Bleeding Red, White, and Blue: Predicting Homicide Rate in 2019

Xinxin Li, Christopher Meng, Jessica Sang, Shikha Shrestha

April 28, 2022

1: Introduction

Crime is important to look at because there are negative effects on both the victims and the communities in which they occur. In a study done by FiveThirtyEight, Majumder (2017) discovered that among U.S. states, income inequality as measured by the Gini index was a notable predictor of hate crimes in 2016. This study corroborates other research that finds a link between income inequality and violence (Hipp 2007). The wealth gap between the richest and poorest people in America has continued to grow (Horowitz, Igielnik, and Kochhar 2020). With this in mind, we were interested in exploring which variables were significantly associated with the homicide rate.

Using FBI data on homicide rate, the Gini index (a measure of income inequality), poverty percentage, median household income, uninsured percentage, high school degree percentage, bachelor's degree percentage, and political side for each U.S. state, we explore the following question: "Which variables have relatively significant effects on the homicide rate in the U.S. in 2019?" In addition to income inequality, other variables like education and the level of social support have also been found to affect crime (Bell, Costa, and Machin 2018). Therefore, we hypothesize variables associated with income inequality, like Gini index, poverty percentage, and bachelor's degree percentage, to have a linear relationship with the homicide rate.

In our exploratory data analysis, we first looked at a scatterplot matrix to get a sense of the relationships between the numerical variables. Almost every variable has some sort of relationship with the other variables; this pattern suggests that we should be cognizant of potential multicollinearity as we are choosing our predictors. Perhaps, most critically, there is a significant outlier that is visible on nearly every scatterplot, which can be identified as Washington D.C.

2: Regression Analysis

Our model predicts homicide rate from the following variables: poverty percentage, unemployment percentage, median household income in thousands of dollars (MHI), bachelor's degree percentage, the interaction between unemployment percentage and bachelor's degree percentage, and the interaction between median household income and bachelor's degree percentage:

$$\text{Homicide rate} = 26.964 + 0.696 * \text{Poverty } \% - 3.946 * \text{Unemployment } \% - 1.036 * \text{Bachelor's } \% - 0.201 * \text{MHI} + 0.131 * \text{Unemployment } \% * \text{Bachelor's } \% + 0.008 * \text{Bachelor's } \% * \text{MHI}$$

Before engaging in the model selection process that led to this regression model, we standardized the units of the potential numeric predictor variables to percentages for ease of interpretability. Our forward selection process started with a simple linear regression with poverty percentage because it had the highest R^2 for the model with one predictor. From there, we moved onto a model with two predictors and this second predictor was determined by seeing which of the remaining variables would increase adjusted R^2 the most. This variable turned out to be bachelor's degree percentage. Next, we included an interaction term between the two to see if there was any increase in R^2. The interaction term ended up being insignificant, so we removed

it and proceeded with including a third predictor variable. We repeated the steps above for choosing the second predictor until we arrived at our model. The final model garnered the highest adjusted R^2 of 0.73 with all of the terms remaining significant, except for median household income, and an ANOVA F-test p-value close to zero. Interpreting the model fit, our final model explains 73% of the variability in the response, and the overall ANOVA F-test suggests that our final model explains significantly more variance in the response than a model without any predictors (which guesses the mean homicide rate for each observation).

Our model satisfies some of the assumptions necessary for inference. We assume zero mean for the errors. While we have no strong reason to suspect that the errors are related to each other, each of the 50 states and D.C. do not exist independently of each other. For example, gun or crime legislation in one state may influence the discussion of policies in another state. In particular, mass shootings, and potentially murders, may be subject to "generalized imitation," where one person's behavior may influence another person's behavior, especially since homicide is a crime that comes with a significant amount of social attention and reporting from the media (Meindl and Ivy 2017). Looking at the residual plot, the linearity and constant variance assumptions are not fully satisfied as we see slight curvature in our data points and nonconstant variance, with the residuals being a lot larger at certain fitted values than others. Looking at the normal quantile plot, normality is met with most of our data points falling on or close to the line, with a couple of outliers at the ends. Finally, with respect to the randomness condition, we do not satisfy it because there is nothing random about the way our observational data was collected; instead, it is like census data.

The coefficients of our model suggest the following. For every 1% increase in the poverty percentage, there's a predicted 0.70 increase in the homicide rate on average. For every 1% increase in the unemployment percentage, there's a predicted 3.95 decrease in the homicide rate on average. For every 1% increase in bachelor degree percentage, there's a predicted 1.04 decrease in the homicide rate on average. For every \$1,000 increase in the median household income, there's a predicted 0.20 decrease in the homicide rate on average. All of our coefficients are significant at $\alpha = 0.05$, except for median household income. Interestingly, unemployment percentage has a negative relationship with homicide rate, which is opposite of what we would expect. However, the significant, positive interaction term between unemployment and bachelor degree percentage suggests that the more educated a population is, the greater the effect of unemployment percentage on homicide rate. Examining the dataset more closely, this interaction term may be overfitting the values of D.C., where 70% of the population has a bachelor's degree (with the next highest being 53% in Massachusetts). The homicide rate in D.C. of 23.4 is also magnitudes higher than the lowest homicide rate of 1.6. A similar phenomenon may be happening for the interaction term between median household income and bachelor degree percentage, where the model may be overcorrecting the relationship between income and homicide rate due to the outlying value of bachelor degree percentage for D.C.

3: Discussion & Limitations

In recent years, a number of studies have pointed to a relationship between economic inequality and crime (both violent and property crime) within and across countries (Kelly 2000; Demombynes and Özler 2005). Such studies have tended to use income inequality as their primary predictor for crime. Blau and Blau (1982) found a strong relationship between economic inequality and violent crimes while studying 125 U.S. metropolitan areas. The main purpose of our study was to identify the key variables among socioeconomic indicators that could predict the homicide rate in the U.S. Four variables proved to be significant in our regression analysis: unemployment percentage, bachelor's degree percentage, poverty percentage, and median household income. Poverty percentage had a positive relationship with homicide rate, while median household income, unemployment percentage, and bachelor's degree percentage had a negative relationship with homicide rate. We believed that having an undergraduate degree would increase the likelihood of both employment and access to a higher income bracket, so we introduced two interaction terms consisting of a pair of these three variables. While both interaction terms proved to be statistically significant, the significance may be due to the unusual predictor and response values for D.C.

There are several limitations to our model. While our model can be used as a descriptive tool, multicollinearity among the predictors can make it difficult to ascertain the individual effects of the predictors on

the response variable. Furthermore, having an interaction term makes descriptive interpretation more complicated by introducing additional caveats to a model. When it comes to prediction using an MLR model, the assumptions of linearity and constant variance are critical. While we attempted different transformations of the response and predictor variables, our residual plots still had some deviations from the assumptions, and our quantile plot indicated some evidence of skewness towards the tails. Finally, in terms of methods and analysis, we were limited by time and what we've learned so far in an introductory linear models course.

Our model would not be suitable for causal inference since our data is derived from observational data. Providing more evidence for the lack of randomness and the consequent inability to generalize findings beyond 2019, cross-validation of our final 2019 model with 2020 data demonstrated a significant 56% shrinkage. Generally, any value of shrinkage over 50% is concerning and indicates that our model may be overfitting the 2019 data and cannot accurately predict homicide rate for other years. However, it's also worth noting that 2020 was both an election year and the onset of the COVID-19 pandemic, two factors that undoubtedly influenced all of our predictor variables and the homicide rate. While our p-values may give us evidence of significant association within our 2019 data, we cannot use probabilistic models to ascertain statistical significance in the overall population of homicide rates in the U.S.

Considering our analysis is from a macro perspective, it could be misleading to apply the results to a local context as the dataset aggregated observations from the state level. While large variation exists between the states regarding socioeconomic and political indicators, there is also variation within the state. A similar limitation is that our analysis cannot be applied to other countries. In recent times, articles have been published regarding the reliability of homicide data from the FBI (our data source). According to this WSJ article, the FBI uses crude estimates to account for the missing data that some police agencies fail to submit (McGinty 2018). For example, the FBI was found to have inflated Indiana's numbers by 9.9%, West Virginia's by 13%, and Mississippi's by 68%. Additionally, the FBI only receives crime data from about eight-in-ten agencies (Horowitz, Igielnik, and Kochhar 2020). This brings into question the reliability of data pertaining to the states for which these estimates were used.

If we were able to redo the project, we would have liked to examine more categorical predictors. Likewise, we would consider placing less emphasis on the significance of the predictor terms while choosing the variables and consider the effects of the predictors on residual and quantile plots. We would also deliberate introducing more variables such as welfare support services, and divorce rates that could possibly explain more variability in homicide rates between the states. Lastly, we would consider using county-level data so that our analysis could be generalized at both micro and macro levels. Having more micro-level data would not only allow us to account for socioeconomic variability within a state, but it would also provide for a more flexible model that could also be used to identify variation among the states by grouping together counties from a particular state.

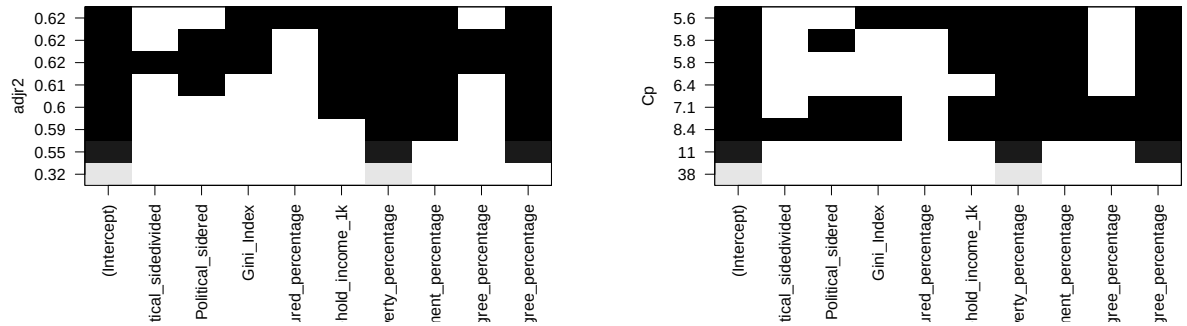
4: Conclusion

Our project focused on finding variables that significantly affected the homicide rate in U.S. states in 2019. Our model uses the predictor terms: poverty percentage, unemployment percentage, median household income, bachelor's degree percentage, an interaction term between unemployment percentage and bachelor's degree percentage, and an interaction term between median household income and bachelor's degree percentage to predict the homicide rate. Our adjusted R^2 of 0.7303 and small overall ANOVA F-test p-value suggest our model's predictors account for some of the variation in homicide rate. Namely, poverty percentage explains the most variability in homicide rate. It is important to note that we should proceed with caution when interpreting the coefficients of our model because the data does not fully satisfy the conditions needed for inference, namely the linearity and constant variance assumptions. Additionally, our data sources may not paint a complete picture of our predictor variables and homicides in the US in 2019. Noting these limitations to our analysis, it is important to continue investigating variables with a significant effect on the homicide rate as crime not only directly negatively impacts the victims, but also the greater community.

5: Additional Work

Predictor selection — adjusted R^2 and Mallow's C_p

For the selection of predictors, we initially attempted to use adjusted R^2 and Mallow's C_p. The results are shown below:



From both plots, the model with Gini index, uninsured percentage, median household income, poverty percentage, unemployment percentage, and bachelor's degree percentage gave the highest adjusted R^2 (0.622) and the smallest Mallow's C_p (5.636). We then checked the p-values for the t-test of the predictor terms and found that many predictors in this model are actually insignificant.

```
##
## Call:
## lm(formula = Homicide_rate ~ Gini_Index + Uninsured_percentage +
##     Median_household_income_1k + Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage, data = homicide_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8143 -1.4156 -0.6073  1.5930  5.3091
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -11.63668    9.43152  -1.234  0.22382
## Gini_Index                   -40.43014   25.78895  -1.568  0.12411
## Uninsured_percentage           0.19559    0.13048   1.499  0.14103
## Median_household_income_1k     0.10174    0.05774   1.762  0.08504 .
## Poverty_percentage             1.27864    0.26481   4.829  1.7e-05 ***
## Unemployment_percentage        1.06309    0.50576   2.102  0.04131 *
## Bachelor_degree_percentage     0.25183    0.07359   3.422  0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.248 on 44 degrees of freedom
## Multiple R-squared:  0.6675, Adjusted R-squared:  0.6221
## F-statistic: 14.72 on 6 and 44 DF,  p-value: 3.878e-09
```

Besides the insignificance in the predictor terms, the function `regsubsets` cannot give any information about the interaction terms, and though its adjusted R^2 (0.622) is not small, we still expected that our model could

explain more variability of the response variable. Therefore, we decided to use forward selection to find the best model.

Predictor selection — forward selection

Single predictor variable (simple linear regression model)

From the adjusted R^2 plot shown in the previous section, we can see that when the model has only 1 predictor (i.e. simple linear regression), `Poverty_percentage` gives the highest R^2. So we keep this predictor and try to add another one so that the adjusted R^2 can have a maximum increase.

Two predictor variables

From the adjusted R^2 plot and p-values for t-test, among all the models with 2 predictors, `Bachelor_degree_percentage` and `Poverty_percentage` gave the largest adjusted R^2 (0.5466) while maintaining predictors significant. Keeping these two predictors in the model, we then check the necessity of including the interaction term between these two variables into the model. We then found that adding the interaction term made all 3 predictor terms insignificant. Therefore, we decided to discard the interaction term and tried to add another predictor variable into the model.

Three predictor variables

From the adjusted R^2 plot, the combination of `Unemployment_percentage`, `Bachelor_degree_percentage`, and `Poverty_percentage` gives the highest adjusted R^2 (0.59) among all the models that have 3 predictors. The predictors also have tiny p-values for their t-tests. Therefore, we decided to keep them and try to add some interaction terms. We tried out each possible combination and found that the best choice was to add the interaction between `Bachelor_degree_percentage` and `Unemployment_percentage`, which raised the adjusted R^2 from 0.59 to 0.6887 while maintaining all terms significant. Other combinations of interaction terms either did not change adjusted R^2 much or made many predictors insignificant.

Four predictor variables

With `Unemployment_percentage`, `Bachelor_degree_percentage`, `Poverty_percentage`, and `Bachelor_degree_percentage:Unemployment_percentage` in the model, we tried to include another predictor variable. Adding `Median_household_income` gave the largest increase in adjusted R^2 (from 0.6887 to 0.7076). If we take $\alpha = 0.05$, the t-test for `Median_household_income` has p-value $= 0.052 > \alpha$. However, since adding this predictor does bring an obvious increase in adjusted R^2, we decided to keep this term and see how the interaction terms could change adjusted R^2. Since 4 predictor variables can bring too many possible combinations, we decided not to try out every possibility but instead terminate on one combination that is relatively satisfactory. It turns out that the interaction between `Median_household_income` and `Bachelor_degree_percentage` can raise adjusted R^2 up to 0.73, which is higher than the adjusted R^2 of any models that we've constructed. Though the term `Median_household_income` in this model has a p-value $= 0.175 > \alpha = 0.05$, we still decide to keep it in the model because the interaction term `Bachelor_degree_percentage:Median_household_income` is significant, and removing `Median_household_income` leads to a decrease in adjusted R^2. Therefore, we decided to end with the model with `Unemployment_percentage`, `Bachelor_degree_percentage`, `Poverty_percentage`, `Bachelor_degree_percentage:Unemployment_percentage`, `Median_household_income`, and `Bachelor_degree_percentage:Median_household_income`.

Five predictor variables

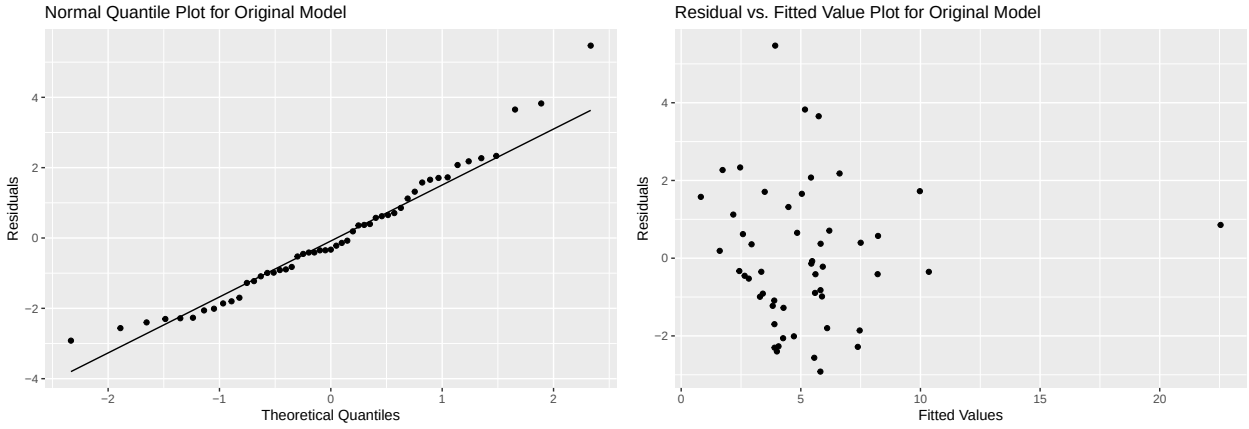
Though the model with the prediction terms stated above is satisfactory enough with regards to adjusted R^2, we still hope to see if it is possible to improve it a bit more. Therefore, we tried to add another predictor variable. However, after we made several attempts, we found that adding more predictor variables could not improve the adjusted R^2 anymore, and sometimes it even made multiple predictor terms insignificant. Therefore, we concluded that using more predictors may not be a good choice for this case.

Model without D.C.

We discovered that Washington D.C. was an extremely high leverage point ($h = 0.92$, with expected leverage of about 0.14). This value suggests that D.C. has the potential to influence the least squares line. Indeed, when we remove D.C. and attempt to fit the same model to the 50 U.S. states, the R^2, the amount of variability explained in the response by the model, decreases by 20%, and only poverty percentage remains as a significant predictor. We attempted to fit a different model to the dataset without D.C., but we found that no predictors in addition to poverty percentage explained significantly more variability in homicide rate than poverty percentage alone. Looking at the residual plot, removing D.C. does not change the violation of the linearity assumption as there is still a clear curved pattern among the points. The constant variance assumption may be slightly better met. However, the normal quantile plot displays further deviations from normality than compared to the proposed model. In particular, the right tail appears to be heavier, with larger values than we would theoretically expect. We decided not to use the model without D.C. because the assumptions are still violated, and conceptually, it feels too simplistic to predict homicide rate on the sole basis of poverty percentage. Directly linking poverty and homicide in this fashion could be problematic and uphold existing conceptions of the relationship between money and crime.

Log transformation for the response variable

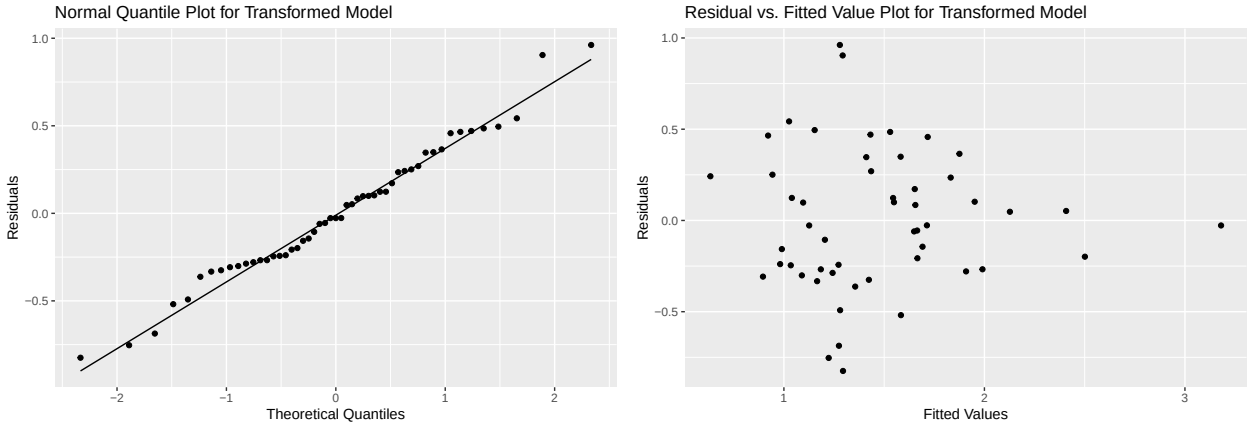
The normal quantile plot for the model that we derived from forward selection has points that align well with the predicted line, so the normality assumption is met. However, in the residual vs fitted value plot, we found that the other points roughly have a concaving-up curvature. Therefore, we decided to make a log transformation for the response variable.



After the log transformation, we found that though the curvature issue was solved, the points in the residual vs fitted value plot show larger residuals for smaller fitted values and smaller residuals for larger fitted values. Besides, from the t-test for slopes, many predictors have p-values that are much larger than $\alpha = 0.05$. This indicates that many predictors become insignificant after the log transformation, and issues with constant variance also occurred.

##

```
## Call:
## lm(formula = log(Homicide_rate) ~ Poverty_percentage + Unemployment_percentage +
##     Bachelor_degree_percentage + Median_household_income_1k +
##     Unemployment_percentage:Bachelor_degree_percentage + Median_household_income_1k:Bachelor_degree_
##     data = homicide_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82471 -0.26775 -0.02743  0.24650  0.96159
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       1.5562263  2.7976093
## Poverty_percentage                                0.1293544  0.0454792
## Unemployment_percentage                          -0.2967074  0.2750805
## Bachelor_degree_percentage                       -0.0821703  0.0629615
## Median_household_income_1k                       -0.0013268  0.0312850
## Unemployment_percentage:Bachelor_degree_percentage 0.0128666  0.0072183
## Bachelor_degree_percentage:Median_household_income_1k 0.0003789  0.0007347
##                                                   t value Pr(>|t|)
## (Intercept)                                         0.556  0.58084
## Poverty_percentage                                  2.844  0.00673 **
## Unemployment_percentage                            -1.079  0.28663
## Bachelor_degree_percentage                         -1.305  0.19865
## Median_household_income_1k                         -0.042  0.96636
## Unemployment_percentage:Bachelor_degree_percentage  1.783  0.08157 .
## Bachelor_degree_percentage:Median_household_income_1k  0.516  0.60862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4069 on 44 degrees of freedom
## Multiple R-squared:  0.5885, Adjusted R-squared:  0.5324
## F-statistic: 10.49 on 6 and 44 DF,  p-value: 3.328e-07
```



Considering the fact that the curvature in the original model is relatively acceptable compared to the issues brought by log transformation, we decided to keep with our original model.

6: Extra Credit

Sun et al. (2021) investigates a novel application of eigenvector spatial filtering (ESF) regression models, which are typically used in urban and regional studies, to local-scale (violent) crime data from New York City. They compare the performance of the "regular," non-spatial ordinary least squares (OLS) regression model to the ESF regression model and another spatial regression model, the spatial error model. The ESF model predicting the log of robbery rate is the following (with coefficients rounded to four decimal places): $ln(\widehat{\text{Robbery Rate}}) = \mathbf{E}_k\hat{\boldsymbol{\beta}}_E + 7.9432 - 0.0577 * \text{Female \%} - 0.0071 * \text{Under 18 \%} - 0.0120 * \text{Over 65 \%} + 0.0077 * \text{Non-white \%} + 0.0241 * \text{Unemployed \%} + 0.0254 * \text{Families Below Poverty \%} + 0.0485 * \text{Vacant Housing Units \%} + 0.0058 * \text{Homeowner Vacancy Rate} + 0.0612 * \text{Rental Vacancy Rate} - 0.0295 * \text{Married-Couple Family \%} - 0.0025 * \text{Population Speaking English Less Than "Very Well" \%} - 0.2871 * \text{Land Use Mix} - 0.0485 * \text{Road Density}$, where $\mathbf{E}_k$ is the set of k eigenvectors chosen from the total set of n eigenvectors from the "spectral decomposition of a transformed spatial weights matrix" via a stepwise procedure and $\boldsymbol{\beta}_E$ are the respective coefficients for each selected eigenvector.

While crime can be analyzed from a bio-psychological perspective, crime data is more accurately modeled by environmental and socioeconomic factors, which depend heavily on the specific geospatial location. Geospatial data, like crime data, tend to have residuals that are autocorrelated with each other, and spatial regression models account for this non-independence. The ESF model outperformed the spatial error and OLS models, offering the highest R^2 values and lowest AIC values. This paper relates to this project as it presents possible new predictors to model violent crime, offers a method to account for autocorrelated residuals (referencing an R package to apply the ESF model ("Spmoran: Moran Eigenvector-Based Spatial Regression Models" (2022))), and examines data from more a more micro-level in one specific urban city.

References

Bell, Brian, Rui Costa, and Stephen Machin. 2018. "Why Does Education Reduce Crime?" Center for Economic Policy Research. https://cepr.org/active/publications/discussion_papers/dp.php?dpno=13162.

Blau, Judith R., and Peter M. Blau. 1982. "The Cost of Inequality: Metropolitan Structure and Violent Crime." *American Sociological Review* 47 (1): 114–29. https://doi.org/10.2307/2095046.

Demombynes, Gabriel, and Berk Özler. 2005. "Crime and Local Inequality in South Africa." *Journal of Development Economics* 76 (2): 265–92. https://doi.org/10.1016/j.jdeveco.2003.12.015.

Hipp, John R. 2007. "Income Inequality, Race, and Place: Does the Distribution of Race and Class Within Neighborhoods Affect Crime Rates?" *Criminology* 45 (3). https://escholarship.org/uc/item/7kw8p7hw.

Horowitz, Juliana Menasce, Ruth Igielnik, and Rakesh Kochhar. 2020. "1. Trends in Income and Wealth Inequality." *Pew Research Center's Social & Demographic Trends Project*. https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/.

Kelly, Morgan. 2000. "Inequality and Crime." *The Review of Economics and Statistics* 82 (4): 530–39. https://www.jstor.org/stable/2646649.

Majumder, Maimuna. 2017. "Higher Rates Of Hate Crimes Are Tied To Income Inequality." *FiveThirtyEight*. https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.

McGinty, Jo Craven. 2018. "The FBI's Crime Data: What Happens When States Don't Fully Report." *Wall Street Journal*, October. https://www.wsj.com/articles/the-fbis-crime-data-what-happens-when-states-dont-fully-report-1539946801.

Meindl, James N., and Jonathan W. Ivy. 2017. "Mass Shootings: The Role of the Media in Promoting Generalized Imitation." *American Journal of Public Health* 107 (3): 368–70. https://doi.org/10.2105/AJPH.2016.303611.

"Spmoran: Moran Eigenvector-Based Spatial Regression Models." 2022. https://CRAN.R-project.org/package=spmoran.

Sun, Yeran, Shaohua Wang, Jing Xie, and Xuke Hu. 2021. "Modeling Local-Scale Violent Crime Rate: A Comparison of Eigenvector Spatial Filtering Models and Conventional Spatial Regression Models." *The Professional Geographer* 73 (2): 312–21. https://doi.org/10.1080/00330124.2020.1844574.