A Modeling Approach to the Relationship Between Violent Crime and Demographic Measurements

Mitchell Ottesen
Geography 6000
Fall 2015
Simon Brewer

Abstract

Violent crime is a social phenomenon that can be influenced by several variables. Many studies have been done exploring socio-economic variables at several scopes. Using violent crime data from the Federal Bureau of Investigation, and demographic data from the U.S. Census Bureau, this study explores how population density affects violent crime. General additive models are flexible models that fit curved relationships to scattered data observations. Violent crime can be modeled with general additive models with significant results.

Introduction

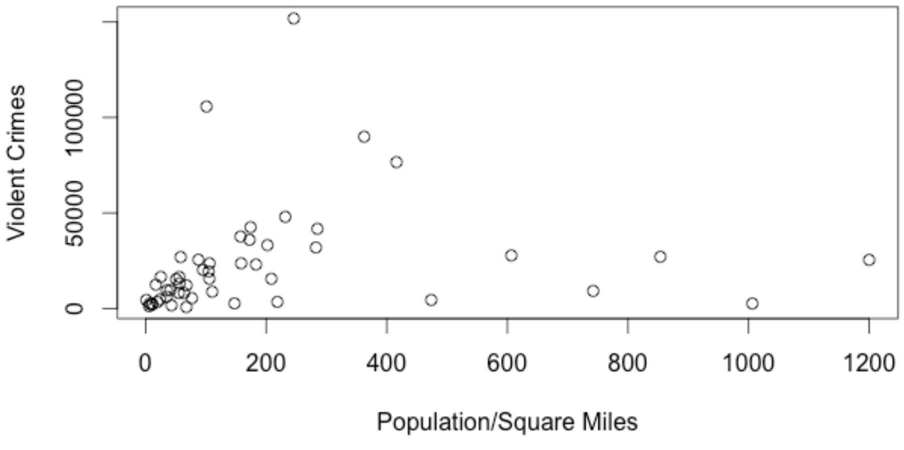
There are many socio-economic factors that can influence violent crime such as income, ethnicity, and household size. In "Neighborhoods and Violent Crime", Robert J. Sampson, a Harvard Social Sciences professor quotes "Violence has been associated with the low socioeconomic status (SES) and residential instability of neighborhoods." (Sampson, 1997). While exploring socio-economic variables to explain the phenomenon of violent crime has been explored repeatedly, the aim of this study is to explore, and model, if any, the relationship between population density and violent crime. Population density is a demographic measurement where a population is measured per unit area. This study was developed with the idea and assumption that humans are more prone to comfort in environments where there is an abundance of personal space. Many psychological studies have been done exploring physiological reactions to inter-personal space variations. "A Methodological Investigation of Personal Space" repeatedly shows that as levels of interpersonal space increase, anxiety and stress levels drop and relaxation increases (Evans, 1973). With these findings in mind, this study explores if there is a linear relationship between violent crime and population density. That is, if geographic units with a more dense population will have an increase in violent crime. It is suspected that units with a dense population will have higher levels of violent crime and violent crime can be modeled based on population density calculations alone. This study focused on two scopes, State-level and City-level.

Data

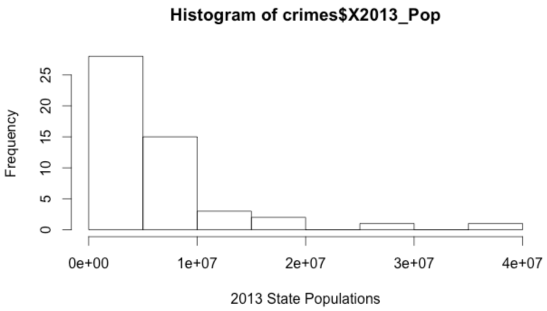
The Federal Bureau of Investigation (FBI) makes violent crime data freely available. An array of criminal offenses is defined as being violent: Murder, Voluntary Manslaughter (Non-negligent), Rape (revised definition), Rape (legacy definition), Robbery, and Aggravated Assault. The violent crime data used in this study came from Table 5 and Table 8, downloaded from the FBI website (www.fbi.gov). Table 5 is a collection of violent crime data divided by state (FBI, 2013). Table 8 is a collection of known criminal offenses divided by state, then divided again by city. State population data and City population data also originated from Table 5 and Table 8. Area data was needed for each U.S. state. This was conveniently found in table format from TheUS50 (Schubach, 2015). A website that publishes simple facts about the U.S. states. The table cites that the land square mileage data was retrieved from the United States Geological Survey. Income data was also used in the study. Median Household Income per State was collected from the 2013 Current Population Survey. The District of Columbia was left out of this study. On a city level, Median Household Income and Income Per Capita data were collected from the 2013 American Community Survey (ACS). The cities used in the study are defined by the U.S. Census Bureau as Metropolitan Statistical Areas (MSAs). Metropolitan Statistical Areas on a general sense are cities or a cluster of cities that have large populations. MSAs were used in this study because the data necessary to fill the variables were readily available. In cases where a MSA encompassed multiple cities and both cities were listed separately on Table 8 from the FBI, the same income value from the MSA found on the 2013 ACS and CPS was compared with the separately listed cities on Table 8. For example, the 2013 ACS combines San Jose and San Francisco California as one MSA. San Jose and San Francisco are listed separately on Table 8. In this case, the single Median Household Income value and the Income Per Capital value from the ACS were both used for San Jose and San Francisco, respectively. The land area for each city used in the study was collected from each city's individual Wikipedia article. All the data collected for the cities used in the study were divided into two different datasets. The dividing factor was the income variable. Cities where Median Household Income was collected from the 2013 ACS was compiled with a total of 29 observations. Cities where Income Per Capital was collected were compiled with a total of 64 observations.

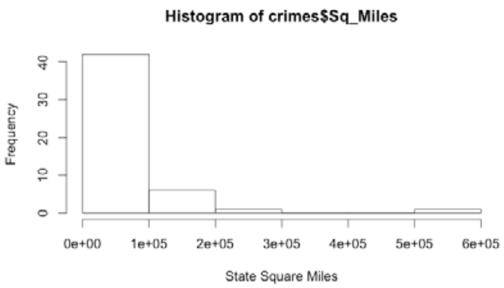
Methods

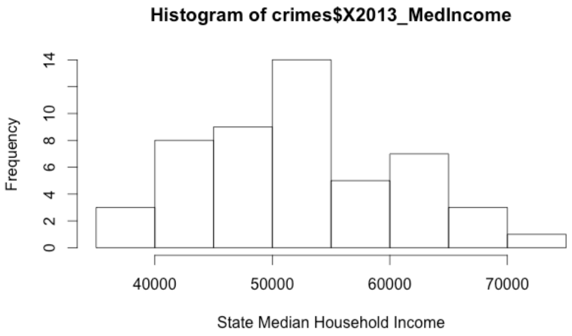
The study began examining the data at the state level with a national scope. As with any study involving statistics, it is first very wise to begin with some Exploratory Data Analysis methods. A basic plot viewing the relationship, if any, between Population Density and Violent Crime can be found below. Population density was calculated by dividing the state population by the square miles.



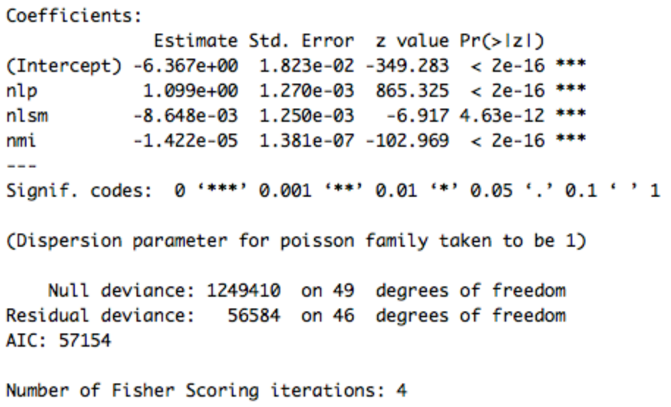
Examining the plot above, there does seem to be a bit of a linear relationship. There are definitely a few outliers. New Jersey has a population density of 1199 persons per square mile. There were approximately 25,000 violent crimes committed in New Jersey in 2013. Rhode Island also has a high population density, however there were only ~1,100 violent crimes committed in Rhode Island in 2013. The response variable in this study is a total amount of violent crimes in each state. This qualifies the quantitative data as being 'count' data. Because the response variable is not continuous, a Generalized Linear Model or Generalized Additive Model will have to be used. Histograms were developed and examined to view normality in the independent variables.





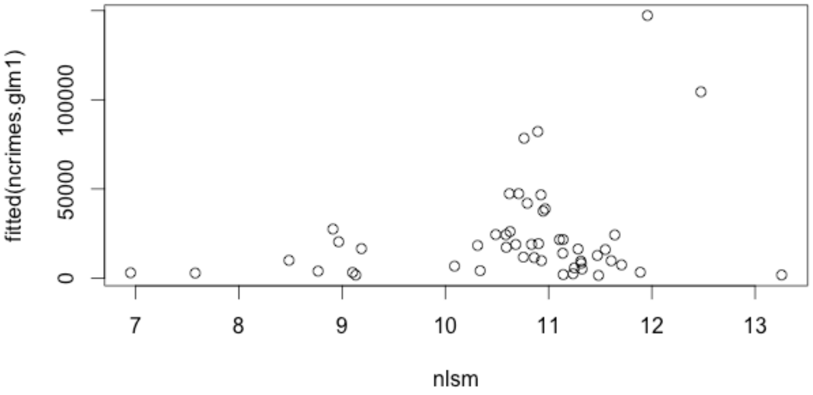


The three independent variables used in the model are: state population, state square miles (land area), and state household median income. Of the three independent variables, two are clearly very heavily right-skewed. For a better-fit GLM/GAM, these variables were log-transformed to more closely resemble a normal relationship. To accommodate the response variable, the GLM was constructed using a "Poisson" distribution and "Log" link function. These act to help make the relationship between the response variable and independent variables linear. A summary of the GLM can be found below.

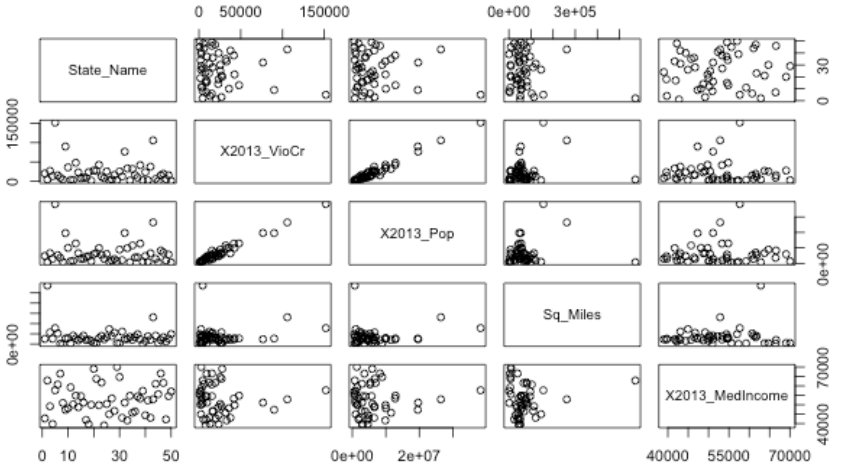


```
Coefficients:
              Estimate Std. Error  z value Pr(>|z|)    
(Intercept) -6.367e+00  1.823e-02 -349.283  < 2e-16 ***
nlp          1.099e+00  1.270e-03  865.325  < 2e-16 ***
nlsm        -8.648e-03  1.250e-03   -6.917 4.63e-12 ***
nmi         -1.422e-05  1.381e-07 -102.969  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1249410  on 49  degrees of freedom
Residual deviance:   56584  on 46  degrees of freedom
AIC: 57154

Number of Fisher Scoring iterations: 4
```

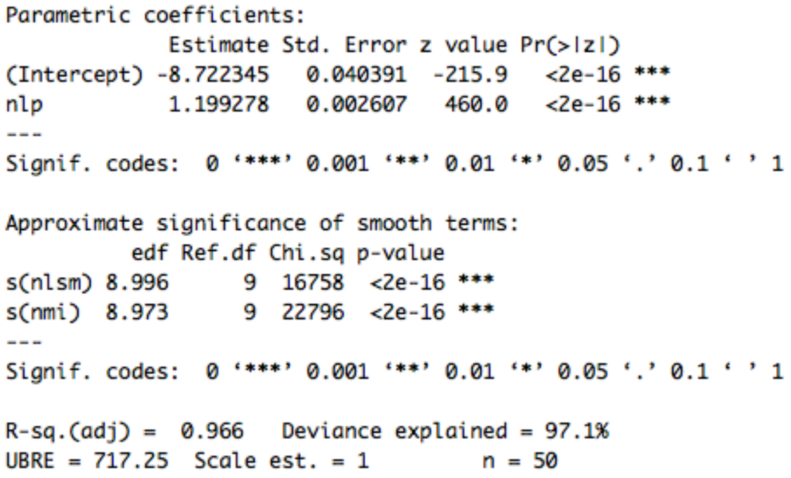
The coefficients in the model above are difficult to interpret. The two variables that were log-transformed were reversed back to non-log form. The population coefficient (nlp in model) is 3 and the square mile coefficient (nlsm) is 0.99. The coefficients all have significant p-values. Once transformed, population density in the form of the square mileage and population of the state do have a positive effect on the model. There is a large reduction of deviance between the Null model and the Residuals of the model. However, the residual deviance is still quite large. This may arise the question as to how well the model fits the data. The plot below does show that the model relationship with the square mileage data isn't actually clearly linear.



On top of the GLM, a GAM model may also be appropriate as well. In comparing how the different independent variables relate with the response variable, if the relationships are not all monotonic, or linear, a General Additive Model may be a better choice. The plots below show that not all the relationships are monotonic.

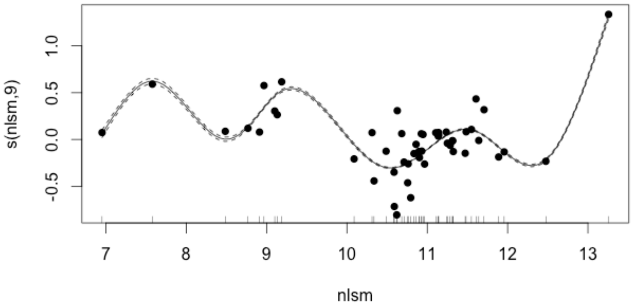


The use of a GAM might be a better choice because it utilizes smoothers to help fit the data. It was shown previously that while the GLM for the state data was significant, there was a high amount of deviance. The same log transformations were used for the GAM. Also, the same syntax was used to build the model. That is, it still utilized the “Poisson” distribution with the “Log” link function to accommodate the integer-form response variable. Smoothers were applied to the income and square mileage variables because their relationships with the response variable were not clearly monotonic. Below is a summary of the General Additive Model.

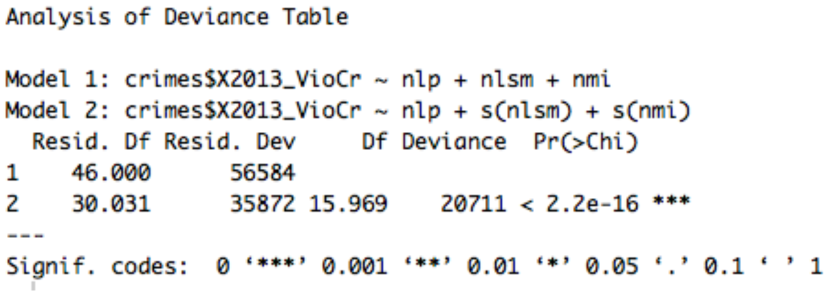


```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.722345   0.040391  -215.9   <2e-16 ***
nlp          1.199278   0.002607   460.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df Chi.sq p-value
s(nlsm) 8.996      9  16758  <2e-16 ***
s(nmi)  8.973      9  22796  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.966   Deviance explained = 97.1%
UBRE = 717.25  Scale est. = 1         n = 50
```

The GAM does seem to fit the data well. However, one can easily be skeptical of the results. The low p-values and R-square value indicate this. Shown below is the smoother for the square mileage variable.

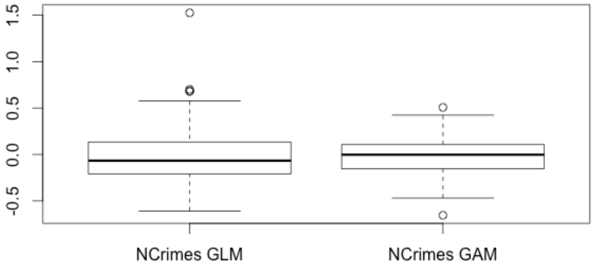


There is a lot of deviation from the smoother. Not the pattern one would hope for. However, the GAM may still be a better fit than the GLM. This can be tested and visualized using an ANOVA and Boxplot.



```
Analysis of Deviance Table

Model 1: crimes$X2013_VioCr ~ nlp + nlsm + nmi
Model 2: crimes$X2013_VioCr ~ nlp + s(nlsm) + s(nmi)
  Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)    
1    46.000      56584                              
2    30.031      35872 15.969    20711 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

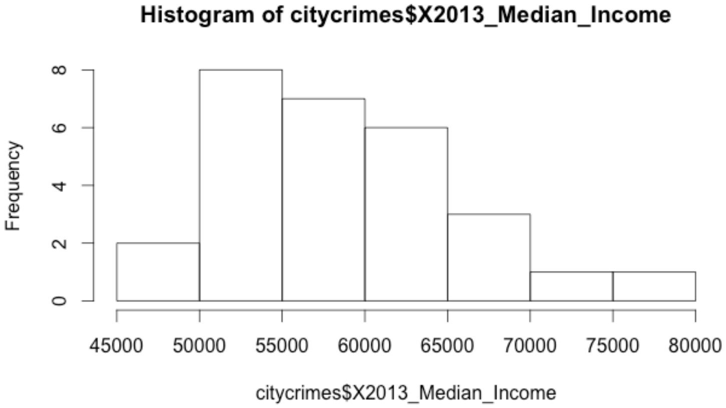
The GAM does appear to be a better fit according to the ANOVA. It has less deviation and the small P-value next to model #2 (GAM) confirms.



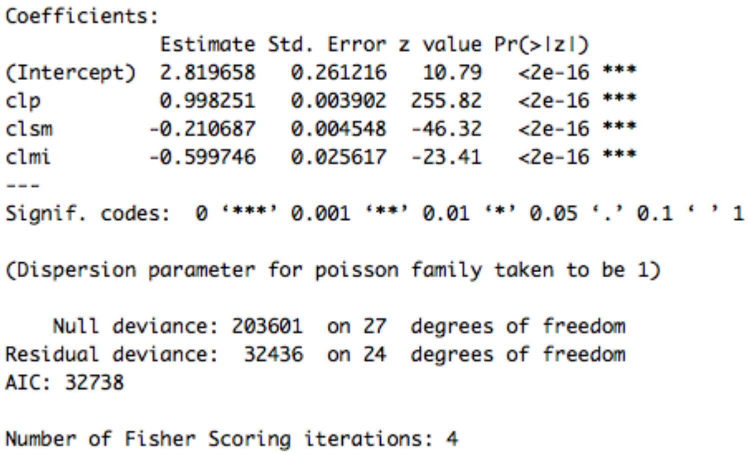
The Box Plot above confirms the ANOVA results. There are less residuals surrounding the GAM.

Methods, cont.

Similar patterns were found in both city datasets. The first dataset examined was the city dataset where the income variable defined is the Median Household Income in each city. This dataset has 28 observations. Unlike the state-level portion of the study, in this dataset, each independent variable (Population, Square mileage, Household Median Income) was heavily right-skewed and log transformed prior to the model formations. The histogram below shows the median income data. The data is right-skewed unlike the state-level dataset.

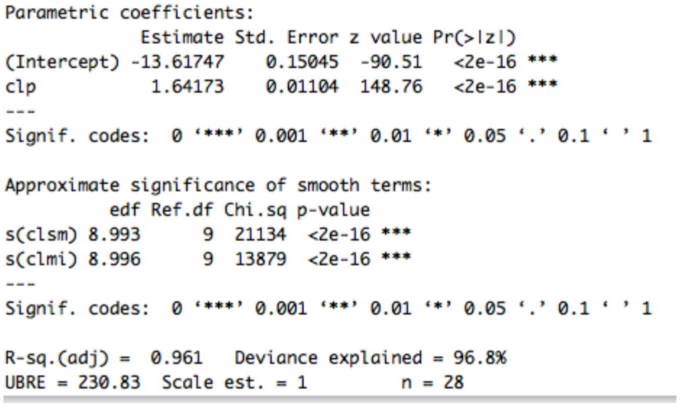


After the transformation of the variables, the GLM was generated. Because the response variable is still measured in counts, the “Poisson” distribution and “Log” link function were still used.



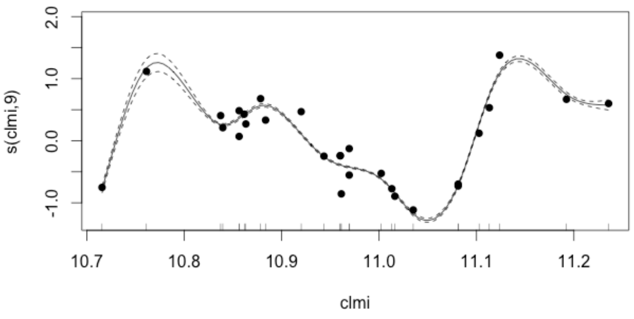
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.819658   0.261216   10.79   <2e-16 ***
clp          0.998251   0.003902  255.82   <2e-16 ***
clsm        -0.210687   0.004548  -46.32   <2e-16 ***
clmi        -0.599746   0.025617  -23.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 203601  on 27  degrees of freedom
Residual deviance:  32436  on 24  degrees of freedom
AIC: 32738

Number of Fisher Scoring iterations: 4
```

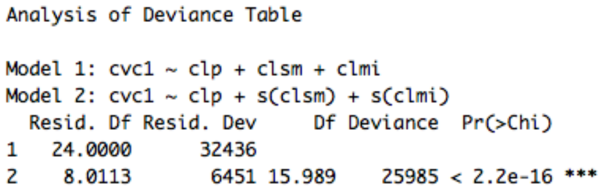
The coefficients are difficult to interpret. When they are transformed back from their log forms, they are all positive. Square mileage (clsm) and Median Income (clmi) both hold positive values less than one: 0.81 and 0.54. As expected, between the Null model and GLM, there is a large reduction in deviance. However, a Residual deviance value of 32,738 is rather large.



```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.61747    0.15045  -90.51   <2e-16 ***
clp           1.64173    0.01104  148.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df Chi.sq p-value
s(clsm) 8.993      9  21134  <2e-16 ***
s(clmi) 8.996      9  13879  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.961   Deviance explained = 96.8%
UBRE = 230.83  Scale est. = 1         n = 28
```

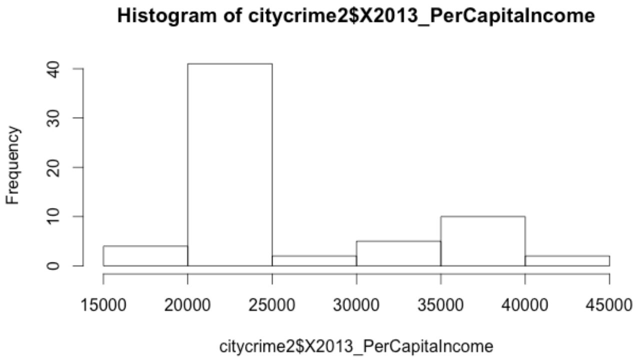
The p-values are small and R-squares are large. This model seems to be a better fit than the GLM. Unlike the state-level GAM, there is not as much deviation around the smoothers with this GAM. See the smoother for the Square Mileage variable below.



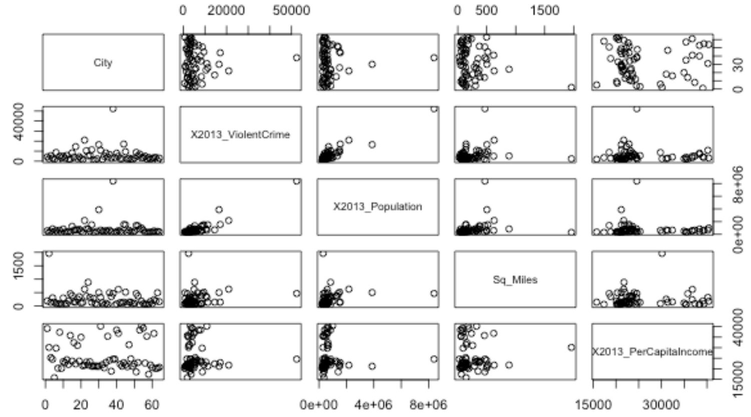


```
Analysis of Deviance Table

Model 1: cvc1 ~ clp + clsm + clmi
Model 2: cvc1 ~ clp + s(clsm) + s(clmi)
  Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
1   24.0000      32436
2    8.0113       6451 15.989    25985 < 2.2e-16 ***
```

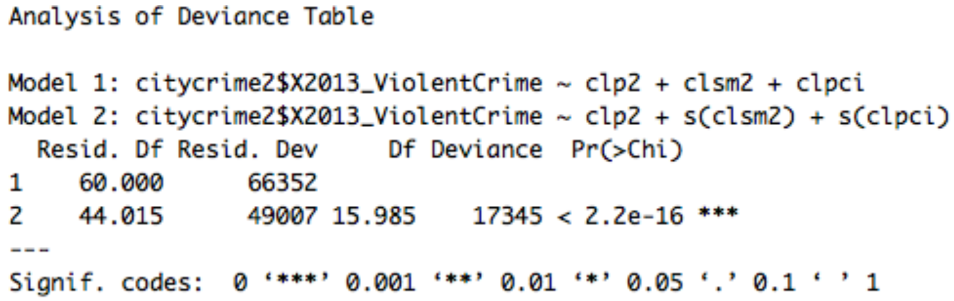
An ANOVA test indicates that the GAM is a better fit than the GLM for this dataset. Utilizing the smoothing capability of the GAM seems to make a big difference. The last models created explored the dataset where the income independent variable is Income Per Capita rather than Median Household Income. This divides the average income throughout the entire population, including children. This brought the values in the observations down significantly. Most of the cities had an Income Per Capita value in the 20-25k range. The histogram can be seen below.



This data was log-transformed as well as the other independent variables before developing the two models. To reiterate, both a GLM and GAM were developed for each dataset because it is unclear whether the relationships between the dependent variable (Violent Crimes reported) and independent variables (Population, Square Mileage, Income) are monotonic. The plots below show that there may be a sign of a linear relationship but it cannot be determined. Both a GLM and GAM are built and then compared against each other to determine which model fits the data best.



The ANOVA test for the second City-level dataset's GLM and GAM models show that the models are very close together in deviance. However, the GAM is chosen as the most significant and better fit.



```
Analysis of Deviance Table

Model 1: citycrime2$X2013_ViolentCrime ~ clp2 + clsm2 + clpci
Model 2: citycrime2$X2013_ViolentCrime ~ clp2 + s(clsm2) + s(clpci)
  Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)    
1    60.000      66352                              
2    44.015      49007 15.985    17345 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results

For this study, General Additive Models were found to be a better modeling selection. Each model was found to be statistically significant in terms of the variables used and their effect on the response variable. Each model showed large amounts of deviation from the actual data values. Smoothers did prove to be a useful tool in fitting the model to the data. However, with some of the smoothed variables, there were still large amounts of deviance.

Discussion

Violent crime is a complex phenomenon. There are several variables that can affect it. To successfully model violent crime in the future, there will have to be more variables. Implementing more socio-economic variables besides income would be worth considering. However, this can be problematic because it is encouraged to make models as simple as possible. Also, changing the geographic unit to smaller scopes would be a good approach. Census tracts would be worth examining because each neighborhood is different in a City and a few neighborhoods will fall within a single Census tract. That would allow enough variation within each Census Tract that the data could be modeled. Unfortunately, finding the data at a Census Tract would be difficult to find.

Conclusion

Population and Population Density are variables that influence violent crime. This phenomenon is difficult to model, especially at large scopes. Finding the perfect model for as large a scope as a populous city or even an entire state would be incredibly difficult. The generalized additive models formulated in this study are good models. They fit the data and are statistically significant. But they are far from perfect. This is due to the high levels of deviation that surround the curvature of the models. Implementing more socio-economic variables at a smaller scope might be a better approach to developing models that have a better fit and do not deviate as much around the observations.

Works Cited

Current Population Survey. “Income of Households by State Using 2-Year-Average Medians. *U.S. Census Bureau,* 2011 to 2014 Annual Social and Economic Supplements, Accessed Nov. 2015

Evans, Gary W. “A Methodological Investigation of Personal Space.” *EDRA 3: Research and Practice*, 2-2-1 to 2-2-8, 1972, Accessed: Dec. 2015

Federal Bureau of Investigation, United States Department of Justice “Crime in the United States 2013.” *Uniform Crime Reports, Table 5,* Accessed Nov. 2015

Federal Bureau of Investigation, United States Department of Justice “Crime in the United States 2013.” *Uniform Crime Reports, Table 8,* Accessed Nov. 2015

Noss, Amanda. “Household Income: 2013”. *American Community Survey Briefs,* ACSBR/13-02 (September 2014), Accessed Dec. 2015

Sampson, Robert J. “Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy.” *Science,* New Series, Vol. 277, No.5328 (Aug. 15, 1997), 918-924. Accessed Dec. 2015

Schubach, Erik. “Fast Facts Study Guide (State Areas).” *TheUS50*. url: www.TheUS50.com/fastfacts/area.php. Accessed Dec. 2015