County Economic Factors Effect on Violence

Brody Erlandson and Danielle Fanizzi

December 2021

1 Summary

Using a longitudinal model, we want to predict what effects the prevalence violence in a school district. We look into what economic factors can help predict this prevalence. These economic factors include government assistance, poverty, and employment. When fitting models, we look at both models that include and exclude year as a predictor and models that do and do not split on sex.

When splitting on sex, we are able to capture an interesting difference in males and females. Furthermore, we see find that although year lowers a model's AIC value, year often does not improve the model's CV-MSE.

2 Introduction

Violence is a prevalent disruption surrounding high school settings, which is often a distraction from learning and a detriment to both a student's mental and physical health. Schools are intended to be a safe space away from home for students to focus on improving his/her self and preparing for his/her future but instead have often become a place of fear or violence for many. Violent actions often are a reaction to or an out lash from a students environment, so we hope to link certain economic situations of a county to the violent severity of schools in that county in order to determine what economic factors play the largest role.

In this document, we will compare three different model : (1) a linear mixed effect model with county as the random effect, (2) a linear mixed effect model with county as the random effect and the data split by sex, and (3) a spline model with the county again as the random effect and the data again split on sex. In all models, a calculated violence score will be our response variable and the economic factors will be used as our predictors.

3 Data Descriptions and Cleaning

We use three different data sets : (1) the Youth Risk Behavior Surveillance System (YRBSS), (2) Supplemental Nutrition Assistance Program (SNAP) data set, and (3) Small Area Income and Poverty Estimates (SAIPE) Program data set.

3.1 Youth Risk Behavior Surveillance Survey (YRBSS)

The YRBSS is a survey from the Centers for Disease Control and Prevention (CDC) given to high school students pertaining to violent behavior, drug use, sexual activity, depression and suicide, and general risky behavior. We will use the responses to the questions surrounding violence to formulate a violence score for each county. These question include topics such as frequency of verbal violence, frequency of fist fitting, and even frequency of carrying a gun in and out of school.

3.2 Supplemental Nutrition Assistance Program (SNAP)

SNAP is a food and nutrition service from the United States Department of Agriculture (USDA). We will use the number of dollars a state spends per participant per year as a predictor. Because SNAP is a state wide effort, all counties within a state will have the same SNAP value.

3.3 Small Area Income and Poverty Estimates (SAIPE)

The SAIPE Program data set is collected by the United States Census and contains information such as poverty rates, unemployment rates and median household income for each county, which will also be used as predictors.

3.4 Imputation and Violence Score

We use four different ways of imputation to account for missing data. The imputations we will consider are (1) the k nearest neighbors (KNN), (2) the mean, (3) the median, and (4) the median. We impute into the violence questions used to create the response. We based the KNN off of the other questions, making sure to *not* based the impute off the predictors. Thus, the response impute is only based off the question given in the survey, and not any of the predictors.

Additionally, our response variable is a calculated violence score using the YRBSS data set. The violence score is calculated by : (1) selecting the question pertaining to violent behavior, (2) assigning point values to each answer based on violence frequency, (3) assigning weights to each question based on violence severity, (4) adding together the points from each question multiplied by the weight of the question for each student in the survey, and (5) adding together the individual violence scores and divide by the saturated violence score for each county. The saturated violence score is calculated by assuming all students selected the most severe answer. Using these calculations, our violence score lies between 0 and 1. If a county has a violence score closer to 1, the county's students have more violent behaviors. For modeling purposes we multiply by 1000, thus the violence score is 0 to 1000.

3.5 Final Data Setups

Our final data set consists of 14 countries and 9 years ranging from 2003 to 2019. The predictors include (1) unemployment rate, (2) SNAP (dollars spent per participant), (3) population size, (4) median household income, (5) poverty percentage for children under 18, and (6) poverty percentage for all ages.

After calculating the violence score, we plotted the violence score against the predictors. Below in figure 1, we can see the violence score against year colored by county. There appears to be a downward tend in violence score as the years become more recent. One explanation for this may be because violence has turned from physical violence to more verbal violence.

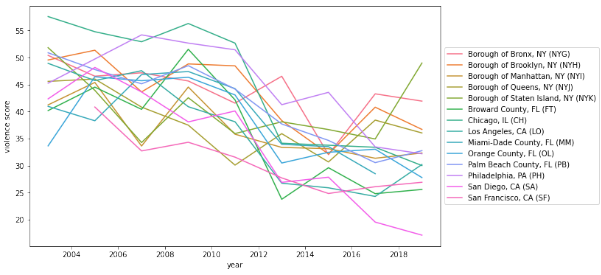

Figure 1: Year vs. Violence Score

Next, we see the violence score against unemployment rate in figure 2. There seems to be a slight positive correlation between violence score and unemployment rate. This trend makes sense because as unemployment rate increase, a student may not be as supported at home which may affect their behavior and out lashes in school.

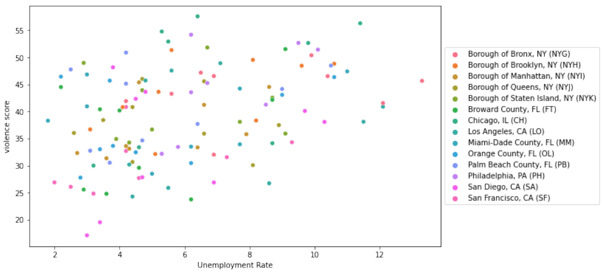

Figure 2: Unemployment Rate vs. Violence Score

Last, we plot the violence score against the SNAP dollar amount per participant over year in figure 3. There appears to be a slight negative correlation between violence score and the SNAP dollar amount. As a student is more nourished and fed, they may feel less drawn towards acting out in violence.

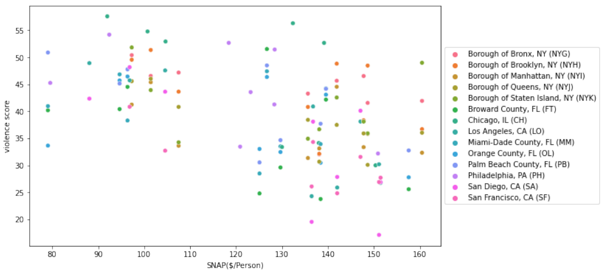

Figure 3: SNAP ($/Participant) vs. Violence Score

4 Methods

In this section, we will be comparing three different methods to analyze what economic factors affect a county's violence. The data used for each model contains the odd years from 2003 to 2019. In all methods, we will be using the variable, county, as a random effect. We will be comparing both the Akaike information criterion (AIC) and cross validation mean square error (CV MSE) among the different models to determine the predictors needed and the best fit model for our data. Where the CV MSE leaves out two counties, fits the model, and finds the MSE of the two left out. We do this for every pair of counties.

We start by analyzing the correlation between predictors as seen in figure 4 that could lead to singularity in the model and misinterpretation of results. High correlations such as the correlation between the poverty percentage of all ages and the poverty percentage of children under 18 must be accounted for when completing variable selection.

4.1 Linear Mixed Effect Model Not Split on Sex

The first section explores the linear mixed effect model where sex of the students was not accounted for. This assumes there is no difference in violence tendencies between male and female students. County is given as a random intercept to allow for model adjustment between counties.

4.1.1 No Higher Order Terms

For variable selection, we use backward elimination comparing AIC resulting in

$$Y = \beta_0 + b_{0,County} + \beta_1 * UnemploymentRate + \beta_2 * Year.$$

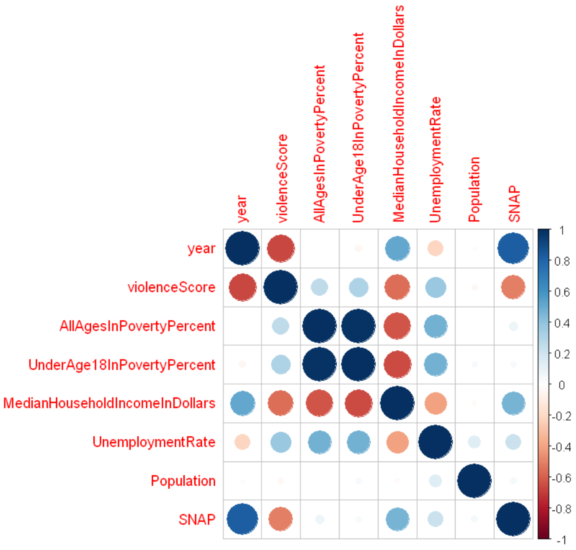

Figure 4: Correlation between Variables

Figure 5 depicts the different intercepts for each county and confirms the model follows the trends of the data.

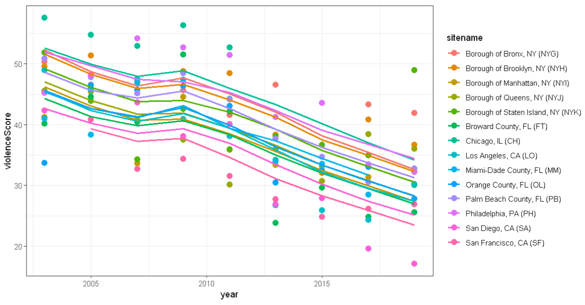

Figure 5: Year vs Violence Score

We also determine the best model without including year as a possible predictor again using backward elimination. This variable selection resulted in

$$Y = \beta_0 + b_{0,County} + \beta_1 * UnemploymentRate + \beta_2 * SNAP.$$

As seen in figure 6, the model using year as a predictor performed slightly better than without using year. There are some increase in trends in the first third of the model that are not present in the data.

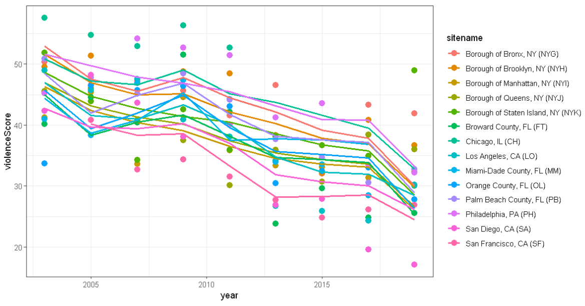

Figure 6: Year vs Violence Score

4.1.2 Higher Order Terms

We now add higher order terms for both year and unemployment rate to determine if we can find a better fit for the violence score. Again, we compare AIC which results in

$$Y = \beta_0 + b_{0,County} + \beta_1 * UnemploymentRate + \beta_2 * Year + \beta_3 * Year^2 + \beta_4 * Year^3.$$

Figure 7 show the higher order terms on year are able to better capture the trend slight upward trend for the most recent years compared to the model without higher order terms. This also signifies that there may not be a complete linear relationship between the violence score and year, which we will explore later in the document.

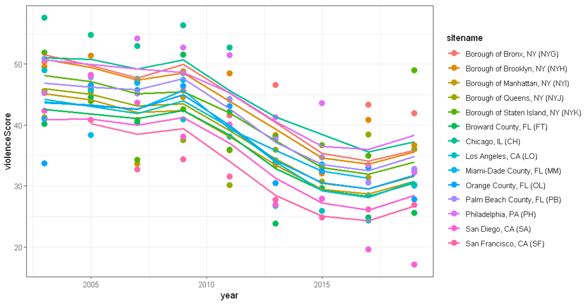

Figure 7: Year vs Violence Score

4.1.3 Imputation Comparisons

We next want to look at the four different ways of imputation to account for missing data. The imputations we will consider are (1) the k nearest neighbors (KNN), (2) the mean, (3) the median, and (4) the median. During this analysis, we will use the linear mixed effect model with only year and unemployment rate as the predictors.

In order to test the models performance on data not included in the data set, we calculate and compare the CV-MSE. Here the testing data set includes one set of data for each county. The training data set includes the additional data not included in the testing set. We also consider the AIC for each. From table 4 we see that the median and mode imputation methods perform the best; however, all methods perform similarly for both AIC and CV-MSE. Because of the similarity, we will continue further imputations using KNN. KNN best captures the surrounding trends that may be present when splitting on county.

Imputation Type	AIC	CV-MSE
KNN	780.8245	31.1571
Mean	778.2661	30.7916
Median	772.6461	29.7111
Mode	772.6461	29.4925

Table 1: Comparison of Performance for Imputation Methods

4.2 Linear Mixed Effect Model Split on Sex

We hypothesized that the violence scores for males and females would be different if split. We want to look into modeling based on this split to see if there is a further interpretation that could be gathered from this split. We split the data before getting the violence score and then calculate a violence score for females and males separately. Below we can see the change of violence scores over the years for males and females separately.

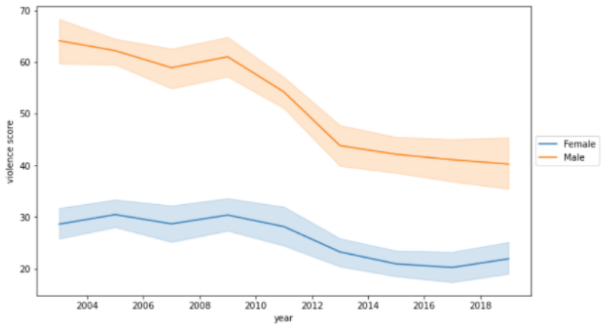

Figure 8: Split on Sex: Year vs Violence Score

We can see a clear distinction in the violence scores between males and females. Not only is the males violence scores larger, but it also seems to have a different rate of change over the years.

4.2.1 No Higher Order Terms

We tried to model this split with both year and no year included in our model due to the year not telling us much about what is affecting the violence to go up or down in a district. Particularly, we want to see if the violence could be captured by not including year just as well as including year.

With all the parameters including year and not including year, we fit a model to every parameter combination. We then retrieve the AIC for each model. After this we found the best model with year is

$$(1)\ Y = \beta_0 + \beta_{0,Male} + b_{0,County} + \beta_1 * UnemploymentRate + \beta_2 * Year + \beta_{2,Male} * Year$$

Also, we found the best model without year is

$$(2)\ Y = \beta_0 + \beta_{0,Male} + b_{0,County} + \beta_1 * UnemploymentRate + \beta_{1,Male} * UnemploymentRate + \beta_2 * SNAP + \beta_{2,Male} * SNAP$$

$b_{0,County}$ is a random intercept for each county. Below we see a table of CV-MSE and AIC for model (1) and (2).

Model	AIC	CV-MSE
(1)	1588.88	50.20365
(2)	1621.03	50.34290

Table 2: Comparison of Performance for Split on Sex Methods

We can see that the model with year has a significantly better AIC; however, the CV-MSE is only slightly better suggesting again that the model with year does no better on Out-of-Sample data than the model without year. While looking into the CV-MSE, we also tried separating the male and female predictions to see if the model predicts them both at the same accuracy. Below we can see the table for the CV-MSE of the female and male separately.

Model	Male CV-MSE	Female CV-MSE
(1)	63.35	37.05
(2)	65.48	35.20

Table 3: Comparison of Performance for Male vs Female

We can clearly see that the Males CV-MSE is much worse than the Females CV-MSE. This Motivated our *Numerical Study (S5)*. From the graph above we can see the male's violence score has a more abnormal change over time, so we also wanted to try higher order terms to see if this could lower the difference in prediction for the two sexes.

4.2.2 Higher Order Terms

We found our model for the higher order terms in a similar manner to what we did above. We fit all parameter combinations to the model and found which one had the lowest AIC value. To make it easier to code, we tried all first order, second order, and third order terms separately. For example, for the second order terms, if we had parameter w and x, then we would try the models $y = x + x^2$, $y = w + w^2$, and $y = x + x^+ w + w^2$. Additionally, we evaluated models containing every term with and without the interaction with sex.

After finding the best models at every order, we tried to lower the order of the terms that were not significant. Again take the example above : if the best model was $y = x + x^+ w + w^2$, but the w^2 was not statistically significant, then we'd drop w^2. We then calculated the CV-MSE for the best reduced models. Despite some models that were overly complicated, we found that all the models with year had higher CV-MSE than those without year. Even the overly complicated ones with year did not do better than the ones without year. Due to this, we only put forward one model:

$$Y = \beta_0 + \beta_{0,Male} + b_{0,County} + \beta_1 * UnemploymentRate + \beta_{1,Male} * UnemploymentRate + \beta_2 * SNAP + \beta_{2,Male} * SNAP + \beta_3 * SNAP^2 + \beta_4 * SNAP^3$$

The performance table is shown below.

Total CV-MSE	Male CV-MSE	Female CV-MSE	AIC
42.41	54.77	30.06	1575.83

Table 4: Performance for Higher Order Model

We can see these higher order terms predict out of sample data better; however the males CV-MSE is still quite a bit higher than the females. We also see some improvement in the *Numerical Study (S5)* results.

4.3 Spline Model Split on Sex

Based on the outcomes in the previous sections, we now look at the use of spline fitting to check for non linearity between the violence score and the predictors. We use the generalized additive model and smoothing terms in order to accomplish this. After comparing the AIC for different combinations of parameters, we determine that adding a smoothing function to both unemployment rate and SNAP is the best model. Again, a random intercept is included for each county. Next, we use the CV-MSE to determine the appropriate number of knots for both SNAP and unemployment rate. The knots of the spline fitting affect the complexity of the model fit. A higher number of knots may lead to over fitting, while a fewer number of knots may lead to too much of a generalization. When calculating our CV-MSE, we use two counties' information for the testing set. The remaining counties' information are used as our training set. After considering the total CV-MSE and the two split CV-MSE for male and female, we determine the best fit model to be

$$Y = \beta_0 + b_{0,County} + s(SNAP, knots = 17, by = sex) + s(UnemploymentRate, knots = 6, by = sex)$$

which include 17 knots for the SNAP smoothing parameter and 5 knots for the Unemployment Rate smoothing parameter as shown in table 5.

5 Numerical Study

Given the male and female CV-MSE were quite different, we wanted to look into how the parameters vary and the distribution of response variance. We look to do this through the bootstrap described below:

Model	AIC	CV-MSE	Male-MSE	Female-MSE
Neither Smooth	2015.223	221.7447	236.421	207.0683
SNAP Smooth	1546.375	38.3705	47.4142	29.3267
Unemployment Rate Smooth	1636.071	58.0098	73.3805	42.63901
Both Smooth : Same Knots (K = 17)	1523.5252	37.9195	47.4138	28.4253
Both Smooth : Best Knots ($K_S = 17$, $K_U = 6$)	1521.8373	36.6879	45.7331	27.6428

Table 5: Comparison of Performance for Spline Models

1. First, we re-sample with replacement the survey answers from each county and year separately. We take 2,000 samples for each county and year separately for the bootstrap sample.

2. Next, we take the bootstrap sample and recalculate the violence scores. Then, we fit the model to these bootstrap violence scores.

3. We then take the parameters from the newly fitted model. We take the variance of the predicted responses and the true responses for each of the years.

Below are the plots of the variance from the bootstraps split on females and males. This first plot is for the model split on sex with no higher order terms. We can see the response bootstrap variance is quite different than the true variance distribution. This suggests the model does not capture the males variance very well or the mean well, although it captures the spread well. Below is the plot for the model split on sex with higher order terms. We see the same problems as above; however, the mean seems to be predicted better than the model without higher order terms. We should look into fixing the model to capture the variance better than it is currently being captured.

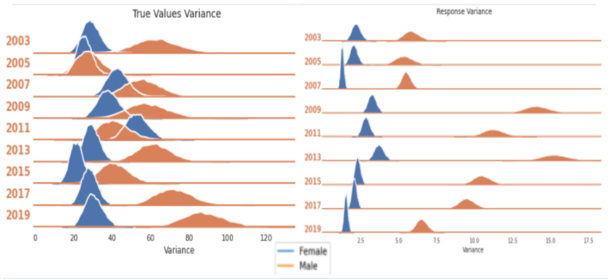

Figure 9: Each Year's Bootstrap Variance Distribution

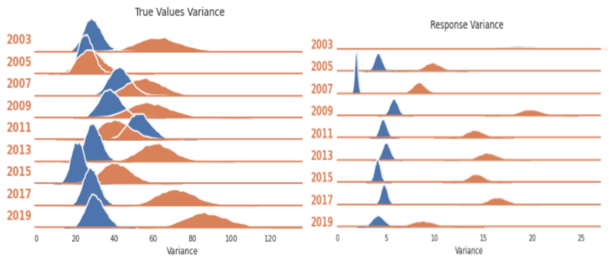

Figure 10: Each Year's Bootstrap Variance Distribution Higher Order

In figure 11, we see the parameter bootstrap distributions. The variances for the males are much larger for all coefficients compared to the variances for the females.

To continue our exploration, we look to see which of the coefficients have the largest CV-MSE. From table 6, we see that the unemployment coefficient has the largest CV-MSE for both the males and the females. This may explain why higher order terms for unemployment rate are not significant, which was also seen with the smaller number of knots used in the spline fitting model in the above section. The higher order terms would capture more of the noise present in unemployment rate.

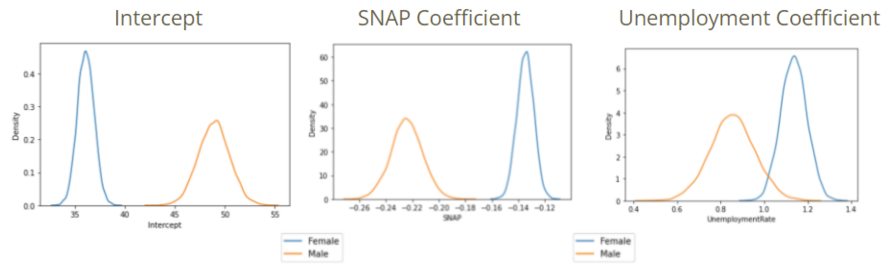

Figure 11: Parameter Bootstrap Distributions

Coefficient	Male : CV-MSE	Female : CV-MSE
Intercept	0.0320	0.0230
SNAP Coefficient	0.0529	0.0470
Unemployment Rate Coefficient	0.1202	0.05256

Table 6: Comparison of Parameter Bootstrap Distributions

6 Conclusion

As seen in the sections above, violence among high schools can be predicted by a counties economic situation. To start, we have found that models using year as a predictor do not perform well for out of sample data. This may be because year captures more noise. Because of this, we prefer selecting models that do not include year. Additionally, we have seen the significance of considering sex when splitting the data. The model predicts male students to be more violent than female students. The male students also have a larger CV-MSE value. This difference is most likely because male students are more likely to be extremely violent or extremely non violent giving the group a larger variance. On the other hand, female students by nature are likely to be not violent and are less likely to be extremely violent giving the group a smaller variance. This difference in violence variance likely would effect the MSE values. Finally, the SNAP dollar amount per participant over a year and the unemployment rate have non linear relationships with the violence score as found when using the smoothing terms. With more support given to SNAP and larger effort to lower unemployment rates for parents with children, we hope to decrease the violent behaviors and minimize these distractions prevalent among high schoolers to create a safer school.

7 Appendix

- GitHub Link
- SNAP Link
- SAIPE Link
- YRBSS Link