Trends in Fatal Police Shootings Since 2015

Lee Panter

ABSTRACT

The proliferation of social media, internet-ready mobile devices, and powerful, easy-to-use camera technology have helped to organize and empower marginalized groups across the world. The effects of these technologies have been apparent in social, and political revolutions including the "Arab Spring" [1], the "Me Too" movement [2], and the "Black Lives Matter" movement[3]. The "Black Lives Matter" movement combined with these new technologies has focused more attention on the activities of police offices, in particular the circumstances surrounding Fatal Police Shootings (FPS). It stands to reason, therefore, that a statistical analysis of FPS incidents might be enlightening when compared against some of the natural biases of the social media platform. Specifically, this study proposes to analyze FPS incidents from 2015-2017 in an effort to identify correlated effects. The study will primarily analyze U.S. Census, firearm, and crime data in an effort to establish factors correlating with FPS incidences in the United States between 2015 and 2017.

INTRODUCTION

The expanded, and combined use of technology and social media has led to two possible, mostly unintended, consequences of social reform. From one perspective, radical and subjective societal perceptions have been exaggerated from highly publicized events, and social media content. From the other perspective, social and political injustice has been exposed through a medium of common access and influence. A major question of interest is determining which of these perspective (if either) is legitimate. Can Statistics be used to analyze the underpinnings of the social media movements in an effort to evaluate their legitimacy? Obviously, analyzing the movements from an ethical perspective must be taken into account as a motivating factor for the fundamental success of each movement. However, when taken purely as a numerical claim, are the motivating factors for these movements legitimate?

The "Black Lives Matter" (BLM) movement became a nationally recognized movement following the street demonstrations organized in response to the 2014 murders of two African American males: Michael Brown, and Eric Garner [4]. The BLM movement is an activist group that "campaigns against violence and systemic racism towards black people" and "holds protests, speaking out against police killings of black people" [5]. The motivation for such protests is that police violence against African Americans occurs in inordinately larger quantities than in comparison to other races and ethnicities. Since the BLM movement is a social media phenomenon, it is important to investigate the motivational factors of the movement. Are there more incidences of police violence in African Americans than in other races or ethnicities? Are there other factors that might be able to explain any increased occurrence in police-involved African American violence?

Approaches to answering these questions have been limited for numerous reasons. Police violence (in particular Fatal Police Shootings) were not recorded in a uniform manner across the United States until 2015. The sources of data which do exist on FPS from before 2015 generally require a Freedom of Information Act petition to obtain and are generally lacking some of the basic factors tracked in modern data. Previous records are also not consistent from state-to-state, or even county-to-county, which leads to extreme difficulty in trying to find data that is actually useful.

The Washington Post has been keeping track of FPS incidences from 2015 in the Police Shootings Database (PSD). This data set represents the most extensive and complete database of FPS incidences yet compiled [6]. Yet any study completed by the Washington Post is limited to summary statistics and inference

[1](Wikipedia, 2018)
[2](Wikipedia, 2018)
[3](Wikipedia, 2018)
[4](Wikipedia, 2018)
[5](Wikipedia, 2018)
[6](washingtonpost.com, 2018)

drawn directly from the data itself [7]. The Washington Post analysis found summary statistics along the line of: in December,2016 the number of FPS incidences in a year had remained essentially unchanged over the past two years (since the start of 2015). A similar summary would be: in FPS occurrences prior to July, 2016, one in five police officers' names went undisclosed.

A more-narrowly focused study was published in the Journal of Criminal Justice in September, 2017. The study specifically focused on whether or not African Americans had a higher FPS incidence than other races or ethnicities. This study also used the Washington Post PSD, and concluded: "Although the data are limited, the patterns are not consistent with the national rhetoric that the police are killing Black people because of their race and that officer-involved shooting fatalities are increasing". [8]

A more comprehensive study was conducted by the Boston University School of Medicine, in which FPS rates in African Americans were compared against measures of pre-existing structural racism across U.S. states. The study concluded: "states with a greater degree of structural racism, particularly residential segregation, have higher racial disparities in fatal police shootings of unarmed victims." [9]

In an effort to better understand the factors contributing to a Fatal Police Shooting this study proposes to compare the data in the Washington Post Police Shootings Database against demographic data obtained from the U.S. Census [10], as well as firearm and crime data obtained from demographicdata.org [11]. The study proposes to answer the question: Is the quantity of FPS incidences in each state related to state demographics, state gun statistics, or state crime statistics?

METHODS

The Original Data

Washington Post Police Shooting Database

The Washington Post Police Shootings Database (WPPSD) is a compilation of the police involved fatalities since January 1st, 2015. The database includes every FPS incident across all 50 States in the United States, and distinguishes those shot in Washington D.C., but does not include shootings in Puerto Rico or Guam. The WPPSD includes only those incidences in which "a police officer, in the line of duty, shoots and kills a civilian". In particular, this eliminates police-related deaths where people were already in police custody, or when people were shot by off duty police officers. The WPPSD keeps tracks of factors related to each of the shootings, including information about the victim, and circumstances under which the shooting took place.

Factors tracked about the victim include:

- Gender: Male/Female/Unknown
- Race: White/Black/Hispanic/Other/Unknown
- Age: Numerical in Years
- Signs of Mental Illness: Yes/No or Unknown

Factors tracked about the shooting circumstances include:

- State of Shooting: Two Letter State Abbreviation
- City of Shooting: Name of City in which shooting took place
- Victim's Weaponry: Gun/Knife/Vehicle/Toy Weapon/Other/Unarmed/Unknown
- Officer Wearing Body Camera?: Yes/No or Unknown

[7] (washingtonpost.com, 2018)
[8] (sciencedirect.com, 2018)
[9] (sciencedaily.com, 2018)
[10] (census.gov, 2018)
[11] (demographicdata.org, 2018)

- Method Victim was Attempting to Flee: Car/Foot/Other/Not Fleeing/Undetermined
- Officer of Record for Death Identified: Yes/No or Unknown

At the time the data sourced from Kaggle [12] (April, 16, 2018) There were a total of 2535 total shooting incidences reported.

United States Census Data

The United States Census Bureau is responsible for conducting surveys every ten years which accumulate relevant statistics relating to demographic, housing, economic, and welfare information. The statistics are compiled on a national, state and city level, and are published for public use on the U.S. Census Bureau website [13]. The primary investigative purpose of this experiment is to attempt to find outside data sources (outside of the WPPSD) that will demonstrate correlated effects with the per-state incidence of FPS measure recorded in the WPPSD. Therefore, several state-level census statistics will be considered for the analysis. These statistics include:

U.S. Census State-Statistics Used in Model

- Population (Number of people per state) (2010 U.S. Census)
- Poverty Level (Percent of people living below Federal Poverty Line in each state) (2016 American Community Survey)
- High School Graduation (Percent of people in each state with a High School degree (or equivalency degree)) (2016 American Community Survey)
- People Under 45 (Percent of people in each state that are below the age of 45) (2010 U.S. Census)
- Race Compositions (Percent of people in each state who are the following race/ethnicities) (2010 U.S. Census)
  - White
  - Black
  - Asian
  - Native American
  - Hispanic
  - Other
- Number of males/100 females ratio (Number of males per 100 females in each state) (2010 U.S. Census)

All U.S. Census data was sourced directly from the U.S. Census Bureau [14], and each of the above variables was sourced for each state corresponding to an observation from the WPPSD database.

Gun and Violence Data

The last source of original data used for this analyis is from DemographicData.org [15]. This is an open-source database that compiles statistics about gun and crime data in the United States, from various open source databases such as the Stanford Libraries, GunViolenceArchive.org and Mother Jones. The data set used for this analysis contained categories reflecting the Gun Murder Rate per 100,000 people in each state in 2010, and the Gun Ownership rate in 2007 (expressed as a percent of the total population) in each state. As was the case in the US Census data, each state corresponding to an observation from the WPPSD had two sourced statistics from the DemographicData.org database.

[12](kaggle.com, 2018)
[13](census.gov, 2018)
[14](census.gov, 2018)
[15](demographicdata.org, 2018)

The Final Data

Since a listing of the occurences of each FPS, as it happened, is not going to be helpful in the contextual analysis of state-level factor analysis, the data needed to be condensed so that a single state recorded the total number of FPS from 2015-2017.

In addition, the modeling variables (regressor variables), were merged into the same data frame so as to make the modeling process easier down the line. A data head is provided below to give an idea of what the final data looks like.

```
head(df)
```

```
##    totshootstate StatePop Poverty   HS White Black Asian NativeAmerican
## AK            15  4779736    18.8 84.3  68.5  26.2   1.1            0.6
## AL            50   710231    10.2 92.1  66.7   3.3   5.4           14.8
## AR            26  6392017    18.2 86.0  73.0   4.1   2.8            4.6
## AZ           118  2915918    19.3 84.8  77.0  15.4   1.2            0.8
## CA           424 37253956    16.3 81.8  57.6   6.2  13.0            1.0
## CO            74  5029196    12.7 90.7  81.3   4.0   2.8            1.1
##    Hispanic Other NumMurders GunOwners Under45 PercMale
## AK      3.9   3.6        2.8     0.517    59.4     94.3
## AL      5.5   9.9        2.7     0.578    64.5    108.5
## AR     29.6  15.5        3.6     0.311    61.7     98.7
## AZ      6.4   5.6        3.2     0.553    59.5     96.5
## CA     37.6  22.3        3.4     0.213    63.7     98.8
## CO     20.7  10.7        1.3     0.347    62.5    100.5
```

Response and Predictors of Interest

The response of interest in this study will be the total number of FPS occurences in each state between the years of 2015 and 2107. Since this is an observational study, no attempt at causal inference will be made.

Descriptive Summary Measures

The data set that this study proposes to analyze concerns a total of 51 observations on 13 variables. The 51 observations correspond to the 51 states in the U.S. and the 13 variables correspond to:

(1) State Population

(2) Poverty Level

(3) H.S. Graduation

(4) Perc. White

(5) Perc. Black

(6) Perc. Asian

(7) Perc. Native American

(8) Perc. Hispanic

(9) Perc. Other

(10) Number of Gun Murders

(11) Perc. Gun Owners

(12) Perc. Under 45

(13) Perc. Male

The Results section that follows will provide summary diagrams and tables which will help to demonstrate how the value of each state's FPS measure varies as the 13 variables are varied. Summaries of such effects will include: numerical 5 number summaries of each of the 13 variables across all the state observations, and graphical demostrations of each variable across all the state obseravations.

Statistical Analysis Methods

The observational nature of the Washington Post Police Shooting Data Base, and the US Census data, leads to the natural conclusion that this study should be approached from the perspective of regression analysis. That is, the most appropriate analysis is one that assumes the least about the underlying observational and measurement structure. The WPPSD database being studied in this analysis is confined to a three year window that captured a "glimpse" into the reality of the true nature of police violence. In comparing the WPPSD to other Cross-Sectional databases and studies (like the U.S. Census and the DemographicData.org database), it is concievably reasonable to further expand the basic conclusions drawn from this analysis to the time periods over which the databases were mutually responsible. However, since the WPPSD reflects observations made from 2015-2017, the US Census data is either from 2010 or 2016, and the Demographic-Data.org data is sourced from 2010 and 2007, the data shares no common date of reference, which relagates the study to purely intellectual and hypothetical purposes.

Nevertheless, an attempt to create a reasonable model based upon the 13 predictors of interest mentioned in the "Desciptive Summary Measures" section will be made. Predictors of interest will be tested for significance using a two sided t-test for the hypothesis test framework:

$$H_0 : \ \beta_i = 0$$

Where

$$i \ \in \ \{1, 2, \ldots, 13\}$$

and each i corresponds to the enumeration in the "Desciptive Summary Measures" section.

Further more, we will define the test statistic:

$$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

The model will also be tested for collinearity using Variance Inflation Factors (removing Regressors with VIF scores above 5).

All tests will be performed with a significance level of $\alpha = 0.1$, in R-Studio version 1.1.383.

RESULTS

Initial Data Analysis

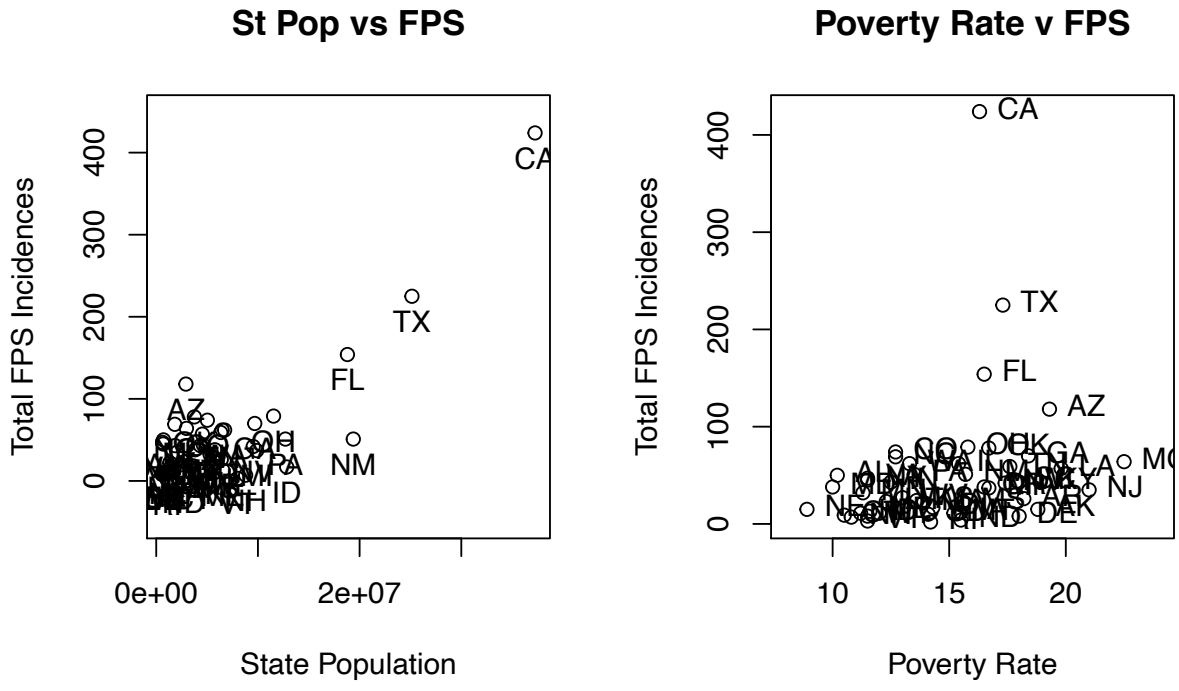
This section is devoted to the summarization and vizualization of the final data set described in the "Methods" section.

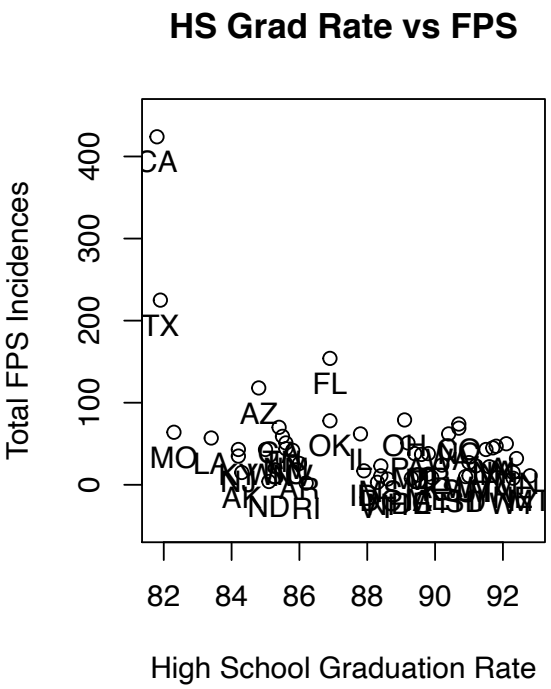
The five number summaries displayed below descibe each of the predictors of interest, as well as the total number of FPS incidences in each state between 2015-2017

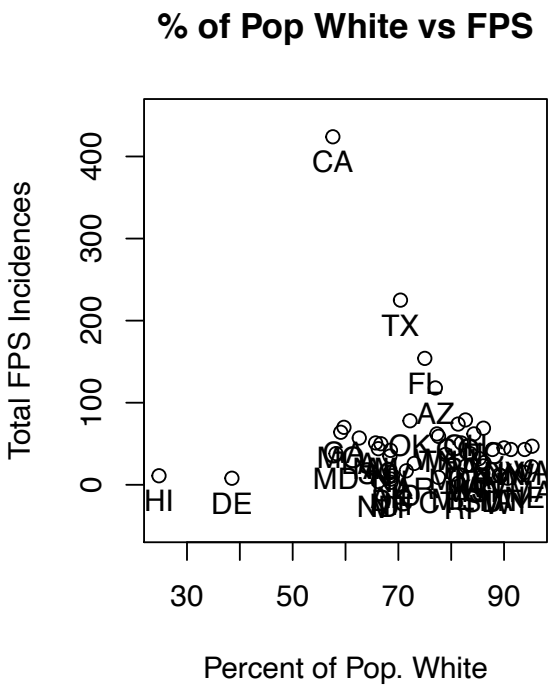
```
##  totshootstate       StatePop          Poverty           HS
##  Min.   :  2.00   Min.   :  563626   Min.   : 8.90   Min.   :81.80
##  1st Qu.: 14.00   1st Qu.: 1696962   1st Qu.:12.15   1st Qu.:85.60
##  Median : 38.00   Median : 4339367   Median :15.20   Median :89.10
##  Mean   : 49.71   Mean   : 6053834   Mean   :14.85   Mean   :88.25
```

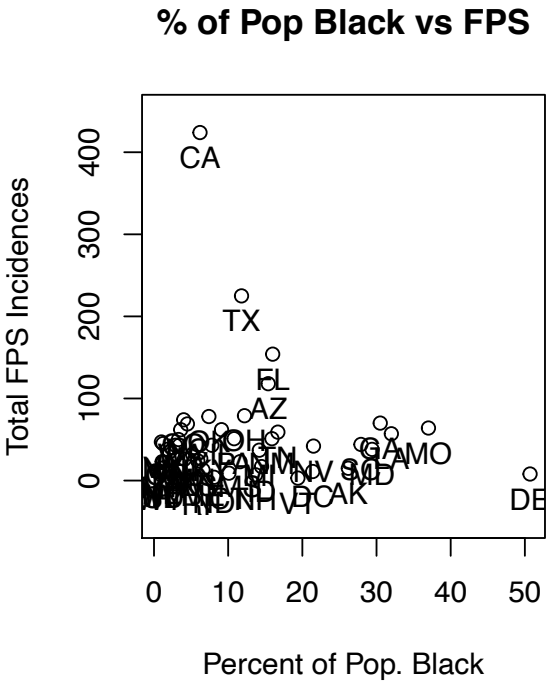
```
##  3rd Qu.: 58.00   3rd Qu.: 6636084   3rd Qu.:17.35   3rd Qu.:90.80
##  Max.   :424.00   Max.   :37253956   Max.   :22.50   Max.   :92.80
##      White           Black           Asian       NativeAmerican
##  Min.   :24.70   Min.   : 0.40   Min.   : 0.60   Min.   : 0.200
##  1st Qu.:68.50   1st Qu.: 3.10   1st Qu.: 1.35   1st Qu.: 0.300
##  Median :77.60   Median : 7.40   Median : 2.30   Median : 0.600
##  Mean   :75.99   Mean   :11.12   Mean   : 3.68   Mean   : 1.712
##  3rd Qu.:86.00   3rd Qu.:15.65   3rd Qu.: 3.80   3rd Qu.: 1.200
##  Max.   :95.30   Max.   :50.70   Max.   :38.60   Max.   :14.800
##     Hispanic         Other          NumMurders       GunOwners
##  Min.   : 1.20   Min.   : 1.800   Min.   : 0.300   Min.   :0.0360
##  1st Qu.: 4.30   1st Qu.: 3.650   1st Qu.: 1.250   1st Qu.:0.3055
##  Median : 8.20   Median : 6.200   Median : 2.700   Median :0.3980
##  Mean   :10.58   Mean   : 7.484   Mean   : 2.794   Mean   :0.3695
##  3rd Qu.:12.05   3rd Qu.: 9.400   3rd Qu.: 3.450   3rd Qu.:0.4400
##  Max.   :46.30   Max.   :34.800   Max.   :16.500   Max.   :0.5970
##     Under45         PercMale
##  Min.   :53.20   Min.   : 89.50
##  1st Qu.:58.45   1st Qu.: 95.10
##  Median :59.60   Median : 96.80
##  Mean   :59.91   Mean   : 97.28
##  3rd Qu.:61.20   3rd Qu.: 98.60
##  Max.   :71.20   Max.   :108.50
```

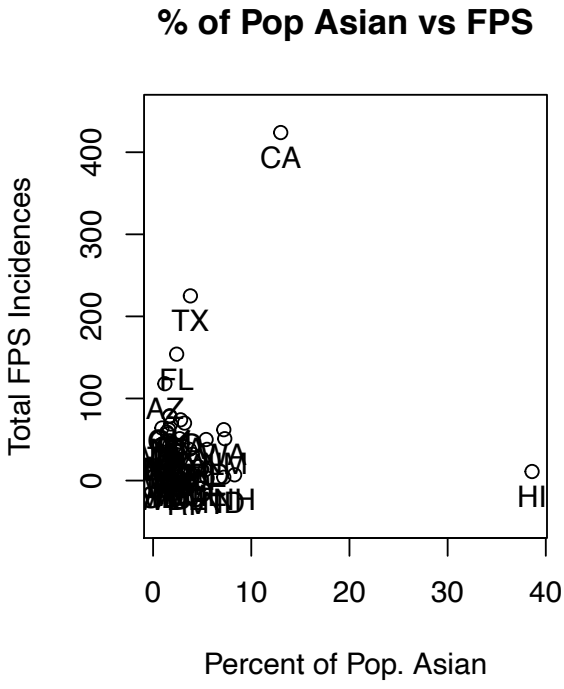
The plots below illustrate the possible correlation effects between the predictors of interest and the total number of FPS incidences in each state between 2015-2017

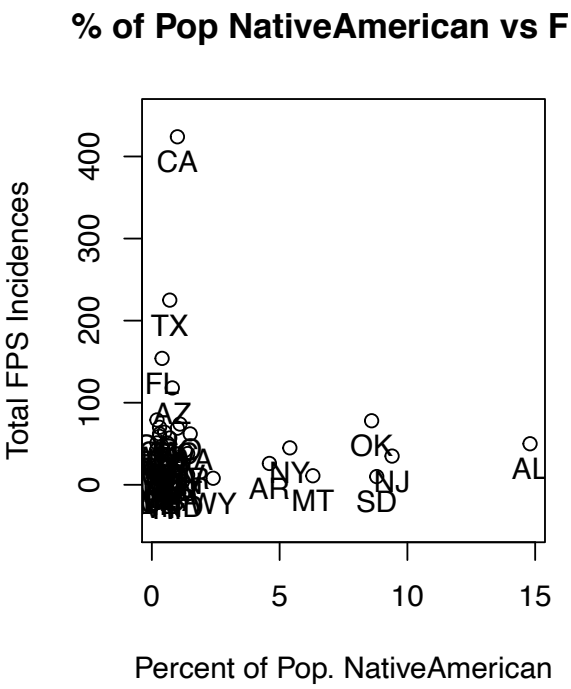


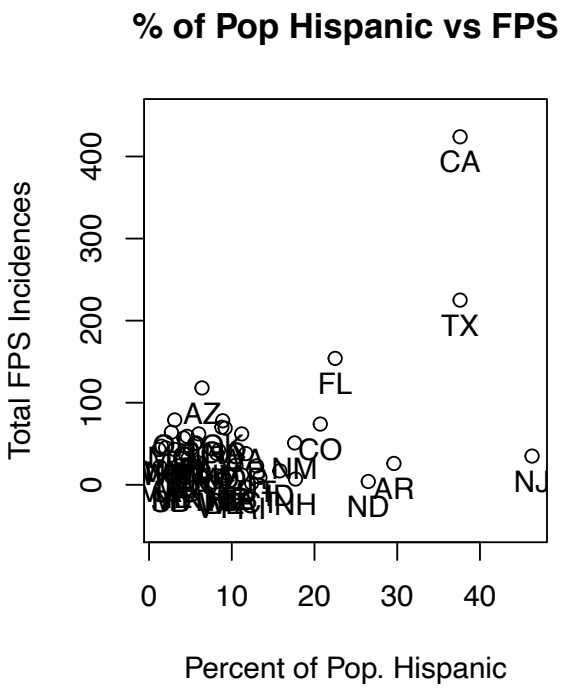

HS Grad Rate vs FPS
Total FPS Incidences
0
100
200
300
400
CA
TX
FL
AZ
MO
OK
82
84
86
88
90
92
High School Graduation Rate

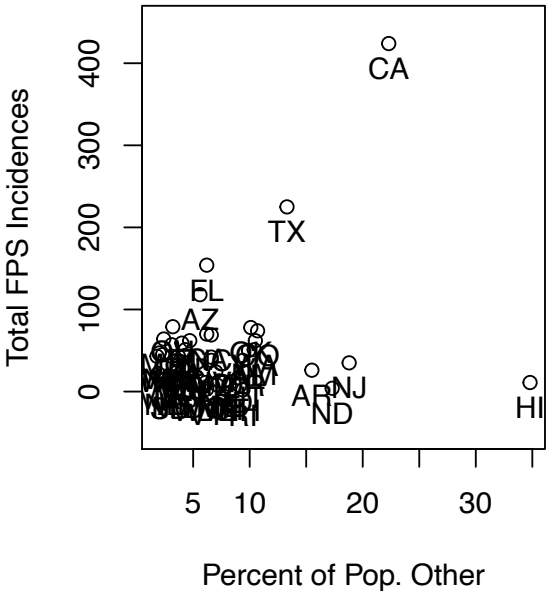

% of Pop White vs FPS
Total FPS Incidences
0
100
200
300
400
CA
TX
AZ
HI
DE
30
50
70
90
Percent of Pop. White

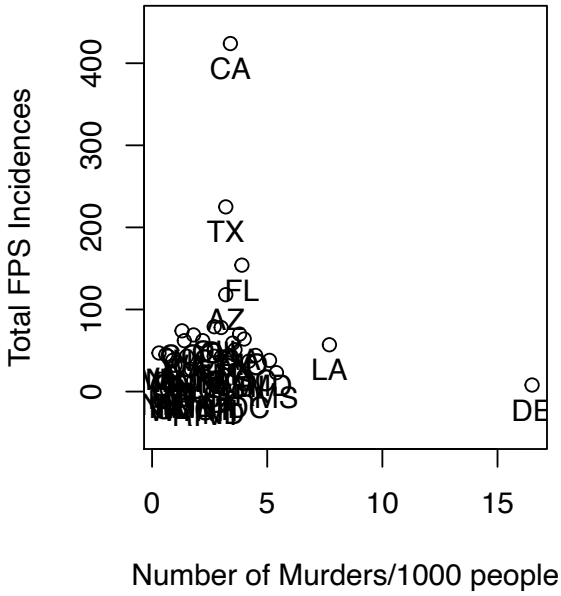

% of Pop Black vs FPS
Total FPS Incidences
0
100
200
300
400
CA
TX
AZ
DE
0
10
20
30
40
50
Percent of Pop. Black

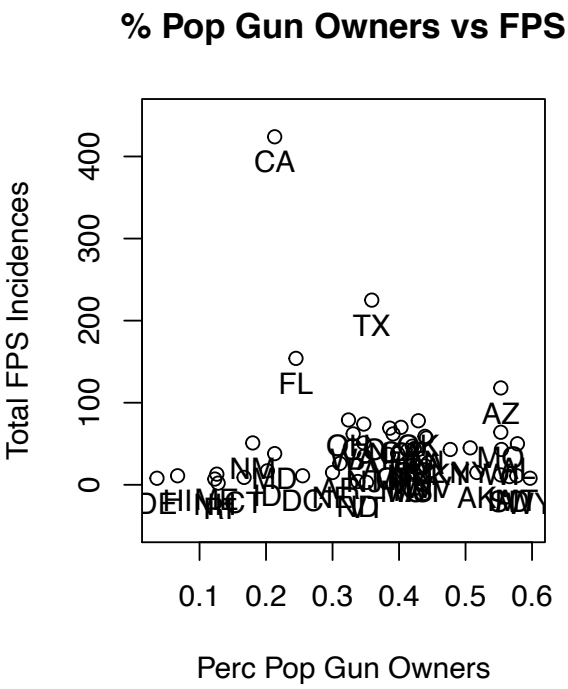

% of Pop Asian vs FPS
Total FPS Incidences
0
100
200
300
400
CA
TX
AZ
HI
0
10
20
30
40
Percent of Pop. Asian

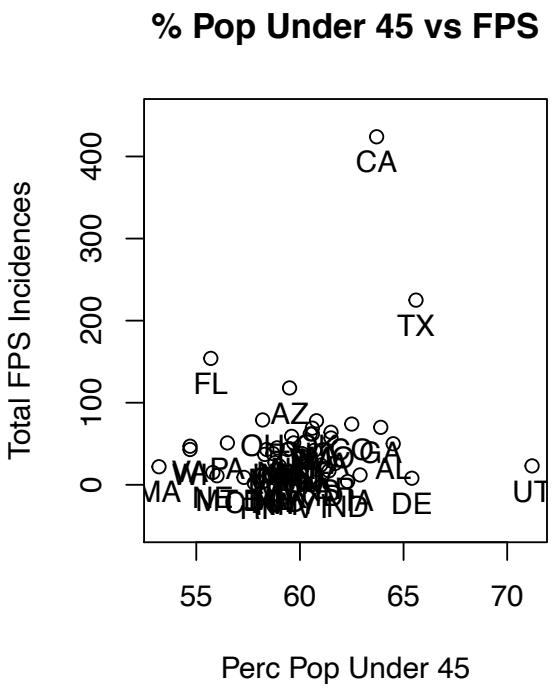

% of Pop NativeAmerican vs FPS
Total FPS Incidences
Percent of Pop. NativeAmerican
CA
TX
FL
AZ
OK
NY
AR
MT
SD
NJ
AL
0
100
200
300
400
0
5
10
15

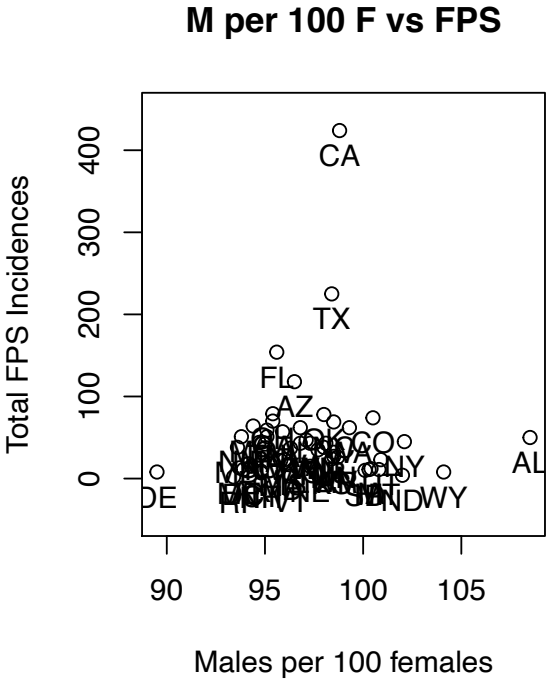

% of Pop Hispanic vs FPS
Total FPS Incidences
Percent of Pop. Hispanic
CA
TX
FL
AZ
CO
NM
AR
ND
NJ
0
100
200
300
400
0
10
20
30
40


% of Pop Other vs FPS
Total FPS Incidences
Percent of Pop. Other
CA
TX
FL
AZ
AR
NJ
ND
HI
0
100
200
300
400
5
10
20
30


# of Murders vs FPS
Total FPS Incidences
Number of Murders/1000 people
CA
TX
FL
AZ
LA
MS
DE
0
100
200
300
400
0
5
10
15







Covariate Considerations

In consideration of the qualitative ananlyis in the graphs above, it it becomes important to consider a possible covariate ajustment before proceeding to fit any additional models. Specifically, upon further examination of the "population" parameter, it would be expected that this variable would be highly (if not perfectly)

correlated with the number of FPS incidences recorded between 2015-2017. This intuition may be justified by runnning an initial regression on the value of state-level FPS vs state population. The anova table listed below gives an indication as to how well the value of state population is correlated with FPS incidence rates.

```
lmodi=lm(totshootstate~StatePop,data = df)
(lmods=summary(lmodi))

##
## Call:
## lm(formula = totshootstate ~ StatePop, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.370  -19.595    1.421   16.796  129.212
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.152e+00  7.470e+00   0.288    0.775
## StatePop    7.855e-06  8.234e-07   9.539 9.41e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.73 on 49 degrees of freedom
## Multiple R-squared:   0.65,Adjusted R-squared:  0.6429
## F-statistic:    91 on 1 and 49 DF,  p-value: 9.406e-13
```

The above anova table demonstrates the fact that state population is essentially perfectly correlated with state-level incidence totals of FPS between the years of 2015-2017.

Based upon the definition of a cavariate being a variable that is related to the outcome (in the case of this study-the response) and not the treatment (in the case of this study-the predictors of interest). It has been verified that the variable "State Population" is correlated with state level FPS incidence totals. In order to verify that "State Population" is NOT correlated with the rest of the predictors of interest, another regression on the value "State Population" as predicted by the remaining predictors can be run. The anova table below represents such a regression.

```
lmodi2=lm(StatePop~.-totshootstate, data = df)
(lmodi2s=summary(lmodi2))

##
## Call:
## lm(formula = StatePop ~ . - totshootstate, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9266404 -2394209  -409619  1516703 14932610
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -41689048 1237473048  -0.034   0.9733
## Poverty            209405     636037   0.329   0.7438
## HS               -1016436     498204  -2.040   0.0483 *
## White              923323   12024353   0.077   0.9392
## Black              685254   12028884   0.057   0.9549
## Asian             2988880   12156487   0.246   0.8071
```

```
## NativeAmerican     972583   12033518   0.081   0.9360
## Hispanic           879370     195223   4.504 6.16e-05 ***
## Other            -1716834   11792735  -0.146   0.8850
## NumMurders          95700     669391   0.143   0.8871
## GunOwners        -9382411   17613716  -0.533   0.5974
## Under45            688577     351101   1.961   0.0572 .
## PercMale            96289     729611   0.132   0.8957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4745000 on 38 degrees of freedom
## Multiple R-squared:  0.6326,Adjusted R-squared:  0.5166
## F-statistic: 5.453 on 12 and 38 DF,  p-value: 2.736e-05
```

Although it seems that the predictors: "HS"(High School Graduation Rate), "Hispanic" (Percent of Population Hispanic), and "Under45"(Percent of Population under age 45) are strongly correlated with the value of "State Population" the effect of this correlation is less alarming at this stage due to the fact that few of the other other predictors of interest are showing correlation, and because further correlation can be identified in the model analysis phase.

Based upon these factors, it seems like a good idea to adjust for "state population" as a covariate prior to fitting a model.

The covariate adjustment is performed according to the transformation:

$$totalFPS'_i = |totalFPS_i - StatePop_i|$$

Model Selection

Now that an appropriate covarite has been adjusted for. Creating a model with the remaining variables of interest remains the goal, in an effort to determine if any are significantly correlated with the covariate adjusted state-level FPS incidence totals. This process begins with fitting the following model in R:

$$\begin{gathered} Y_i = \mu + \beta_{\texttt{Proverty}} + \beta_{\texttt{HS}} + \beta_{\texttt{White}} + \beta_{\texttt{Black}} + \beta_{\texttt{Asian}} + \beta_{\texttt{NativeAmerican}} + \beta_{\texttt{Hispanic}} \\ +\beta_{\texttt{Other}} + \beta_{\texttt{NumMurders}} + \beta_{\texttt{GunOwners}} + \beta_{\texttt{Under45}} + \beta_{\texttt{PercMale}} \\ \text{where } i \in \{AL,\ AK,\ AZ, \ldots,\ WY\} \end{gathered}$$

The next step will be to evaluate the model for collinearity. Specifically, regressors with variance inflation factors (VIF) greater than 5 will be removed one at a time until no more vifs are greater than five. The process results in the the regression parameters: Asian, White, GunOwners, and Black being removed from the model. The resulting model is:

$$\begin{gathered} Y_i = +\mu + \beta_{\texttt{Proverty}} + \beta_{\texttt{HS}} + \beta_{\texttt{NativeAmerican}} + \beta_{\texttt{Hispanic}} + \beta_{\texttt{Other}} \\ +\beta_{\texttt{NumMurders}} + \beta_{\texttt{Under45}} + \beta_{\texttt{PercMale}} \\ \text{where } i \in \{AL,\ AK,\ AZ, \ldots,\ WY\} \end{gathered}$$

It is possible to further reduce the model by looking at the summary table of t-tests. The test statistics, and p-values for the hypothesis frameworks:

$$\begin{gathered} H_0 :\ \beta_{\texttt{i}} = 0 \\ \text{where } i \in \{Poverty,\ HS, \ldots, PercMale\} \end{gathered}$$

using the test statistic:

$$t_i = \frac{\hat{\beta}_{\texttt{i}}}{se(\hat{\beta}_{\texttt{i}})}$$

This table is displayed below:

```
lmod1s

##
## Call:
## lm(formula = scaled.shot ~ Poverty + HS + NativeAmerican + Hispanic +
##     Other + NumMurders + under45 + PercMale, data = df2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -11399913  -3034980   -279894   2678078  17285512
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    105543627   57402309   1.839  0.07304 .
## Poverty          -714451     439423  -1.626  0.11146
## HS              -1195455     502223  -2.380  0.02191 *
## NativeAmerican   -550454     360605  -1.526  0.13439
## Hispanic          327621     119905   2.732  0.00916 **
## Other            -152431     185489  -0.822  0.41584
## NumMurders         73205     467100   0.157  0.87621
## under45            -5416     331382  -0.016  0.98704
## PercMale          157747     453795   0.348  0.72986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5485000 on 42 degrees of freedom
## Multiple R-squared:  0.4573,Adjusted R-squared:  0.3539
## F-statistic: 4.423 on 8 and 42 DF,  p-value: 0.000609
```

The table clearly shows that the only variables of interest that are significant at the $\alpha = 0.1$ level are: the intercept, the High School Graduation Rate, and the Percentage of the Population that is Hispanic. Using this fact, we may reduce the model to:

$$Y_i = \mu + \beta_{\texttt{HS}} + \beta_{\texttt{Hispanic}}$$
$$\text{where } i \in \{AL,\ AK,\ AZ, \ldots,\ WY\}$$

Checking Model Assumptions

In order for the coefficient estimates and the inferential conclusions found in a summary table of the final model to be legitimate, it is necessary to check the regression diagnostics. There are two different categories of potential problems which need to be addressed, and each category has several specific assumptions which need to be verified. These categories, and assumptions are:

- Checking Error Assumptions
  - Constant Variance
  - Normality
  - Uncorrelated Errors
- Finding Unusual Observations
  - Leverage Points
  - Outliers
  - Influential Observations

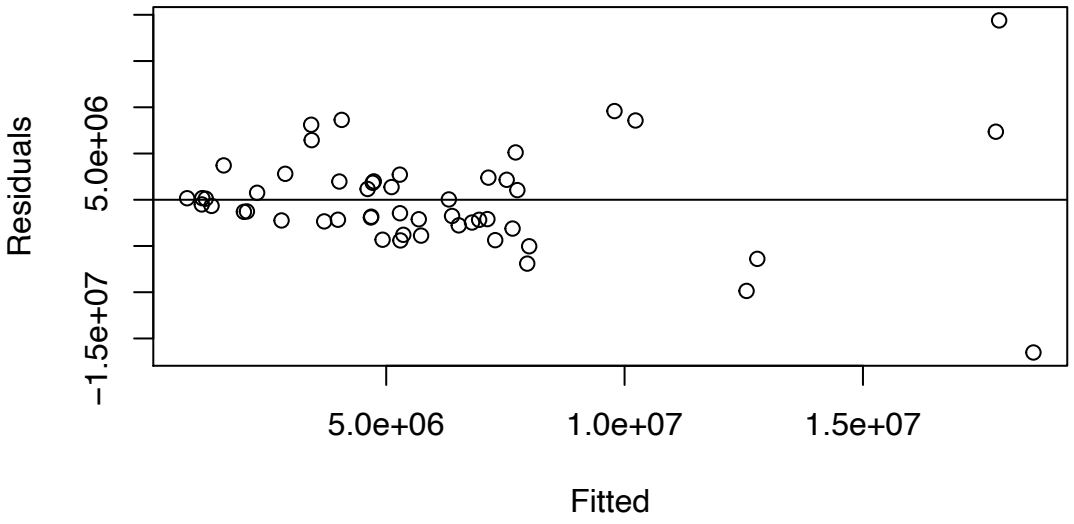
The process of verifying model assumptions will progress systematically through these categories, changing (if possible) the model where appropriate. It should be noted that the final model being evaluated is given by:

$$Y_i = \mu + \beta_{\texttt{HS}} + \beta_{\texttt{Hispanic}}$$
$$\text{where } i \in \{AL,\ AK,\ AZ, \ldots,\ WY\}$$

Checking Error Assumptions

Constant Variance

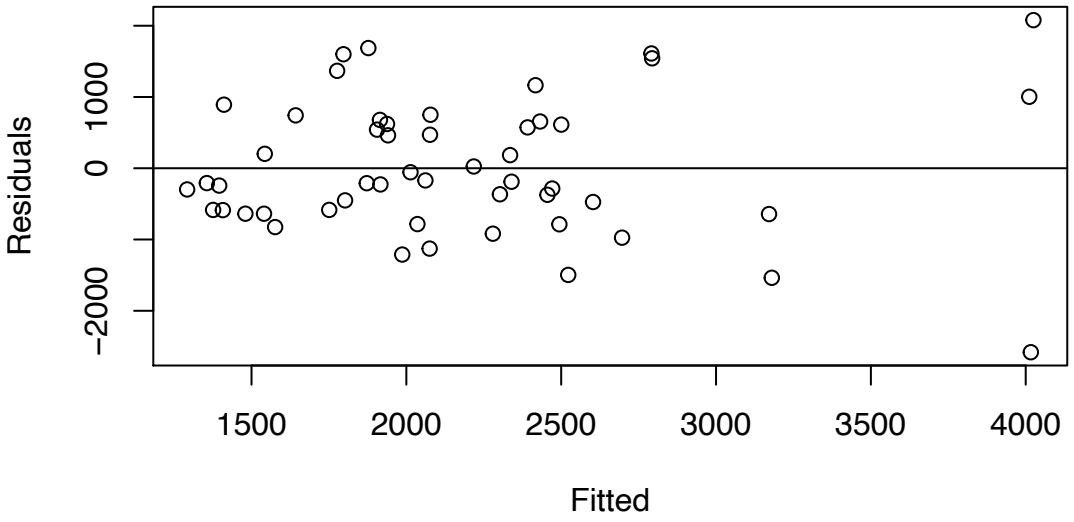
The assumption of constant error variance can be checked by looking at the plot of fitted value vs residual value. Ideally, the plot should look like a homogenous scatter about the x-axis.



This is a clear case of heteroscedasticity, and may be partially rectified by performing a sqare root transformation on the outcome variable. This implies that the final model is now:

$$\sqrt{Y_i} = \mu + \beta_{\texttt{HS}} + \beta_{\texttt{Hispanic}}$$
$$\text{where } i \in \{AL,\ AK,\ AZ, \ldots,\ WY\}$$

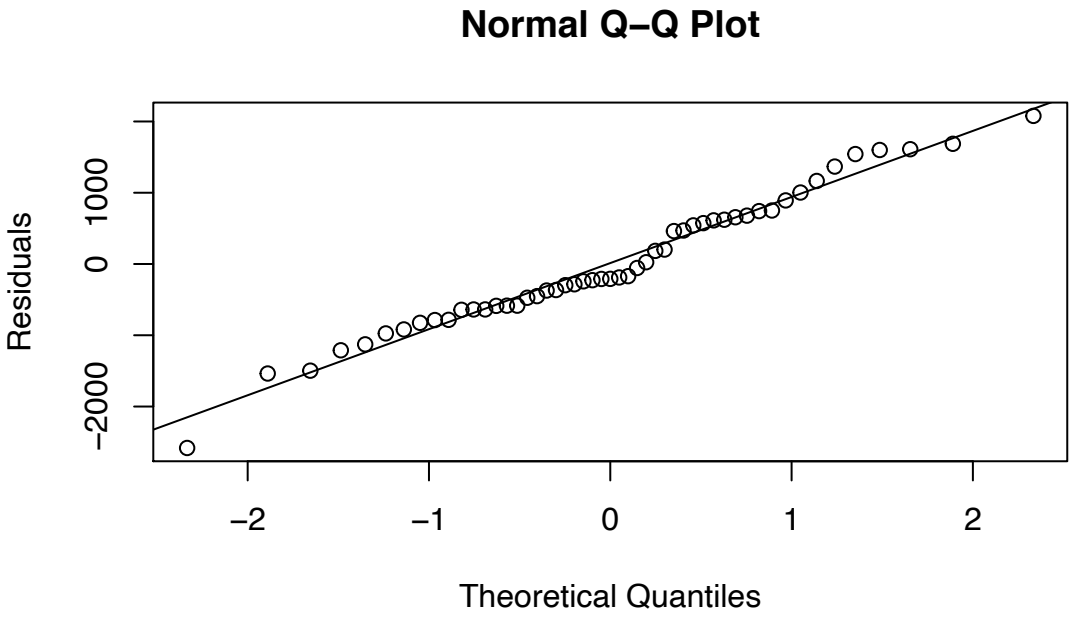
re-plotting the new model on the same plot produces:



The new model still exhibits signs of heteroscedasticity, but is clearly better.

Normality of Errors

The process of verifying that the errors are approximately normal can again be verified using plots. In this case the errors are plotted in a histogram and on a Q-Q plot, to determine their normality.



There is not indication for extreme non-normality in the errors as indicatied by either of the plots.

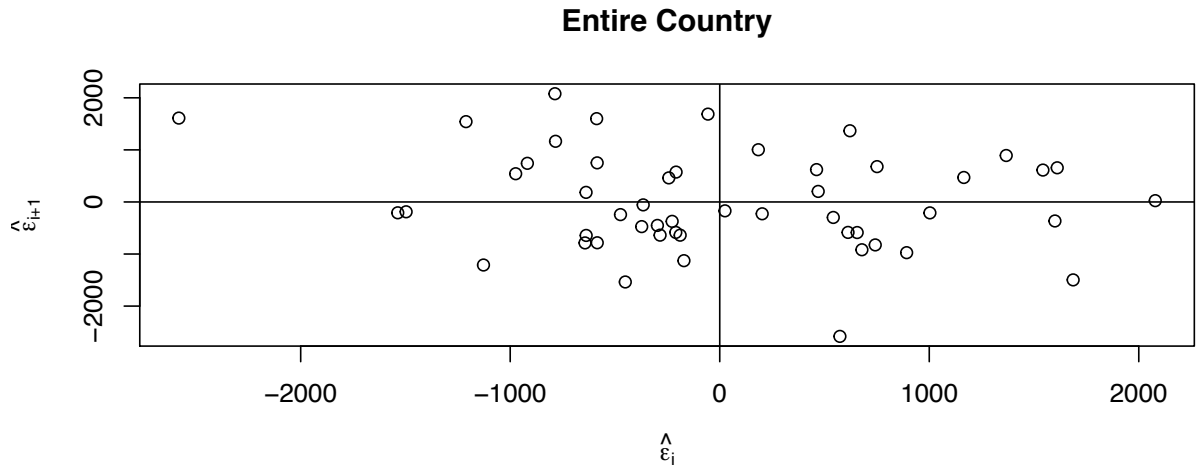
Correlated Errors

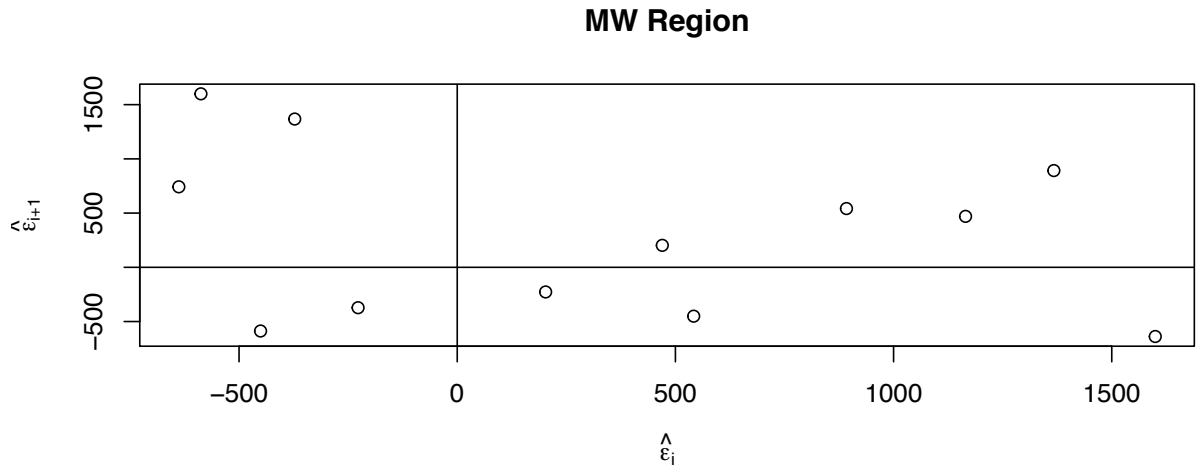
Checking for correlation in the errors is more difficult, and should be addressed from a geographical-location perspective. In the sense that a particular state of the United States might have a higher total FPS incidence due to the fact that it is next to a state that ALSO has a high total FPS incidence (or similar for low total FPS incidence). This is certianly a question worthwhile addressing; however, due to the complexity of such an analysis, a standard "successive residual" diagnostic plot is used to demonstrate the random scatter of successive residuals, and then the country is broken down into 5 sub-groups as classified by The Guardian Newpaper in the artice "Gun laws in the US, state by state" [16]. The classifications are:

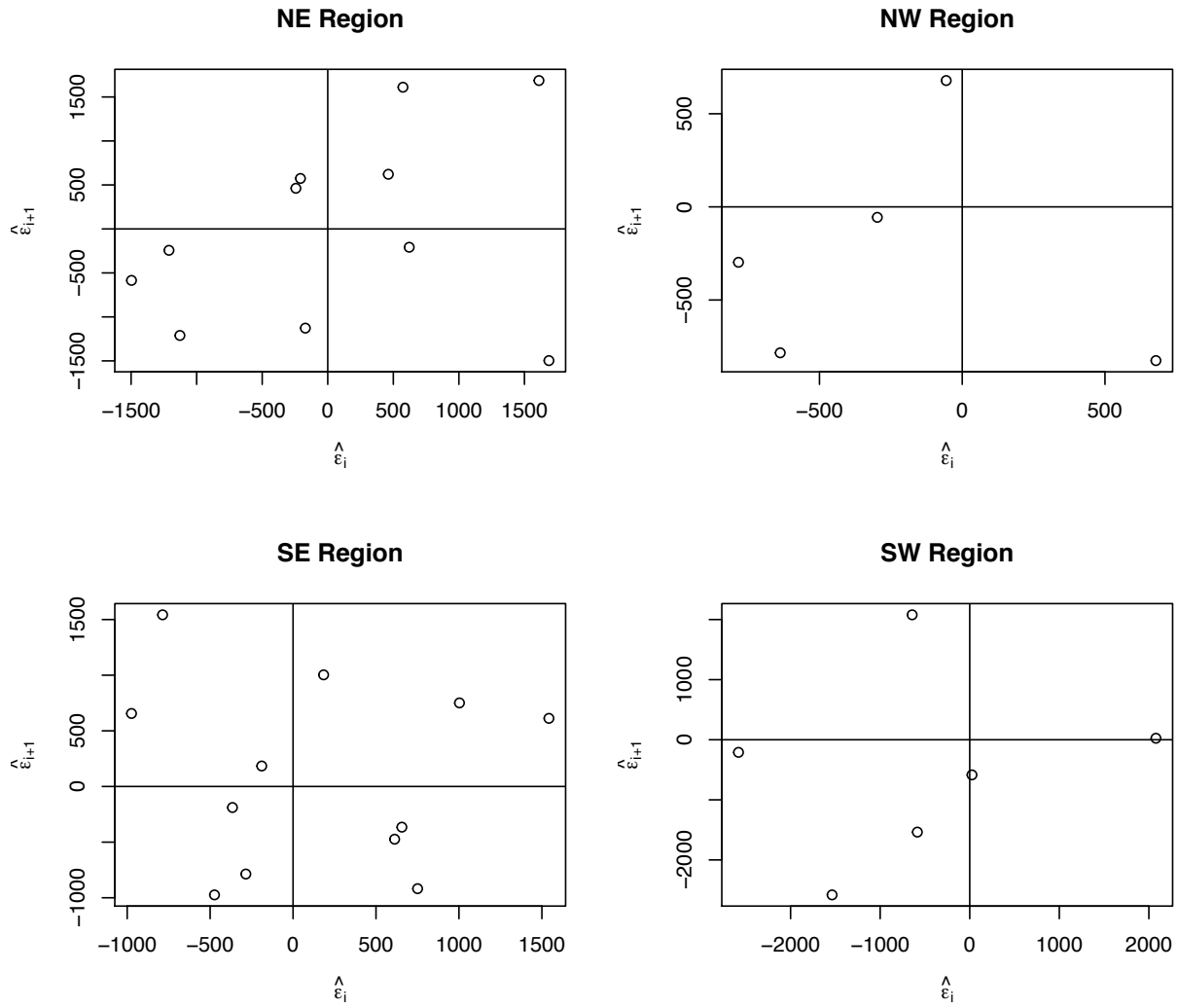
- NW: WY, ID, AK, MT, OR, WA
- SW: CA, HI, CO, UT, AZ, NV, NM
- NE: VT, RI, PA, NY, NJ, NH, DC, MA, MD, ME, DE, CT
- SE: TX, OK, AR, MS, AL, LA, FL, GA, NC, SC, TN, VA, WV
- MW: NE, SD, ND, KS, WI, MO, MN, MI, KY, OH, IA, IN, IL

The successive residuals within each sub-group is then plotted to demonstrate random scatter on a macro-geographical level.

[16] Guardian Article


Entire Country
2000
0
−2000
−2000
−1000
0
1000
2000
$\hat{\varepsilon}_{i+1}$
$\hat{\varepsilon}_i$


MW Region
1500
500
−500
−500
0
500
1000
1500
$\hat{\varepsilon}_{i+1}$
$\hat{\varepsilon}_i$

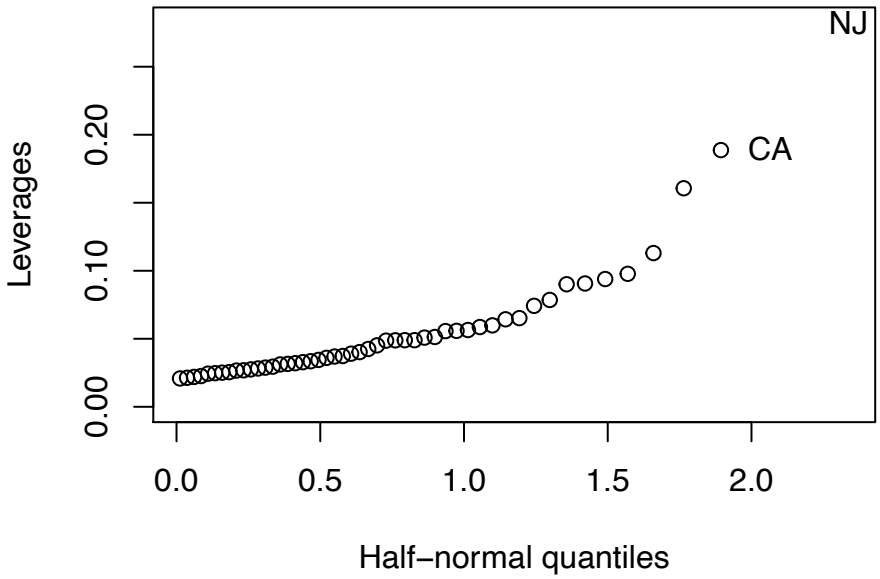


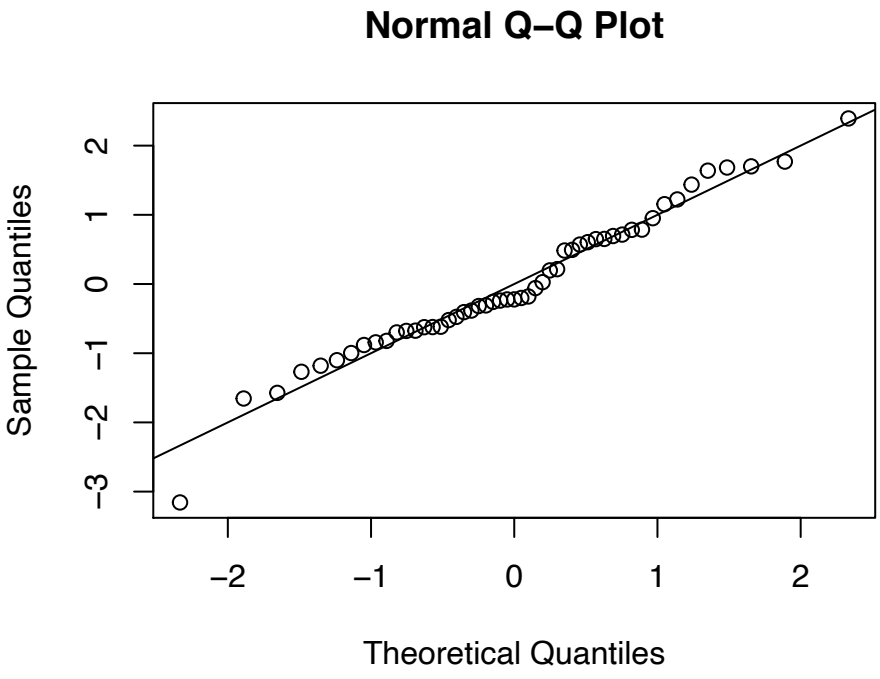
Each of the successive residual plots demonstrates that the residuals are sufficiently uncorrelated to succesfully verify that the errors are uncorrelated.

Finding Unusual Observations

Leverage Points

We may check for points with significant leverage by plotting: the model hat values for each state, and the standardized residuals.





It is clear from the leverage plot that New Jersey and California have larger than average leverage. However, the plot of the standardized residuals indicated that these values may not be as influential as indicated in the leverage plot. Nonetheless, careful note of these observations will be made, in case action need to be taken later in the model analysis.

Outliers

We may check for outliers by performing an outlier test, which looks for Bonferroni-corrected p-values indicating that Studentized residuals are abnormal.

```
outlierTest(lmod)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## NJ -3.509218           0.001002     0.051104
```

Under the pre-determined significance value of $\alpha = 0.1$, the above test has determined that the New Jersey observation IS and outlier. Due to the extremis of this value, it is wise to consider two models. The first model will remain unchange; however, the second model will remove the New Jersey observation in an effort to achieve a better fit. Displayed blow are the summary tables for each of the descrbed models. The first is the original model (including New Jersey), the second is the updated model that has New Jersey Removed.

```
##
## Call:
## lm(formula = sqrt(scaled.shot) ~ HS + Hispanic, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2581.6  -612.4  -208.4   638.6  2078.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13663.48    4434.67   3.081  0.00341 **
## HS           -134.44      49.41  -2.721  0.00904 **
## Hispanic       36.14      15.31   2.360  0.02239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 965.3 on 48 degrees of freedom
## Multiple R-squared:  0.3215,Adjusted R-squared:  0.2932
## F-statistic: 11.37 on 2 and 48 DF,  p-value: 9.071e-05
##
## Call:
## lm(formula = sqrt(scaled.shot) ~ HS + Hispanic, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2019.2  -509.3  -157.1   601.4  1742.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12840.11    3996.24   3.213 0.002374 **
## HS           -127.58      44.49  -2.867 0.006178 **
## Hispanic       63.37      15.81   4.008 0.000217 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 868.4 on 47 degrees of freedom
```

```
## Multiple R-squared:  0.4577,Adjusted R-squared:  0.4346
## F-statistic: 19.83 on 2 and 47 DF,  p-value: 5.694e-07
```

Quantification of what the removal of the New Jersey observation caused is possible by calculating the percent change in the estimated regression coefficients. The percent change in the estimated coefficeients for the intercept, High School Graduation, and Hispanic regressors in listed below.

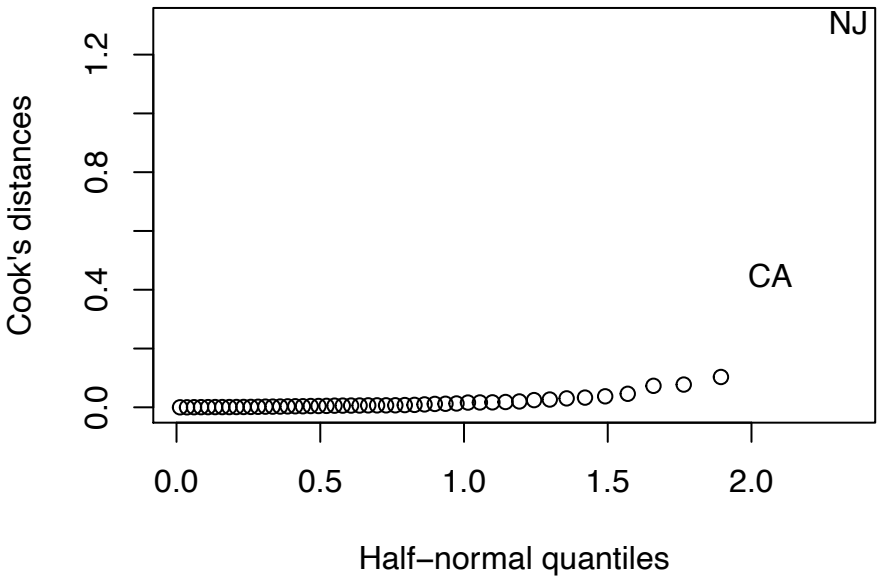
```
perc.change

## [1]  0.06026090  0.05106296 -0.75356058
```

Examining these changes tells us that New Jersey plays a significan role in the estimation of the Hispanic regression coefficient. An approximately 75% change in the estimation of this coefficient based upon one observation demonstrates the estimate's instability, especially since a clarification on the test significance level needed to be made in order to determine the New Jersey observation's qualification as an Outlier.

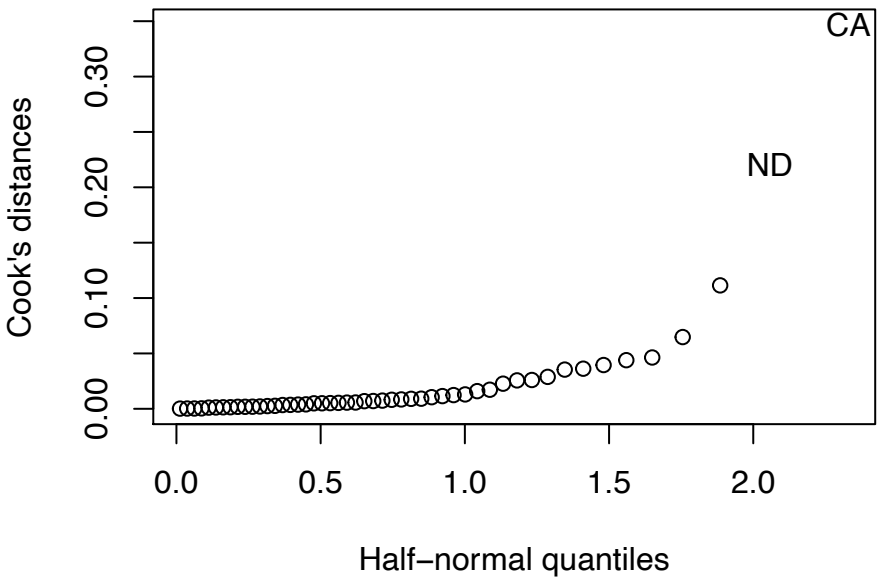
Officially, a decision between the two models (including the NJ observation or not) cannot be made without a contextual factor being present. From the perspective of an unbiased analysis it is important to consider both models.

Influential Observations

One of the most popular measures for infuence is the Cook's Distance of an observation. A plot of the Cook's distance is plotted below



A similar plot for the modified model from above is also displayed.



Although the plots look similar, the percent change in range of Cook's distances between the first and second plot is an approximately 74% decrease. This means that the Cook's distance between New Jersey, and the next highest value (California) accounted for approximately 74% of the variation in the Cook's distances. This implies that the removal of New Jersey as an observation from the data is a significant motivation for removing influential observations. Again, a legitimate reason for this removal cannot be made apart from contextually. It has become even more important to consider a two-model approach when considering a final model.

DISCUSSION

Main Conclusions

The primary conclusions of this analysis are restricted to inference about the particular variables of interest that were tested. The analysis found that many of the cross-sectional measures of social, political and economic status that might be expected to correlate with the total state-level incidences of fatal police shootings, did not correlate at a significant level.

In fact, of the 13 variables of interest that were tested, only two tested significant at the $\alpha = 0.1$ level. These variables were: percentage of population with a high school education, and the percentage of population that were Hispanic. Upon running a model on the standardized versions of the variables, the formula of which is given by:

$$x_i^{std} = \frac{x_i - \bar{x}}{sd(x)}$$
$$\text{for } i \in \{AL,\ AK, \ldots,\ WY\}$$
$$\text{and } x = \{TotFPS,\ HS,\ Hispanic\}$$

It is possible to get a percentage estimate of the regression coefficients. The standardized regression coefficient estimates are:

Variable	Estimate
Intercept	0.45551
HS.std	-0.07106
Hispanic.ste	0.01069

The interpretation of the above coefficients are:

- Each unit percentage increase in standardized population High School Graduation rate is correlated with a 7.1% decline in covariate-adjusted, standardized, total state-level fatal police shootings at a significance level of $\alpha = 0.1$.
- Each unit percentage increase in standardized Hispanic population percentage is correlated with a 1.1% increase in covariate-adjusted, standardized, total state-level fatal police shootings at a significance level of $\alpha = 0.1$.

Unfortunately, further, more relatable, interpretations of these coefficients will be harder to create due to the fact that the initial response variable was transformed by a convariate prior to modeling. In particular, interpretations which relate the estimated coefficients to the original response variable of: total state-level fatal police shootings, would be extrordinarily hard to achieve. Nevertheless, the net correlation (direction) can be estimated, even if the magnitude is harder to interpret.

Concerns and Future Analyses

The data studied in this analysis represents an amalgomation of data pulled from several sources. This data is inteded to represent a homogeneous cross-section of the United States, and is intended to model the Washington Post Police Shootings Database. Unfortunately, as was described in the "Statistical Analysis Methods" section, the cross-sectional data of the variables of interest do not align with the the WPPSD data cross-section. This is concerning, in that any conclusions drawn from the analysis cannot rule out confounding effects of a temporal nature. The conclusions made in this analysis were specifically stated as "correlated effects" and not causal, and this point cannot be understated. No causal conclusions can be concluded from this analysis.

It should be noted that an $\alpha = 0.1$ significance level was used as a method of searching for possible variables of significance, and not for the purpose of proving correlation. This study was primarily inteded as a searching divice, designed to highlight possible variables on which to focus more attention for possible fututre study.

Cross-sectional data on social, economic and political standings can suffer from severe biases due to temporal, interviewer, and misclassification errors. Although these effects cannot be eliminated, other study designs can help to limit their effects, or at least compensate for them. Based upon the high-stakes nature of the response being modeled in this analysis, it would be advisable to begin a prospective, or longitudinal study which will help to determine the underlying causes of fatal police shootings, with the end goal of supporting those variables that reduce FPS incidences and reducing the variables that increase FPS incidences.

APPENDIX

Refferences

- "Arab Spring." Wikipedia, Wikimedia Foundation, 5 May 2018, $en.wikipedia.org/wiki/Arab_Spring$.[1]
- "Me Too Movement." Wikipedia, Wikimedia Foundation, 6 May 2018, $en.wikipedia.org/wiki/Me_Too_movement$.[2]
- "Black Lives Matter." Wikipedia, Wikimedia Foundation, 3 May 2018, $en.wikipedia.org/wiki/Black_Lives_Matter$.[3][4][5]
- "2015 Washington Post Database of Police Shootings." The Washington Post, WP Company, $www.washingtonpost.com/graphics/national/police-shootings/$.[6][7]

- "The Prevalence of Fatal Police Shootings by U.S. Police, 2015–2016: Patterns and Answers from a New Data Set." Egyptian Journal of Medical Human Genetics, Elsevier, 10 May 2017, *www.sciencedirect.com/science/article/pii/S0047235217301344*.[8]
- "Police Shootings Reflect Structural Racism, Study Finds." ScienceDaily, ScienceDaily, 5 Feb. 2018, *www.sciencedaily.com/releases/2018/02/180205134232.htm*.[9]
- Data Access and Dissemination Systems (DADS). American FactFinder, 5 Oct. 2010, *factfinder.census.gov/faces/nav/jsf/pages/error.xhtml*.[10][13][14]
- "Gun Ownership Statistics by State." Demographic Data, 6 Sept. 2016, *demographicdata.org/facts − and − figures/gun − ownership − statistics/*.[11][15]
- Wullum, Karolina. "Fatal Police Shootings in the US Kaggle." Countries of the World Kaggle, 22 Sept. 2017, *www.kaggle.com/kwullum/fatal − police − shootings − in − the − us*.[12]

R-Code

```
##################################################################################################
#Load Packages
##################################################################################################
library(gdata)
library(reshape2)
library(ggplot2)
library(gridExtra)
library(faraway)
library(perturb)
library(car)
library(leaps)
##################################################################################################



##################################################################################################
#Import Data Sets
##################################################################################################
#General State Information
data_State=read.csv("/Users/lee/Desktop/FINAL PROJECT/
                    Project_Data_Official/State_Data.csv")

#WP Shooting Data
data_shootings=read.csv("/Users/lee/Desktop/FINAL PROJECT/
                        Project_Data_Official/fatal-police-shootings-in-the-us/
                        PoliceKillingsUS_Rev1.csv")

#US Census State-Poverty Data
data_poverty=read.csv("/Users/lee/Desktop/FINAL PROJECT/Project_Data_Official/
                      Census_Poverty_Data/Poverty_Data.csv")

#US Census State-High School Graduation Data
data_HS=read.csv("/Users/lee/Desktop/FINAL PROJECT/Project_Data_Official/
                 Census_HS_Data/Census_HS_Data_Rev.csv")
```

```
#US Census State-Race Composition Data
data_Race=read.csv("/Users/lee/Desktop/FINAL PROJECT/Project_Data_Official/
                   Census_Race_Data/Census_Race_Data_Rev.csv")

#US Census State-Population Composition
data_population=read.csv("/Users/lee/Desktop/FINAL PROJECT/Project_Data_Official/C
                         ensus_Population_Data/Census_Data_Populations.csv")

#US Census State-Age Composition
data_age=read.csv("/Users/lee/Desktop/FINAL PROJECT/Project_Data_Official/
                  Census_Age_Data/Census_age_data.csv")

#Violent Crime data
data_ViolentCrime=read.csv("/Users/lee/Desktop/FINAL PROJECT/Project_Data_Official/
                           Violent_Crime_Stats.csv")
##############################################################################################



##############################################################################################
#Creating Covariate Matrix
##############################################################################################
#create "Other" race classification
data_Race_Other=data_Race$PacificIslander+data_Race$TwoorMore+data_Race$Other

#Store Values related to Violent Crime
num.murders=data_ViolentCrime$Gun.Murder.Rate.per.100K..2010.
gun.owners=data_ViolentCrime$Gun.Ownership..2007.

#Store values related to age
under45=c()
for(i in 1:51)
{
  under45[i]=data_age$Under.18[i]+data_age$X18.to.24[i]+data_age$X25.to.44[i]
}

#Store values related to gender
perc.male=c()
for(i in 1:51)
{
  perc.male[i]=data_age$Males.per.100.females...All.ages[i]
}


#Create Covariate Matrix
Covariates=data.frame(data_State$SateAbrev, data_population$Population,
                      data_poverty$Percent,data_HS$Percent,
                      data_Race$White,data_Race$BlackOnly,data_Race$Asian,
                      data_Race$NativeAmerican,
                      data_Race$Hispanic,data_Race_Other, num.murders, gun.owners,
                      under45, perc.male)

names(Covariates)=c("StateAbbrev","StatePop","Poverty","HS","White","Black","Asian",
```

```
                    "NativeAmerican","Hispanic","Other","NumMurders",
                    "GunOwners", "Under45", "PercMale")

#################################################################################################



#################################################################################################
#Data Cleaning WPPSD
#################################################################################################
#Create Age Groups
age=data_shootings$age
ageNA=unknownToNA(x=age, unknown=c("",NA,0))
data_shootings$AgeGroup<-NA
data_shootings$AgeGroup[ageNA<=9]<-"0s"
data_shootings$AgeGroup[ageNA>=10 & ageNA<=19]<-"10s"
data_shootings$AgeGroup[ageNA>=20 & ageNA<=29]<-"20s"
data_shootings$AgeGroup[ageNA>=30 & ageNA<=39]<-"30s"
data_shootings$AgeGroup[ageNA>=40 & ageNA<=49]<-"40s"
data_shootings$AgeGroup[ageNA>=50 & ageNA<=59]<-"50s"
data_shootings$AgeGroup[ageNA>=60 & ageNA<=69]<-"60s"
data_shootings$AgeGroup[ageNA>=70 & ageNA<=79]<-"70s"
data_shootings$AgeGroup[ageNA>=80 & ageNA<=89]<-"80s"
data_shootings$AgeGroup[ageNA>=90 & ageNA<=99]<-"90s"
data_shootings$AgeGroup[is.na(data_shootings$age) |
                          data_shootings$age==""]<-"Unknown"


#Rename Race & Gender Levels
data_shootings$race=unknownToNA(x=data_shootings$race, unknown=c("",NA,0))
data_shootings$race=NAToUnknown(x=data_shootings$race, unknown = "Unknown")
levels(data_shootings$race)=c("Asian", "Black", "Hispanic", "NativeAmerican"
                              , "Other", "Unknown","White")
levels(data_shootings$gender)=c("Female","Male")


#Store Critical Values
CaseNO=data_shootings$CaseNO
AgeGroup=data_shootings$AgeGroup
Gender=data_shootings$gender
Race=data_shootings$race
State=data_shootings$state
#################################################################################################



#################################################################################################
#Setting Data Types
#################################################################################################
#Covariates DF
Covariates$StateAbbrev=factor(Covariates$StateAbbrev)

#Shootings DF
```

```
data_shootings$gender=factor(data_shootings$gender)
data_shootings$race=factor(data_shootings$race)
data_shootings$state=factor(data_shootings$state)
data_shootings$AgeGroup=factor(data_shootings$AgeGroup)
#############################################################################################



#############################################################################################
#Create Rawdata Data Frame
#############################################################################################
RawData=data.frame(CaseNO, AgeGroup, Gender, Race, State)

#Modify RawData Dataframe with Poverty Covariate
RawData$Poverty=NA
for(i in 1:2535)
{
  RawData$Poverty[i]=Covariates$Poverty[RawData$State[i]==Covariates$StateAbbrev]
}

#Modify RawData Dataframe with HS Covariate
RawData$HS=NA
for(i in 1:2535)
{
  RawData$HS[i]=Covariates$HS[RawData$State[i]==Covariates$StateAbbrev]
}

#Modify RawData Dataframe with Race Variables
RawData$White=NA
RawData$Black=NA
RawData$Asian=NA
RawData$NativeAmerican=NA
RawData$Hispanic=NA
RawData$Other=NA

for(i in 1:2535)
{
  RawData$White[i]=Covariates$White[RawData$State[i]==Covariates$StateAbbrev]
  RawData$Black[i]=Covariates$Black[RawData$State[i]==Covariates$StateAbbrev]
  RawData$Asian[i]=Covariates$Asian[RawData$State[i]==Covariates$StateAbbrev]
  RawData$NativeAmerican[i]=Covariates$NativeAmerican[RawData$State[i]==
                                                      Covariates$StateAbbrev]
  RawData$Hispanic[i]=Covariates$Hispanic[RawData$State[i]==Covariates$StateAbbrev]
  RawData$Other[i]=Covariates$Other[RawData$State[i]==Covariates$StateAbbrev]
}

#Modify RawData Dataframe with Population Covariate
RawData$Pop=NA
for(i in 1:2535)
{
  RawData$Pop[i]=Covariates$StatePop[RawData$State[i]==Covariates$StateAbbrev]
}
```

```
RawData$Count=1
#################################################################################################




#################################################################################################
#Create Count Data, in FPS data frame
#################################################################################################
FPS=aggregate(RawData$Count, by=list(State=RawData$State, AgeGroup=RawData$AgeGroup,
                                     Race=RawData$Race,Gender=RawData$Gender), FUN=sum)

FPS=data.frame(FPS$x,FPS$State,FPS$AgeGroup,FPS$Race,FPS$Gender)
names(FPS)=c("Shoot","State","AgeGroup","Race","Gender")

totShootState=tapply(FPS$Shoot,FPS$State,sum)
totShootAgeGroup=tapply(FPS$Shoot,FPS$AgeGroup,sum)
totShootRace=tapply(FPS$Shoot,FPS$Race,sum)
totShootGender=tapply(FPS$Shoot,FPS$Gender,sum)
#################################################################################################



#################################################################################################
#Create Create Final(non-covariate-scaled) Data Frame
#################################################################################################
df.totshootState=data.frame(totShootState)
df=data.frame(df.totshootState,Covariates$StatePop,Covariates$Poverty,
              Covariates$HS,Covariates$White,Covariates$Black,Covariates$Asian,
              Covariates$NativeAmerican,Covariates$Hispanic,Covariates$Other,
              Covariates$NumMurders,Covariates$GunOwners,Covariates$Under45,
              Covariates$PercMale)
names(df)=c("totshootstate","StatePop","Poverty","HS","White","Black",
            "Asian", "NativeAmerican", "Hispanic", "Other", "NumMurders", "GunOwners",
            "Under45", "PercMale")
head(df)

#################################################################################################



#################################################################################################
#Initial Data Analysis
#################################################################################################
plot(df$totshootstate~df$StatePop, xlab="State Population", ylab="Total FPS Incidences",
     main="FPS vs State Population")
with(df, text(df$totshootstate~df$StatePop, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$Poverty, xlab="Poverty Rate", ylab="Total FPS Incidences",
     main="Poverty Rate vs State Population")
with(df, text(df$totshootstate~df$Poverty, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$HS, ylim=c(-50,450), xlab="High School Graduation Rate", ylab="Total FPS Incid
```

```
     main="High School Graduation Rate vs State Population")
with(df, text(df$totshootstate~df$HS, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$White, ylim=c(-50,450), xlab="Percent of Pop. White",
     ylab="Total FPS Incidences",
     main="Percent of Pop White vs State Population")
with(df, text(df$totshootstate~df$White, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$Black, ylim=c(-50,450), xlab="Percent of Pop. Black",
     ylab="Total FPS Incidences",
     main="Percent of Pop Black vs State Population")
with(df, text(df$totshootstate~df$Black, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$Asian, ylim=c(-50,450), xlab="Percent of Pop. Asian",
     ylab="Total FPS Incidences",
     main="Percent of Pop Asian vs State Population")
with(df, text(df$totshootstate~df$Asian, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$NativeAmerican, ylim=c(-50,450),
     xlab="Percent of Pop. NativeAmerican", ylab="Total FPS Incidences",
     main="Percent of Pop NativeAmerican vs State Population")
with(df, text(df$totshootstate~df$NativeAmerican, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$Hispanic, ylim=c(-50,450), xlab="Percent of Pop. Hispanic",
     ylab="Total FPS Incidences",
     main="Percent of Pop Hispanic vs State Population")
with(df, text(df$totshootstate~df$Hispanic, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$Other, ylim=c(-50,450), xlab="Percent of Pop. Other",
     ylab="Total FPS Incidences",
     main="Percent of Pop Other vs State Population")
with(df, text(df$totshootstate~df$Other, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$NumMurders, ylim=c(-50,450),
     xlab="Number of Murders/1000 people", ylab="Total FPS Incidences",
     main="Number of Murders vs State Population")
with(df, text(df$totshootstate~df$NumMurders, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$GunOwners, ylim=c(-50,450), xlab="Perc Pop Gun Owners",
     ylab="Total FPS Incidences",
     main="% Pop Gun Owners vs State Population")
with(df, text(df$totshootstate~df$GunOwners, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$Under45, ylim=c(-50,450), xlab="Perc Pop Under 45",
     ylab="Total FPS Incidences",
     main="% Pop Under 45 vs State Population")
with(df, text(df$totshootstate~df$Under45, labels = row.names(df), pos = 1))

plot(df$totshootstate~df$PercMale, ylim=c(-50,450), xlab="Males per 100 females",
     ylab="Total FPS Incidences",
     main="Males per 100 females vs State Population")
with(df, text(df$totshootstate~df$PercMale, labels = row.names(df), pos = 1))
```

```
#############################################################################################



#############################################################################################
#Create Covariance Scaled Data Frame
#############################################################################################
df
lmodi=lm(totshootstate~StatePop,data = df)
(lmods=summary(lmodi))

lmodi2=lm(StatePop~.-totshootstate, data = df)
(lmodi2s=summary(lmodi2))


df2=df
df2$scaled.shot=NA
for(i in 1:51)
{
  df2$scaled.shot[i]=abs(df2$totshootstate[i]-df2$StatePop[i])
}
df2$scaled.shot
#############################################################################################


#############################################################################################
#Modeling,Selection, and Collinearily
#############################################################################################
lmod1=lm(scaled.shot~Poverty+HS+White+Black+Asian+NativeAmerican+Hispanic+Other+
           NumMurders+GunOwners+under45+PercMale,data=df2)
(lmod1s=summary(lmod1))

vif(lmod1)
#remove Asian

lmod1=lm(scaled.shot~Poverty+HS+White+Black+NativeAmerican+Hispanic+Other+NumMurders+
           GunOwners+under45+PercMale,data=df2)
(lmod1s=summary(lmod1))
vif(lmod1)
#Remove White

lmod1=lm(scaled.shot~Poverty+HS+Black+NativeAmerican+Hispanic+Other+NumMurders+
           GunOwners+under45+PercMale,data=df2)
(lmod1s=summary(lmod1))
vif(lmod1)
#Remove GunOwners

lmod1=lm(scaled.shot~Poverty+HS+Black+NativeAmerican+Hispanic+Other+NumMurders
         +under45+PercMale,data=df2)
(lmod1s=summary(lmod1))
vif(lmod1)
#Remove Black
```

```
lmod1=lm(scaled.shot~Poverty+HS+NativeAmerican+Hispanic+Other+NumMurders
         +under45+PercMale,data=df2)
(lmod1s=summary(lmod1))
vif(lmod1)
#Nothing to remove


lmod1=lm(scaled.shot~HS+Hispanic
           ,data=df2)
(lmod1s=summary(lmod1))
vif(lmod1)


lmodf=lm(scaled.shot~HS+Hispanic, data = df2)
lmodfs=summary(lmodf)
##############################################################################################



##############################################################################################
#Checking Model Assumptions
##############################################################################################

#Constant Variance
plot(fitted(lmodf),residuals(lmodf),xlab = "Fitted", ylab="Residuals")
abline(h=0)

lmodfsqrt=lm(sqrt(scaled.shot)~HS+Hispanic, data = df2)
lmodfsqrts=summary(lmodfsqrt)

plot(fitted(lmodfsqrt),residuals(lmodfsqrt),xlab = "Fitted", ylab="Residuals")
abline(h=0)

#Normality
qqnorm(residuals(lmodfsqrt),ylab="Residuals")
qqline(residuals(lmodfsqrt))

hist(residuals(lmodfsqrt), xlab="Residuals")


#Uncorrelated Errors
n=length(residuals(lmodfsqrt))
plot(tail(residuals(lmodfsqrt),n-1)~head(residuals(lmodfsqrt),n-1),
     xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0)

df.reg=data.frame(df2,data_State$Region)
levels(df.reg$data_State.Region)

lmodfsqrt=lm(sqrt(scaled.shot)~HS+Hispanic, data = df.reg)
lmodfsqrts=summary(lmodfsqrt)
region.resid.df=data.frame(data_State$Region, lmodfsqrts$residuals)
```

```
mw.states=data_State$SateAbrev[data_State$Region=="MW"]
mw.region.resid.df=region.resid.df$lmodfsqrts.residuals[
  region.resid.df$data_State.Region=="MW"]
n.mw=length(mw.region.resid.df)
plot(tail(mw.region.resid.df,n.mw-1)~head(mw.region.resid.df,n.mw-1),
     xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0)

ne.states=data_State$SateAbrev[data_State$Region=="NE"]
ne.region.resid.df=region.resid.df$lmodfsqrts.residuals[region.resid.df$data_State.Region=="NE"]
n.ne=length(ne.region.resid.df)
plot(tail(ne.region.resid.df,n.ne-1)~head(ne.region.resid.df,n.ne-1),
     xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0)


nw.states=data_State$SateAbrev[data_State$Region=="NW"]
nw.region.resid.df=region.resid.df$lmodfsqrts.residuals[region.resid.df$data_State.Region=="NW"]
n.nw=length(nw.region.resid.df)
plot(tail(nw.region.resid.df,n.nw-1)~head(nw.region.resid.df,n.nw-1),
     xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0)


se.states=data_State$SateAbrev[data_State$Region=="SE"]
se.region.resid.df=region.resid.df$lmodfsqrts.residuals[region.resid.df$data_State.Region=="SE"]
n.se=length(se.region.resid.df)
plot(tail(se.region.resid.df,n.se-1)~head(se.region.resid.df,n.se-1),
     xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0)


sw.states=data_State$SateAbrev[data_State$Region=="SW"]
sw.region.resid.df=region.resid.df$lmodfsqrts.residuals[region.resid.df$data_State.Region=="SW"]
n.sw=length(sw.region.resid.df)
plot(tail(sw.region.resid.df,n.sw-1)~head(sw.region.resid.df,n.sw-1),
     xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0)


#Leverage Values
lmod=lm(sqrt(scaled.shot)~HS+Hispanic, data = df2)
lmods=summary(lmod)

hatv=hatvalues(lmod)
states=row.names(df2)
halfnorm(hatv,labs=states,ylab = "Leverages")
qqnorm(rstandard(lmod))
abline(0,1)

#outliers
outlierTest(lmod)
```

```
df3=df2[-32,]
lmod.NJ=lm(sqrt(scaled.shot)~HS+Hispanic, data = df3)
lmod.NJs=summary(lmod.NJ)


perc.change=c()
for(i in 1:3)
{
  perc.change[i]=(lmods$coefficients[i]-lmod.NJs$coefficients[i])/lmods$coefficients[i]
}


#Influential Observations
cook=cooks.distance(lmod)
halfnorm(cook,labs = states,ylab ="Cook's distances")
r1=range(cook)[2]-range(cook)[1]

cook.nj=cooks.distance(lmod.NJ)
halfnorm(cook.nj,labs = states,ylab ="Cook's distances")
r2=range(cook.nj)[2]-range(cook.nj)[1]

perc.change=(r1-r2)/r1

##############################################################################################

#final model
lmod=lm(sqrt(scaled.shot)~HS+Hispanic, data = df2)
lmods=summary(lmod)

##############################################################################################
#Standardized Model
##############################################################################################
totshootstate.std=(df$totshootstate-mean(df$totshootstate))/sd(df$totshootstate)
statepop.std=(df$StatePop-mean(df$StatePop))/sd(df$StatePop)

scaled.shot.std=abs(totshootstate.std-statepop.std)
HS.std=(df2$HS-mean(df2$HS))/sd(df2$HS)
Hispanic.std=(df2$Hispanic-mean(df2$Hispanic))/sd(df2$Hispanic)

df.std=data.frame(scaled.shot.std,HS.std,Hispanic.std)

lmod.std=lm(scaled.shot.std~HS.std+Hispanic.std, data=df.std)
lmod.stds=summary(lmod.std)
```