Predicting Violent Crimes in US Communities

Team 7

Ryan Jaeger and Tatum Bair

Fall 2018 Semester

Abstract

While violent crimes are a social issue that plagues all parts of the country, it is well established that some communities suffer from higher crime rates than others. The goal of this analysis is to understand how data surrounding socioeconomic, demographic, and crime factors can be used to predict which communities are most susceptible to violent crime. Using data on almost 2,000 US communities from the 1990s provided by the UC Irvine Machine Learning Repository, we develop a multiple linear regression to predict the number of violent crimes per capita. To best fit the data, we employ a number of regression techniques, including normalization, weighted least squares, and backward elimination. Our final model is able to explain 66% of the variability in violent crimes per capita, a surprisingly high amount for real socioeconomic data. To better explore which predictors are related to number of violent crimes per capita, we employ a permutation test as a nonparametric method of conducting inference, and we find that 6 of the 10 considered predictors are significant. Future analysis on this topic could include testing many of the other predictors available in this dataset or testing the same model on an updated dataset.

Introduction

This dataset combines socio-economic data, law enforcement data, and crime data from communities across the United States in the 1990s. It was assembled by Michael Redmond of La Salle University by joining data from the 1990 US Census, 1995 FBI Uniform Crime Report, and the 1990 Law Enforcement Management and Administrative Statistics Survey. Data was collected from 1,994 communities. The 15 variables in consideration were: Population, Percent of Population that is Caucasian, Age Percentage Between 16 and 24, Median Family Income, Percent of Population Under Poverty, Percent Unemployed, Percent of People Who Do Not Speak English Well, Percent of People in Dense Housing, Number of Homeless People, Percent Born in the Same State, Total Requests for Police, Police Per Population, Percent of Police that are Caucasian, Number of Different Kinds of Drugs Seized, Population Density, and Police Operating Budget. We selected Violent Crimes Per Population as our response variable.

The research questions in consideration are:

1) How can we predict the number of violent crimes using data from across the United States?
2) Which of these variables is strongly related with the number of violent crimes?

These questions are important in our analysis as we are trying to determine if it is possible to use certain variables to predict crime in communities around the United States, and figure out which variables should be taken into consideration.

Exploratory Data Analysis

As described above, we are looking at data regarding socioeconomic, crime, and law enforcement factors. Because many of these predictors are on different scales, we started by using the normalized dataset provided by the UCI Machine Learning Repository. To ensure they are all on the same scaled, each predictor is standardized, taking on values between 0 and 1.

The basic summary statistics for each predictor of interest can be found in Table 1 below. Because each of the underlined predictors are missing in 1,675 observations, we decided to remove these variables from our analysis. Our final dataset had 1,994 observations of 12 variables.

Table 1. Summary Statistics of Predictors

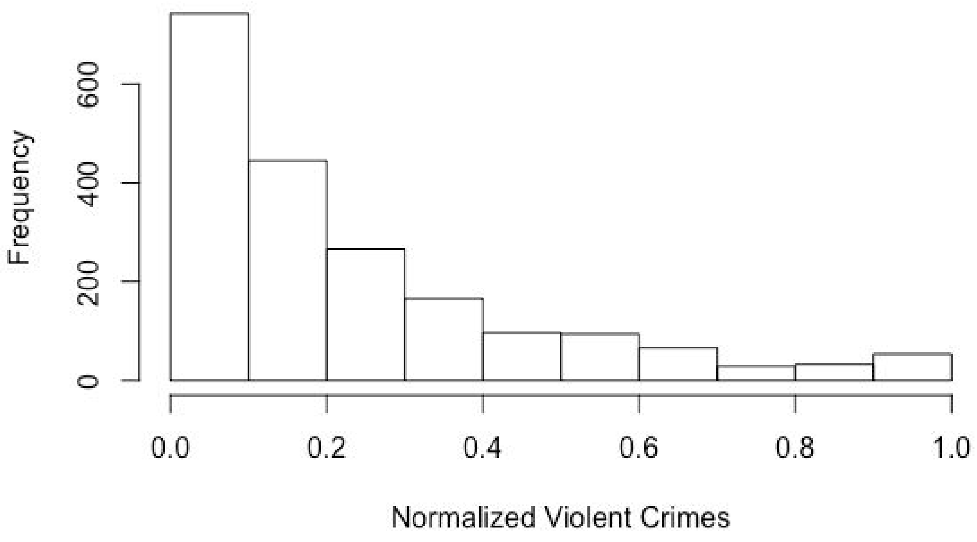
Variable	Min	Max	Mean	SD	Correl	Median	Mode	Missing
Pop	0	1	0.06	0.13	0.37	0.02	0.01	0
RacePctWhite	0	1	0.75	0.24	-0.68	0.85	0.98	0
AgePct16-24	0	1	0.34	0.17	0.1	0.29	0.29	0
MedFamIncome	0	1	0.38	0.2	-0.44	0.33	0.25	0
PctPopUnderPoverty	0	1	0.3	0.23	0.52	0.25	0.08	0
PctUnemployed	0	1	0.36	0.2	0.5	0.32	0.24	0
PctNotSpeakEngWell	0	1	0.15	0.22	0.3	0.06	0.03	0
PctPersDenseHous	0	1	0.19	0.21	0.45	0.11	0.06	0
NumStreet	0	1	0.02	0.1	0.34	0	0	0
PctBornSameState	0	1	0.61	0.2	-0.08	0.63	0.78	0
PolicePerPop	0	1	0.22	0.16	0.15	0.18	0.2	1675
PctPoliceWhite	0	1	0.73	0.22	-0.44	0.78	0.72	1675
NumDiffDrugsSeiz	0	1	0.56	0.2	0.13	0.57	0.57	1675
PopDensity	0	1	0.23	0.2	0.28	0.17	0.09	0
PolicOperBudget	0	1	0.08	0.14	0.34	0.03	0.02	1675
ViolentPerPop	0	1	0.24	0.23	1	0.15	0.03	0

Next, we considered pairwise correlations between all of the remaining predictors (shown in Table 2), and of all pairs, we only found high correlations between Number of Homeless Individuals and Population, Percent Population Under Poverty and Percent Unemployed, and Percent Speak English Poorly and Percent Persons living in Dense Housing. While the first two of these are intuitive, the relationship between english aptitude and dense housing is less obvious, but potentially due to immigrants living in dense housing.

Table 2. Pairwise Correlations

	Crimes	Pop	%White	16-24	MedInc	%UnderPov	%Unemploy	%SpeakBadEng	%Dense	NumStreet	%SameState
Pop	-0.05										
%White	0.02	-0.17									
16-24	-0.02	0.02	-0.12								
MedInc	0.02	-0.05	0.33	-0.21							
%UnderPov	-0.01	0.09	-0.53	0.46	-0.73						
%Unemploy	0.02	0.08	-0.52	0.16	-0.64	0.77					
%SpeakBadEng	0.04	0.12	-0.43	0.05	-0.15	0.28	0.43				
%Dense	0.01	0.11	-0.59	0.13	-0.29	0.41	0.5	0.89			
NumStreet	-0.04	0.92	-0.1	0	-0.02	0.05	0.05	0.09	0.08		
%SameState	-0.03	-0.07	0.11	-0.03	-0.24	0.15	0.15	-0.28	-0.25	-0.05	
PopDens	-0.02	0.21	-0.32	0.03	-0.03	0.07	0.17	0.55	0.43	0.2	-0.22

Figure 1: Violent Crimes Per Capita



While a full description of each of the predictors is beyond the scope of this report, we have included Figure 1 above which offers a histogram of Violent Crimes Per Capita, our response variable. While the data has been standardized, it is obvious that it is not distributed normally, so we will keep this in mind as we consider the normality of errors assumption of linear regression later in our analysis.

Methodology & Results

We organize this section by what analysis can be used to answer each of our two research questions. We begin with our first research objective: constructing a linear model to predict violent crimes in communities based on the data at hand.

Our first step in this analysis was recognizing the issues that may arise by using the raw, non-standardized data. When we first fitted the model on the original data, we had very small estimates for our predictors, on the order of 10^{-4} in some cases, and a very low R^2 value of 0.00965, meaning that less than 1% of the variability in violent crimes can be explained with this model. This is a very weak model, and we realized that this might be due to the differences in scale for our predictors, so we made the choice to use standardized data. Luckily, the UCI Machine Learning Repository offered a standardized data set ready to go, so the remainder of our analysis uses that. The initial variables we considered were: Percent of Population that is Caucasian, Age Percentage Between 16 and 24, Median Family Income, Percent of Population Under Poverty, Percent Unemployed, Percent of People Who Do Not Speak English Well, Percent of People in Dense Housing, Number of Homeless People, Percent Born in the Same State, and Population Density. Population is not included since we employ it in a different capacity later.

Next, in order to choose which predictors we should include in our analysis, we decided to employ backward elimination with $\alpha = 0.01$. Because we wanted to consider how many different factors impacted our regression model, backward elimination was a good choice because it could include many predictors, but only those that were significant at that level. Table 3 below describes each our steps in the backward elimination process.

Table 3: Backward Elimination

Step	Predictor Removed	P-Value
1	PctUnemployed	0.4344
2	PctNotSpeakEnglWell	0.0325

At this point, we had all the remaining predictors significant by t-tests at the 0.01 significance level, so the next factor we had to consider was collinearity. Because many of these variables might be related, we run the risk of suffering from collinearity, which would make our estimation and inference faulty. To understand this, we looked at variance inflation factors for each of the remaining predictors, shown below in Table 4.

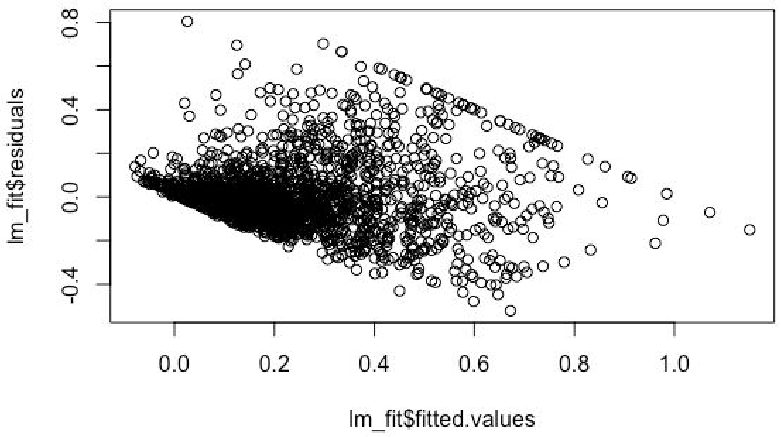
Table 4: Variance Inflation Factors

Original Model		Removing PctPopUnderPov	
Predictor	VIF	Predictor	VIF
racePctWhite	2.103037	racePctWhite	1.811753
AgePct16t24	1.297132	AgePct16t24	1.085839
medFamInc	3.15281	medFamInc	1.388959
PctPopUnderPov	4.724481	PersDenseHous	1.861954
PersDenseHous	1.894195	NumStreet	1.359382
NumStreet	1.364582	PctBornSameState	1.414625
PctBornSameState	1.493684	PopDens	1.458664
PopDens	1.459205		

The left hand side of the table shows the VIF values for the model after backward elimination. From this, we can see that the VIF for PctPopUnderPov is greater than 4, which is a typically used cut off, so we decide to remove this from our model. This also makes intuitive sense because the economic data captured by PctPopUnderPov could also be understood from median family income. After removal, we can see the updated VIF values in the right hand side of the table, and all these updated VIFs are less than 4, so we see no evidence of collinearity as an issue.

To understand how well this model is fitting our data, we examined a residuals vs. fitted values plot, shown in figure 2.

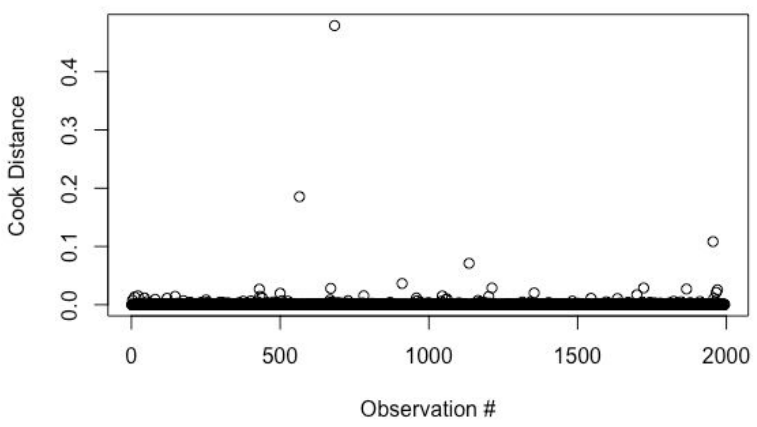
Figure 2: Residuals vs. Fitted Values



Based on the plot above, we see that there are issues with the equivariance assumption of our linear model. To handle this, we employ the population variable as a weight for weighted least squares. Since the response variable is per capita violent crimes, we can trust the values that correspond to communities with higher populations, therefore it is reasonable to use this as a weight. When we weight our model by population, our variance of errors decreases significantly as shown in figure 4 later.

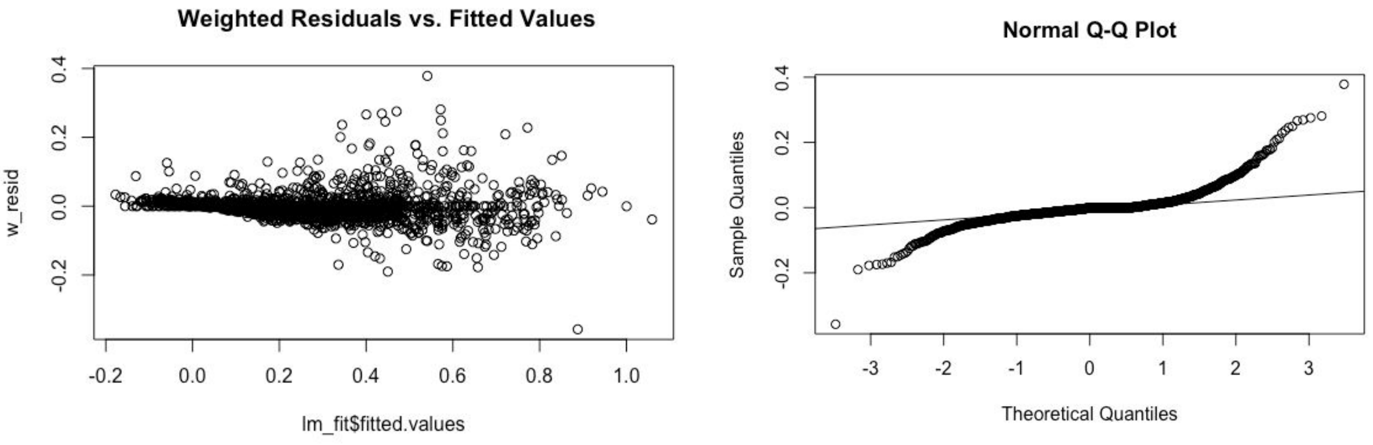
Now that the equivariance assumption is satisfied, the next issue that we had to deal with was extreme values. When looking at a plot of Cook's Distance (figure 3), there is one point, corresponding to Philadelphia, significantly different than the others and therefore influential on the model, so we decide to remove this from the analysis.

Figure 3: Cook Distance



Once removed, we have reached our final model for this analysis; this model uses 7 variables as predictors, and population as a weight, with all variables being significant at the 0.01 level. Before we interpret our model, let us take a final look at model diagnostics to ensure that the assumptions of linear regression are satisfied.

Figure 4: Model Diagnostic Plots (Weighted Residuals/Q-Q Plot)


Weighted Residuals vs. Fitted Values

Normal Q-Q Plot

Based on the weighted residuals plot above, we can see that the linearity and equivariance assumptions are satisfied with this model. We do not know the specifics of data collection, but we must assume independence. The normal Q-Q plot deviates significantly from the line given, so this means that our errors are not normal; a Shapiro-Wilk test confirms this (p-value < 2.2e-16). However, since we have a large dataset (n=1993), we can rely on the Central Limit Theorem and not worry about this violation of normality.

Given that our model satisfies the conditions of regression, the estimates for the predictors are shown in Table 5.

Table 5: Estimates for Predictors

Predictor	Estimate
Intercept	1.06894
racePctWhite	-0.59859
agePct16t24	-0.1221
medFamInc	-0.55453
PctPersDenseHous	-0.15816
NumStreet	0.25174
PctBornSameState	-0.14292
PopDense	0.06102

Because we are using normalized data, we cannot make meaningful interpretations about the value of each of our estimates, but rather we can just use the sign. For example, since population density has a positive estimate, we know that higher population density leads to higher violent crimes per capita, but we cannot make further conclusions. We can make the same kind of binary increase or decrease statement for each of these predictors. Our R^2 value is 0.6638, meaning that 66.38% of the variability in violent crimes per capita can be explained by our model. We are very satisfied with this value, since it means our model is relatively good given the difficulty of working with real socioeconomic data. With this, we have answered our first research question of generating a linear model.

Our next research question concerns which predictors are significantly related to violent crimes per capita. While we can use the results of our linear model, our inference may not be perfect since our normality assumption is not satisfied, but rather estimated by Central Limit Theorem. Instead, we can test the significance of each of our predictors in context of the full model using a permutation test, a nonparametric method of inference that generates our sampling distribution by shuffling the observed data.

The results of our permutation test are shown in Table 6 below:

Table 6: Permutation Test Results (significant predictors in bold)

Predictor	P-Value
racePctWhite	0
agePct16t24	0.0156
medFamInc	0.0029
PctUnderPoverty	0.0247
PctUnemployed	0.9124
PctNotSpeakEnglWell	0.5732
PctPersDenseHous	0.2951
NumStreet	0
PctBornSameState	0.0026
PopDense	0.1422

At the α = 0.05 significance level, we can see that 6 of the 10 predictors considered are related to violent crimes per capita by the permutation test. Because this inference does not rely on the assumption of normality, we can trust these results more than the significance that might come from the R output of a linear model. With this, we have successfully answered our second and final research question.

Conclusion

In summary, after an series of steps to curate our linear regression, we have developed a model to predict violent crimes per capita in US communities that performs relatively well ($R^2 =$ 0.6638). Furthermore, we conducted a permutation test to conduct inference on the significance of 10 selected predictors and found 6 of them to be significantly related to violent crimes per capita. While we had to limit our analysis to exclude crime and law enforcement data, our model could potentially be employed by a government body to predict the number of violent crimes in a given community and inform their action to mitigate that crime. Future analysis for this project might include expanding to some of the other predictors used in the dataset, working with the original (non-standardized data), or updating the model with data from the 2010s rather than the 1990s.

Team Member Contributions

Ryan found the dataset and wrote the description. Tatum wrote and formatted the short report, including the research questions and interpreted the exploratory data analysis. Ryan fitted the model and performed model selection. Finally, Ryan and Tatum both prepared figures and tables for the final report. All members contributed to the in-class presentation.

References

Communities and Crime Normalized Data Set from the UC Irvine Machine Learning Repository - http://archive.ics.uci.edu/ml/datasets/communities+and+crime