PILOT RESEARCH PROJECT – POLICE KILLINGS IN US DURING 2015

Barbara Annalisa, Morandini Maria

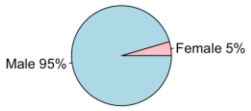
{January 2022}

1. INTRODUCTION

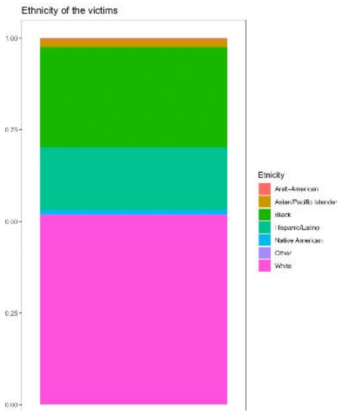
During these years there have been numerous movements in the US against unjustified police violence which very often led to tragic consequences.
In this paper, we thus want to analyse the social and economic backgrounds that the victims seem to have in common, trying to recover the underlying reasons that have led to such a dramatic result in the last years.
More specifically, we will focus on some variables, such as gender, ethnicity, level of wealth and level of schooling of the victims.
Before doing this, however, it is necessary to make some clarifications regarding the limits of our study and the data sets used.
The main source that was used for our research is a data sets built by the Guardian which, thanks to numerous testimonies, the work of some journalists and some official data provided by the government, has reconstructed the story of the people dead in 2015 due to this cause. Although numerous sources have been crossed for the realization of this project by the Guardian, the number of victims does not reflect the real number of them as many have never been officially declared by the police as deaths due to this cause. Although, we do not have a data set that includes all the victims, we still have over 1000 cases available that will allow us to have insights, not on the spread of the phenomenon quantitatively in absolute terms, but mostly on the economic and social background recurring among the victims of the sample.
Moreover, parts of our studies will also refer to another data set, for which we leave the references in the appendix, that is an elaboration and an extension of the Guardian's data set as more variables are specified, trough data crossing with the available data from Census Official American website. Unfortunately, in this last case the sample is reduced in size, as it was updated on 2 June 2015, so the killings that took place after that date are neglected, hence it only covers the first half of the year 2015.
As a last remark we also mention that, before using the data set in R, we have cleaned it deleting the rows where NA entries were present.
Note: in the work it will be mentioned when the larger data set is used, with an asterisk (*), so in general it can be safely assumed that we are referring to the smaller one.

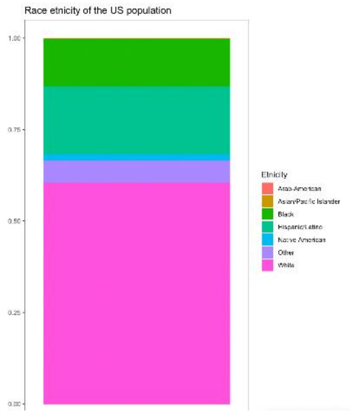
2. PLOTTING

First, to understand in which direction to focus our interest we made some plots using different variables. It is quite clear that most of the victims, exactly 95% are males (*).

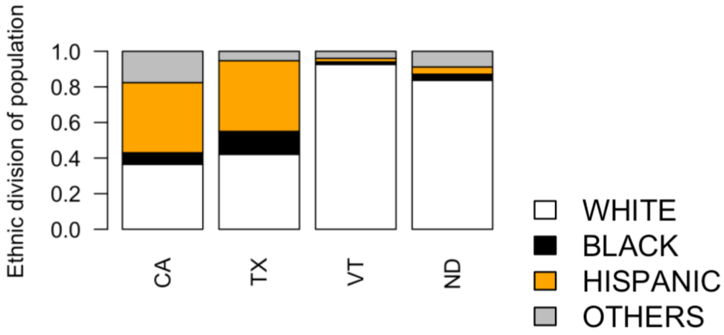


Regarding instead the ethnic group that the dyed people belonged to, we can see through a pie chart (*) that 52% of the victims are white people, followed by 27% of Black people and 17% of Hispanic/Latino people with finally a minority consisting of 2% of Asian people. Although white people would seem to be the most affected, however, it should be emphasized that the total percentage of white people in the US is about 60% against 13% of Black people and 18% of Hispanic / Latino ones; it would therefore be more reasonable to say that, Black people and Hispanic / Latino people are more easily persecuted than white people compared to their total proportion in the US population but it is still too early to draw hasty conclusions...





It also seems that in the states where there is a higher percentage of Black and Hispanic / Latino people there is a higher number of police killings, as the bar plot (*) of the ethnicity proportions of the most and least affected states shows. Over the whole year, we have that the states where more killings have occurred are California and Texas with respectively 211 and 112 killings while Nevada and Vermont were the less affected ones with only one killing recorded. From the graphs we can deduce that there is a bond between the higher percentages of ethnicity different from the white one and the higher number of deaths, but we will come back to these results later.



Concerning the age (*) at which the victims were killed, the most frequent age is 24 years with an average of 37 years, and a median of 34 on the whole year-based sample.

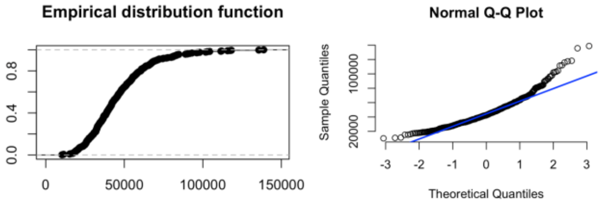
We have also made some other plots which, despite leading to fairly obvious conclusions, we mention briefly. From the data it emerges a negative relationship (correlation coefficient of -0.512) between the victim's house income and the unemployment rate in that county and a positive relationship (correlation coefficient of 0.665) between the victim's house income and the percentage of people who went to college within the county.

Shifting the attention instead to the cause of death (*) we immediately see that death by gunshot is the main cause while death by taser or because the victim had been hit by a police vehicle or held in custody represent only rare causes of death.

Although we can already graphically deduce interesting relationships, in the following sections we will go to see if there are indeed statistical evidence that confirm our hypotheses or if our assumptions should remain only such.

3. HYPOTHESIS TESTING

To begin we want to test whether the mean house income of the areas where the victims lived is bigger or lower than the mean US house income in 2015. First, we plot the empirical distribution function, and we see that it could resemble the CDF of a normal distribution.



Secondly, we perform a QQ-plot to compare the sample we have with the normal distribution.

We also decided to perform a Kolmogorov-Smirnov test on the sample, to test the null hypothesis H_0:sample~N(μ, σ^2) with μ=samplemean(X) and σ ^2=sd(X), but the results that we obtain show that the p-value is very small (0.005320532<0.05): we thus have to reject the null hypothesis, that shows that the sample is not normally distributed, actually this is something quite common for large data set coming directly from "real life".

We should stop here with our test, since the basic assumption to perform a Gauss Test (the sample should be normally distributed) does not hold, but since the QQ-plot does not show a bad fitting from the normal distribution we still try do it and comment the results.

Running the test we obtain a very low p-value that allow us to conclude that if our household income sample would have been normally distributed, then μ_0 ≠ μwhere μ_0 stands for the American average household income.

Additionally, just comparing the mean of the average house income of the areas where the victims lived and the average US mean income, we obtain an income of 46379 against 56516 dollars of the mean US income and from this we can consider plausible that victims tend to live in poorer areas.

4. REGRESSIONS

4.1 R2: age, urate and gender

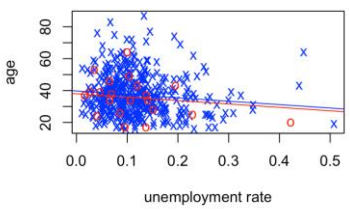
As a first regression we wanted to investigate a possible correlation between the unemployment rate in the geographic area of belonging and the gender as predictor variables while the age of death of the victim as dependent variable.

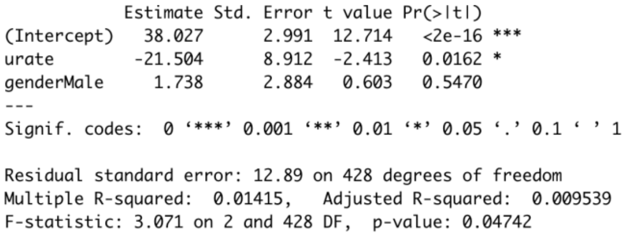
We have modelled the gender as a dummy variable. The employment rate was calculated on the basis of the areas defined as tract-levels by the American census that are small areas with similar socio-economic characteristics.

Before commenting on the output of R, however, it is necessary to premise that, since the number of data available for female victims is much lower than that available for males, the analysis we are doing cannot be considered conclusive in any case as we have too little data available on female victims.

Looking at the output of R we see that the *urate* has as an estimate of the coefficient with a p-value equal to 0.0111 while the male gender has a coefficient equal to 0.7507 with a p-value of 0.8490.

We therefore deduce that gender does not seem to be correlated with the age of death so we repeat the linear regression by removing the gender and leaving only the unemployment rate as predictor variable obtaining that unemployment rate seems to negatively influence the age of death but being the p- value obtained by testing the estimate of the parameter *urate*, equal to 0.0104 , we try to run again the linear regression without considering the gender variable in order to decide if retaining the null hypothesis that would mean that unemployment rate is not correlated in determining the age of death of the victim or reject it.





```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.027      2.991  12.714   <2e-16 ***
urate         -21.504      8.912  -2.413   0.0162 *
genderMale      1.738      2.884   0.603   0.5470
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.89 on 428 degrees of freedom
Multiple R-squared:  0.01415,	Adjusted R-squared:  0.009539
F-statistic: 3.071 on 2 and 428 DF,  p-value: 0.04742
```

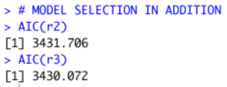
4.2 R3: deleting the gender

However , after having removed the gender variable, the p-value related to the urate coefficient is still larger than 0.5, hence probably also the unemployment rate is not a crucial variable in determining the age of death of the victims.

Brief excursus on model selection:

Before giving conclusions, we must remark that the driving line of our project has not been to find the best fitting model, as we wanted to explore the factors under many points of view. For some of them we wanted to see the influence they have on some quantitative variables, such as age, and for some others the influence they had on some categorical ones. Thus, our purpose is not in general to find the best fitting model, but to explore some relevant relations.

Nevertheless, we can perform a Model Selection to the regressions r2 and r3 that take into consideration two factors that might influence the age of death of the victims. Using the AIC criterion, as it is useful in comparative terms and not in absolute terms we obtain :



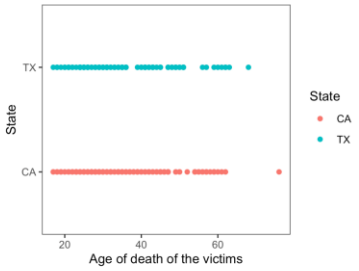
```
> # MODEL SELECTION IN ADDITION
> AIC(r2)
[1] 3431.706
> AIC(r3)
[1] 3430.072
```

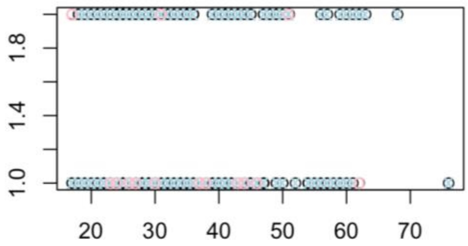
This implies that the r3 model is slightly more fitting, hence the gender variable can be removed as already mentioned above. It should be noticed, however, that in this case the AIC method is not very sensible since the two models differ by only one dimension and in general AIC works better when the dimensions of the model you are testing are higher.

4.3 R4: age, state, gender

We then performed a multi-dimensional regression (*) on the age of the victim, taking care of the parameters *state* and *gender*.

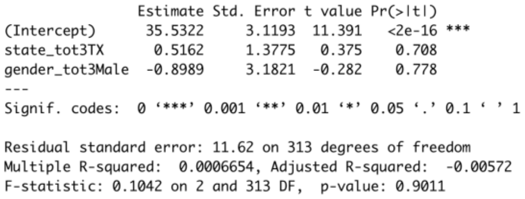
For simplicity we will consider only California and Texas, to avoid the use of 49 dummy variables.





We plot in the graph the variables to have an initial idea, and we use different colours for the points for different genders, so that 2 dimensions are enough for the plot.

Performing the regression, we obtain that the intercept is 35.53 and that the coefficient for being in Texas is approximately of 0.51 while the one for being a male is negative equal to -0.8989.



```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      35.5322     3.1193  11.391   <2e-16 ***
state_tot3TX      0.5162     1.3775   0.375    0.708
gender_tot3Male  -0.8989     3.1821  -0.282    0.778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.62 on 313 degrees of freedom
Multiple R-squared:  0.0006654, Adjusted R-squared:  -0.00572
F-statistic: 0.1042 on 2 and 313 DF,  p-value: 0.9011
```

Unfortunately all the p-values that we obtain for the regression are clearly too high to show a significance of the considered factors.

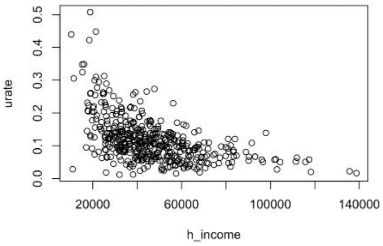
4.4 R5: nonlinear regression

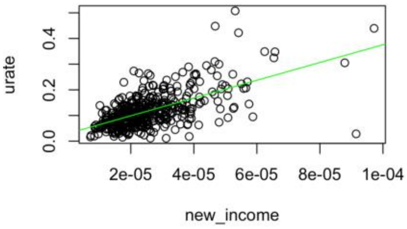
Looking at the scatter plot of the house income against the unemployment rate we notice that there could be a *y=1/x* relationship, where x represents the house income and y the urate, that better fits our data.

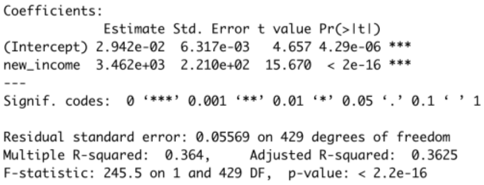
In fact, doing a linear regression:

x <- (1 / (h-income))

r6 <-lm (urate ~ x, data = police-killings



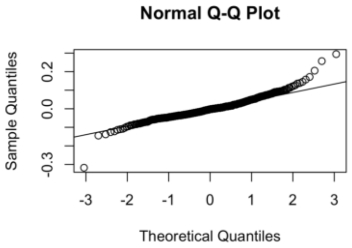


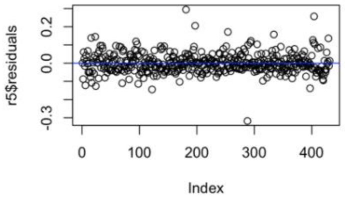


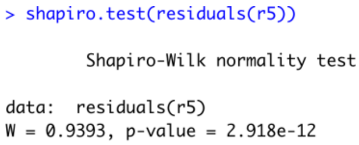
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.942e-02  6.317e-03   4.657 4.29e-06 ***
new_income  3.462e+03  2.210e+02  15.670  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05569 on 429 degrees of freedom
Multiple R-squared:  0.364,	Adjusted R-squared:  0.3625
F-statistic: 245.5 on 1 and 429 DF,  p-value: < 2.2e-16
```

we obtain an intercept equal to 2.942e-02 with and an estimate of the beta coefficient equal to 3462e+03 (as a side note the coefficient is so high because the scales of the explanatory variable x is completely different from the urate scale which can be only between 0 and 1 since it is a percentage). We therefore can be quite confident in both the estimate of the intercept and the coefficient as the p-values obtained are extremely low, hence concluding that indeed there is a non-linear regression between the house income and the unemployment rate which is well expressed by the curve y = 1 / x.

Before continuing, however, we must remember to check that the residuals are distributed as a normal with mean zero and constant variance. Plotting the residuals, it emerges that they have zero mean and by doing a QQ-plot we can also verify that the normality assumption can be reasonably assumed to be satisfied, despite the qq-line does not perfectly fits the data. The fact that the residuals does not come *exactly* from a gaussian is also confirmed by the Shapiro-Wilk normality test that gives a very small p-value hence the null hypothesis H0: “residuals come from a normal distribution” must be rejected.







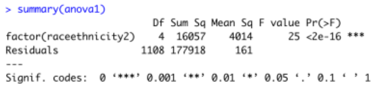
```
> shapiro.test(residuals(r5))

	Shapiro-Wilk normality test

data:  residuals(r5)
W = 0.9393, p-value = 2.918e-12
```

5. STATISTICAL TEST

5.1. *Anova 1*

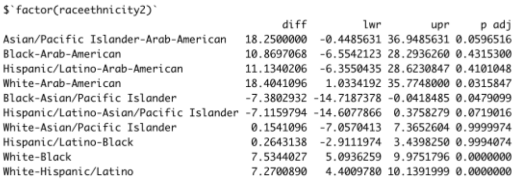
Our interest now shifts to verifying whether the victim's age of death could be influenced by his or her racial ethnicity. Since the racial ethnicity is a categorical variable, we run an ANOVA (*) but before running it we clean our data from the rows where the victim's age or his ethnicity were unknown and then we proceed. We have also decided for this specific task to eliminate the following ethnic groups: Asian/Pacific Islander, Native American and "Other" as the vast majority of victims did not fall into one of these categories and therefore to simplify the analysis we only focused on the following three ethnicity: white, black and Hispanic/Latino.

As grand mean we obtain 22 while the main effects given by the ethnicity are respectively the following: 10.86 for black people, 11.13 for Hispanic/Latino and 18.40 for White people. Due to the result of the F test, we can conclude that the factor *raceetnicity* has a statistical significance (p-value of <2e-16) in determining the age of death of the victim.



```
> summary(anova1)
                        Df Sum Sq Mean Sq F value Pr(>F)    
factor(raceethnicity2)    4  16057    4014      25 <2e-16 ***
Residuals              1108 177918     161                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Looking more carefully at the R output, we also note that the main effect given by the white ethnicity is bigger than both that of the black ethnic group and that of the Hispanic/Latino one, thus deducing that the white victims tend to be older than those of the others two ethnicity. To have an additional confirmation of this fact we use the Tukey command and in fact we see that White-Black is 7.53 and White-Hispanic/Latino is 7.27 and in both cases the difference has a 95% confidence interval which lies on the positive axis of the real line with a p-value very very close to zero, hence from the data at our disposal we can conclude that non-White people are younger when they die.



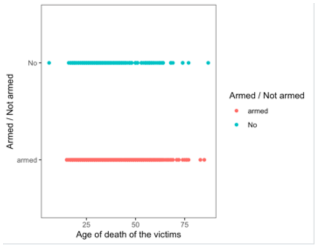
```
$`factor(raceethnicity2)`
                                             diff         lwr        upr     p adj
Asian/Pacific Islander-Arab-American    18.2500000  -0.4485631 36.9485631 0.0596516
Black-Arab-American                     10.8697068  -6.5542123 28.2936260 0.4315300
Hispanic/Latino-Arab-American           11.1340206  -6.3550435 28.6230847 0.4101048
White-Arab-American                     18.4041096   1.0334192 35.7748000 0.0315847
Black-Asian/Pacific Islander            -7.3802932 -14.7187378 -0.0418485 0.0479099
Hispanic/Latino-Asian/Pacific Islander -7.1159794 -14.6077866  0.3758279 0.0719016
White-Asian/Pacific Islander             0.1541096  -7.0570413  7.3652604 0.9999974
Hispanic/Latino-Black                    0.2643138  -2.9111974  3.4398250 0.9994074
White-Black                              7.5344027   5.0936259  9.9751796 0.0000000
White-Hispanic/Latino                    7.2700890   4.4009780 10.1391999 0.0000000
```

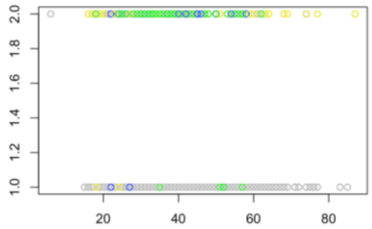
5.2 *Anova 2*

We also decided to perform an another ANOVA (*), to check how much the age of the victim can be predicted, knowing if the victim was armed and the cause of the death.

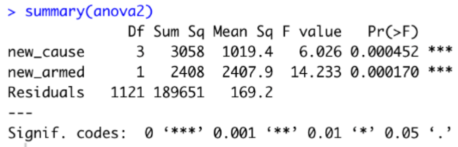
To do so these parameters are converted into factors, and in particular, we let the variable *armed* to be binary.

For the variable *cause* we have instead 4 possible cases.



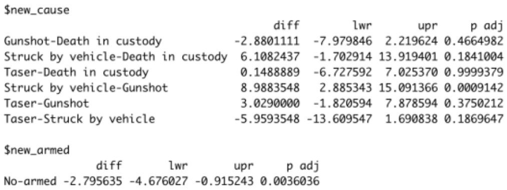


Running the ANOVA, we first notice that the p-values obtained give a very strong significance. Thus, we can move on and focus on the coefficients obtained: the grand mean is of 43 years approximately and the effects are:



```
> summary(anova2)
              Df Sum Sq Mean Sq F value   Pr(>F)    
new_cause      3   3058  1019.4   6.026 0.000452 ***
new_armed      1   2408  2407.9  14.233 0.000170 ***
Residuals   1121 189651   169.2                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

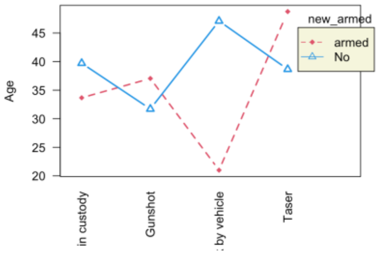
To observe the interactions effects, we can plot the graph and use the TuKey command. Although the results are not very interesting from a practical point of view for our research we still comment on them.



```
$new_cause
                                      diff        lwr       upr     p adj
Gunshot-Death in custody          -2.8801111  -7.979846  2.219624 0.4664982
Struck by vehicle-Death in custody 6.1082437  -1.702914 13.919401 0.1841004
Taser-Death in custody             0.1488889  -6.727592  7.025370 0.9999379
Struck by vehicle-Gunshot          8.9883548   2.885343 15.091366 0.0009142
Taser-Gunshot                      3.0290000  -1.820594  7.878594 0.3750212
Taser-Struck by vehicle           -5.9593548 -13.609547  1.690838 0.1869647

$new_armed
              diff       lwr       upr     p adj
No-armed -2.795635 -4.676027 -0.915243 0.0036036
```

Regarding the difference between the *cause* factors the highest one is given by the Struck by vehicle-Gunshot (diff: 8.98), as it can also be seen from the plot. Instead, the highest gap between the age of the armed and not armed case is attained under the cause of death of struck by vehicle, moving from a mean of 45 to a mean of 20.

In general from the plot we see that *armed* has an ambiguous behaviour: in the case of gunshot and taser caused death being armed is associated with an higher age of the victim with respect to the unarmed case. Instead the opposite happens in the cases of struck by vehicle and death in custody.

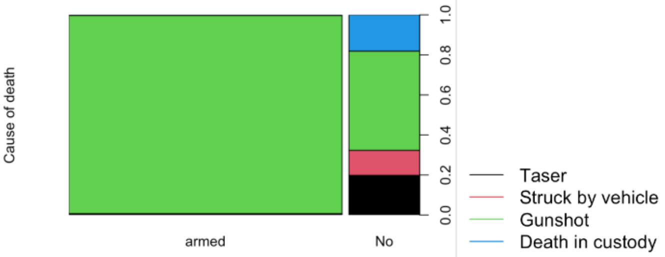


6. LOGISTIC REGRESSION

6.1. *logistic regression 1: cause and armed*

Finally, using a logistic regression (*), we want to verify if there is a relationship between the cause of death and whether the victim was armed or not as binary variable.

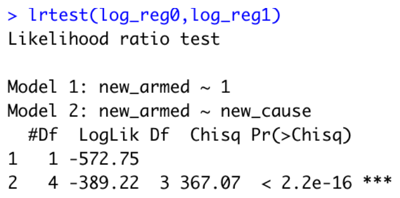
In fact, plotting an histogram, at first glance it appears that if the victim was armed then almost for sure he or she had been killed by a gunshot while, if the victim was not armed this was no longer the case; in fact not-armed victims had been killed by gunshot but also by taser, hit by a police vehicle or while they were kept in custody.



Running a logistic regression, we can see in fact that being killed by a gunshot is correlated with the fact that the victim was armed, used as the binary variable, since the *gunshot* categorical variable has a p-value equal to 1.12e-14.

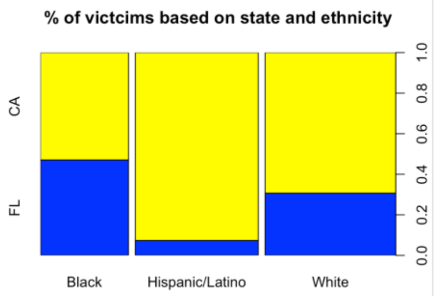
Then we perform the likelihood ratio test comparing the larger model with the *cause* factor added, to the smaller one without such a factor.

The output that we obtain gives a p-value of 2.2e-16, that is clearly small enough to reject the null hypothesis that the coefficient given by the cause is significantly different from zero.



```
> lrtest(log_reg0,log_reg1)
Likelihood ratio test

Model 1: new_armed ~ 1
Model 2: new_armed ~ new_cause
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -572.75
2   4 -389.22  3 367.07  < 2.2e-16 ***
```

6.2 logistic regression 2: state and ethnicity



Finally we have compared California and Florida, the two states with more deaths, and the race ethnicity composition of these two and tested if, through a logistic regression (*), in California a specific ethnicity was more affected by this phenomenon than in Florida and it emerged that Hispanic/Latino ethnicity was the only that seemed to play a role in establishing the probability of being killed between the two states. Since with respect to this variable the p-value is 1.73e-07 we can conclude that if you live in California and you are Hispanic/Latino you are more likelihood to be killed by the police with respect that if you live in Florida.

As a side note it is important to underline that this last result does not imply that there is casual relationship between the race-ethnicity and the state where you are killed since for instance, in California we could have more Hispanic/Latino victims because the percentage of Hispanic people in California is much higher in California with respect to Florida (see Census data on population estimates 2015) and this would *explain* why in California the probability of being killed by the police increase if you are Hispanic .

7. CONCLUSIONS

To conclude, retracing the points addressed previously, it emerges that the Black and Hispanic / Latino ethnic groups tend to be the most affected compared to their proportion in the total population; in fact, the states that present a minor number of deaths during the year are also the states that present a higher percentage of white people.

On the other hand, with respect to the economic context, it seems to emerge that the victims who lived in areas with a higher house income belonged to areas where the unemployment rate was lower and the level of education higher. By subsequently testing the hypothesis that the victims' mean house income was lower than the US mean house income, we were thus able to verify that on average the victims actually lived in poorer areas with a consequent higher level of unemployment and a lower level of education. A further analysis could therefore be made to see if the areas with a lower house income also have a higher level of criminality and from here one could therefore deduce a causal relationship between the crime rate and the number of police killings but unfortunately, we have not got to deepen this part.

Using linear regression, we also discarded the hypothesis that gender, unemployment rate or the state were related to the age of death but in any case none of these analysis can be considered conclusive since the residuals were not *perfectly* normally distributed.

By applying a non-linear transformation (1/x) to house income, on the other hand, we were able to verify with a significant p-value that there is a negative relation between house income and the unemployment rate.

Thanks to ANOVA we then saw that race-ethnicity relates to the age of death of the victim as white victims are older than the others. Also in this case it would have been interesting to investigate whether the fact that the white victims are older than those of the black and Hispanic /Latino people was determined by the fact that in the latter two ethnic groups juvenile crime is much more widespread and therefore it would be reasonable that on average non-white victims are younger but even in this case we have not been able to actually confirm this assumption.

Finally, through the two logistic regressions it emerges that being killed by a firearm increased the probability that the victim was also armed at the time of the killing and that not in all States the ethnic groups of the victims were equally affected; for example in California the probability of being killed increased for Hispanic / Latino people compared to Florida.

8. REFERENCES

- Guardian's database on police killings, 2015, (www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-map-us-police-killings) (referred to as (*)
- 2. \ Census American Site, (www.census.gov)
- 3. CENSUS + THE COUNTED ELABORATION:\\ https://github.com/fivethirtyeight/data/tree/master/police-killings

9. CODE

```
install.packages("lmtest")
install.packages("ggpubr")
library(lmtest)
install.packages("Rcmdr")

library(ggplot2)
library(scales)
library(ggpubr)


#insert june_killings.csv
police_killings=read.csv(file.choose(),sep=";",header = T)
attach(police_killings)
names(police_killings)
View(police_killings)


#insert annual_killings.csv
tot_killings=read.csv(file.choose(),header=T, sep=";",col.names = c( "null" ,"names_tot","age_tot","gender_tot","raceethnicity_tot","month_tot","day_tot","year_tot","streetaddress_tot","city_tot","state_tot","cause_tot","lawenforcementagency_tot","armed_tot"))
attach(tot_killings)
names(tot_killings)
View(tot_killings)

# step 0:  cleaning the data set
tot_killings<-tot_killings[complete.cases(tot_killings), ]
police_killings<-police_killings[complete.cases(police_killings),]
tot_killings<-na.omit(tot_killings)
police_killings<-na.omit(police_killings)


#1: GRAPHS:
# as we are working with a large amount of data, it is necessary to have a preliminary
# overall idea on the data we are working with. It is thus useful to plot some graphs
# to start familiarize with the results and to gain some insights on possible tests and
# regressions that could be later studied.

# 1.1 PIE CHART OF GENDERS (tot annual)
x<-as.data.frame(prop.table(table(gender_tot)))
labels_GENDER=paste0(names(table(gender_tot))," = ",c(100 *x[2]) ,"%")
x <- x[-c(3),] #deleting misleading row
q<-round(100*x[2])
q
bp<- ggplot(x, aes(x="", y=Freq, fill=gender_tot))+
  geom_bar(width = 1, stat = "identity")
bp

graph2 = pie(x$Freq, labels = c("Female 5%","Male 95%"), col = c("pink", "lightblue"))

#1.2 MONTHS BAR PLOT
z<-table(month_tot)
z<-z[c(5,4,8,1,9,7,6,2,12,11,10,3)]
z<-data.frame(z)

p<-ggplot(data=z, aes(x=z[,1], y=z[,2])) +
  geom_bar(stat ="identity", fill = "lightblue")+
  theme_bw() +
  labs (y = "Number of killings", x = "Months") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
p

#1.2 CHART OF ETHNICITY
w<-as.data.frame(prop.table(table(raceethnicity_tot)))
labels_ethnicity=paste0(names(table(raceethnicity_tot))," = ",c(100 *w[2]) ,"%")
y<-round(100*w[2])
y

b<- ggplot(w, aes(x="", y=Freq, fill=names(table( raceethnicity_tot))))+
  geom_bar(width = 1, stat = "identity") +
  labs (x = "", y = "", title = "Ethnicity of the victims") +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  guides(fill=guide_legend("Etnicity"))

 b


w2<-c(0.001,0.001,0.13,0.185,0.018,0.06, 0.605)
c<- ggplot(as.data.frame(w2), aes(x="", y=w2, fill=names(table(raceethnicity_tot))))+
  geom_bar(width = 1, stat = "identity") +
  labs (x = "", y = "", title = "Race etnicity of the US population") +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  guides(fill=guide_legend("Etnicity"))
c

final = ggarrange(b, c + font("x.text", size = 10), #To stick the two graph together
                           ncol = 2, nrow = 1)
final


#1.3HISTOGRAM AGE
plot(table(age_tot),xlab= "age", ylab= "occurrencies")
mean(as.numeric(age_tot))
plot(h_income, urate)#1
cor(h_income, urate)#2
plot(h_income, college, xlab="Income", ylab="College")
cor(h_income,college)

#1.4 CAUSE OF DEATHS HISTOGRAM

barplot(table(cause_tot),names.arg=names(table(cause_tot)), las=2, ylab = "Number of killings")

#1.5 GRAPHS ON ETHNICITIES OF MOST HIT STATES
tbl_etn <- matrix(c(  0.365,0.065,0.394,0.176,    0.421, 0.129,0.397,0.053,    0.926,0.014,0.02,0.04,       0.837,0.034,0.041, 0.088,                    0.601, 0.134,0.185,0.08),ncol=4,nrow=4, byrow=FALSE)
colnames(tbl_etn) <-c("CA","TX","VT", "ND")
rownames(tbl_etn)<-c("WHITE","BLACK","HISPANIC", "OTHERS")
mycol<-c("white", "black", "orange","grey")
tbl_etn
barplot(tbl_etn, col= mycol, las=2, ylab = "Ethnic division of population")
legend(x = "right", legend = rownames(tbl_etn), col = mycol, fill=mycol)


#2 STATISTICAL TESTS : INCOME
data <- police_killings$h_income
```

```
mean(data)
median(data)
plot(ecdf(h_income), main = "Empirical distribution function", ylab = "")

qqnorm(h_income, pch = 1, frame = FALSE)
qqline(h_income, col = "blue", lwd = 2)
#2nd we run a  Kolmogorov-Smirnov test and since the p-value is less than 0.05
#we have to  reject H_0..hence h_income in not "normally distributed"

alpha = .05
ks_test2<-ks.test(data, "pnorm", mean = mean(data), sd = sd(data))
ks_test2$p.value< alpha

#3rd we test the hp that mu=mu_0 vs mu!=mu_0 using a t-test where mu_0 is the mean
#house income in US during 2015 and we reject H_0 hence the h_income of the victims
#differs significantly from the average US house income
n = length(h_income)
alpha = 0.05
mu_0 =56516
t_stat = sqrt(n) * (mean(police_killings$h_income) - mu_0) / sd(police_killings$h_income)
c_alpha = qt(p = 1-alpha/2, df = n-1)
abs(t_stat) > c_alpha
p_value = 2 * min(pt(t_stat, n-1), pt(t_stat, n-1, lower.tail = F))
p_value < alpha



#3 REGRESSIONS:
#3.1 : TRICKY REGRESSION: UNEMPLOYEMENT ~SHARE OF BLACK POPULATION
#not covered in the paper because the discoveries are not very deep

attach(police_killings)
share_new<- as.numeric(share_black)/100
plot(share_new~urate)
cor(share_new, urate)
r1<-lm(urate~share_new, data=police_killings)
r1
summary(r1)
abline(a=r1$coefficients[1],b=r1$coefficients[2],col="green")

plot(r1$residuals, ylab= "residuals", xlab = "")
abline(a=0, b=0, col="blue")
qqnorm(residuals(r1))
qqline(residuals(r1))
shapiro.test(residuals(r1))


#3.2 DUMMY VARIABLES/MUTLTIPLE REGRESSION : AGE~URATE+ GENDER

plot(urate,age, col = "white" , xlab = "unemployment rate")
points(urate[gender=="Male"],age[gender=="Male"], col="blue", pch="x")
points(urate[gender=="Female"],age[gender=="Female"], col="red", pch="o")
police_killings$age <- as.numeric(police_killings$age)
r2<-lm(age~urate+gender, data=police_killings)
r2
summary(r2)
r2$coefficients
abline(a=(r2$coefficients[1]+r2$coefficients[3]),b=r2$coefficients[2], col="blue")
abline(a=r2$coefficients[1],b=r2$coefficients[2], col="red")

#R3:here we decrease the dimension, using only the urate
r3<-lm(age~urate, data=police_killings)
r3
summary(r3)
r3$coefficients
cor(urate, as.numeric(age))


# MODEL SELECTION IN ADDITION
AIC(r2)
AIC(r3)

#R4 MULTIPLE REGRESSION: AGE~STATE+GENDER
attach(tot_killings)
take3<-which((state_tot=="CA"|state_tot=="TX") & age_tot!="Unknown")
state_tot3<-state_tot[take3]
gender_tot3<-gender_tot[take3]
age_tot3<-age_tot[take3]

df = data.frame(age_tot3, state_tot3) #Creating the dataframe
#----
plot3 = ggplot(data = df, aes(x = age_tot3, y = state_tot3, color = state_tot3 )) +
  geom_point() +
  labs( x = "Age of death of the victims", y = "State", color = "State" ) +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

plot3
#----
plot(age_tot3, factor(state_tot3))
points(age_tot3[gender_tot3=="Male"],factor(state_tot3[gender_tot3=="Male"]), col="lightblue", pch="x")
points(age_tot3[gender_tot3=="Female"],factor(state_tot3[gender_tot3=="Female"]), col="pink", pch="o")

#here at level one is reported the ages of victims in  California, and at level 2
#the ones of Texas. The color of the points stands for the gender

r4<-lm(age_tot3~state_tot3+gender_tot3)
r4
summary(r4)
plot(r4$residuals)
abline(a=0, b=0, col="blue")
qqnorm(residuals(r4))
qqline(residuals(r4))
shapiro.test(residuals(r4))


#3.5 : NON LINEAR REGRESSION: URATE~NEW INCOME
plot(h_income, urate)
new_income<-(1/(h_income))
plot(new_income, urate)
r5<-lm(urate~new_income)
r5
summary(r5)
abline(a=r5$coefficients[1],b=r5$coefficients[2], col="green")

plot(r5$residuals)
abline(a=0, b=0, col="blue")
qqnorm(residuals(r5))
qqline(residuals(r5))
shapiro.test(residuals(r5))


#4: ANOVA
#4.1: anova on race and age
del=which(raceethnicity_tot=="Unknown"|age_tot=="Unknown"| raceethnicity_tot==" Asian/Pacific Islander" | raceethnicity_tot=="Native American" | raceethnicity_tot=="Other")
del=which(raceethnicity_tot=="Unknown"|age_tot=="Unknown"| raceethnicity_tot==" Asian/Pacific Islander"| raceethnicity_tot=="Native American"| raceethnicity_tot=="Other")
raceethnicity2=factor(raceethnicity_tot[-del])
raceethnicity2
age1=as.numeric(age_tot[-del])
mean(age1)
l<-lm(age1~factor(raceethnicity2))
```

```
anova1=aov(age1 ~factor(raceethnicity2))
anova1$coefficients
summary(anova1)
TukeyHSD(anova1)



#ANOVA 2
#----- variables intermezzo
new_armed<-c()
for (x in armed_tot) {
  if (x != "No"){
    new_armed<-c(new_armed,"armed")
  }
  else{
    new_armed<-c(new_armed,"No")
  }
}
del1=which(cause_tot =="Other")
new_cause<-factor(cause_tot[-del1])
new_armed<-factor(new_armed[-del1])
new_age<-as.numeric(age_tot[-del1])
#----------
df2 = data.frame(new_age, new_armed)

plot4 = ggplot(data = df2, aes(x = new_age, y = new_armed, color = new_armed )) +
  geom_point() +
  labs( x = "Age of death of the victims", y = "Armed / Not armed", color = "Armed / Not armed" ) +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

plot4

plot(new_age,new_armed, xlab="age", ylab="armed/not armed")
points(new_age[new_cause=="Gunshot"],new_armed[new_cause=="Gunshot"], col="grey")
points(new_age[new_cause=="Death in custody"],new_armed[new_cause=="Death in custody"],col="blue")
points(new_age[new_cause=="Struck by vehicle"],new_armed[new_cause=="Struck by vehicle"],col="yellow")
points(new_age[new_cause=="Taser"],new_armed[new_cause=="Taser"],col="green")
# here armed is level 1, not armed is  level 2

anova2<-aov(new_age~new_cause+new_armed)
anova2
anova2$coefficients
summary(anova2)
TukeyHSD(anova2)
interaction.plot(new_cause, new_armed, new_age, ylab = "Age", xlab = "", type="b", col=c(2,4),
                 leg.bty="o", leg.bg="beige", las=2, lwd=2, pch=c(18,24,22))



#5: USE LOGISTIC REGRESSION TO STUDY CATEGORICAL VARIABLES

#5.1 LOGISTIC REGRESSION: ARMED~CAUSE

df5 = data.frame(new_armed, new_cause)
ggplot(data = df5, aes(x = new_armed, y = new_cause))


plot(new_armed,new_cause, las=1, las=2, ylab = "Cause of death", xlab = "", legend = "", col = c(1, 2,3,4))
legend(x = "topleft", legend = c("Taser","Struck by vehicle","Gunshot","Death in custody"), col = c(1, 2, 3, 4), lty = c(1, 1, 1, 1))
logistic_regression1<- glm(new_armed ~new_cause,family=binomial(link='logit'),data=police_killings)
summary(logistic_regression1)

log_reg1<- glm(new_armed ~ new_cause,family=binomial(link='logit'),data=police_killings)
log_reg0<- glm(new_armed ~1,family=binomial(link='logit'),data=police_killings)
lrtest(log_reg0,log_reg1)

#5.2LOGISTIC REGRESSION: STATE~RACEETHNICITY
new_ethnicity<-factor(subset(raceethnicity_tot,(state_tot=="CA"|state_tot=="FL" )&(raceethnicity_tot=="White"|raceethnicity_tot=="Black"|raceethnicity_tot=="Hispanic/Latino")))
new_state<-factor(subset(state_tot, (state_tot=="CA"|state_tot=="FL")&(raceethnicity_tot=="White"|raceethnicity_tot=="Black"|raceethnicity_tot=="Hispanic/Latino")))
new_state<-na.omit(new_state)
new_ethnicity<-na.omit(new_ethnicity)
plot(new_ethnicity, new_state, ylab = "", xlab = "" , col = c("blue", "yellow"), main =  "% of victims based on state and ethnicity")

logistic_regression2<- glm(new_state ~new_ethnicity,family=binomial(link='logit'),data=police_killings)
summary(logistic_regression2)

log_reg1b<- glm(new_state ~ new_ethnicity,family=binomial(link='logit'),data=police_killings)
log_reg0b<- glm(new_state ~1,family=binomial(link='logit'),data=police_killings)
lrtest(log_reg0b,log_reg1b)
```