Community Factors on Violent Crime Rates

Abstract
Violent crime is a social issue that causes social, emotional, physical, and economic harm on both the community and individual levels. In this study, we aim to identify the most significant predictors of violent crime rates across communities. Specifically focussing on socio-economic and law enforcement data from the 1990s. We built a multiple linear regression model in R, using the stepwise procedure with the BIC criterion. Our final model utilizes a box-cox transformation and shows that demographic factors, housing conditions, racial factors, and police force factors are all significant predictors of violent crime.

I. Background and Significance

Violent crime is a large social issue that causes significant harm on both an individual and community level. On the individual level, it causes physical, emotional, and economic harm. On the community level, it harms the community's sense of trust, cohesion, and economic development. In our study, we aim to identify the most significant predictors of violent crime rates across communities. The results have direct policy implications, such as which populations should be targeted when deciding where and how the resources are distributed. Possible policy implications could be increasing funding for income support programs, job-searching programs and opportunities, and restorative justice programs that seek to improve the social, economic, and environmental factors.

II. Data

A. Data Source

The dataset of interest is Communities and Crime Unnormalized from the US Irvine Machine Learning Repository. The dataset combines socio-economic, law enforcement, and crime data from the 1990 Census, the 1990 U.S. LEMAS survey, and the 1995 FBI UCR. The data contains 18 target variables and 129 predictor variables. The target variables are violent crime rate(per 100k people), non-violent crime rate (per 100K people), and the crimes that make up violent and nonviolent crimes (murders, robberies, auto theft, etc.). For our study, the response variable will be violent crimes (Violentperpop).

B. Data Cleaning

First, unclear column identifiers in the dataset were renamed to their corresponding variable names based on the data dictionary. Then entries with missing values for the response variable (Violentperpop) were removed. To continue with data cleaning, variables were recategorized into correct data types. To deal with other missing values, variables missing over 10% of observations were removed. 22 law enforcement variables were removed as 83% of the observations were missing. Additionally, identifier variables that would not meaningfully contribute to the analysis were removed. After trimming, 10 variables had missing values. These missing values were imputed with the median as they all had a right-skewed distribution.

Before data analysis, further data cleaning is necessary to remove multicollinearity issues. First, a full model regression with all the variables was run to identify and remove perfectly linearly dependent variables. To detect multicollinearity issues, we used the variance inflation test (VIF) with a threshold of 10 to remove highly collinear variables. In this step, 53 variables were removed. Then we removed the variables representing subcategories of violent crimes, as they are already counted for in the response variable. Also, the categorial variable "state" was changed from states to regions, as a few states only had a few observations. The regions were: West, Midwest, South, and Northeast. The cleaned and trimmed data set has 1993 observations and 57 variables.

III. Methods and Results

A. Model Selection

Considering the large number of variables in the final dataset, we used automatic selection to find the best model instead of an all-subset comparison. Furthermore, we compared the models selected with the backward stepwise procedure using the AIC criterion with the model selected with the BIC criterion. We used the 10-fold cross-validation (CV) score to evaluate the prediction performance of each model. The 10-fold cross-validation score is obtained by randomly dividing the dataset into 10 folds. Then 9 folds are used to train the model and the remaining fold is used to test the model by comparing the predicted response from the trained model with the responses from the remaining fold. This process is repeated 10 times for each fold. This measure represents the sum of squared prediction errors. Additionally, taking the square root of this measure gives us a representation of the prediction error in the same unit as the response variable, which ranges from 0 to 4300. We labeled this the “Original-Units CV”.

Table 1. AIC Model and BIC Model Comparison

	AIC	BIC	Adjusted R^2	CV	Original-Units CV	# Predictors
AIC Model	**29269.29**	29465.19	**0.6366**	**143702.8**	**379.08**	34
BIC Model	29300.18	**29412.13**	0.628	145371.6	381.28	**19**

Although the AIC model has a higher adjusted R^2 value and a lower prediction error, the difference in values between the two models is small given the range in Violentperpop. The difference in prediction error is only 2.2. As the BIC model is more parsimonious, with 15 fewer predictor variables, we move forward with the model selected using the BIC criterion.

B. Model Refinement

A residual analysis will be conducted to evaluate the selected model. A residual plot (Figure 1) shows that the residuals violate the constant variance and linearity assumption necessary for multiple linear regression. The Q-Q plot (Figure 2) is heavy-tailed, showing that the normal residual assumption is also violated.

First, influential outliers were considered to improve the model. Using standardized residuals above 2, 59 outliers were identified. To determine if these outliers were influential, we calculated the cook’s distance. These values were compared to the 50th percentile of the F distribution (df1=10, df2=1974). Observations with a cook’s distance above the threshold (0.965), are considered influential. We found that there were no influential outliers (Figure 3).

To improve the model, we decided to conduct a transformation in the response variable (Violentperpop). We first considered a box-cox transformation, however there was one observation of 0 in the response variable. Although removing the observation limits the spectrum of the response variable and introduces bias, removing the observation would be reasonable as it is only one out of nearly 2000 observations. The box-cox transformation showed an optimal transformation of 0.26 (Figure 4). For ease of interpretation, the resulting transformation was $\text{Violentperpop}^{0.25}$. The residual plot (Figure 5) shows that the constant variance assumption is satisfied. While the linearity assumption is not perfectly satisfied, it is an improvement from the original model. In this Q-Q plot (Figure 6), the tails are much less heavy than in the original model, meaning the residuals are more normally distributed.

Another solution considered was an adjusted log transformation. The transformed response variable is log(1+Violentperpop). Again, the residual plot (Figure 7) shows a violation of the linearity and constant variance assumption. However, the Q-Q plot (Figure 8) shows improvements in residual normality, as only the lower tail is heavy.

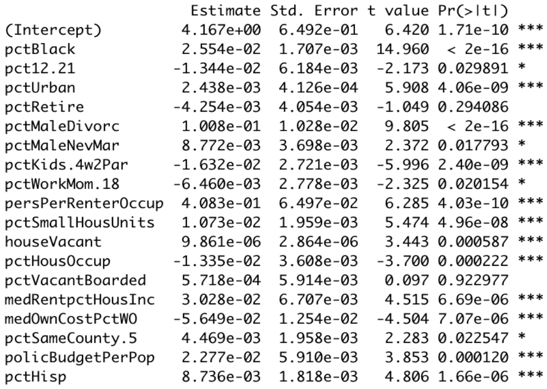
Table 2. Transformed Models Comparison

	AIC	BIC	Adjusted R^2	CV	Original-Units CV
BIC Model	29300.18	29412.13	0.628	145371.6	381.28
Box-Cox Model	4216.79	4328.73	**0.662**	0.485	0.235
Adjusted Log Model	**4196.14**	**4308.09**	0.612	0.467	0.981

C. Result

The box-cox model was chosen as it best satisfies the multiple linearity assumptions. Additionally, the box-cox model has the highest adjusted R^2 value and the lowest prediction error. For ease of interpretation, we compared the original-units CV. We calculated this by performing the inverse of the model transformation on the square-root CV for each respective model: $(\sqrt{CV})^4$ for the box-cox model and $(e^{\sqrt{CV}} - 1)$ for the adjusted log model. The box-cox model has the lowest original-units CV, meaning it has the lowest prediction error.

The final model includes the predictors listed on the right. (Variable descriptions can be found in the appendix). Out of the 19 predictors, 17 variables are statistically significant. The interpretation for the coefficients is as follows: for a 1 unit increase in the predictor, Violentperpop$^{0.25}$ increases by the coefficient value, holding all other variables constant.



```
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         4.167e+00  6.492e-01   6.420 1.71e-10 ***
pctBlack            2.554e-02  1.707e-03  14.960  < 2e-16 ***
pct12.21           -1.344e-02  6.184e-03  -2.173 0.029891 *  
pctUrban            2.438e-03  4.126e-04   5.908 4.06e-09 ***
pctRetire          -4.254e-03  4.054e-03  -1.049 0.294086    
pctMaleDivorc       1.008e-01  1.028e-02   9.805  < 2e-16 ***
pctMaleNevMar       8.772e-03  3.698e-03   2.372 0.017793 *  
pctKids.4w2Par     -1.632e-02  2.721e-03  -5.996 2.40e-09 ***
pctWorkMom.18      -6.460e-03  2.778e-03  -2.325 0.020154 *  
persPerRenterOccup  4.083e-01  6.497e-02   6.285 4.03e-10 ***
pctSmallHousUnits   1.073e-02  1.959e-03   5.474 4.96e-08 ***
houseVacant         9.861e-06  2.864e-06   3.443 0.000587 ***
pctHousOccup       -1.335e-02  3.608e-03  -3.700 0.000222 ***
pctVacantBoarded    5.718e-04  5.914e-03   0.097 0.922977    
medRentpctHousInc   3.028e-02  6.707e-03   4.515 6.69e-06 ***
medOwnCostPctWO    -5.649e-02  1.254e-02  -4.504 7.07e-06 ***
pctSameCounty.5     4.469e-03  1.958e-03   2.283 0.022547 *  
policBudgetPerPop   2.277e-02  5.910e-03   3.853 0.000120 ***
pctHisp             8.736e-03  1.818e-03   4.806 1.66e-06 ***
```

IV. Discussion and Other Considerations

A. Results

The model shows that significant predictors of violent crimes across communities were demographic factors, housing conditions, racial factors, and police force factors. We found that more disadvantaged socio-economic states, such as higher housing costs and higher divorce rates, are associated with a higher number of violent crimes. Additionally, we found that communities with a higher percentage of African American and Hispanic populations also have a positive association with violent crimes. Interestingly, the model shows that the police budget also has a positive association. These findings can help shape policies to support families, close economic gaps, and improve housing. With a specific focus on reducing violent crime, the results can guide policymakers in identifying subpopulations to direct attention and resources.

While discussing the findings, it is important to note that these findings reflect social and economic inequalities shaping these patterns, not direct causation. The broader implication of our research is that violent crime is part of a complex web of social and economic factors. It is important to interpret these results carefully to avoid reinforcing biases, stereotypes, or stigmatizing communities.

B. Limitations

The largest limitation of our research is the accuracy of the data. Part of the dataset is from the US Census, which is self-reporting. It is hard to verify the accuracy of this data, and it only represents individuals who decided to participate in the 1990 Census. Additionally, there could be an underreporting of violent crimes, especially rapes. Due to the stigmatization of rape, victims might not report the crime for fear of backlash.

Another limitation is how the dataset was created. The crime reporting data from the FBI UCR is from 1995 while the socio-economic and law enforcement data are from 1990. This mismatch of years assumes that crime reports in 1990 and 1995 are equivalent. While investigating, we found that the earliest FBI UCR was from 1995 and is the closest estimate for crime reporting for 1990. Additionally, as this data is outdated and not generalizable to the present, it should not be used for current policy making decisions.

C. Future Research

To better understand the causes of violent crime, we need to include social science research to understand underlying factors and how they interact to shape violent crime rates. Additionally, other factors could be considered by finding other data sources. Law enforcement data was a topic of interest but was unfortunately removed due to large missingness. To have results that could be applied in policy, more recent data should be used. Additionally, a potential topic of future research is the direct impact of policy on violent crime rates. Specifically focusing on the types of policy and targets of policy.

References

Redmond, M. (2009). Communities and Crime Unnormalized [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC8X.

Appendix

Figure 1. BIC Model - Residual Plot

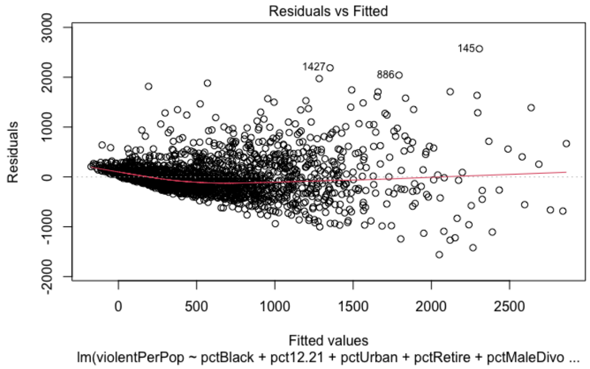


Figure 2. BIC Model - Q-Q Plot

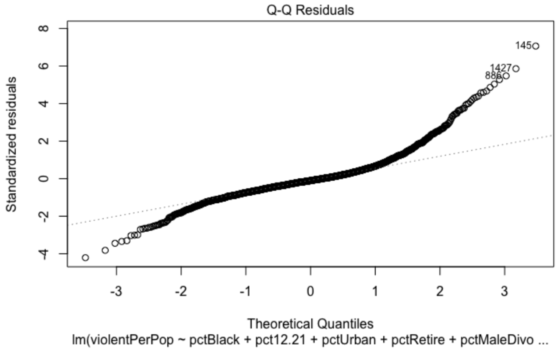


Figure 3. Outlier Analysis - Cook's Distance

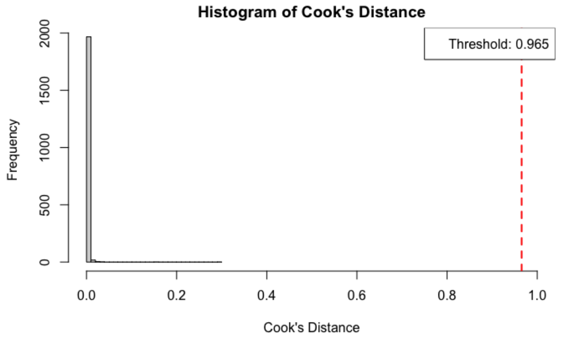


Figure 4. Optimal Box-Cox Transformation

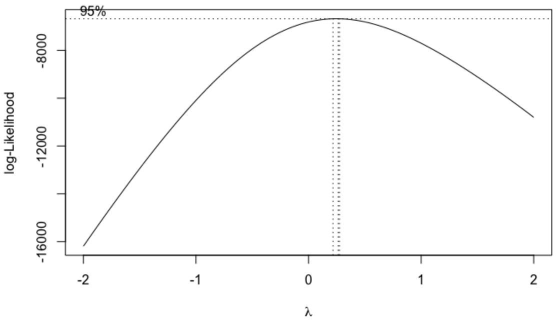


Figure 5. Box-Cox Model - Residual Plot

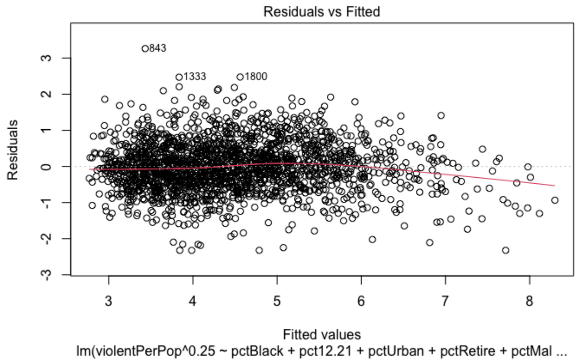


Figure 6. Box-Cox Model - Q-Q Plot

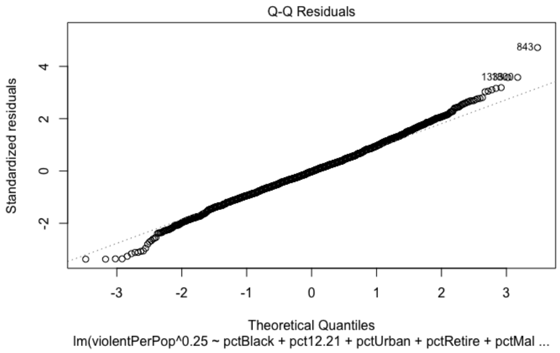


Figure 7. Adjusted Log Model - Residual Plot

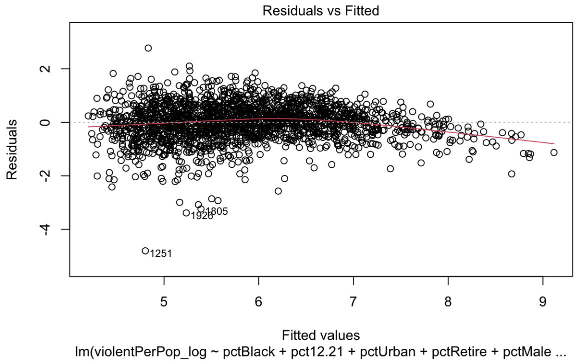


Figure 8. Adjusted Log Model - Q-Q Plot

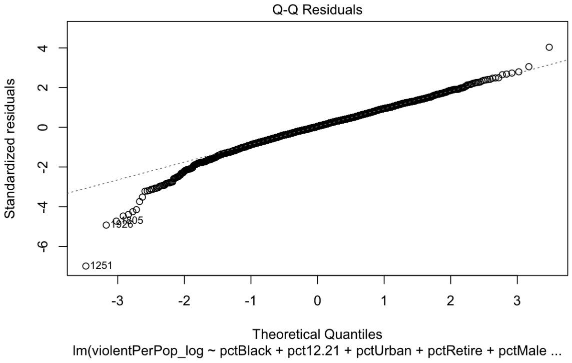


Table 3. Description of Model Predictors

Variable	**Description**
pctBlack	Percentage of the population that is African American
pct12.21	Percentage of the population that is 12-21 years old
pctUrban	Percentage of people living in areas classified as urban
pctRetire	Percentage of households with retirement income in 1989
pctMaleDivorc	Percentage of males who are divorced
pctMaleNevMar	Percentage of males who have never married
pctKids.4w2Par	Percent of kids 4 and under in two-parent households
pctWorkMom.18	Percentage of moms of kids under 18 in the labor force
persPerRenterOccup	Mean persons per rental household
pctSmallHousUnits	Percent of housing units with less than 3 bedrooms
houseVacant	Number of vacant households
pctHousOccup	Percent of housing occupied
pctVacantBoarded	Percent of vacant housing that is boarded up
medRentpctHousInc	Median gross rent as a percentage of household income
medOwnCostPctWO	Median owners cost as a percentage of household income (for owners without a mortgage)
pctSameCounty.5	Percent of people living in the same city as in 1985 (5 years before)
policBudgetPerPop	Police operating budget per population
pctHisp	Percentage of the population that is of Hispanic heritage