Regression Analysis
of
Violent Crimes in the United States

Arora, Mansi	Bu, Nan	Kholgade, Nitish
mansi.arora@gatech.edu	nanb@gatech.edu	nitish.kholgade@gatech.edu

ISyE 6414

Georgia Institute of Technology

Professor Nicoleta Serban

December 11th, 2017

Table of Contents

I. Introduction

The Communities and Crime Unnormalized Data Set was taken from the UCI Machine Learning Repository. The data combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law.

Violent crime in the United States refers to murder, homicides, rape and sexual assault, and armed robbery. Such violent crimes severely affect the lives of the victims who survive as well as the families of those who lose their lives. The families of the victims experience physical and emotional challenges, and also face diminished earning potential due to the loss of an earning member.

Violent crime also impose large costs on communities through lower property values, higher insurance premiums, and reduced investment in high-crime areas. In addition, violent crimes impose significant costs on taxpayers, who bear the financial burden of maintaining the police personnel and operations, courts, jails, and prisons directed toward these crimes and their perpetrators.

By most measures, violent crime continues to impose significant costs on Americans and their communities. The costs borne by the American public for such high levels of criminal activity are significant. Moreover, the costs of the pain and suffering borne by the victims of violent crimes is several times greater than the more direct costs of those crimes. As a result, successful efforts to reduce violent crime can produce substantial economic benefits for individuals, communities, and taxpayers.

In today's tight fiscal and economic environment, the leaders of every city, along with state and the federal governments, are searching for ways to use their resources more efficiently to reduce violent crimes. The common challenge is what key factors to focus on. This analysis aims to provide a solution to the problem of violent crime in many communities in the US by identifying those critical factors that lead to high violent crime rates. The major findings of this analysis were to focus on creating employment opportunities to reduce crime, and encourage the importance of family life to improve the well-being of the society.

II. Data Source

The Communities and Crime Unnormalized Data Set was taken from the UCI Machine Learning Repository. The data combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR.

The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units.

The per capita crimes variables were calculated using population values included in the 1995 FBI data (which differ from the 1990 Census values) and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault.

The dataset contains 147 variables and 2215 observations. Among all the variables, there are 4 categorical variables and 125 numeric variables. There are 18 potential crime variables that can be used as response variables. For this analysis, the Per Capita Violent Crimes was taken as the response variable.

The link to the data is:
http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized

Citations

1. U. S. Department of Commerce, Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a (Computer Files),
2. U.S. Department of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)
3. U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management and Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)
4. U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)

III. Methods

a. Lasso Regression

Lasso is a regression analysis method that performs variable selection. Lasso is a least absolute shrinkage and selection operator that is able to improve model prediction accuracy and interpretability of the model it produces. The theory of lasso is that it forces the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero, and thus it produces a simpler reduced model. The algorithm is defined as

$$\boldsymbol{argmin_{\beta}} \sum_{\boldsymbol{i}} (\boldsymbol{y_i} - \boldsymbol{\beta' x_i})^2 + \boldsymbol{\lambda} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} |\boldsymbol{\beta_k}|$$

We can see from the equation that lasso uses L1 norm, and lambda is a tuning parameter that penalize the absolute size of the regression coefficients. The larger the penalty applied, the further estimates are shrunk towards zero.

b. Elastic net Regression

Elastic net is a regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. Elastic net algorithm is defined as

$$argmin_{\beta} \sum_{i} (y_i - \boldsymbol{\beta' x}_i)^2 + \lambda_1 \sum_{k=1}^{K} |\beta_k| + \lambda_2 \sum_{k=1}^{K} \beta_k^2$$

The equation shows that elastic net penalizes the size of the regression coefficients based on both their L1 norm and their L2 norm, where L1 norm penalty generates a sparse model (force some betas to be zero) and L2 norm penalty minimizes the estimate that account for multicollinearity. L2 norm also removes the limitation on the number of selected variables, encourages grouping effect and stabilizes the L1 norm regularization path, while it does not perform variable selection.

However, elastic net method can overcome both the limitations of ridge and lasso regression because it performs ridge and lasso regression simultaneously. First, for each fixed lambda 2, elastic net finds the ridge regression coefficients, and then does a lasso type shrinkage. This method recommended to use for variable selection especially when

dealing with multicollinearity. As for prediction accuracy, elastic net also performs better than lasso.

c. Stepwise Regression

Stepwise regression is model selection process by successively adding or removing variables from the set of explanatory variables based on some prespecified criterion. Usually, the criterion are F-tests or t-tests, adjusted R-squared, Akaike information criterion, Bayesian information criterion, Mallow's Cp, PRESS, or false discovery rate. Stepwise regression is used when dealing with large numbers of potential independent variables in a model from which you wish to extract the best subset for your predicting model. The main approaches of stepwise regression are forward, starting with no variables in the model and then add one variable at a time; backward, starting with all variables in the model and remove one variable at a time; hybrid, a combination of forward and backward, which test at each step for variables to be included or excluded.

Stepwise regression algorithm is that, at each stage in the process, it performs a test to check if some variables can be deleted without appreciably increasing the residual sum of squares. The procedure terminates when the measure is maximized, or when the available improvement falls below some critical value.

d. Multiple Linear Regression

Multiple linear regression is a method to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line. The multiple linear regression model with n data observations is described as below:

Data: $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$

Model: $y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$

Assumptions:

- Linearity/Mean Zero Assumption: E(ei) = 0
- Constant Variance Assumption: Var(ei) = s2
- Independence Assumption: {e1,…, en} are independent random variables
- Normality: ei ~ Normal

Linear regression is one of the most important tools to describe possible relationships between variables and it's widely used in many disciplines.

IV. Procedure

a. Data Cleaning

- From the summary of variables in the dataset, it was found that some variables had ~90% missing values. There were 40 such variables, and hence these variables were removed from the dataset.
- Some variables such as *countyCode*, *communityCode*, *communityname*, and *state* were removed, because these are present in the dataset for identification, and will not help in explaining the variability in the response variable.
- Since this analysis is focused on the *ViolentCrimesPerPop* variable as the response, the other 17 crime related variables were removed.
- Some data points a missing value for the response variable. There were 222 such rows, and hence were removed from the dataset.
- Variables were converted to the correct format, i.e. numeric so that they are not treated as categorical variables while performing regression analysis.
- The final dataset after cleaning consisted of 1993 rows and 102 variables. critical value.

b. Data Processing

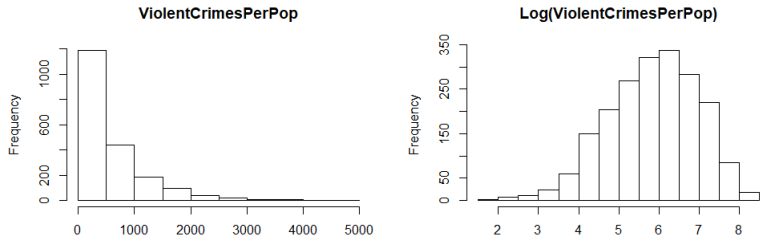
- Some variables in the dataset are in absolute terms, whereas some are in percentage terms. For uniformity and better interpretation, these variables were converted into percentages by dividing the value by population. Hence,
    - *NumStreet* → *PctInStreet*
    - *NumInShelters* → *PctInShelters*

- *NumImmig* → *PctImmig*

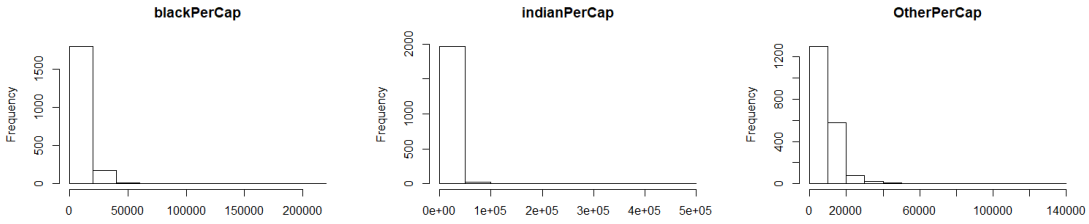
- Some variables were available in absolute, as well as percentage terms. Hence, the former kind were removed from the dataset. Variables included
  - *numbUrban*
  - *NumUnderPov*
  - *NumKidsBornNeverMar*
- Since our response variable *ViolentCrimesPerPop* has already been controlled for population, the *population* variable was removed from the dataset after the above-mentioned steps.

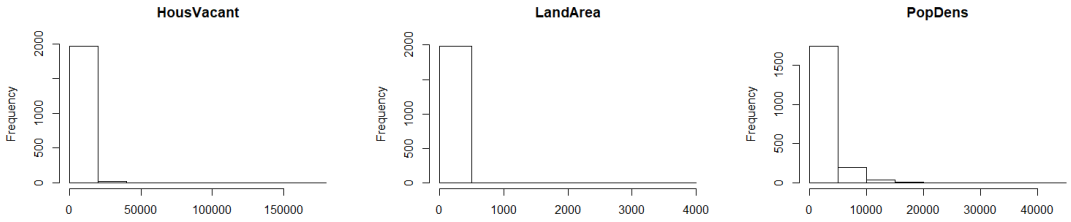
c. Exploratory Data Analysis

- An important assumption to perform statistical inference using multiple linear regression is normality. It was found that the response variable was highly skewed. On applying a logarithmic transformation on the response variable, the histogram fit well to a normal distribution.



- The predicting variables were plotted against the Log(*ViolentCrimesPerPop)*, and there were no specific trends (e.g. quadratic, logarithmic, etc.). Hence, it is safe to assume that the linearity assumption holds.
- There were, however 6 predicting variables that were highly skewed. The plots are shown below.

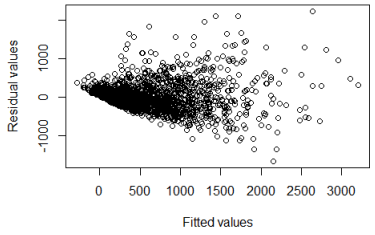




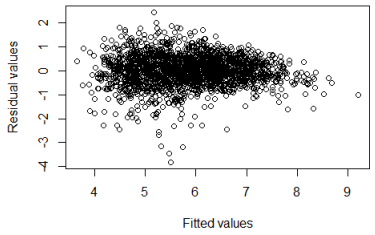
- A correlation plot was created to show the correlation between the predicting variables (Figure in Appendix). Clearly, there are a lot of variables that are highly correlated. Hence, it is essential to perform variable selection to select variables.

d. Preliminary Model fitting

- A basic linear model was fitted using *ViolentCrimesPerPop* as the response variable. Residual analysis was performed, and a clear violation of the constant variance assumption was found.



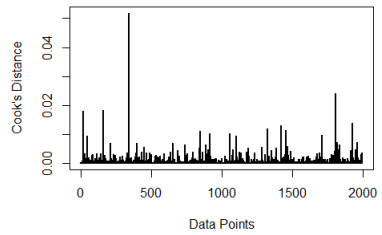
- Moving forward, a basic linear model was fitted using the logarithmic transformation of *ViolentCrimesPerPop.* Residual analysis was performed, and the plots improved considerably.



- Hence, it was decided to move forward with the logarithmic transformation of the response variable to perform any further analysis. The response variable will now be referred to as *LogViolentCrimesPerPop*.

e. Outlier Removal

- Using the above-mentioned model with *LogViolentCrimesPerPop* as the response variable, the cook's distance was calculated to identify in case of any outliers in the dataset.



- Data points with cook's distance > 0.02 were removed.

f. Variable Selection

- Lasso and Elastic net regression were performed for 2 datasets – one with taking log of 6 variables mentioned in the Exploratory Data Analysis, and one without.
- The data was scaled to perform variable selection.
- For the dataset without log transformation of the 6 variables:
    - LASSO removed 49 variables
    - Elastic net removed 48 variables

  For the dataset with log transformation of the 6 variables:
    - LASSO removed 68 variables
    - Elastic net removed 63 variables

  Based on these results, and human judgement, the LASSO model which removed 68 variables was chosen.
- Taking the variables selected from LASSO, the step regression method was used. This reduced the variables to 23.

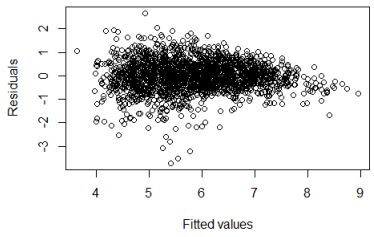
g. Multiple Linear Regression

- Using the 23 variables selected, a multiple linear regression model was fitted on the unscaled dataset, using *LogViolentCrimesPerPop* as the response variable.
- The Multiple R-squared is 0.6596, and the Adjusted R-squared: 0.6556
- The F-statistic is 165.8 on 23 and 1968 Degrees of Freedom. The p-value is < 2.2e-16, confirming that the regression is significant.
- The estimated model equation is as follows:

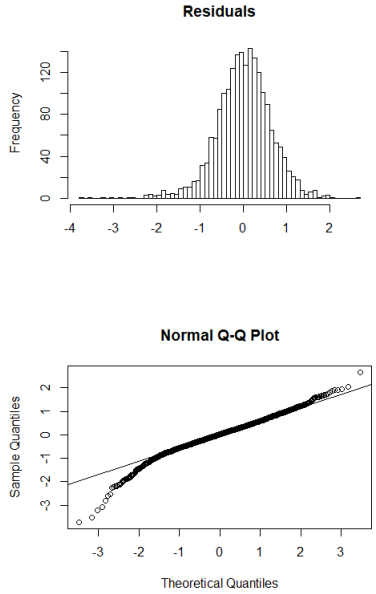
$$Log(ViolentCrimesPerPop) = 9.2026 - 0.0265 * PctKids2Par + 0.1330 * LogHousVacant + 0.1779 * PersPerRentOccHous - 0.0132 * racePctWhite + 0.0018 * pctUrban - 0.0250 * pctWInvInc - 0.0034 * PctVacMore6Mos + 0.0011 * RentQrange + 0.0331 * MalePctDivorce - 0.0053 * PctEmplManu - 0.0216 * MedOwnCostPctIncNoMtg + 0.0158 * LogOtherPerCap - 0.2590 * householdsize + 0.0448 * LogPopDens - 0.0095 * PctUsePubTrans - 0.0335346 * PctUnemployed + 0.0033 * racePctHisp - 0.0105 * pctWWage - 0.0752981 * PctWOFullPlumb + 0.3815 * PctStreet + 0.0079 * PctSameState85 - 0.0028 * PctBornSameState + 0.1913 * PctInShelters$$

h. Goodness of fit analysis

- Residuals v/s Fitted values were plotted to check the linearity assumption and constant variance assumption. As can be seen from the plot below, these assumptions hold.



- The histogram and Normal Q-Q plot confirm that the normality assumption of the residuals seems to hold.



V. Model Results and Interpretation

Given all other predicting factors holding constant,

- A one percentage point increase in percent of kids in family housing with two parents decreases the Violent Crimes per 100k population by 2.62%
- A one-unit increase in the mean persons per rental household increases the Violent Crimes per 100k population by 19.47%
- A one percentage point increase in percentage of population that is Caucasian decreases the Violent Crimes per 100k population by 1.31%
- A one percentage point increase in percentage of people living in areas classified as urban increases the Violent Crimes per 100k population by 0.18%
- A one percentage point increase in the percentage of households with investment / rent income in 1989 decreases the Violent Crimes per 100k population by 2.47%
- A one percentage point increase in percent of vacant housing that has been vacant more than 6 months decreases the Violent Crimes per 100k population by 0.34%
- A one-unit increase in rental housing (difference between upper quartile and lower quartile rent) increases the Violent Crimes per 100k population by 0.11%

- A one percentage point increase in percentage of males who are divorced increases the Violent Crimes per 100k population by 3.37%
- A one percentage point increase in percentage of people 16 and over who are employed in manufacturing decreases the Violent Crimes per 100k population by 0.53%
- A one percentage point increase in median owners cost as a percentage of household income - for owners without a mortgage decreases the Violent Crimes per 100k population by 2.14%
- A one-unit increase in mean people per household decreases the Violent Crimes per 100k population by 22.82%
- A one percentage point increase in percent of people using public transit for commuting decreases the Violent Crimes per 100k population by 0.94%
- A one percentage point increase in percentage of people 16 and over, in the labor force, and unemployed decreases the Violent Crimes per 100k population by 3.3%
- A one percentage point increase in percentage of population that is of hispanic heritage increases the Violent Crimes per 100k population by 0.33%
- A one percentage point increase in percentage of households with wage or salary income in 1989 decreases the Violent Crimes per 100k population by 1.05%
- A one percentage point increase in percent of housing without complete plumbing facilities decreases the Violent Crimes per 100k population by 7.25%
- A one percentage point increase in percent of homeless people counted in the street increases the Violent Crimes per 100k population by 46.45%
- A one percentage point increase in percent of people living in the same state as in 1985 increases the Violent Crimes per 100k population by 0.8%
- A one percentage point increase in percent of people born in the same state as currently living decreases the Violent Crimes per 100k population by 0.28%
- A one percentage point increase in percent of people in homeless shelters increases the Violent Crimes per 100k population by 21.08%
- A one percentage point increase in number of vacant households increases the Violent Crimes per 100k population by 0.13%
- A one percentage point increase in the per capita income for people with 'other' heritage increases the Violent Crimes per 100k population by 0.015%

- A one percentage point increase in population density in persons per square mile increases the Violent Crimes per 100k population by 0.044%

VI. Conclusion and Discussions

- Based on the model results, it is evident that two of the most important factors contributing to violent crimes are presence of homeless people on the streets as well as those living in homeless shelters. There have been efforts to provide shelters for the homeless; however, it hasn't been able to solve the problem of violent crimes. Hence, an important insight based on this model is that using the government budget on shelters for the homeless people may not be the best solution to address crime.
- A recommended solution to this problem would be to employ these homeless people in the labor force, as we can see from the model results that a one percentage point increase in the percentage of people 16 and over, in the labor force, and unemployed decreases the Violent Crimes per 100k population by 3.3%
- It can be seen from the model results that an increased number of households with wage or salary income can help address crime. This again reinforces the importance of increasing employment to reduce crime. Hence, focusing on creating employment opportunities should be the topmost priority of states to address violent crime.
- From the model results, we can see that an increase in percent of kids in family housing with two parents decreases violent crime. Moreover, an increase in percentage of males who are divorced increases violent crime. Hence, states must encourage and spread awareness about the importance of family, as it improves the well-being of society.
- The model can be used to predict for one unit or one percent increase/decrease in a predicting variable given all other predictors in the model, what the percentage change of violent crimes per 100k population is. From our model, we can see that violent crimes are associated with 23 factors. These 23 predictors include demographic, geographical, income, race, employment factors as expected. Interestingly, we found that crime rates also associate with real estate related factors, such as rental, vacant housing and rent income. Residency and number of people using public transportation also unexpected factors we found that affect crime rate.

- We also noticed that some of the predictors are not statistically significant at 95% confidence level. For example, percent of housing without complete plumbing facilities, percent of homeless people counted in the street and percent of people born in the same state as currently living. Such predictors do not have predictive power in the model.

VII. Future Work

We recommend performing the same analysis for the current violent crime rates. Since violent crimes have gone down since the 1980's, it would be useful to model the crime rates for the year 2015, and see how the beta coefficients of the predicting variables have changed over time, and which factors played a role in reducing crime. It would also be insightful to see if apart from the predicting variables in this model, if any other additional variables turn out to be more significant in a different time period.

ISyE 6414 Project R Code

Regression Analysis of Violent Crimes in the United States

In [1]:

```
# Clear log
rm(list=ls())

# Set working directory
setwd('C:\\Users\\mansiarora\\Documents\\ISyE 6414\\Project')
```

I. Data Cleaning

In [2]:

```
data_raw = read.csv("Team2_Dataset.csv")
#summary(data_raw)
```

As we can see, some columns have 1872 missing values. We're going to remove those columns. Also, we have selected ViolentCrimesPerPop, so we have deleted other measures of crime rate. We have also deleted the descriptors - communityname, state and fold columns.

In [3]:

```
# Removing columns
droplist = c("countyCode",      "communityCode",        "LemasSwornFT", "LemasSwFTPerPo
p",     "LemasSwFTFieldOps",
             "LemasSwFTFieldPerPop",    "LemasTotalReq",        "LemasTotReqPerPop",
        "PolicReqPerOffic",     "PolicPerPop",
             "RacialMatchCommPol",      "PctPolicWhite",        "PctPolicBlack",
        "PctPolicHisp", "PctPolicAsian",
             "PctPolicMinor",   "OfficAssgnDrugUnits",  "NumKindsDrugsSeiz",    "PolicA
veOTWorked",    "PolicCars",
             "PolicOperBudg",   "LemasPctPolicOnPatr",  "LemasGangUnitDeploy",  "LemasP
ctOfficDrugUn",
             "PolicBudgPerPop", "murders",      "murdPerPop",   "rapes",        "rapesP
erPop", "robberies",    "robbbPerPop",
             "assaults",        "assaultPerPop",        "burglaries",   "burglPerPop",
        "larcenies",    "larcPerPop",   "autoTheft",
             "autoTheftPerPop", "arsons",       "arsonsPerPop", "nonViolPerPop", "commu
nityname",      "state",        "fold")

data_raw = data_raw[, !(colnames(data_raw) %in% droplist)]
```

There are 221 rows where we don't have the value of our response value, and one row where the value our response is 0. These rows have been deleted.

In [4]:

```
data_raw = data_raw[!(data_raw$ViolentCrimesPerPop %in% c("?",0)),]
# One column of OtherPerCap is  ?, changing this value to 0
data_raw[data_raw$OtherPerCap == "?",]$OtherPerCap = 0
```

In [5]:

```
# 2 columns - OtherPerCap and ViolentcrimesPerPop - are convered to numeric
data_raw$OtherPerCap = as.numeric(as.character(data_raw$OtherPerCap))
data_raw$ViolentCrimesPerPop = as.numeric(as.character(data_raw$ViolentCrimesPerPop))
dim(data_raw)
data = data_raw
```

```
    1993  102
```

II. Data Transformation

We have transformed some of the variables from absolute to a percent of population This will give us a better interpretation. Also, there are some variables which are available in percent as well as absolute form. We are keeping the ones which are in percent form.

In [6]:

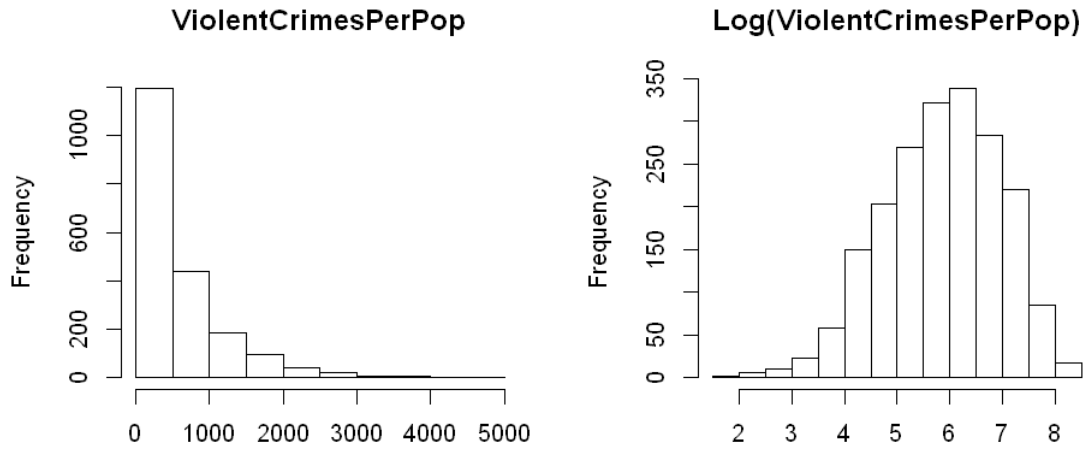
```
# Converting NumStreet, NumInShelters, NumImmig to Pct
data$PctInShelters = data$NumInShelters/data$population*100
data$PctStreet = data$NumStreet/data$population*100
data$PctImmig = data$NumImmig/data$population*100

# Removing numUrban, NumUnderPov, NumKidsBornNeverMar, and the 3 columns we transformed
 above
data$numbUrban = NULL
data$NumUnderPov = NULL
data$NumKidsBornNeverMar = NULL
data$NumInShelters = NULL
data$NumStreet = NULL
data$NumImmig = NULL

# Since our response variable (ViolentCrimesPerPop) has already been controlled for pop
ulation,
# we don't need to keep 'population' as a predicting variable.
data$population = NULL
```

III Exploratory Data Analysis

Our response variable shows a very strong right tailed skew, hence we have used a log transformation to satisfy the normality assumption.

In [7]:

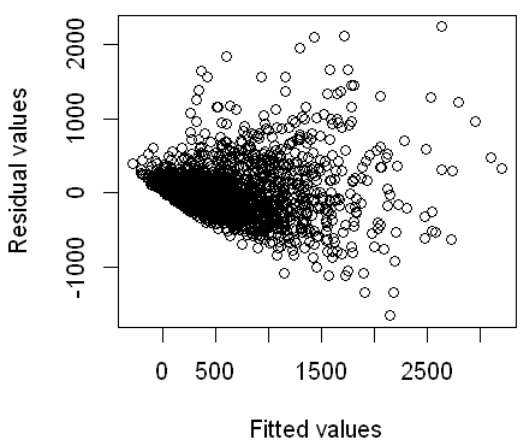
```
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
hist(data$ViolentCrimesPerPop, main = "ViolentCrimesPerPop", xlab="")
hist(log(data$ViolentCrimesPerPop), main = "Log(ViolentCrimesPerPop)", xlab="")

# Taking log of the response variable
data$LogViolentCrimesPerPop = log(data$ViolentCrimesPerPop)
```

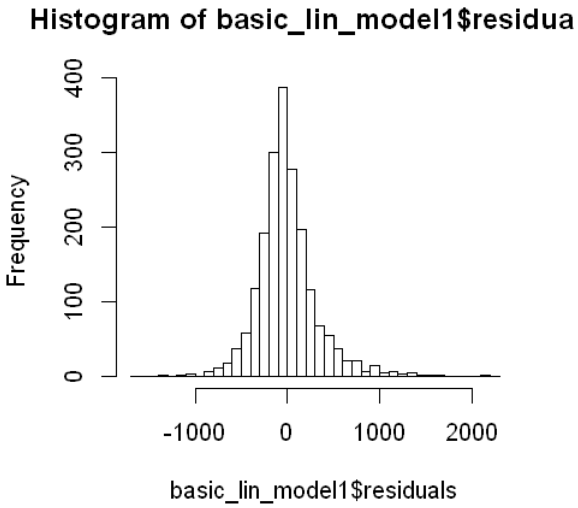


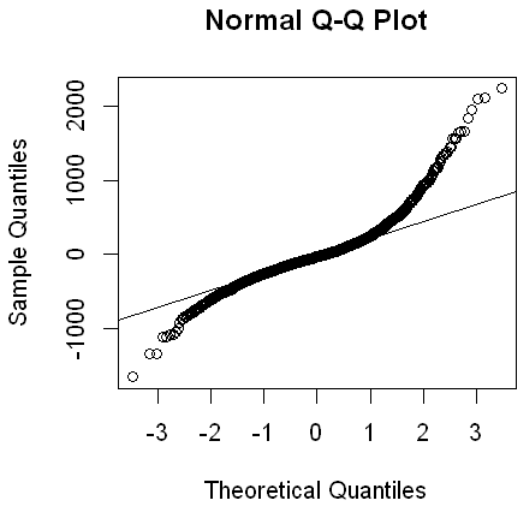
IV Preliminary Model Fitting

In [8]:

```
############################## Without Log ###################################
basic_lin_model1 = lm(ViolentCrimesPerPop ~ ., data=data[,!(colnames(data) %in% "LogVio
lentCrimesPerPop")])
par(mfrow=c(1,2))
# 1. Check constant variance
plot(basic_lin_model1$fitted,basic_lin_model1$residuals, xlab="Fitted values", ylab="Re
sidual values")  # Constant variance assumption violated

# 2.Check normality of residuals
hist(basic_lin_model1$residuals,breaks=50) # Fairly symmetric, normal distribution
qqnorm(basic_lin_model1$residuals)
qqline(basic_lin_model1$residuals)
```

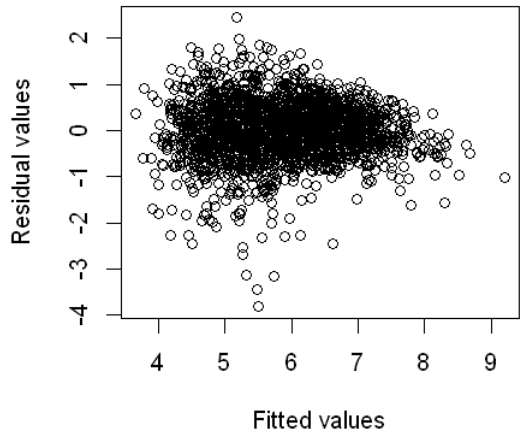


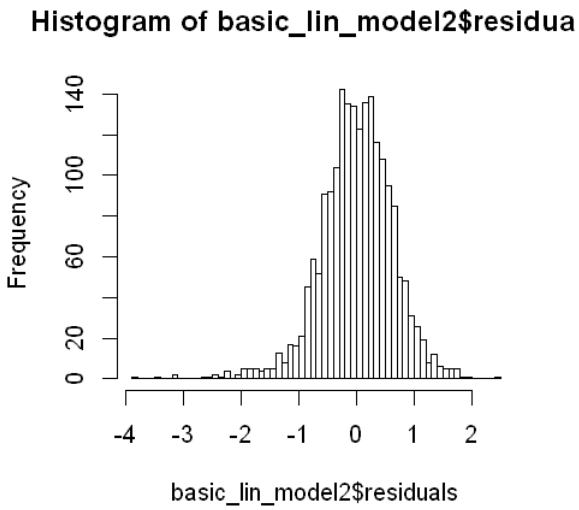


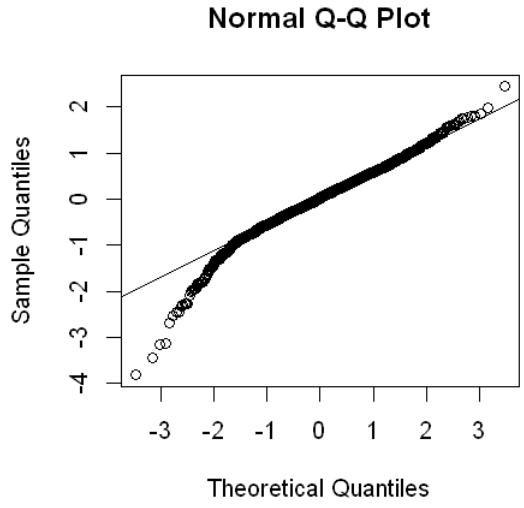


In [9]:

```
#################################### With Log ####################################

# Retry linear regression basic model with Log
basic_lin_model2 = lm(LogViolentCrimesPerPop ~ ., data=data[,!(colnames(data) %in% "Vio
lentCrimesPerPop")])
par(mfrow=c(1,2))
# 1. Check constant variance
plot(basic_lin_model2$fitted,basic_lin_model2$residuals, xlab="Fitted values", ylab="Re
sidual values")
# Constant variance assumption holds though still marginally questionable due to skew o
n the right but acceptable

# 2.Check normality of residuals
hist(basic_lin_model2$residuals,breaks=50) # Fairly symmetric, normal distribution
qqnorm(basic_lin_model2$residuals) # QQ plot traces the theoretical quantiles' line
qqline(basic_lin_model2$residuals)

# Conclusion = Taking Log makes sense. We are going to remove the ViolentCrimesPerPop r
esponse
data$ViolentCrimesPerPop = NULL
```







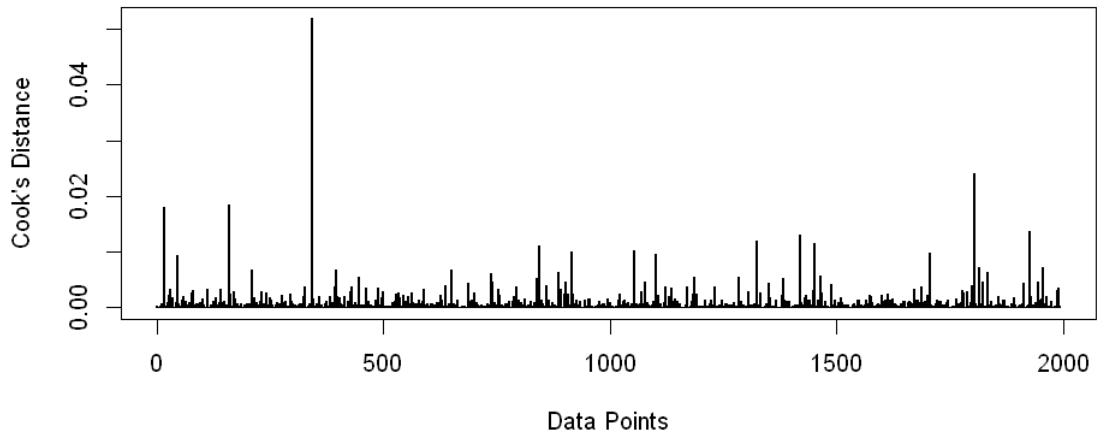
In [10]:

```
# Checking relationship between predicting variables and response
# This portion has been commented out to avoid having too many plots on display
# for(i in colnames(data[,!(colnames(data) %in% c("ViolentCrimesPerPop", "LogViolentCri
mesPerPop"))])){
#  plot(data[,i], data$LogViolentCrimesPerPop, main = i)
# }
```

V Outlier Removal

In [11]:

```
####### IV. Outlier removal using Cook's distance (Response = log(ViolentCrimesPerPop))
######

model = lm(LogViolentCrimesPerPop ~ ., data = data)
cook = cooks.distance(model) # Vector with Cook's distance values
plot(cook,type="h",lwd=2, xlab="Data Points", ylab = "Cook's Distance") # Plot of Coo
k's distance to identify outliers
```



In [12]:

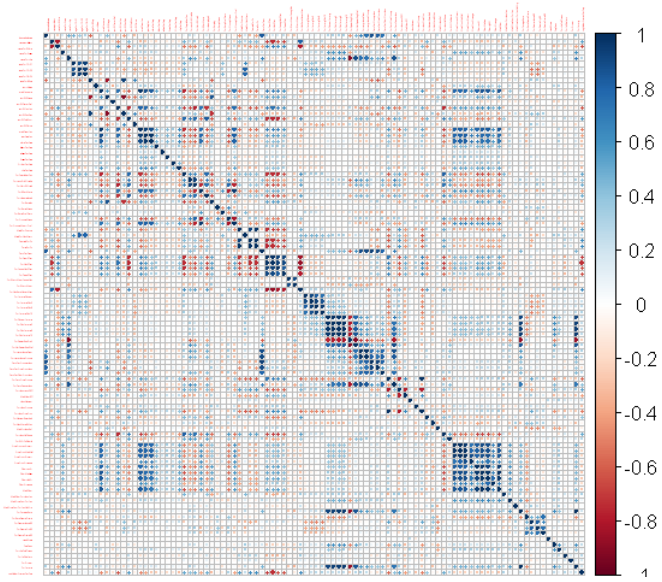
```
# Removing outliers
data_wo_outliers = data[-as.integer(row.names(as.data.frame(cook[cook>0.02]))),] # Dele
te outlier points
data_wt_outliers = data
dim(data_wo_outliers)
```

1992 98

In [13]:

```
############ Checking for correlation between predicting variables ##################
#install.packages("corrplot")
library(corrplot)
corrplot(cor(as.matrix(data_wo_outliers)), tl.cex = .1)
#cor(data_wo_outliers)
# As we can see, there are a lot of correlated variables, and hence we must perform var
iable
# selection to account for the same
```

```
Warning message:
"package 'corrplot' was built under R version 3.4.2"corrplot 0.84 loaded
```



In [14]:

```
###################### Checking if taking log of some variables makes sense #############
#

#for(i in colnames(data_wo_outliers)) {
#  par(mfrow=c(1,2))
#  hist(data_wo_outliers[,i], main = i, xlab = "")
#  hist(log(data_wo_outliers[,i]), main = paste0("Log(",i, ")"), xlab = "")
#}

# Log-Transformed 6 predictor variables, which show a skewed distribution
data_wo_outliers_log = data_wo_outliers
data_wo_outliers_log$LogblackPerCap = log(data_wo_outliers_log$blackPerCap+1)
data_wo_outliers_log$LogindianPerCap = log(data_wo_outliers_log$indianPerCap+1)
data_wo_outliers_log$LogOtherPerCap = log(data_wo_outliers_log$OtherPerCap+1)
data_wo_outliers_log$LogHousVacant = log(data_wo_outliers_log$HousVacant+1)
data_wo_outliers_log$LogLandArea = log(data_wo_outliers_log$LandArea+1)
data_wo_outliers_log$LogPopDens = log(data_wo_outliers_log$PopDens+1)
data_wo_outliers_log$blackPerCap = NULL
data_wo_outliers_log$indianPerCap = NULL
data_wo_outliers_log$OtherPerCap = NULL
data_wo_outliers_log$HousVacant = NULL
data_wo_outliers_log$LandArea = NULL
data_wo_outliers_log$PopDens = NULL
```

VI Variable Selection

In [15]:

```
###### Scaling #######
# Required prior to Variable Selection
# R 'scale' function carries out standardization based on mean and SD
data_wo_scaled_log = as.data.frame(sapply(data_wo_outliers_log, scale))
data_wo_scaled = as.data.frame(sapply(data_wo_outliers, scale))
data_wt_scaled = as.data.frame(sapply(data_wt_outliers, scale))
```

In [16]:

```
##### Regularized Regression Approaches - LASSO & Elastic Net #####

# Define set.seed(100) for all steps to get the same output
set.seed(100)

library(glmnet) # Import glmnet for LASSO, Elastic Net

# Define the linear regression model for non-log predicting variables
basic_lm_scaled1 = lm(LogViolentCrimesPerPop ~ ., data=data_wo_outliers)
X_1 = model.matrix(basic_lm_scaled1); X_1 = X_1[,-c(1)]
y_1 = as.vector(data_wo_outliers[,'LogViolentCrimesPerPop'])

# Define the linear regression model for log predicting variables
basic_lm_scaled2 = lm(LogViolentCrimesPerPop ~ ., data=data_wo_outliers_log)
X_2 = model.matrix(basic_lm_scaled2); X_2 = X_2[,-c(1)]
y_2 = as.vector(data_wo_outliers_log[,'LogViolentCrimesPerPop'])
```

```
Warning message:
"package 'glmnet' was built under R version 3.4.2"Loading required packag
e: Matrix
Loading required package: foreach
Loaded glmnet 2.0-13
```

In [17]:

```
##### LASSO for non-log predicting variables #####
model_cv_lasso_1 = cv.glmnet(X_1, y_1, alpha=1, family = "gaussian", nfolds=10) # 10-fo
ld cross-validation, k=10 standard
model_lasso_1 = glmnet(X_1, y_1, alpha=1, family = "gaussian", lambda = model_cv_lasso_
1$lambda.min)
sum(model_lasso_1$beta == 0) # 47 coefficients with 0 beta values found
```

```
49
```

In [18]:

```
##### Elastic Net for non-log predicting variables #####
model_cv_elnet_1 = cv.glmnet(X_1, y_1, alpha=0.5, family = "gaussian", nfolds=10) # 10-
fold cross-validation, k=10 standard
model_elnet_1 = glmnet(X_1, y_1, alpha=0.5, family = "gaussian", lambda = model_cv_elne
t_1$lambda.min)
sum(model_elnet_1$beta == 0) # 21 coefficients with 0 beta values found
```

```
48
```

In [19]:

```
##### LASSO for log predicting variables #####
model_cv_lasso_2 = cv.glmnet(X_2, y_2, alpha=1, family = "gaussian", nfolds=10) # 10-fo
ld cross-validation, k=10 standard
model_lasso_2 = glmnet(X_2, y_2, alpha=1, family = "gaussian", lambda = model_cv_lasso_
2$lambda.min)
sum(model_lasso_2$beta == 0) # 47 coefficients with 0 beta values found
```

67

In [20]:

```
##### Elastic Net for log predicting variables #####
model_cv_elnet_2 = cv.glmnet(X_2, y_2, alpha=0.5, family = "gaussian", nfolds=10) # 10-
fold cross-validation, k=10 standard
model_elnet_2 = glmnet(X_2, y_2, alpha=0.5, family = "gaussian", lambda = model_cv_elne
t_2$lambda.min)
sum(model_elnet_2$beta == 0) # 21 coefficients with 0 beta values found
```

63

In [21]:

```
# Variables Removed by Lasso and ElasticNet for non-log predicting variables
droplist_lasso_1 = row.names(model_lasso_1$beta)[as.vector(model_lasso_1$beta == 0)]
droplist_elnet_1 = row.names(model_elnet_1$beta)[as.vector(model_elnet_1$beta == 0)]

# Variables Removed by Lasso and ElasticNet for log predicting variables
droplist_lasso_2 = row.names(model_lasso_2$beta)[as.vector(model_lasso_2$beta == 0)]
droplist_elnet_2 = row.names(model_elnet_2$beta)[as.vector(model_elnet_2$beta == 0)]
```

In [22]:

```
# Output from Regularized Regression Step for log predicting variables
data_fnl_regr_log = data_wo_outliers_log[,!(colnames(data_wo_outliers_log) %in% droplis
t_lasso_2)]
# Final dataset after LASSO has 1992X30 dimensions
dim(data_fnl_regr_log)
```

1992 31

In [28]:

```
###### Stepwise Regression #######
full_2 = lm(LogViolentCrimesPerPop ~ ., data = data_fnl_regr_log) # Define full model w
ith all variables
null_2 = lm(LogViolentCrimesPerPop ~ 1, data = data_fnl_regr_log) # Define null model w
ith no variables
#model_both_2 = step(null_2, scope = list(upper=full_2), data=data_wo_outliers_log, dir
ection="both")
summary(model_both_2)
length(model_both_2$coefficients)
# 23 variables, Multiple R-squared:  0.6596,    Adjusted R-squared:  0.6556
```

```
Call:
lm(formula = LogViolentCrimesPerPop ~ PctKids2Par + LogHousVacant +
    PersPerRentOccHous + racePctWhite + pctUrban + pctWInvInc +
    PctVacMore6Mos + RentQrange + MalePctDivorce + PctEmplManu +
    MedOwnCostPctIncNoMtg + LogOtherPerCap + householdsize +
    LogPopDens + PctUsePubTrans + PctUnemployed + racePctHisp +
    pctWWage + PctWOFullPlumb + PctStreet + PctSameState85 +
    PctBornSameState + PctInShelters, data = data_fnl_regr_log)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7322 -0.3697  0.0200  0.3966  2.6728

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.2026661  0.5859294  15.706  < 2e-16 ***
PctKids2Par          -0.0265454  0.0037150  -7.145 1.26e-12 ***
LogHousVacant         0.1330692  0.0178587   7.451 1.38e-13 ***
PersPerRentOccHous    0.1779239  0.0782379   2.274 0.023065 *
racePctWhite         -0.0132172  0.0017399  -7.596 4.68e-14 ***
pctUrban              0.0018240  0.0004124   4.423 1.03e-05 ***
pctWInvInc           -0.0250019  0.0028745  -8.698  < 2e-16 ***
PctVacMore6Mos       -0.0034257  0.0013563  -2.526 0.011624 *
RentQrange            0.0011488  0.0002242   5.124 3.29e-07 ***
MalePctDivorce        0.0331365  0.0104224   3.179 0.001499 **
PctEmplManu          -0.0052879  0.0020434  -2.588 0.009729 **
MedOwnCostPctIncNoMtg -0.0216137  0.0123092  -1.756 0.079262 .
LogOtherPerCap        0.0157686  0.0067557   2.334 0.019689 *
householdsize        -0.2590054  0.0898010  -2.884 0.003967 **
LogPopDens            0.0447743  0.0244228   1.833 0.066909 .
PctUsePubTrans       -0.0094843  0.0041592  -2.280 0.022696 *
PctUnemployed        -0.0335346  0.0101965  -3.289 0.001024 **
racePctHisp           0.0032753  0.0016760   1.954 0.050817 .
pctWWage             -0.0105054  0.0031134  -3.374 0.000755 ***
PctWOFullPlumb       -0.0752981  0.0432492  -1.741 0.081835 .
PctStreet             0.3815164  0.2482291   1.537 0.124466
PctSameState85        0.0079544  0.0035776   2.223 0.026301 *
PctBornSameState     -0.0027627  0.0017646  -1.566 0.117591
PctInShelters         0.1912918  0.1278970   1.496 0.134900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6503 on 1968 degrees of freedom
Multiple R-squared:  0.6596,    Adjusted R-squared:  0.6556
F-statistic: 165.8 on 23 and 1968 DF,  p-value: < 2.2e-16
```

VII Multiple Linear Regression Model

In [24]:

```
vars_select_2 = c((names(model_both_2$coefficients[2:length(model_both_2$coefficients
)])),'LogViolentCrimesPerPop')
# Unscaled dataset consisting of 23 selected predictor variables and 1 response variabl
e
data_mlr_unscaled_2 = data_wo_outliers_log[,vars_select_2]
dim(data_mlr_unscaled_2)
# 1992 x 24
```

1992 24

In [25]:

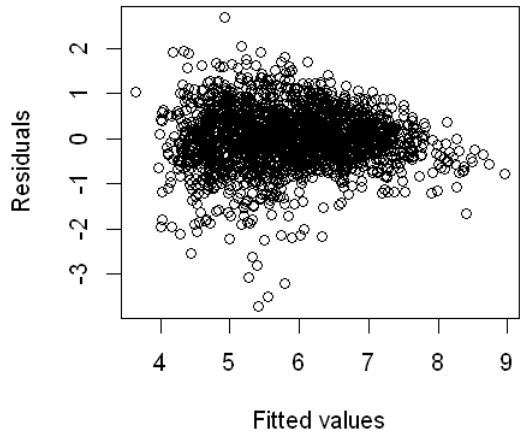
```
# Multiple Linear Regression model fitted
mlr_model_2 = lm(LogViolentCrimesPerPop ~., data= data_mlr_unscaled_2)
summary(mlr_model_2)
```

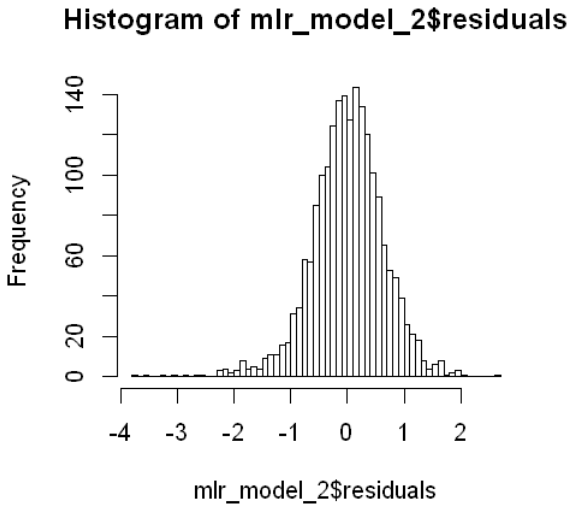
```
Call:
lm(formula = LogViolentCrimesPerPop ~ ., data = data_mlr_unscaled_2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7322 -0.3697  0.0200  0.3966  2.6728

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.2026661  0.5859294  15.706  < 2e-16 ***
PctKids2Par          -0.0265454  0.0037150  -7.145 1.26e-12 ***
LogHousVacant         0.1330692  0.0178587   7.451 1.38e-13 ***
PersPerRentOccHous    0.1779239  0.0782379   2.274 0.023065 *
racePctWhite         -0.0132172  0.0017399  -7.596 4.68e-14 ***
pctUrban              0.0018240  0.0004124   4.423 1.03e-05 ***
pctWInvInc           -0.0250019  0.0028745  -8.698  < 2e-16 ***
PctVacMore6Mos       -0.0034257  0.0013563  -2.526 0.011624 *
RentQrange            0.0011488  0.0002242   5.124 3.29e-07 ***
MalePctDivorce        0.0331365  0.0104224   3.179 0.001499 **
PctEmplManu          -0.0052879  0.0020434  -2.588 0.009729 **
MedOwnCostPctIncNoMtg -0.0216137  0.0123092  -1.756 0.079262 .
LogOtherPerCap        0.0157686  0.0067557   2.334 0.019689 *
householdsize        -0.2590054  0.0898010  -2.884 0.003967 **
LogPopDens            0.0447743  0.0244228   1.833 0.066909 .
PctUsePubTrans       -0.0094843  0.0041592  -2.280 0.022696 *
PctUnemployed        -0.0335346  0.0101965  -3.289 0.001024 **
racePctHisp           0.0032753  0.0016760   1.954 0.050817 .
pctWWage             -0.0105054  0.0031134  -3.374 0.000755 ***
PctWOFullPlumb       -0.0752981  0.0432492  -1.741 0.081835 .
PctStreet             0.3815164  0.2482291   1.537 0.124466
PctSameState85        0.0079544  0.0035776   2.223 0.026301 *
PctBornSameState     -0.0027627  0.0017646  -1.566 0.117591
PctInShelters         0.1912918  0.1278970   1.496 0.134900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6503 on 1968 degrees of freedom
Multiple R-squared:  0.6596,    Adjusted R-squared:  0.6556
F-statistic: 165.8 on 23 and 1968 DF,  p-value: < 2.2e-16
```

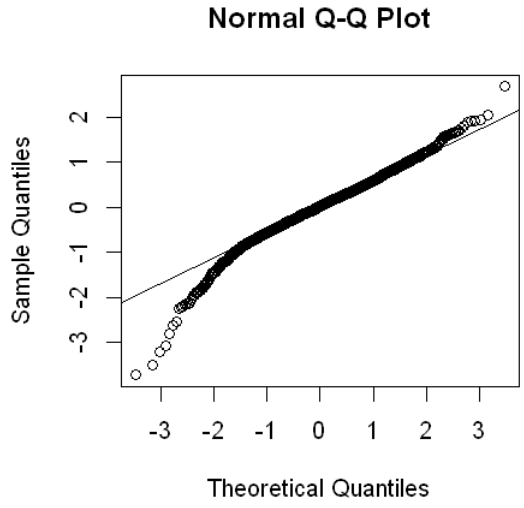
VIII Goodness of Fit

In [26]:

```
# Constant variance & Independence
par(mfrow=c(1,2))
plot(mlr_model_2$fitted, mlr_model_2$residuals, xlab="Fitted values",ylab="Residuals")
# Constant variance assumption holds though a narrowing is observed as we increase the
 fitted values
# Normality of residuals
hist(mlr_model_2$residuals, breaks=50)
qqnorm(mlr_model_2$residuals)
qqline(mlr_model_2$residuals)
```







In [27]:

```
### End
```