**Analysis of Population growth and Gun Violence incidents in United States**

Vivek bhat hosmat and Deepika jantz

**Table of Contents**

[**Overview:** 2](#_Toc531615946)

[**Dataset Description:** 2](#_Toc531615947)

[**Research questions:** 3](#_Toc531615948)

[**Planned statistical methods:** 4](#_Toc531615949)

[**Inference on Population Variances:** 5](#_Toc531615950)

[**Simple Linear Regression:** 5](#_Toc531615951)

[**Variable selection and Multiple Linear Regression:** 5](#_Toc531615952)

[**Time Series Analysis:** 5](#_Toc531615953)

[**Forecasting Model:** 5](#_Toc531615954)

[**Test Results, Analysis and Conclusions:** 6](#_Toc531615955)

[**Comparing the variance of two population:** 6](#_Toc531615956)

[**Simple linear regression:** 6](#_Toc531615957)

[**Variable selection method:** 10](#_Toc531615958)

[**Model Building Regression Analysis Forward Selection** 11](#_Toc531615959)

[**Backward Elimination** 11](#_Toc531615960)

[**Stepwise Regression** 12](#_Toc531615961)

[**Best Subsets Regression with the Criteria of R-squared** 12](#_Toc531615962)

[**Time series analysis:** 14](#_Toc531615963)

[**Time Series Regression Model with Seasonality and Trend Components** 16](#_Toc531615964)

[**Multiplicative Time Series Decomposition** 18](#_Toc531615965)

[**Summary and Conclusion:** 21](#_Toc531615966)

[**References:** 22](#_Toc531615967)

## **Overview:**

The issue of gun violence is of paramount importance in the United States with an increase in number of cases in uncommon areas like the Schools, Apartments and other public places. Many numbers of innocent people which includes kids have fallen victim for these acts of violence. For example, America has six times as many firearm homicides as Canada and nearly 16 times as many as Germany. Our data set is originally from *Gun Violence Archive* formed in 2013, which is an online platform that records all the gun related violence occurrence in the United States and has made it available to the General Public with accurate information. However, we found a partially cleansed data set on *Kaggle.com* and we opted this data set for our analysis. The data set under consideration has roughly 240,000 records (incidents). These incidents are dated between January 2013 and March 2018.

As part of data preparation, we had to leave out the data for the year of 2013 as incidents were recorded by Gun Violence Archive only when the number of victims (killed or injured) was greater than four. For this reason, the number of incidents recorded are very few and it gives an impression that there were fewer incidents in the year 2013 which is not the case. We may also opt to leave out the data for the year of 2018 in some cases to avoid a false perception of number of incidents declining for 2018.

To aid our analysis of gun violence with respect to an increase in population, we added a secondary data set of population estimates for each state between 2010 and 2017. The data under consideration are ***population estimates*** and not the actual census (the last census was held in the year 2010). However, the data is from United States census department which gives us enough assurance to go ahead with the data set with respect to credibility of the data.

## **Dataset Description:**

The variables in the gun violence data set are, an Incident ID to uniquely identify each incidence of gun violence, Date of occurrence of the incident, Geographic information of the incident such as city, state, address, location, latitude and longitude, congressional district, state house district and senate house district. The data set also includes the source of the information about the incident, characteristics of the incident, and notes about each incident. The next set of variables state the information related to the kind of guns used in the incident. It also highlights details if the guns are stolen, if yes, it gives the count of stolen guns used. The dataset also gives in detail information about the participant involved - the age of the victims, the age group, participant gender. Please note, the participant here involves both the victim and offender.

The variables in the population estimates data set are, states in United states and their population estimates for the years between 2010 and 2017. We also have the population census for the year of 2010, as we do not have the pertaining data for gun violence, we would be considering only 2013-2017 in our analysis.

In our original dataset of Gun violence incidents, we did not have enough variables to run the variable selection method, so we merged another dataset which had number of permits issued every month, number of hand guns and number of long guns.

## **Research questions:**

The objective of this project is to find answers to the questions stated and understand the pattern of gun violence occurrence in the United States.

* How does the gun violations vary with the growth in population of a state? Our assumption to start with is, states with higher population growth will have higher gun violations.
* Is there a seasonality influence on the number of incidents?
* How does the number of guns in a state compared with number of victims?

#### **Descriptive statistics:**

The below table summarizes the variables of interest in our data set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Total Killed** | **Total Injured** | **Average Killed** | **Average Injured** | **SD\_Killed** | **SD\_Injured** |
| 2013 | 317 | 979 | 1.14 | 3.52 | 1.53 | 2.31 |
| 2014 | 12557 | 23002 | 0.24 | 0.44 | 0.5 | 0.7 |
| 2015 | 13484 | 26967 | 0.25 | 0.5 | 0.51 | 0.72 |
| 2016 | 15066 | 30580 | 0.26 | 0.52 | 0.54 | 0.75 |
| 2017 | 15511 | 30703 | 0.25 | 0.5 | 0.51 | 0.71 |
| 2018 | 3533 | 6171 | 0.26 | 0.45 | 0.53 | 0.67 |

As stated earlier, the number of incidents in the year 2013 is extremely low as compared to the previous year. From the statistics we can also observe that there is a steady increase in number of incidents pertaining to gun violence. Through our analysis we would like to attribute this increase to probable factors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **Mean Population** | **Median Population** | **Std Deviation** | **Minimum Population** | **Max Population** |
| 2013 | 2480270 | 17596484 | 28169493 | 2329364 | 153389532 |
| 2014 | 24990002 | 17641660 | 28465669 | 2333336 | 154805112 |
| 2015 | 25179595 | 17688228 | 28765016 | 2344408 | 156129776 |
| 2016 | 25365171 | 17744452 | 29028604 | 2339640 | 157185904 |
| 2017 | 25546602 | 17816756 | 29267053 | 2317260 | 158146612 |

As anticipated, there is a steady increase in the overall population across states in United States. This data will help us in understanding the impact of increase in population on gun violence incidents.

## **Planned statistical methods:**

We have arrived upon the following models based on the exploratory analysis of the data.

## **Inference on Population Variances:**

Using this method, we intend to see if the population growth variances for the state with highest population and least population year-on-year is same. This gives us an idea on whether the rate of increase in population is uniform across these states.

## **Simple Linear Regression:**

We are planning on building a Simple Linear Regression model on how the number of gun violence varies with respect to the overall growth in population of State. The hypothesis for this would be

*H₀: ẞ₁ = 0*

*Hₐ: ẞ₁ ≠ 0*

## **Variable selection and Multiple Linear Regression:**

We would like to see if there is any influence of gun violence based on the number of permits issued by the states, number of hand guns and long guns possessed in a state. To understand this, we would first run a variable selection method to check if the three independent variables stated are significant and then we would fit a Multiple linear regression model for the variables selected. The hypothesis is-

*H₀: ẞ₁ = ẞ₁ = … ẞ₈ = 0*

*Hₐ: One or more of the parameters is not equal to zero.*

## **Time Series Analysis:**

We would like to understand seasonal patterns using the time series analysis of number of gun violence (people injured and killed) over the years under consideration.

## **Forecasting Model:**

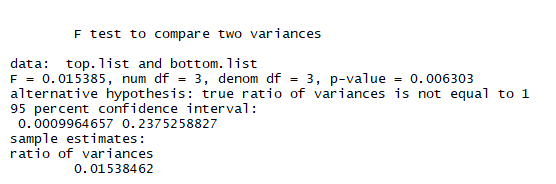
Based on the dataset we have from 2013-2018, we would like to build a forecasting model to predict the number of gun violence induced incidents over the next year. We would also like to analyze the accuracy of the forecasting model being built to reproduce the time-series data that we have for the occurrence of gun violence incidents.

## **Test Results, Analysis and Conclusions:**

For the below tests, we used R as the statistical tool for running programs to test the hypothesis and analyze the results.

## **Comparing the variance of two population:**

We are assuming in our hypotheses that the population variance for the state with highest population and for the state with lowest population is equal.

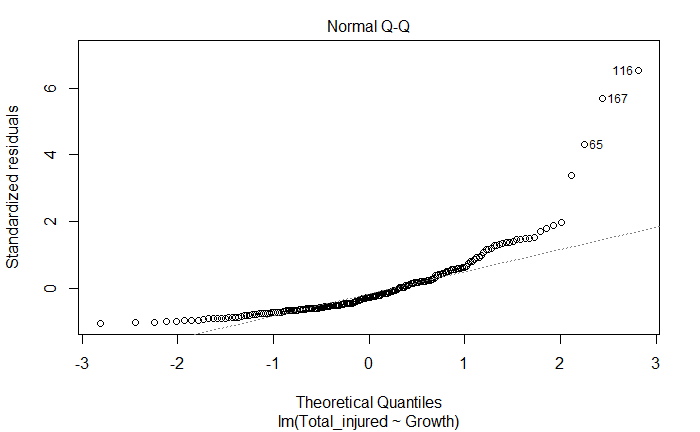


From the above summary of F test, we see that at 95% confidence interval p-value is less than the **value considered for the F test and hence we can reject the null hypothesis H₀ which signifies that the population growth variance for the states with highest and least population are not the same. This in-turn states that the rate at which the population is growing in the state of California (highest population) and Wyoming (least population) is not the same.

## **Simple linear regression:**

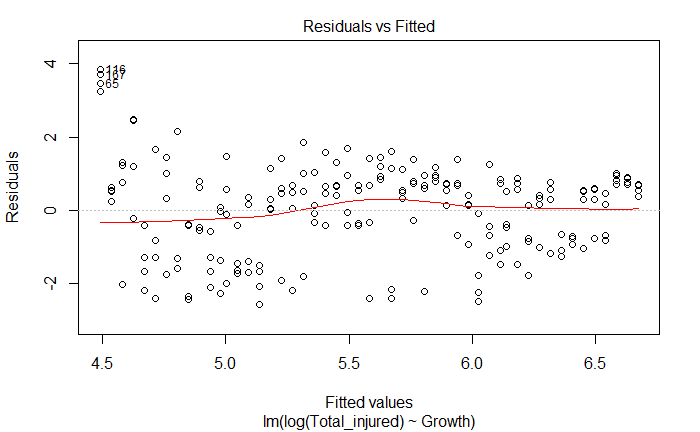
Our general assumption is that higher the population, higher is the crime rate. We wanted to test this by testing if there is an increase in number of gun violence related incidents with increase in population growth. To understand this, we considered the population growth estimates for all the states in United States and scaled it to percentile values ranging from 0 to 1 to make it comparable with one another. For this analysis, we considered the population growth estimates to be the independent variable and the total number of people injured due to gun violence as the dependent variable.

By plotting a simple linear regression by considering the population growth and the number of people injured, we see that the linear model under-estimates at lower and higher values of population growth estimates.

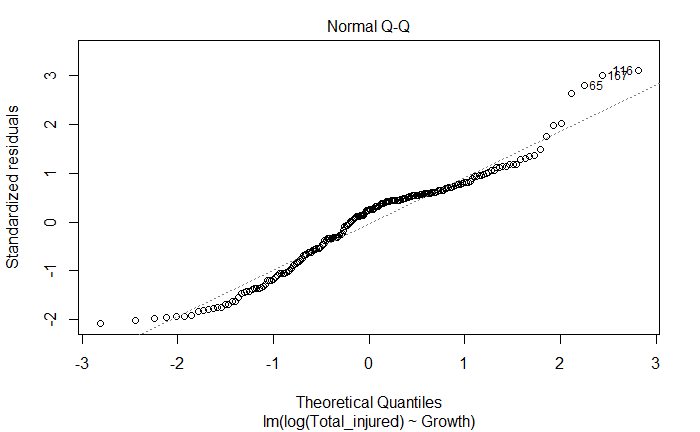


For a better fitting model, we considered taking log of the dependent variable and hence arriving upon the equation,

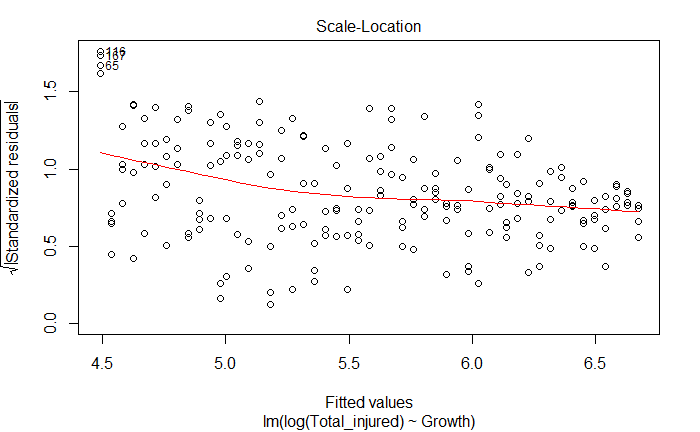
Log(Total Injured) = 4.4470 + 2.22243\*(Population growth percentile)



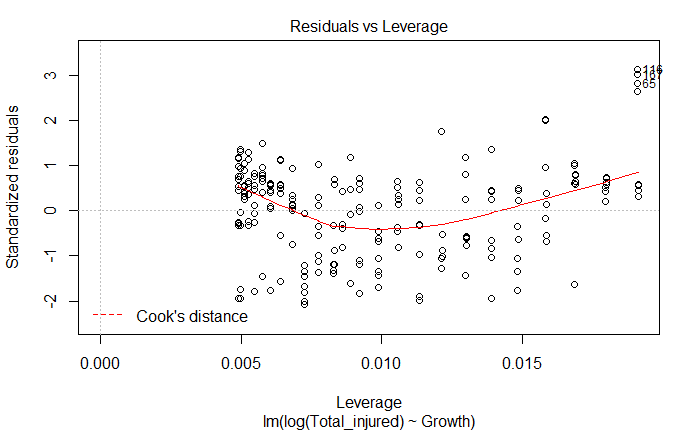
This plot shows if residuals have non-linear patterns. We see that the values are equally spread residuals around the horizontal line without distinct patterns, this is a good indication we don’t have non-linear relationships.



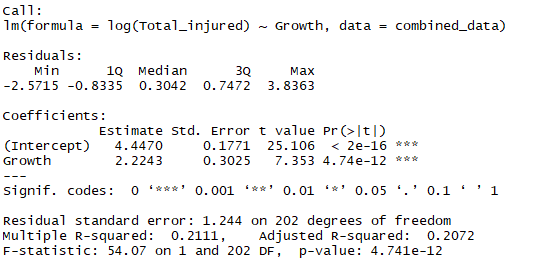
This plot shows if residuals are normally distributed. We see that majority of the values lie on the dotted line except for few at higher values of population growth estimates.



This plot shows if residuals are spread equally along the ranges of predictors. We see that the line is nearly horizontal which signifies that residuals are spread equally along the ranges of predictors.



This plot helps us in identifying outliers and leverage points. We see that 116,16 and 65 are outliers but are not leverage points.

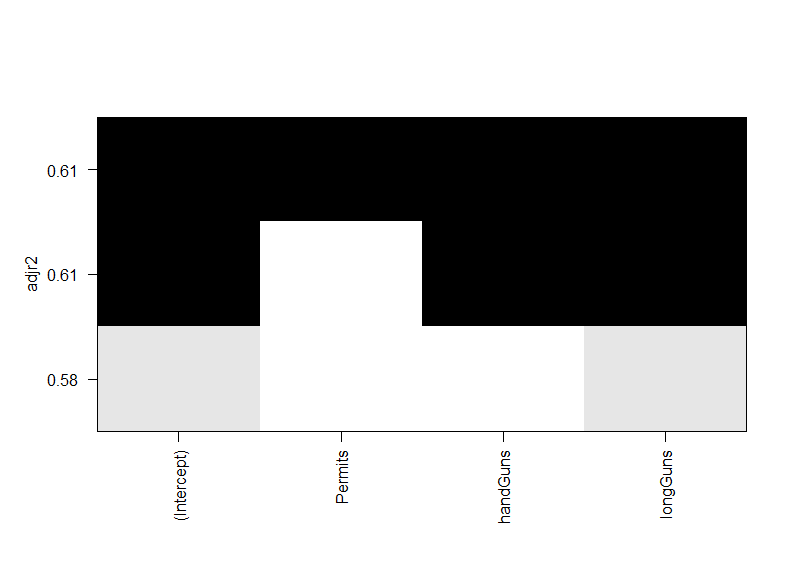


From the summary, we see that variables involved are highly significant and the p-value is again highly significant with a value of 4.741e-12. However, the adjusted R-squared has the value of 0.2072 which signifies that the model is not a great fit and there is no significant relationship between population growth and the number of people injured due to gun violence.

## **Variable selection method:**

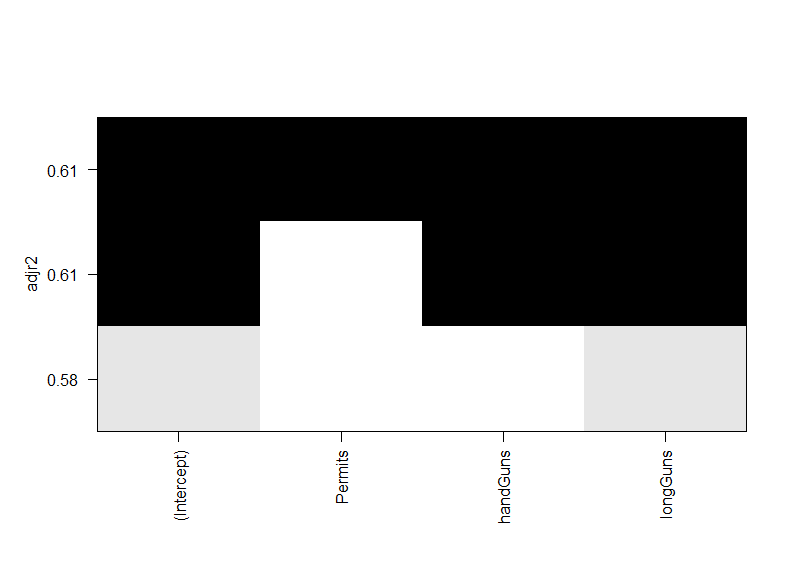
Our next step was inspired by the need for understanding how the availability of guns varies with the number of victims due to gun violence. To facilitate this, we had to merge another dataset as stated earlier.

##### **Model Building Regression Analysis Forward Selection**



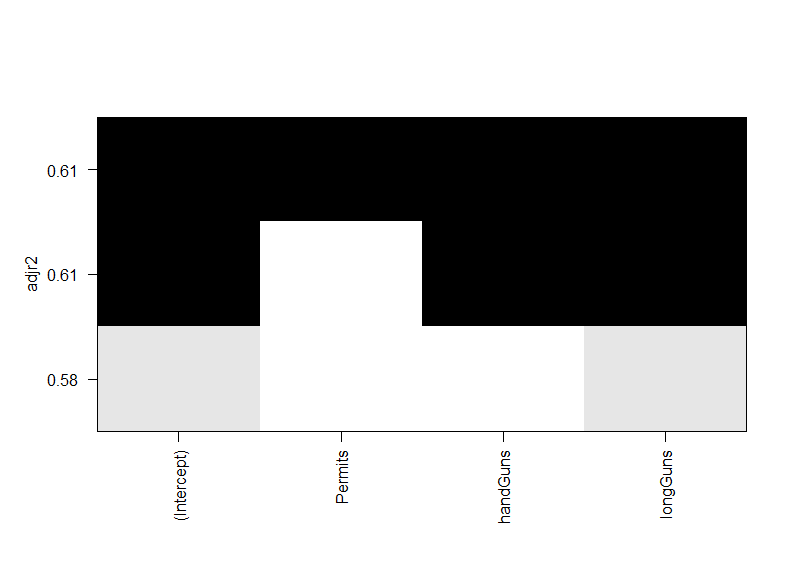
The method of Forward selection reports that the best model to predict the total number of victims is the one that includes all three variables Permit, hand guns and long guns are significant.

##### **Backward Elimination**



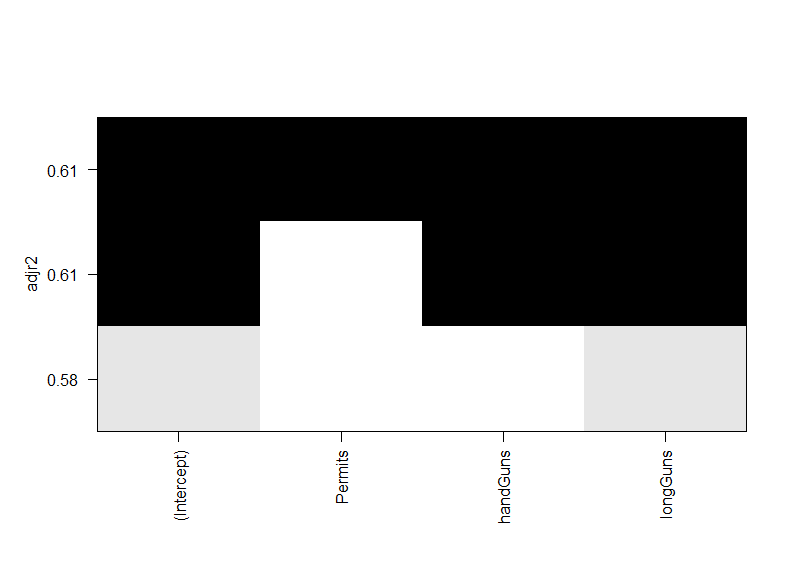
The method of backward elimination reports that the best model to predict the total number of victims is the one that includes all three variables Permit, hand guns and long guns are significant.

##### **Stepwise Regression**



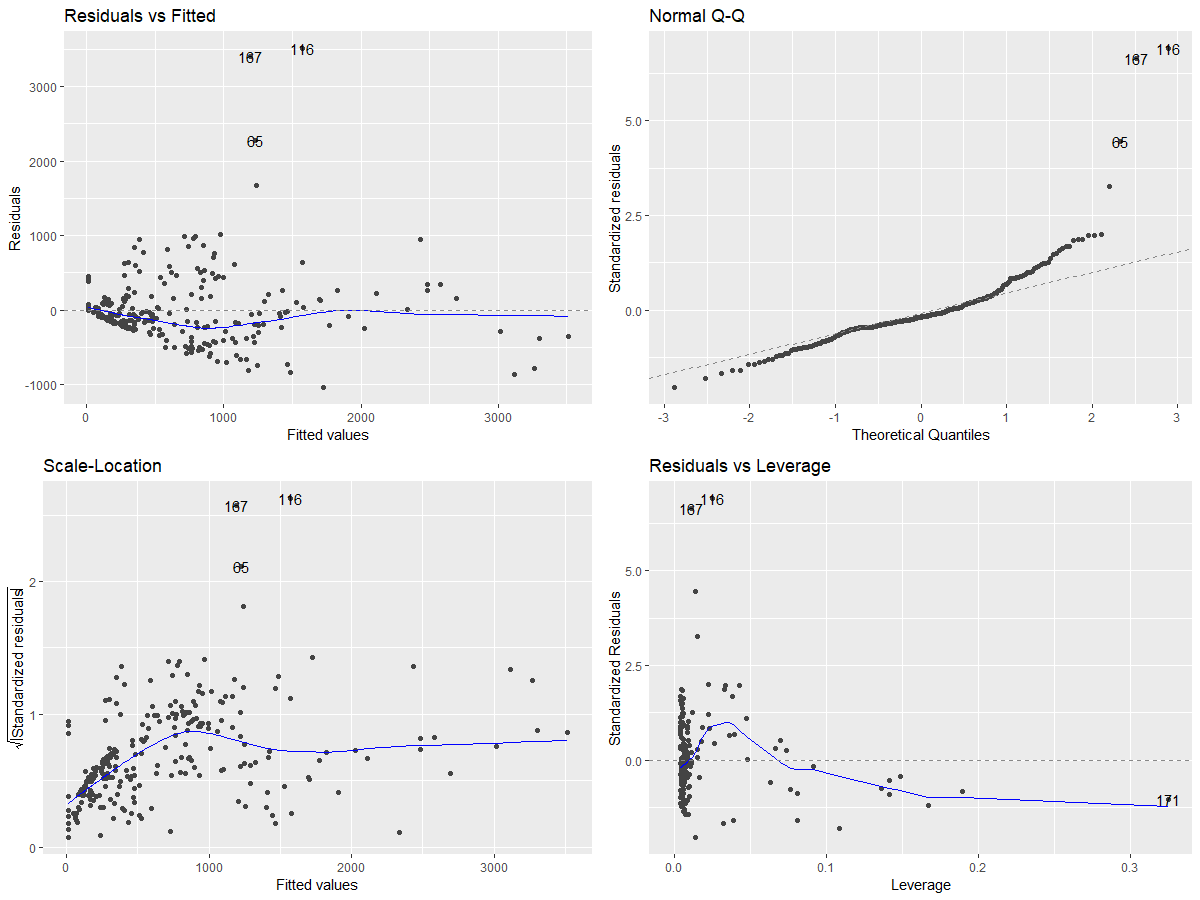
The method of stepwise regression reports that the best model to predict the total number of victims is the one that includes all three variables Permit, hand guns and long guns are significant.

##### **Best Subsets Regression with the Criteria of R-squared**



The method of Best subsets regression with the criteria of R-squared reports that the best model to predict the total number of victims is the one that includes all three variables Permit, hand guns and long guns are significant.

From all the models we considered, it is extremely evident that all 3 variables are to be considered to build the best model. Let’s build a multilinear model with Number of victims being the dependent variable and Number of permits, hand guns and long guns being the independent variables.



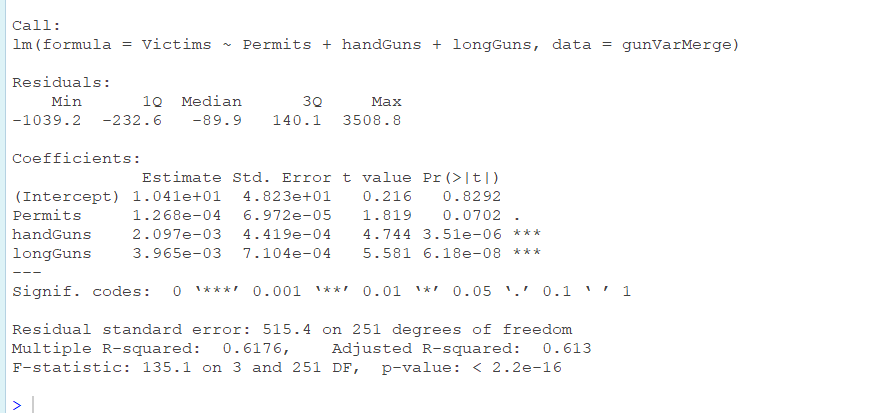
This Residuals vs fitted plot shows if residuals have non-linear patterns. We see that the values are equally spread residuals around the horizontal line without distinct patterns, this is a good indication we don’t have non-linear relationships.

The Normal Q-Q plot shows that the observations are not normally distributed as we see many of them away from the dotted line.

The scale-location plot indicates if the variance is constant, at the lower end we do see an increasing standard residual however after 1000 the variance seems to be constant.

In the residual vs Leverage plot, we do see 167 and 116 which are outliers as well as leverage points.

The summary of the model is displayed below,



We see that number of gun permits for a month in a state is significant at 90% confidence interval and the other two variable number of hand gun and long gun are extremely significant at 99% confidence as well. The p-value of F-statistic at 2.2e-16 is extremely significant at given value of alpha. The Adjusted R-squared is decent at 61.3% but is not conclusive to state that the model correctly predicts the number of victims.

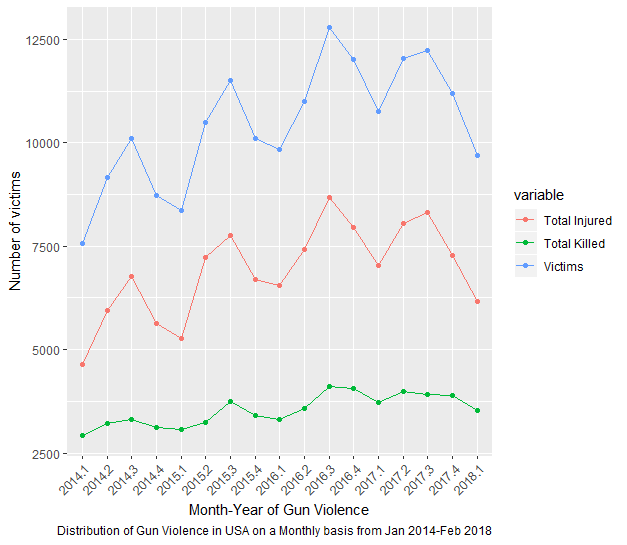
The equation for the arrived model is given below,

NumberOfVictim[Estimate] = 1.041\*101 +1.268\*10-4(Permits) + 2.097\*10-3 (handGuns)+3.965\*10-3(longGuns)

From the equation, we see that the null hypothesis can be rejected since atleast one of the co-efficients of the parameters is not equal to 0.

## **Time series analysis:**

Since the data under consideration is over a certain time period, we would plot a time series plot to understand how the data varies with time. We were interested in two variables in our dataset, number of people killed, and number of people injured. To better visualize the data, we divided the time into quarters. We have considered the data from 1st quarter of 2014 to 1st quarter of 2018. The third variable represented as victims is a sum of total number of people killed and injured.

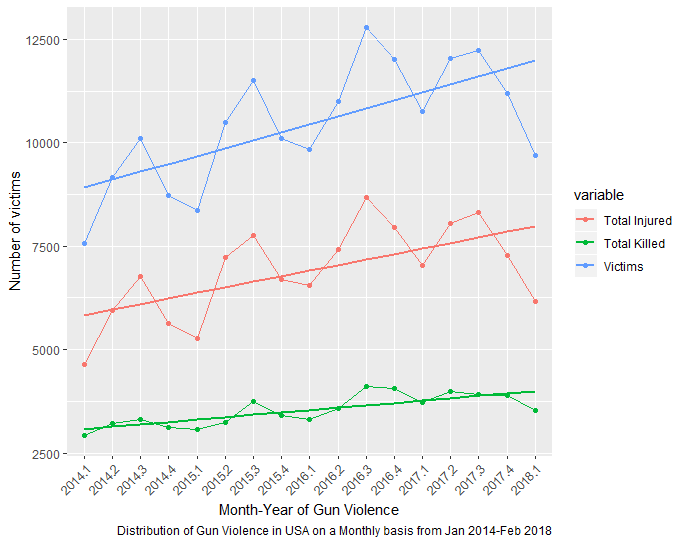


We can clearly see that, there are significantly higher number of people killed and injured in the 3rd quarter, this is represented by the peaks in the 3rd quarter. Apart from the seasonality component we also observe a trend of increase in number of people injured and killed. The upward trend is even more evident in the number of victims’ plot. Based on this plot, it is conclusive that we must arrive upon a forecasting model that has the capability of dealing with both seasonality and the trend components. To forecast the total number of victims, we would consider Time Series Regression with Seasonality and Trend Components and a Multiplicative Time Series Decomposition method.

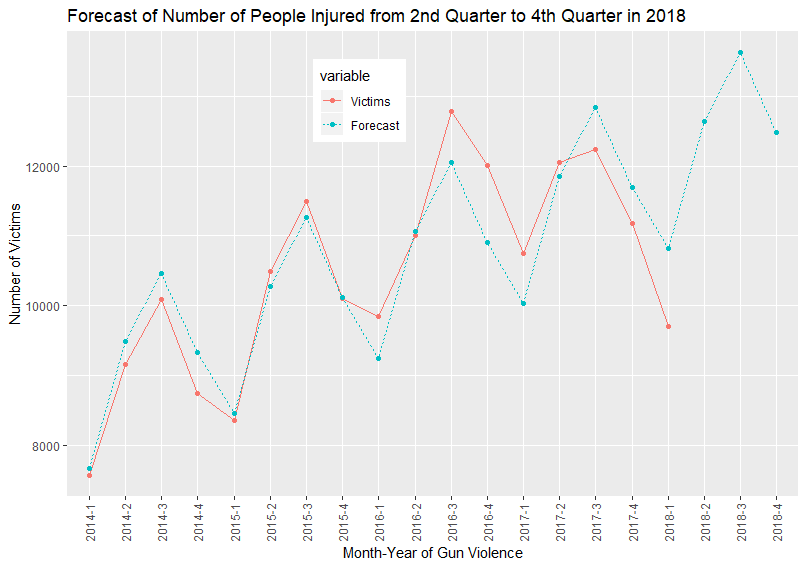
## **Time Series Regression Model with Seasonality and Trend Components**

To make it comparable in the further analysis we would be using Total number of victims (sum of number of people killed and injured) as the variable changing over time.

Based on the time series plot, an upward trend was very evident for total number of victims over the years.

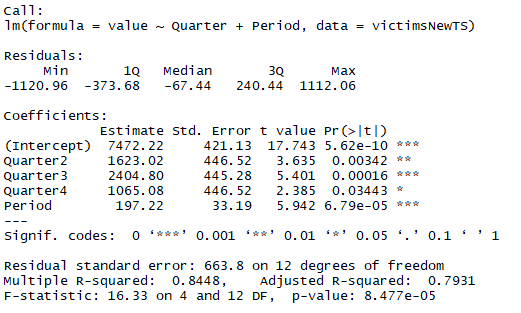


From the trend lines, as noted earlier there is an upward trend in total number of victims between the years 2014 and 2018.



The equation for the time series regression model is stated below,

EstimatedVictimCount= 7472.22+   1623.02  (Quarter2) +2404.80  (Quarter3) + 1065.08  (Quarter4)+197.22  (Time)



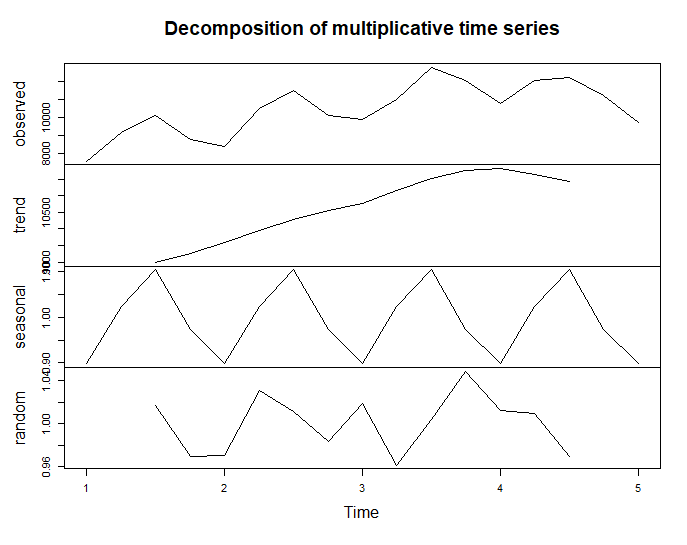
We see that all the variables in the equation are significant at 95% confidence. The F statistics p-value = 8.477e-05 is very significant at any significance level which suggests that all the variables in the group are significant in predicting the number of victims due to gun violence incidents. The model shows a decent R-squared and adjusted R-squared value (0.84 and 0.79 respectively) which indicates that there is a relationship between Quarter and period.

Using the method above, the forecasted values for the remaining quarters of 2018 obtained are displayed in the table,

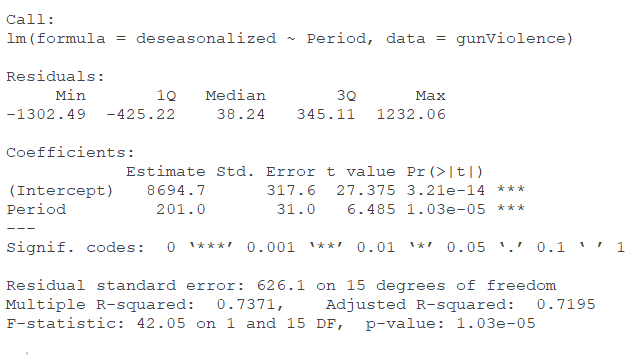
|  |  |  |
| --- | --- | --- |
| **Year-Quarter** | **Victims** | **Forecasted Victims** |
| 2018-1 | 9704 | 10824.96 |
| 2018-2 |  | 12645.2 |
| 2018-3 |  | 13624.2 |
| 2018-4 |  | 12481.7 |

## **Multiplicative Time Series Decomposition**

Decomposition methods assume that the actual time series value at period t is a function of three components: trend, seasonal, and irregular. The relationship between these three components is best described by a multiplicative model in this case, since the seasonal fluctuations grow larger as the Number of victims increases over time due to the upward trend observed.



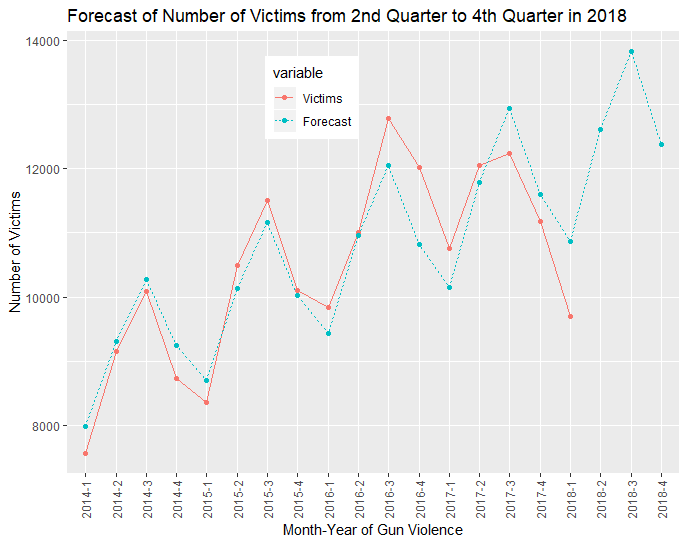
As part of multiplicative time series decomposition, we must de-seasonalize the data to identify trend. Based on the de-seasonalized total number of victims, a simple linear regression model with the time period t as independent variable will be created.



Using the least squares method: Tt = b0 + b1t

De-seasonalized number of victims t = 8694.7 + 201.0 t

Using the de-seasonalized total number of victims, we will estimate the number victims for the next 3 quarters of 2018.



The forecasted number of victims for the next 3 quarters of 2018 are,

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Year-Quarter** | **Victims** | **Forecasted Victims** |
| 2018-1 | 9704 | 10873.325 |
| 2018-2 |  | 12607.258 |
| 2018-3 |  | 13820.778 |
| 2018-4 |  | 12382.04 |

Finally, we would select the better model between the two by calculating the Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). As the number involved is quite big, the MSE can be disregarded in this case and if we compare the MAPE for the Time series regression with Seasonality and Trend components and Multiplicative Time Series Decomposition, they are almost the same with Time Series Regression having slightly lower MAPE. On this basis, we would pick Time series regression model for forecasting the number of victims for the next 3 quarters in 2018.

|  |  |  |
| --- | --- | --- |
| **Method** | **MSE** | **MAPE** |
| Time Series Regression with Seasonality and Trend Components | 311024.5 | 0.04221286 |
| Multiplicative Time Series Decomposition | 321883.9 | 0.04442178 |

## **Summary and Conclusion:**

For the inference of two population variances, we wanted to understand if the rate at which the population is growing is uniform across all the states. We came up with a suitable hypothesis, through our F test we found that the variances in terms of year-on-year growth for the state with highest and lowest population is not the same.

In the case of linear regression models, we observed that there is a logarithmic transformation in the dependent variable, Number of Injured people due to Gun Violence increases by [exp (2.2243) + exp (4.4470)] 94.62 units, for one unit increase in the population growth in the states over the duration. We would hence conclude that the number of people injured due to gun violence is dependent on the population growth in a state. This analysis helps us in answering the first research question, Is there a relationship between number of victims(injured) and population growth?

For multiple linear regression model, we used the variable selection methods to find out if the independent variables under consideration – (Number of Permits in a year for a State, Number of Shotguns, Number of Handguns), have any influence in determining the number of Victims for a given year in a given State. With all the 4-model building methodology, we see that all the three variables have influence in determining the Number of Victims affected due to gun violence.

|  |  |
| --- | --- |
| Independent Variable | Estimate of Number of Victims |
| Permits | NumberOfVictim[Estimate] =10.41 |
| Hand Guns | NumberOfVictim[Estimate] = 10.4101268 |
| Long Guns | NumberOfVictim[Estimate] = 10.4140918 |

The multilinear model helps us in answering the third research question of whether the number of guns possessed in a state has an affect on the number of victims.

For the time series models, we observed the presence of seasonality and trend components in the data. We developed two models which takes the above stated components into consideration. The Time Series Regression with Seasonality and Trend Components had marginally better MAPE compared to the Multiplicative Time Series Decomposition model, we decided to go with the former for our predictions. The time series analysis also answers the second research question of a presence of seasonality in the data we have.

## **References:**

1. Ko, James. “Gun Violence Data.” *Kaggle: Your Home for Data Science*, 15 Apr. 2018, [www.kaggle.com/jameslko/gun-violence-data](http://www.kaggle.com/jameslko/gun-violence-data).
2. “Gun Violence Archive.” *Gun Violence Archive*, www.gunviolencearchive.org/.
3. *Index of /Programs-Surveys/Popest/Datasets/2010-2017/State/Asrh*, www2.census.gov/programs-surveys/popest/datasets/2010-2017/state/asrh/.
4. *Total Population by Country 2018*, worldpopulationreview.com/states/.
5. US Census Bureau. “Data.” *Census Bureau QuickFacts*, United States Census Bureau, 4 Sept. 2018, www.census.gov/data/tables/2017/demo/popest/state-total.html#par\_textimage.