UNITED STATES MILITARY ACADEMY

PROJECT REPORT

MA478: GENERALIZED LINEAR MODELS

HOUR H2

COL NICHOLAS CLARK

By

CADET JOSHUA WONG '24, CO I3

WEST POINT, NEW YORK

07 MAY 2024

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

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

SIGNATURE: _____

Exploring the Impact of Young Males on Burglary Rates in Chicago Census Blocks

CDT Joshua Wong

Abstract

This study delves into urban crime's intricate dynamics by focusing on Chicago's burglary incidents. We use interdisciplinary approaches, statistical methodologies, and criminological insights to inform effective strategies for combating burglary crime and pave the way for future crime-based data collection and analysis. The study explores the influence of demographic factors on burglary rates, particularly the proportion of young males within census block groups. The study uncovers significant associations between demographic variables and burglary incidence through comprehensive data analysis and modeling techniques, including generalized linear models (GLMs) and generalized linear mixed-effects models (GLMMs). Findings indicate a positive correlation between the proportion of young males and burglary rates, highlighting the importance of demographic factors in understanding crime patterns. The GLMM emerges as the preferred model, demonstrating superior performance in capturing the underlying data structure and accurately estimating burglary counts. While emphasizing the significance of demographic factors, the study acknowledges limitations and suggests avenues for future research to enhance model robustness and expand understanding of crime dynamics. Overall, the study offers insights for law enforcement agencies to effectively tailor strategies and interventions and future data collection and interpretation for decision-makers, furthering the ultimate goal of enhancing public safety and reducing crime in urban areas.

Keywords: urban crime, burglary incidents, demographic factors, generalized linear models (GLMs), generalized linear mixed-effects models (GLMMs)

1 Introduction

Urban crime, particularly burglary, poses a persistent challenge in cities worldwide, encompassing economic losses and social instability [1]. Chicago, a city renowned for its cultural diversity and economic vibrancy, grapples with significant crime-related issues, with burglary remaining a focal concern for policymakers and law enforcement agencies looking for a data-driven approach [2, 3]. Often, traditional data-driven approaches to crime modeling overlook nuanced socio-demographic factors that may significantly shape criminal activity patterns [4, 5]. While prior research has explored various determinants of burglary rates, such as socio-economic status and neighborhood characteristics, the specific influence of demographic variables [6], however, the proportion of young males within census block groups, remains more unexplored.

This paper addresses the central problem of understanding the influence of the proportion of young males within census block groups on burglary rates in Chicago. We hypothesize that areas with a higher proportion of young males may exhibit higher burglary rates due to various socio-economic and behavioral factors associated with this demographic group. Moreover, recognizing the dynamic nature of crime patterns, we emphasize the importance of incorporating multiple-year

temporal data to capture the evolving dynamics of burglary incidents and their association with demographic characteristics.

The subsequent sections of this paper are structured to provide a comprehensive examination of the relationship between young male demographics and burglary rates in Chicago. In the literature review section, we synthesize existing research findings to contextualize our study within the broader academic discourse on urban crime dynamics. Following this, the methodology section outlines our approach to data collection, variable selection, and statistical techniques employed in the analysis. The experimentation and results section presents empirical findings, including descriptive statistics, regression analyses, and findings from our models. Finally, the discussion and conclusions section offers insights into the theoretical implications of our findings, identifies potential next steps, and suggests avenues for future research. Through this structured approach, we aim to contribute valuable insights to criminology, urban studies, and public policy, facilitating a deeper understanding of the complex dynamics of urban crime in Chicago and beyond.

2 Literature Review

Analyzing burglary patterns through statistics and data science has garnered significant attention in criminological research. Understanding the temporal and spatial dynamics of burglary events is crucial for developing effective crime prevention strategies and allocating resources efficiently. Researchers have employed various methodologies to delve into the intricacies of burglary crime, aiming to identify patterns, elucidate influencing factors, and ultimately contribute to the broader discourse on crime prevention and control.

Temporal analysis methods have emerged as fundamental tools for unraveling burglary patterns. Bolt, for instance, contributes to this area by meticulously evaluating the accuracy of existing temporal analysis methods and introducing a novel approach [7]. By emphasizing the importance of precise temporal information in crime analysis, this study aligns with the broader landscape of applied statistics, where techniques for analyzing temporal data have widespread application in diverse domains, including criminology.

In parallel, spatial and temporal hotspot analysis has gained prominence for its utility in identifying areas of concentrated burglary activity. Cai delves into this realm by conducting spatial and temporal hotspot analysis of burglary crime, offering valuable insights for law enforcement agencies to devise targeted prevention and control strategies [8]. This approach resonates with existing data science and spatial statistics research, where hotspot analysis and spatial interpolation methodologies have been extensively utilized to discern crime patterns and inform resource allocation efforts.

Predictive modeling, particularly through logistic regression, has been another cornerstone of burglary crime analysis. Antolos employs logistic regression modeling to investigate factors influencing burglary occurrence probability, shedding light on significant predictors while acknowledging the model's limitations in pinpointing specific locations of criminal activity [9]. Logistic regression, a staple method in applied statistics for modeling binary outcomes, such as criminal events, underscores the interdisciplinary nature of crime analysis, where statistical methodologies intersect with criminological theories to elucidate complex phenomena.

These studies contribute to the burgeoning literature on burglary crime analysis, offering diverse methodologies to scrutinize burglary patterns from temporal, spatial, and predictive perspectives. While each paper tackles specific facets of burglary crime analysis, they converge on common objec-

tives: understanding patterns, identifying influential factors, and informing evidence-based crime prevention strategies. The interdisciplinary nature of these endeavors underscores the imperative of collaboration across disciplines in addressing real-world challenges in crime prevention and control.

Nevertheless, despite the strides made in burglary crime analysis, challenges persist, including data limitations, model assumptions, and the generalizability of findings. Further research is warranted to validate and extend the insights gleaned from these studies across diverse contexts and populations. This underscores the ongoing imperative for interdisciplinary approaches to crime analysis, where statistical methodologies and criminological insights converge to inform effective strategies for combating burglary crime.

3 Methodology

Our study delves into the complex dynamics of urban crime, focusing on the problem of modeling burglary incidents in Chicago. The overarching goal is to elucidate the influence of demographic factors on burglary rates, particularly the proportion of young males within census block groups. Our methodology first explores the initial insights within the data to see what distribution the data would most likely come from. Our choice of distribution will be the foundation for choosing our generalized linear model (GLM) to fit the data.

Before modeling, we conducted exploratory data analysis to gain insights into the dataset's structure and characteristics. We imported various datasets, including crime incidents over time and demographic variables (population, unemployment rate, and wealth) grouped by census block. This heterogeneous data was integrated and processed to create a unified dataset for analysis. After unifying the data set, we looked into various explained and unexplained variations to consider the various external effects on burglary rates other than the data that we have.

3.1 Source of Variation Diagram

The sources of variation diagram in Figure 1 presented in this inferential analysis provides a comprehensive overview of the factors influencing burglary rates in Chicago.

<table>
<tr>
<td><u>Observational Units</u>
• Chicago census block groups per month</td>
<td rowspan="2"><u>Explained Variance</u>
• Population
• Seasonality
• Wealth
• Unemployment
• Young Males</td>
<td rowspan="2"><u>Unexplained Variance</u>
• Education levels
• Economic state
• Known home presence
• Gang presence
• Gentrification
• Police Presence</td>
</tr>
<tr>
<td><u>Response Variable</u>
• # of burglaries in Chicago block groups in each month</td>
</tr>
</table>

Figure 1: Sources of Variation Diagram

Figure 1 shows the sources of variation in the number of burglaries occurring in Chicago census block groups every month. The first column specifies the observational units and the Chicago census block groups and indicates monthly data collection. The second column identifies the factors

contributing to the explained variance in burglary rates, which include population size, seasonal variations, economic wealth, unemployment rates, and the presence of young males in the area. These factors collectively explain a portion of the variability in burglary rates. The third column delineates sources of unexplained variance, such as education levels, overall economic conditions, known home presence (likely referring to occupancy rates), gang activity, gentrification trends, and police presence. These variables may contribute to the unpredictability of burglary rates beyond the factors accounted for in the explained variance. This table provides a structured overview of the multifaceted nature of Chicago burglary rates, highlighting the known determinants and the residual variability that remains unexplained by the listed factors.

3.2 Data Methodology

In addition to the sources of variation diagram, we explored temporal patterns of burglary incidents over 72 months, identifying fluctuations and seasonality in crime rates. Additionally, we examined the distributions of key variables, such as population, unemployment rate, and the proportion of young males, to understand their variability across census block groups and over time.

Data preprocessing played a crucial role in preparing the dataset for modeling. We cleaned data, handled missing values, and standardized variable names and formats. Moreover, we engineered new features, such as the percentage of young males within census block groups, by aggregating and transforming raw data to get at our response variable. Finally, we conducted correlation analyses among demographic variables to mitigate potential multicollinearity issues and identify highly correlated predictors. The informed decisions we made about variable selection and transformation of our data were to allow our models to fit and to preserve the ability to learn about the association between our response, the number of burglaries, and our explanatory variable of interest, the proportion of young males.

3.3 Modeling Approch

Our modeling approach used various GLMs to capture the relationship between burglary rates and demographic predictors. We initially fitted Poisson GLMs to account for the count nature of burglary data. We explored various model specifications incorporating demographic variables, allowing our temporal data to be represented by explicit fixed effects.

We then considered zero-inflated Poisson (ZIP) models. These models incorporate a mixture of a Poisson distribution and a point mass at zero, allowing for more flexible modeling of zero-inflated count data. In the case of burglaries, we considered the binomial portion of the ZIP model that generates the extra zeros to represent the factor of whether or not burglaries were reported in a specific census block. A lack of burglary reports could be one of the reasons there were so many census blocks and zero burglaries recorded.

Finally, we explored the utility of generalized linear mixed-effects models (GLMMs) to account for the temporal portion of the data. By incorporating mixed effects to represent the changes over time, we hoped to distill better young males' effect on the number of burglaries, compared to our original Poisson GLM, which used fixed effects for our temporal portion of the model.

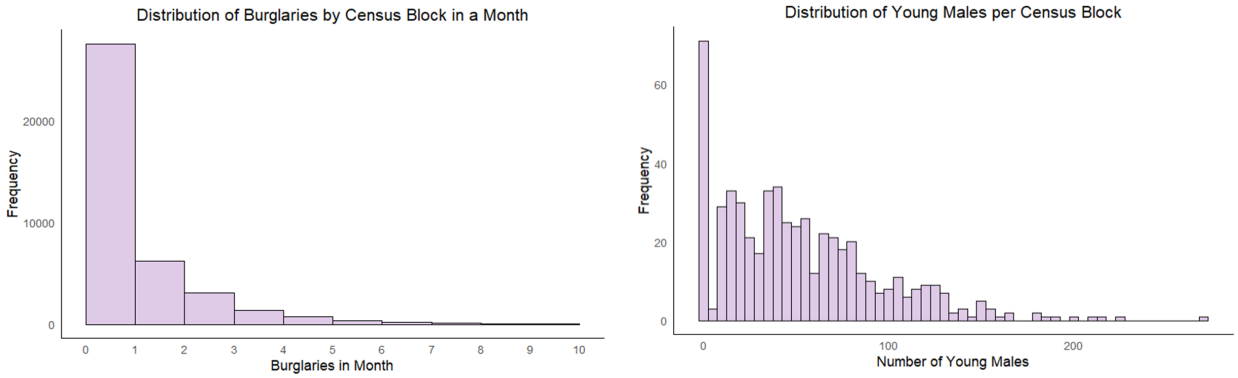
Our model selection involved comparing competing models' goodness of fit and predictive performance. We employed techniques such as likelihood ratio tests and the Akaike Information Criterion (AIC) to evaluate model adequacy and complexity. Our methodology encompasses modeling burglary rates in Chicago, leveraging advanced statistical techniques and careful data analysis by systematically exploring the relationships between demographic factors and crime incidence.

4 Experimentation and Results

4.1 Data Exploration

The dataset we used contained 39744 observations and eight different variables. The variables include the census block number, the population by census block group, the percentage unemployed by census block group, the centered and scaled average family income by census block group in 2015 dollars, the number of young males by census block group aged 15-20-year-olds, the month, the year, and the number of burglaries that took place on that month of the year. To understand how the burglaries in the dataset were distributed over time and space, we first looked at the distribution of burglaries for each census block month. Figure 2a shows a histogram of burglaries per each census block month, demonstrating a larger number of lower values, skewed right.

Next, we looked into the distribution of our main explanatory variable, the number of young males in a census block. Figure 2b shows the histogram of the number of young males over all of the census blocks. Again, we notice many zero values, denoting that many census blocks reportedly have zero young males.



(a) Bugalies by Census Block Month

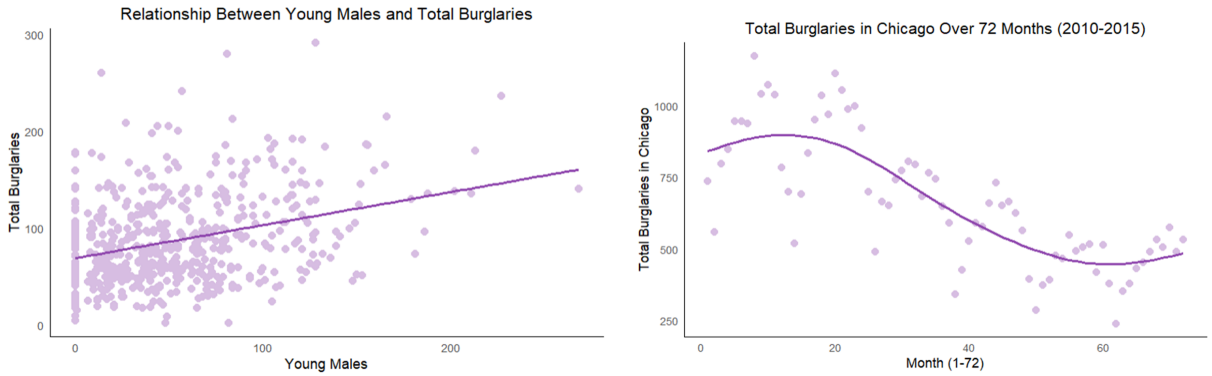
(b) Young Males per Census Block

Figure 2: Histogram of Response and Explanatory Variables

We next investigated the relationship between total burglaries in Chicago and the number of young males in a census block. We also look at the total burglaries over time. While total burglaries in Chicago are not the response variable, the following figures provide a better idea of how the burglaries in Chicago while controlling for time or census block. Figure 3a comes from a dataset where the burglaries are summed over all the months so that we only look at the total burglaries in each census block over the 72 months. However, the response variable differs. We observe a generally positive trend between total burglaries and the number of young males in each census block, providing insight on a different level about the relationship between young males and burglaries.

Figure 3b provides another look at total burglaries; rather than looking at total burglaries over time, Figure 3b looks at the total burglaries in Chicago over 72 months. In this case, we see a seasonal and yearly trend of the total burglaries in Chicago over time. We observe an overall decrease in burglaries over the 72 months and a non-linear curvature pattern with each of the seasons, where the total burglaries peak in the summer.

Finally, Figure 4 provides a correlation matrix between the numerical variables in the data set.



(a) Young Males vs Total Burgalies

(b) Total Burgalies Over 72 Months

Figure 3: Plots of Total Burgalies Against Young Males, and Over Time

The figure shows that Burgalires experiences a generally positive correlation with the numerical variables wealth, unemployment, population, and the number of young males. Statistical models have yet to determine whether these correlations are significant associations. Additional notable findings are that, from the correlation plot, we find population and total number of burglaries to have a positive correlation. We also find that population and wealth are also highly positively correlated.

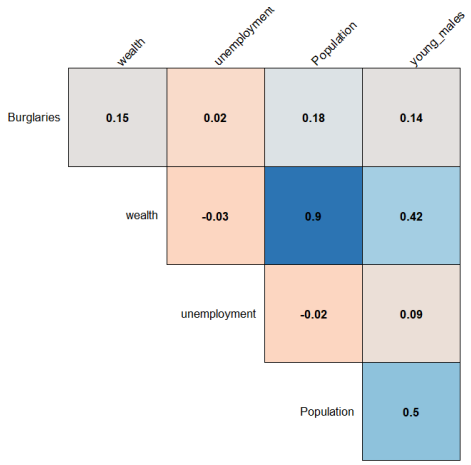


Figure 4: Correlation Matrix of Numerical Variables

4.2 Data Preparation

To prepare the data, we combined the information about each census block (wealth, population, number of young males, unemployment) with the burglary data so that each observation represented the number of burglaries in a given census block in a given month. We also created two other variables for a year and month for each census block month observation. Additionally, because our initial variable for young males represented the total count of young males in a census block, we transformed the variable into our response variable by dividing it by the population of the census

block. Therefore, our response variable became the proportion of young males in a census block relative to the population of the census block. Finally, we standardized all of your variables to include the proportion of young males in a census block, the population, and the unemployment rate in a census block. We did not standardize wealth as it was our understanding that the wealth variable came as a centered and scaled family average income by census block. In this case, there were no missing values to impute.

4.3 Model Building and Selection

For our models, we looked at three general types of GLM to fit our data: a Poisson, a Zero-Inflated Poisson, and a Generalized Mixed Effects Model (GLMM), where our mixed effects accounted for the temporal portion of our data. Consider the following covariates for our models:

Consider the following response variable:

$$Y_{i,j,k} = \text{burglaries in census block i, year j, and month k.}$$

Now for the covariates, let $x_{5,1}, x_{5,2}, \ldots, x_{5,12}$ be indicator variables for each month:

$$x_{5,k} = \begin{cases} 1 & \text{if month is } k \\ 0 & \text{otherwise} \end{cases}$$

Now, the representation becomes:

$$\begin{aligned} x_{1,i} &= \text{percent young males in population of census block i} \\ x_{2,i} &= \text{population of census block i} \\ x_{3,i} &= \text{unemployment in census block i} \\ x_{4,j} &= \text{year j} \\ x_{5,1}, x_{5,2}, \ldots, x_{5,12} &= \text{indicator variables for each month} \end{aligned}$$

The month is treated as a categorical variable, while the rest are numerical. Also, please note that we decided to remove wealth as an explanatory variable to reduce colinearity with the population. In the Poisson model portion of this report, we explain why we chose population over wealth when selecting our initial best Poisson model.

4.3.1 Poisson Model

First, we build out our Poisson models as they are suited for counting data analysis, making them ideal for investigating burglary rates. It is also the simplest model we can train, making it a good baseline for our other models. We trained three different models for the Poisson family: the model represented above, a model with wealth, and a model with wealth and without population. We did this to compare the significance of wealth vs. population in the model. We compared our three Poisson models with and without covariates population and wealth using an ANOVA Chi-squared test. The ANOVA Chi-squared test for nested models provided a p-value greater than 0 when

comparing models with and without wealth.05, meaning that we cannot reject the null hypothesis that the simpler model is better and conclude that the variable wealth is insignificant. On the other hand, when comparing models with and without population using the ANOVA Chi-square test, we got a p-value less than 0.01, showing that population was a significant covariate in our model. Therefore, due to the colinearity between population and wealth variables, we included only the population variable and removed the wealth variable from this and future analyses. Therefore, our final Poisson model is written in Equation 1, with all variables except wealth. For this reason, we also do not include wealth in our future models. Table 2 in the Appendix shows the coefficient values for our selected Poisson GLM.

$$\begin{aligned} Y_{i,j,k} &\sim \text{Po}(\lambda_{i,j,k}) \\ \log(\lambda_{i,j,k}) &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,1} + \beta_6 x_{5,2} + \ldots + \beta_{16} x_{5,12} \end{aligned} \tag{1}$$

4.3.2 Zero-Inflated Poisson Model

The next model we built was the ZIP model. In urban settings like Chicago, certain census blocks may exhibit an excess of zero burglary counts, possibly due to factors not captured by the Poisson model. Therefore, the Zero-Inflated Poisson model provides a more nuanced approach by accounting for excess zeros in the data. The model comprises two components: a binary process (logit function) determining the likelihood of observing zero burglaries versus any positive count and a Poisson process modeling the counts of burglaries for non-zero observations. The intuition behind using a ZIP model is that there may be many instances of zero burglaries reported within our data. In this case, we hypothesize that an additional mechanism may motivate people in a specific census block month not to report burglary cases. Therefore, we model this potential explanation by using a ZIP model.

We compare ZIP models with and without the covariates month and young males in the zero-inflated portion of the model using an ANOVA Chi-squared test. We thought that possibly the proportion of young males in a census block and the month might not affect whether or not burglaries were reported in an area. However, both variables proved statistically significant, with p-values less than 0.01 when comparing nested ZIP models with and without both covariates. Our chosen ZIP model is represented in Equation 2, containing all of the covariates in the Poisson portion and the zero-inflated binomial portion of the models. Additionally, Tables 3 and 4 in the Appendix show the coefficient outputs for the Poisson and binomial parts of the ZIP model, respectively.

$$\begin{aligned} Y_{i,j,k} &\sim \text{ZIP}(\pi_{i,j,k}, \lambda_{i,j,k}) \\ \text{logit}(\pi_{i,j,k}) &= \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{3,i} + \gamma_4 x_{4,j} + \gamma_5 x_{5,1} + \gamma_6 x_{5,2} + \ldots + \gamma_{16} x_{5,12} \\ \log(\lambda_{i,j,k}) &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,1} + \beta_6 x_{5,2} + \ldots + \beta_{16} x_{5,12} \end{aligned} \tag{2}$$

4.3.3 Genralized Mixed Effects Model

The Generalized Linear Mixed Effects Model (GLMM) extends the Poisson model by capturing potential unobserved variation across census blocks over time. We first introduce random effects at the census block level (i), represented by γ_i. With a random effect at the block level, we capture the additional variation between census blocks not captured by unemployment, population, and

proportion of young males. Additionally, by introducing random effects at the temporal month-year level (j,k), the model acknowledges that burglary rates may vary between census blocks and over different years. Therefore, we introduce the random effect to the model $\delta_{j,k}$ to account for this variation over different month-years. This approach allows us to control for potential temporal variations in burglary rates while examining the impact of explanatory variables, more specifically, the association between the proportion of young males and burglaries.

However, to select the final GLMM, we compared three different variations of the GLMM. The first GLMM only had the random effect of month-year ($\delta_{j,k}$), the second GLMM only had the random effect of census block (γ_i), and the third model hand random effects for both the month-year and the census block. We fit all of our models via Laplace Approzximation to approximate the maximum likelihood and compared each model's calculated Akaike Information Criterion (AIC) to select the best GLMM for later comparison with the other models. In this case, our third GLMM fit the best data, with an AIC of 111489.2, lower than the AICs for the first and second GLMMs, with AICs of 119533.5 and 116567.0, respectively. Therefore,e for our GLMM, we build a GLMM with two random effects for month-year and census block; equation 3 is the final GLMM model we chose, with both random effects. We removed the month and year covariates with the understanding that the variation between months and years was captured within the random temporal effect of the model. The tables 6 and 5 demonstrate the summary statistics and coefficients for the fixed effects and random effects of our selected GLMM, respectively.

$$\begin{aligned} Y_{i,j,k} &\sim \text{Po}(\lambda_{i,j,k}) \\ \log(\lambda_{i,j,k}) &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \gamma_i + \delta_{j,k} \\ \gamma_i &\sim \mathcal{N}(0, \sigma^2_{\gamma_i}) \\ \delta_{j,k} &\sim \mathcal{N}(0, \sigma^2_{\delta_{j,k}}) \end{aligned} \tag{3}$$

4.4 Model Evaluation

The final part of our analysis compares our three chosen models within each general category: Poisson GLM, ZIP GLM, and GLMM. The initial assessment involved a Chi-squared goodness-of-fit test on the basic Poisson GLM, revealing that the Poisson GLM failed to capture the response variable's distribution adequately. The goodness of fit test failed, with a p-value less than 0.01, meaning we reject the null hypothesis that the model is not a bad fit for the data. Additionally, we look at the mean-to-variance ratio of the model and get an estimate of 1.5. One of the assumptions that come with a Poisson model is that the mean-to-variance ratio is 1. However, a ratio of 1.5 demonstrates that our model is slightly overdispersed. The failure of the goodness of fit test and the overdispersion suggested limitations on the Poisson model's ability to represent the underlying data structure accurately.

Addressing the challenge of predicting zero counts within the dataset, we employed the ZIP GLM due to its capability to handle excess zeros compared to the Poisson distribution. However, upon evaluation, the ZIP GLM did not outperform the Poisson GLM in accurately estimating predicted zeros seen in the values listed in Table 1. This outcome raised doubts about the ZIP GLM's effectiveness in addressing excess zeros in our data. We also compared the AIC values to assess further model performance, as seen in Table 1, with lower AIC values indicating better-fitting models than others in the comparison. Analysis of the AIC values revealed that the GLMM exhibited the lowest AIC value among the three models, followed by the ZIP GLM and then the

Poisson GLM. This finding indicated that the GLMM best balanced model fit and complexity.

	Poisson GLM	ZIP GLM	GLMM
AIC	119603.6	117321.5	111489.2
Predicted Zeros	13373	13310	14752
Actual Zeros	16044	16044	16044

Table 1: Comparison of AIC and Predicted Zeros for Different Models

Considering the goodness-of-fit test results, accuracy in predicting zeros, and AIC values, we concluded that the GLMM is the most suitable model for our dataset. Despite its inherent complexity as a mixed model, the GLMM demonstrated superior performance in capturing the underlying data structure while providing the most accurate estimation of zero counts. Therefore, we recommend the GLMM as the preferred model for further analysis and interpretation of our data.

5 Discussion and Conclutions

This study evaluated the suitability of three different models, Poisson GLM, ZIP GLM, and GLMM, for analyzing burglary data in Chicago census blocks over six years from 2010 to 2015. Our analysis revealed that the GLMM best-balanced model fit and complexity, outperforming the Poisson GLM and ZIP GLM based on goodness-of-fit tests and AIC values.

Utilizing the GLMM, we identified the percentage of young males in a census block as a statistically significant explanatory variable in estimating the number of burglaries in a block during a specified period. Our analysis revealed that for each percent increase in the proportion of young males in a census block, the number of burglaries increased by a factor of $e^{0.06}$, which can be interpreted as for every one standard deviation increase in the proportion of young males, there is an associated 1.82 increase in the number of burglaries in a given census block month. This finding underscores the importance of demographic factors in understanding patterns of criminal activity within urban areas.

However, based on our findings, we do not necessarily recommend that law enforcement agencies, such as the Chicago Police Department (CPD), prioritize their presence and resource allocation in census blocks where the percentage of young males is on the rise. While we observe an association between the proportion of young males in a census block and more burglaries, we do not necessarily believe this calls for immediate action but rather additional research. Within our data, we observed many zeros concerning reported burglaries and the number of young males reported in the given census block. In this study we did not investigate potential reasons or associations behind these zero, values beyond the ZIP models, however we belive future work should look more into the effect of potentially lower instances of reporting, even concerning police presence.

In addition to many zero values potentially representing areas of low reporting, we also want to address some additional limitations. Firstly, our analysis is limited to specific information for each census block, which may not capture all relevant factors influencing burglary rates; even with a random effect, it would be better to look to add more specific data to get at more of the unexplained variation. Additionally, our dataset spans only six years, from 2010 to 2015, limiting the temporal scope of our findings. Furthermore, generalizing our results to regions outside of Chicago and beyond the specified period may be challenging and requires caution.

Additional future research could address some of the limitations identified in this study. Investigating additional approximation methods for GLMM, such as Gauss-Hermite Quadrature or Bayesian methods like Integrated Nested Laplace Approximation, could provide more robust estimates and enhance model performance. Furthermore, considering an auto-regressive structure on the time series random component could help account for temporal dependencies in burglary data, offering a more comprehensive understanding of spatiotemporal patterns of criminal activity.

In conclusion, our study highlights the effectiveness of GLMM in analyzing burglary data and underscores the significance of demographic factors in predicting criminal behavior. By understanding the relationship between demographic characteristics and crime rates, law enforcement agencies can better tailor their strategies and interventions to enhance public safety and reduce crime incidence in urban areas. However, further research is needed to address the limitations of our study and advance our understanding of complex interactions influencing crime dynamics.

Works Cited

[1] Joong-Hwan Oh. Social disorganizations and crime rates in U.S. central cities: Toward an explanation of urban economic change. *The Social Science Journal*, 42(4):569–582, December 2005. Publisher: Routledge _eprint: https://doi.org/10.1016/j.soscij.2005.09.008.

[2] Julia Burdick-Will. School Violent Crime and Academic Achievement in Chicago. *Sociology of Education*, 86(4):343–361, October 2013. Publisher: SAGE Publications Inc.

[3] Ayidh alqahtani, Ajwani Garima, and Ahmad Alaiad. Crime Analysis in Chicago City. In *2019 10th International Conference on Information and Communication Systems (ICICS)*, pages 166–172, June 2019. ISSN: 2573-3346.

[4] Dhaifallah M. Alghamdi. *A Data Mining Based Approach For Burglary Crime Rate Prediction.* thesis, University of Illinois at Chicago, May 2017.

[5] Jun Luo. Multi-spatiotemporal patterns of residential burglary crimes in Chicago: 2006-2016. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, October 2013.

[6] Wim Bernasco and Richard Block. Robberies in Chicago: A Block-Level Analysis of the Influence of Crime Generators, Crime Attractors, and Offender Anchor Points. *Journal of Research in Crime and Delinquency*, 48(1):33–57, February 2011. Publisher: SAGE Publications Inc.

[7] Martin Boldt and Anton Borg. Evaluating Temporal Analysis Methods Using Residential Burglary Data. *ISPRS International Journal of Geo-Information*, 5(9):148, September 2016. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Dong Cai and Zimiao Shi. Analysis of Temporal Characteristics of Burglary Crime. *CONVERTER*, pages 697–707, July 2021.

[9] Daniel Antolos, Dahai Liu, Andrei Ludu, and Dennis Vincenzi. Burglary Crime Analysis Using Logistic Regression. In Sakae Yamamoto, editor, *Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business,*, pages 549–558, Berlin, Heidelberg, 2013. Springer.

6 APPENDIX A: Model Output Summary Tables

Poisson GLM with Log Link				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	0.406590	0.017813	22.826	$< 2 \times 10^{-16}$***
perc_ym	0.054195	0.004644	11.670	$< 2 \times 10^{-16}$***
pop	0.197534	0.004117	47.981	$< 2 \times 10^{-16}$***
unemp	0.029073	0.004677	6.216	5.11e-10***
year	-0.165125	0.002745	-60.164	$< 2 \times 10^{-16}$***
month02	-0.362852	0.026377	-13.756	$< 2 \times 10^{-16}$***
month03	-0.057578	0.024245	-2.375	0.0176*
month04	0.035613	0.023684	1.504	0.1327
month05	0.166834	0.022958	7.267	3.67e-13***
month06	0.194710	0.022813	8.535	$< 2 \times 10^{-16}$***
month07	0.231388	0.022626	10.226	$< 2 \times 10^{-16}$***
month08	0.322882	0.022185	14.554	$< 2 \times 10^{-16}$***
month09	0.238160	0.022593	10.542	$< 2 \times 10^{-16}$***
month10	0.268734	0.022442	11.974	$< 2 \times 10^{-16}$***
month11	0.208709	0.022741	9.178	$< 2 \times 10^{-16}$***
month12	0.124390	0.023185	5.365	8.09e-08***

Table 2: Poisson GLM with Log Link

Count Model Coefficients (Poisson with Log Link)				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	0.581993	0.023884	24.368	$< 2 \times 10^{-16}$***
perc_ym	0.037473	0.006425	5.833	5.45e-09***
pop	0.156604	0.005646	27.737	$< 2 \times 10^{-16}$***
unemp	0.044247	0.006458	6.852	7.29e-12***
year	-0.132422	0.003716	-35.636	$< 2 \times 10^{-16}$***
month02	-0.308522	0.037821	-8.157	3.43e-16***
month03	-0.064270	0.032935	-1.951	0.05101*
month04	0.030160	0.031609	0.954	0.34001
month05	0.158685	0.030086	5.274	1.33e-07***
month06	0.163560	0.030047	5.443	5.22e-08***
month07	0.189650	0.029727	6.380	1.77e-10***
month08	0.282419	0.028773	9.816	$< 2 \times 10^{-16}$***
month09	0.244268	0.029416	8.304	$< 2 \times 10^{-16}$***
month10	0.226017	0.029273	7.721	1.15e-14***
month11	0.195333	0.029742	6.568	5.11e-11***
month12	0.091581	0.030900	2.964	0.00304**

Table 3: Poisson Component of ZIP model with Log Link

Zero-Inflation Model Coefficients (Binomial with Logit Link)				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-1.68841	0.09281	-18.192	$< 2 \times 10^{-16}$***
perc_ym	-0.06562	0.02551	-2.573	0.0101*
pop	-0.24800	0.02750	-9.019	$< 2 \times 10^{-16}$***
unemp	0.07491	0.02345	3.195	0.0014**
year	0.16476	0.01400	11.769	$< 2 \times 10^{-16}$***
month02	0.25247	0.13148	1.920	0.0548.
month03	-0.02310	0.12583	-0.184	0.8544
month04	-0.01356	0.11977	-0.113	0.9098
month05	-0.01844	0.11383	-0.162	0.8713
month06	-0.16631	0.11826	-1.406	0.1596
month07	-0.23122	0.11901	-1.943	0.0520.
month08	-0.19074	0.11360	-1.679	0.0931.
month09	0.03155	0.10954	0.288	0.7733
month10	-0.22044	0.11662	-1.890	0.0587.
month11	-0.05145	0.11342	-0.454	0.6501
month12	-0.18276	0.12296	-1.486	0.1372

Table 4: Binomial Component of ZIP model with Logit Link

Random Effects			
Groups	**Name**	**Variance**	**Standard Deviation**
block	(Intercept)	0.2440	0.494
month-year	(Intercept)	0.1218	0.349

Table 5: GLMM Random Effects

Fixed Effects				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-0.008316	0.046524	-0.179	0.85813
perc_ym	0.060408	0.021820	2.769	0.00563**
pop	0.209524	0.021659	9.674	$< 2 \times 10^{-16}$***
unemp	0.015006	0.021904	0.685	0.49331

Table 6: GLMM Fixed Effects

APPENDIX B: Final Project Code

CDT Joshua Wong

2024-04-14

Import Libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# for glms
library(faraway)
# for zip glms
library(pscl)
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
# linear mixed effects models
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
##
## The following object is masked from 'package:lme4':
##
```

```
##     lmList
##
## The following object is masked from 'package:dplyr':
##
##     collapse
```

Import Data

```
# Must Set First, WD Reset for every chunk
setwd("C:/Users/Joshua.Wong/OneDrive - West Point/AY24-2/MA478/Final Project/Data")
all <- read.csv("alldata.csv")
crime <- read.csv("crime.csv")
#Population by Census Block Group
pop <- read.csv("pop.csv")
#Percentage Unemployed by Census Block Group
unemp <- read.csv("unemp.csv")
#Centered and Scaled Average Family Income by Census Block Group (2015 Dollars)
wealth <- read.csv("wealth.csv")
#Number of Young Males by Census Block Group (15-20 yr olds)
ym <- read.csv("ym.csv")
neigh <- readMM("neighborhood.mtx")

pop <- pop %>% rename(pop = x)
wealth <- wealth %>% rename(wealth = x)
unemp <- unemp %>% rename(unemp = x)
ym <- ym %>% rename(ym = x)

head(all)
```

```
##   Column1 count.201001 count.201002 count.201003 count.201004 count.201005
## 1       1            0            1            0            0            0
## 2       2            0            0            1            0            0
## 3       3            0            0            0            0            0
## 4       4            0            0            1            0            0
## 5       5            0            0            0            0            0
## 6       6            0            0            0            0            0
##   count.201006 count.201007 count.201008 count.201009 count.201010 count.201011
## 1            1            1            0            1            1            0
## 2            0            0            0            0            0            0
## 3            0            0            0            0            1            0
## 4            1            0            0            0            0            0
## 5            0            1            0            0            0            0
## 6            0            0            1            0            0            0
##   count.201012 count.201101 count.201102 count.201103 count.201104 count.201105
## 1            0            0            0            0            0            0
## 2            0            0            0            0            0            0
## 3            0            0            0            0            0            0
## 4            0            0            0            0            0            0
## 5            0            0            0            0            0            0
## 6            0            2            0            0            0            0
##   count.201106 count.201107 count.201108 count.201109 count.201110 count.201111
## 1            0            0            0            0            1            0
## 2            1            0            0            0            0            0
## 3            0            1            0            0            0            0
```

```
## 4            0            0            0            0            0            0
## 5            0            0            0            0            0            0
## 6            1            0            0            0            0            0
##   count.201112 count.201201 count.201202 count.201203 count.201204 count.201205
## 1            0            0            0            0            0            0
## 2            1            0            1            0            0            0
## 3            0            0            0            0            0            0
## 4            0            0            0            0            0            1
## 5            0            0            0            0            0            0
## 6            0            0            0            0            1            0
##   count.201206 count.201207 count.201208 count.201209 count.201210 count.201211
## 1            0            0            3            0            1            0
## 2            0            0            0            0            0            0
## 3            0            0            2            0            0            2
## 4            1            0            1            1            0            1
## 5            0            0            0            0            0            0
## 6            0            0            1            0            1            0
##   count.201212 count.201301 count.201302 count.201303 count.201304 count.201305
## 1            0            0            0            0            0            0
## 2            0            0            0            0            0            0
## 3            0            0            0            2            0            0
## 4            2            0            2            0            0            1
## 5            0            0            0            0            0            0
## 6            0            0            0            1            1            0
##   count.201306 count.201307 count.201308 count.201309 count.201310 count.201311
## 1            0            0            0            1            1            1
## 2            0            0            0            0            0            0
## 3            0            0            1            0            1            0
## 4            1            0            0            0            0            0
## 5            0            0            0            0            0            0
## 6            0            0            0            0            0            0
##   count.201312 count.201401 count.201402 count.201403 count.201404 count.201405
## 1            0            0            0            0            0            0
## 2            0            0            0            0            0            0
## 3            0            0            0            0            0            0
## 4            0            0            0            0            0            0
## 5            1            0            0            0            0            0
## 6            2            0            0            1            0            1
##   count.201406 count.201407 count.201408 count.201409 count.201410 count.201411
## 1            0            1            0            0            0            1
## 2            0            0            0            0            0            0
## 3            0            0            0            0            0            0
## 4            0            0            0            3            0            0
## 5            0            0            0            0            0            0
## 6            0            1            0            0            0            0
##   count.201412 count.201501 count.201502 count.201503 count.201504 count.201505
## 1            0            0            0            0            0            0
## 2            0            0            0            0            0            0
## 3            0            0            0            1            0            0
## 4            1            0            0            0            0            0
## 5            0            0            0            0            0            0
## 6            1            0            0            0            0            1
##   count.201506 count.201507 count.201508 count.201509 count.201510 count.201511
## 1            1            1            0            0            0            0
```

```
## 2            0            0            0            0            0            0
## 3            0            3            0            1            1            1
## 4            0            0            0            0            0            0
## 5            0            0            0            0            0            0
## 6            1            0            0            0            0            0
##   count.201512 young_males      wealth unemployment Population
## 1            0          36 -0.04277439   0.21259843       1141
## 2            0           0 -1.32527784   0.07875895        850
## 3            2          33 -0.50217861   0.10022779        719
## 4            0           0 -0.86587362   0.11206897        458
## 5            0          48 -0.32990203   0.09822485        991
## 6            1          21  1.04831063   0.09104404       2239
```

Data Processing

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

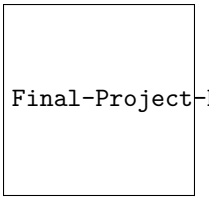
```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
long <- melt(all, id.vars=c("Column1", "wealth", "unemployment", "Population", "young_males"))
colnames(long)[colnames(long) == 'value'] <- 'Burglaries'
colnames(long)[colnames(long) == 'Column1'] <- 'Block'

# Add Month and Year and date Variables
long$year <- substr(long$variable, 7, 10)
long$month <- substr(long$variable, 11, 13)
long$date <- substr(long$variable, 7, 13)


# Rename other census block variables
long <- long %>%
  rename(pop = Population) %>%
  rename(unemp = unemployment) %>%
  rename(ym = young_males) %>%
  rename(target = Burglaries) %>%
  rename(block = Block)


# Add percent Young males variable
long$perc_ym <- long$ym / long$pop

# Take away uneeded col
long <- subset(long, select = -ym)
long <- subset(long, select = -date)
# long <- subset(long, select = -variable)

# make variables right type
long$year <- as.integer(long$year) - 2010
long$month <- as.factor(long$month)
```

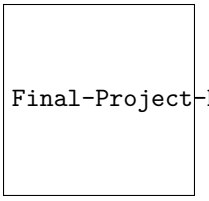
```r
# scale variables
long[, c("pop", "unemp", "perc_ym")] <- scale(long[, c("pop", "unemp", "perc_ym")])

head(long)
```

```
##   block      wealth      unemp        pop     variable target year month
## 1     1 -0.04277439  1.2098838  0.2142364 count.201001      0    0    01
## 2     2 -1.32527784 -0.6059003 -0.4280530 count.201001      0    0    01
## 3     3 -0.50217861 -0.3146352 -0.7171935 count.201001      0    0    01
## 4     4 -0.86587362 -0.1539873 -1.2932675 count.201001      0    0    01
## 5     5 -0.32990203 -0.3418088 -0.1168406 count.201001      0    0    01
## 6     6  1.04831063 -0.4392300  2.6377199 count.201001      0    0    01
##       perc_ym
## 1 -0.48227912
## 2 -1.31164595
## 3 -0.10518060
## 4 -1.31164595
## 5 -0.03844373
## 6 -1.06510163
```

```r
crime2 <- apply(crime, 2, sum)

# Sum over census blocks (see variation over month)
crime3 <- data.frame(val=crime2)
crime3 <- slice(crime3, -1)
crime3$name <- row.names(crime3)

# Add an index column
crime3 <- crime3 %>% mutate(index = row_number())

# Add Month and Year and date Variables
crime3$year <- as.integer(substr(crime3$name, 7, 10))
crime3$month <- as.integer(substr(crime3$name, 11, 13))

graph1 <- ggplot(crime3, aes(x = index, y = val)) +
  geom_point() +
  labs(title = "Total Burglaries in Chicago Over 72 Months",
       x = "Month (1-72)",
       y = "Total Burglaries in Chicago")

graph2 <- ggplot(crime3, aes(x = month, y = val)) +
  geom_point() +
  labs(title = "Total Burglaries in Chicago Per Year",
       x = "Month (1-12)",
       y = "Total Burglaries in Chicago")

graph1
```

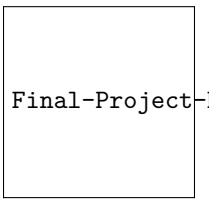


```
graph2
```



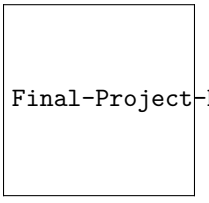

```
# Find columns starting with "count."
count_cols <- names(all)[startsWith(names(all), "count.")]

# Sum the values across selected columns for each row
all$sum_count <- rowSums(all[count_cols])

all2 <- all[c("Column1", "sum_count", "young_males", "wealth", "unemployment", "Population")]

plot1 <- all2 %>% ggplot(aes(x=young_males, y=sum_count)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

plot2 <- all2 %>% ggplot(aes(x=wealth, y=sum_count)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

plot3 <- all2 %>% ggplot(aes(x=unemployment, y=sum_count)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

plot4 <- all2 %>% ggplot(aes(x=Population, y=sum_count)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

plot1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



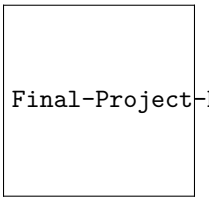

```
plot2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



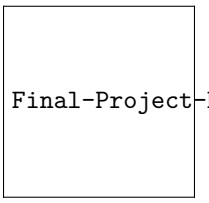
```
plot3
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



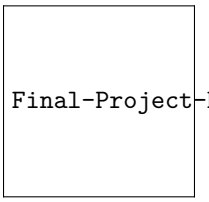

```
plot4
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



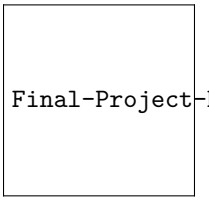

```
plot1 <- all2 %>% ggplot(aes(x=Population, y=wealth)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

plot2 <- long %>% ggplot(aes(x=pop, y=wealth)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

plot1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
plot2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Models

Poisson GLM

```r
# With and Without Wealth
pois_glm1 <- glm(target~perc_ym+pop+unemp+year+month+wealth, data=long, family=poisson)
pois_glm2 <- glm(target~perc_ym+pop+unemp+year+month, data=long, family=poisson)

# With and Without Population
pois_glm3 <- glm(target~perc_ym+pop+unemp+year+month+wealth, data=long, family=poisson)
pois_glm4 <- glm(target~perc_ym+unemp+year+month+wealth, data=long, family=poisson)

anova(pois_glm1, pois_glm2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ perc_ym + pop + unemp + year + month + wealth
## Model 2: target ~ perc_ym + pop + unemp + year + month
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     39727      61063
## 2     39728      61066 -1  -3.4459  0.06341 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(pois_glm3, pois_glm4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ perc_ym + pop + unemp + year + month + wealth
## Model 2: target ~ perc_ym + unemp + year + month + wealth
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     39727      61063
## 2     39728      61573 -1  -510.19 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: In the poisson model, we test weather the variable wealth, or population is significant, given that they are both highly correlated and we wanted to mitigate the potential for multi-coliniearity within our data. We found that while population is significant, wealth is not.

```r
# this is for best pois model with population, and without wealth
# Goodness of fit test for poisson glm 1 (our model compared to saturated)
1-pchisq(deviance(pois_glm2), df.residual(pois_glm2))
```

```
## [1] 0
```

Note: We accept the null hypothesis that the model is not a good fit for our data. Does not pass goodness of fit test.

```r
# Estimate phi
deviance(pois_glm2) / df.residual(pois_glm2)
```

```
## [1] 1.537108
```

```r
sum(residuals(pois_glm2, type="pearson")^2)/pois_glm2$df.res
```

```
## [1] 1.576132
```

Note: We test for over dispersion. The residuals are slightly over dispersed but not significantly so. Therefore we rule out the need for a negative binomial for count data. We then move on to a Zip GLM because of the apparent intuition for having a significant number of zeros in the response.

```r
# Take model and look at lambda (mean) values and simulate data with model
# (generate data from fitted model and see if it looks like actual data)

lambda.fitted <- predict(pois_glm2, type="response")
sim_data <- rpois(nrow(long), lambda.fitted)

sum(sim_data==0)
```

```
## [1] 13292
```

```r
sum(long$target==0)
```

```
## [1] 16044
```

```r
summary(pois_glm2)
```

```
##
## Call:
## glm(formula = target ~ perc_ym + pop + unemp + year + month,
##     family = poisson, data = long)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.406590   0.017813  22.826  < 2e-16 ***
## perc_ym      0.054195   0.004644  11.670  < 2e-16 ***
## pop          0.197534   0.004117  47.981  < 2e-16 ***
## unemp        0.029073   0.004677   6.216 5.11e-10 ***
## year        -0.165125   0.002745 -60.164  < 2e-16 ***
## month02     -0.362852   0.026377 -13.756  < 2e-16 ***
## month03     -0.057578   0.024245  -2.375   0.0176 *
## month04      0.035613   0.023684   1.504   0.1327
## month05      0.166834   0.022958   7.267 3.67e-13 ***
## month06      0.194710   0.022813   8.535  < 2e-16 ***
## month07      0.231388   0.022626  10.226  < 2e-16 ***
## month08      0.322882   0.022185  14.554  < 2e-16 ***
## month09      0.238160   0.022593  10.542  < 2e-16 ***
## month10      0.268734   0.022442  11.974  < 2e-16 ***
## month11      0.208709   0.022741   9.178  < 2e-16 ***
## month12      0.124390   0.023185   5.365 8.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 68412  on 39743  degrees of freedom
## Residual deviance: 61066  on 39728  degrees of freedom
## AIC: 119604
##
## Number of Fisher Scoring iterations: 5
```

Zip GLM ???

```r
# complex model
zip_glm1 <- zeroinfl(target ~ perc_ym + pop + unemp + year + month | perc_ym + pop + unemp + year + mon
                     data=long,
```

```
                    dist="poisson")

# simpler model
zip_glm2 <- zeroinfl(target ~ perc_ym + pop + unemp + year + month | pop + unemp + year + month,
                    data = long,
                    dist = "poisson")

logLik(zip_glm1) # 32 df
```

```
## 'log Lik.' -58628.75 (df=32)
```

```
logLik(zip_glm2) # 20 df
```

```
## 'log Lik.' -58632.22 (df=31)
```

```
# use difference of 12 for df to compare nested models
# 2 * (likelihood of complex model - likelihood of simple model)

1-pchisq(2 * (logLik(zip_glm1) - logLik(zip_glm2)), df = 1)
```

```
## 'log Lik.' 0.00845963 (df=32)
```

```
# complex model
zip_glm1 <- zeroinfl(target ~ perc_ym + pop + unemp + year + month | perc_ym + pop + unemp + year + mon
                    data=long,
                    dist="poisson")

# simpler model
zip_glm2 <- zeroinfl(target ~ perc_ym + pop + unemp + year + month | perc_ym + pop + unemp + year,
                    data = long,
                    dist = "poisson")

logLik(zip_glm1) # 32 df
```

```
## 'log Lik.' -58628.75 (df=32)
```

```
logLik(zip_glm2) # 20 df
```

```
## 'log Lik.' -58641.63 (df=21)
```

```
# use difference of 12 for df to compare nested models
# 2 * (likelihood of complex model - likelihood of simple model)

1-pchisq(2 * (logLik(zip_glm1) - logLik(zip_glm2)), df = 11)
```

```
## 'log Lik.' 0.007032543 (df=32)
```

```
summary(zip_glm1)
```

```
##
## Call:
## zeroinfl(formula = target ~ perc_ym + pop + unemp + year + month | perc_ym +
##     pop + unemp + year + month, data = long, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.8054 -0.8223 -0.2651  0.4839 15.6108
##
```

```
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.581993   0.023884  24.368  < 2e-16 ***
## perc_ym      0.037473   0.006425   5.833 5.45e-09 ***
## pop          0.156604   0.005646  27.737  < 2e-16 ***
## unemp        0.044247   0.006458   6.852 7.29e-12 ***
## year        -0.132422   0.003716 -35.636  < 2e-16 ***
## month02     -0.308522   0.037821  -8.157 3.43e-16 ***
## month03     -0.064270   0.032935  -1.951  0.05101 .
## month04      0.030160   0.031609   0.954  0.34001
## month05      0.158685   0.030086   5.274 1.33e-07 ***
## month06      0.163560   0.030047   5.443 5.22e-08 ***
## month07      0.189650   0.029727   6.380 1.77e-10 ***
## month08      0.282419   0.028773   9.816  < 2e-16 ***
## month09      0.244268   0.029416   8.304  < 2e-16 ***
## month10      0.226017   0.029273   7.721 1.15e-14 ***
## month11      0.195333   0.029742   6.568 5.11e-11 ***
## month12      0.091581   0.030900   2.964  0.00304 **
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.68841    0.09281 -18.192   <2e-16 ***
## perc_ym     -0.06562    0.02551  -2.573   0.0101 *
## pop         -0.24800    0.02750  -9.019   <2e-16 ***
## unemp        0.07491    0.02345   3.195   0.0014 **
## year         0.16476    0.01400  11.769   <2e-16 ***
## month02      0.25247    0.13148   1.920   0.0548 .
## month03     -0.02310    0.12583  -0.184   0.8544
## month04     -0.01356    0.11977  -0.113   0.9098
## month05     -0.01844    0.11383  -0.162   0.8713
## month06     -0.16631    0.11826  -1.406   0.1596
## month07     -0.23122    0.11901  -1.943   0.0520 .
## month08     -0.19074    0.11360  -1.679   0.0931 .
## month09      0.03155    0.10954   0.288   0.7733
## month10     -0.22044    0.11662  -1.890   0.0587 .
## month11     -0.05145    0.11342  -0.454   0.6501
## month12     -0.18276    0.12296  -1.486   0.1372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 42
## Log-likelihood: -5.863e+04 on 32 Df
```

```
AIC(pois_glm2)
```

```
## [1] 119603.6
```

```
AIC(zip_glm1)
```

```
## [1] 117321.5
```

```
# 119533.5
```

```
lambda.fitted <- predict(zip_glm1, type="response")
sim_data <- rpois(nrow(long), lambda.fitted)
```

```r
sum(sim_data==0)
```

```
## [1] 13183
```

```r
sum(long$target==0)
```

```
## [1] 16044
```

Generalized Mixed Effects Models

Linear Mixed Effects Model

```r
start_time <- Sys.time()
pois_glmer1 <- glmer(target~perc_ym+pop+unemp+(1|variable), nAGQ=1, family=poisson, data=long)
summary(pois_glmer1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: target ~ perc_ym + pop + unemp + (1 | variable)
##    Data: long
##
##      AIC      BIC   logLik deviance df.resid
## 119533.5 119576.4 -59761.7 119523.5    39739
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9356 -0.9317 -0.2958  0.5480 16.3489
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  variable (Intercept) 0.1215   0.3485
## Number of obs: 39744, groups:  variable, 72
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.103792   0.041367   2.509   0.0121 *
## perc_ym     0.054195   0.004642  11.674  < 2e-16 ***
## pop         0.197534   0.004115  47.998  < 2e-16 ***
## unemp       0.029073   0.004676   6.218 5.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) prc_ym pop
## perc_ym -0.006
## pop     -0.022  0.015
## unemp   -0.003 -0.108  0.032
```

```r
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 11.52174 secs
```

```r
start_time <- Sys.time()
pois_glmer2 <- glmer(target~perc_ym+pop+unemp+(1|block), nAGQ=1, family=poisson, data=long)
```

```
summary(pois_glmer2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: target ~ perc_ym + pop + unemp + (1 | block)
##    Data: long
##
##      AIC      BIC   logLik deviance df.resid
## 116567.0 116610.0 -58278.5 116557.0    39739
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9946 -0.8831 -0.3031  0.5585  9.7480
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  block  (Intercept) 0.2436   0.4935
## Number of obs: 39744, groups:  block, 552
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05037    0.02166   2.325  0.02005 *
## perc_ym      0.06040    0.02180   2.771  0.00559 **
## pop          0.20952    0.02164   9.683  < 2e-16 ***
## unemp        0.01501    0.02188   0.686  0.49290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) prc_ym pop
## perc_ym -0.004
## pop     -0.011 -0.017
## unemp    0.000 -0.108  0.022
```

```
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 14.34216 secs
```

```
start_time <- Sys.time()
pois_glmer3 <- glmer(target~perc_ym+pop+unemp+(1|block)+(1|variable), nAGQ=1, family=poisson, data=long
summary(pois_glmer3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: target ~ perc_ym + pop + unemp + (1 | block) + (1 | variable)
##    Data: long
##
##      AIC      BIC   logLik deviance df.resid
## 111489.2 111540.8 -55738.6 111477.2    39738
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.2296 -0.8261 -0.3319  0.5649 11.7679
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  block    (Intercept) 0.2440   0.494
##  variable (Intercept) 0.1218   0.349
## Number of obs: 39744, groups:  block, 552; variable, 72
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.008316   0.046524  -0.179  0.85813
## perc_ym      0.060408   0.021820   2.769  0.00563 **
## pop          0.209524   0.021659   9.674  < 2e-16 ***
## unemp        0.015006   0.021904   0.685  0.49331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) prc_ym pop
## perc_ym -0.002
## pop     -0.005 -0.017
## unemp    0.000 -0.108  0.022
```

```
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 36.73216 secs
```

```
lambda.fitted <- predict(pois_glmer1, type="response")
sim_data <- rpois(nrow(long), lambda.fitted)

sum(sim_data==0)
```

```
## [1] 13310
```

```
sum(long$target==0)
```

```
## [1] 16044
```

```
lambda.fitted <- predict(pois_glmer2, type="response")
sim_data <- rpois(nrow(long), lambda.fitted)

sum(sim_data==0)
```

```
## [1] 13906
```

```
sum(long$target==0)
```

```
## [1] 16044
```

```
lambda.fitted <- predict(pois_glmer3, type="response")
sim_data <- rpois(nrow(long), lambda.fitted)

sum(sim_data==0)
```

```
## [1] 14766
```

```
sum(long$target==0)
```

```
## [1] 16044
```