50 Most Dangerous Colleges

Albert Wong
Joshua Gordon

December 4, 2011

1 Introduction

In the following pages we will analyze data on the 50 most dangerous colleges in the United States. We will discuss our data selection process, variable selection, methods, results, and the limitations of our analysis. We were motivated to pursue this analysis because UCLA is a national leader in academia and sports. We were surprised to find that UCLA is also a leader in safety, as we do not appear on this list. Nevertheless, we are interested in investigating the relationship of certain violent crimes at the 50 most dangerous universities in the United States.

2 Data Selection

We found a data set detailing the 50 most dangerous colleges online from a free data aggregator called info chimps. The data was collected and compiled by the Daily Beast. The Daily Beast is an American news reporting and opinion website. To collect the data, staff of the Daily Beast "pored over the three most recent calendar years of campus security and crime data (2006-2008) compiled by the U.S. Department of Education, as well as the FBI and the Secret Service, in conjunction with the Clery Act, the federal mandate requiring all schools that receive federal funding to disclose crime information annually." [1] A quick web search shows that data set is heavily cited around the web. While a data set exists showing the 50 safest colleges in the United States, it is quiet boring to analyze due to a lack of observations.

We do not consider the most dangerous rank of each school in our analysis but it would be interesting to look at. Among the UC's, only Berkeley and Riverside made it on to the list. Other prestigious schools such as Harvard, Stanford, Columbia, Brown, and Yale all made it onto this list! While we sometimes hear stories in the media of crimes occurring on college campuses, we have no idea that it was taking place at this extreme level at some of America's top universities.

The data itself is split into nine categories when ignoring name and the rank provided by the Daily Beast. The columns are as follows: Murder, Negligent Homicide, Forcible Rape, Non-Forcible Rape, Robbery, Aggravated Assault, Burglary, Car Theft, and Arson. These are all violent crimes,

[1] Fastenberg, Dan. "The Other College Rankings: Most Dangerous U.S. Campuses." Time, 17 Sept. 2010. Web. 04 Dec. 2011. http://newsfeed.time.com/2010/09/17/the-other-college-rankings-most-dangerous-u-s-campuses/.

some far worse than others. Before we continue, an important distinction is required. Forcible Rape exists in the presence of violence whereas Non Forcible Rape is not violent. For example, statutory rape is an example of Non-Forcible Rape. Another important distinction is between Burglary and Robbery. According to laws.com, burglary results when a person enters a structure without permission by breaking and entering. The offender does not necessarily have to participate in theft. Robbery is the result of theft of money or property through violence. With burglary the victim will never encounter the perpetrator but in robbery, the victim must be present. [2]

A preliminary examination of the data reveals three key details. First and most obvious, is that the Virginia Tech massacre is accounted for in the murders column. This is an extreme outlier when compared to the rest of the data. As will be discussed in detail below,the murder column is not statistically significant in our analysis. Second, there is little or no data in the Non-Forcible rape column. Third, interestingly there are no Negligent homicides at any of the 50 most dangerous universities.

Given these details, we are left with eight categories of data to analyze. We decided that we wanted to use our data to predict Forcible Rape. This made sense for a few reasons. Its frequency of occurrence as well as how often is highlighted in the media. It is the worst and most extreme variable left in our analysis. There is data from nearly every school. And due to a lack of reported rapes in general, we can be fairly certain that any estimate that we provide will be extremely conservative.

3 Summary Statistics

In the following section we provide summary statistics of all categories.

The range of forcible rapes among the 50 most dangerous universities is 0 to 128. The standard deviation is relatively high at 24 Forcible Rapes with a mean of only 22.5. We are 95% confident that the true mean lies between 15.653 and 29.347 Forcible Rapes per institution, with 49 degrees of freedom.

Variable	Mean	Standard Deviation	Range	95% CI
Murder	1.12	4.675773	0 to 33	(-0.2088401, 2.4488401)
Negligent Homicide	0	0	0	N/A
Non-Forcible Rape	0.12	0.435187	0 to 2	(-0.003678787, 0.243678787)
Robbery	50.36	47.81288	0 to 192	(36.77173, 63.94827)
Aggravated Assault	41.46	41.09978	1 to 225	(29.77957, 53.14043)
Burglary	213.5	178.216	17 to 909	(162.8516, 264.1484)
Car Theft	68.94	77.93048	0 to 384	(46.7924, 91.0876)
Arson	5.52	9.118808	0 to 54	(2.928464, 8.111536)

[2] "Burglary Vs Robbery." Criminal Laws -From Assault to Robbery Criminal Laws. Web. 04 Dec. 2011. http://criminal.laws.com/burglary/burglary-vs-robbery.

4 Methods

To explain the data, we will use simple linear regression, kernel density estimation, beta removal analysis, and kernel regression. Simple linear regression aids our analysis by helping in variable selection for our model as well as residual analysis. Kernel density estimation helps us understand the spread of our data and its distribution. Beta removal lets us know which schools in our analysis have the most weight or leverage on our regression. Finally, Kernel regression allows us to detect non-linear relationships between our variables.

4.1 Linear Regression

We first begin with a model using Forcible Rapes as our response variable and Murder, Non-Forcible Rape, Robbery, Burglary, Aggravated Assault, Car Theft, and Arson as our predictors. We find that Aggravated Assault, Burglaries, Non-Forcible Rape, and Arson are statistically significant at the 5% level (See Appendix A.1 For Regression Summary). Murder, Robbery, and Car Theft are not significant at the 10% level (Car Theft has a p-value of .128) and we did not include Negligent Homicides due to a lack of any such occurrences in our data.

For our second model, we removed Murders, Robbery, and Car Theft. An interesting result is that Aggravated Assault has a p-value that is basically zero (See Appendix A.2 For Regression Summary). This leads us to believe that something else may be going on. We will analyze below what is happening. Removing Car Theft from our model, we find that the remaining variables are significant at the 5% level. Our final model is as follows: $F orcib\hat{l}eRape$ = -6.43899 + 9.89144 Non Forcible Rape +.44073 Aggravated Assault + .03024 Burglary + .547677 Arson. The adjusted R-squared suggest that about 80 percent of the variation is accounted in the model. In practice, this implies that holding all else constant, each Non Forcible Rape adds 9.89144 Forcible Rape to a college. The other variables can be interpreted in a similar fashion.

Analyzing the correlation between Aggravated Assault and Forcible Rape shows that there is an 81.54% correlation between them! This may be due to the fact that when someone is charged with Forcible Rape, they are more likely to be charged also with Aggravated Assault. We will further analyze this relationship in the kernel regression section 4.5.

4.2 Residual Analysis

The next step is to analyze the residuals of our final model. A plot of the residuals vs fitted values shows a relatively constant variance (See Appendix A.3 For Plots). There are a few outliers but they are not of that much concern because it looks pretty random. The normal QQ plot shows a relatively normal distribution of the residuals as indicated by the near straight line.

4.3 Kernel Density Estimation

In this part of analysis, we would like to see how the shape of the distribution and the 95 percent confidence bound for the Forcible Rape incidents in the 50 colleges listed in the data set. The density estimate obtained by kernel density estimation technique assumes a bell-shaped curve similar to the Gaussian distribution except for a bump at the far right of the distribution (See Appendix A.4 For Plot). The estimated distribution shows that it is quiet common for the schools listed to have about 0 to 40 forcible rape incidents over the three calendar years. Looking back at the data, we noticed that there are two schools, Harvard and University of Michigan, with very high numbers of forcible rape incidents, 128 and 119 respectively. Obviously these two schools provide the explanation for the bump we observed in the estimated density.

Given that we only find these two unusual occurrences among the total of 50 observations, we believe that it is pretty safe to consider those as outliers. With those two schools removed, it seems that one may safely infer that the distribution is approximately normal and centered around 20. Despite the pretty high number of the Forceful Rape incidents, it seems that this number is not too depressing if one remembers the relatively large sizes of the schools in our data set and if one also takes into a consideration that this data is taken over a three-year period.

4.4 Simulation of Burglary(i) removal from model

We are also interested in analyzing how influential each school in our simple multivariate regression. In particular, we wanted to identify influential schools on the regression estimate of the coefficient of burglaries in our final model. The result shows that most but 7 of the schools in our dataset exhibit no or very little influence on the estimated burglaries coefficient (See Appendix A.5 For Plot).

We observed that schools with unusually high or low numbers in burglary incidents do not necessarily suggest that the corresponding schools would show much influence on our estimated burglaries coefficient. We also noticed that the direction of the influence (positive or negative) is not so easy to determine. For example, the numbers of burglary incidents in both Harvard and Temple Universities are both much higher than the mean (909 and 747, respectively), but our graph suggests that removing Harvard University from our data has a negative influence on our fitted burglary coefficient while removing Temple University has a positive influence. This is not surprising, however, since we know that leverage points of burglaries (those points that are far from the mean) would not necessarily imply that they would have large effect on the outcome of the fitted value of our burglary coefficient.

4.5 Kernel Regression

In this following section, we will perform multiple kernel regression estimates between our variables (See Appendix A.6 For Plot). The goal is to estimate the non-linear relationships between our predictors and our response with a 95% confidence bound.

First, we looked at the relationship between Forcible Rape and Robberies. The relationship is generally positive. However, there is not that large a trend of increase which possibly indicates that Robberies do not have that significant an effect on Forcible Rape.

Next, we looked at the relationship between Car Theft and Forcible Rape. As expected, in the initial area of interest, we see a positive increasing relationship. However, as Car Thefts increase, the number of Forcible Rapes actually tends to decrease. Both universities with over 200 Car Thefts do not have noticeably more Forcible Rapes than most universities below 200.

Comparing Aggravated Assaults to Forcible Rape, the region of interest is between 0 and 100 Aggravated Assaults. Outside of this region, the regression estimate becomes unreliable There is a relatively clear linear relationship between Aggravated Assaults and Forcible Rape. However, it appears that Forcible Rapes taper off in comparison to Aggravated Assaults. Aggravated Assaults continue to increase while Forcible Rape does not.

Burglaries and Forcible Rapes also shows a generally positive increase initially. But as Burglaries increase, Forcible Rapes do not increase that dramatically. In fact, at over 400 Burglaries , the instances of Forcible Rape tend to decrease in general. This result may explain why the coefficient of Burglary is so small in our final model.

5 Sensitivity Analysis and Future Research

One concern is that 80-90% of rapes go unreported! [3]. This means that any estimate that we produce is without a significant underestimate of the true number of rapes occurring. If some schools have a higher report rate, then this could potentially skew our analysis. We hope that this lack of available data does not invalidate our analysis. It is also possible that other crimes are underestimated.

As mentioned previously, the Virginia Tech Massacre is included in the data set, which presents us with a significant outlier when compared to other schools. An interesting research pursuit would be to compare nation wide FBI data to see the robustness of our conclusions. It would be interesting to see if the academic rank had an effect on the crime rate, or the presence of greek life, or the relative rate of women to men at a university. In addition, we would like to see if non violent crimes, such as alcohol and drug abuse, could be used to predict violent crimes at universities.

6 Conclusion

The use of simple linear regression, kernel density estimation, beta removal analysis, and kernel regression have given us a better idea of the relationship between violent crimes on college campuses. One needs to take all reasonable precautions to protect oneself since criminal activity happens even at the most prestigious universities in the United States.

[3] Storck, MD, Susan. ”Rape: MedlinePlus Medical Encyclopedia.” National Library of Medicine - National Institutes of Health. 30 Mar. 2010. Web. 04 Dec. 2011. http://www.nlm.nih.gov/medlineplus/ency/article/001955.htm.

Appendices

A Graphics

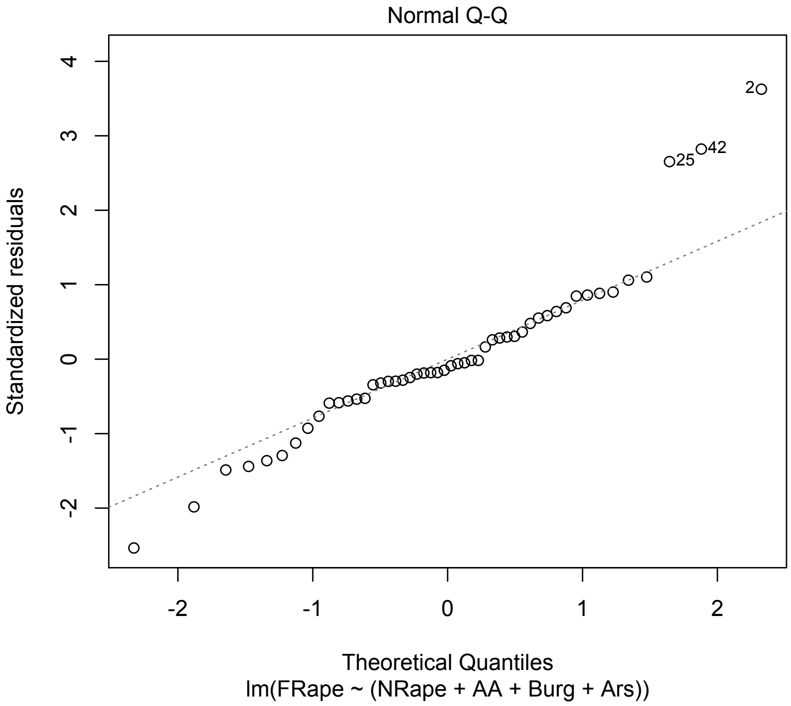
A.1 First Model

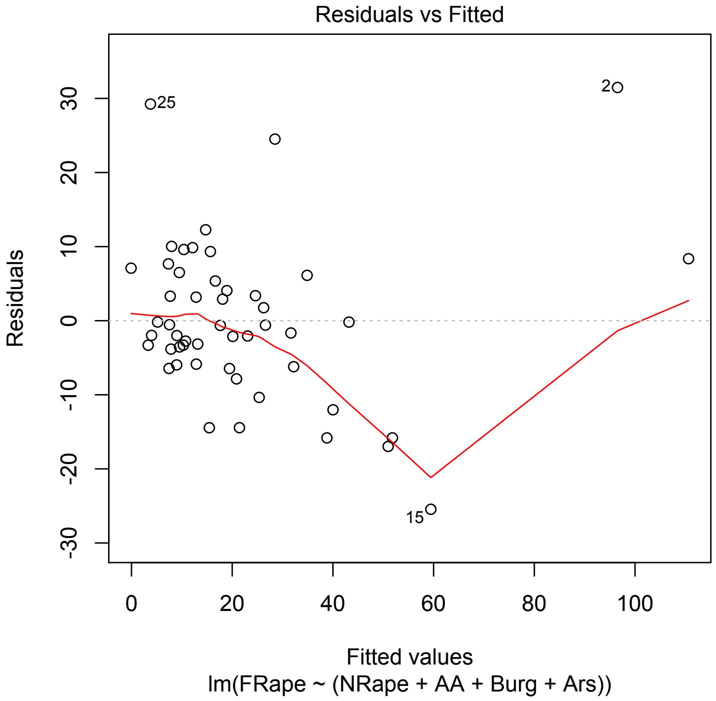
```
Call:
lm(formula = FRape ˜ (Mur + NRape + Rob + AA + Burg + Car + Ars))

Residuals:
    Min      1Q  Median      3Q     Max 
-22.956  -4.735  -1.313   3.998  30.690 

Coefficients:
            Estimate Std. Error t value        Pr(>|t|)    
(Intercept) -7.82795    3.23808  -2.417         0.02005 *  
Mur          0.18087    0.35119   0.515         0.60925    
NRape        9.43085    3.93922   2.394         0.02120 *  
Rob         -0.01358    0.04462  -0.304         0.76229    
AA           0.42689    0.04988   8.559 0.0000000000942 ***
Burg         0.03119    0.01002   3.112         0.00334 ** 
Car          0.03546    0.02286   1.551         0.12840    
Ars          0.52100    0.18341   2.841         0.00691 ** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.38 on 42 degrees of freedom
Multiple R-squared: 0.8089,       Adjusted R-squared: 0.777 
F-statistic: 25.39 on 7 and 42 DF,  p-value: 0.0000000000003627
```

A.2 Second Model

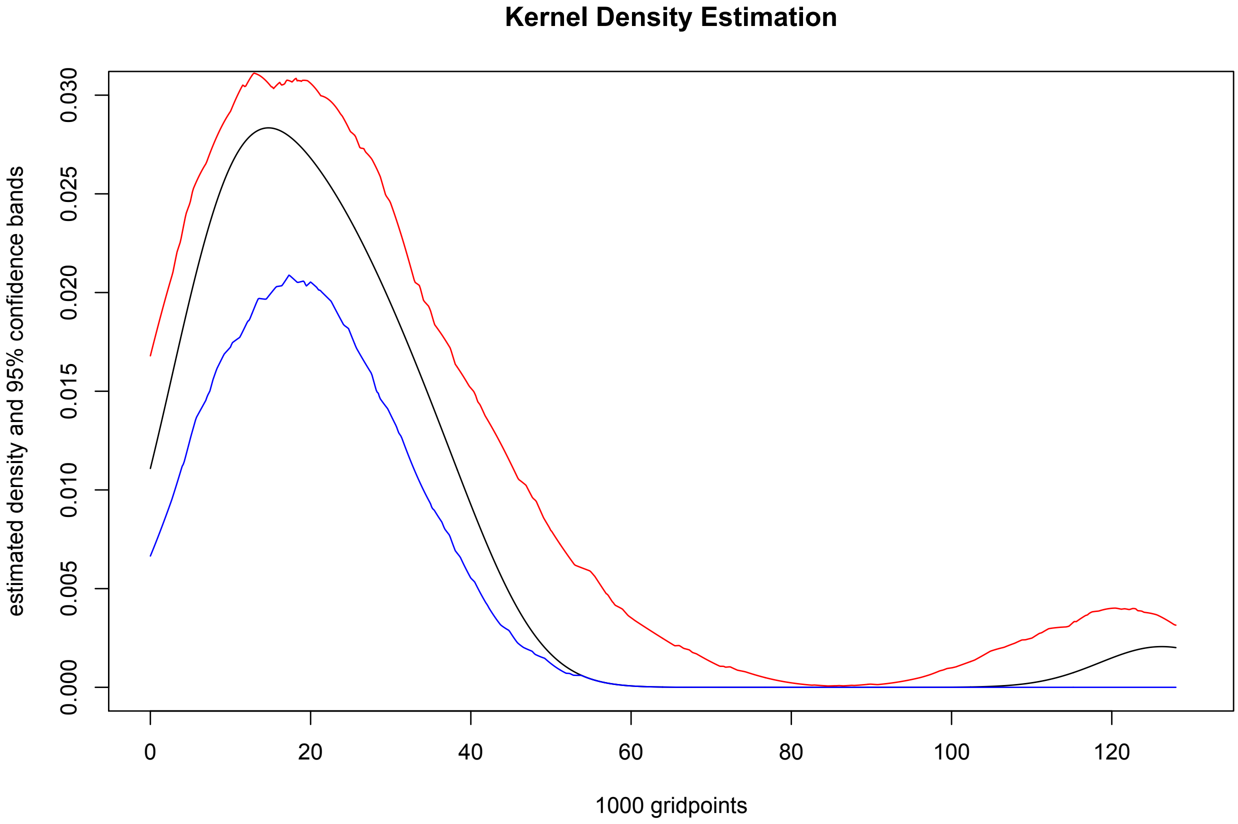
```
Call:
lm(formula = FRape ˜ (NRape + AA + Burg + Ars))

Residuals:
    Min      1Q  Median      3Q     Max 
-25.451  -5.942  -1.147   5.931  31.477 

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)    
(Intercept) -6.439852   2.904833  -2.217           0.03172 *  
NRape        9.891442   3.805954   2.599           0.01260 *  
AA           0.440728   0.042623  10.340 0.000000000000181 ***
Burg         0.030244   0.009791   3.089           0.00344 ** 
Ars          0.547677   0.178412   3.070           0.00362 ** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.32 on 45 degrees of freedom
Multiple R-squared: 0.7972,       Adjusted R-squared: 0.7792 
F-statistic: 44.23 on 4 and 45 DF,  p-value: 4.85e-15
```

A.3 Residual Analysis

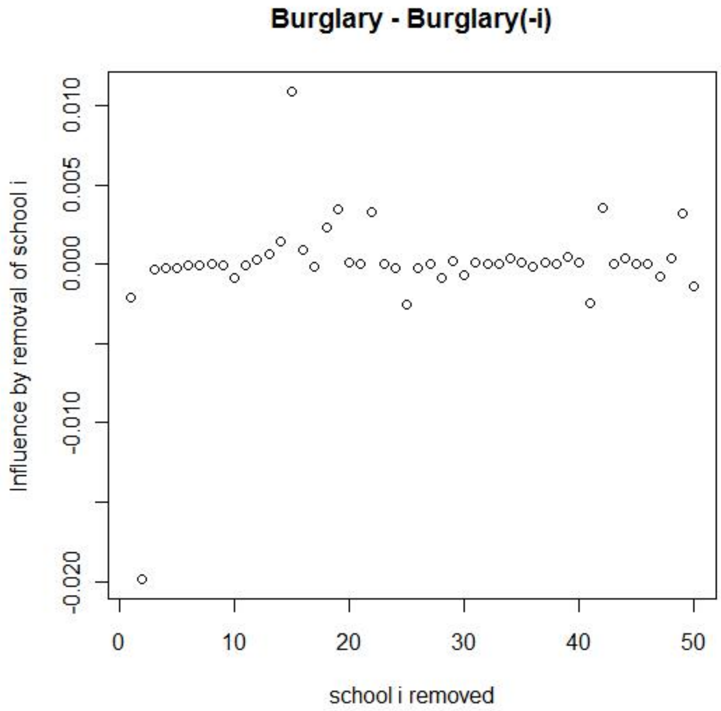




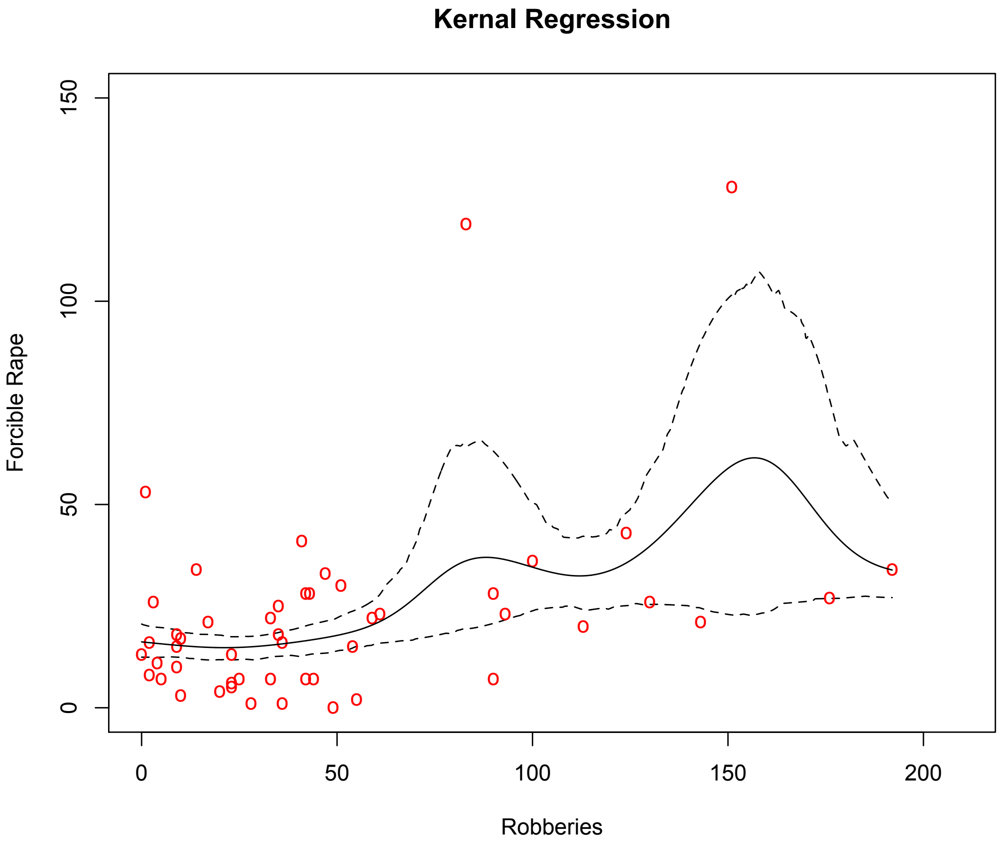
A.4 Kernel Density

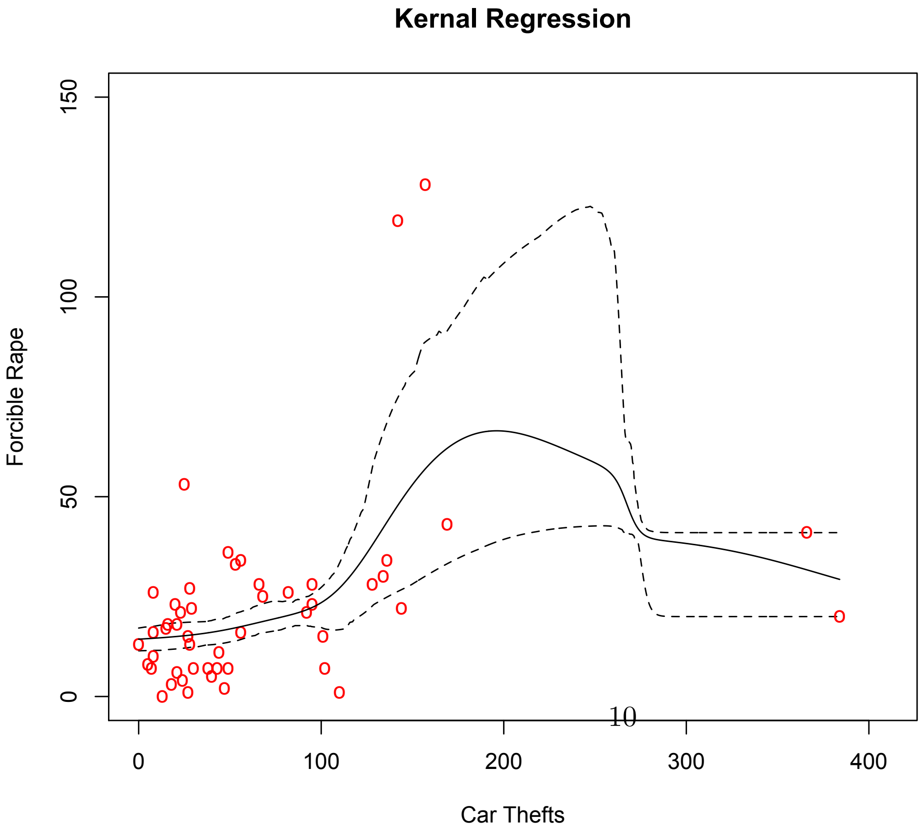


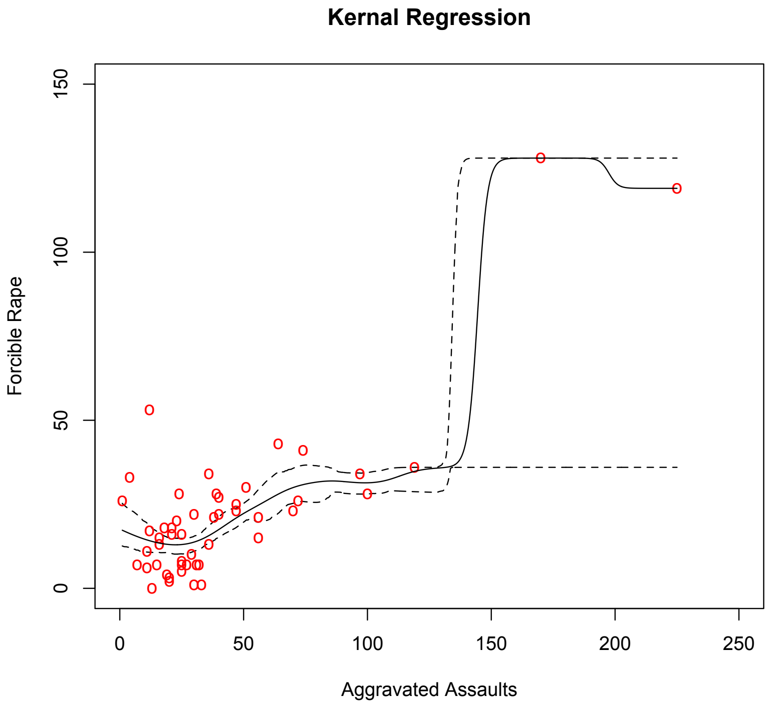
A.5 Influence

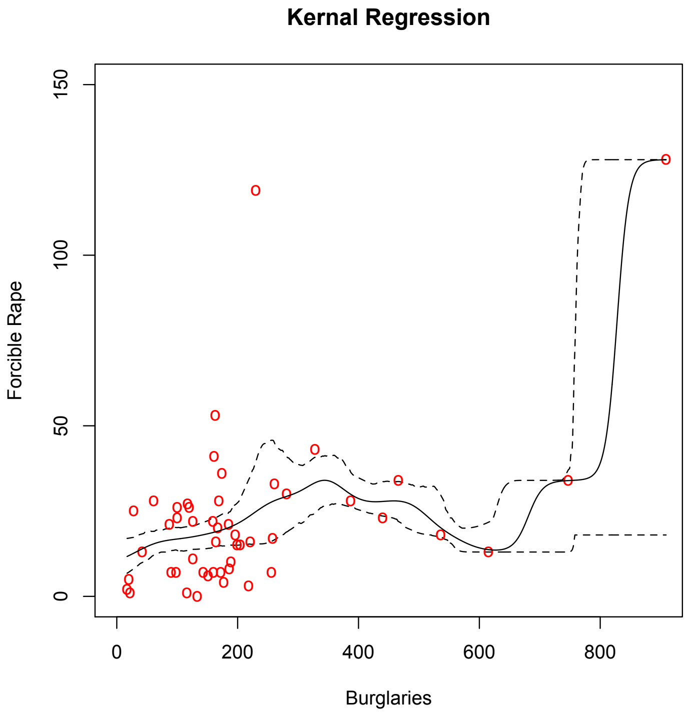


A.6 Kernel Regression






Kernal Regression
Forcible Rape
0
50
100
150
0
50
100
150
200
250
Aggravated Assaults


Kernal Regression
Forcible Rape
0
50
100
150
0
200
400
600
800
Burglaries