Final Project Write-Up

CDT Karly Parcell

May 7, 2024

Abstract

Using generalized linear models under a Poisson distribution to estimate count data of burglaries in Chicago, with variables population, wealth, precipitation as well as random effects. Building two separate Poisson mixed effect models and one Poisson random effects model, we compare the performance of the models by using AIC, a measure to determine how well a model fits the data. The model with the lowest AIC was the Poisson random effects model, meaning this model best fit the data of the three models we had to compare. Throughout the project we go through the process of cleaning data, determining significant factors to use, fitting models and comparing unnested models to choose the one that best fits your data.

1 Key Words

Chicago, Burglary, Poisson, Precipitation, Socio-economic Factors

2 Introduction

The motivation behind this project was to build a statistical model to capture the impact of various socio-economic factors and weather on burglaries in Chicago. Many researchers have done work on predicting burglary counts or crime rates in various settings. In this project, we assess the factors, precipitation, wealth, population, and effect of block and month, to predict the incidence rate ratio for predicting burglaries in Chicago. What stands out for this project is the use of precipitation as a predictor, as most other sources do not acknowledge this effect, and, if a weather factor is mentioned, it is typically always temperature, not precipitation. This is a factor that has not been assessed with this type of data often.

The goal of these estimates is to be able and predict where and when burglaries may occur in Chicago, so that way the forces in charge can take preventative action, rather than just addressing burglaries after they occur. Being able to look across different census block groups, which are also a variety of wealth's and populations, experiencing different levels of precipitation over the year, we hope to help give law enforcement helpful information on where may be best to reinforce based on all of the factors previously mentioned.

3 Literature Review

3.1 Previous Research

The ST-AR model for predicting medium and long term crime rate estimation [Sho13]. For the study I am assessing, ST-AR modeling was used at the State and National level, where it outperformed other models [Sho13]. These researchers use data from the past 50 years, and categorize the crime based on type of crime committed, with violent crime being separate from property crime [Sho13]. ST-AR uses an estimation parameter of ϕ_{ij} with i representing region and j representing lag [Sho13]. ST-AR also uses a parameter to make the coefficients for cross-regional impacts the same across all regions [Sho13]. A limitation of the ST-AR model is that it is better at predicting property crime than violent crime and this model works better on a larger scale, the article mentions use for a more local population, but that State and National level are where it performs the best [Sho13].

Another approach taken to look at burglaries in Wuhan City, China was a Spatio-temporal Bayesian model [HZDG18]. Using historical data and various socio-economic variables, this researcher noted that while spatial analysis is common, the connection between spatial and temporal is not as commonly used, especially when looking at their particular problem set [HZDG18]. This model uses spatial and temporal random effects, this is a common way to look at the spread of diseases, so they change it for use in a analysis of crimes [HZDG18]. For the spatial grouping, they look at villages split into sub-districts. In depth information was also taken on those who committed the crimes. The general components of this model are area, time and area*time [HZDG18]. This researcher used the past 7 months of data to predict it's placement or burglary rate for the 8th month. This study addresses limitations with generalizing and interpreting results, as having more bars or WiFi can be interpreted a variety of ways, where one direction is more students may come to study and are not so likely to increase crime [HZDG18]. Or the other direction, where those who cannot afford WiFi, go to where bars are in hope of free WiFi, and if you already have very little to nothing, you are likely to steal for more [HZDG18]. Another limitation of this study is that time is recorded when the individual reports the crime, not when the burglary truly occurred, because the burglary is not noticed, often until the resident arrives back to their residence, which effects the time that the model is taught to believe burglaries occur [HZDG18].

The last state-of-the-art model I looked at was using just spatial influence and analysis. An interesting point about this research is how they address the anchor points of convicted offenders [BB11]. Noting that there are areas that are more easily accessible and those areas are considered more at risk than others, due to their spatial layout [BB11]. This would be a factor difficult to account for without using a spatial model [BB11]. Another example given by the researchers of these offender anchors is a subway stop, where the population is greater and more diverse, that the proximity to a location with a subway stop is more probable to have crime occur [BB11]. This method uses many variables labeled as crime attractions, additionally the population data was taken from a block census [BB11]. Something interesting that this researcher walks us through that others have not, is using a spatial and negative binomial model together [BB11]. The combination of these models was used in deciding which offender anchors to account for which facilities were correlated with different than average burglary counts/reports [BB11].

3.2 My Approach vs State-Of-The-Art

Comparing our model first to the ST-AR model, the ST-AR model takes into consideration the different levels of crime committed, as well as the events going on around the region it is looking to predict. A difference in favor of our models is that for estimations, we are able to look at the effects on a much lower scale, which the ST-AR model is not made to perform well on. This is important to note because while the ST-AR model does address considerations that increase it's performance at the larger scale in comparison to our model, knowing the trends of a region, does not help to spread out resources and take preventative measures to stop the increases that are seen in crime.

The variables used for the Spatio-temporal Bayesian model are much more specific than the variables for our model. They look at levels of internet (bars), hotels, residential zones, all variables that were not addressed in our data. This researcher is not separating the events occurring by region or sub-group, but still looking at individual records of those who committed a crime. This is different from our analysis, in that we are more interested in the areas that crime is being committed, then looking at the individuals committing the crimes. However a similarity between our model and this model is that ultimately, we are looking at assessing how the trends seen can assist in directing preventative action taken by placement of police and other law enforcement. This researcher uses crime rate which means crime that has occurred in a given area is expressed as a rate per population, this is something my models do not consider, while it does have both population and count of crime. Both our model and this state of the art model use a Poisson distribution for the family of the model. We also do not have household population available, so we use population by block, which does not account for either the spatial model or our model in how household population affects burglaries.

The third approach, a spatial approach, along with the use of negative binomial models for determining the anchor hot spots, differs from the approach we take, in that it accounts for spatial variables. Additionally this model accounts for a larger variety and multitude of variables than we account for in our model.

Something that our model accounts for that has not been included in any of the models above, is weather, while others do account for date, which presumably has a relationship to temperature, they do not directly look at the relationship between temperature and burglary or crime counts/rates. Something that most of the researchers account for that we do not account for in our model is looking at a combined variable for population and parameter.

4 Methodology

We started out with a multitude of socio-economic factors, such as burglary count (separated by census block, year and month), population, wealth, youth_males, unemployment and were told we could add in other other factors, such as factors to account for weather. We choose to look at precipitation as our weather factor, adding it to the list of explanatory variables for determining burglary count which was our response variable. We are tasked with building a generalized linear model to account for some to all of the given explanatory variables listed above. All of the data given is real-world data, not given for easy use for building models as each factor was given as a different csv file and the count of burglaries and precipitation data found, were not given in a format to fit any of the other csv's. We made the majority of changes to our data set in excel, prior to bringing it over in R. And then we loaded the data as one large csv containing all of the data to R to be used in building our models and visuals of the data.

The models we will use for our analysis are two separate Poisson mixed effects models and a Poisson random effects model. We chose to use a Poisson distribution for our data, because we are dealing with counts of the data, as our response variable is burglary count. When deciding on the models above specifically, we decided to use them mostly because we were the most comfortable with these models after using them in class. They gave us the opportunity to explore models that we were semi-comfortable with, comfortable enough to understand the basics of what we were doing, but still allowed us to get out of our comfort zone to explore new effects in the models and solidify our understanding for what we were creating and the results we were producing.

The first model I will discuss is the first Poisson mixed effects model. We chose to include wealth, precipitation and an offset for population, as factors for this model in predicting burglary count. For this model we also included an effect for block and month. We presented our model like this:

$$MODEL\ 1:$$

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_{1i} + log(p_i) + \gamma_i + \beta_2 x_{2j}$$

$$i = block, j = month$$

$$x_{1i} = wealth, x_{2j} = weather, p_i = population$$

$$Y_{ij}\ Po(\lambda_{ij}), Y_i\ N(0, \sigma_\gamma^2)$$

The second model that we present is a Poisson random effects model, this model allows us to look at the unique aspects of each block during each month, where we again asses the effect of wealth and for an offset of population on burglary count. This model does not include a factor for precipitation, this was purposeful, so that we can asses the models performance without precipitation.

$$MODEL\ 2:$$

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_{1i} + \upsilon_j + \epsilon_{ij} + \gamma_i + log(p_i)$$

$$i = block, j = month$$

$$x_{1i} = wealth\ per\ block, p_i = population\ per\ block, \upsilon_j = unique\ aspects\ of\ each\ month$$

$$\epsilon_{ij} = unique\ aspects\ of\ each\ block\ i\ during\ month\ j, \gamma_i = unique\ aspects\ of\ block\ i$$

$$Y_{ij}\ Po(\lambda_{ij})$$

The final model that we have made to show the effect of wealth and population pulls similarities from both our first and second model. Structurally, this model is built similar to our first model, but a couple of key differences are that this model does not contain a factor for precipitation and this model assumes that the impact of wealth is different each block. The factor for precipitation was originally present in the first Poisson mixed effects model, and the assumption in creating our model that wealth impacts different blocks differently is a new addition to this model. This model also does not factor in month, just block, wealth and population, in determining burglary count.

$$MODEL\ 3:$$

$$n_i = \beta_0 + \upsilon_{0i} + (\beta_1 + \upsilon_{1i}x_i) + log(p_i)$$

$$i = block, j = month$$

$$x_i = wealth, p_i = population$$

$$Y_i\ Po(X_i), \log(\lambda_i) = n_i$$

$$\upsilon_{0i}\ N(0, \sigma^2_{\upsilon_{0i}}), \upsilon_{1i}\ N(0, \sigma^2_{\upsilon_{1i}})$$

discuss the key aspects of your problem, data set and generalized linear model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for generalized linear modeling, your process for building appropriate regression models, and your model selection.

5 Experimentation and Results

describe the specifics of what you did (data exploration, data preparation, model building, model selection, model evaluation, etc.), and what you found out (statistical inference and/or predictive modeling, interpretation and discussion of the results, etc.). Be sure to have good data visualizations throughout

Prior to running our models, we believed that an increase in precipitation would be correlated to a decrease in burglaries. We assumed this because we thought more people would be home, leaving less opportunity for burglaries, and then with the weather un-optimal, it would make more sense to not conduct as many burglaries as they would probably be harder to cover up. Upon further exploration of our data and looking at how these variables interact, we could see our initial predictions were incorrect. While there does seem to be an increase in precipitation around the spike in burglary count when looking at it over the years, the spike does not overlap with the highest days per month when it rains.

From figure 1 shown below, we can also see that the count of burglaries has been following a similar pattern each year, but over the years, the burglary count has decreased each year. While we do not account for the decreasing trend by year, we do include the month in our models to account for the cycle you can see occurring below, where the beginning of the year begins at a low burglary count, increasing until getting to the middle part of the year and then steadily decreasing again to then start the next year.

	Estimate	Standard Error	P-value
Intercept	-5.196	0.032	< 2e-16
Wealth	-0.226	0.023	< 2e-16
Precipitation	0.025	0.003	2.26e-15

Table 1: Coefficients for Poisson Mixed Effects Model 1.

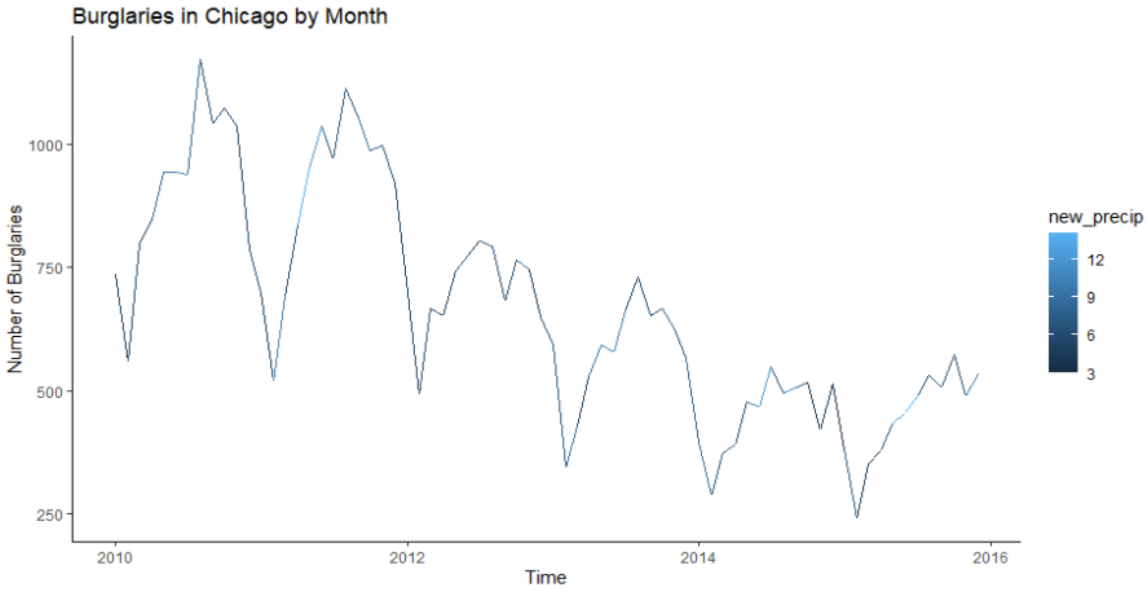

Figure 1: Data Exploration: Burglary and Precipitation

	Variance	Standard Deviation
γ_i	0.273	0.522

Table 2: Random Effects for Poisson Mixed Effects Model 1.

All of the coefficients have a very small p-value, this is indicating that these coefficients are significant for the model. The intercept being negative, for all of our models, is saying that when all other variables are 0, there are less than 0 burglaries, this is impossible, however, it is also not possible that we would get a scenario where all other coefficients are 0. Wealth is also a negative value for all models, which means that as wealth increases, the number of burglaries decreases. This is not what you would predict would happen right away, but after thinking through the situation, likely those with more money can set in place more measures to keep burglaries from occurring, there is also the possibility that there is already a larger police presence in those areas which would serve as a confounding variable. The coefficient of precipitation is only present in the first model, and based on the coefficient estimate, as the precipitation increases, there is also an increase in burglaries. This is not what we and originally predicted, but it does follow the graph from figure 1 that was made during our data exploration.

For the effect γ represents the random effect associated with each block in the census. γ is assumed under a normal distribution, with mean 0 and variance σ^2

	Estimate	Standard Error	P-value
Intercept	-5.042	0.057	$<$ 2e-16
Wealth	-0.226	0.023	$<$ 2e-16

Table 3: Coefficients for Poisson Random Effects Model.

	Variance	Standard Deviation
υ_j	0.032	0.180
ϵ_{ij}	0.030	0.172
γ_i	0.271	0.520

Table 4: Random Effects for Poisson Random Effects Model.

The significance of each of the effects for model 2 are υ_j being the unique aspect of each month, so looking at each month individually, JAN - DEC. ϵ_{ij} is the unique aspect of each block i, during the month, j. This effect contains the change due to the interaction between block and month, and then γ_i. γ_i is the effect for the unique aspects of each block, i. The difference in the standard deviation of these effects, with two close to 0.1 and γ_i much larger standard deviation, indicates that this effect is more different from the other two effects than they are from each other.

	Estimate	Standard Error	P-value
Intercept	-5.029	0.022	< 2e-16
Wealth	-0.210	0.028	3.97e-14

Table 5: Coefficients for Poisson Mixed Effects Model 2.

	Variance	Standard Deviation
υ_{0i}	0.093	0.305
υ_{1i}	0.234	0.484

Table 6: Random Effects for Poisson Mixed Effects Model 2.

Model 3 assumes that the effect of each variable differs by block, the effects are also assuming that the variable wealth, has a different impact depending on the block. Not all blocks are impacted the same way.

Aside from looking at our coefficients and effects given from running our models and interpreting them to understand which factors effect the burglary count per block for each of our models, we also compared the models against one another using AIC. AIC is used to compare unnested models, meaning none of the models we are comparing is fully contained within another one of the models we are using AIC as the comparison value for. After running each of our model summary's to obtain the values you can see in the tables above, as well as the AIC values, we also plotted the residuals of the values given from our model. This was done to check to ensure our models were are a good fit for the data. You will notice that looking at the residual plots from model 1 and model 3, they are very similar, this is because the models were a similar structure and contain similar variables. None of the residuals that can be found within the appendix code are reason for concern, there are no obvious patterns or apparent outliers. From the table below (Table 7) we can see that the AIC with the lowest value is model 2, so this is our preferred model for predicting burglaries out of the 3 models built.

MODEL	AIC
1	34815.7
2	33465.6
3	34965.4

Table 7: AIC Values for all Models

6 Discussion and Conclusion

6.1 Ethical Considerations

A possible ethical issue that can come from acting off of the results given here, is that by reinforcing some areas, this law enforcement has to come from somewhere, and likely we are not hiring more people, but rather redistributing who is already in the department. Multiple issues can arise with this, the biggest concern I have is that by taking the resources away from an area deemed more secure, potentially that area was more secure because of the already present law enforcement. Therefore by redistributing those police elsewhere, you are saying that the people who were being protected are not worth the protection. Another possible conflict with this is that we could be unintentionally giving fewer resources to certain demographics. This could tarnish the view of the public on the police department making them less trusted among the civilians as well.

6.2 Findings, Limitations and Future Work

The findings from this project are that our second model is the one that best fits our data. This is based on comparing AIC values between our models and seeing that the second model has the lowest AIC value. Thus indicating the best fit of that model, using the factors we were presented with, to our data and predicting burglary count using only those coefficients.

The limitations for our models that were tested, is frist, none of the models was a spatial model, so geographical location was not a consideration in determining burglary count, for instance, we have no idea if being closer to a more unemployed block, my affect your likelihood of being the victim of burglary. Another limitation is that we do not take into account the downward trend over the years in total burglary count. A third limitation for these models is that we only looked at precipitation as a weather factor, using temperature with precipitation or with precipitation and snowfall, would give a better full picture of the weather. The last limitation I will mention is that these results are only generalized to the districts during the years the data was collected, we are not able to take this model and use it to determine burglaries in another city in the US.

An idea for future work is to use a combination of population and youth males, or population and wealth, as variables, rather than looking at them all separately. These factors are likely related to one another, but we have no way of seeing that other than our pairs plot to look at co-linearity. Another point for future work is to see about getting week report days and the time of the burglary. For the time, this would likely be the time of the report, not the time of the burglary, as individuals are infrequently home when they are victim of burglary.

References

[BB11] 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, 2011.

[HZDG18] Tao Hu, Xinyan Zhu, Lian Duan, and Wei Guo. Urban crime prediction based on spatio-temporal bayesian model. *PLOS ONE*, 13(10):1–18, 10 2018.

[Sho13] Gary L. Shoesmith. Space–time autoregressive models and forecasting national, regional and state crime rates. *International Journal of Forecasting*, 29(1):191–201, 2013.