Predicting Counties with Higher/Lower Risk of a School Shooting Incident

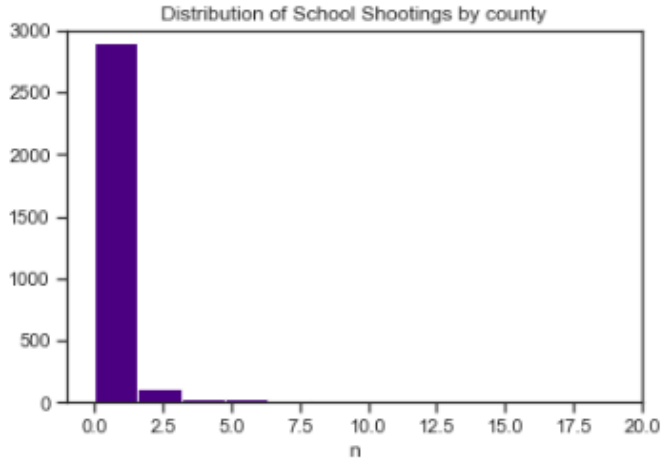
Alex Bass, Connie Cui, Tulsi Ratnam

Introduction

Recently, gun violence has taken center stage in national news coverage with Congress recently passing a bipartisan gun bill. This study focuses on deepening our understanding of the types of counties in which gun violence in schools is prevalent. Our report aims to answer questions like: What are the characteristics of a county that are more susceptible to gun violence? Are gun violence incidents concentrated in particular regions in America? If so, which? How do state laws play a role in deterring gun violence in schools? In the end, our motivation is to provide more context, information, and understanding on a major debate in American politics today using predictions from Bayesian regression modeling.

Methodology

Our analysis begins with a joined dataset created by combining four datasets: 1) a compilation of data obtained from different sources of over 2000 school shootings in school counties from 1970 to present (2020) from the Center for Homeland Defense and Security[1], 2) census data that provides demographics such as percent gender, race, level of schooling, etc[2], at the county level, 3) a county level crime dataset that provides percentages of gun owners, as well as if there are loose or strict gun laws within the state[3], and 4) a score of strictness of gun laws from fivethirtyeight[4]. Our final dataset contains 3141 rows of school counties, with 23 variables[5] containing county demographic variables, gun law variables, as well as our response variable, the total count of school shootings within each county from 1970 to 2020. We also standardized all of our non-categorical variables and one hot encoded these to ensure any categorical variables were binary. We had a few issues getting the urban, suburban, rural classifications to estimate proper distributions for some of our models, so we decided to drop these variables for the purposes of this study, and given we had a population variable that was highly correlated, we felt confident about removing these.



Our response variable has a Poisson distribution with high counts of 0, indicating that many of the counties did not experience school shootings. Therefore, for our analysis, we decided to explore six different Bayesian models to help determine the most prominent features among our data that contribute to predicting the number of school shootings in a county. Our six models include a pooled Poisson regression model with all 18 feature variables (excluding the response variable 'n' and 'STATE' variable), a pooled Poisson regression model with only feature variables that are deemed 'significant' out of the original 18, a pooled Zero Inflated Poisson regression model with all 18 feature variables, a pooled Zero Inflated Poisson regression model with feature variables that are deemed 'significant' out of the original 18, a hierarchical Poisson regression model with 19 feature variables (including the 'STATE' variable), and a hierarchical Zero Inflated Poisson regression model with all feature variables. We decided to use the Zero Inflated Poisson primarily because of the large number of counties that had no school shootings which is an imbalance in our dataset. We will detail the methodology for each of the six models below.

Pooled Poisson + Selected Pooled Poisson Models:

The following equations are foundational to all of our models described here:

$$\theta = exp(\beta X)$$

$$Y_{school\_shootings} = Poisson(\theta)$$

We are estimating the number of school shootings in each county by using a poisson distribution. We are estimating the lambda of our poisson distribution using an exponentially-linked regression equation (Theta). For the parameter estimates for our 18 feature variables and model intercept, we selected Normal distributions for all of the priors since Normal distributions are a good default for weakly informative priors on real number values. For all Normal distributions, we chose hyperparameters of mu = 0 and standard deviation = 5. We set the value of mu for these distributions to zero because we are unsure of the true value of our parameter estimates and whether they are of significant impact to our model overall. Because all of our non-categorical variables are all standardized, we set our standard deviation to 5 for these distributions because after standardization, we do not see our coefficients being greater than this range, especially since we will have to exponentialize our final regression equation to ensure that the count estimate we receive is always positive for our process.
We also consulted the PYMC3 documentation example of a poisson regression which used 10 for the sigma values.

The first model had all 18 variables. The second model had only a selected number of the original variables[6]. Factors that contributed to removing these variables were having a very wide credible interval, and having zero included in said interval.

Pooled Zero Inflated Poisson + Selected Pooled Zero Inflated Poisson Models:

We also explored two different pooled Zero Inflated Poisson regression models, one with all 18 feature variables (excluding the STATE variable as that one is only used for the hierarchical models) and one with a select group of feature variables from the original 18 that

we believe contribute to the model more, which will be determined after our initial analysis of the full model.

Zero Inflated Poisson regressions are often used to model the number of events occurring in a fixed period of time when the times at which events occur are independent and there is an excess amount of zeros within the counts variable. This fits our school shootings problem since we hope to create a model to predict the number of school shootings within the next 50 year period, but a majority of our school counties have not had a school shooting before or did not report their shootings, leading their counts to all be 0. What separates a Zero Inflated Poisson regression model from the regular Poisson regression model is the additional step prior to running a Poisson process, which is a different underlying process to first determine whether or not the count should simply be zero or non-zero. If the count is determined to be non-zero, we then move onto the regular Poisson process to predict the actual non-zero value based on our model. We are exploring this model in addition to our regular Poisson regression in case the excess amount of zeros affects the regular Poisson regression parameter estimates. The full equation for our two Zero Inflated Poisson models will look like:

$$\hat{y} = 0 \ when \ P(y_i = 0) = psi_i + (1 - psi_i) * e^{-\lambda_i}$$

$$\hat{y} = 1, 2, 3... \ when \ P(y_i = k) = (1 - psi_i) \frac{e^{-\lambda_i} * \lambda_i^{k}}{k!}$$

$$where \ k \ is \ observed \ school \ shootings, \ k > 0, \ and$$

$$\lambda_i = e^{x_i \beta} \ (x = intercept \ and \ feature \ variables, \ \beta = regression \ coefficients)$$

For the Zero Inflated Poisson regression model with all feature variables, we had to select priors and corresponding hyperparameters for our 18 features variables' parameter coefficients, our model intercept, as well as our 'psi' variable. This 'psi' variable represents the expected proportion of Poisson variates in our model compared to the excess zeros and will always range between 0 and 1. For our 'psi' variable, we selected a Beta distribution for the prior, with an alpha and beta value of 1. A Beta distribution by nature is a continuous probability distribution defined on the interval [0, 1], so we could ensure that psi would be between 0 and 1. We selected 1 for both values in order to simulate an 'uninformed prior' situation since we are not sure what the true proportion of excess zeros to Poisson variates is for school shootings in the United States. Having a Beta(1,1) distribution will keep our distribution for this uninformed 'psi' variable uniform. For the other priors in this model, we used the same priors and reasoning as in the previous models since they are nearly the same (excluding the 'psi' variable).

After analyzing the parameter estimates in our full Zero Inflated Poisson model, our second Zero Inflated Poisson regression model consists of our intercept, psi variable, and 13 of the original 18 features used in our full model[7]. The primary group of feature variables that were removed included county education demographics and were removed by similar methods as before. All priors for the intercept, the 13 feature variables[7], and the psi remained the same as the full model with similar justification.

Hierarchical Poisson + Hierarchical Zero Inflated Poisson:

Lastly, we layered a hierarchical model onto the poisson regression and zero-inflated poisson regression models. A hierarchical model, or partial pooling, allows us to cluster data around a group mean and variance when parameters in the data are assumed to be dependent on each other. In our study, we have state-level and county-level variables. By using this type of model, we are able to see how the same gun laws across all states impact the potential risk of gun violence in schools. Additionally, we are able to include the census data variables that vary across counties and are independent of each other.

In both hierarchical models, we used the full 18 variables plus the 'state' variable for clustering. For the group-level variables, we selected hyperpriors with a Normal(0, 10) distribution, and a Half Cauchy(0, 5) for the sigma values for both hierarchical models. Since we have quite a large number of groups (50 states), we chose weaker and more broad hyperprior parameters such that it will be informative but not overly influence the posterior inference. For the Hierarchical Zero Inflated Poisson, we kept the same psi value with a Beta(1, 1) distribution. The remaining non-hierarchical priors were kept the same with a Normal(1,1) distribution.

Results

Table 1 - All models with WAIC scores

	WAIC Values
Full Poisson Regression	-2640.131213
Selective Poisson Regression	-2816.587878
Full Zero Inflated Poisson	-2337.220527
Selective Zero Inflated Poisson	-2450.220935
Full Hierarchical Poisson Regression	-2428.352785
Full Hierarchical Zero Inflated Poisson Regression	-2225.573280

Once we computed the models, we performed model diagnostics with trace plots to verify convergence, and outputted a summary table that gave us r-hat values, means, standard deviations, and credible intervals to ensure that all our models were performing well. We then utilized the Widely-Applicable Information Criterion (WAIC)/ Leave-One-Out method, a fully Bayesian criterion for estimating out-of-sample expectation, to compare and determine the best performing model.

The results from WAIC/LOO indicated that the Full Hierarchical Zero Inflated Poisson Regression was the best model out of the six models we explored, with the highest WAIC/LOO value. This is not terribly surprising that this model fits our data format best because:

1. The nature of our dataset is hierarchical, given we have state and county level variables that are generally better suited for a hierarchical model.
2. Our dependent variable is unbalanced with a lot more zeros than other values, so a Zero-Inflated model seems to be more appropriate than a regular poisson.

Findings

Our findings from the Full Hierarchical Zero Inflated Poisson regression model provide more context around the rising concern of gun violence. Below is a table of selected coefficients that had a noticeable impact on the prevalence of school shootings.

Table 2 - Selected Coefficient Table for Hierarchical Zero Inflated Poisson regression[1]

	mean	sd	hdi_3%	hdi_97%	r_hat
background_checks	0.509	0.258	0.031	0.997	1
gun_permit_law	-0.662	0.322	-1.274	-0.062	1.01
White	0.229	0.136	-0.029	0.478	1
Black	0.663	0.122	0.443	0.902	1

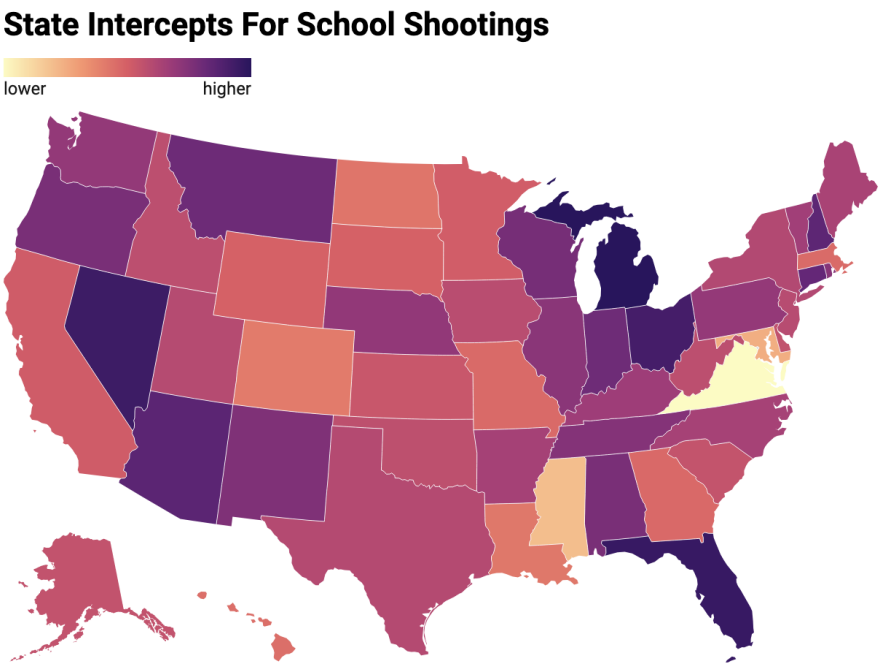
The columns cannot be directly interpreted as is because we first need to choose and specify values for each predictor variable, multiply them by their respective coefficients, and then exponentiate the sum. However, in this format, it is easier to distinguish the direction (positive or negative) and magnitude of each coefficient (a larger coefficient estimate has a larger impact on school shootings with one standard deviation increase of the predictor).

The most significant indicator of gun violence in schools was the percentage of Black people in a county. Though we are not exponentiating this coefficient to get a ratio, we can see that the credible interval for this variable is well within the positive region meaning that we are confident that as the percentage of Black people increases in a county, the number of school shooting incidents also increases. Additionally, since the coefficient for 'Black' is the highest, our model predicts that increasing one standard deviation in the proportion of a Black population in a county will increase the risk of school shootings more than a standard deviation increase in any other predictor. This contrasts with the 'White' predictor variable, as the credible interval for increasing the percentage of White people in a particular county includes zero which may suggest that there isn't a meaningful relationship between the percentage of White people in a county and the number of school shootings. It's important to note that the results of this study do not speak to causality and only assess the prevalence of gun violence in schools. This means that the model predicts that more school shootings occur in counties with a larger Black population, than in other counties where the racial makeup may be different.

Another substantive finding was the relationship between different gun laws and school shootings. In our model, we have two binary variables for two common types of gun laws: gun permit laws and universal background checks. For states that require gun permits, our model suggests these states are less likely to have school shootings – which is not too surprising. However, states that implement universal background checks actually seem to increase the number of school shootings. Both of these coefficient estimates have credible intervals that do not include zero that suggests gun permit laws are more effective at curbing gun violence in schools.

[1] For the sake of brevity, not all coefficients are included in this table (see full table in appendix[9])

Lastly, the map below illustrates where our model predicted that gun violence would be more or less prevalent, using the intercepts from different states.



While we don't specify the coefficients directly in the chart, the full range of un-exponentiated coefficients are within this interval [-2.895, -0.191] (see notebook on github for full list of coefficients). Controlling for all other variables, Florida, Michigan and Nevada are on average more likely to experience school shootings whereas Virginia, Maryland, Mississippi, and other states in the Midwest appear less likely.

Conclusion / Limitations

Given the timing limitations of this study, there were a few more things we wanted to explore in this paper:

1) We got a few errors with our WAIC score suggesting that they could be unreliable due to the estimated shape parameter of the Pareto distribution being greater than 0.7, so another avenue would be to explore other model evaluation techniques.

2) We also wanted to explore a few more models; specifically, selected hierarchical poisson and selected hierarchical zero inflated poisson that only focus on the variables that were significant from the full variable models.

3) There are other variables that we could have included in our dataset such as race of the shooter and victims, political party affiliation, safety measures already taken by schools, among others. Further studies could explore adding more such features to our models or exploring questions that we brought up here: Why do counties with a higher percentage of Black people more likely to experience school shootings? Besides background checks and gun permit laws, what other laws or preventative measures help deter gun violence in schools?

Appendix

Datasets (all data used in github repository):

1) https://github.com/acbass49/school-shootings/blob/main/data_w_county_match.csv
2) https://github.com/acbass49/school-shootings/blob/main/county_level_data_raw/co-est2021-alldata.csv
3) https://www.kaggle.com/datasets/mikejohnsonjr/united-states-crime-rates-by-county
4) https://fivethirtyeight.com/features/gun-laws-stop-at-state-lines-but-guns-dont/

Final Dataset Variables and Descriptions:

5) STATE = State FIPS
   gun_own = Estimated percentage of gun owners by state. Averaged from the GSS and PEW estimates.
   hunt_license = estimated percent of residents with hunting license
   background_checks = whether state has law for universal background checks
   gun_permit_law = whether state requires a gun permit before gun purchase
   under40 = Percent of county under 40 according to census data
   Male = Percent of county that is male according to census data
   White = Percent of county that is white according to census data
   Black = Percent of county that is African America according to census data
   Asian = Percent of county that is Asian according to census data
   Hispanic = Percent of county that is Hispanic according to census data
   Unemployment_rate_2021 = unemployment rate
   Median_Household_Income_2020 = median hh income
   ba_plus = percent of residents with college education
   less_than_hs = percent of residents with less than high school level education
   hs = percent of residents with high school level education
   some_col = percent of residents with some college education
   urban = 1 if the county is considered urban, 0 otherwise
   suburban = 1 if the county is considered suburban, 0 otherwise
   rural = 1 if the county is considered rural, 0 otherwise
   population = population
   n = number reported school shootings in county from 1970 to 2020
   gun_strictness = rating by 538 rating states across 22 different gun laws and giving a gun strictness score where low is less strict and high is more strict.

Model Methodology

6) The 14 variables used for the second Pooled Poisson regression model are Background Checks, Gun Permit Law, Under 40, Male, White, Black, Asian, Hispanic, Median Household Income 2020, Population, Gun Law Strictness, Gun Ownership, and Hunting License, Unemployment_rate_2021.
7) The 13 variables used for the second Zero Inflated Pooled Poisson regression model are Background Checks, Gun Permit Law, Under 40, Male, White, Black, Asian, Hispanic,

Median Household Income 2020, Population, Gun Law Strictness, Gun Ownership, and Hunting License.

8) Traceplots and coefficient plots for all models, and the WAIC comparison are all contained in this notebook: https://github.com/acbass49/school-shootings/blob/main/scripts/combined_poisson.ipynb
9) Our final Model full coefficient table:

	mean	sd	hdi_3%	hdi_97%	r_hat
gun_own	0.243	0.168	-0.067	0.563	1.01
background_checks	0.509	0.258	0.031	0.997	1
gun_permit_law	-0.662	0.322	-1.274	-0.062	1.01
gun_strictness	0.262	0.143	-0.013	0.522	1.01
hunt_license	-0.196	0.139	-0.46	0.058	1.01
under40	0.504	0.044	0.425	0.59	1
Male	-0.699	0.078	-0.852	-0.559	1
White	0.229	0.136	-0.029	0.478	1
Black	0.663	0.122	0.443	0.902	1
Asian	0.07	0.039	0.001	0.146	1
Hispanic	0.282	0.051	0.189	0.382	1
Unemployment_rate_2021	0.136	0.051	0.045	0.235	1
Median_Household_Income_2020	0.137	0.051	0.042	0.23	1
ba_plus	0.236	6.648	-11.712	13.213	1
less_than_hs	-0.23	4.142	-7.764	7.758	1
hs	-0.155	5.057	-9.2	9.727	1
some_col	0.06	3.664	-6.61	7.144	1
population	0.122	0.006	0.111	0.132	1

Sources:
All code and data can be found at this repository https://github.com/acbass49/school-shootings