Predicting Violent Crime

Jesus I. Aguilar

5/2/2022

Introduction

The objective of this study was to build a linear regression model that could effectively predict violent crimes per population: total number of violent crimes per 100,000. The data set provided was Communities & Crime Unnormalized from the UCI Machine Learning Repository. This data set includes US data from the 1990 Census, 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR.

The following is a list of the initial predictors selected for the linear regression model:

- *PolicBudgPerPop*: police operating budget per population
- *RacialMatchCommPol*: a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar
- *medIncome*: median household income
- *PctFam2Par*: percentage of families (with kids) that are headed by two parents

Initial predictor selection was based on current discussions in the US in regards to crime and its relationship to law enforcement budget, relationship between law enforcement and the communities policed, and income.

Exploratory data analysis was performed to get an insight into a possible correlation between each predictor and violent crime. A linear regression model was fit to the data and model selection was performed using two different techniques. One utilized the p-value of each predictor as an exclusion criterion while the second method performed stepwise selection of the optimal model using the AIC value.

Mathematical assumptions of the model were checked through various diagnostic plots and the data was checked for outliers and influential observations using two different mathemical values calculated from the data, standardized residuals and *Cook's Distance*. After further tuning and refining the model, confidence and prediction intervals were calculated using the final model.

Exploratory Data Analysis

Each of the four predictors initially selected were plotted against violent crime to provide a visual representation of the possible correlation between predictor and violent crime, showing a negative correlation. The negative correlation with violent crime was obvious from the scatterplot for the following two predictors in particular:

- *RacialMatchCommPol*: a measure of the racial match between the community and the police force
- *PctFam2Par*: percentage of families (with kids) that are headed by two parents

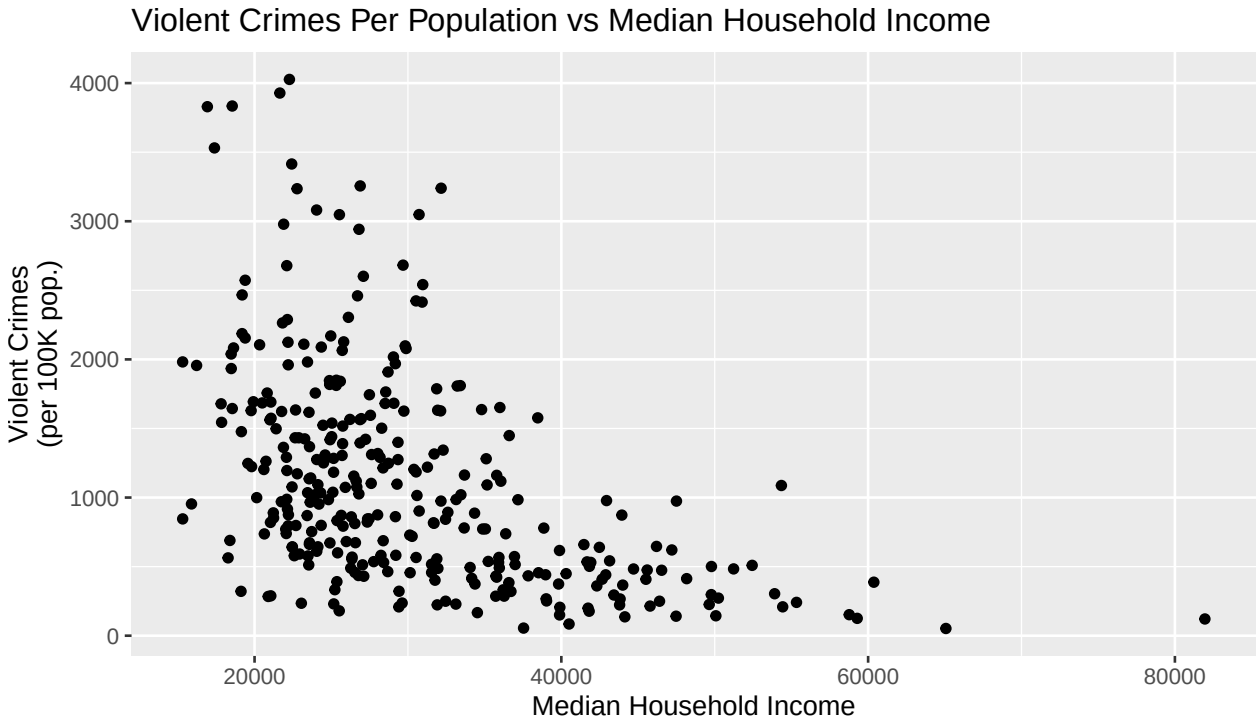

Figure 1: Potential negative correlation between median household income and violent crime

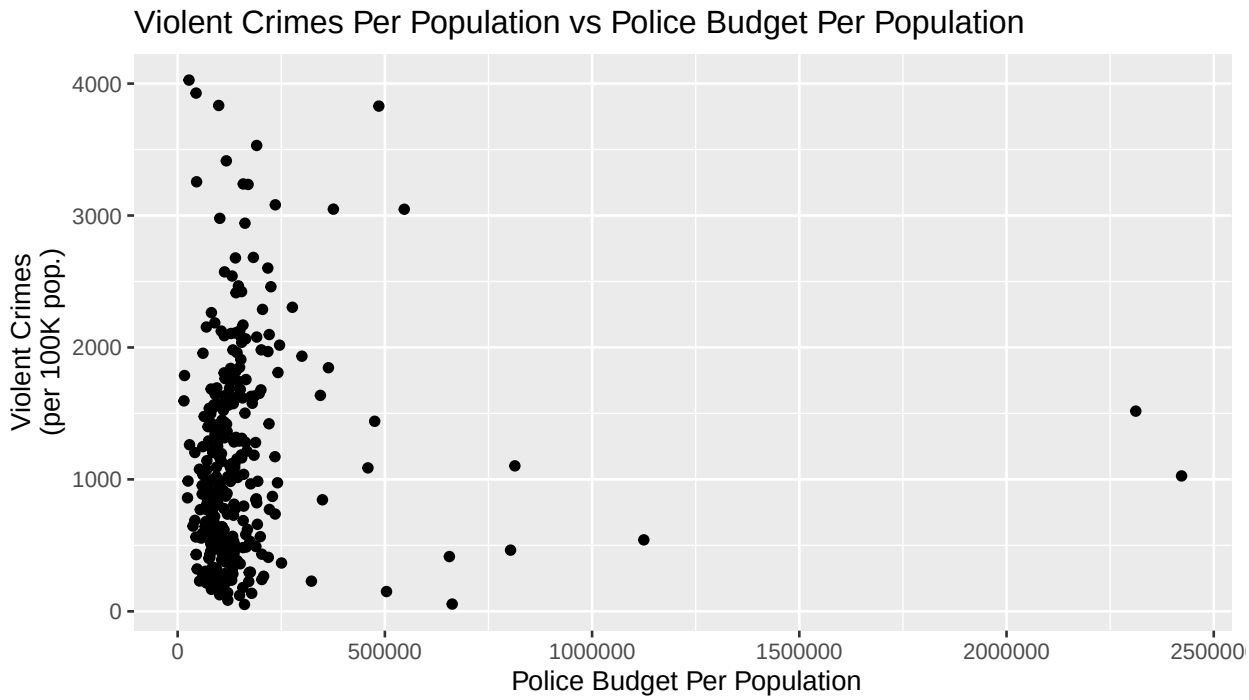

Figure 2: Potential negative correlation between police budget per population and violent crime, with two outliers having a large police budget.

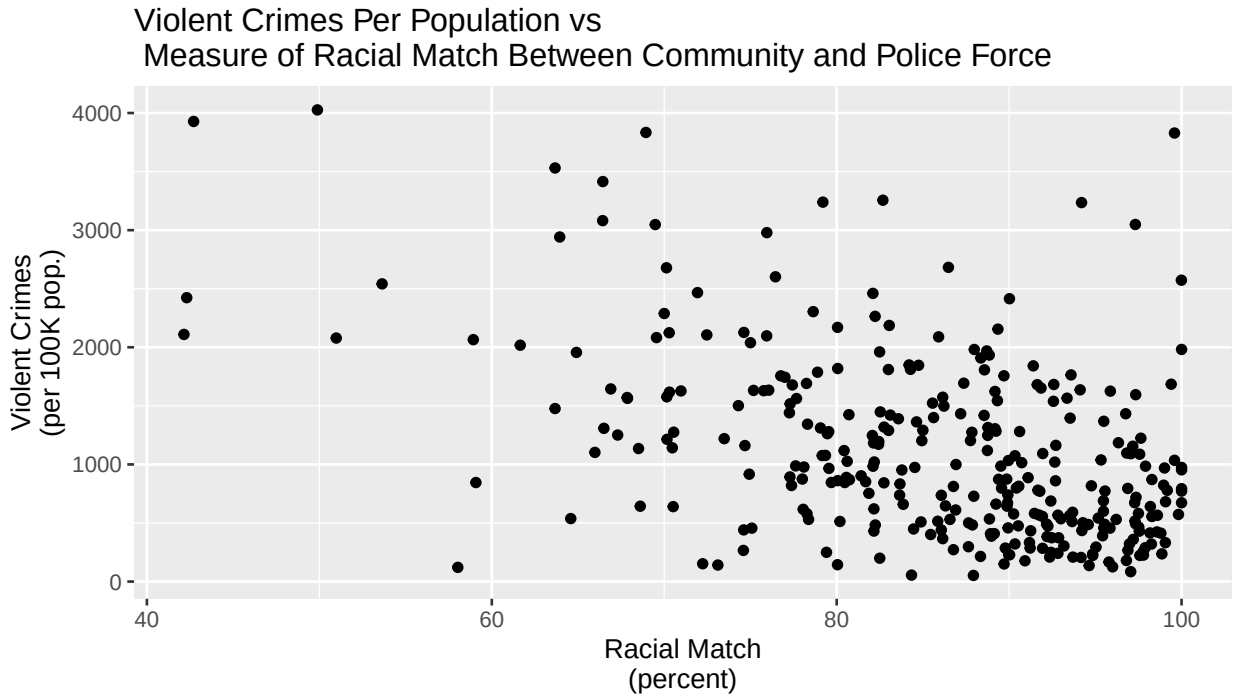

Figure 3: Potential negative correlation between violent crimes and a high racial match between community and police force

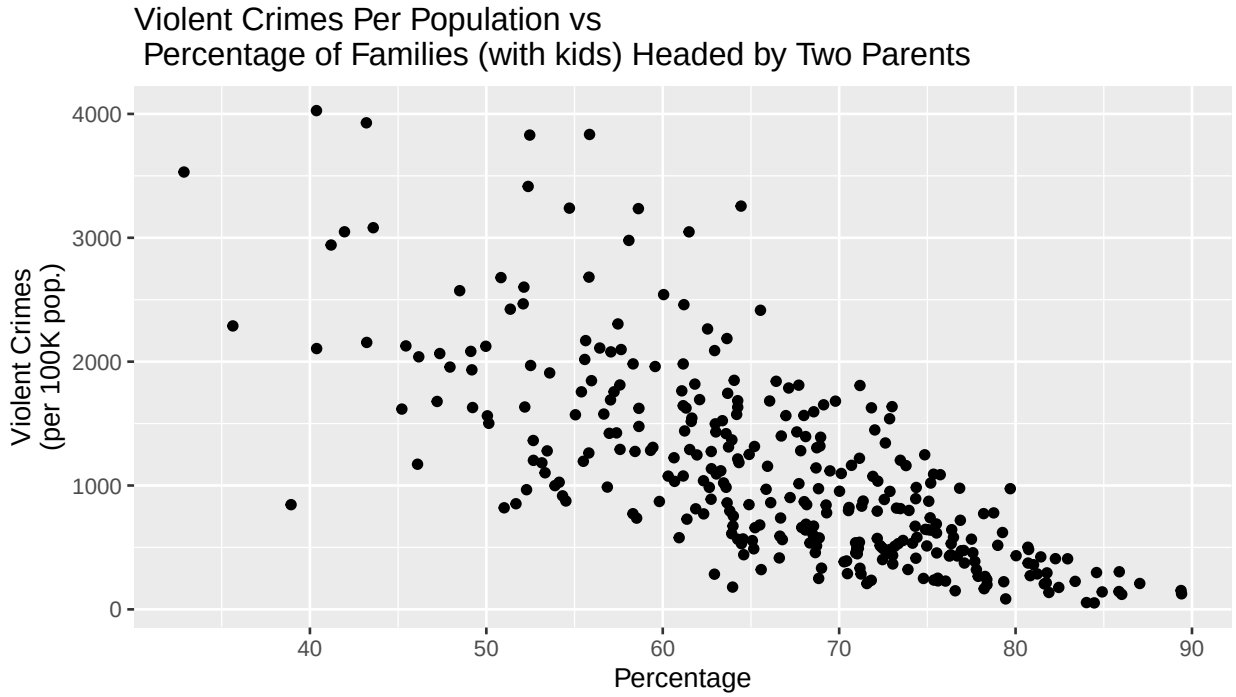

Figure 4: The scatter plot appears to show a negative correlation between violent crimes and percentage of families headed by two parents.

Correlation between all selected predictors and violent crime was expected, but the potential negative correlation between *PolicBudgPerPop*: police operating budget per population and *medIncome*: median household

income was not obvious in the plots. This stood out because police budgets in particular have become a contentious point of discussion when addressing crimes in the US. The association of income and crime is perhaps one that was expected to be a strong one, considering discussions in US discourse in which a correlation between the two is often implied.

In this study median household income was used as a predictor but perhaps it is not an ideal variable to represent poverty and income inequality. Both *PolicBudgPerPop* and *medIncome* had outliers that possibly affect data visualization of the correlation in the two plots or perhaps it is a much more complicated subject than political discourse makes it seem.

Fit A Linear Model

A linear regression model was fit using all of the four predictors. In Table 1 we can see a summary of the linear regression model.

Table 1: Linear Regression Model Summary

```
Call:
lm(formula = ViolentCrimesPerPop ~ PolicBudgPerPop + RacialMatchCommPol +
    medIncome + PctFam2Par, data = comm_crime)

Residuals:
     Min       1Q   Median       3Q      Max
-1920.03  -357.28   -73.74   252.13  2222.59

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.377e+03  2.556e+02  21.041  < 2e-16 ***
PolicBudgPerPop    -5.417e-05  1.450e-04  -0.374 0.708926
RacialMatchCommPol -1.251e+01  3.237e+00  -3.864 0.000136 ***
medIncome          -3.031e-03  4.601e-03  -0.659 0.510471
PctFam2Par         -4.665e+01  4.758e+00  -9.805  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 532.2 on 314 degrees of freedom
Multiple R-squared:  0.5464,    Adjusted R-squared:  0.5406
F-statistic: 94.57 on 4 and 314 DF,  p-value: < 2.2e-16
```

The p-value of each predictor is under the *Pr(>|t|)* column in Table 1. At $\alpha = 0.05$, *PctFam2Par* and *RacialMatchCommPol* appeared to be statistically significant predictors. The p-value for these predictors are listed below:

- *RacialMatchCommPolm*: p-value = 0.000136
- *PctFam2Par*: p-value = (< 2e-16)

The results in the model summary supported what was observed in the correlation plots, specifically Figure 3 and Figure 4. These plots showed that both *RacialMatchCommPolm* and *PctFam2Par* had a strong correlation to violent crime. $R^2 = 0.5464$ for the model and it was considered to be a value that indicated the model had significant predictive power. This value is represented as *Multiple R-squared* in Table 1.

Model Selection

The *fastbw()* and *stepAIC()* methods were applied to address model simplicity and select a model in which only variables with significant predictive power were used. The *fastbw()* method utilized the p-value of each predictor as an exclusion criterion. The *stepAIC()* method performed stepwise selection of the optimal model using the AIC value.

Table 2 shows that the fastbw() selection method excluded *PolicBudgPerPop* and *medIncome*, listed under *Deleted*. Only *RacialMatchCommPol* and *PctFam2Par* were selected as predictors.

Table 2: Model selection using *fastbw()* method

```
 Deleted         Chi-Sq d.f. P      Residual d.f. P      AIC   R2
 PolicBudgPerPop 0.14   1    0.7087 0.14     1    0.7087 -1.86 0.546
 medIncome       0.53   1    0.4677 0.67     2    0.7164 -3.33 0.545

Approximate Estimates after Deleting Factors

                      Coef    S.E.  Wald Z         P
Intercept          5364.64 252.027  21.286 0.0000000
RacialMatchCommPol  -11.88   3.087  -3.847 0.0001195
PctFam2Par          -48.78   3.252 -15.003 0.0000000

Factors in Final Model

[1] RacialMatchCommPol PctFam2Par
```

In selection by *stepAIC()* a predictor is removed at each step if there is a decrease in the AIC value of the model when the predictor is removed. Table 3 shows the initial model with all predictors had an AIC=4009.67 and the final model had an AIC=4006.34. The selected predictors for the model with the lowest AIC were *RacialMatchCommPol* and *PctFam2Par* under the *Coefficients* column. Both methods utilized for model selection resulted in *RacialMatchCommPol* and *PctFam2Par* as the two best predictors.

Table 3: Model selection using *stepAIC()* method

```
Start:  AIC=4009.67
ViolentCrimesPerPop ~ PolicBudgPerPop + RacialMatchCommPol +
    medIncome + PctFam2Par

                     Df Sum of Sq       RSS    AIC
- PolicBudgPerPop     1     39537  88966940 4007.8
- medIncome           1    122938  89050341 4008.1
<none>                             88927403 4009.7
- RacialMatchCommPol  1   4227742  93155145 4022.5
- PctFam2Par          1  27226193 116153596 4092.9

Step:  AIC=4007.81
ViolentCrimesPerPop ~ RacialMatchCommPol + medIncome + PctFam2Par

                     Df Sum of Sq       RSS    AIC
- medIncome           1    149365  89116305 4006.3
<none>                             88966940 4007.8
- RacialMatchCommPol  1   4284581  93251521 4020.8
- PctFam2Par          1  27740889 116707829 4092.4
```

```
Step:  AIC=4006.34
ViolentCrimesPerPop ~ RacialMatchCommPol + PctFam2Par

                     Df Sum of Sq       RSS    AIC
<none>                             89116305 4006.3
- RacialMatchCommPol  1   4191707  93308012 4019.0
- PctFam2Par          1  63746433 152862738 4176.5

Call:
lm(formula = ViolentCrimesPerPop ~ RacialMatchCommPol + PctFam2Par,
    data = comm_crime)

Coefficients:
       (Intercept)  RacialMatchCommPol          PctFam2Par
           5364.64              -11.88              -48.78
```

Model Diagnostics

A linear regression model was fit using *RacialMatchCommPol* and *PctFam2Par* as the only predictors based on the results for model selection and model diagnostics was performed to check the following linear regression model assumptions:

Constant Error Variance

Figure 5 showed heteroscedasticity was present in the model, indicating that the assumption of constant error variance was not upheld.

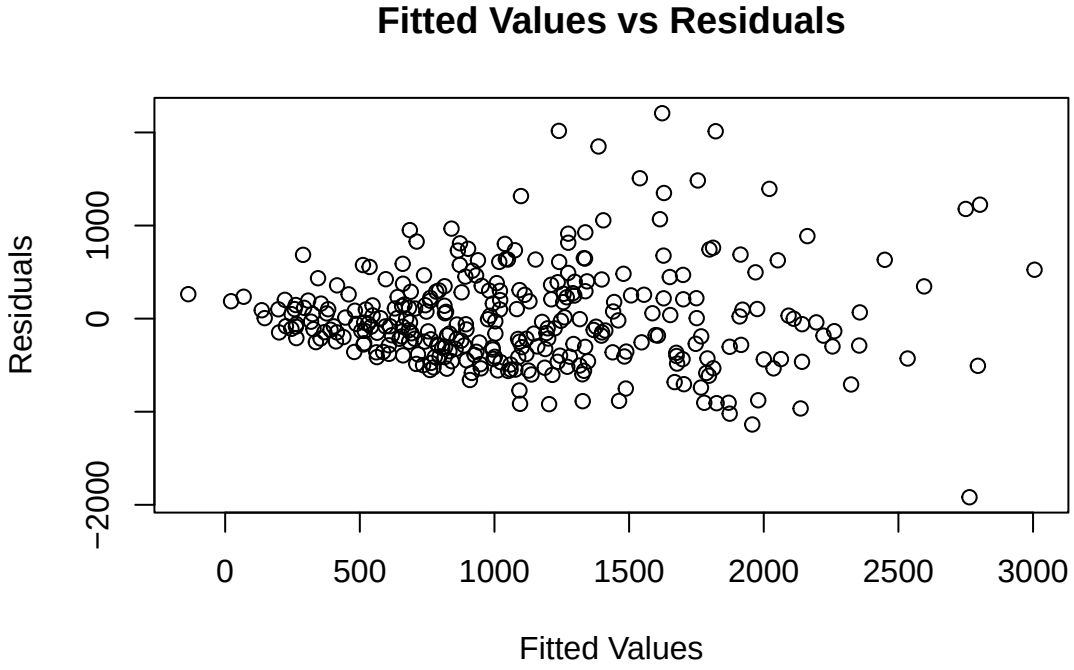

Figure 5: Heteroscedasticity in the initial model.

Normality of Errors

The QQ Plot in Figure 6 showed that there was no gross violation of linearity indicating that the assumption of normality of erros was upheld.

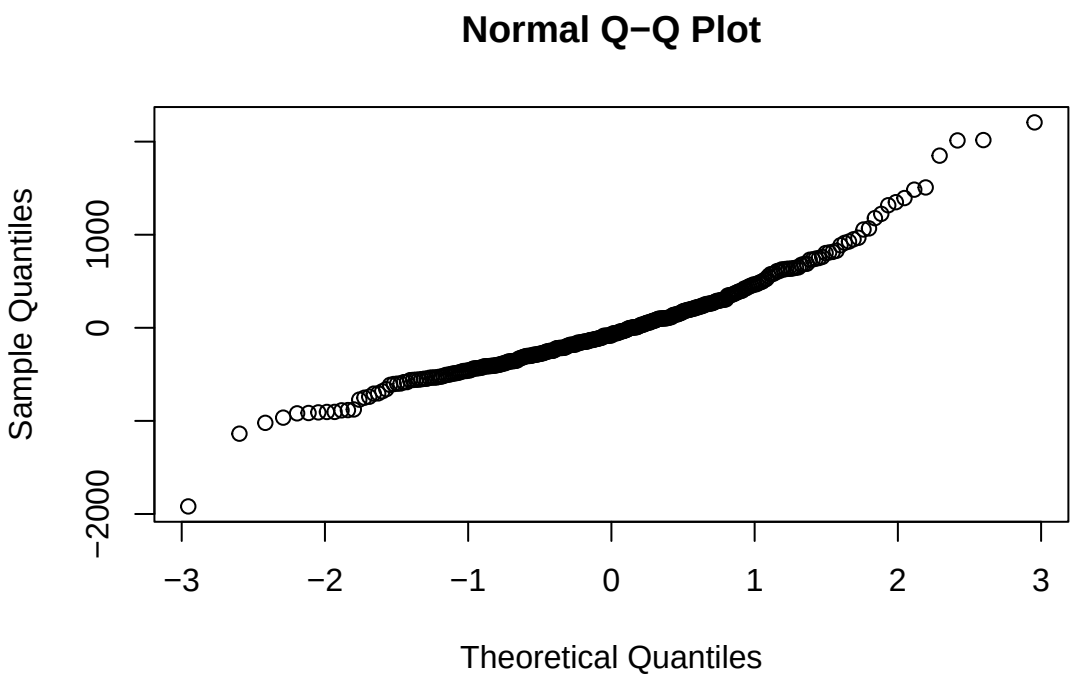

Figure 6: QQ Plot for normality of errors.

Independence of Errors

As observed in the lagged residual plot in Figure 7, there is no correlation between the residuals. This plot indicated that the assumption of error independence was upheld.

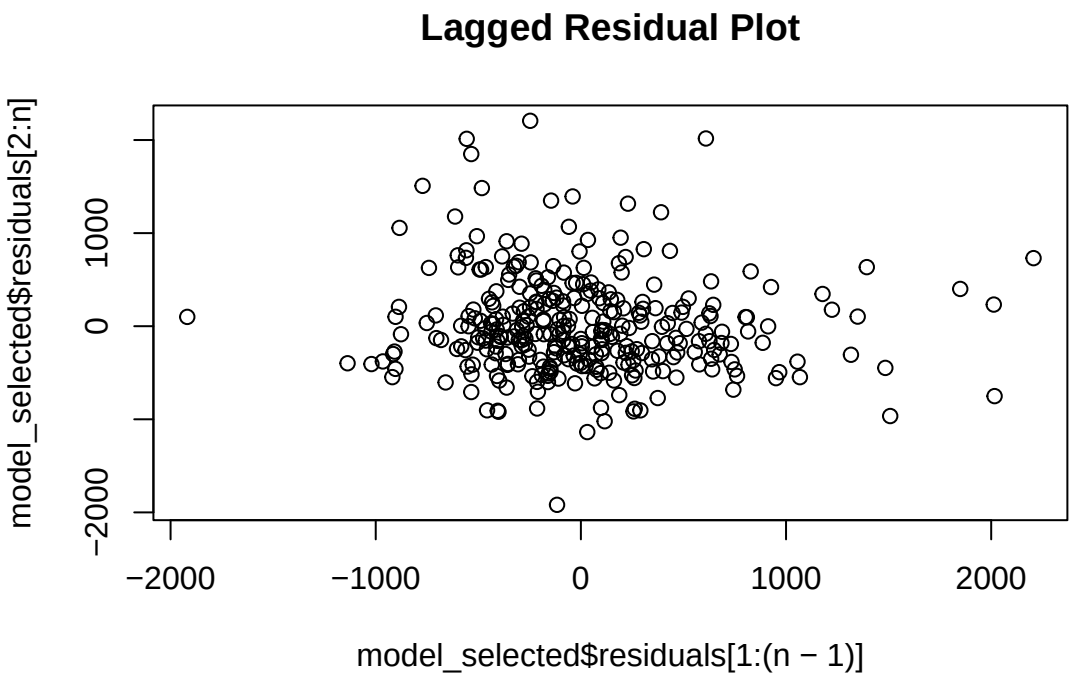

Figure 7: Lagged residual plot for independence of errors.

Investigate Fit for Individual Observations

The data was checked for outliers and influential points that could affect model fit. There were outliers in the data but no influential points so no data points were removed. The standardized residuals values were calculated and only three data points exceeded the threshold of 3, refer to Table 4, indicating that these were outliers in the data. *Cook's Distance* for each data point was calculated to check if any exceeded the F-statistic threshold. No Cook's distances were found to exceed the F-statistic threshold indicating there were no influential observations affecting model fit.

Table 4: Standardized Residuals Exceeding Threshold

	stan_resids
24	3.499805
79	3.803587
134	3.810371
190	-3.669071
266	4.204479

Model Transformations

It was determined that the response variable could be transformed through the Box-Cox transformation to address heteroscedasticity in the model, Figure 5. The goal of this transformation was to find a tuning parameter, λ, that would provide the most optimal approximation for normal distribution of the response variable in an attempt to fix heteroscedasticity. Figure 8 shows the 95% confidence interval for the value of λ. The value of $\lambda = 0.30$ and zero is not in the confidence interval denoted by the left and right vertical dashed lines in Figure 8. Therefore the response varible was transformed by raising it to the value of λ. The transformed response variable took the form of $ViolentCrimesPerPop^{\lambda}$.

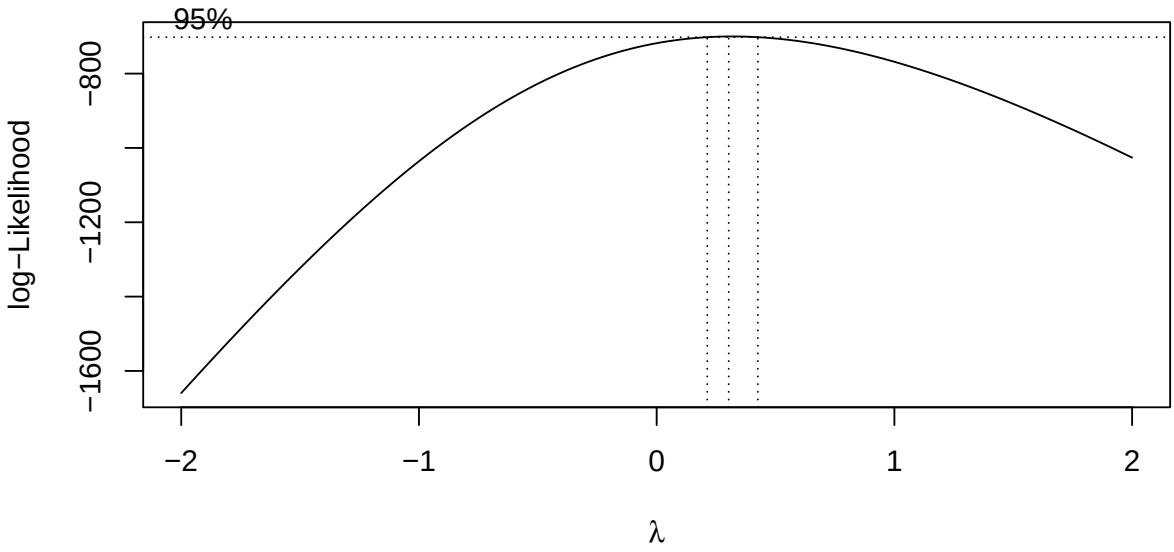

Figure 8: Confidence interval for λ

After transformation of the predicted variable the linear model assumptions were checked for the transformed model through the use of plots. All of the previous upheld assumptions, normality of errors and independence of errors were also upheld in the transformed model as observed in Figure 9 and Figure 10. Most importantly, Figure 11 shows that there was no heteroscedasticity present after transformation of the model. This indicates that the assumption of constant error variance was upheld too.

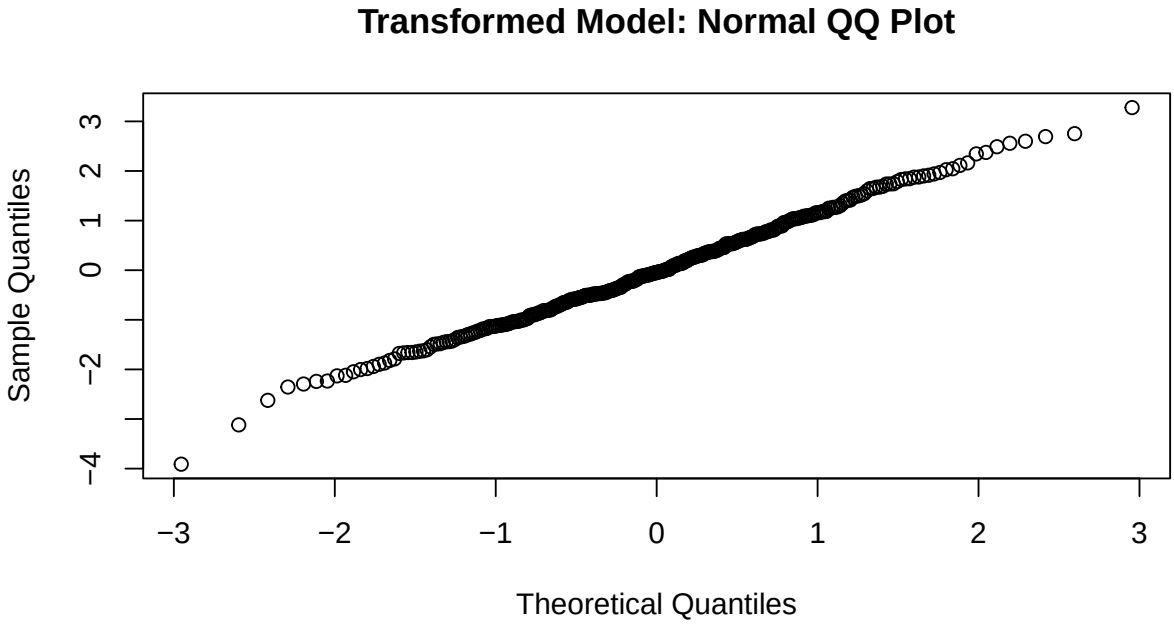

Figure 9: QQ Plot for transformed model.

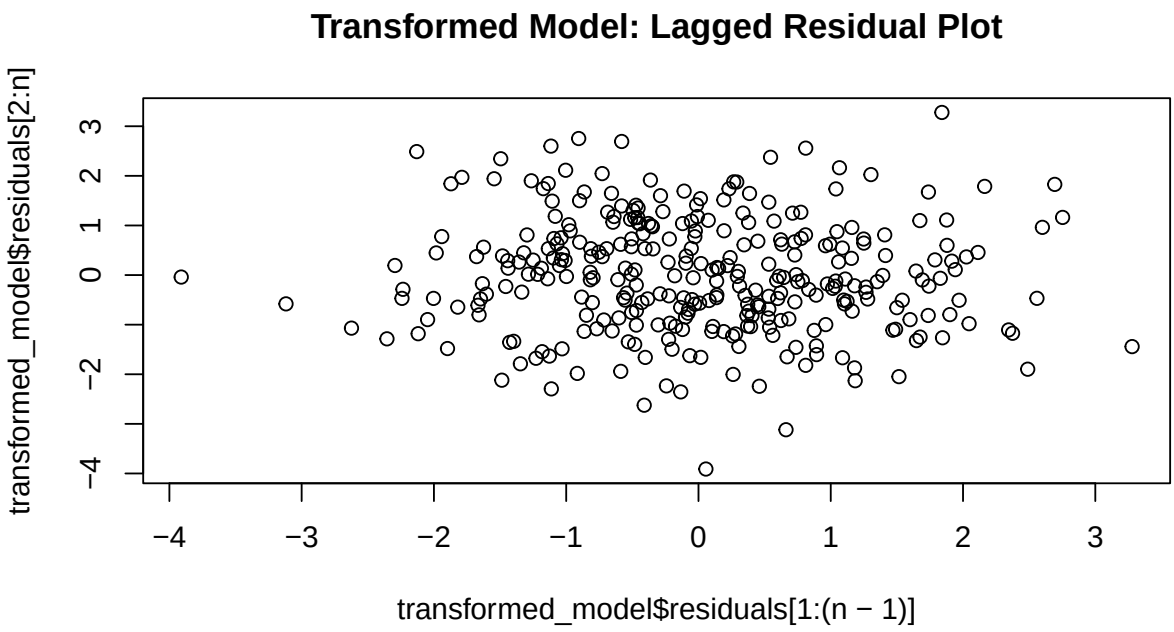

Figure 10: Lagged residual plot for transformed model.

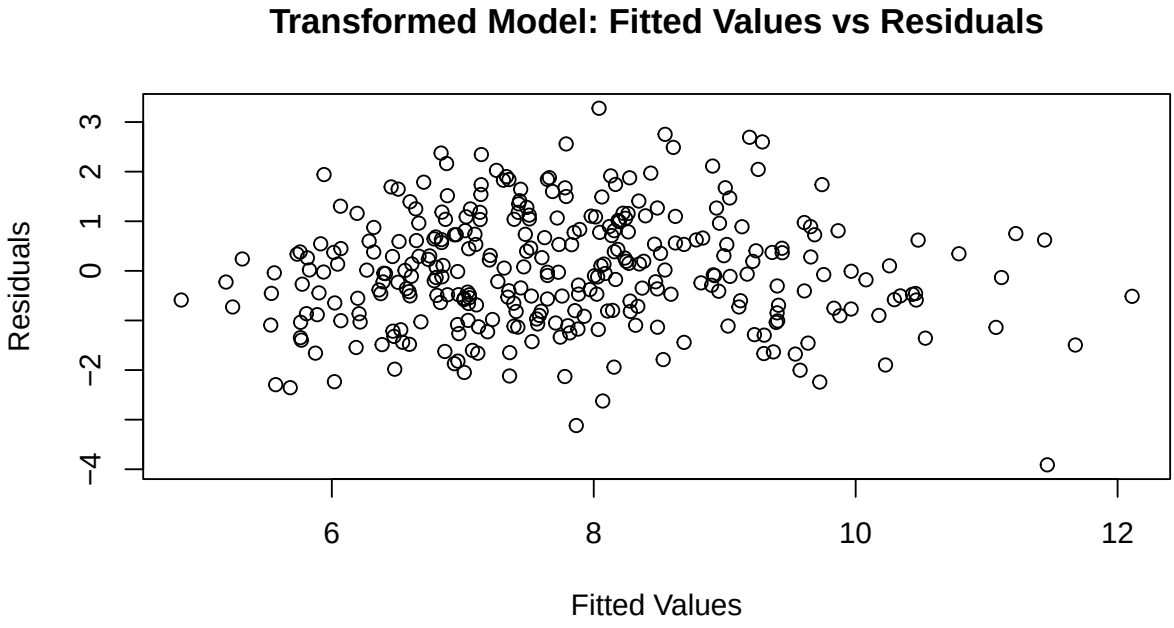

Figure 11: Absence of heteroscedasticity in the transformed model.

Report Inferences and Make Predictions Using The Final Model

The final model selected was: $ViolentCrimesPerPop^{\lambda} = \beta_1 RacialMatchCommPol + \beta_2 PctFam2ParX_2 + 17.52$

Table 5 lists the parameter estimates and p-values for the selected final model. The $R^2 = 0.576$ for the final

model, indicating significant predictive power.

Table 5: Parameter Estimates and *p*-values For The Final Model

```
        Predictor Estimate p.value
1    Racial Match  -0.0169   0.014
2 Two Parent Home  -0.1230  <2e-16
```

95% confidence interval for *PctFam2Par*

PctFam2Par appears to be the most important predictor based on the p-value. A 95% confidence interval for percentage of two parent home was calculated to be (-0.13, -0.11)

95% confidence interval for a prediction.

The median values for *RacialMatchCommPol* and *PctFam2Par* were selected to calculate a 95% confidence interval for a prediction. We are 95% confident that the true mean value of violent crime for all communities with *RacialMatchCommPol* = 87.95 and *PctFam2Par* = 67.22 lies in the interval (881.76, 989.01).

95% prediction interval for a particular observation

The median values for *RacialMatchCommPol* and *PctFam2Par* were also used to calculate a 95% prediction interval for a particular observation and we are 95% confident that value of violent crime for a particular community with *RacialMatchCommPol* = 87.95 and *PctFam2Par* = 67.22 lies in the interval (285.84, 2235.17).

Conclusion

Different statistics and diagnostics were used to determine model predictive power. Initially the model with all four predictors had an $R^2 = 0.5464$, this improved for the final transformed model as it increased to $R^2 = 0.576$. This shows that our transformed model had more predictive power than the first model. Another value worth noting is the Adjusted R^2 of the models. This value was used to assess model predictive powere versus model simplicity. Adjusted $R^2 = 0.5406$ in the first model and increased to Adjusted $R^2 = 0.5735$ in the transformed model. Both the R^2 and Adjusted R^2 values supported the conclusion that the final transformed model outperformed the first. The transformed model was a simpler model with more predictive power.

Exploratory data analysis showed the significant potential correlation between *RacialMatchCommPol* and *PctFam2Par* to violence through scatterplots. This was further supported by the results of model selection techniques *fastbw()* and *stepAIC()* in which the two variables selected as having significant predictive power were *RacialMatchCommPol* and *PctFam2Par*. This provides support that in the context of this study, the most appropriate model was selected.

Though the most optimal model was selected, there are limitations to this study. There is potential bias in initial variable selection as it was based on intuition considering current discourse in the US in regards to violence. There is a possibility that other predictors in the data were more optimal for the model but they were not selected for to fit the initial model. The result of this study showed a negative correlation between the predictors and violence, but this result should be put in context. It would not be ideal to base actions to address violence solely on the negative correlation observed, but it would be ideal to address potential underlying reasons that lead to this negative correlation. For example, in this study we could erroneously assume that the "traditional" family structure leads to less violence, but this assumption is biased and out of context. It could be ignoring non-traditional two parent households such as same-sex parents. *RacialMatchCommPol* showed that a high correlation in terms of race between police force and community had a negative correlation to violent crime. This could potentially be used to enact short-term solutions such as hiring more police from the community without addressing underlying reasons behind the correlation, such as race-relations. These two examples provide insight into further limitations of the study. Despite having an optimal model, the results should always be used in context.

Appendix

Table 4: Standardized Residuals

```
#STEP 1: EXPLORATORY DATA ANALYSIS (EDA)
suppressPackageStartupMessages(library(tidyverse)
suppressPackageStartupMessages(library(rms))
suppressPackageStartupMessages(library(MASS))

#Read in data
communities_crime <- read.table('~/Documents/LinearModels_SP22/DS64510_data_sets/CommViolPredUnnormaliz
#selected and renamed columns to be used for EDA
comm_crime <- communities_crime %>% dplyr::select(V146, V129, V112, V18, V49)
comm_crime <- comm_crime %>% rename(ViolentCrimesPerPop=V146, PolicBudgPerPop=V129, RacialMatchCommPol=
#removed NA values and converted data types as needed
comm_crime[comm_crime == '?'] <- NA
comm_crime <- comm_crime %>% drop_na()
comm_crime <- comm_crime %>% mutate_if(is.character, as.double)

#scatter plot Violent Crimes Per Population vs Median Household Income
ggplot(data = comm_crime, aes(x=medIncome, y=ViolentCrimesPerPop)) +
  geom_point() +
  labs(title = "Violent Crimes Per Population vs Median Household Income",
       x ='Median Household Income',
       y = 'Violent Crimes\n(per 100K pop.)') +
  theme(plot.margin = margin(0, 0, 0, 0, "cm"))

#scatterplot Violent Crimes Per Population vs Police Budget Per Population
ggplot(data = comm_crime, aes(x=PolicBudgPerPop, y=ViolentCrimesPerPop)) +
  geom_point() +
  labs(title = "Violent Crimes Per Population vs Police Budget Per Population",
       x =' Police Budget Per Population',
       y = 'Violent Crimes\n(per 100K pop.)')

#scatterplot Violent Crimes Per Population vs Measure of Racial Match Between Community and Police Forc
ggplot(data = comm_crime, aes(x=RacialMatchCommPol, y=ViolentCrimesPerPop)) +
  geom_point() +
   labs(title = "Violent Crimes Per Population vs\n Measure of Racial Match Between Community and Polic
       x ='Racial Match\n(percent)',
       y = 'Violent Crimes\n(per 100K pop.)')

#scatterplot Violent Crimes Per Population vs Percentage of Families (with kids) Headed by Two Parents
ggplot(data = comm_crime, aes(x=PctFam2Par, y=ViolentCrimesPerPop)) +
  geom_point() +
  labs(title = "Violent Crimes Per Population vs\n Percentage of Families (with kids) Headed by Two Par
       x ='Percentage',
       y = 'Violent Crimes\n(per 100K pop.)')


#STEP 2: FIT A LINEAR MODEL
#used lm to fit a linear regression model with selected predictors
model_initial <- lm(ViolentCrimesPerPop~PolicBudgPerPop+RacialMatchCommPol+medIncome+PctFam2Par, data =
#summary of the model
summary(model_initial)
```

```r
#STEP 3: MODEL SELECTION
#fit a linear model using ols and used fastbw() for variable selection
model1.ols <- ols(ViolentCrimesPerPop~PolicBudgPerPop+RacialMatchCommPol+medIncome+PctFam2Par, data = co
fastbw(model1.ols, rule = 'p', sls = 0.05)

#used stepAIC for model selection
stepAIC(model_initial)

#STEP 4: MODEL DIAGNOSTICS
#fit linear regression model using variables selected through the model selection process
model_selected <- lm(ViolentCrimesPerPop~RacialMatchCommPol+PctFam2Par, data = comm_crime)

#scatterplot of Fitted Values vs Residuals
plot(model_selected$fitted.values, model_selected$residuals,
     main="Fitted Values vs Residuals",
     xlab="Fitted Values",
     ylab="Residuals")

#QQ plot for normality of errors
qqnorm(model_selected$residuals)

#lagged residual plot for independence of errors
n <- dim(comm_crime)[1]
plot(model_selected$residuals[1:(n-1)], model_selected$residuals[2:n],
     main = "Lagged Residual Plot")

#STEP 5: INVESTIGATE FIT FOR INDIVIDUAL OBSERVATIONS

#calculated standardized residual and checked if any exceeded threshold value of 3.
stan_resids <- rstandard(model_selected)
data.frame(stan_resids) %>% filter(abs(stan_resids) > 3)

#calculated cooks distances and checked if any exceeded Fthreshold
distances <- cooks.distance(model_selected)
n <- dim(model.matrix(model_selected))[1]
p <- dim(model.matrix(model_selected))[2]
Fthreshold <- qf(0.5, p, n-p)
which(distances > Fthreshold)

#STEP 6: MODEL TRANSFORMATION

#used boxcox() to find the value of lambda
bc_mod <- boxcox(model_selected)
lambda <- bc_mod$x[which.max(bc_mod$y)]
#transformed the predicted variable using lambda
transformed_model <- lm((ViolentCrimesPerPop^0.30)~RacialMatchCommPol+PctFam2Par, data = comm_crime)

#QQ plot for transformed model
qqnorm(transformed_model$residuals, main = 'Transformed Model: Normal QQ Plot')

#lagged residual plot for transformed model
n <- dim(comm_crime)[1]
plot(transformed_model$residuals[1:(n-1)], transformed_model$residuals[2:n],
```

```
     main = 'Transformed Model: Lagged Residual Plot')

#fitted values vs residuals for transformed model
plot(transformed_model$fitted.values, transformed_model$residuals,
     main='Transformed Model: Fitted Values vs Residuals',
     xlab='Fitted Values',
     ylab='Residuals')



#STEP 7: INFERENCES AND PREDICTIONS WITH FINAL MODEL
#summary of the final transformed model
summary(transformed_model)

#created table fir parameter estimates and p-values for the final model
table <- data.frame(Predictor = c('Racial Match', 'Two Parent Home'),
                    Estimate = c(-0.0169, -0.123),
                     'p-value' = c(0.014, '<2e-16'))
table

#calculated 95% confidence interval for PctFam2Par
statistic <- qt(0.975, 316)
 -0.122710 + c(-1,1)*statistic*0.007217

#calculated 95% confidence interval for a prediction
racialmatch_median <- median(comm_crime$RacialMatchCommPol)
pctfam2par_median <- median(comm_crime$PctFam2Par)
predict(transformed_model, new = data.frame(RacialMatchCommPol = racialmatch_median, PctFam2Par = pctfa

#Applied inverse transformation function to the endpoints of calculated confidence interval
7.649^(10/3);7.917^(10/3)

#95% prediction interval for a particular observation
racialmatch_median <- median(comm_crime$RacialMatchCommPol)
pctfam2par_median <- median(comm_crime$PctFam2Par)
predict(transformed_model, new = data.frame(RacialMatchCommPol = racialmatch_median, PctFam2Par = pctfa

#Applied inverse transformation function to the endpoints of calculated prediction interval
5.455537^(10/3); 10.11098^(10/3)
```