An Analysis of Factors Contributing to Violent Crime

Mean_Girls: Woongkyu Ham, Hyunji Kim, Aaron Wan, Qianni Zhang, Yeeun Yang

Table of Contents

Abstract
Methods
<Graph1 - Violent Crimes By state on US Map> (Linked)
<Graph2 - correlation between two explanatory variables ("NumUnderPov" and "NumInShelters")>
<Graph 3 - original response variable>
<Graph 4 - logit response variable>
Results
<Test Result>
<Coefficients Chart>
Discussion
Conclusion
References

Abstract

The purpose of this research is to identify the main factors affecting violent crimes. The analysis presented provides some statistical inferences regarding correlations between the rate of violent crimes and many socioeconomic factors in the United States. These correlations are made using a compiled government data set, which was analyzed using linear and logistic regression models and visualized with graphic tools in both R and JavaScript. As a result of logistic regression of the data, both strong positive and strong negative correlations with violent crimes were found. Generally, the most prominent factors were discovered in economic, domestic, and cultural categories. In particular, percentage of illegal immigration as well as number below poverty had a strong, positive correlation with violent crime, while the percentage of household kids with two parents had a strong, negative correlation. These kinds of results may be useful for governments facing rampant issues with violent crime.

Methods

We used "Communities and Crime Data Set." This data contains 1827 Communities within the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.[1] The response variable is violent crimes per 100k population, and there are 122 explanatory variables. However, we removed 21 explanatory variables based on the fact that each of them has more than 1000 missing values. We also replaced one missing value of the variable named OtherPerCap with the mean of this variable. The number of the explanatory variables is thus reduced to 101 after the data clearing.

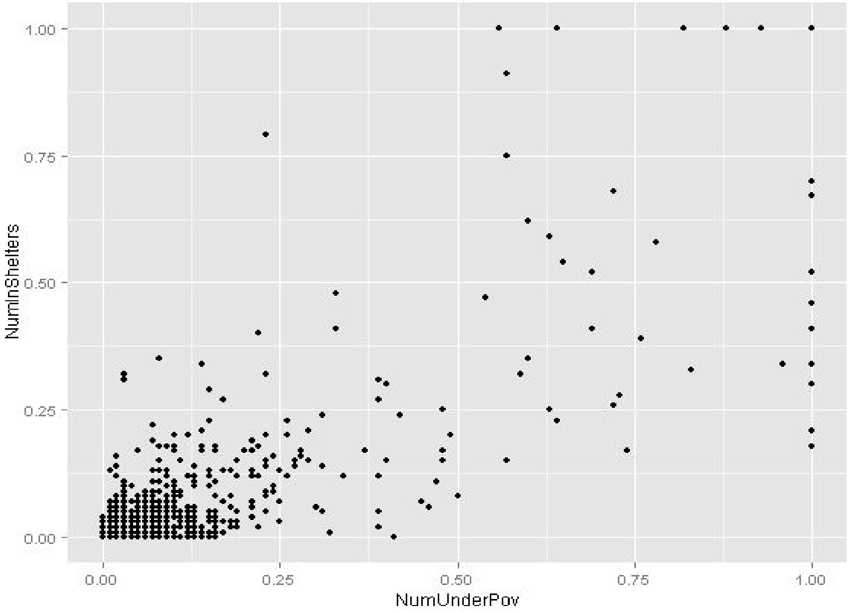
<Graph1 - Violent Crimes By state on US Map> (Linked)

Violent Crimes By State

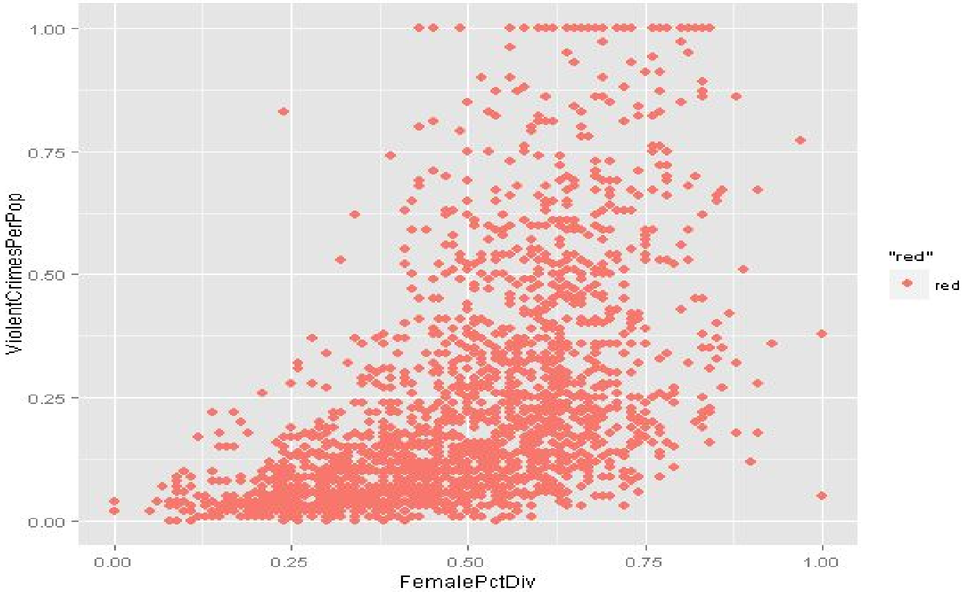
Before handling the data, we drew a plot of response variable on a map, by state. Firstly, We applied Ordinary Linear Model. However, there are some non-linearity relationships between explanatory variables and the response variable. In addition, multicollinearity exists, i.e. explanatory variables themselves are highly correlated. Graph1 shows a multicollinearity between "NumUnderPov" and "NumInShelters."

[1] https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

<Graph2 - correlation between two explanatory variables ("NumUnderPov" and "NumInShelters")>

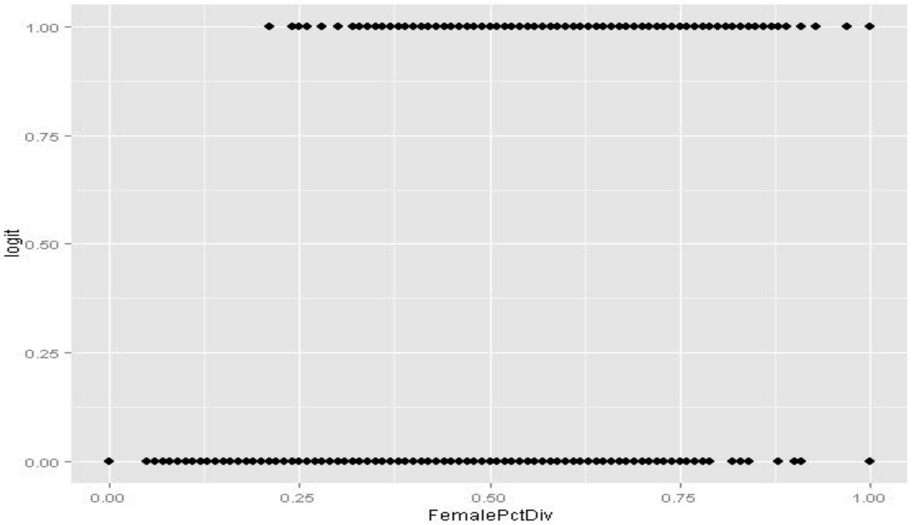

1.00
0.75
0.50
0.25
0.00
NumInShelters
0.00
0.25
0.50
0.75
1.00
NumUnderPov

<Graph 3 - original response variable>


1.00
0.75
0.50
0.25
0.00
ViolentCrimesPerPop
0.00
0.25
0.50
0.75
1.00
FemalePctDiv
"red"
● red

Considering the two aforementioned problems, non-linearity and multicollinearity, we changed our analysis method to logistic method, which is more flexible and allows us to use the linear model's property as well. However, the response variable in logistic regression can either take the value of "0" or "1." We thus defined a new response variable named "logit." The response variable will take the value of "1," if the original response variable, ViolentCrimesPerPop, is greater than 0.24. On the other hand, the response variable will take the value of "0," if the original response variable is less than 0.24. Graph 2 reflects the relationship between the original response variable and FemalePctDiv, whereas Graph 3 plots the relationship between the new "logit" response variable and the same explanatory variable.

<Graph 4 - logit response variable>



In logistic regression, we firstly divided our original data into 2 separate groups, training data and testing data. The first 70% of the original data was chosen as training data to build formula. And the remaining 30% served as testing data and was used to test whether the formula performs well. Then, we started to build full-model with every explanatory variable, and we used Stepwise selection within BIC(Bayesian Information Criterion) to remove the explanatory variables barely related to the response variable. The reason why we used BIC as criteria is that it imposes more penalties on variables.

Results

The Stepwise selection using BIC allowed us to make a reduced-model with 11 variables out of 100 variables. A statistic for reduced-model on likelihood ratio test is 865.6987 on 11 degree of freedom with p-value=0, n=1396, and a statistic on Wald test is 319.6714 on 11 degree of freedom with p-value=0. As a result, we could say that our reduced model is reliable with 95% confidence level. The chart below shows the exact coefficient estimates for each variable. Finally,we used testing data to evaluate our reduced model made with training data. A cut point to predict the observation whether has relatively high or low crime rates is at 0.5 (median).

<Test Result>

threshold	AUC	omission.rate	sensitivity	specificity	prop.correct	Kappa
0.5	0.81514	0.2792793	0.7207207	0.9095745	0.8394649	0.647007

The reduced-model's prediction power is about 83. 9%. Especially, the probability to predict "0" as "0" is 90.9% accuracy.

<Coefficients Chart>

Coefficients:	**Estimate**	**Std.**	**Error**	**z-value**	**Pr(>|z|)**
(Intercept)	7.8958	1.168	6.76	1.38E-11	***
pctWFarmSelf	1.5292	0.4924	3.105	0.0019	**
pctWInvInc	-6.2308	0.9681	-6.436	1.23E-10	***
NumUnderPov	10.1802	2.0442	4.98	6.36E-07	***
PctPopUnderPov	-3.7541	0.8096	-4.637	3.53E-06	***
PctKids2Par	-6.1354	1.2026	-5.102	3.36E-07	***
PctWorkMom	-2.2919	0.6168	-3.716	0.000202	***
PctIlleg	4.2088	0.878	4.794	1.64E-06	***
PctNotSpeakEnglWell	-3.0342	0.9082	-3.341	0.000835	***
PctHousOccup	-2.0396	0.4617	-4.417	9.99E-06	***

MedOwnCostPctIncNoMtg	-1.5791	0.4995	-3.161	0.001572	**
PctForeignBorn	3.6062	0.8416	4.285	1.83E-05	***

The result indicates that 11 among 100 variables are main factors of the violent crime rates. The variables pctWFarmSelf, NumUnderPov, PctIlleg, PctForeignBorn are positively related to the crime rate, i.e. the increase in these variables raises the crime rate. In particular, the relationship between poverty and crime rate is considerable. Conversely, pctWInvInc, PctPopUnderPov, PctKids2Par, PctWorkMom, PctNotSpeakEnglWell, PctHousOccup, MedOwnCostPctIncNoMtg, especially the first two variables,are negatively related to crime rate, i.e. the increase in these variables lowers the crime rate.

Discussion

Many studies have posed similar questions and have already discovered many of the common characteristics associated with crime. According to Pallini's 'Factors Affecting Crime Rates Across the United States,' ordinary linear regression models were able to identify a few major economic factors affecting crime rates. The factors identified include police expenditures per crime from the previous year, the per capita income, the unemployment rate, and the percentage of people living under the poverty line. We've also found the two significant variables related to poverty have some relationship with crime rates, and we have discussed a lot about these variables.

There is a negative correlation between 'the percentage of population in poverty' and crime rates, and a positive correlation between 'the number of people in poverty' and crime rates. It seems like contradiction, but they are just in trade-off relationships. It brings calculation procedure obviously. We could predict the response logit by calculating 'Estimates x Value', so we should calculate like 'exp(10.1802)*number of people in poverty in 100k people' and 'exp(-3.7541)*percentage of people in poverty in population' for each variable and plus them with other 'Estimates x Values'. Therefore we can make a guess that the effect of 'the number of people in poverty' overwhelm the effect of 'the percentage of people in poverty in population'. Especially the number of people in poverty was divided by 100K, the positive relationship is more reliable as same as Pallini said in his paper.

However, we found some other variables also in other section that economics and we used logistic regression, but pallini worked with ordinary linear regression. In preliminary research, we also found there is high multicollinearity within variables related to income and poverty. Therefore, we can point out the difference and academic shortcomings in Pallini's paper and our papers. Pallini used coefficient to check out multicollinearity instead of VIFs, and that could be a problem.

Our study has some limitations. Some factors cannot or cannot easily be captured in a data set such as ours. Thus, there may still be underlying factors that we are unable to study and account for. Second, there are limitations in methods. Since we used a logistic regression model, we cannot find exact relationships between actual crime rate and these factors but only tendencies. To be more specific, the communities was simply divided as binomial, so the estimates only could say the difference between communities which have relatively higher or lower crime rates. Therefore, some informations could be missed out from the ordinary crime rates data in continuous numbers. Furthermore, the conclusions gathered from this study are only applicable in the scope of the United States due to the origin of the data. Other countries may have largely different tendencies, but we hope the methods and findings may serve as guidelines and benchmarks for studies conducted for other countries.

Conclusion

Goal for this research is finding prominent factors that are correlated to violence crimes may help government authorities allocate the appropriate resources to combat crime. After research is done, we can conclude “Percentage of households with farm or self-employment income in 1989” , “number of people under the poverty level”, “percentage of kids born to never married”, and “percent of people who do not speak English” well have positive relation with odds that Violent crime rate is greater than 0.24. That is, those factors increase violent crimes possibility. Especially, “number of people under the poverty level” increases a unit(0.01 person for 100K), the possibility of outbreaking violent crime increases increase 2637.7(exp(10.1802)) ford. Close look on our findings, economic status is one of the main factor for committing crimes. Also, single-parent family and a language barrier are another main reasons for violent crimes.

For further analysis, we can consider analyzing other countries’ data and compare with this research to figure out common factors for committing violent crimes between different countries. Also, mash-up with other data sets that have different variables which might have relation with the crimes.

References

Pallini, Brian, Factors Affecting Crime Rates Across the United States:, A Regression Analysis < www.marietta.edu/~khorassj/econ421/crime2.doc>

Committee on Understanding Crime Trends, Committee on Law and Justice, Division of Behavioral and Social Sciences and Education, “Understanding Crime Trends:: Workshop Report” , 30.June.2008 <https://books.google.co.kr/books?hl=ko&lr=&id=PvpjAgAAQBAJ&oi=fnd&pg=PT28&dq=factors+in+crime+in+the+US&ots=M0k7s8unIT&sig=Y27TkD45SBMECNw51af4cVMLCIU#v=onepage&q&f=false>

David G. Kleinbaum, Mitchel Klein , Springer , Logistic Regression p 389-428, 24.Mar.2010