UNITED STATES MILITARY ACADEMY

FINAL PROJECT

MA478: GENERALIZED LINEAR MODELS

SECTION H2

COL NICHOLAS CLARK

By

CADET COOPER KLEIN, '24, CO B3

WEST POINT, NEW YORK

7 MAY 2024

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

_____ I DID NOT USE ANY SOURCES OR ASSISTANCE REQURING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE:____

Chicago PD: Understanding Burglaries

Cooper Klein

May 2024

Abstract

Burglaries in Chicago are a trend plaguing the city's neighborhoods. Our study looks at burglaries by zip code, and leverages Census statistics to include population, median earnings, level of education, and gender ratio. We also consider if day of the week has any impact on burglaries. Ultimately, we build and evaluate three models to predict burglaries, recommend varying police presence by day, and suggest collecting more covariates on zip codes.

***Keywords* :** Crime, Chicago, Burglaries by zip code, Inferential models

1 Introduction

Our research is working to better allocate Chicago PD (CPD) resources throughout the city. Previous research has largely investigated how to reduce public violence and homicide. [RS05] However, this study will approach the crime issue by considering how to prevent burglaries.

To help inform the CPD, we will investigate three guiding questions throughout this research:

1. Are socioeconomic factors associated with the number of daily burglaries?
2. Is there an effect of zip code on daily burglaries that the other features do not account for?
3. Is there an impact of day of week on burglaries? Should additional officers be hired part time?

We will build a unique model predicting count of burglaries across zip codes and days to answer each question. For each model, we will evaluate its performance, and draw any inference. Lastly, we will make formal recommendations for updated CPD policy.

2 Literature Review

Approaching the Chicago crime problem, researchers often take one of three main approaches:

- An analysis of social trends supported by descriptive statistics.
- Building inferential models to understand effects of spatial, temporal or social covariates in Chicago.
- Building predictive models (often using machine learning) to make predictions on crime related activity.

Early research approaching Chicago crime often took the first approach, with researchers at the University of Illinois at Chicago looking at homicide rates in conjunction to changes in strategies from the CPD [RS05]. Another researcher working for the CPD looked at how crime statistics could be effectively mapped to better inform legislature on trends[BM98].

Recent research has shifted toward a modeling approach, with studies investigating the spatial effects of Chicago crime. One example is (Ha et al., 2021) where the effect of green space distribution on crime was investigated [HD24]. Another example is the advisor of this study, (Clark and Dixon, 2019), who investigated different spatial-temporal models to capture trends on burglaries [CD19]. Additionally, the expansion of machine learning techniques has led to sophisticated predictive models. An example of this is (Safat et al., 2021) who used methods such as k-nearest neighbors and XGBoost to predict the overall crime rate of Chicago [SA21].

This research will follow the second approach. With our goal to inform CPD of different covariate effects, we will build inferential models similar to what is seen in (Clark and Dixon, 2019).

3 Methodology

To best inform the CPD on effective policing, we sourced the most recent data available. We first will explore trends present in the data, and discuss the statistic methods used to address these trends. We then will present three models that each directly address one of our guiding questions.

3.1 Data Exploration

Our dataset looks at burglaries across the 58 zip codes in Chicago. We gathered 2022-2023 crime data from the City of Chicago Data Portal [oC]. Specifically for this research, we filtered to only look at crimes listed as a burglary. For each zip code, we collected data from the United States Census Bureau.[Bur]. The statistics collected on each zip code are listed in Table 1:

Table 1: Zip Code Statistics

Statistic	Mean	1st quartile	3rd quartile
Population Size	47535	30222	65898
Median Earnings	$56212	$37799	$73494
Degree Ratio (Proportion with degree above high school)	0.37734	0.20650	0.54615
Gender Ratio (Proportion male)	0.4790	0.4494	0.5044
Degree Ratio (Proportion with degree above high school)	0.37734	0.20650	0.54615
Total Burglaries in 2022-23	257.9	116.2	335.2

Ultimately, the variable we are trying to predict is the count of burglaries by zip code (displayed in Figure 1) and burglaries day of the week (displayed in Figure 2). Burglaries by day of the week is the count of burglaries in a zip code aggregated across the seven respective days recorded.

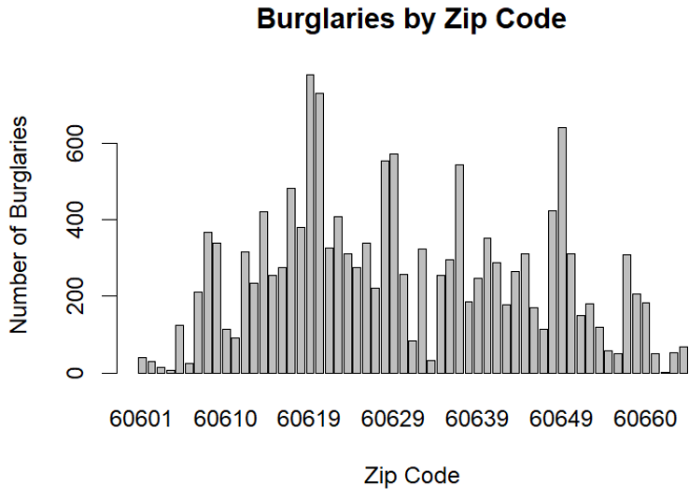

Figure 1: Count of Burglaries by Zip Code

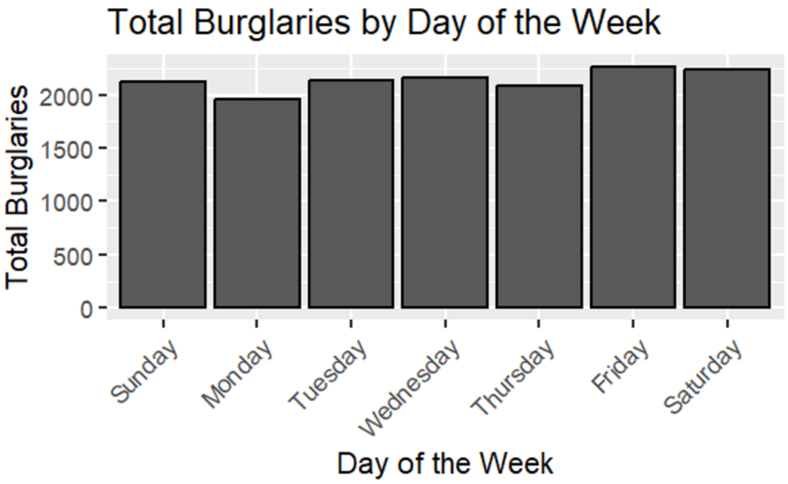

Figure 2: Count of Burglaries by Day of the Week

From Figure 1, we see a clear discrepancy in the count of burglaries occurring in different zip codes across two years. Models 1 and 2 investigate if this variance is caused by a known covariate, or random variance between zip codes. Figure 2 displays a potential discrepancy between burglaries on different days of the week. Monday and Thursday have lower burglary activity while Friday and Saturday appear to have more burglary activity. Model 3 will investigate this significance.

Additional data exploration is provided in the appendices. Figure 3 shows a scatterplot of burglary count by population. The positive linear relationship shown in Figure 3 justifies the use of an offset term for population in each of the models presented. Intuitively, it makes sense that zip codes with more residents will have more burglaries. Figure 4 shows a scatterplot of burglary count by the median earnings of the zip code. Displayed in Figure 5 is a negative correlation between burglary count and the number of males per 100 females. Displayed in Figure 6 is a negative correlation between burglary count and the ratio of residents with a degree above high school. It is important to note, that following this data exploration, all covariates are standardized on the same scale before fitting the models in the following section.

3.2 Model 1: Poisson with Covariates

Model 1 seeks to answer the question, Are socioeconomic factors associated with the number of burglaries? This model will use the three covariates provided by the census, median earnings, sex proportion, and degree proportion.

Model 1, like all three models, predicts on the number of burglaries per zip code per day of the week. This ultimately being a count variable, all three models will assume a poisson distribution.

An offset term, as seen in all three models, is a method of adjusting for the population differences of each zip code. As seen in Figure 3, zip codes exposed to higher population tend to have more burglaries. This is an effect we will control for, as we do not want this to interfere with any inference of our other covariates.

$$\log \lambda_{i,j} = -7.23 - 0.36 \times \text{MedianEarnings} - 0.22 \times \text{SexRatio} + 0.28 \times \text{DegreeRatio} + \log Population \tag{1}$$

$$i = \text{Zip Code}$$
$$j = \text{Day of Week}$$
$$\text{burglaries in zip code by day of week} \sim \text{Po}(\lambda_{i,j})$$

3.3 Model 2: Poisson with Random Effect for Zip Code

Model 2 seeks to answer the question, Is there an effect of zip code on daily burglaries that the other features do not account for? Instead of using the Census covariates, this model will instead consider a potential random effect of each zip code. This random effect creates 58 unique groups, one for each zip code, and considers if there is any spatial trend not captured by the covariates in Model 1.

If we do not find significance in this random effect, we know the three covariates used in Model 1 can adequately inform the CPD of risk factors associated with zip codes. However, if we do not find significance we may recommend looking at additional covariates.

It is important to note that we assume a normal distribution for the random effect of zip code, as we do not have any additional information that would lead us to use a different distribtion.

$$\log \lambda_{i,j} = -7.25 + \psi_i + \log Population \tag{2}$$

$$i = \text{Zip Code}$$

$$j = \text{Day of Week}$$
$$\text{burglaries in zip code by day of week} \sim \text{Po}(\lambda_{i,j})$$
$$\psi_i \sim N(0, \sigma_u^2)$$

3.4 Model 3: Poisson with Day Fixed Effect

Model 3 seeks to answer the question, Is there an impact of day of the week on burglaries? Should additional officers be hired part time? In Model 3, we look to temporal trends to see if we can capture any additional variance in burglary count. This model assumes a fixed effect for each day. Assessing each day of week through a fixed effect will allow for direct comparison of burglary trends on a Monday with those of Tuesday. Understanding these fixed effects could lead us to recommend additional officer presence aligning with temporal trends.

It is important to note that we use Friday as the baseline day of the week. Therefore, when a day has a negative parameter, it can be interpreted as having a lower chance of burglary than on a Friday.

$$\log \lambda_{i,j} = -7.20 + \psi_i - 0.144 \times \text{monday} - 0.053 \times \text{tuesday} - 0.043 \times \text{wednesday} - 0.080 \times \\ \text{thursday} - 0.007 \times \text{saturday} - 0.063 \times \text{sunday} + \log \mathit{Population}$$

$$i = \text{Zip Code}$$
$$j = \text{Day of Week}$$
$$\text{burglaries in zip code by day of week} \sim \text{Po}(\lambda_{i,j})$$
$$\psi_i \sim N(0, \sigma_u^2)$$
$$\text{day} = \begin{cases} 1 & \text{if it is respective day} \\ 0 & \text{otherwise} \end{cases}$$

4 Results

For Model 1, the fitted parameters for median earnings and sex ratio follow what was observed in the data exploration. As the median earnings and number of males of a zip code increases, the count of burglaries decrease. Alternatively, as the level of education of a zip code increases, the count of burglaries is expected to increase according to Model 1. This trend does not follow what is expected intuitively or from the data exploration.

For Model 3, the fitted parameters for each day follow what is expected from Figure 2. Friday was automatically set as the baseline as it is the day burglaries are most likely to occur. From there, Saturday, Wednesday, Tuesday, Sunday, Thursday, and Monday all respectively follow as the most prominent days for burglaries.

4.1 Model Assessment

Assessing the models overall, we will look at the residuals vs fitted plots for each model. Figures 7, 8, 9 of the appendices display these plots. None of the models show alarming trends. Models 2 and 3 have similar looking plots, potentially indicating a lack of significance from the fixed effect of day. This is something we will further investigate with AIC.

4.2 Model Evaluation

With all three models predicting on the same observational units, we can use AIC to test across models.

Table 2: Performance of Models Quantified by AIC

Model Number	**AIC**
Model 1	4617.3
Model 2	2787.7
Model 3	2770.3

From Table 2, we see Model 3 having the best performance. With this improvement of Model 3 over Model 2, this leads us to believe that day of the week is a significant factor for CPD to consider.

Additionally, Model 2, considering only random effects of zip code outperforms Model 1. This leads us to believe our covariates are not adequate in capturing variance in zip codes. Alternatively, we suggest either treating each zip code as a random effect, or collecting and testing more covariates.

5 Discussion and Conclusion

The largest recommendation we will make to the CPD decision makers is to vary police presence by day of the week. Friday and Saturday are shown to have a significantly higher rates of burglary. This will allow for more effective use of CPD resources and lead to a safer Chicago. Another recommendation we can make is to collect more information on the characteristics of different zip codes.

5.1 Limitations

The largest limitation in this research is the lack of covariates. Increasing the number of predictors in Model 1 may have led to an increase in significance, and possibly led to Model 1 outperforming Model 2. Additionally, data collected from the Census are all estimates within a margin of error. Therefore, the covariates in Model 1 were highly informed estimates, but not exact statistics.

5.2 Ethical Considerations

To approach this problem appropriately, it is important we consider ethical considerations to include gentrification. Currently, we do not have any covariates that will outperform the random effect of zip code. Without clear trends identified from covariates, we do not believe it is ethically sound to vary police presence due to an observed random effect. Therefore, without clear indications of specific covariates, we do not recommend CPD vary police presence by any spatial trends. This is approach aligns with our use of an inferential model, while a purely predictive model may not allow for this human oversight for bias or racism.

Additionally, we would also like to vouch for collection of data within ethical guidelines. While more data on zip codes will help build more sophisticated models, this data must be collected in a fair process.

6 Appendices

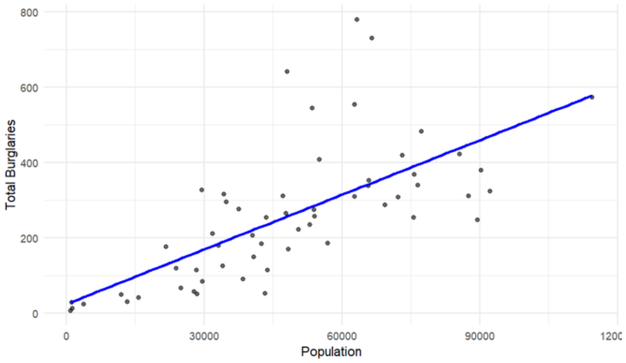

Figure 3: Positive Linear Relationship of Population and Burglaries

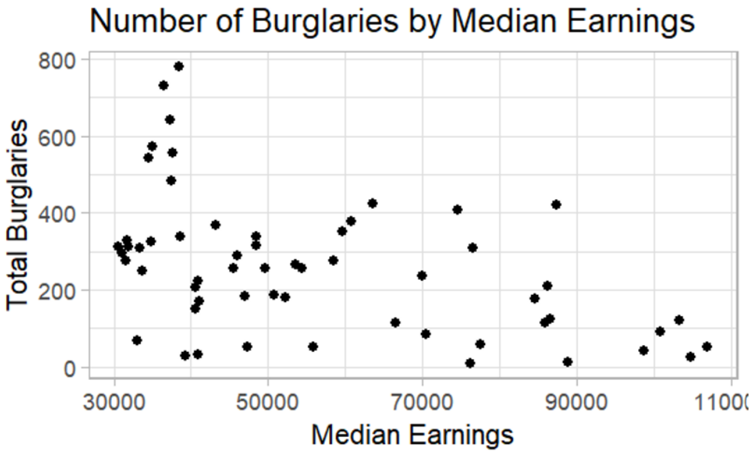

Figure 4: Number of Burglaries by Median Earnings

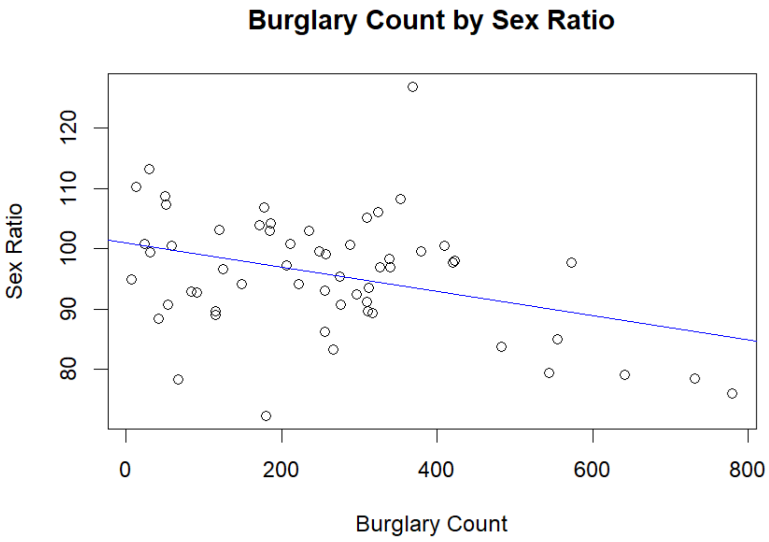

Figure 5: Number of Burglaries by Sex Ratio

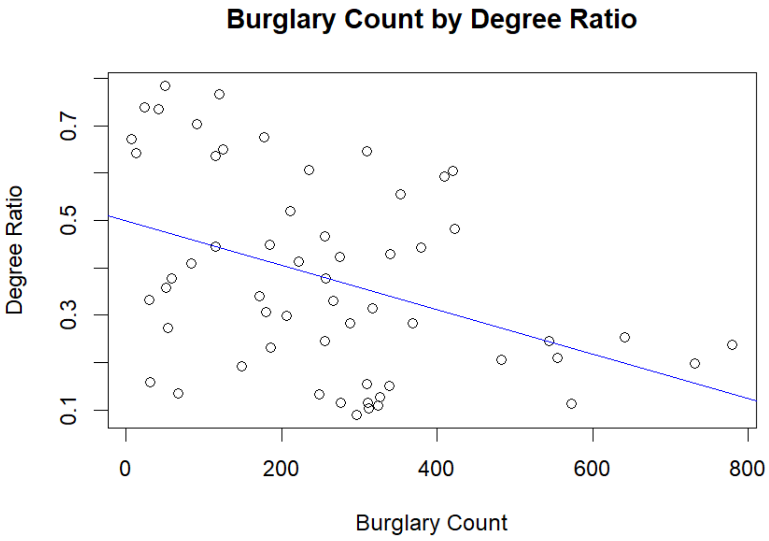

Figure 6: Number of Burglaries by Degree Ratio

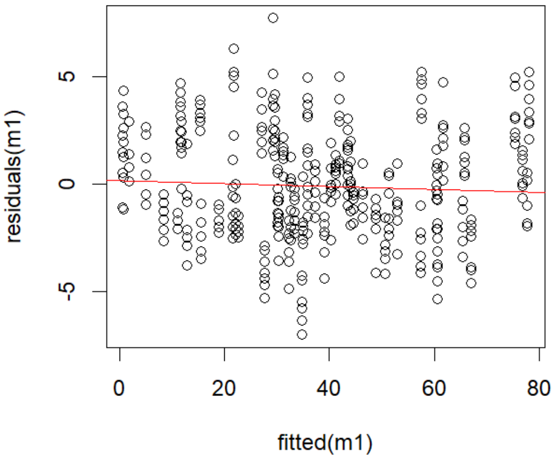

Figure 7: Residuals vs. Fitted Plots: Model 1

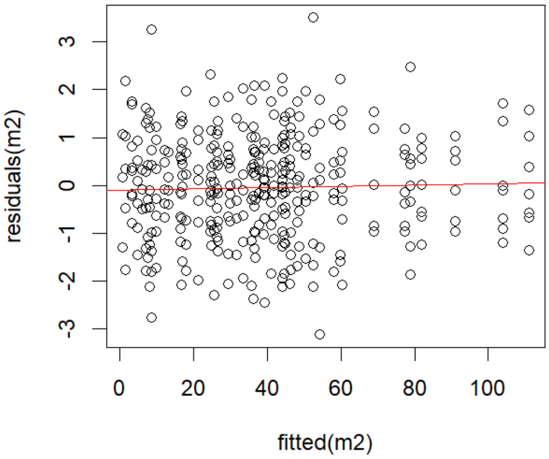

Figure 8: Residuals vs. Fitted Plots: Model 2

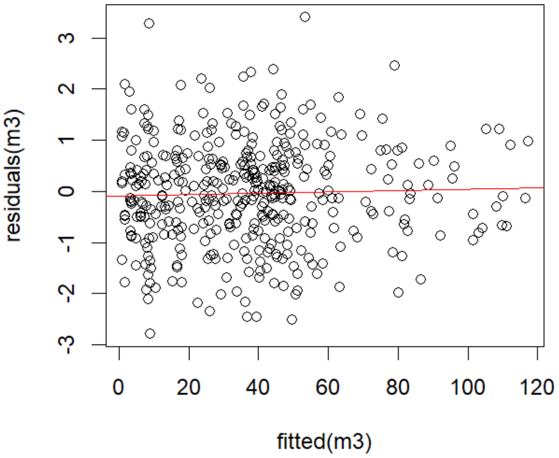

Figure 9: Residuals vs. Fitted Plots: Model 3

7 R-Code

***Please note: ChatGPT was used to merge and clean the datasets downloaded from the Census Bureau. [Ope]

```
library(tidyverse)
library(sf)
library(faraway)



burglary_data <- read_csv('burglary_long.csv')
burglary_short <- read_csv('burglary_agg.csv')

#sex ratio is males per 100f

#exploration of burg by month



# Plot histogram
#barplot(burglary_counts, main = "Chicago Burglaries per Month", xlab = "Month", ylab =
"Number of Burglaries")

#exploration of burg by zip

# Step 1: Subset the 'data_with_zip' dataframe for burglaries
#burglary_data <- data_with_zip[data_with_zip$`Primary Type` == 'BURGLARY', ]

write_csv(burglary_data, "chicago_burglaries.csv")

# Step 2: Count burglaries per zip code
#burglary_counts <- table(burglary_data$zip)

# Step 3: Plot histogram
#barplot(burglary_counts, main = "Burglaries by Zip Code", xlab = "Zip Code", ylab =
"Number of Burglaries")



###distributions###
barplot(burglary_data$population, main = 'Population by Zip Code', ylab = 'Population',
xlab = 'Zip')

fit <- lm(Sex_Ratio ~ total_burglary, data = burglary_short)
plot(burglary_short$total_burglary, burglary_short$Sex_Ratio, main = 'Burglary Count by
Sex Ratio', xlab = 'Burglary Count',ylab = 'Sex Ratio')
abline(fit, col = "blue")

fit <- lm(degree_ratio ~ total_burglary, data = burglary_short)
plot(burglary_short$total_burglary, burglary_short$degree_ratio, main = 'Burglary Count by
Degree Ratio', xlab = 'Burglary Count',ylab = 'Degree Ratio')
abline(fit, col = "blue")


###standarizing covariates###

#earnings
mean_value <- mean(burglary_data$Median_earnings_dollars)
sd_value <- sd(burglary_data$Median_earnings_dollars)
burglary_data$std_Median_earnings_dollars <- (burglary_data$Median_earnings_dollars -
mean_value) / sd_value

#sex ratio
mean_value <- mean(burglary_data$Sex_Ratio)
sd_value <- sd(burglary_data$Sex_Ratio)
burglary_data$std_Sex_Ratio <- (burglary_data$Sex_Ratio - mean_value) / sd_value
```

```
#degree ratio
mean_value <- mean(burglary_data$degree_ratio)
sd_value <- sd(burglary_data$degree_ratio)
burglary_data$std_degree_ratio <- (burglary_data$degree_ratio - mean_value) / sd_value


summary(burglary_data$Median_earnings_dollars)

####Model 1####

m1 = glm(total_burglaries ~ std_Median_earnings_dollars + std_Sex_Ratio + std_degree_ratio
+ offset(log(population)), data = burglary_data ,family = poisson)

summary(m1)

#comparing model to null model -> model is more effective than nothing
1-pchisq(3119.7-2110.7,3)

#comparing model to saturated model -> model does not adequately capture variance
1-pchisq(deviance(m1), df.residual(m1))

halfnorm(residuals(m1))

deviance(m1)/ df.residual(m1) #variance is bigger than mean

#try negative binomial

library(MASS)

m1nb <- glm.nb(total_burglary ~ Median_earnings_dollars + Sex_Ratio + degree_ratio +
offset(log(population)), data = burglary_data)
summary(m1nb)

m1nb <- glm.nb(total_burglaries ~ Median_earnings_dollars + Sex_Ratio + degree_ratio +
offset(log(population)), data = burglary_long)
summary(m1nblong)

#checking dispersion of poisson vs nb
logLik(m1)
logLik(m1nb)

#dispersion of nb is higher

plot(fitted(m1), residuals(m1))
abline(lm(residuals(m1) ~ fitted(m1)), col = "red")


plot(fitted(m1nblong), residuals(m1nblong))
abline(lm(residuals(m1nblong) ~ fitted(m1nblong)), col = "red")


####model 2####

#lmer
library(lme4)
?lme4

m2 <- glmer(total_burglaries ~ offset(log(population)) + (1|zip), data = burglary_data,
family = poisson)
summary(m2)

m2null <- glm(total_burglaries ~ offset(log(population)), data = burglary_long, family =
poisson)
summary(m2null)
```

```
plot(fitted(m2), residuals(m2))
abline(lm(residuals(m2) ~ fitted(m2)), col = "red")


####model 3####

m3 <- glmer(total_burglaries ~ offset(log(population)) + (1|zip) + day_of ,data =
burglary_data, family = poisson )
summary(m3)


plot(fitted(m3), residuals(m3))
abline(lm(residuals(m3) ~ fitted(m3)), col = "red")
```

References

[BM98] Marc Buslik and Michael Maltz. Power to the people: Mapping and information sharing in the chicago police department. *University of Illinois at Chicago*, 1998.

[Bur] United States Census Bureau. Zip code statistics.

[CD19] Nicholas Clark and Philip Dixon. A class of spatially correlated self-exciting statistical models. *Spatial Statistics*, 2019.

[HD24] Jaeyoung Ha and Lindsay Darling. Is the spatial distribution of urban green space associated with crime in chicago? *Urban Forestry Urban Greening*, 2024.

[oC] City of Chicago. Chicago data portal.

[Ope] OpenAI. chatgpt.

[RS05] Dennis Rosenbaum and Cody Stephens. Reducing Public Violence and Homicide in Chicago: Strategies and Tactics of the Chicago Police Department. *Center for Research in Law and Justice University of Illinois at Chicago*, 2005.

[SA21] Wajiha Safat and Sohail Asghar. Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE Explore*, 2021.