Submitted to the Annals of Applied Statistics

IDENTIFYING THE DOMINANT PREDICTORS OF VIOLENT CRIMES

BY MINGXIN (IVY) LIN

Faculty of Mathematics, University of Waterloo

1. Introduction.

1.1. *Motivation.* In the United States, violent crimes are defined as incidents involving force or the threat of force. The main offences reported under violent crime are murder and non-negligent manslaughter, rape and sexual assault, robbery, and aggravated assault (FBI,2023)[1]. Compared to other high-income nations, U.S violent crime rates were 7.0 times higher, primarily driven by a staggering gun homicide rate that was 25.2 times higher (Grinshteyn & Hemenway, 2010)[2]. Clearly, there exists a significant firearm problem within the United States compared to its peers, resulting in the issue of gun laws becoming a heavily debated topic in the country. Concealed handgun carry, commonly referred to as "shall-issue" laws, is permitted in numerous states. We were inspired by the original researcher, John R. Lott Jr., who tried to determine whether or not citizens should be allowed to carry a concealed handgun, and if so, will it reduce violent crimes? He concluded that shall-issue laws have no effect on violent crimes, and thus we will be taking a different approach.

1.2. *Research question.* Our research question is to determine the 3 covariates with the largest level of influence on violent crime rates. To address the research question, our objective is to build an inference model that best explains the relationship between the violent crime rate and other covariates based on our existing observations. Once we identify the best fitted model, we will rank the covariates based on the magnitude of a one unit change in violent crime rate when holding other variables constant, and select the top three. This analysis aims to support governments in initiating discussions on policy making and implementing targeted programs that could effectively reduce crime rates.

2. Data.

2.1. *Description of the Dataset.* We have chosen the "Guns" dataset under the R package "AER" (CRAN, 2022)[3]. The dataset is compiled by the U.S. Department of Justice to analyze the effect of gun-control laws on violent crimes. It features a balanced panel of data from 1977–1999 on 50 U.S. states, plus the District of Columbia, for a total of 51 states. Each observation is a given state in a given year. Given the relatively low number of rows containing NA values, we dropped those rows, resulting in a dataset consisting of 51 states × 23 years = 1173 observations. Moreover, the dataset contains information on various demographic and socioeconomic factors that may be potential predictors of violent crime. We formally define all 13 variables within our dataset: `state`, factor indicating state; `year`, factor indicating year; `violent`, violent crime rate (incidents per 100,000 members of the population); `murder`, murder rate (incidents per 100,000); `robbery`, robbery rate (incidents per

Keywords and phrases: .

[1] https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime

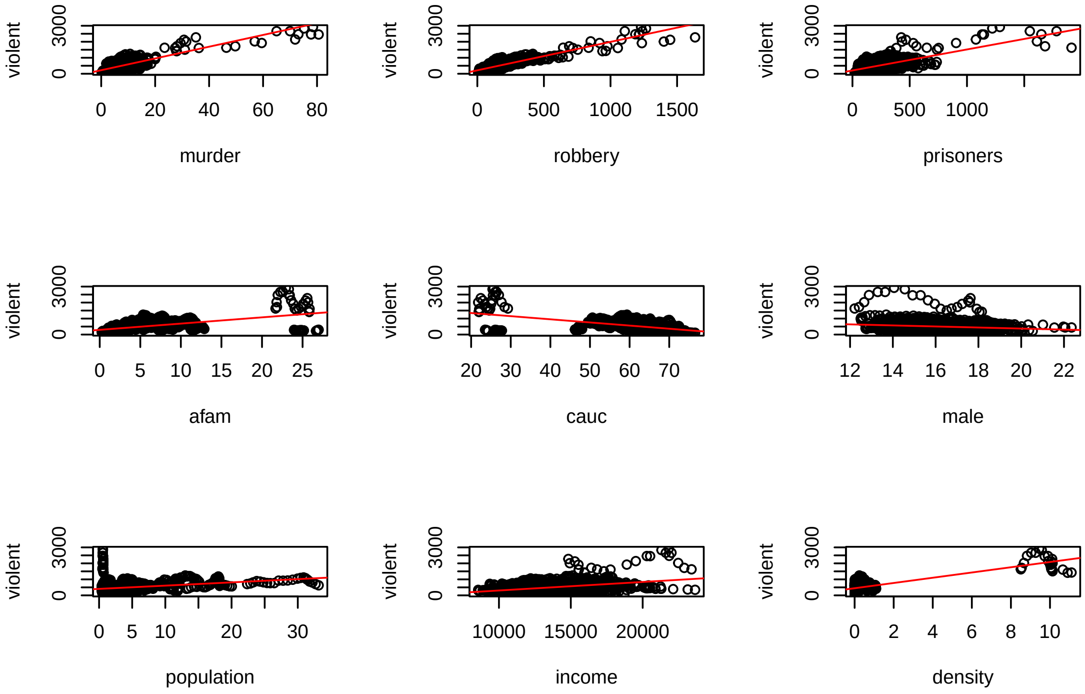
[2] https://doi.org/10.1016/j.amjmed.2015.10.025

[3] https://search.r-project.org/CRAN/refmans/AER/html/Guns.html

100,000); prisoners, incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents); afam, percent of state population that is African-American, ages 10 to 64; cauc, percent of state population that is Caucasian, ages 10 to 64; male, percent of state population that is male, ages 10 to 29; population, state population, in millions of people; income, real per capita personal income in the state (US dollars); density, population per square mile of land area, divided by 1,000; law, factor for "Does the state have a shall carry law in effect in that year?"

We will be modeling violent crime rates (vio) against nine continuous covariates (murder, robbery, prisoners, afam, cauc, male, population, income and density). We exclude law from the model as the original study has shown that it is not associated with vio. We also exclude state from the model as our focus is not on determining state-specific predictors of violent crime rates.

2.2. *Initial Exploratory Analysis.* Upon plotting each covariate against violent crime rates and adding linear regression lines using lm, we noticed that most plots display a non-linear relationship. However, for murder and robbery, a nearly linear relationship is evident, indicating that multiple linear regression could be a reasonable starting point, but it will likely not provide the best fit for our data. In addition, the plots for murder, robbery, prisoners, afam, and density display two distinct clusters. One cluster contains the majority of the observations, while the other cluster comprises only 10 to 20 observations. The presence of outliers may require special treatment during model fitting. Furthermore, the regression lines for male and income are nearly flat, suggesting that these covariates may not have a significant impact on violent crime rates.



3. Methods.

3.1. *Data Preperation.* To build a good inference model, we ensure that our model selection process is independent of the information contained in the inference set. Therefore, we randomly sample from the original dataset, allocating approximately 90% (1063 observations) of the data for training 10% of the data for testing (110 observations).

3.2. *Multiple Linear Regression.* We start our analysis by fitting a multiple linear regression model to all 9 covariates. Let $x_1 =$ murder, $x_2 =$ robbery, $x_3 =$ prisoner, $x_4 =$ afam, $x_5 =$ cauc, $x_6 =$ male, $x_7 =$ population, $x_8 =$ income, $x_9 =$ density. Our coefficients are estimated by $\beta = (X^T X)^{-1} X^T y$, and thus we define the full linear model as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \epsilon, \epsilon \sim N(0, \sigma^2)$$

3.2.1. *Checking for Collinearity.* A notable finding was the evaluation of the Variance Inflation Factor (VIF). We observed that the variables `afam` and `cauc` exhibit exceptionally high VIF values (Table 1). This indicates strong collinearity between these two covariates, which can affect the reliability and interpretation of the regression results.

TABLE 1
VIF for Full Linear Model

Covariate	Murder	Robbery	Prisoners	Afam	Cauc	Male	Population	Income	Density
VIF	6.0050	6.1079	4.1380	42.2080	42.0050	2.4588	1.6896	2.5363	4.3158

Thus, we use a ridge regression model to further check for collinearity between `afam` and `cauc`. We still have the same linear model, but now we introduce a penalty term $\lambda > 0$ and estimate β using:

$$\hat{\beta}(\lambda) = \text{argmin} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{i=1}^{n} \beta_j^2$$

When we increased the penalty from $\lambda = e^1$ to $\lambda = e^{5.5}$, the coefficient associated with `cauc` shrunk to 0, while the coefficient associated with `afam` has a magnitude of 1.63. As a result, we have dropped `cauc` in all future models. We refit our multiple linear regression model without `cauc` and noticed that the VIF decreased significantly (Table 2). Thus, we have 2 candidate models to score in our model selection process: the full model, and the reduced model after removing `cauc`.

TABLE 2
VIF for Reduced Linear Model

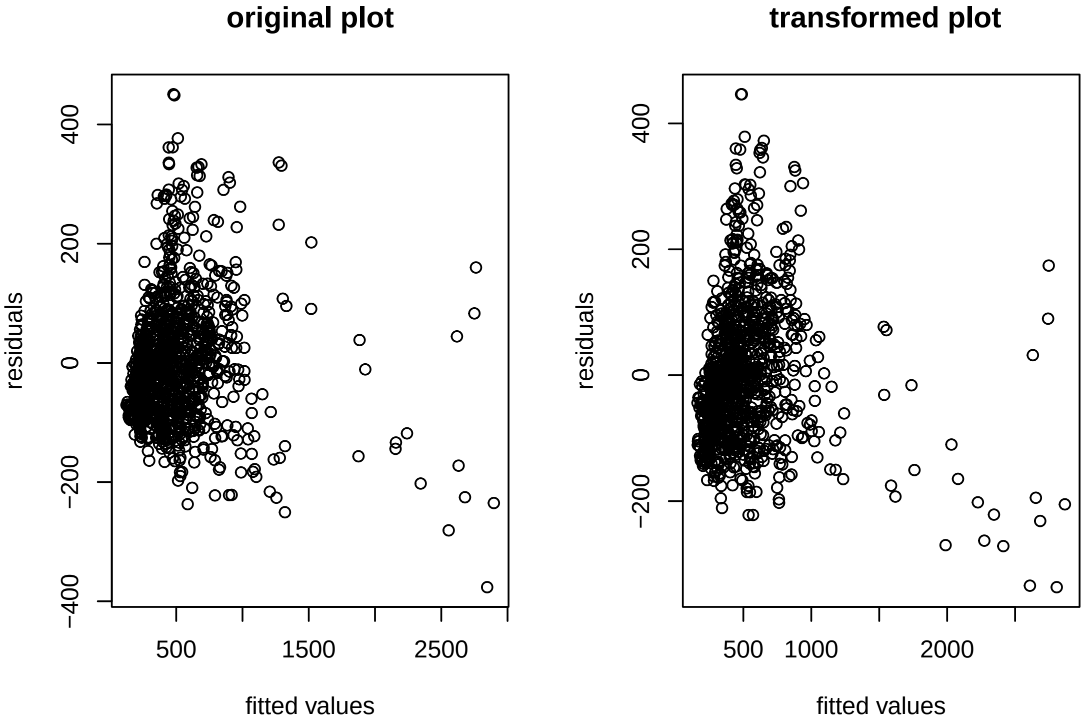
Covariate	Murder	Robbery	Prisoners	Afam	Male	Population	Income	Density
VIF	5.9802	6.1058	4.1232	1.8272	2.2155	1.6412	1.8606	4.1484

3.2.2. *Checking Assumptions.* We now check for the independence of covariates and the constant variance assumption. We see that the independence assumption is not met well. Although we have addressed the multicollieartity issue, we still see some relatively big correlation coefficients (Table 3). Note that we have about 26 observations for each U.S state, but we have reason to suspect that some of the observations within each state might originate from geographically close regions. Thus, it provides an explanation as to why our covariates might not be entirely independent.

TABLE 3
Correlation Matrix

	Murder	Robbery	Prisoners	Afam	Male	Population	Income	Density
Murder	1.00	0.80	0.71	0.60	0.01	0.10	0.22	0.75
Robbery	0.80	1.00	0.57	0.58	-0.09	0.32	0.41	0.78
Prisoners	0.71	0.57	1.00	0.53	-0.45	0.10	0.46	0.56
Afam	0.60	0.58	0.53	1.00	0.02	0.06	0.26	0.54
Male	0.01	-0.09	-0.45	0.02	1.00	-0.10	-0.53	-0.06
Population	0.10	0.32	0.10	0.06	-0.10	1.00	0.22	-0.08
Income	0.22	0.41	0.46	0.26	-0.53	0.22	1.00	0.34
Density	0.75	0.78	0.56	0.54	-0.06	-0.08	0.34	1.00

Moreover, the residual vs. fitted values plot below displays a funnel pattern and suggests that the model has a non-constant variance. Hence, we suspect that some transformation on the predictor variables should be performed to conform to the homoscedastic assumption of regression. We tried a logarithmic and square transformation on suspicious covariates (e.g. `log(density)`), however no significant change was observed. Since the independence assumption and the constant variance assumption does not hold well, it suggests that a linear multiple regression model may not be a good choice. We now explore models that can handle non-linear associations.



Lastly, we suspected that `income` and `male` may not have a significant impact on violent crime rates. Upon evaluating the summary outputs of the multiple linear regression models, we see that both covariates are statistically significant (p-value <0.5).

3.3. *Additive Spline Model.* Let $f_l(x) = \sum_{j=1}^{d_l} f(x_j)\beta_{l_j}, l = 1,..,p$ be a cubic spline curve and $\epsilon_i \sim N(\mu_i, \sigma^2)$. An additive model is mathematically defined as,

$$y_i = \alpha + f_1(x_{i1}) + ... + f_p(x_{ip}) + \epsilon_i$$

We begin by building an initial additive model with 8 covariates: `robbery`, `murder`, `prisoners`, `afam`, `male`, `population`, `income`, and `density`. We utilize the `gam` function to automatically select the tuning parameter λ based on the generalized cross validation score of the training set.

3.3.1. *Checking for Interactions.* We have identified a potential for interactions among the covariates `murder`, `robbery`, and `prisoners`, as these crimes share a similarity in nature. Consequently, we introduce three interaction terms to account for these possible interactions in our analysis: $\gamma_{1,2} =$ murder × robbery, $\gamma_{1,3} =$ murder × prisoners, and $\gamma_{2,3} =$ robbery × prisoners.

For each interaction term, there are two options: it can either be included in the model or excluded. Therefore, we have a total number of $2^3 = 8$ possible models to score as part of our model selection process. Namely, we will have 1 model with no interaction terms, 1 model with all 3 interaction terms, 3 models with 1 interaction terms, and 3 models with 2 interaction terms.

3.3.2. *Checking for Outliers.* Referring back to our explanatory analysis, we saw there was a cluster of outliers across almost every covariate. Analyzing the original data, we see that about 25 observations are outliers and most of them are from Washington DC. This can be attributed to DC's reputation as America's "Murder Capital." To determine the importance of these outliers, we will remove observations associated with DC state from just the training set and refit to all possible additive models defined above. This gives us 8 new models, and thus we have a total of $8 + 8 + 2 = 18$ candidate models to score in our model selection process.

4. Results.

4.1. *Model Selection Process.* To evaluate the performance of our 18 models, we employ a squared loss function against the test set, selecting the model with the lowest squared loss as the best fit. Initially, we score all models against the unmodified dataset. The results indicate that the additive model with interactions `murder*robbery` and `prisoner*robbery` achieves the best score, while the reduced linear model exhibits the poorest performance (Table 4).

TABLE 4
Score of Linear Models and Additive Models (Rounded)

model	mlr1	mlr2	am1	am2	am3	am4	am5	am6	am7	am8
score	11635	11657	7729	7206	6895	7452	6841	18803	6735	8880

Subsequently, we evaluate our models using a dataset from which observations associated with the DC state have been removed. All 8 fitted additive models display notably lower scores compared to the models trained on the unmodified dataset (Table 5). This suggests that the outliers associated with the DC state are important observations that should be used for model training.

TABLE 5
Score of Additive Models without DC Observations (Rounded)

model	amod1	amod2	amod3	amod4	amod5	amod6	amod7	amod8
score	22071278	25047296	22293766	2942279	74107464	33130319	16958854	53219252

4.1.1. *Covariates with the Largest Level of Influence.* We have identified the additive model with interactions `murder*robbery` and `prisoner*robbery` is the best model (Model A). We use this model to determine the 3 covariates with the largest level of influence on violent crime rates. To assess the level of influence of each covariate, we standardize them using the formula $\frac{x_i-\mu_i}{\sigma_i}$ and refit the model with this standardized data (Model B). Standardizing the covariates ensures that each one contributes equally to the model and serves as a fair benchmark for comparisons. To measure the level of influence, we examine the magnitude of a one unit increase for each covariate on the response variable. Specifically, we take the difference between the predictions of Model B and Model A using the `predict.gam` function. After analyzing all possible one unit increases of each covariate, we observe that `density` (28427.22), `afam` (14621.94) and `robbery*prisoners` (13069.72) exhibit the largest magnitudes.

5. Conclusions.

Based on the squared loss scores on the test set, our analysis reveals that additive models outperformed multiple regression models, and is a good method to model nonlinear associations. In addition, additive models with interactions performed better across the board, thus suggesting important interactions between specific covariates. Moreover, models trained without outliers demonstrated poor model performance, and thus implying that these observations are important in the training process. In addition, the additive models with interactions performed better across the board, delivering the most favorable outcomes in explaining violent crime rates.

To summarize, our analysis identifies `density`, `robbery*prisoners`, and `afam` as the three most influential covariates affecting violent crime rates. These findings suggest that regions with higher murder rates and prisoner populations, a higher proportion of African American residents, and greater population density may experience elevated levels of violent crime compared to other areas. As a critique to the original researcher, these insights suggests that rather than solely relying on laws and regulations, focusing on targeted programs that enhance American neighborhoods, and ensure safety for minorities may be a more effective and sustainable approach to reduce violent crime rates.

5.1. *Limitations and Future Work.* Our analysis relies on historical data collected between 1977 and 1999. We must acknowledge that using older data may not accurately reflect the current state of violent crimes, nor the covariates being measured. As a result, the insights provided in this report may not be a good representative of present-day conditions, leading to lower data reliability. In future work, we would look to using an updated dataset from sources near or up to year 2023, and possibly increase the number of observations per state. We could also try to identify influential covariates on violent crime rates at a state-specific level, rather than making generalizations across all states. For instance, we recognized that the District of Columbia (DC) possesses unique characteristics, and thus, the factors contributing to violent crime rates in DC may differ from those in a relatively safe state. By conducting state-specific analyses, we can gain deeper insights into the specific drivers of violent crime within each state, and consider how diverse socioeconomic and demographic patterns affect crime rates.