MAT 453 Regression Analysis

Dr. Olcay Akman

**Term Project**

**Study of Impact Factors on the Number of Violent Crimes in the United States**

Macie Hanneken

mehanne@ilstu.edu

Tanja Kohlruss

tkohlru@ilstu.edu

Anja Kohlschreiber

akohlsc@ilstu.edu

**December 1, 2017**

**Abstract**

In 2016 alone, there were 1,248,185 violent crimes reported across the United States. This was a 4.1 percent increase from the reported crimes in 2015 (Federal Bureau of Investigation 2017). Whenever these violent crimes occur, the media is quick to show the public the story in the way that best suits the media’s interest, which is often biased. This can lead to general biases in the population that may or may not have already been there. These biases, whether they would be from the media or not, influence parts of the population to believe that certain factors such as race and socioeconomic status are responsible for high crime rates. With the data from the 2007 U.S census and methods in multiple regression analysis we determine if race, socioeconomic factors, education, wealth, cultural background, family history, government expenditures, and other population related factors have a significant impact on the number of violent crimes in the United States. The goal of this study is to find the factors that have the greatest impact on the number of violent crimes in the United States. This information will be helpful in future crime prevention tactics. Furthermore, we hope to invalidate some common prejudices that are prevalent in the general population.

**Introduction**

Need more. There are four different types of crime that are encompassed in the violent crime category: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. These violent crimes are plaguing our communities and any effort to reduce them should be considered and implemented. Without any data on what factors are contributing to these crimes, understanding and implementing procedures to reduce these violent crimes is impossible. With this study, we hope to uncover some of those factors that influence violent crime, both positively and negatively. We aim to better understand the impacts of violent crime in the United States and how we might be able to implement different programs or policies in high risk areas to reduce the violent crime rate.

**Methods**

In order to address the question as to whether or not certain population factors are impacting crime rates in the United States, we retrieved data from the 2007 United States Census. After retrieving this data, we selected factors from those categories collected that we thought would have an impact on violent crime in the United States (Table1), these include: race, education, government spending, health care, presidential vote, median age, income, and race. After choosing these variables that we assumed may have an impact on crime, we decided to create three different models in order to see which one had the best fit and predictive power. Those models will be denoted as the Variance Inflation Factor Model, the Partial Correlation Model, and the Neural Network Model. The first two models were chosen and named after the method used to reduce and eliminate multicollinearity in the models. Since we are dealing with socioeconomic factors, we assumed that there would be a fair amount of multicollinearity between variables. In order to verify that assumption, before we began any variable selection methods, we checked the variance inflation factor (VIF) for each of our independent variables. According to Montgomery et al. (2012) a variance inflation factor larger than 10 indicates that the regression coefficients are not well estimated due to multicollinearity. All of our independent variables had a VIF well over ten, proving our assumption right. Therefore, we chose to deal with the multicollinearity in multiple different ways in order to increase our chances of obtaining the best model. For all of the models, we first used R to split our data into three different sets: Training, Test, and Validation. We split our data into three sets, because we also use a neural network model, so in order to accurately compare models, we wanted to use the same data sets for each of the different models.

*Variance Inflation Factor Model*

To address the confirmed multicollinearity issue, using the variance inflation factor (VIF) model, we first used backwards elimination, a method of variable selection. Instead of using the p-value or AIC though, we eliminated variables based on which ones had the largest VIF. We continued with this variable selection until all variables included in the model had a VIF that was below ten. After we eliminated all variables with a VIF greater than ten, we then ran a t-test to determine the significance of each variable that was in the model. We then continued using backwards elimination, now using the p-values, until all non-significant variables were removed from the model. Non-significance meaning that p > .05.

*Partial Correlation Model*

The next model was built using a partial correlation matrix. The partial correlation matrix was used in order to determine the correlation between the independent variables in the model. We began with all 36 variables, and then used a partial correlation matrix to begin eliminating those most correlated with each other. According to Dr. Akman, no two variables should be more than 75% correlated. With this information, we found which independent variables occurred the most in the partial correlation matrix that were more than 75% correlated with other variables. The variable that occurred the most for each round of elimination was deleted. When each variable only began to appear once, the variable with the highest values of multicollinearity was deleted. After all variables were correlated less than 75%, we used backwards elimination, using p-values, to eliminate any non-significant variables that were still left in the model. Non-significance in this case is described as p > .05. Just as in the VIF model, these p-values were determined using a t-test.

*Neural Network Model*

After the partial correlation and the VIF model were built, we needed to see which model had the best fit, and the best predictive power. To do this, we compared r-squared values, adjusted r-squared values, F-statistic, p-value, and the PRESS statistic for each of the models. After we determined which model was best for our data, we needed to test the assumptions of multiple regression: a linear relationship between the independent and response variables, the residuals must be normally distributed, constant error variance, and no multicollinearity. To test for a linear relationship between the independent and response variables, we used component residual plots which can be viewed in Figure 1 and Figure 2, for the partial correlation model and the VIF model respectively. These plot the response variable against each of the independent variables. The red and green lines represent the component and the residual lines. If there is a significant difference between the two lines, this shows that there is not a linear relationship between the given independent variable, and the response variable. Figures 1 and 2 are the component residual plots for the partial correlation model, and the VIF model, respectively. After viewing these figures, it is noticeable that several of the independent variables do not have a linear relationship with the response. This would normally call for a transformation of the variables that are showing a non-linear relationship with the response variable. In order to test the second assumption, that the residuals are normally distributed, we constructed a QQ-plot for each model to determine their normality. If the data points are close to the regression line, then we can assume normality in the data. If we look at Figure 3, we can see that for both the VIF model and the partial correlation model, the data points tend to follow the regression line quite well outside of the tails. This implies that the data is normally distributed and this assumption is met. The next assumption is constant error variance. To determine if our models had constant error variance, we plotted the studentized residuals against the fitted values of the model. We chose to use the studentized residuals versus the residuals themselves, because when using the studentized residuals, we may use the same plot to help with outlier detection. If a data point is greater than three or negative three, it may be flagged as a potential outlier and will require further analysis. In order for the error variance to be constant, we should expect to see no trend in the data points, and no particular shapes arising from the points either. In Figure 4, we can see that for the partial correlation model, the data points seem well spread out with no particular trend or shape. This leads us to believe that this model has constant error variance. The VIF model, also in Figure 4, seems to have data points that increase as the fitted values increase. This would not allow the VIF model to meet the assumption of constant variance and just as with the linearity issue, we would need to transform the data in order to meet the assumption. As far as outlier detection goes, Figure 4 clearly shows that both models have potential outliers. These points would need to be flagged for further analysis to determine if they truly are outliers and if so, how to treat them. The last assumption is that there is no multicollinearity in the model. We initially took care of this problem when we set out to make the models. We did this by using the partial correlation matrix and ensuring that no two variables were correlated more than 75% (Dr. Akman). For the model where we used the variance inflation factor, we took care of the multicollinearity issue by using the VIF as described above.

Due to the number of variables in this model, we did not make any data transformations in order to meet the assumptions. The results presented below are given with that in mind. If this was for something outside of the classroom, all assumption would have to be met before running the regression analysis and making any complete conclusions.

**Results**

*Variance Inflation Factor Model*

After the analysis was done on the VIF model, we found that the fitted VIF model is as follows: . Table 2 shows the coefficient analysis for the variables in the given model, using the t-test for significance. The model quality factors, such as PRESS, r-squared, adjusted r-squared, F statistic, and p-value can be found in Table 3. The ANOVA table for the model, Table 4, shows the significance level of each of the independence variables, as well as other statistics. The ANOVA analysis uses the F-test to calculate the significance of variables, whereas the summary uses the t-test. This is why x33 appears to not be significant in the ANOVA table, but was found significant for our analysis (see Table 2). This model found factors such as: Median age, vote cast for president, government spending, health insurance, income, and race to have a significant impact on violent crime rates. Of these factors, median household income, and resident Asian population had a negative significant impact on crime, with the resident Asian population having the largest negative significant impact on violent crime rates. Factors such as: median age of the population, vote cast for president, government spending, people under 18 without health insurance, and the American Indian, Alaska native, Black, and Pacific Island races having a significant positive impact on violent crime. The factor with the largest, positive significant impact on crime is median age of the population, the coefficient of this variable is over 100 times larger than that of the other variables, implying that it has a very large impact on violent crime rates.

*Partial Correlation Model*

After the analysis was done on the partial correlation model, we found that the fitted partial correlation model is:

. Table 5 shows the coefficient analysis for the variables in this model. This analysis used the t-test for significance. The ANOVA table for this model can be found in Table 6. The ANOVA table shows the significance of each independent variables, as well as other statistics for the model. The significance values for the ANOVA (Table 6) and the summary of coefficients (Table 5) vary because the summary uses the t-test to determine significance whereas the ANOVA uses the F-test. This is why x13, x21, and x22 are included in our model, even though the ANOVA table shows them to be non-significant. This model found factors such as: school enrollment, labor force, vote cast, health insurance, income, marital status, and race significant. Of these, school enrollment, presidential vote cast, lack of healthcare for children under 18, median household income, and the Asian population had a negative impact on crime. With females enrolled in graduate or professional school being the factor to most negatively impact violent crime rate. Things such as employment status, lack of healthcare for people over 18, personal income, females never married or divorced, and the Black, Pacific Islander, and Hispanic population all having a significant positive effect on violent crime. With divorced females having the largest positive impact on violent crime rates.

*Model Comparison*

Table 2 shows the r-squared values, adjusted r-squared values, F-statistic, p-value, and the PRESS statistic for the variance inflation factor model and the partial correlation model. If you look at the values for the model quality determinants, we can see that the r-squared and adjusted r-squared values are not nearly as good for the variance inflation factor model as they are for the partial correlation model. Although, the F-statistic and PRESS statistic for the variance inflation factor model are much better than those of the partial correlation model. The PRESS statistic gives the predictive power of the model, so this implies that the variance inflation factor will better predict other data making it the better model. We can visually see this if we compare the two different plots in Figure 5. We can see that the first plot in Figure 1, the partial correlation model, may be effected by a potential outlier. Although no official analysis was done, it appears if we were to take out this potential outlier, the slope of the line may change, making the point an influential point. The second plot in Figure 1 is the for the VIF model. It appears to predict the data better. Although, it has the same potential outlier, the model seems to be much less affected by it, or even unaffected. The results from the PRESS statistic for the two models reiterates this claim.

As far as the variables included in each model, Table 1 has the list of variables included in each model highlighted in different colors for easy comparison. The two models contain almost completely different variables but do overlap in a few areas, such as health insurance, median household income, and race. Based on the model quality statistics, it appears that the VIF model is the best model we built.

The neural network model…

**Discussion**

**Appendix**

Table 1. Description of the independent variables being used in this project. Item Id is the name of the variable in the data set, Item Description explains the variable and unit explains the unit that the data was collected in. ABS is the absolute value of the number, YRS is years, and TH$ is thousands of dollars. Values highlighted in yellow are those of the VIF model, those highlighted in green are in the partial correlation model, and those highlighted in purple are in both models.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Item Id | Item Description | Unit |
| y | CRM110207D | Number of violent crimes known to police 2007 | ABS |
| x1 | AGE050210D | Resident population: Median age (April 1 - complete count) 2010 | YRS |
| x2 | CLF020207D | Civilian labor force employment 2007 | ABS |
| x3 | CLF030207D | Civilian labor force unemployment 2007 | ABS |
| x4 | EDU430209D | School enrollment - persons 3 years and over, male, enrolled in grades 9 to 12: 2005-2009 | ABS |
| x5 | EDU436209D | School enrollment - persons 3 years and over, male, enrolled in college, undergraduate years: 2005-2009 | ABS |
| x6 | EDU442209D | School enrollment - persons 3 years and over, male, enrolled in graduate or professional school: 2005-2009 | ABS |
| x7 | EDU448209D | School enrollment - persons 3 years and over, male, not enrolled in school: 2005-2009 | ABS |
| x8 | EDU478209D | School enrollment - persons 3 years and over, female, enrolled in grades 9 to 12: 2005-2009 | ABS |
| x9 | EDU484209D | School enrollment - persons 3 years and over, female, enrolled in college, undergraduate years: 2005-2009 | ABS |
| x10 | EDU490209D | School enrollment - persons 3 years and over, female, enrolled in graduate or professional school: 2005-2009 | ABS |
| x11 | EDU496209D | School enrollment - persons 3 years and over, female, not enrolled in school: 2005-2009 | ABS |
| x12 | ELE010208D | Vote cast for president - total 2008 \*\*\*Subject to copyright\*\*\* | ABS |
| x13 | ELE020208D | Vote cast for president - Democratic 2008 \*\*\*Subject to copyright\*\*\* | ABS |
| x14 | ELE030208D | Vote cast for president - Republican 2008 \*\*\*Subject to copyright\*\*\* | ABS |
| x15 | FED110207D | Federal Government expenditure - total FY 2007 | TH$ |
| x16 | HEA730207D | All persons 18 to 64 years without health insurance 2007 | ABS |
| x17 | HEA710207D | All persons under 18 years without health insurance 2007 | ABS |
| x18 | IPE010207D | Median household income 2007 | DOL |
| x19 | IPE110207D | People of all ages in poverty - number 2007 | ABS |
| x20 | IPE310207D | Related children age 5 to 17 in families in poverty - number 2007 | ABS |
| x21 | IRS140207D | Wages and salaries income 2007 | TH$ |
| x22 | PEN010207D | Personal income 2007 | ML$ |
| x23 | POP430209D | Males 15 years and over - total 2005-2009 | ABS |
| x24 | POP440209D | Males 15 years and over - never married 2005-2009 | ABS |
| x25 | POP480209D | Males 15 years and over - divorced 2005-2009 | ABS |
| x26 | POP530209D | Females 15 years and over - total 2005-2009 | ABS |
| x27 | POP540209D | Females 15 years and over - never married 2005-2009 | ABS |
| x28 | POP580209D | Females 15 years and over - divorced 2005-2009 | ABS |
| x29 | PST045207D | Resident total population estimate (July 1) 2007 | ABS |
| x30 | PST150208D | Components of change - births for July 1, 2007 to July 1, 2008 | ABS |
| x31 | RHI120207D | Resident population: White alone (July 1 - estimate) 2007 | ABS |
| x32 | RHI220207D | Resident population: Black alone (July 1 - estimate) 2007 | ABS |
| x33 | RHI320207D | Resident population: American Indian and Alaska Native alone (July 1 - estimate) 2007 | ABS |
| x34 | RHI420207D | Resident population: Asian alone (July 1 - estimate) 2007 | ABS |
| x35 | RHI520207D | Resident population: Native Hawaiian and Other Pacific Islander alone (July 1 - estimate) 2007 | ABS |
| x36 | RHI720207D | Resident population: Hispanic or Latino Origin (July 1 - estimate) 2007 | ABS |

Table 2. The summary of the coefficient analysis for the VIF model. This summary used the t-test to test for significance of variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | t-value | Pr >| t | |
| Intercept | -18.13 | 146.8 | -0.124 | 0.901715 |
| x1 | 6.498 | 3.029 | 2.145 | 0.032171 |
| x14 | .009901 | .0008241 | 12.015 | < 2e-16 |
| x15 | .0001174 | .00001071 | 10.962 | < 2e-16 |
| x17 | .05943 | .003025 | 19.647 | < 2e-16 |
| x18 | -.009274 | .001577 | -5.879 | 5.55e-09 |
| x32 | .01566 | .0006788 | 23.072 | < 2e-16 |
| x33 | .01827 | .005183 | 3.525 | 0.000442 |
| x34 | -.02018 | .0008894 | -22.693 | < 2e-16 |
| x35 | .178 | .01859 | 9.574 | < 2e-16 |

Table 3. The Variance Inflation Factor (VIF) model and constituents that give an idea of model quality.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variance Inflation Factor Model | Partial Correlation Model | Neural Network Model |
|  | .9225 | .9573 |  |
| Adjusted | .9218 | .9566 |  |
| F-Statistic | 1371 | 1281 |  |
| PRESS Statistic | 455275482 | 524552007 |  |
| p-value | < 2.2e-16 | < 2.2e-16 |  |

Table 4. Analysis of variance (ANOVA) table for the VIF model. x33 was not initially removed because we used model summary to eliminate non-significant variables which uses the t-test versus the ANOVA that uses the F-test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum of Squares | Mean Sum of Squares | F Value | Pr(>F) |
| x1 | 1 | 117192278 | 117192278 | 500.1501 | < 2.2e-16 |
| x14 | 1 | 1979383307 | 1979383307 | 8447.5604 | < 2.2e-16 |
| x15 | 1 | 181233109 | 181233109 | 773.4619 | < 2.2e-16 |
| x17 | 1 | 392264999 | 392264999 | 1674.0983 | < 2.2e-16 |
| x18 | 1 | 11207360 | 11207360 | 47.8305 | 8.115e-12 |
| x32 | 1 | 87139787 | 87139787 | 371.8929 | < 2.2e-16 |
| x33 | 1 | 718719 | 718719 | 3.0673 | 0.08018 |
| x34 | 1 | 100125014 | 100125014 | 427.3109 | < 2.2e-16 |
| x35 | 1 | 21478941 | 21478941 | 91.6673 | < 2.2e-16 |
| Residuals | 1037 | 242983819 | 234314 |  |  |

Table 5. The summary of the coefficient analysis for the partial correlation model. This summary used the t-test to test for significance of variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | Estimate | Std. Error | t-value | Pr( > | t | ) |
| Intercept | .01603 | .5190 | 3.089 | 0.002064 |
| x2 | .008187 | .002274 | 3.600 | 0.000333 |
| x3 | 7.617e-02 | .01523 | 5.002 | 6.67e-07 |
| x4 | -.1526 | .02305 | -6.618 | 5.86e-11 |
| x9 | -.03314 | .009366 | -3.538 | 0.000420 |
| x10 | -.2791 | .03885 | -7.184 | 1.30e-12 |
| x11 | -.04285 | .003062 | -13.992 | < 2e-16 |
| x13 | -.01014 | .001405 | -7.221 | 1.00e-12 |
| x16 | .01607 | .003055 | 5.259 | 1.76e-07 |
| x17 | -06264 | .008703 | -7.198 | 1.18e-12 |
| x18 | -.004276 | .001254 | -3.410 | 0.000674 |
| x21 | .0002344 | .00003314 | 7.074 | 2.79e-12 |
| x22 | .003461 | .01042 | 3.320 | 0.000933 |
| x27 | .06022 | .006485 | 9.286 | < 2e-16 |
| x28 | .2020 | .009095 | 22.209 | < 2e-16 |
| x32 | .01481 | .0009964 | 14.866 | < 2e-16 |
| x34 | -.02860 | .001156 | -24.731 | < 2e-16 |
| x35 | .1442 | .01577 | 9.145 | < 2e-16 |
| x36 | .007768 | .0008191 | 9.483 | < 2e-16 |

Table 6. ANOVA table for the Partial Correlation model. The highlighted values were not initially removed because we used the t-test to remove non-significant values initially and the ANOVA uses the F-test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | Sum of Squares | Mean Sum of Squares | F-value | Pr(>F) |
| x2 | 1 | 2281790571 | 2281790571 | 17538.4189 | < 2.2e-16 |
| x3 | 1 | 57423419 | 57423419 | 441.3709 | < 2.2e-16 |
| x4 | 1 | 1468608 | 1468608 | 11.2881 | 0.0008087 |
| x9 | 1 | 28853605 | 28853605 | 221.7761 | < 2.2e-16 |
| x10 | 1 | 22489142 | 22489142 | 172.8572 | < 2.2e-16 |
| x11 | 1 | 13850391 | 13850391 | 106.4576 | < 2.2e-16 |
| x13 | 1 | 224044 | 224044 | 1.7221 | 0.1897209 |
| x16 | 1 | 227004390 | 227004390 | 1744.8131 | < 2.2e-16 |
| x17 | 1 | 50011758 | 50011758 | 384.403 | < 2.2e-16 |
| x18 | 1 | 2105687 | 2105687 | 16.1848 | 6.17E-05 |
| x21 | 1 | 3127 | 3127 | 0.024 | 0.8768209 |
| x22 | 1 | 425458 | 425458 | 3.2702 | 0.0708422 |
| x27 | 1 | 34297452 | 34297452 | 263.6189 | < 2.2e-16 |
| x28 | 1 | 163227049 | 163227049 | 1254.6043 | < 2.2e-16 |
| x32 | 1 | 37230484 | 37230484 | 286.1629 | < 2.2e-16 |
| x34 | 1 | 58537968 | 58537968 | 449.9376 | < 2.2e-16 |
| x35 | 1 | 9339036 | 9339036 | 71.7822 | < 2.2e-16 |
| x36 | 1 | 11699865 | 11699865 | 89.9281 | < 2.2e-16 |
| Residuals | 1028 | 133745278 | 130102 |  |  |

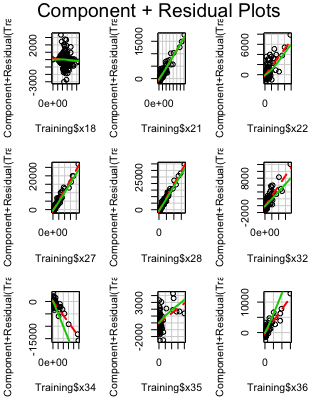
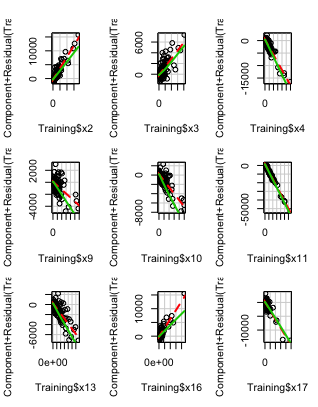
****

Figure 1. The component residual plots for the independent variables in the partial correlation model using the Training data set. These plots will give us an idea of the relationship between the response variable and each independent variable. The green and the red lines are component and residual lines. A large difference between the two lines will imply a nonlinear relationship between the given independent variable, and the response. Some plots that are showing a possibility of non-linearity are: x9, x16, x32, x34, and x36

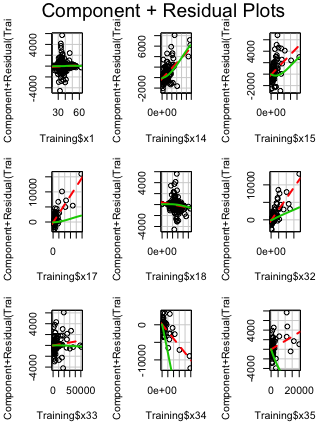


Figure 2. The component residual plots for the independent variables in the Variance Inflation Factor model using the Training data set. These plots will give us an idea of the relationship between the response variable and each independent variable. The green and the red lines are component and residual lines. A large difference between the two lines will imply a nonlinear relationship between the given independent variable, and the response. A few of these plots show a non-linear relationship: x15, x17, x32, x34, and x35.

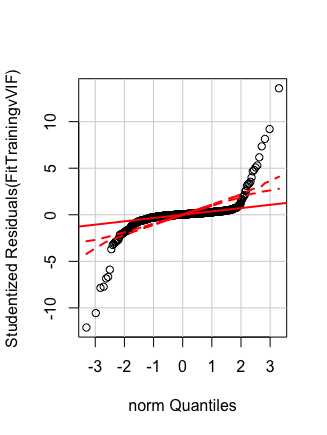
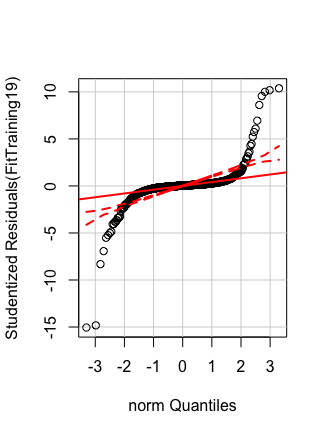
 

Figure 3. The QQ plots for the variance inflation factor model and the partial correlation model, respectively. Look up what dotted lines mean. These QQ plots help us to understand if the data is normally distributed or not. Since all of the data points follow the regression line (the solid red line) besides the tails for both models, we can conclude that the data is normally distributed.

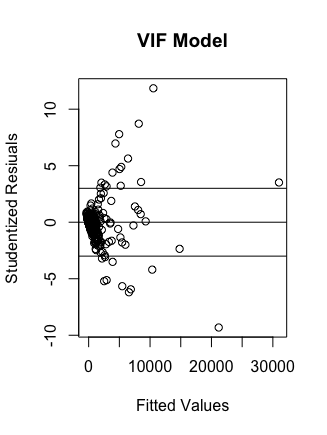


Figure 4. The studentized residuals versus the fitted values for the partial correlation and VIF model. These plots will help us determine if the data has constant variance. If the plots have random data points with no particular shape or trend, we can assume non-constant variance. The lines are at zero, three, and negative three. The markers at three and negative three are to help with outlier detection. Any point past three or negative three needs to be flagged as a possible outlier.

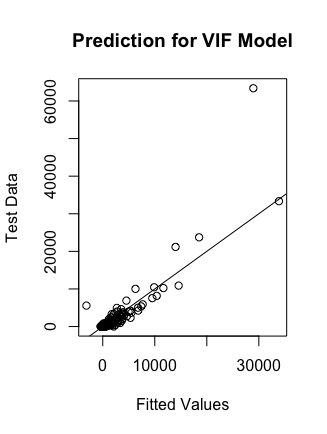
****

Figure 5. Comparing the Test data to the fitted values for the Partial Correlation model and the VIF model, respectively. Although both models have outliers, it appears that the partial correlation model may be more effected by the outlier which may effect its predictive power.

**References:**

Federal Bureau of Investigation. Uniform Crime Report: Crime in the United States, 2016. United States Department of Justice. Web. September 27, 2017.

Montgomery, D.C, Peck, E.A., Vining, G. G. 2012. Introduction to Linear Regression Analysis: Fifth Edition. John Wiley & Sons, Inc. Hoboken, New Jersey.