Statistical Analysis on Crime Characteristics Within Communities

Milani Lawrence, Caiyun Zhu, Prarthana Bhattarai

12/15/2014

Abstract

The Justice System in America has been under critical social analysis this year because of the increase in police brutality against civilians, especially African American, and working class citizens. Many Americans believe police are reacting violently to these groups of people because of internalized racial stereotyping. However, how are these racial stereotypes formulated, does the racial and economic construct of a community attribute any significant influence on crime status? In this paper, we created several multivariate linear regression models that aim to understand the factors that affect crime status of a community (ie: either violent or nonviolent) in America by using multivariate regression methods.

Introduction

Crime in the United States has been present since colonization. Crime rates have varied over time, with a sharp rise after World War II, before peaking between the 1970s and early 1990s especially in urban communities. Currently, due to new technologies, such as data mining, statistical regression modeling and other techniques Americans are able to identify patterns and detect future criminal actions therefore helping to decrease the crime rates. However, our justice system is not perfect and there are many problems we still need to fix. For example, scientists, statisticians and government officials are spending time studying crime and criminal behaviors in order to understand the characteristics of the crime and to discover crime patterns in order to decrease crime rates.

The primary purpose of our project is to create several multivariate linear regression models that aim to understand the factors that affect crime status (ie: either violent or nonviolent) in American cities. We want to be able to model several different characteristics such as the different races in a society, income groups, age groups, family structure (single, divorced, married), level of education, the racial match between police allocated to a locality and the community and the number of employed and unemployed people.

We also wanted to test a few hypotheses in order to understand how specific certain characteristic of a community could affect crime rates. Some examples of hypothesis test we wanted to run are: how does median rent, racial structure of the community, the structure of the police force (ie: measure of the racial match between the community and the police force) , and family structure affect crime status within the community.

Data

Our data set, “Communities and Crime Unnormalized” data set, was accessible from the UCI Machine Learning Repository. Our data set focuses on American cities, and combines the socio-economic data from the 1990 US Census, law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics survey, and crime data from the 1995 US FBI Uniform Crime Report.

The raw data set consisted of 2215 total observations and 147 variables for communities. The 147 variables were split between 125 predictive variables, 4 non-predictive variables and 18 potential goal attributes. The observational unit is one community, which falls under the jurisdiction of one police department. The population of our model is all communities in different states in America that are under the jurisdiction of one police department, for a rough total population estimate of $12,575$. The states are represented in the form of number, every number representing its respective American state. Our response variables were represented as a rate of the number of violent crimes per $100,000$ population and the number of nonviolent crimes per $100,000$ population. Also included in our data set was the measurement of violent crimes, which are murder, rape, robbery, and assault. Our predictor variables included information across a diversity of crime-related facts, ranging from the percentage of racially matched officers assigned to a community, household density, percentage of the population that live in an urban environment to the median rent of an apartment within a community.

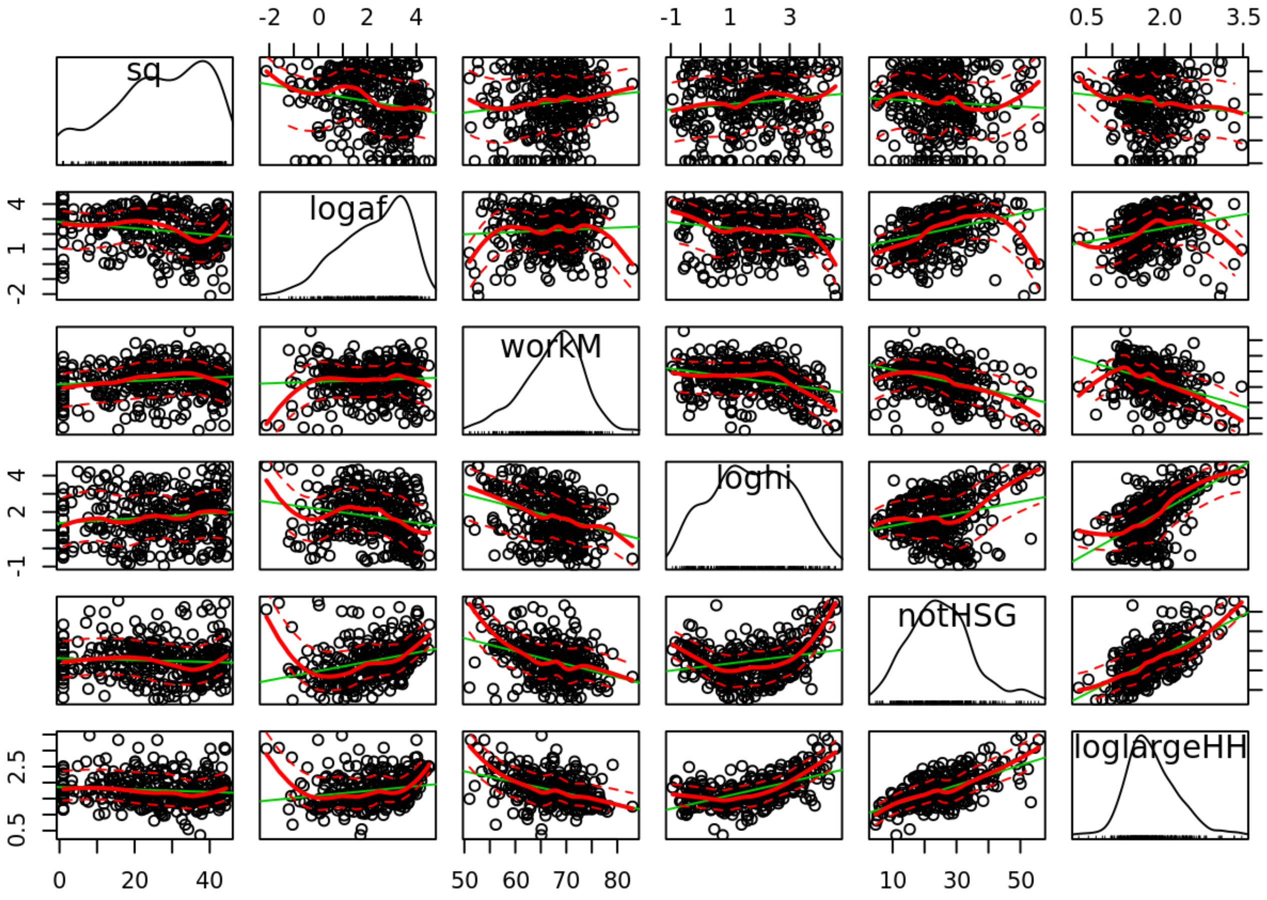
The first step for data cleaning was using subsetting methods in R in order to omit the incorrect fields, or rows that contained missing values. After that, we were left with 323 total observations. Then, we also read a number of articles that gave us some basic knowledge about factors that affect crime rates. With a general idea of the basic characteristics that influence criminal status, we were able to shift through all of our observations and label 25 predictive variables as relevant with a corresponding 323 observational units by removing entries without any values. We applied logarithmic transformation and quadratic transformation for some of our predictors because of exponential relationship and quadratic relationships between dependent variable and independent variables.

Models and Approach

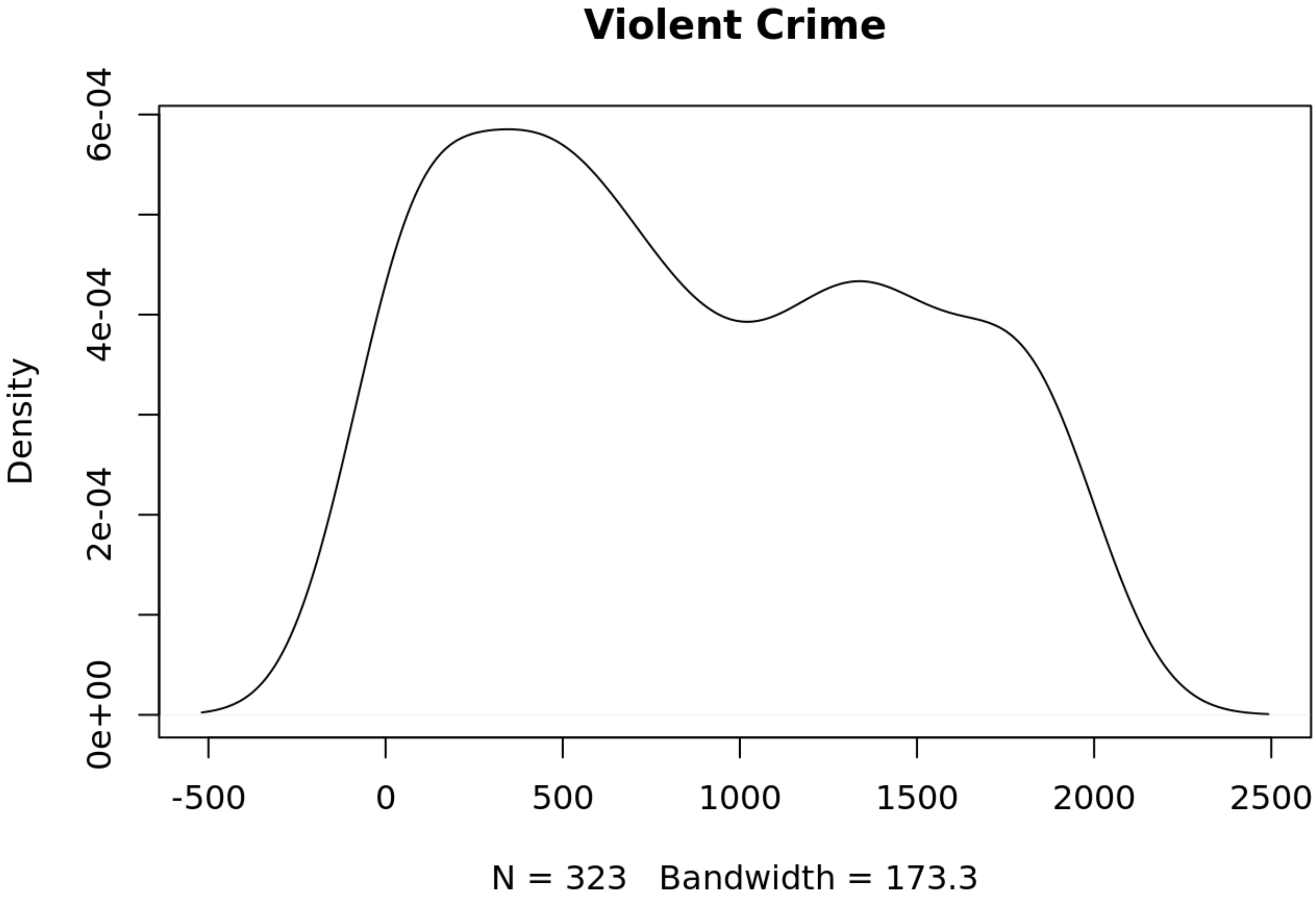
We developed separate models for violent and nonviolent crimes as we believed that these two crime rates are influenced by different factors. Our final data set contained 25 variables. We performed exploratory data analysis, looking at the scatter plots of the response variables against each of the predictors to determine which variables need to be transformed before building a linear regression model.

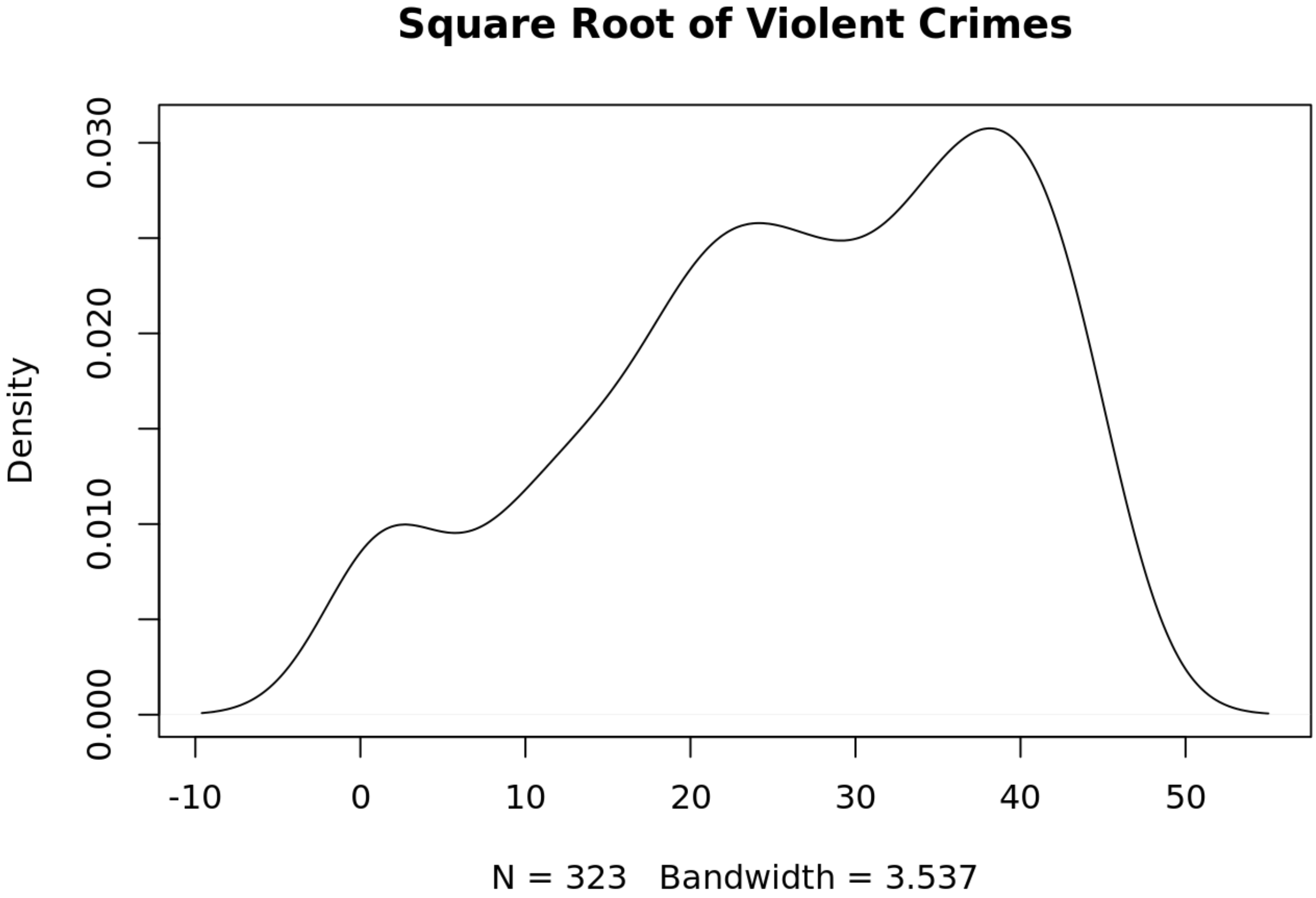
Model for Violent Crimes

Although we did not see much structure in the initial scatter plots, after a few transformations we noticed structures suggesting some association between the response variables and the predictors. This can be observed from the scatter plot matrix of the final model.



The distribution of the response variable (rate of violent crimes) did not follow a normal trend. We noticed that it resembled the shape of a bi-modal distribution. We transformed the response variable using square root and obtained a slight improvement in the shape of the distribution. After the transformation, the bi-modal trend looks less prominent.


Violent Crime
Density
6e-04
4e-04
2e-04
0e+00
-500
0
500
1000
1500
2000
2500
N = 323   Bandwidth = 173.3

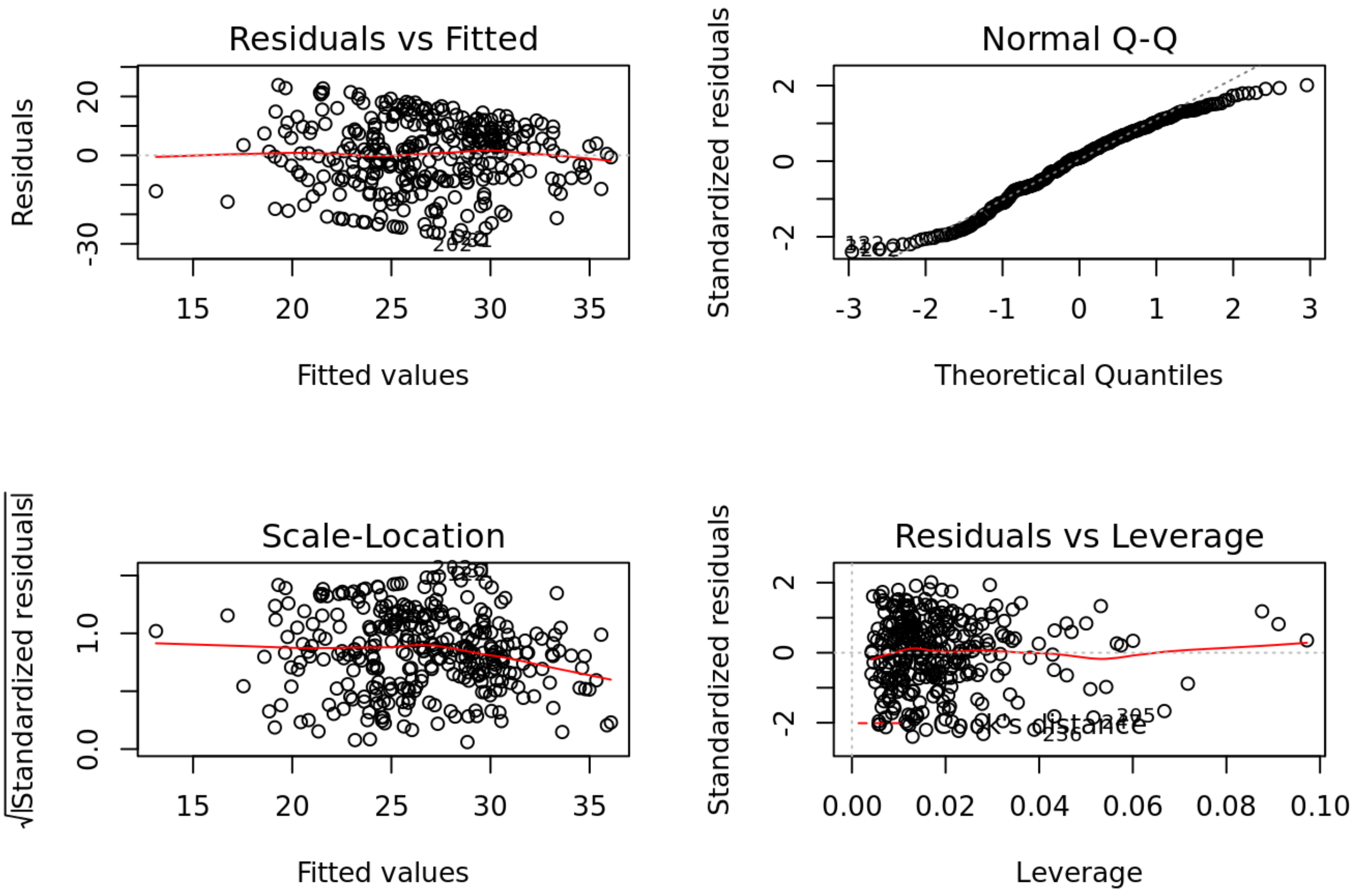

Square Root of Violent Crimes
Density
0.030
0.020
0.010
0.000
-10
0
10
20
30
40
50
N = 323   Bandwidth = 3.537

To narrow down the number of variables and to find the best model, we used AIC and BIC, in both forward and backward directions. We looked at the models obtained in each step of the model selection process through AIC and BIC. To select the best model, first, we looked at the quartet of diagnostic plots to assess model validity. Second, we eliminated the models in which the three assumptions (linearity, normality and constant variance of errors) for building a multiple linear regression model did not seem reasonable. Then, we selected the best model based on the value of adjusted-R^2.

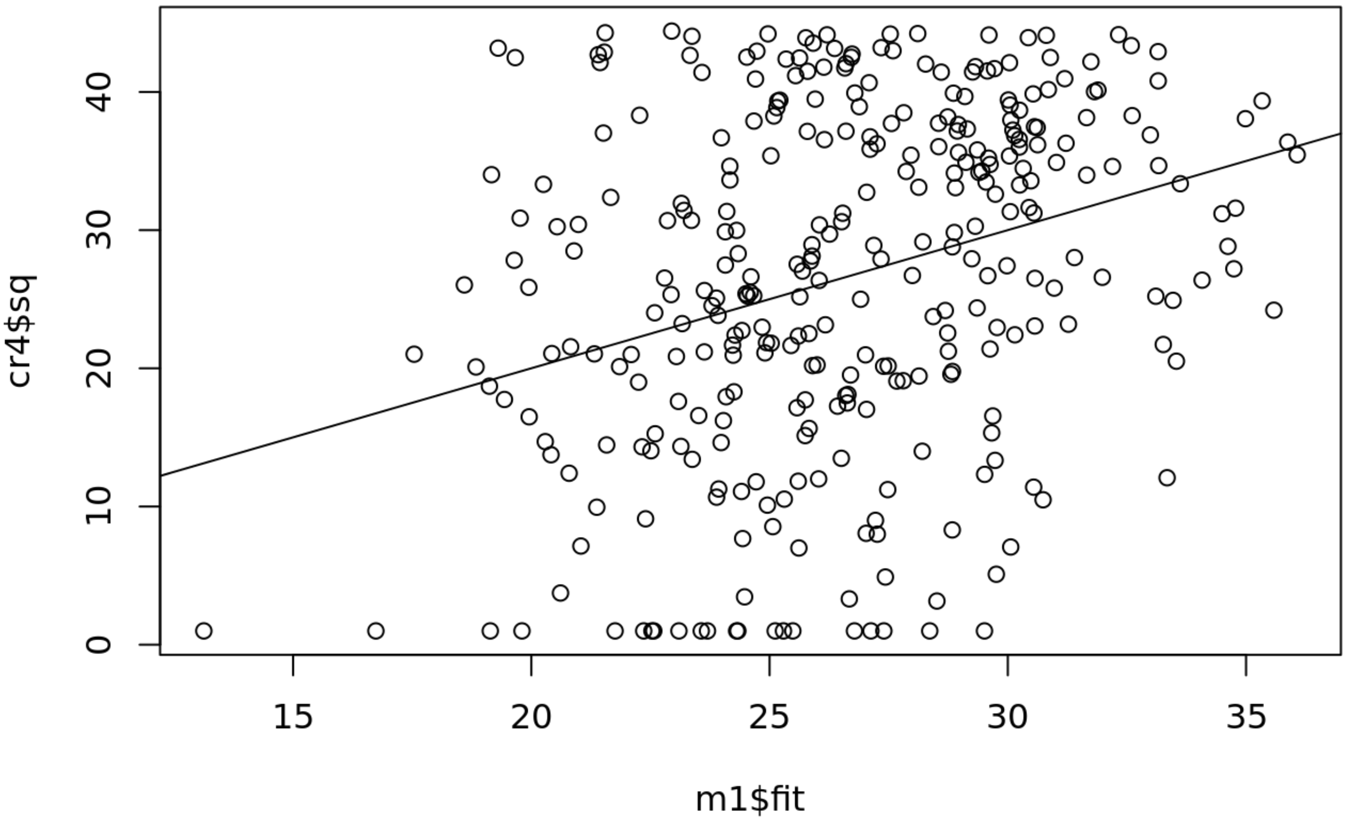
The final model for the rate of Violent Crimes had five predictors, out of which four predictors turned out to be statistically significant at 95% confidence level:

$$\widehat{Violent}\ \beta_0 + \beta_1 logaf + \beta_2 workM + \beta_3 loghi + \beta_4 notHSG + \beta_5 loglargeHH + \epsilon.$$

The diagnostic plot for the model shows that the assumptions of linearity, normality and constant variance are reasonably met. Although the Normal Q-Q plot shows that the distribution of errors is not quite normal, particularly in the extreme tails, it is not too concerning. In addition, there are no influential points with a cook's distance greater than 0.5. However, one striking feature of the residual vs. fitted values graph is the distinct bin of residual values. The response variable in this model is expressed as rate, which means that there are sharp cut off values for the response. The value of response variable is always within a range of 0 to $100,000$. This results in a bin-shaped spread of residual values. A similar trend is noticeable in the $Y\ vs.\ \hat{Y}$ graph; the communities with very low rates of crime (close to 0) appear at the bottom, distinctly away from the rest of the data. The adjusted-R^2 value of this model is quite low; only 8% of the total variation in the rate of violent crimes is explained by the predictors in the model, suggesting that the model does not have a strong explanatory power.



The four predictors that were statistically significant (at 95% confidence level) were percentage of the population that is African American, percentage of the population that is Hispanic, percentage of mothers with kids under 18 in labor force and the percentage of family households that have 6 or more people. The the predictors such as the percentage of the population that is African American and the percentage of large family households were negatively correlated, whereas the percentage of the population that is Hispanic and the percentage of mothers with kids under 18 who were working were positively correlated with the rates of violent crimes. The percentage of people of age 25 and over who were not high school graduates was correlated positively with the rate of violent crimes, only at 90% confidence level.



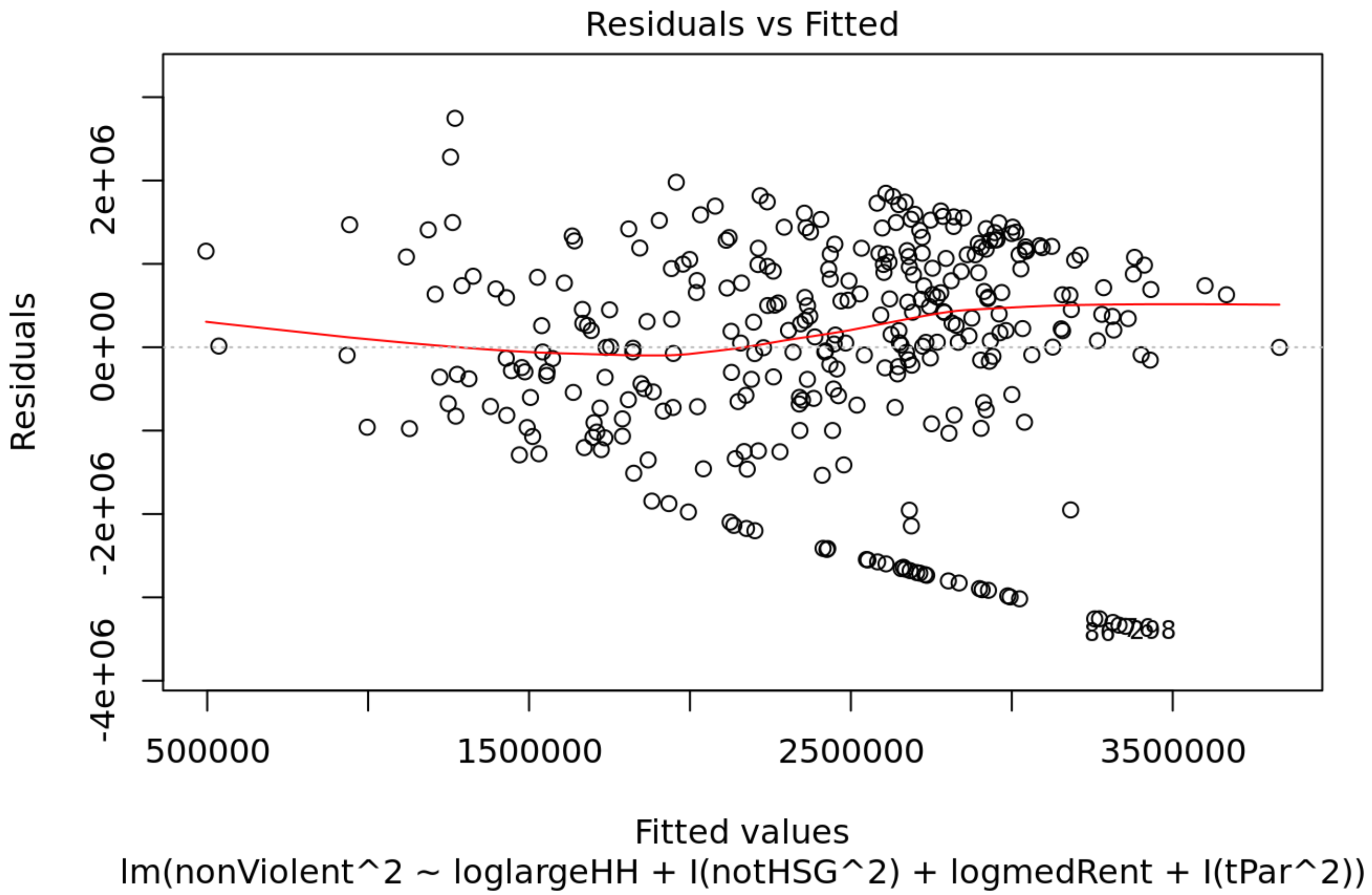
```
##
## Call:
## lm(formula = sq ~ logaf + workM + loghi + notHSG + loglargeHH,
##     data = cr4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.510  -8.018   1.192   9.096  23.870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.21436   10.42284   0.788 0.431221
## logaf       -1.42892    0.57004  -2.507 0.012687 *
## workM        0.35730    0.13475   2.652 0.008413 **
## loghi        2.34436    0.68868   3.404 0.000749 ***
## notHSG       0.18306    0.09501   1.927 0.054903 .
## loglargeHH  -6.28852    2.34131  -2.686 0.007614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.96 on 317 degrees of freedom
## Multiple R-squared:  0.09528,    Adjusted R-squared:  0.08101
## F-statistic: 6.677 on 5 and 317 DF,  p-value: 6.248e-06
```

Models for Nonviolent Crimes

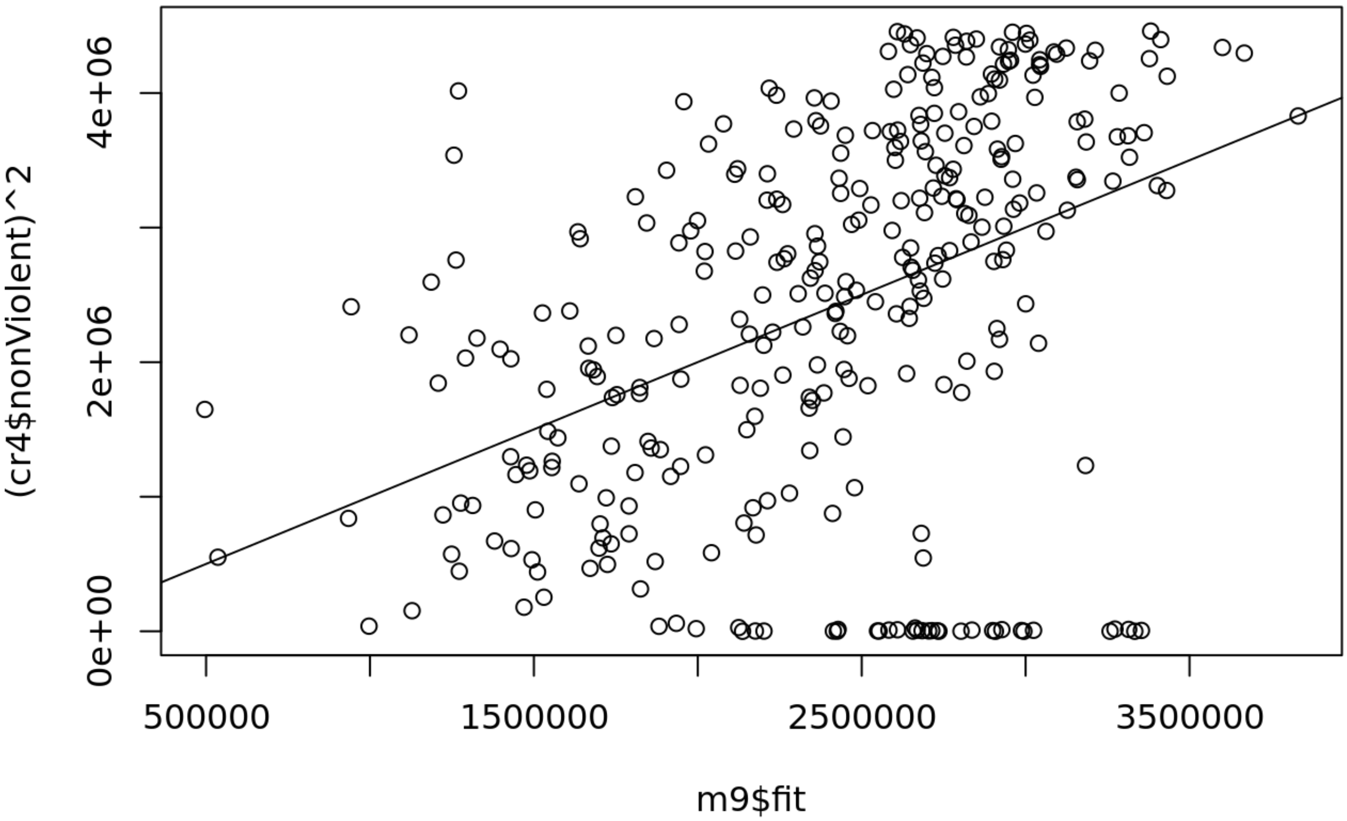
We then did a very similar data analysis on the number of nonviolent crimes per $100,000$ people within a community. First, we realized that a quadratic transformation of nonviolent rate was necessary because this improved the linear relationship between nonviolent crime rates and the predictors. We also included quadratic terms of some of the predictors. Secondly, after running Forward Selection Algorithm on all the 25 predictors that could potentially affect nonviolent crime using AIC as the guideline, we obtained the following model with 4 significant predictors for nonviolent crimes: proportion of households in a community that have family size greater than or equal to 6, population proportion of people of age 25 and over who were not high school graduates, median rent of an apartment in a community, and proportion of children in a community that have two parents in one household.

$$\widehat{nonVio}^2 = \beta_0 + \beta_1 loglargeHH + \beta_2 notHSG^2 + \beta_3 logmedRent + \beta_4 tPar^2 + \epsilon$$

This model approximately meets 2 modeling assumptions (normality in residuals and no obvious outliers) with the exceptions in the residual plot and scale-location plot where there is a prominent negative linear pattern underneath the zero residual line and the change in variance.

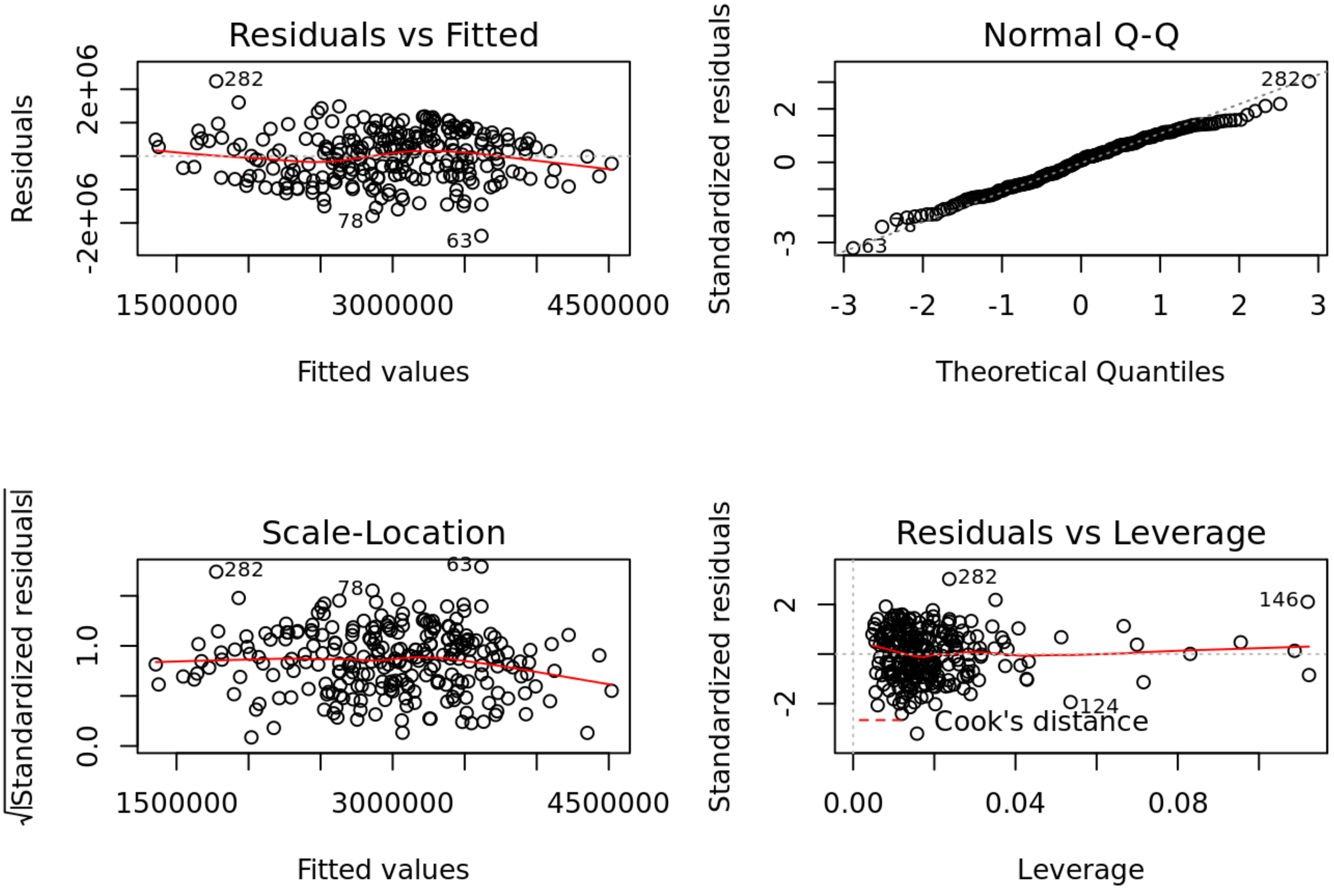


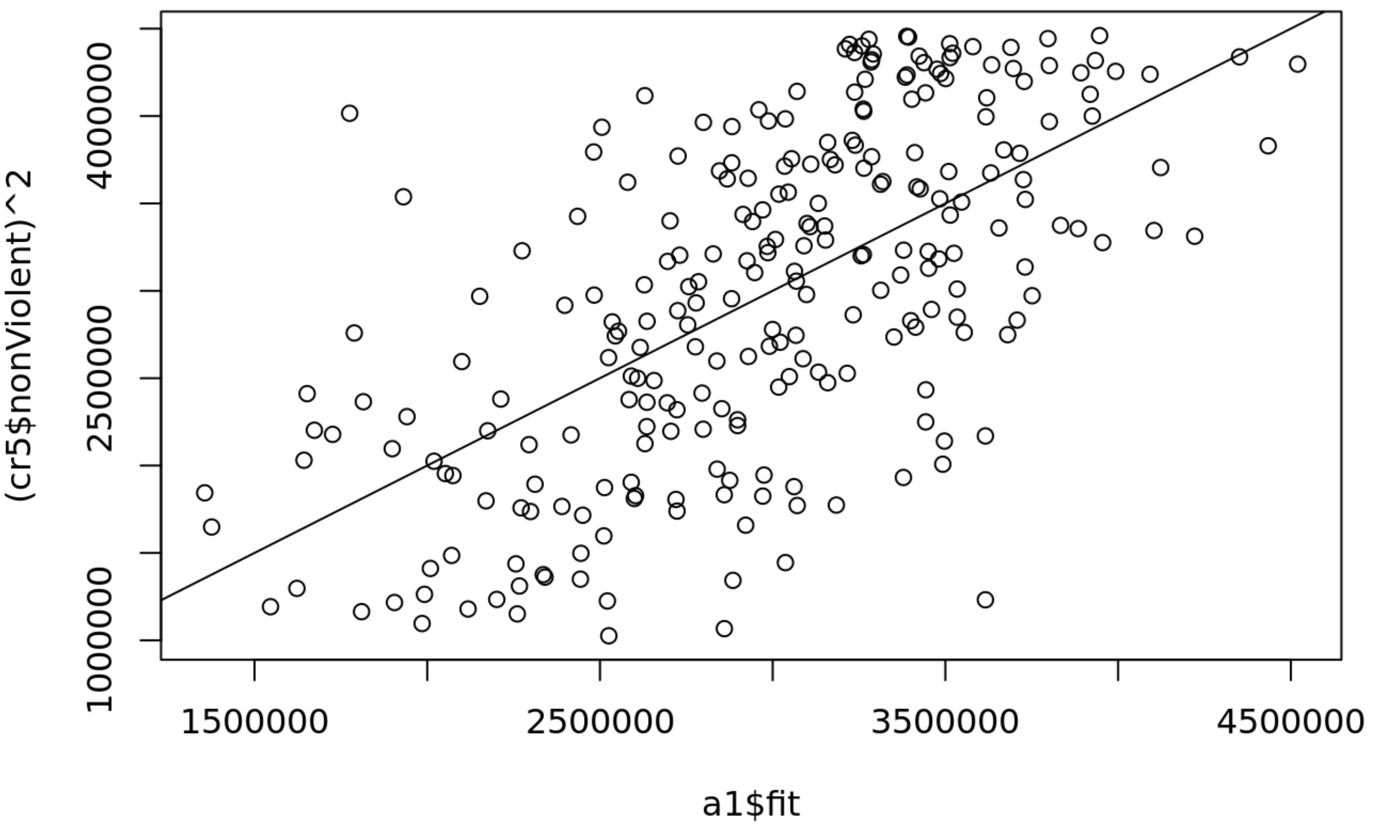
Looking at the $Y\ vs.\ \hat{Y}$ plot, we saw that the predicted nonviolent crime rates from our model does not fit the data very well. Those data points with low nonviolent crime rates cause the slope of the $Y = X$ line to decrease. They are good explanations for low adjusted-R^2 of 0.1741.



Then we decided to remove those communities that have low nonviolent crime rates and run the same model for the 253 communities left over. As a result, we obtained a model that met the 4 modeling assumptions (normality, randomness, constant variance, no outliers in residuals) better and had a higher adjusted-R^2 of 0.3879 with all the same 4 predictors remaining significant. At the same time, the $Y = X$ line is closer to capture the trend in this data set and all the variance inflation factors of this model are below 2, which implies little to no multicolinearity among the predictors. Therefore, we concluded that this is the best model for rates of non-violent crimes.

```
##
## Call:
## lm(formula = nonViolent^2 ~ loglargeHH + I(notHSG^2) + logmedRent +
##     I(tPar^2), data = cr5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2383773  -570609    59975   535161  2240508
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9254341.69  931457.13   9.935  < 2e-16 ***
## loglargeHH   633277.28  137138.72   4.618 6.23e-06 ***
## I(notHSG^2)    -653.53     128.26  -5.095 6.90e-07 ***
## logmedRent  -895398.60  174897.40  -5.120 6.15e-07 ***
## I(tPar^2)      -356.26      47.17  -7.552 8.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 748400 on 248 degrees of freedom
## Multiple R-squared:  0.3976, Adjusted R-squared:  0.3879
## F-statistic: 40.92 on 4 and 248 DF,  p-value: < 2.2e-16
```





Discussion

In the violent model, we found that the factors like percentage of community population that is African American and the percentage of family households that has 6 or more people had negative correlation with the rate of violent crimes, whereas the factors like percentage of the population that is Hispanic and percentage of mothers with kids under 18 in the labor force were positively correlated with the rate of violent crimes.

In the nonviolent model, we found that the factors like the percentage of family households that are large positively correlates with the rate of nonviolent crimes. The factors that are negatively correlated with nonviolent crime rate are population proportion of people of age 25 and over who were not high school graduates, median rent, and proportion of kids in a community that have two parents in one household.

One weakness of our multiple regression model is that the crime rates are measured in the number of cases per $100,000$ population, thus there are minimum and maximum cutoff values that lead to a prominent band in the residual plot of the violent crime model. Although the second model gives us much higher predictive power in terms of adjusted-R^2, we obtained this only after removing 70 of the communities that have low non-violent crime rate among the total of 343 communities. Therefore, it is important to note that our model might not be a good fit for the communities with very low rate of nonviolent crimes.

Conclusion:

In sum, we created two models that aimed to analyze the factors affecting the rates of crimes, both violent and nonviolent. We created these models in order to understand how specific characteristics of communities affect crime rates. We found that in the two models, there were overlapping predictors such as large households and high school graduation rates. Most interestingly, we found that the racial composition of the community was a significant predictor in our violent crime model but not in our nonviolent crime model. Moreover, since there are many confounding factors such as the occupations of community members which is associated with crime rate and some predictors, we were not able to establishing causal links between our response variables and predictors. We believe that Logistic Regression analysis would have been a better resource for modeling our data set and could have improved the explanatory power of our models. Therefore, we believe that further research and different modeling tools are needed before we could recommend any policy changes to reduce crime rates.