Understanding Gun Violence and the Effectiveness of Laws via Fine-Grained Data Analysis

Xin Bian

Abstract

Gun violence in the United States is becoming a substantial public health concern due to recent high-profile cases that grabbed national attention. Over the past two years, many studies have been done on explaining, preventing, and predicting gun violence at the individual level, community level, and federal level. However, many studies did not consider some variables that might affect gun violence, including average income, crime rate, unemployment rate, education level, and race. Moreover, these studies were conducted at the state level, while different counties of a state can vary in all the aforementioned variables. In this study, we aim to analyze the problem using the firearm death data along with factors including unemployment rate, education level, crime rate, poverty rate, median household income and the strictness of law regulations at fine-grained county level. The multiple linear regression model shows the effectiveness of law on reducing firearm death rate is statistically significant but fairly limited. Moreover, we build three machine learning models to predict the firearm death rate with a mean percentage error about 4-5%.

Introduction

Gun violence in the United States is a substantial public health concern. Every day in the United States, more than 200 people are murdered or assaulted with a firearm. Number of mass shooting in the US is on the rise.[1] Over the past couple years, many studies have been working on explaining, preventing, and predicting gunshot violence at the individual level, community level, and federal level (Green, Horel, and Papachristos 2017; Kalesan et al. 2016; Fleegler et al. 2013; Lee et al. 2017).

A study compared 25 different state gun laws and concluded that implementing three state laws at federal level could reduce the rate of US gun deaths by more than 80% (Kalesan et al. 2016). It is questionable that implementing three relatively modest gun restrictions could have a huge impact on gun death and they didn't provide a convincing validation. Also, at state level, a legislative strength score was created. The association between firearm-related fatalities and the strength score were measured (Fleegler et al. 2013). However, as noted in this paper, the legislative strength score has not been validated, and some other factors such as nonfatal firearm injury, exploitation of loopholes are not considered. 34 articles investigating the relationship between firearm laws and firearm homicides from 1970 to 2016 are summarized (Lee et al. 2017). It is found that certain stronger gun policies are associated with reducing firearm homicide rates. Also, the paper suggested that legislation was not the only way to help reduce firearm tragedies.

To our best knowledge, the literature studying the effect of gun-control laws did not consider enough variables. There are other variables that might affect gun violence. For example, a region's average income, crime rate, unemployment rate, education level, race, etc., might be associated with firearm crimes. Many other factors may need to be considered. Moreover, most work studied impact of gun-control laws at a state-level. However, within a state, different regions can be very different in terms of crime rate, economics, education, etc. Thus different regions can have different gun rates even though they are in the same state and having the same strictness of law. For example, New York City is very different compared to cities in Upstate New York. New York City might have more immigrants, stronger economics, and higher crime rates. Conclusions drawn from the state level rather than city or country level may not be valid or persuasive enough to explain gun violence.

In this study, we are aiming to analyze the problem using fine-grained county-level data. We first try to find the relation, if any, between the strictness of law regulations and firearm death rate using multiple linear regression. Then three machine learning models are built to predict the firearm death rate at county level.

Dataset and Preprocessing

The county-level firearm mortality data is from Centers for Disease Control and Prevention[2]. The dataset has fire accidents recorded by government and local body. Data is based on death certificates for US residents. The data of

[1] https://thesocietypages.org/socimages/2015/12/31/mass-shootings-in-the-u-s-what-makes-so-many-american-men-dangerous/

[2] https://wonder.cdc.gov

mortality caused by firearm from 1999 to 2015 is used. In this study, gun violence is measured by firearm mortality which includes both suicide and homicide. The data of poverty rate, median household income, unemployment rate and education level is obtained from US Department of Agriculture Economic Research Service[3]. The data of poverty rate, median household income, and unemployment rate is from the year of 2015. Education level data is measured by using the percentage of adults completing college or associate's degree from 2011 to 2015. State gun-control laws information is obtained from the website of Brady Campaign to Prevent Gun Violence[4]. It provides data regarding gun control laws by state in the year of 2013 and a grading system. We measure the strictness of law regulations in a state by using the grading system.

For preprocessing, about 1500 tuples with unreliable data of firearm mortality are removed. Also, a linear regression analysis is applied to study the relation between crime rate and firearm rate, as shown in Figure 1. The correlation coefficient between them is 0.42 with p-value much less than 0.0001. The relatively strong correlation reminds us that the crime rate and firearm death rate have overlaps, to some extent. Thus, we drop the attribute of crime rate in the following study. Furthermore, the min-max normalization is used to rescale numerical data to [0, 1]. At last, the dataset has 1496 tuples (counties) and 6 attributes including median household income, unemployment rate, education level, poverty rate, curved score for laws, and firearm death rate (hereinafter referred to short term as Income, Unemployment, Education, Poverty, Law, and Firearm, respectively).

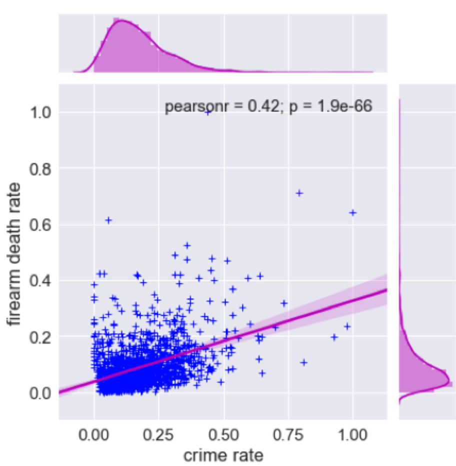

Figure 1: Linear regression between crime rate and firearm death rate

[3]https://www.ers.usda.gov/data-products/county-level-data-sets/

[4]https://www.bradycampaign.org/

Results

The Effect of Law Regulations

The probability distribution function of firearm death rate is shown in Figure 2. It can be noted that most counties have low firearm death rate(< 0.2). Only very few counties have firearm death rate greater than 0.4. Figure 3 shows the firearm death rate in a map. The black regions in the map are not used in our study due to the removal of unreliable data. The map suggests that many counties in southern United States have a lot of firearm accidents. The majority of other areas have a low firearm death rate, which is in the range of 0 to 0.2. The top 5 counties that have largest firearm death rate are Orleans Parish, LA, Baltimore city, MD, St. Louis city, MO, Coahoma County, MS, and Richmond city, VA.

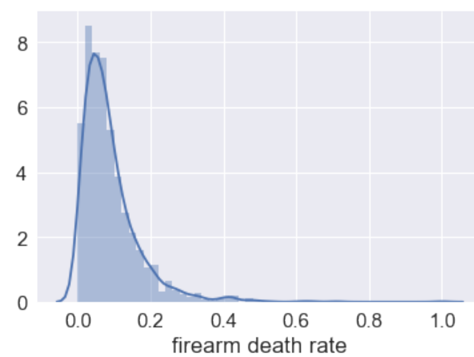

Figure 2: Probability distribution of firearm death rate

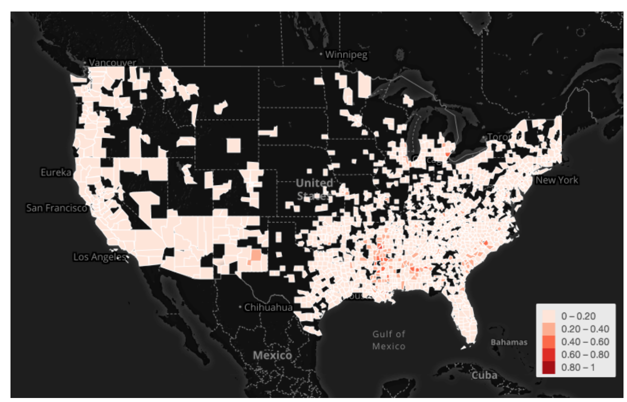

Figure 3: Map of normalized firearm death rate (only contiguous US is shown)

To study the effectiveness of gun-control laws, multiple linear regression is applied. The multiple linear regression results are summarized in Table 1. The attribute of education level has a large p-value, so we can't conclude anything about the effect of education level. For other attributes, the p-value is very small. Poverty rate has the largest regression coefficient, followed by unemployment rate, median household income, and law strictness. One

Table 1: Multiple linear regression summary

	Coefficient	t-statistic	p-value
Unemployment	0.148	5.454	< 0.0001
Income	0.0969	4.175	< 0.0001
Education	-1.629e-05	-0.001	0.999
Poverty	0.354	14.978	< 0.0001
Law	-0.0445	-5.631	< 0.0001

surprising observation is the positive coefficient of median household income. A possible explanation might be that higher household income reflects better economic conditions in the region, which attracts people from other places to seek their fortune. It leads to a higher transient population, which brings more safety issues. It's easy to understand the sign of correlation coefficient of unemployment rate, poverty rate, and law strictness. Generally, stricter laws should have positive effect, but the result shows that the effect is weak with a very small regression coefficient of -0.0445. This means that the firearm death rate can only be reduced by 0.89% when the score of gun-control laws is increased by 20%. In contrast, by reducing unemployment rate or poverty rate by 20%, the firearm death rate could be reduced by 2.97% or 7.08%, respectively. Obviously, the effect of poverty rate and unemployment rate is much stronger than gun-control laws.

Since we might incorrectly conclude the relationship between firearm death rate and law strictness by using the individual p-value (James et al. 2013), a hypothesis test is performed to check this.

H_0: the coefficient of law strictness, $\beta_{law} = 0$
H_a: the coefficient of law strictness, $\beta_{law} \neq 0$

The F-statistic is:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}, \tag{1}$$

where RSS is the sum of squared residuals for the original model, RSS_0 is the sum of squared residuals for a second model without the use of law attribute, $q = 1$, $n = 1496$, $p = 5$.

The result is $F = 31.713$, corresponding to a p-value < 0.0001. The null hypothesis H_0 is rejected. It can be concluded that the law strictness does affect the firearm death rate. However, the impact is limited according to the multiple linear regression coefficient.

Predicting Firearm Death Rate

We build three models to predict firearm death rate. The goal is to predict a county's firearm death rate by its attributes. We build and compare three machine models: SVM, random forest (RF) and neural network (NN) by using LinearSVR, RandomForestRegressor and MLPRgressor in scikit-learn (Buitinck et al. 2013). Five attributes: Income, Unemployment, Education, Poverty and Law are used to predict firearm death rate. For linear SVM, penalty parameter is selected as 1.0. Bootstrap samples are used in random forest regression model. Rectified linear unit is used as activation function in our neural network. The neural network has two hidden layers. Each hidden layer has 100 neurons.

Table 2: Regression model accuracy

	SVM	RF	NN
mean absolute error	0.0404	0.0467	0.0417
R^2 score	0.270	0.0329	0.265

We report our model accuracy by measuring mean absolute error and R^2 score using 10-fold cross validation, as shown in Table 2. The results of similar mean absolute error indicate that the three regression models have similar accuracy. However, if we look at R^2 score, we can find SVM and neural network are more accurate than random forest. In summary, SVM and neural network have better performance with mean absolute error around 0.4 and R^2 score around 0.27.

The regression models are tested on a test dataset. To visualize the result, the probability density function of percentage error is shown in Figure 4. For the three models, the percentage error is below 10% in most data points. Specifically, SVM and neural network have more data points falling in the region of percentage error less than 6%, compared to random forest distribution. The results of random forest show more data points in the region of percentage error from 12% to 20%. Accordingly, the random forest model results in a much smaller R^2 score, even though the mean absolute errors of the three models are nearly the same.

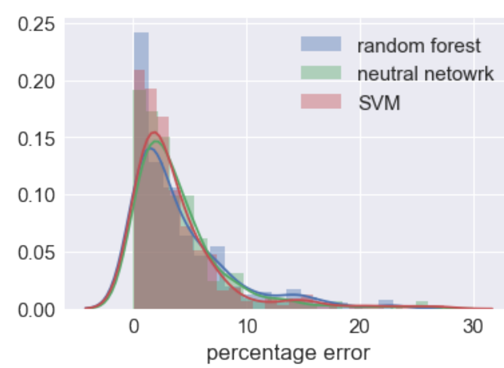

Figure 4: Probability density function of percentage error

We also visualize the percentage error in a map. For example, the result of neural network is shown in Figure 5. Darker red region corresponds to larger percentage error. The maps for random forest and SVM results (not shown here) are similar to Figure 5.

For neural network and SVM, the top five counties that have largest percentage error are shown in Table 3 and 4. The counties are the same for the two models and they suffer from relatively high firearm death rate. Since the firearm death rate distribution is positively skewed, the

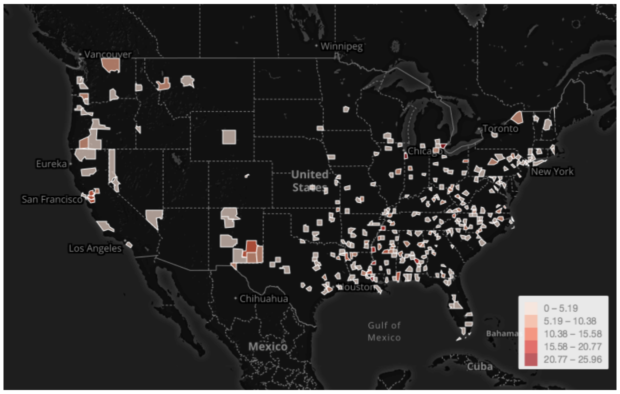

Figure 5: Percentage error of neural network model

Table 3: Counties with largest percentage error in NN

	Firearm death rate	Percentage error
Wayne County, MI	0.410	25.961%
Lake County, IN	0.350	25.260%
East Baton Rouge Parish, LA	0.334	23.438%
Essex County, NJ	0.293	21.198%
Shelby County, TN	0.320	19.968%

amount of data is not enough to train models at high firearm death rate level, leading to low performance in this region. Another possible reason might be that some mass shooting incidents lead to a spike in firearm death rate, while the regression models are incapable of predicting these accidents. In contrast, the results of random forest, as shown in 5, seem to be more complicated. In random forest model, both high firearm death rate counties, such as Lake County and Essex County, and counties with very low firearm death rate, such as Oktibbeha County, appear.

Conclusions

In this paper, we have analyzed the effects of different attributes of an area, including the unemployment rate, education level, median household income, poverty rate, gun control laws strictness, on gun violence (measured by firearm death rate in this paper) using fine-grained county-level data. Multiple linear regression results show the effect of gun control laws is weak, with a correlation coefficient of -0.0445. Median household income has the strongest impact on gun violence, followed by unemployment rate, poverty rate, laws regulations.

We also build three regression models, which are based on SVM, neural network, and random forest. The regression models predict firearm death rate at county level when given the specific attributes of a county. SVM and neural network model show relatively better performance, with a mean percentage error of 4.04% and 4.17%, respectively. The two models do not work well for counties with a high firearm death rate partially because of the insufficient data in a high firearm death rate region. In the future, we plan to investigate the same issues at an international level by including fine-grained data from other nations.

Table 4: Counties with largest percentage error in SVM

	Firearm death rate	Percentage error
Wayne County, MI	0.410	27.419%
Lake County, IN	0.350	26.221%
East Baton Rouge Parish, LA	0.334	24.324%
Essex County, NJ	0.293	23.302%
Shelby County, TN	0.320	21.216%

Table 5: Counties with largest percentage error in RF

	Firearm death rate	Percentage error
Oktibbeha County, MS	0.0846	22.943%
Essex County, NJ	0.293	22.815%
Kings County, NY	0.130	19.501%
Lake County, IN	0.350	17.606%
Shelby County, TN	0.320	17.602%

References

Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; and Varoquaux, G. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Fleegler, E. W.; Lee, L. K.; Monuteaux, M. C.; Hemenway, D.; and Mannix, R. 2013. Firearm legislation and firearm-related fatalities in the united states. *JAMA internal medicine* 173(9):732–740.

Green, B.; Horel, T.; and Papachristos, A. V. 2017. Modeling contagion through social networks to explain and predict gunshot violence in chicago, 2006 to 2014. *JAMA internal medicine* 177(3):326–333.

James, G.; Witten, D.; Hastie, T.; and Tibshirani, R. 2013. *An introduction to statistical learning*, volume 112. Springer.

Kalesan, B.; Mobily, M. E.; Keiser, O.; Fagan, J. A.; and Galea, S. 2016. Firearm legislation and firearm mortality in the usa: a cross-sectional, state-level study. *The Lancet* 387(10030):1847–1855.

Lee, L. K.; Fleegler, E. W.; Farrell, C.; Avakame, E.; Srinivasan, S.; Hemenway, D.; and Monuteaux, M. C. 2017. Firearm laws and firearm homicides: a systematic review. *JAMA internal medicine* 177(1):106–119.