University of California, Santa Barbara

PSTAT 126: Regression Analysis
Final Project

Sam Guo (6865141, Ke Wang R 1-1:50)
Jackie Kajisa (3838430, Ke Wang R 1-1:50)

20 November 2019

Regression Analysis of Per Capita Violent Crimes

I. Introduction

This project will revolve around the analysis of Per Capita Violent Crimes in various United States communities, as well as how it pertains to various traits of each community provided within the data set “Communities and Crime (Unnormalized)” provided by the UC Irvine Machine Learning Repository. We will examine the effect of the following variables on predicting crime:

- population
- median income
- percentage of population that is 12-21 in age
- percentage of population that is 12-29 in age
- percentage of population that is 16-24 in age
- percentage of population that is 65 and over in age
- unemployment rate
- population density
- mean household size
- poverty rate

In addition to determining whether any of these values can predict the true number of per capita violent crimes, we will be looking into which combination of them has the most correlation and impact on predicting the response.

II. Questions of Interest

1. Is there a correlation between the percentage of individuals within a certain age range in a community and the number of per capita violent crimes?

2. Can a community’s number of per capita violent crimes be predicted by population size, median income, unemployment rate, population density, mean household size, and percentage of people under the poverty level?

III. Regression Method

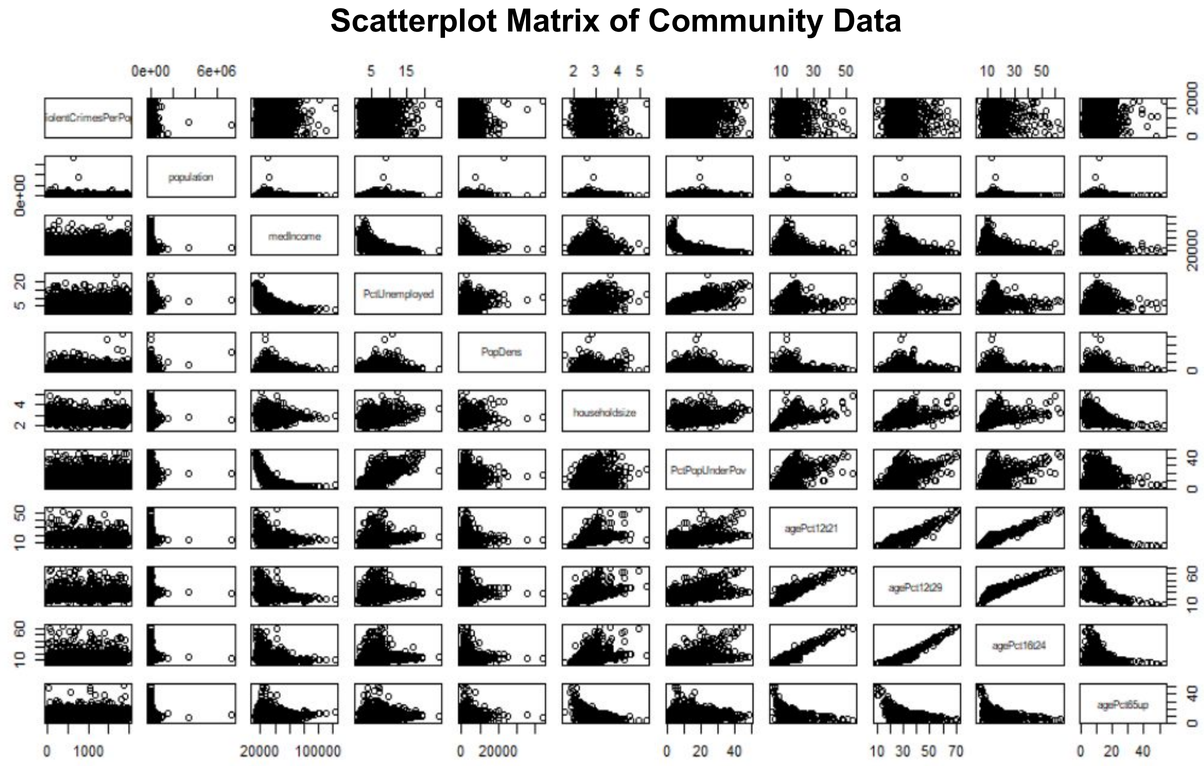
In order to perform any regression analysis, a proper model must be built that fulfills all the LINE conditions. To do this, we will take our predictors and perform residual analysis to determine whether or not they are suitable for the linear model in their current state. If not, appropriate transformations will be made to fix any problems of linearity, normality, or unequal variance. Additionally, each multivariate model will undergo stepwise regression to ensure that we are using the model with the best relationship.

The first question of interest can be answered by observing the fit between each age range percentage and the number of per capita violent crimes in each community. We will then observe the coefficient of determination and the p-value of the F-test to determine whether or not the age ranges have a significant effect on the response.

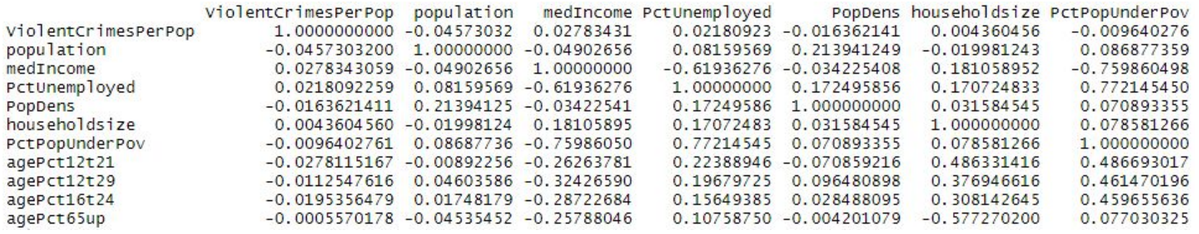
The second question of interest can be answered similarly. We will again create a linear model with population size, median income, unemployment rate, population density, mean household size, and percentage of people under the poverty level as predictors, with per capita violent crimes as the response. T-tests and F-tests will be conducted to determine whether or not each of these variables have a significant effect on the response. Finally, a confidence interval will be generated to show results.

IV. Regression Analysis, Results and Interpretation

We start by cleaning the data. Out of the 147 variables given in the data set, we are only concerned with 10. Using select() and na.omit(), we're able to obtain an organized dataframe with only the predictors we need. Next, we generate the scatterplot matrix for the response and all 10 predictors.



The scatterplot matrix doesn't reveal any trends between ViolentCrimesPerPop and any of the predictors. Multicollinearity is visible between the age variables, but this is to be expected as they are essentially overlapping variables. A look at the correlation matrix shows there is no other multicollinearity present.



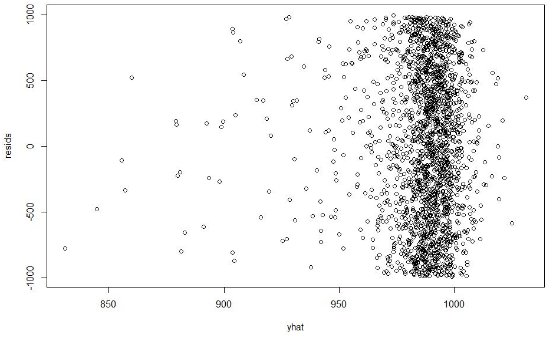
```
                  ViolentCrimesPerPop  population   medIncome PctUnemployed      PopDens householdsize PctPopUnderPov
ViolentCrimesPerPop      1.0000000000 -0.04573032  0.02783431    0.02180923 -0.016362141   0.004360456   -0.009640276
population              -0.0457303200  1.00000000 -0.04902656    0.08159569  0.213941249  -0.019981243    0.086877359
medIncome                0.0278343059 -0.04902656  1.00000000   -0.61936276 -0.034225408   0.181058952   -0.759860498
PctUnemployed            0.0218092259  0.08159569 -0.61936276    1.00000000  0.172495856   0.170724833    0.772145450
PopDens                 -0.0163621411  0.21394125 -0.03422541    0.17249586  1.000000000   0.031584545    0.070893355
householdsize            0.0043604560 -0.01998124  0.18105895    0.17072483  0.031584545   1.000000000    0.078581266
PctPopUnderPov          -0.0096402761  0.08687736 -0.75986050    0.77214545  0.070893355   0.078581266    1.000000000
agePct12t21             -0.0278115167 -0.00892256 -0.26263781    0.22388946 -0.070859216   0.486331416    0.486693017
agePct12t29             -0.0112547616  0.04603586 -0.32426590    0.19679725  0.096480898   0.376946616    0.461470196
agePct16t24             -0.0195356479  0.01748179 -0.28722684    0.15649385  0.028488095   0.308142645    0.459655636
agePct65up              -0.0005570178 -0.04535452 -0.25788046    0.10758750 -0.004201079  -0.577270200    0.077030325
```

To answer the first question, we will reduce the model to just two of the age predictors and determine whether or not a correlation exists.

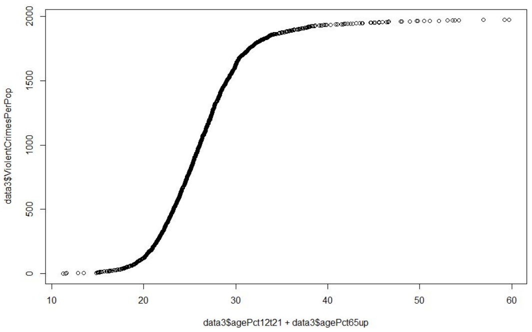
Question 1: Violent Crimes vs. Age Percentage

We constructed a simple linear regression model fitting ViolentCrimesPerPop to AgePct12t21 and AgePct65 up. AgePct12t29 and AgePct16t24 overlap heavily and are dropped from the model to avoid multicollinearity. The LINE conditions are checked next, using the residuals vs. fit plot, the Q-Q plot, and a histogram of the residuals.

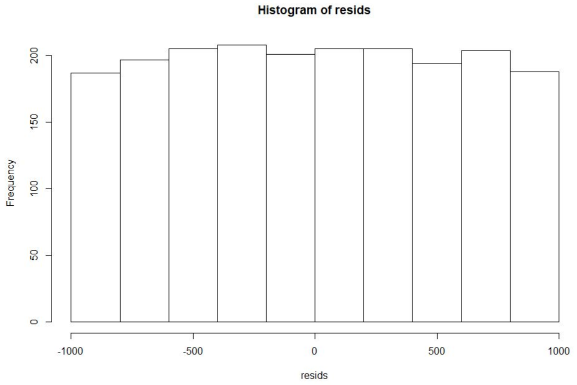
Residuals vs. Fit



Q-Q Plot

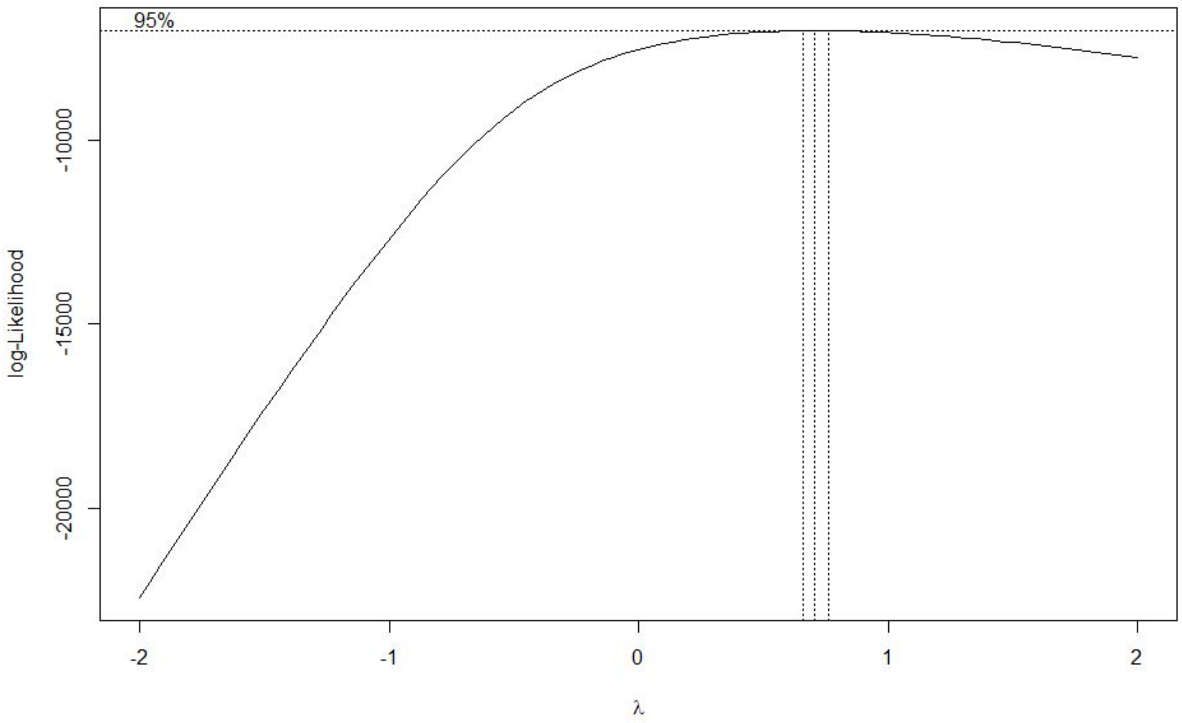


Histogram of Residuals

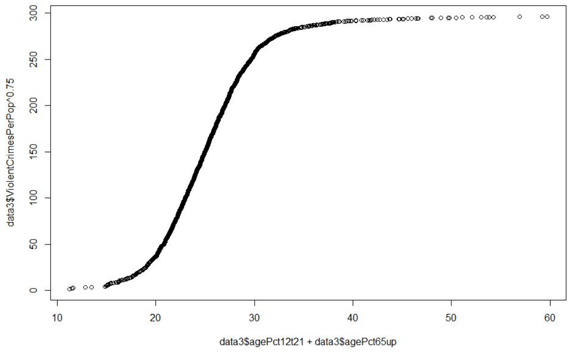


It is clear that there is a serious non-normality error, as can be seen in the Q-Q plot and histogram of residuals. The data seems to be extremely heavy-tailed. To attempt to combat this, we use a BoxCox transformation with a lambda value of 0.75.

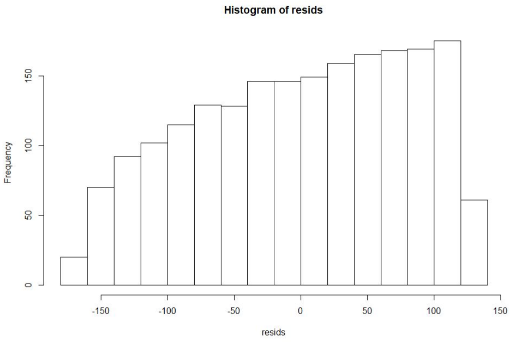
BoxCox Lambda Confidence Graph



Residuals vs. Fit

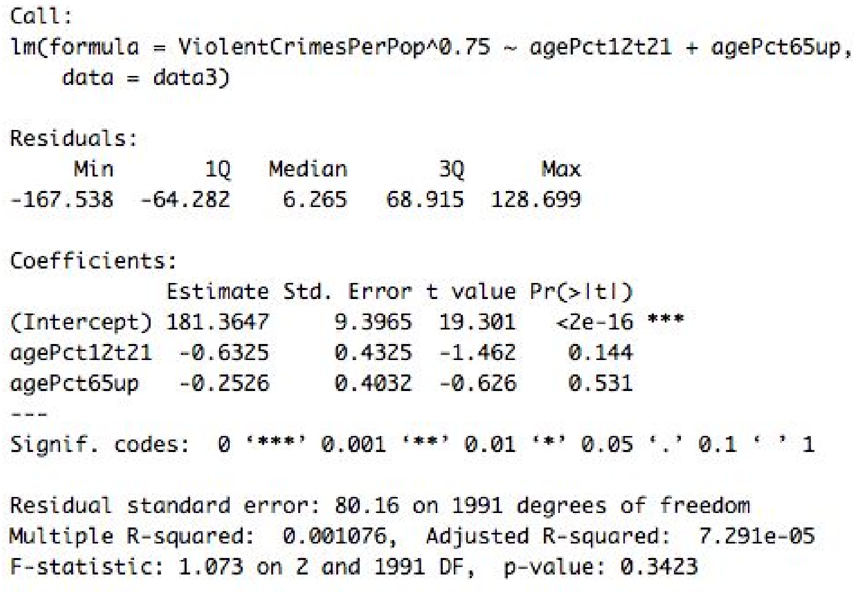


Histogram of Residuals



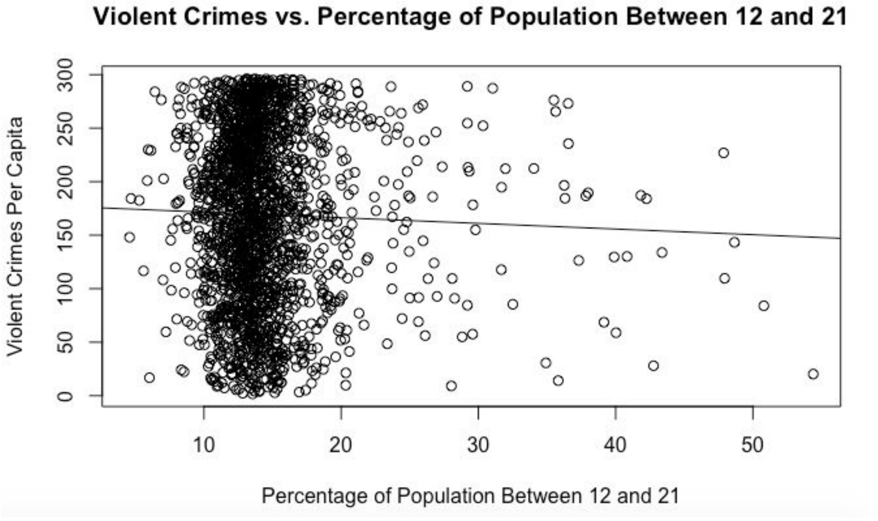
The issue of non-normality is not fixed with the BoxCox model. Seeing as the residual vs. fit graph is relatively spread around the 0 value (and we also are out of options to fix the non-normal data), we will ignore the issue of non-normality and proceed to performing tests on the model.

Stepwise regression is not necessary as there are only two predictors. As this is a relatively simple model, we will simply use t-tests and a global F-test to determine whether age percentage has an effect on predicting violent crimes per capita.



```
Call:
lm(formula = ViolentCrimesPerPop^0.75 ~ agePct12t21 + agePct65up,
    data = data3)

Residuals:
     Min       1Q   Median       3Q      Max 
-167.538  -64.282    6.265   68.915  128.699 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 181.3647     9.3965  19.301   <2e-16 ***
agePct12t21  -0.6325     0.4325  -1.462    0.144    
agePct65up   -0.2526     0.4032  -0.626    0.531    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.16 on 1991 degrees of freedom
Multiple R-squared:  0.001076,	Adjusted R-squared:  7.291e-05 
F-statistic: 1.073 on 2 and 1991 DF,  p-value: 0.3423
```

It appears there is very little correlation between percentage of people within a certain age range and violent crimes per capita. The p-values for the individual t-tests are not significant by the alpha value of 0.05, and the p-value for the global F-test is below 0.05 as well. Therefore we can conclude that neither the percentage of people between 12 and 21 and the percentage of people above the age of 65 have any noticeable effect on violent crimes per capita. This is also confirmed by the R^2 value, which tells us that the model with these two predictors explains 0.1% of the variance in violent crimes. A final graph of the transformed Y vs. one of the predictors shows almost no correlation, with the estimated regression line accounting for almost none of the response.

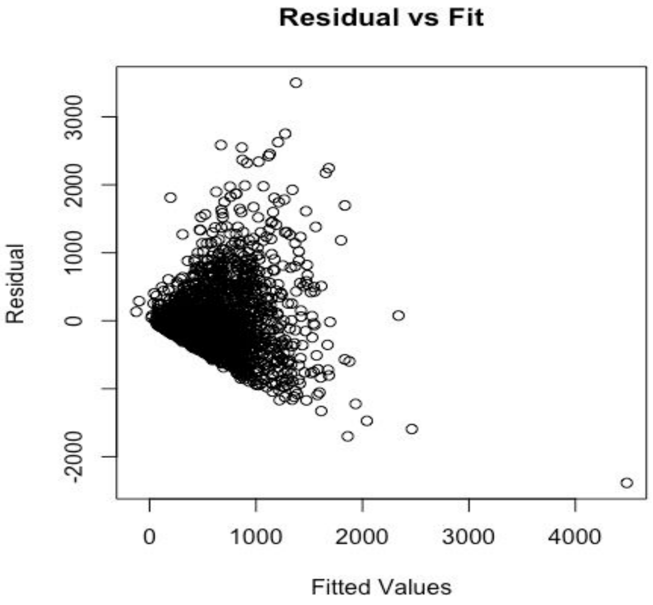


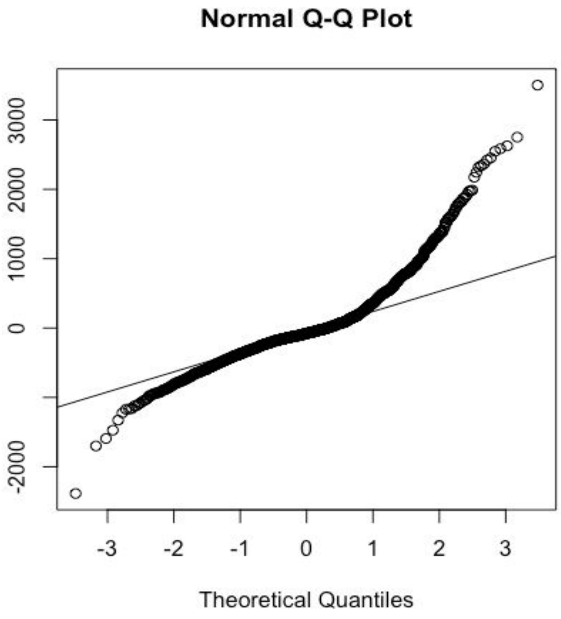
Question 2: Violent Crimes vs. Other Predictors

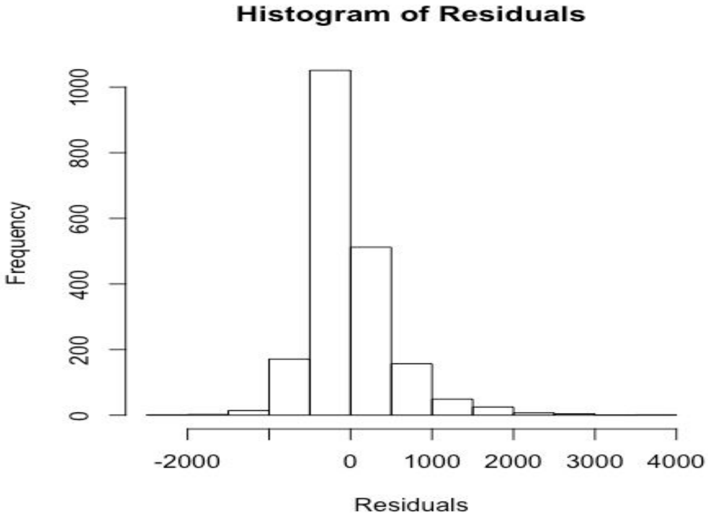
First, we used a general F test to indicate whether or not the remaining predictors had any significant effect on the response. The null hypothesis is that the reduced model is more effective so that **β**1=**β**2=**β**7=**β**8=**β**9=**β**10=0. The alternative hypothesis is that at least one of the slope parameters is not equal to 0. The resulting ANOVA table is given below.
Analysis of Variance Table

```
Model 1: Y ~ 1
Model 2: Y ~ x1 + x2 + x7 + x8 + x9 + x10
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   1993 753274276
2   1987 497836970  6 255437306 169.92 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

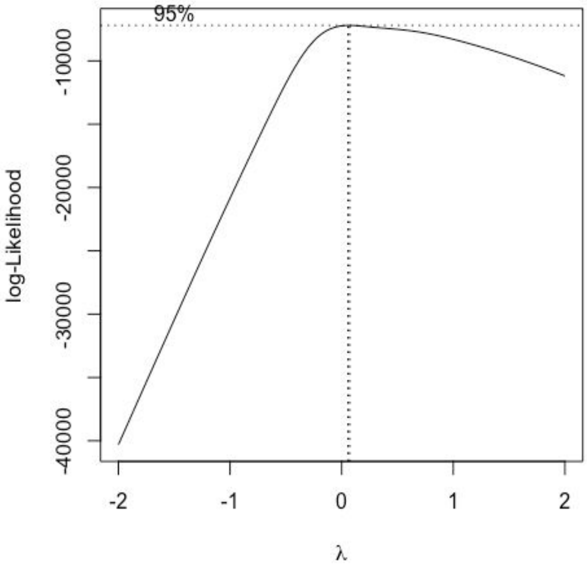
The F statistic for this test is 169.92 and the corresponding P-value is 2.2e-16. We have enough information to reject the null hypothesis since the P-value is <.05. This proves that the data for at least one of the predictors correlates to that of the response variable. Now, we must check the LINE conditions for the full model and decide whether we need to define a new set of predictors to reach the “best” model.

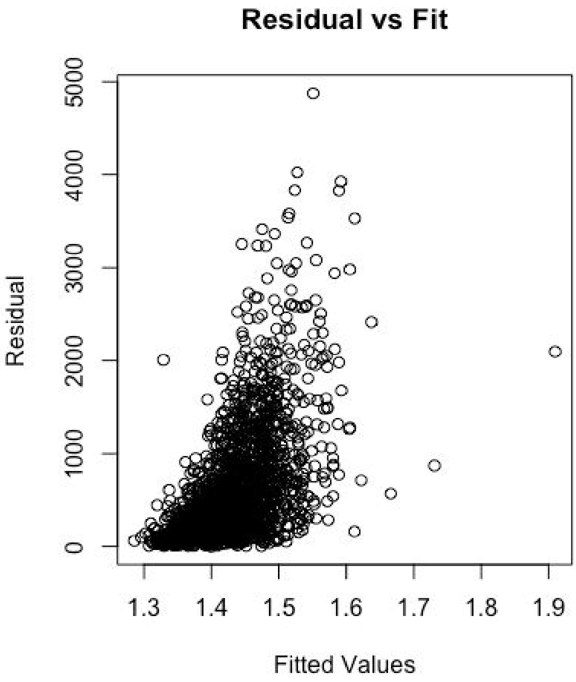


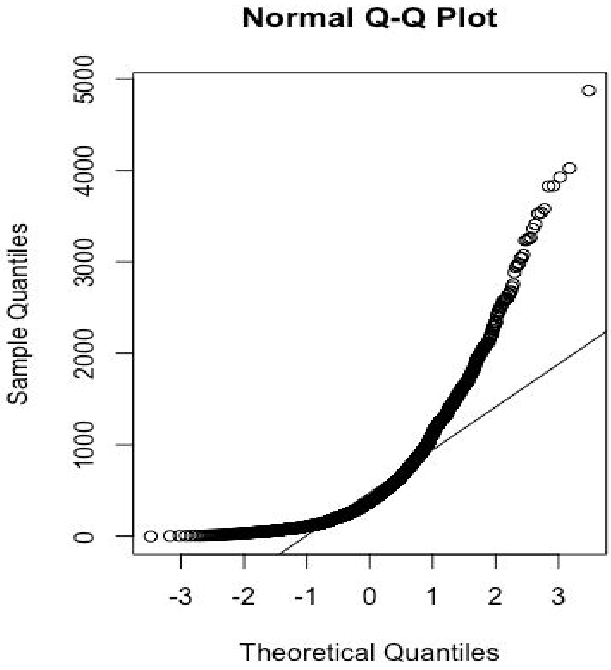


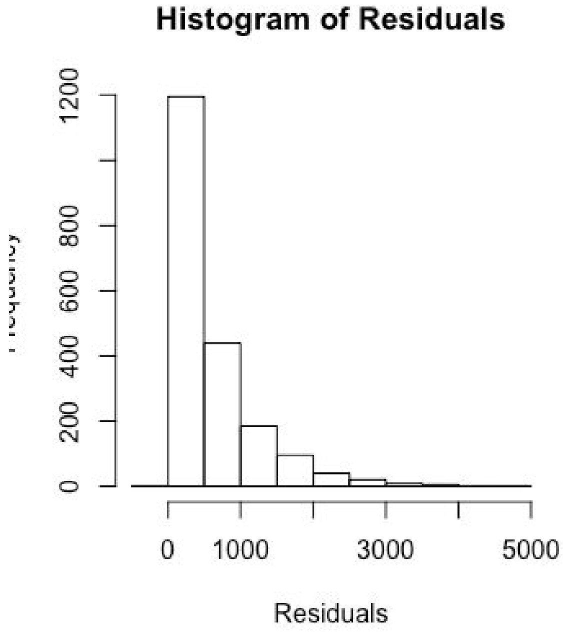


Looking at the residual vs fit graph there is a clear problem with variance due to the fact that there is a fanning of data points. The Q-Q plot also indicates a possible problem with normality because the points stray away from the line on both ends of the interval. The histogram shows a slight problem in normality as well, but not as severe as that of the age predictors. In order to combat these problems, we used boxcox in order to find a transformation of Y which would yield a stronger linear relationship.









```
Call:
lm(formula = tran2y ~ x1 + x2 + x7 + x8 + x9 + x10)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64331 -0.04824  0.00717  0.05234  0.25666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.441e+00  1.464e-02  98.450  < 2e-16 ***
x1           4.303e-08  8.732e-09   4.927 9.02e-07 ***
x2          -9.065e-07  2.156e-07  -4.205 2.73e-05 ***
x7           6.594e-03  1.039e-03   6.349 2.68e-10 ***
x8           5.934e-06  6.142e-07   9.662  < 2e-16 ***
x9          -2.817e-02  5.550e-03  -5.077 4.20e-07 ***
x10          3.062e-03  3.913e-04   7.824 8.22e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07668 on 1987 degrees of freedom
```

Using the boxcox transformation yields a transformation of log(y), and an analysis of the residuals vs fit plot and the Q-Q plot shows that some of the issues are rectified. Visually there is a higher potential for non-normality than the original graphs. Despite this, the adjusted R^2 value is larger, so there is an increase in strength of linear relationship. Overall the transformation does not indicate a significant positive effect on the data, so we attempted to find a better model by running stepwise regression with AIC, and best subsets regression on R^2 and Mallow's Cp.

```
Step:  AIC=24793.68
Y ~ x10 + x8 + x1 + x7 + x9

       Df Sum of Sq       RSS   AIC
<none>              497956677 24794
+ x2    1    119707 497836970 24795
- x9    1   5025900 502982577 24812
- x7    1  11191509 509148186 24836
- x1    1  11718423 509675100 24838
- x8    1  21010368 518967045 24874
- x10   1  33520664 531477341 24922

Call:
lm(formula = Y ~ x10 + x8 + x1 + x7 + x9)

Coefficients:
(Intercept)          x10           x8           x1           x7           x9  
  3.150e+02    2.423e+01    3.668e-02    3.896e-04    4.470e+01   -1.491e+02
```

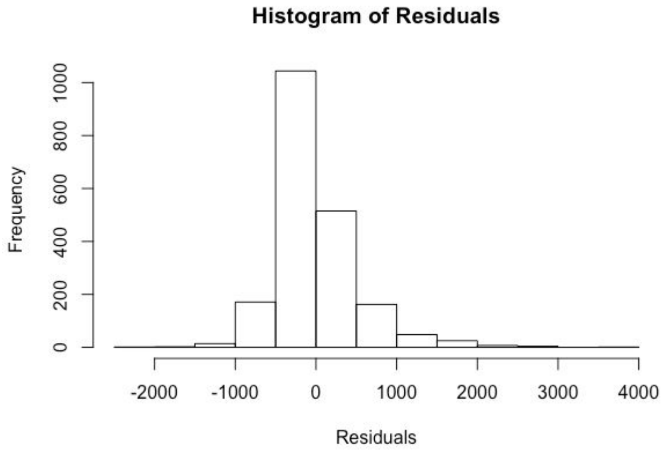
Using stepwise regression with AIC, the model deemed to be the best is one including the percent of the population under the poverty level,population density, population size, percent unemployed, and household size. The AIC alone may not provide enough proof that a subset is the best model, so next we did best subsets regression on R^2.

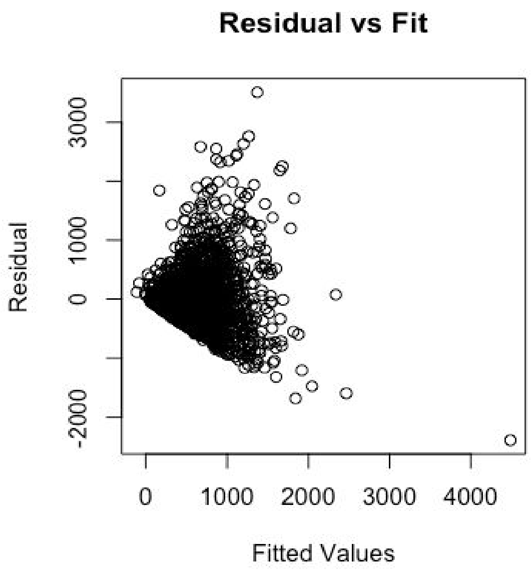
```
  (Intercept)    x1    x2    x7    x8    x9  x10
1        TRUE FALSE FALSE FALSE FALSE FALSE TRUE
2        TRUE FALSE FALSE FALSE  TRUE FALSE TRUE
3        TRUE  TRUE FALSE FALSE  TRUE FALSE TRUE
4        TRUE  TRUE FALSE  TRUE  TRUE FALSE TRUE
5        TRUE  TRUE FALSE  TRUE  TRUE  TRUE TRUE
6        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE TRUE

[1] 0.2550040 0.3038295 0.3192556 0.3309288 0.3372811 0.3371070
```

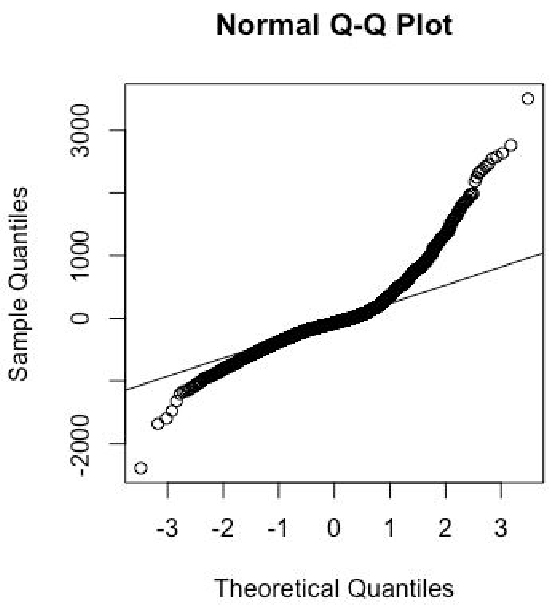
Considering best subsets regression with adjusted R^2 the model with the highest R^2 would be the same as the one indicated by stepwise regression with AIC. The largest increase in R^2 indicates the model with population density and percent of the population under the poverty level. All of the R^2 values are rather low; however, this is to be expected since we are only analyzing a small portion of the data set due to its size. The last test we ran was stepwise regression on Mallow’s Cp.

```
   (Intercept)   x1     x2     x7     x8    x9    x10
1     TRUE  FALSE FALSE FALSE FALSE FALSE TRUE
2     TRUE  FALSE FALSE FALSE  TRUE FALSE  TRUE
3     TRUE  TRUE FALSE FALSE  TRUE  FALSE  TRUE
4     TRUE  TRUE FALSE  TRUE  TRUE  FALSE  TRUE
5     TRUE  TRUE FALSE  TRUE  TRUE   TRUE   TRUE
6     TRUE  TRUE  TRUE  TRUE  TRUE   TRUE   TRUE

[1] 248.720192 102.948837  57.589717  23.537489   5.477783   7.000000
```

According to best subsets regression with Mallow's Cp the "best" model is, again, the model which excludes only median income. To be sure, we ran stepwise regression as well and came up with the same model. We then analysed the residual vs fit graph, Q-Q plot, and histogram of residuals under the new model.



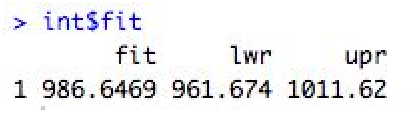




The residual vs fit model maintains a problem with variance, due to obvious fanning, and possible outliers. There is also still a problem with normality since the data points on the Q-Q plot stray from the line on the right-hand side. Although the data supports that this is the best model for the given predictors, it still does not indicate a strong linear relationship with the statistics for violent crimes per population.

```
Call:
lm(formula = Y ~ x1 + x7 + x8 + x9 + x10)

Residuals:
    Min      1Q  Median      3Q     Max 
-2390.5  -248.5   -74.6   143.8  3505.2 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.150e+02  8.996e+01   3.501 0.000474 ***
x1           3.896e-04  5.696e-05   6.840 1.05e-11 ***
x7           4.470e+01  6.688e+00   6.684 3.00e-11 ***
x8           3.668e-02  4.005e-03   9.159  < 2e-16 ***
x9          -1.491e+02  3.329e+01  -4.479 7.91e-06 ***
x10          2.423e+01  2.095e+00  11.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 500.5 on 1988 degrees of freedom
Multiple R-squared:  0.3389,	Adjusted R-squared:  0.3373 
F-statistic: 203.9 on 5 and 1988 DF,  p-value: < 2.2e-16
```

The p-values are all significant for the alpha value 0.05, so we reject the null hypothesis that all intercepts are 0. The correlation coefficient for this data states that the model is able to explain approximately 34% of the variance in violent crimes per capita. We conclude that there is some correlation between the predictors and response and that the six remaining predictors in the model are able to predict violent crimes per capita to some extent. However, the lack of proper LINE conditions and low R^2 value would dictate that using this model would not be very accurate, and prediction using it is not recommended. Even so, we construct a 95% confidence interval of a supposed community with all average values to conclude the study.



```
> int$fit
       fit     lwr     upr
1 986.6469 961.674 1011.62
```

According to the model, we are 95% confident that the mean value of violent crimes per capita of an average community lies between 986.65 and 1011.62.

V. Conclusion

In conclusion, the regression analysis shows that violent crimes per capita in United States cities and communities is not influenced by the percentage of people within a certain age range. However, tests revealed that it is able to be predicted by a combination model of population, unemployment rate, poverty rate, population density, and household size.

For the first question, we determined that the likelihood of the relationship between age range and violent crimes per capita being formed by random chance was extremely higher than usual, with the highest being a 53% probability, far higher than the usually recognized standard of 5%. Additionally, statistical analysis showed that the age range model was only accountable for a little over 0.1% of the spread of data for violent crimes. This forced us to conclude that there was no significant relationship between the age range percentages and violent crimes per capita.

For the second question, initial analysis determined that median income lowered the effectiveness of the model at predicting violent crimes per capita, so it was dropped from the model. Refining the resulting model, we determined that it was capable of explaining around 34% of the spread of violent crimes per capita. We also calculated that we can be 95% confident that the average violent crimes per capita of an average United States community would be between approximately 987 and 1011 crimes.

It is important to note that the first model suffered from severe non-linearity and would probably be better suited for a different type of model. Additionally, severe non-normality was detected in both models that was unable to be completely fixed, which suggests that the data and its models may not be accurate at determining a truly linear relationship. We are unsure if there were more advanced methods to fix and transform the data that would prove useful beyond the scope of this report.

Finally, the data set contains over 100 predictors, 10 of which we selected in the interest of answering specific research questions. Perhaps the best predictors of violent crimes per capita are among the unused predictors, but that is beyond the scope of the research questions behind this research project. We accept the results and conclude the study here.

VI. Appendix

```
#Part 1
library(dplyr)
data1 <- select(crimedata, ViolentCrimesPerPop, population, medIncome, PctUnemployed,
        PopDens, householdsize, PctPopUnderPov, agePct12t21, agePct12t29, agePct16t24,
agePct65up)
#removing missing values
idx <- data1 == "?"
is.na(data1) <- idx
data1 <- na.omit(data1)

#assigned values (Y changed after BoxCox)
Y=data3$ViolentCrimesPerPop ^ 0.75
x1=data3$population
x2=data3$medIncome
x3=data3$agePct12t21
x4=data3$agePct12t29
x5=data3$agePct16t24
x6=data3$agePct65up
x7=data3$PctUnemployed
x8=data3$PopDens
x9=data3$householdsize
x10=data3$PctPopUnderPov

#matrix and df stuff
data2 <- data.matrix(data1)
crimedata <- data.matrix(crimedata)
typeof(data2)
data3 <- data.frame(data2)
typeof(data3)

#scatter and correlation matrix
pairs(data2, main = "Scatterplot Matrix of Community Data")
cor(data2)
typeof(data1)

#residual vs fit
fit <- lm(data3$ViolentCrimesPerPop ~ data3$PctUnemployed)
resids = resid(fit)
yhat = fitted(fit)
plot(resids ~ yhat)
summary(fit)

#LINE tests
age <- lm(ViolentCrimesPerPop ~ agePct12t21, data = data3)
resids <- resid(age)
plot(resids ~ data3$ViolentCrimesPerPop)

#more LINE tests
age2 <- lm(Y ~ x1)
resids <- resid(age2)
hist(resids)
```

```
yhat <- fitted(age2)
plot(resids ~ yhat)
summary(age3)
anova(age3)

#BoxCox and tests
age3 <- lm(ViolentCrimesPerPop^0.75 ~ agePct12t21 + agePct65up, data = data3)
resids <-resid(age3)
yhat <- fitted(age3)
plot(resids ~ yhat)
plot(log(ViolentCrimesPerPop) ~ agePct65up, data = data3)
library(MASS)
boxcox(age3)

#residual vs fit again
resids <- resid(AIC)
yhat <- fitted(AIC)
plot(resids ~ yhat)

#t-tests and F-tests
summary(AIC)

#visuals
qqplot(data3$agePct12t21 + data3$agePct65up, data3$ViolentCrimesPerPop^0.75)
plot(ViolentCrimesPerPop ~ agePct12t21, data = data3)
plot(ViolentCrimesPerPop ~ agePct12t29, data = data3)
plot(ViolentCrimesPerPop ~ agePct16t24, data = data3)
plot(ViolentCrimesPerPop ~ agePct65up, data = data3)

#Part 2

#general F-test
reduced=lm(Y~1)
fullmodel=lm(Y~x1+x2+x7+x8+x9+x10)
anova(reduced,fullmodel)

#Check LINE conditions
full=lm(Y~x1+x2+x7+x8+x9+x10)
Yhat=fitted(full)
efull=Y-Yhat
plot(Yhat, efull, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')

qqnorm(efull)
qqline(efull)

hist(efull, main="Histogram of Residuals", xlab="Residuals")

#boxcox
```

```
library(MASS)
boxcox.trans=boxcox(fullmodel,data=data3,lambda= seq(-2, 2, length = 10))
best.lam=boxcox.trans$x[which(boxcox.trans$y==max(boxcox.trans$y))]
tran2y=Y^.0606061
fit3=lm(tran2y~x1+x2+x7+x8+x9+x10)
yhat3=fitted(fit3)
e3=Y-yhat3
plot(yhat3, e3, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')

qqnorm(e3)
qqline(e3)

hist(e3,main="Histogram of Residuals", xlab="Residuals")
summary(fit.model)
summary(fit3)

#AIC Test
mod0=lm(Y~1)
mod.upper=lm(Y~x1+x2+x7+x8+x9+x10)
step(mod0,scope=list(mod0, upper=mod.upper))

#subsets on R^2
install.packages("leaps")
library(leaps)
model=regsubsets(cbind(x1,x2,x7,x8,x9,x10),Y)
summary.model=summary(model)
summary.model$which
summary.model$adjr2

#Mallow's Cp
summary.model$cp
summary.model$which

#stepwise regression
mod0=lm(Y~1)
add1(mod0,~.+x1+x2+x7+x8+x9+x10, test='F')
mod1=update(mod0,~.+x10)
add1(mod1,~.+x1+x2+x7+x8+x9, test='F')
mod2=update(mod1,~.+x8)
summary(mod2)
add1(mod2,~.+x1+x2+x7+x9, test='F')
mod3=update(mod2,~.+x1)
summary(mod3)
add1(mod3,~.+x2+x7+x9, test='F')
mod4=update(mod3,~.+x7)
summary(mod4)
add1(mod4,~.+x2+x9, test='F')
mod5=update(mod4,~.+x9)
```

```
summary(mod5)
add1(mod4,~.+x2, test='F')

#residual vs fit
fit.model=lm(Y~x1+x7+x8+x9+x10)
yhat=fitted(fit.model)
e=Y-yhat
plot(yhat, e, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')

qqnorm(e)
qqline(e)

hist(e,main="Histogram of Residuals", xlab="Residuals")

summary(fit.model)
summary(fullmodel)

#confidence interval
x11 = mean(data3$population)
x71 = mean(data3$PctUnemployed)
x81 = mean(data3$PopDens)
x91 = mean(data3$householdsize)
x01 = mean(data3$PctPopUnderPov)
new = data.frame(x1 = x11, x7 = x71, x8 = x81, x9 = x91, x10 = x01)
int <- predict(fit, new, se.fit = TRUE, interval = 'confidence', level = 0.95, type = 'response')
int$fit
```

- END -