firstly, unemployment exerts a stronger motivational force on property crime among youth and young adults; secondly, the relationship between unemployment and crime fluctuates over time, displaying a more sporadic pattern rather than a consistent trend.[4]. With research being conclusive about young male effect in general, it becomes our basis for the use of Young Male population and unemployment for our research.

What is heavily noted as well is how all research accounts for some kind of fluctuation in crime. Some dictates it as random and some says its cyclical but not established what causes it. To both account for and explore it, we include time as well for our variation.

Finally, violent crime is often studied with individual level variables, using population characteristics as predictors. Plenty of studies attempt to predict the amount of variability in violent crime using the variable—population density—in a single U.S. city. Using data aggregated to the census block group level and basing relationships non-linear relationship between crime and density, it was conversely hypothesized that density would have a positive predictive effect on violent crime in the suburban areas [5]. The analyses support the hypotheses for the urban areas, but fail to support the hypotheses for the suburban areas, providing insight into an elusive relationship—and the effects of environments on behavior patterns [6]. Since population and wealth are heavily implicated as explanatory variables, both factors are included [7].

Within the unexplained variation in Table 1, we have other points that may put together more of the puzzle of this analysis and future research may point to the addition of them to the study.

3 Method(s)

The dataset consists of 552 observations with the number of burglaries from different locations in Chicago. The observations are the burglaries records in Chicago from January 2010 to December 2015. Table 2 has the values regarding the main blocks. Notably, wealth is normalized around 0.

3.1 Hypothesis

Based on our initial assumptions, literature research, our hypotheses are:

1. *Areas with high unemployment rates and higher levels of wealth are related to a higher incidence of burglaries in Chicago. This is based on the idea that the stark contrast between the two economic levels would invite crime at a higher rate.*

2. *The hotter the weather is, the more burglaries are present. The better the weather in Chicago, the more likely and more often people will be outdoors and physically active. This would lead to more empty homes and more burglaries in theory.*

Variable	**Population**	**Unemployment**	**Wealth**	**Young Male Count**
Minimum	12	0	-2.2537	0
1st Quartile	743	0.07379	-0.6768	18
Median	973	0.1123	-0.1289	44
Mean	1044	0.12342	0	52
3rd Quartile	1259	0.16495	0.5243	76
Maximum	3057	0.5433	4.427	268

Table 2: Descriptive Statistics of Population, Unemployment, Wealth, and Young Male Count

Checking correlation between all the variables shows very high association between Wealth and Population, confirming what our prior research has stated [7]. Figure 1 shows this relationship while it informs how our models may behave.

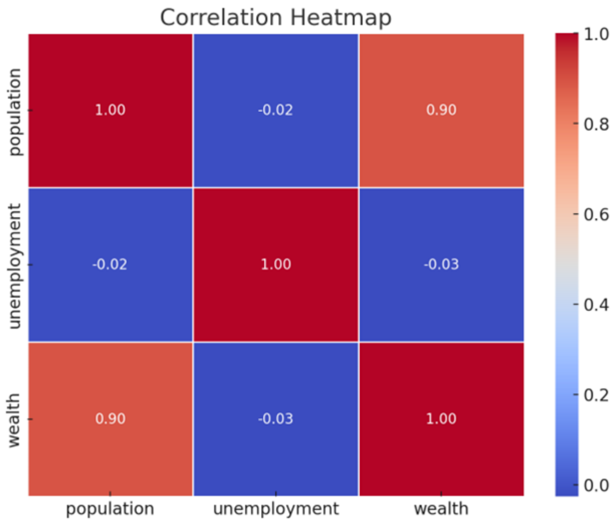

Figure 1: Correlation Heatmap of the Main Demographic Variables

3.2 Data Dissection

In an effort to explore the data in as many ways as possible, we adopted 3 formats of it:

- Count Data based on each Block each Month. This supports it so that Observational Units is each Month in every Census Block in Chicago and our response variable would be Burglary Count per Block per Month. The distribution of it can be seen in the right plot of Figure 2. Population, unemployment rate, wealth, and young male count were merged here.
- Aggregate Data 1 (We will call this AG1) based on the block. This would format the Observational Units so that it is every Census Block in Chicago and our response is the Total Number of Burglaries in each Block. The distribution of it can be seen in the left plot of Figure 2. Population, unemployment rate, wealth, and young male count were merged here.
- Aggregate Data 2 (We will call this AG2) based on the month. This would format the Observational Units so that it is every Month in Chicago and our response is the Total Number of Burglaries in each Month. The distribution of it is the same as the AG1. Only temperature was merged here since it was the only one based on monthly data.

With the Count Data, we have the variety of Poisson and Poisson variants to choose from. Specifically, we are interested in the standard Poisson Model and the Zero-Inflated Poisson Model. This is due to the vast amount zeros that are present in the distribution of the Crime Counts. In the right plot of Figure 2, it is the ideal distribution for both options. For this reason, we start with them for the count data.

With AG1 and AG2, we can use the standard Gaussian distribution to ascertain a multi-Linear model. This allows us to attack the solution with a standard idea. This also has the added benefit of interpretablity to our stakeholders for implementation. For further examination, a linear mixed effects model is also considered for this project but only for AG2 as we aim to use month as the random effect. This is due to the cyclical nature of crime in Chicago and temperature. This can

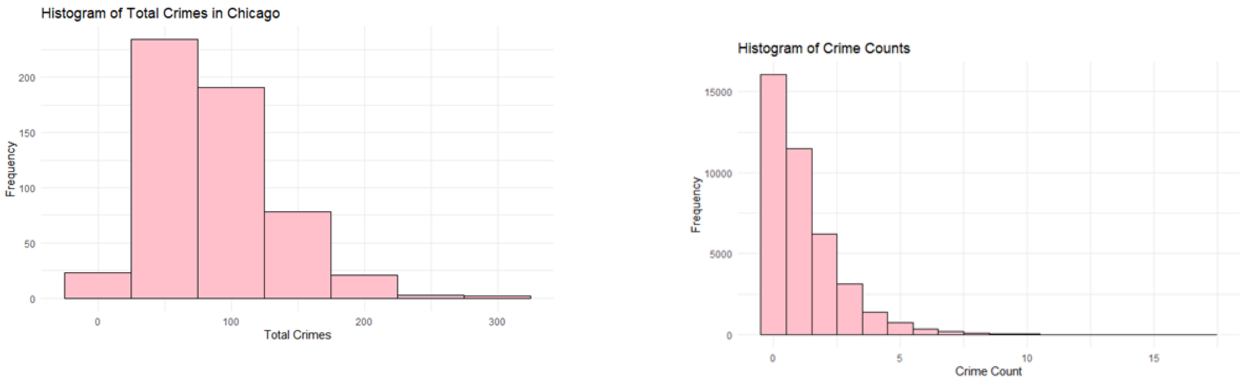

Figure 2: Our Two Histograms of Crime in Chicago.

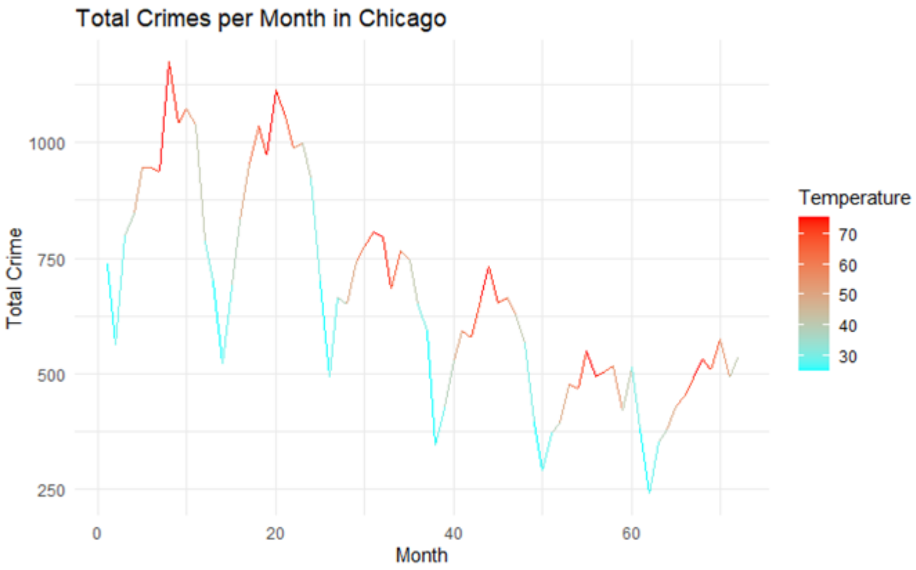

Figure 3: Total Crimes per Month in Chicago cross referenced with Temperature

be seen in Figure 3. AG2 accounts for both pieces of information so we hope to explore the effect that Temperature has on Crime while fixing time.

4 Experimentation and Results

4.1 Linear Regression Models

Using AG1 in a standard Linear Regression but testing all options provides us with the results shown in Table 3 in the prediciton of total crime per block. The lowest AIC is accounted as the linear model with Population and Young Male Count only with a value of 5666.8. Wealth proves to only be useful when population is not a part of the model. This is reasonable with the information that the two are correlated but counter to our original thoughts. Equation 1 shows the distribution and Equation 2 shows the resulting model.

AIC	Population	Unemployment	Wealth	YM_Count
5666.8	0.033763			0.167674
5667.3	0.034193	29.30192		0.160886
5668.3	0.039115		-2.658488	0.165555
5668.8	0.039483	29.202957	-2.628593	0.158814
5677.4	0.041956	39.19565		
5678.1	0.0418			
5678.6	0.048922	38.891726	-3.528435	
5679.3	0.048967		-3.620412	
5684.9			12.12432	0.22551
5685.7		25.77175	12.27342	0.22006
5707.6		39.309	16.338	
5708.1			16.263	
5720.5				0.3415
5722.2		14.92299		0.33917
5782.3		33.596		

Table 3: AIC and Predictor Coefficients (Green shows significant variables in the model, Red shows insignificance.)

$$y_i \sim \mathcal{N}(\mu_i, \sigma) \quad \text{for each block } i \tag{1}$$

$$y_i = \beta_0 + \beta_1(\text{Population}_i) + \beta_2(\text{Young Male Count}_i) \tag{2}$$

4.2 Linear Mixed Effects Models

Using AG2 for the Mixed Effects Models became the next step of our exploration. Our first model assumes each observation is treated independently, with the response variable assumed to follow a normal distribution with mean determined by a fixed linear predictor and constant variance. The linear predictor comprises fixed effects, specifically intercept and the month. This model overlooks any shared characteristics or patterns that may exist within groups or clusters present in the data, potentially leading to underestimation of variability and inaccurate estimates of model parameters. This can be seen in Equations 3-4 and Equations 6-7.

In contrast, the second model introduces a random effect term to account for potential clustering within Temperature. By incorporating this random effect term, the model acknowledges the hierarchical structure of the data, allowing for variability in the response variable attributable to group-level characteristics. This model captures both fixed effects, representing overall trends across all months, and random effects, representing variability specific to each month. This can be seen in Equations 3 and Equations 5-8.

The standard linear models predict total crime per block while the mixed effect models predict total crime per month so we cannot compare the AIC between the models. However the first Mixed Effect model resulted in a lower AIC in 980.6277 and the second achieved 972.0033. These give us a direction to the significance of the time of the year and temperature with crime in Chicago.

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{for each month}_i \tag{3}$$

$$\text{Linear Predictor 1: } \eta_i = \beta_0 + \beta_1(\text{Temperature}_i) \tag{4}$$

$$\text{Linear Predictor 2: } \eta_i = \beta_0 + \beta_1(\text{Temperature}_i) + u_{\text{month}[i]} \tag{5}$$

$$\text{Link Function: } \eta_i = e_i \tag{6}$$

$$e_i \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

$$u_{\text{month}[i]} \sim \mathcal{N}(0, \sigma_u^2) \tag{8}$$

We can say adequately that our second GLMM is the best model so far since the spread of the residuals is equally spread and unpatterned than our best LM. This can be seen in Figure 4.

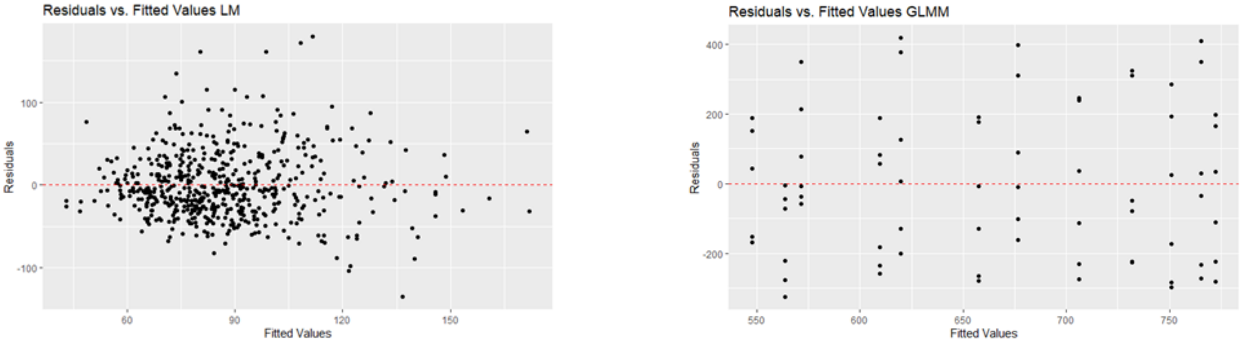

Figure 4: Residuals vs Fitted Values for the best Linear Model and best GLMM

4.3 Poisson Model Variants

Finally, to observe the count data, we sought the Poisson Model and ZIP Model. This proved to be subpar as they failed the goodness of fit test horribly. This led to the testing of Negative Binomial and even Quasi-Poisson Model. In the end, all failed the goodness of fit test except Quasi-Poisson. The quasi-Poisson model, a generalized form of the Poisson model, accommodates overdispersion by introducing a dispersion parameter. By relaxing the assumption of equal mean and variance, it better fits data variability. If the quasi-Poisson model passes the goodness-of-fit test, it indicates improved fit by addressing overdispersion [8].

For the same reason as before, we cannot compare any previous models directly since this model is predicting the number of burglaries in a block per month. The best Poisson Model and the best ZIP Model had 124526 and 121012 respectively. All other forms of these models had an AIC higher than these but again, they failed the goodness of fit tests.

Going through the Quasi-Poisson Model, we understand that there is no likelihood since it estimates the dispersion parameter, which quantifies the degree of overdispersion present in the data. To do this, it ignores likelihood to scale the variance of the Poisson distribution to match the observed variance [8]. Equations 9-12 showcases how the best model turned out. We determined

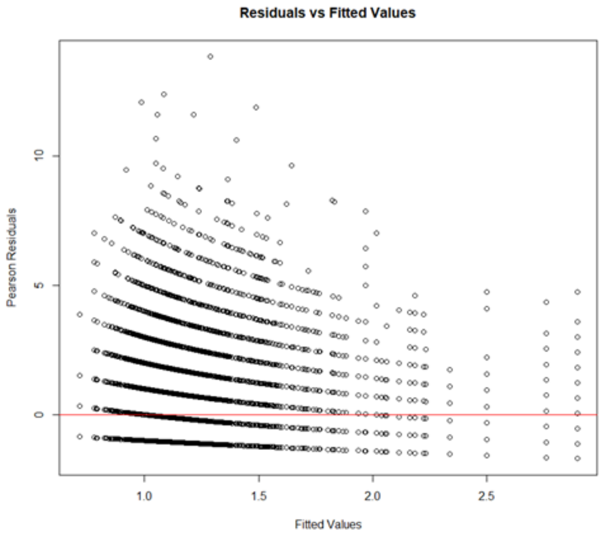

Figure 5: Residuals Vs Fitted Values for the Quasi-Poisson Model.

the best model based on the Residuals vs Fitted plot as well as you can see in Figure 5. Notably, it isn't has a very distinct pattern and is not centered around 0 (the red line). This appeared to be the best option of the other Quasi-Poisson models.

$$i = \text{month per block} \tag{9}$$

$$y_i \sim \text{Poisson}(\lambda_i) \tag{10}$$

$$\log(\mu_i) = \eta_i; Var(y_i) = \phi\mu_i \tag{11}$$

$$\eta_i = \beta_0 + \beta_1(\text{young male}_i) + \beta_2(\text{population}_i) + \beta_3(\text{unemployment}_i) \tag{12}$$

Variable	**Estimate**	**Std. Error**	**Pr($>\|t\|$)**
(Intercept)	0.0035874	0.0139110	0.797
Unemployment	0.3276173	0.0839355	9.51×10^{-5}
Wealth	0.1320384	0.0062183	$< 2 \times 10^{-16}$
YM_Count	0.0023009	0.0001409	$< 2 \times 10^{-16}$

Table 4: Coefficients

5 Conclusions and Discussion)

Our analysis reveals significant correlations between the incidence of burglaries in Chicago and two key variables: the proportion of young males within the population and the standard population. This underscores the influence of demographic composition on crime rates within the city. Moreover, our examination of monthly data suggests that temperature patterns play a crucial role in predicting burglary occurrences. Specifically, we find that variations in weather conditions can serve as reliable indicators of fluctuations in burglary rates throughout the year, highlighting the importance of considering environmental factors in crime analysis. Finally, our investigation into the distribution of burglary counts indicates evidence of over-dispersion when the model is either null or saturated, underscoring the need for robust statistical methods to account for the complex variability inherent in crime data. These findings collectively contribute to a deeper understanding of the multifaceted dynamics underlying burglary patterns in urban settings like Chicago.

5.1 Limitations

The biggest limitation is that we cannot compare our models directly due to differing response variables. This makes it hard to know exactly which model is the best outside of more subject means. Also, the pattern in Figure 3 showed a total decrease over time. We did not explore the impact of year within the cycles of the crimes in Chicago and it may have been more impactful in the GLMMs we were using.

We recommend exploring spatial effects on Chicago crime as 56–65% of the total variability in violent crime incidents can be attributed to street segments in Chicago [9]. We did not explore it but it is clear that it may account for our overdispersion.

5.2 Ethical Implications

”Stop and frisk” is a policing tactic used by policing agencies, primarily in low-income, high-crime areas. In the early 2010s, over 684,000 stop-and-frisk encounters were conducted in one year in NYC, with the majority of those stopped found innocent. Furthermore, 87 percent of those targeted were Black or Latinx [10]. This practice perpetuates a cycle where policing actions generate data justifying further policing, resulting in the incarceration of many individuals, often from impoverished neighborhoods, and disproportionately affecting Black and Hispanic communities due to the strong correlation between geography and race in segregated cities [10].

The Chicago Police Department mission statement states that it is ”committed to protect the lives, property, and rights of all people, to maintain order, and to enforce the law impartially” [11]. The ”Stop and Frisk” is just one example of how policy using the data and models we have can do more harm than good for the community by being partial. The profession of the Police Department is contingent on the actions of Chicago PD and the reactions of the people they serve. As a public service organization, it is important they focus on the impact of our research and how it impacts policy and decisions so that it avoids creating ethical dilemmas regarding fairness, justice, and the protection of individual rights.

5.3 Recommendations

With the context of the limitations and the ethical impacts of our research, our recommendations are:

1. Policy makers can develop their community programs to engage positively with the young male population in Chicago and reallocate resources for rising burglary rates in the summer.

2. Examine ways to combat the dissonance between a rising population and crime. There is a factor that is decreasing crime over time despite population increasing nonstop [12]. It is in our best interests to know what is the cause.

3. For further caution, the overdispersion in burglary count insinuates hotspots or particular surges of burglaries in Chicago. This could mean urban development, political centers, housing stability, etc. These can all impact how the crime is distributed.

6 Appendix A

```r
{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
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)

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")


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))
```

```r
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))

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")

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)
)


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

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

plot(block_data$population, block_data$wealth, xlab = "Population", ylab = "Wealth", main = "Population vs Wealth")


hm <- glm(total_crimes ~ 1, data = block_data, family = gaussian(link = identity))
#Goodness of fit test

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

#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
    )
  }
}

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)

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

```

```
glm5 <- glm(total_crimes ~ ym_count + population, data = block_data, family = gaussian(link = identity))
#Goodness of fit test

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



#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 = "blue", color = "white") +
  labs(title = "Histogram of Crime Counts", x = "Crime Count", y = "Frequency") +
  theme_minimal()


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

#None of these models work so far regarding the goodness of fit. We will try to use a negative binomial model instead. Also, there is a mass amount of zeroes present in the data, so we will try to use a zero-inflated poisson model instead.

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

#Goodness of fit test

1-pchisq(nb$deviance, nb$df.residual)
zip_null <- zeroinfl(count ~ 1, data = pdf, dist =  "poisson")


zip <- zeroinfl(count ~ ym_count + population + unemployment + wealth, data = pdf, dist =  "poisson")
summary(zip)


zip_1 <- zeroinfl(count ~ ym_count + population + unemployment, data = pdf, dist =
```

```r
"poisson")
summary(zip_1)
#Goodness of fit test

zip_2 <- zeroinfl(count ~ ym_count + population, data = pdf, dist =  "poisson")
summary(zip_2)


logLik(zip)


logLik_full <- logLik(model)  # Log-likelihood of the full model
model_null <- zeroinfl(count ~ 1, data = pdf, dist = "poisson")  # Fit a null model
logLik_null <- logLik(model_null)  # Log-likelihood of the null model

pseudo_R2 <- 1 - (logLik_full/logLik_null)
pseudo_R2


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)


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

```r

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")
```

Works Cited

[1] Steve Hendershot. The inequality of safety. *Crain's Chicago Business*, October 24 2022.

[2] Simon Field. The effect of temperature on crime. *The British Journal of Criminology*, 32(3):340–351, 1992.

[3] Keith R. Ihlanfeldt. Neighborhood crime and young males' job opportunity. *The Journal of Law and Economics*, 49(1):249–283, 2006.

[4] Chester L. Britt. Reconsidering the unemployment and crime relationship: Variation by age group and historical period. *Journal of Quantitative Criminology*, 13:405–428, 1997.

[5] Wendy C. Regoeczi. The impact of density: The importance of nonlinearity and selection on flight and fight responses. *Social Forces*, 81(2):505–530, 2002.

[6] Brian Christens and Paul W. Speer. Predicting violent crime using urban and suburban densities. *Behavioral Sciences and the Law*, 14(2):113–128, 2005.

[7] R. Ram. Population increase, economic growth, educational inequality, and income distribution: some recent evidence. *Journal of Development Economics*, 14(3):419–428, 1984.

[8] Jay M Ver Hoef and Peter L Boveng. Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88:2766–2772, 2007.

[9] Christina Schnell, Anthony A. Braga, and Eric L. Piza. The influence of community areas, neighborhood clusters, and street segments on the spatial variability of violent crime in chicago. *Journal of Quantitative Criminology*, 33:469–496, 2017.

[10] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group (NY), 2016.

[11] City of Chicago. Police - Mission. `https://www.chicago.gov/city/en/depts/cpd/auto_generated/cpd_mission.html`, September 2009. Accessed on 7 May 2024.

[12] MacroTrends. Chicago metro area population 1950-2024. https://www.macrotrends.net/global-metrics/cities/22956/chicago/population. Retrieved 2024-05-06.