Revealing the common factors between the victims from police killing in the US.

Writer: Samuel Chen

Project Category: Police Killing

Date: 2021/4/18

Data source:

<https://github.com/washingtonpost/data-police-shootings/blob/master/fatal-police-shootings-data.csv>

## 1 Introduction

Police killing is an inevitable problem, especially after the Killing for George Floyd, this problem has brought attentions of Police killing to many people across the world.

This project aims to revealing the common factors (such as race, mental) of Police killing in the US using multiple different models, such as linear model or logistic model and trying to predict what kind of person will most likely be shot by the police.

## 2 Related work

The Original report:

<https://www.washingtonpost.com/graphics/investigations/police-shootings-database/>

In this original report, the article first plots the data using Months (from Jan. to Dec.) vs. cumulative total shootings through. The plot shows from 2015 to 2021, the number of killing has been increasing as the time goes in that year. However, the year 2020 has most police killing reports, it has slightly higher increasing rate than other years, it may because the protests happened during the year 2020.

The next plot, the report uses the percentage to show the how many victims were killed vs. their race in million. This plot shows us the rate at which black Americans are killed by police is more than twice as high as the rate for white Americans.

The third plot shows the age group which the victims were belonged to. It clearly states that from 0 increment by 5-years most victims were within the four age group from 20 to 39, they were young male.

## 3 The DATASET

This dataset is accessed from <https://www.washingtonpost.com/> on 2021/4/18. In 2015, The Washington Post began to log [every fatal shooting](https://www.washingtonpost.com/graphics/national/police-shootings-year-end/) by an on-duty police officer in the United States.

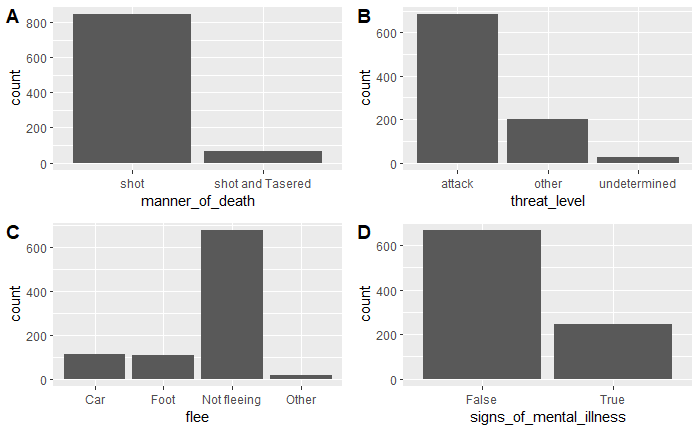
This dataset contains 6214 observations(victims) and 17 features, such as age, gender, race, armed and manner of death etc. After filtering out the observations with missing value and picking the meaningful features for this project, the dimension of the data set is 4817 by 9.

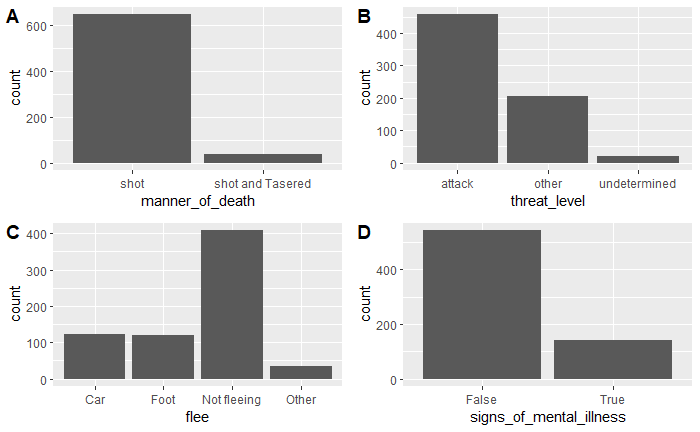
The features kept at this step are:

("manner\_of\_death","armed","gender","race","signs\_of\_mental\_illness","threat\_level","flee").

This project will only focus the police killing happened in the year of 2015 and 2020.

This is small summary of both subset year of 2015 and 2020:





The summary shows for both years, the factors of police killing has not been changed, most of the victims were shot instead of tasered, most victims attacked the police, most victim were not fleeing and most victim were showing no signs of mental illness.

## 4 Models

**The potential drawback of the original model:**

However, this model in the original report has a huge drawback due to the biased portion of different races in the data set, the amount of RaceB is much higher than anyother race in this dataset.

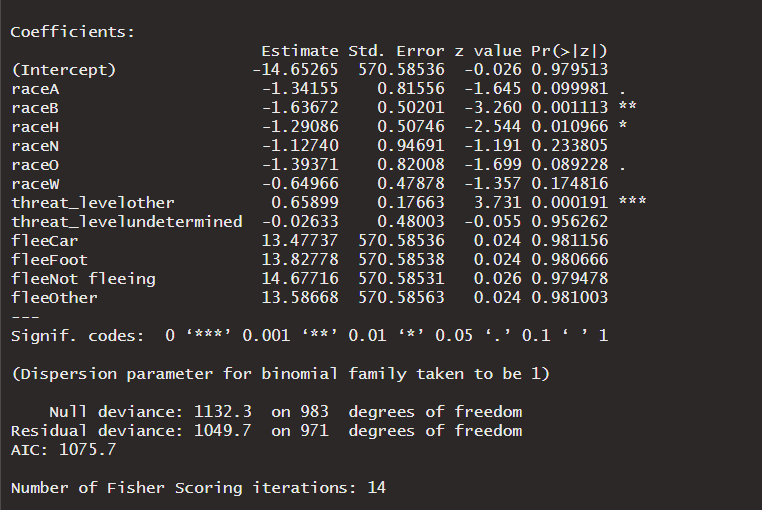
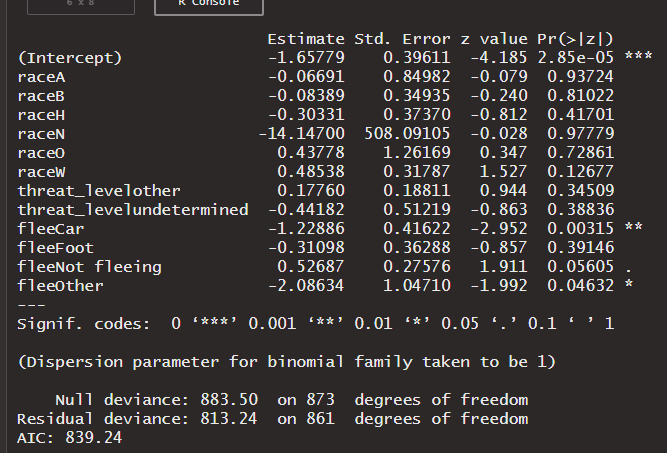
The first model of this project is using logistic regression model to predict/interpret the relationship between the factor signs\_of\_mental\_illness compare to other features (race + threat\_level + flee) :

Based on the summary of this logistic model, the result showing the factor `RaceB` and `Threadt\_levelother` are significant to the first logistic model in 2015.

The model for the 2020 data suggests a different result. This model can show the most significant factors for the features.

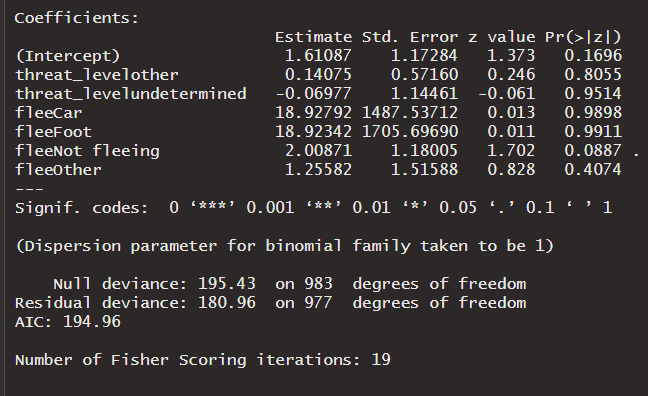
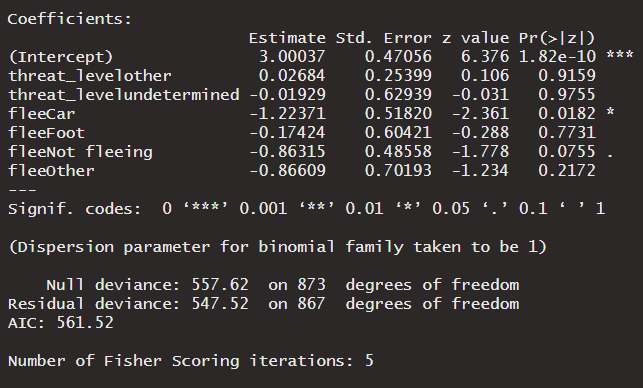
**Pro & con:**

There is also a huge drawback for this model, the format of the dataset and the type of the variable being factor and character/string has a huge impact on the prediction of the signs\_of\_mental\_illness of the victim.

Then the project will focus on the feature: races vs. the features: threat\_level + flee:

This is the result of the summary of the model:

**Questions:**

The original reports show the proportions of the factors in the same categories, but it did not show the correlation of the features such as the relationship between the victims’s fleeing method and the manner of death, will the victim have more chance getting shot by guns instead pf taser if he/she tried to escape with the car, or will the victim have more chance getting tasered or a fatal shot.

## 5 Results

Those results from above suggest there is no significant feature that can determine/predict the type of the victim, since the significance of those factors are small enough to drop in most case, the prediction from this model will also be in accurate, it will have a lot of estimate errors.

## 6 DISCUSSION AND CONCLUSIONS

Due to the format and having almost no numeric data, this dataset is hard to interpret without proper conversion from factors to numeric, and due to an issue in this dataset, the prediction of those model does not work as intended to. It is reporting an error saying there’s NA data in the dataset after the Missing Value and empty var is removed from the data set.

It is fascinating to learn about the data science behind an event, as before the writer thought the most victims had fight with police and tried to flee before they got fatal shots, but it turns out most of the victims were not fleeing and there were also a small number of victims that were not even armed.

This dataset is much better to do a unsupervised learning instead of a supervised learning, there is not a lot of data we can get from the prediction due to the small significance of the factors, but we can try to grouping all the observation together using a unsupervised learning methods to study the pattern behind the factors of the victims of the police killing in order to reduce the police killing.

## Works CITED:

<https://github.com/washingtonpost/data-police-shootings/blob/master/fatal-police-shootings-data.csv>

<https://www.washingtonpost.com/graphics/investigations/police-shootings-database/>