Predicting Violent Crime Rates in the United States

Briana Nguyen

2024-03-15

Introduction

Although the general trend of violent crime rate in California has decreased over the past decade, there has been a gradual increase from 2014 of about 391 violent crimes per 100,000 residents to 495 violent crime per 100,000 residents in 2022, resulting in an overall 26.4% increase. These violent crimes mainly consisted of a large percentage of aggravated assaults followed by robberies, rapes and homicides. In order to implement more effective policies to mediate and combat violent crimes, it is crucial to understand its causes. This study aimed to explore different demographic features and crime rates in various communities within not just California, but the entire United States. A Kaggle dataset from the year 2022 was utilized to develop a multiple linear regression (MLR) model to predict the number of violent crimes per population. This dataset contained 146 different variables and 2018 observations (n = 2018), but this study mainly focused on the following variables:

Table 1. Descriptive statistics for model variables

Variable	mean	sd	var	min	median	max
medIncome	33901	13409	179794509	12908	31264	123625
RentMedian	432.86	174.65	30504	139	399.50	1001
PctUnemployed	6.02	2.71	7.34	1.32	5.47	23.83
PctNotHSGrad	22.68	11.07	122.58	2.09	21.56	73.66
PctPopUnderPov	11.68	8.48	71.85	0.64	9.48	48.82
PctFam2Par	74.01	10.30	106.08	32.24	74.89	93.60
PctKidsBornNeverMar	3.12	3.06	9.38	0.03	2.07	24.19
TotalPctDiv	10.87	3.01	9.09	2.83	11.04	19.11
nonViolPerPop	4946	2786	7759268	116.79	4481.38	27119.76
ViolentCrimesPerPop	584.24	608.36	370103	6.64	370.94	4877

Predictor variables: percent of kids born with non-married parents (PctKidsBornNeverMar), median income (medIncome), percent of population under poverty (PctPopUnderPov), percent of divorce (TotalPctDiv), median rent (RentMedian), percent of 2 parent families (PctFam2Par), percent of non-high school graduates (PctNotHSGrad), percent of unemployed (PctUnemployed), number of non-violent crimes per population (nonViolPerPop)

Outcome variable: number of violent crimes per population (ViolentCrimesPerPop)

Methods

Data Cleaning

The data was filtered based on the features of interest and cleaned by removing incomplete cases.

Variable Correlation

A correlation matrix was used to evaluate the relationship between pairs of variables, specifically between each predictor variable and the outcome variable with PctKidsBornNeverMar, PctFam2Par and nonViolPerPop

having the highest correlations of 0.74, -0.70 and 0.68 to the outcome variable. It was also observed there were some predictor variables that appeared to have high correlations with each other, specifically, medIncome and RentMedian (0.85) and PctKidsBornNeverMar and PctFam2Par (-0.84), which implied that there were also multicollinearities that must be addressed.

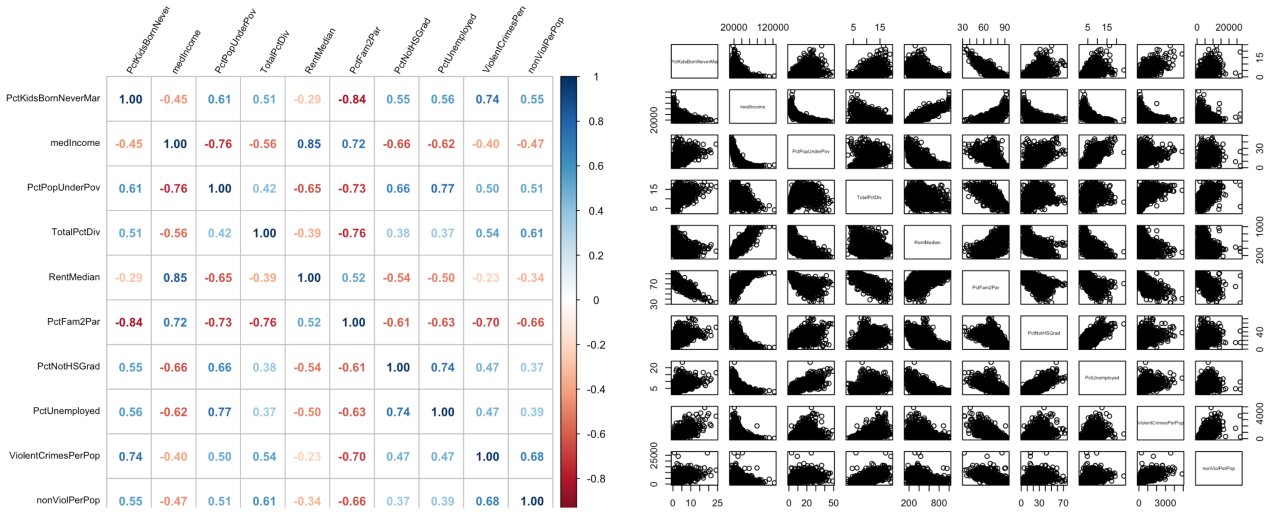

Figure 1. Pairwise correlation matrix

Initial Linear Model

Table 2. Summary of MLR model with all 9 variables

Coefficients	Estimate	Std. Error	t-val	p-val
(Intercept)	-8.584e+02	2.524e+02	-3.401	0.000685 ***
PctKidsBornNeverMar	9.675e+01	5.912e+00	16.365	< 2e-16 ***
medIncome	-3.496e-04	1.650e-03	-0.212	0.832244
PctPopUnderPov	-1.688e+00	2.045e+00	-0.825	0.409261
TotalPctDiv	2.032e+01	4.942e+00	4.111	4.11e-05 ***
RentMedian	4.221e-01	9.302e-02	4.538	6.04e-06 ***
PctFam2Par	2.529e+00	2.731e+00	0.926	0.354486
PctNotHSGrad	4.363e+00	1.214e+00	3.595	0.000333 ***
PctUnemployed	1.364e+01	5.381e+00	2.535	0.011327 *
nonViolPerPop	8.105e-02	4.044e-03	20.041	< 2e-16 ***
Residual standard error: 353.5	1890 DF			
Multiple R-squared: 0.664	Adjusted R-squared: 0.6624			
F-statistic: 415.1	9 and 1890 DF	p-value: < 2.2e-16		

A MLR model was created with all 9 predictor variables with an adjusted R-squared value of 0.6624, a significant p-value for the F-statistic and significant p-values for multiple predictor variables. This indicated that the linear regression model can be suitable for this data, but must be further refined with variable selection. Furthermore, the standardized residual plots and model diagnostic plots showed normality and constant variance assumption violations.

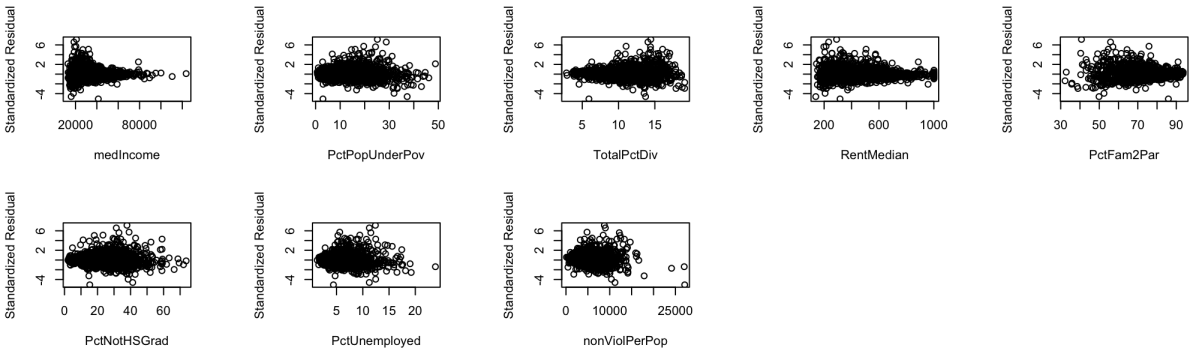

Figure 2. Standardized residual for each predictor variable

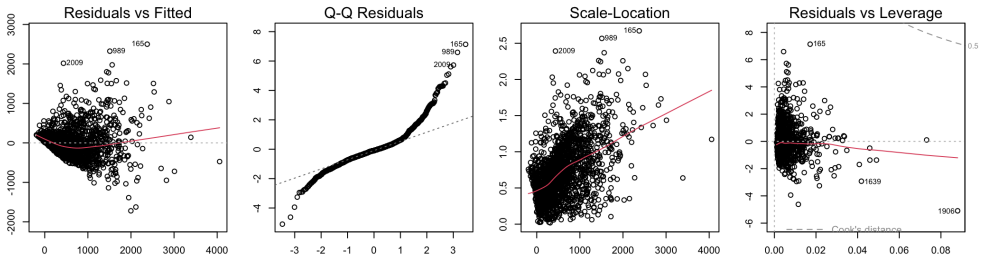

Figure 3. Initial model diagnostic plots

Box-Cox Transformation

A box-cox transformation was conducted on the skewed data with a calculated lambda value of around 0.2626, which showed major improvements in the normality shown in the Q-Q plot and the homoscedasticity shown in the Scale-Location plot. However, there appears to be a downward trend for the residual plots as a result of outliers.

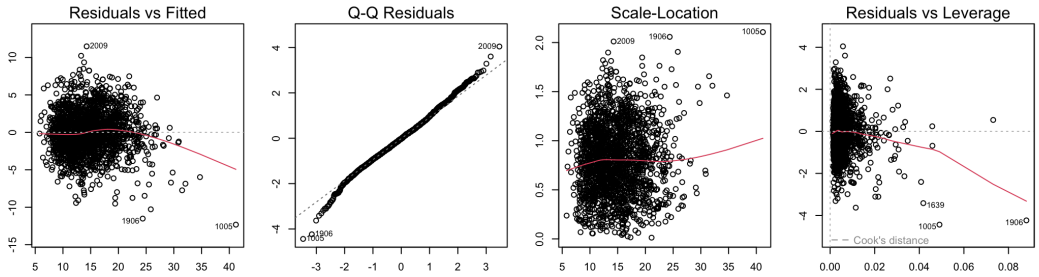

Figure 4. Box-cox transformed model diagnostic plots

Multi-collinearity and Variable Selection

Table 3. VIF values for predictor variables

Variable	VIF < 5	Variable	VIF > 5
nonViolPerPop	1.929086	medIncome	7.440398
PctNotHSGrad	2.744866	PctFam2Par	12.023016
PctUnemployed	3.229985		
TotalPctDiv	3.372803		
RentMedian	4.011729		
PctPopUnderPov	4.568561		
PctKidsBornNeverMar	4.981763		

Multi-collinearity was evaluated by calculating the VIF for each predictor variable and variables with a VIF greater than 5 was considered to have extreme multicollinearity. In addition, added variable plots were used to further visualize the marginal influence of predictor variables with high VIFs. A horizontal line indicates that the variable is highly correlated with another variable and therefore, does not have much marginal impact on the outcome and should be removed. MedIncome and PctFam2Par had extreme VIFs of around 7.44 and 12.02 and to address this, a backward stepwise regression with the Akaike Information Criterion (AIC) to measure the goodness of fit was used. This eliminated the variables, PctPopUnderPov and PctFam2Par.

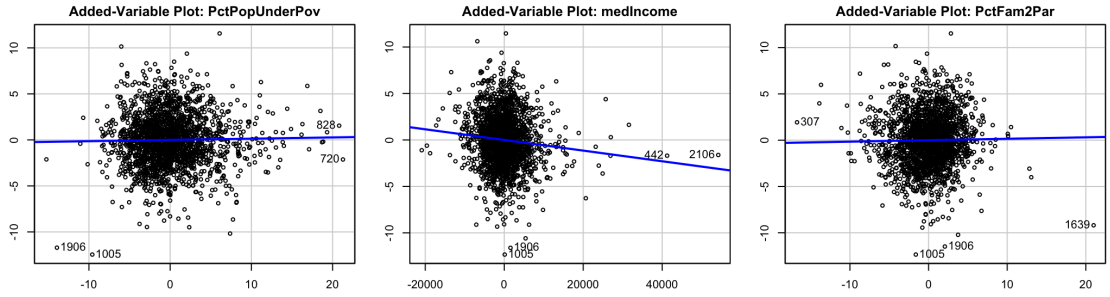

Figure 5. Added-Variable plots for PctPopUnderPov, medIncome and PctFam2Par

However, the variable medIncome still had a high VIF of 5.964. In order to determine if medIncome should be included in the model, a partial F-test was implemented.

Table 4. Partial F-test with and without the medIncome variable

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Full Model	1893	15528				
Reduced Model	1892	15343	1	185.25	22.843	1.893e-06 ***

The significant p-value from the f-test revealed that there was evidence against the reduced model in favor of the full model and medIncome should be included in the final model.

Outliers

Outliers were then removed by identifying bad leverage points and by evaluating standardized residual and cook's distance. Leverage points are points whose hat values exceeded $\frac{2(p+1)}{n}$. It is considered a bad leverage and an outlier if the point had a standardized residual value outside of the range (-2, 2) and if the point had a cook's distance greater than $\frac{4}{n-2}$.

Results and Limitations

Table 5. Summary of MLR for the final model

Coefficients	Estimate	Std. Error	t-val	p-val
(Intercept)	2.941e+00	5.082e-01	5.787	8.44e-09 ***
PctKidsBornNeverMar	5.650e-01	2.784e-02	20.291	< 2e-16 ***
medIncome	-4.962e-05	1.054e-05	-4.706	2.72e-06 ***
TotalPctDiv	3.013e-01	2.656e-02	11.343	< 2e-16 ***
RentMedian	6.379e-03	6.383e-04	9.995	< 2e-16 ***
PctNotHSGrad	5.031e-02	8.410e-03	5.982	2.67e-09 ***
PctUnemployed	1.015e-01	3.371e-02	3.011	0.00264 **
nonViolPerPop	7.657e-04	3.074e-05	24.912	< 2e-16 ***
Residual standard error: 2.322	1762 DF			
Multiple R-squared: 0.7713	Adjusted R-squared: 0.7704			
F-statistic: 849.1	7 and 1762 DF	p-value: < 2.2e-16		

The summary shows that the adjusted R squared is 0.77, the p-value for the f-statistic is significant and all the predictor variables have a significant p-value. The model was able to explain about 76.77% of the variance in the outcome variable and the f-test result indicates that there was an association between the outcome and at least one of the predictor variables. Furthermore, the standardized residual plots for each variable does show a slight pattern for a few variables, but this may be due to a small sample size since most of the points showed no general pattern. The model residual plot also does not show a distinct pattern with a relatively straight horizontal line centered at around 0, the Q-Q plot shows a relatively constant line with light tails, the scale-location plot does not show a pattern and the residuals vs leverage plot does not have any extreme outliers. As a result, the linearity, normality homoscedasticity assumptions have been fulfilled. However, the model is still far from ideal and the sample size may have also resulted in a slight down-ward pattern in the residual plots and light tails in the Q-Q plot from more extreme data. Further studies should aim to incorporate a larger sample size in order to make the model more generalizable to real world data and additional demographic features should be explored including gender and percentages of incarcernated family members.

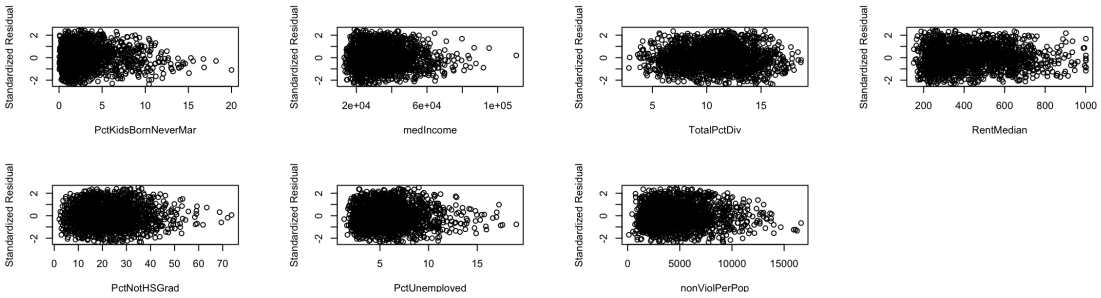

Figure 6. Standardized residual for each predictor variable

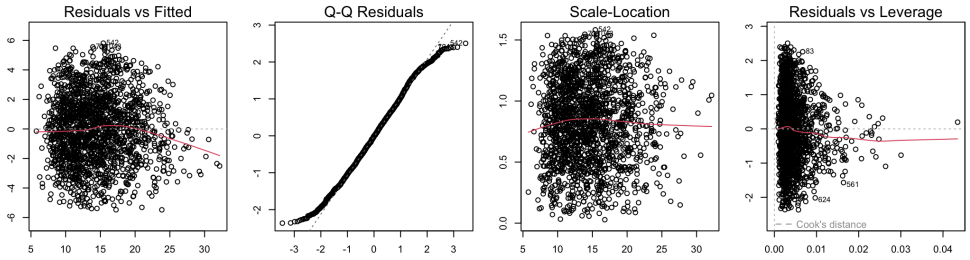

Figure 7. Final model diagnostic plots

Discussion

Overall, the MLR model was appropriate in predicting the number of violent crimes. From the feature weights, it can be observed that only median income has a negative linear association with violent crime, which could be due to a better quality of life in wealthier communities or more funds allocated towards the police. On the other hand, the percentage of divorce and non-married parents showed a positive association with violent crime, revealing the importance of family dynamic. This aligns with many personal testimonies from criminals, who often grew up in unstable families without a father figure or grew up with very young inexperienced parents. As a result, these children grew up without proper discipline. Furthermore, more violent crimes tend to be committed in communities with less education (PctNotHSGrad) and less employment (PctUnemployed). However, it is interesting to note that rent also has a positive association with violent crime, which could be due to a lower median income making it difficult to purchase a home. In conclusion, it can be seen that crime can rise from the environment that an individual is raised in and therefore, it is crucial to ensure that education, employment and family counseling opportunities are provided. Crime plagues many countries throughout the world and although it cannot be completely eliminated, it can be alleviated through the implementation of effective public policy focusing on human welfare.