Does Improving Diversity Among Police Officers Reduce Racial Profiling

November 16th, 2017

### Revisions on Part I

On August of 2014, an African American male, Michael Brown at only eighteen years of age was killed by a police officer after Brown was accused of strong-arm robbing a convenience store. Michael Brown was confirmed as being unarmed, but was still fatally shot after the officer claimed that Brown had allegedly attacked him in order to try and take control of his firearm, until it was fired. However, the officer shot a total of twelve rounds in the whole altercation and at the time he was the only witness. The officer that shot Brown was not charged with murder and was released after the U.S. Department of Justice concluded that the officer shot Brown in self defense (New York Times, 2017). Also, there have been many cases where police officers are accused of racially profiling African Americans, Hispanic and other minorities. The incident that motivated us to take on this topic were the killing of four Hispanic males in Salinas, CA by police in only the span of five months, in the year of 2014. One of our main interests is to identify any type of unfairness between the police officers and minorities. According to Nancy Krieger, a professor of social epidemiology at Harvard T.H. Chan School of Public Health, “Black men, compared to white men, were from five to 19 times at greater risk of a law enforcement-related death over the past 50 years”. There is a significant disparity in the risk of death when comparing a minority and a white individual, giving us some evidence to why such topic is important and analyze this issue with our own statistical knowledge. The data we will analyze does not contain information about people getting killed by police. We will only try to provide evidence to issues regarding racial profiling from traffic stop information. Therefore, the first question we want to analyze is if police officers regardless of ethnicity target more Hispanics over whites. In the study, “Black and Hispanic Men Perceived to Be Large Are at Increased Risk for Police Frisk, Search, and Force” they look at a similar scenario where the results showed that indeed Black and Hispanic suspects were at a greater risk of being frisked or searched compared to whites. Next, we want to determine if the result of the stop for a minority is less severe on average when the police officer is a minority. In recent years, research explains that about six percent more of the stops were between a white officer and black citizen over a white officer and white citizen being stopped(Criminology, 2012). Finally, we would like to see the results of being stopped leading to being searched is higher for African Americans and Hispanics than for white citizens. In a 2011 report, the Leadership Conference on Civil Rights found evidence of widespread racial profiling, showing that African Americans and Hispanics are disproportionately likely to be stopped and searched by police, even though they’re less likely to be found possessing contraband or committing a criminal act. For example, in Illinois, Hispanic and black drivers were twice as likely to be searched after being stopped by police compared to white drivers, but the surprising thing was that white drivers were twice as likely to have contraband in their possession (Washington Post, 2014). ###Partial Methods

The traffic stop data was downloaded from Data.Gov. The Records from all 79885 reported traffic stops was for a large Southeastern city in 2016. There is no information on how the data was collected. For all of our questions we will use multinomial regression or logistic regression. The variables are defined in the following way: Variables/Names Month 1-12 (Month) Reason for Stop (RsnStop) Officer Race (OffRace) Officer Male (OffMale) Officer Years Service (OffYrsSrv) Driver Race (DrvRace) Driver Hispanic (DrvHisp) Driver Male (DrvMale) Driver Age (DrvAge) Search Vehicle (SrchVhcl) Result of Stop (RsltStop) Levels for categorical variables Reason for Stop: 1=CheckPoint, 2=DWI, 3=Investigation, 4=Other, 5=Safe Movement, 6=SeatBelt, 7=Speeding, 8=StopLight/Sign, 9=VehicleMovement, 10=VehicleRegistry Officer race: Blank=NotSpecified, 1=AmericanIndian/NativeAlaska/Hawaii, 2=Asian/PacificIslander, 3=Black/AfricanAmerican, 4=Hispanic/Latino, 5=White Driver Race: 1=Asian, 2=Black, 3=NativeAmerican,4=Other/Unknown,5=White Result of Stop: 1=NoActionTaken, 2=VerbalWarning,3=WrittenWarning, 4=Citation, 5=Arrest Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot. ###Partial results

In the traffic stop data set from a large southeastern city of the United States, we found that the mean age of the drivers was about 36 years old. Also, we could observe that the mean number of service years for the police officers in this data set was around 11 years. As we continued to explore the data we found that the proportion of drivers who are males is about 58 percent while the percent of female drivers was about 42 percent. Our first question focuses on if Hispanics are targeted more than whites regardless of the gender of the police officer. In the data set there are 7578 drivers who were identified as Hispanics. We now explored more specific summary statistics for each question. Our first question is about observing how severe is the result of a stop for a driver who is a minority when the police officer is also a minority. Since our variables are mostly categorical we decided to create tables. According to Figure 1A in the appendix, we can observe a table with crucial information about a drivers race and a police officers race. For example, the total number of times that an Asian driver got stopped by a white officer is 1072 times. On the other hand, when the driver was black we can observe that the number of times stopped by a white officer actually is drastically larger with 30993 traffic stops. Another great representation of the data was through a proportions table where we can observe the proportion of traffic stop by an officer’s race and the drivers race. Our second question focuses on if Hispanics are targeted more than whites regardless of the gender of the police officer. In the data set there are 7578 drivers who were identified as Hispanic. In figure 2a, the table summarizes the distribution of Hispanics among the rest of the ethnicities. Now from this information we observe that 5305 traffic stops happened when the driver was Hispanic and the officer was white. The proportion of Hispanic drivers stopped by the different officer race’s can be observed in figure 2b.We also conducted a multinomial regression model for each of the first two questions and the results for the models will be interpreted in the next part of the project. For our third part of the project, we would like to see the results of being stopped leading to being searched is higher for minorities like African Americans and Hispanics than for other citizens like whites? For our summary statistics we compare the result of being stopped versus the binary outcome of being searched or not in a table. Then we also compare the binary event of being searched versus the drivers race and calculate the proportions. The proportions are as follows asian drivers that were searched was 0.0116, black drivers that were searched was 0.0576, native american drivers that were searched was 0.0322, and white drivers that were searched was 0.0206. Not suprisingly, we see that the proportion of black drivers being searched was the highest and the lowest proportions searched were when the drivers were asian and white, respectively. However, this is only from this sample and we need more test to see if there is a statistically significant difference between these races. We also conducted a simple logistic regression model with a vehicle being searched being the qualitative response variable and the drivers race being the qualitative response variable. We find that without adjusting for confounding variables, that the races asian, black and white are statistically significant predictors of the response, but the race titled as other/unknown is not statistically significant predictor. In the next part we will adjust this model to include the explanatory variables we believe could have predictive power over our response like Reason for Stop, Officers Race, Officer Male, Officer Years Service, Driver Male and the Drivers Age.

### Partial Appendix

General Summary Statistics

library(readr)  
trafficstop <- read\_csv("~/Desktop/ucsc/trafficstop.csv")

## Parsed with column specification:  
## cols(  
## Month = col\_integer(),  
## RsnStop = col\_integer(),  
## OffRace = col\_integer(),  
## OffMale = col\_integer(),  
## OffYrsSrv = col\_integer(),  
## DrvRace = col\_integer(),  
## DrvHisp = col\_integer(),  
## DrvMale = col\_integer(),  
## DrvAge = col\_integer(),  
## SrchVhcl = col\_integer(),  
## RsltStop = col\_integer()  
## )

attach(trafficstop)  
mean(DrvAge)

## [1] 35.88215

mean(OffYrsSrv)

## [1] 11.1789

men<-sum(DrvMale==1)  
men\_prop<-men/length(DrvAge)  
men\_prop

## [1] 0.5795153

women<-sum(DrvMale==0)  
women\_prop<-women/length(DrvAge)  
########### summary Statistics question 1  
  
#install.packages("expss")  
#install.packages("matrixStats")  
library(expss)  
library(matrixStats)  
trafficstop1 = apply\_labels(trafficstop,  
 OffRace = "Officers Race",  
 OffRace = c("American Indian/Native Alaska/Hawaii" = 1,  
 "Asian/Pacific Islander" = 2  
 ,"Black/African American"=3,"Hispanic/Latino"=4,"White"=5  
   
   
   
 ),  
 DrvRace = "Drivers Race",  
 DrvRace = c("Asian" = 1,  
 "Black"=2,  
 "Native American"=3,  
 "Other"=4,  
 "White"=5  
   
 )  
   
)  
#Figure 1A  
#cro(trafficstop1$DrvRace,trafficstop1$OffRace)###### Frequency table for first question  
  
  
#Figure 1B  
#cro\_cpct(trafficstop1$OffRace, list(total(), trafficstop1$DrvRace))### proportions  
  
  
  
########################### Summary Stats for question 2  
  
#Based on that we know that Hispanics are a minority in the United States, we will take the lowest sum of the categorical variable DrvHispanic and define it as Hispanic.   
sum(trafficstop$DrvHisp==1)

## [1] 7578

sum(trafficstop$DrvHisp==2)

## [1] 72306

### Now we have some certainty that the categorical value 1 of the variable DrvHispanic is for the drivers who were hispanic.   
  
  
  
trafficstop2 = apply\_labels(trafficstop,  
 OffRace = "Officers Race",  
 OffRace = c("American Indian/Native Alaska/Hawaii" = 1,  
 "Asian/Pacific Islander" = 2  
 ,"Black/African American"=3,"Hispanic/Latino"=4,"White"=5  
   
   
   
 ),  
 DrvHisp = "Hispanic Drivers",  
 DrvHisp = c("Hispanic" = 1,  
 "Other"=2  
   
 )  
   
)  
#cro(trafficstop2$DrvHisp,trafficstop2$OffRace)###### Frequency table for Second question  
  
  
#cro\_cpct(trafficstop2$OffRace, list(total(), trafficstop2$DrvHisp))### proportions  
  
########################### Summary Stats for question 3  
  
  
#Comparing the results of the stop versus if the person got searched  
table(SrchVhcl, factor(RsltStop))

##   
## SrchVhcl 1 2 3 4 5  
## 0 1846 38852 3370 32063 545  
## 1 68 1217 25 863 1035

#Comparing the drivers race versus if the person got searched  
table(SrchVhcl, factor(DrvRace))

##   
## SrchVhcl 1 2 3 4 5  
## 0 1444 40492 60 2389 32291  
## 1 17 2478 2 33 678

#Calculate what proportions of drivers got searched for each race  
prop.asian.searched <- 17 / (1444 + 17)  
prop.asian.searched

## [1] 0.01163587

prop.black.searched <- 2478 / (2478 + 40492)  
prop.black.searched

## [1] 0.05766814

prop.native.searched <- 2 / 62  
prop.native.searched

## [1] 0.03225806

prop.other.searched <- 33 / ( 33 + 2389)  
prop.white.searched <- 678 / (678 + 32291)  
prop.white.searched

## [1] 0.02056477

############################# Modeling question 1  
library(readr)  
#install.packages("VGAM")  
library(VGAM)

## Loading required package: stats4

## Loading required package: splines

offrace <- factor(trafficstop$OffRace)  
contrasts(offrace)=contr.treatment(levels(offrace),base=5)  
contrasts(offrace)

## 1 2 3 4  
## 1 1 0 0 0  
## 2 0 1 0 0  
## 3 0 0 1 0  
## 4 0 0 0 1  
## 5 0 0 0 0

q\_Hispanic<-vglm((factor(DrvHisp)~factor(offrace)),data=trafficstop,family = multinomial())  
summary(q\_Hispanic)

##   
## Call:  
## vglm(formula = (factor(DrvHisp) ~ factor(offrace)), family = multinomial(),   
## data = trafficstop)  
##   
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max  
## log(mu[,1]/mu[,2]) -0.4096 -0.3181 -0.3181 -0.3091 3.324  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.40243 0.23350 -10.289 < 2e-16 \*\*\*  
## factor(offrace)2 0.46360 0.23794 1.948 0.05137 .   
## factor(offrace)3 0.05425 0.23543 0.230 0.81777   
## factor(offrace)4 0.61734 0.23930 2.580 0.00989 \*\*   
## factor(offrace)5 0.11185 0.23394 0.478 0.63258   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Number of linear predictors: 1   
##   
## Name of linear predictor: log(mu[,1]/mu[,2])   
##   
## Residual deviance: 49523 on 79158 degrees of freedom  
##   
## Log-likelihood: -24761.5 on 79158 degrees of freedom  
##   
## Number of iterations: 5   
##   
## Reference group is level 2 of the response

############################ Modeling Question 2  
library(readr)  
#install.packages("VGAM")  
library(VGAM)  
  
#Alana's code  
offrace <- factor(trafficstop$OffRace)  
contrasts(offrace)=contr.treatment(levels(offrace),base=5)  
contrasts(offrace)

## 1 2 3 4  
## 1 1 0 0 0  
## 2 0 1 0 0  
## 3 0 0 1 0  
## 4 0 0 0 1  
## 5 0 0 0 0

q12<-vglm((factor(RsltStop)~factor(offrace)),data=trafficstop,family = multinomial())  
summary(q12)

##   
## Call:  
## vglm(formula = (factor(RsltStop) ~ factor(offrace)), family = multinomial(),   
## data = trafficstop)  
##   
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max  
## log(mu[,1]/mu[,5]) -3.467 -0.08817 -0.08817 -0.07245 10.4571  
## log(mu[,2]/mu[,5]) -4.833 -0.62684 -0.50939 0.82966 0.9945  
## log(mu[,3]/mu[,5]) -3.512 -0.14652 -0.12234 -0.11468 15.0451  
## log(mu[,4]/mu[,5]) -4.937 -0.46989 -0.46989 1.07313 1.4254  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -1.09861 0.81650 -1.346 0.17846   
## (Intercept):2 3.34990 0.41535 8.065 7.31e-16 \*\*\*  
## (Intercept):3 -1.79176 1.08012 -1.659 0.09715 .   
## (Intercept):4 2.31911 0.42786 5.420 5.95e-08 \*\*\*  
## factor(offrace)2:1 0.59896 0.83452 0.718 0.47292   
## factor(offrace)2:2 -0.24870 0.42925 -0.579 0.56233   
## factor(offrace)2:3 0.34765 1.10703 0.314 0.75350   
## factor(offrace)2:4 0.89164 0.44130 2.020 0.04334 \*   
## factor(offrace)3:1 1.91071 0.82114 2.327 0.01997 \*   
## factor(offrace)3:2 0.07041 0.42184 0.167 0.86744   
## factor(offrace)3:3 2.68462 1.08355 2.478 0.01323 \*   
## factor(offrace)3:4 1.28667 0.43413 2.964 0.00304 \*\*   
## factor(offrace)4:1 1.23969 0.83365 1.487 0.13700   
## factor(offrace)4:2 -0.12861 0.43390 -0.296 0.76693   
## factor(offrace)4:3 1.85060 1.09367 1.692 0.09063 .   
## factor(offrace)4:4 0.49339 0.44623 1.106 0.26887   
## factor(offrace)5:1 1.20109 0.81746 1.469 0.14175   
## factor(offrace)5:2 -0.14536 0.41638 -0.349 0.72700   
## factor(offrace)5:3 2.63387 1.08067 2.437 0.01480 \*   
## factor(offrace)5:4 0.59723 0.42887 1.393 0.16375   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Number of linear predictors: 4   
##   
## Names of linear predictors:   
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])  
##   
## Residual deviance: 159271.4 on 316632 degrees of freedom  
##   
## Log-likelihood: -79635.7 on 316632 degrees of freedom  
##   
## Number of iterations: 7   
##   
## Reference group is level 5 of the response

#### Addding Service years to the model to make it more significant  
q13<-vglm((factor(RsltStop)~factor(offrace)+OffYrsSrv),data=trafficstop,family = multinomial())  
summary(q13)

##   
## Call:  
## vglm(formula = (factor(RsltStop) ~ factor(offrace) + OffYrsSrv),   
## family = multinomial(), data = trafficstop)  
##   
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max  
## log(mu[,1]/mu[,5]) -6.078 -0.09667 -0.08695 -0.06759 10.457  
## log(mu[,2]/mu[,5]) -7.294 -0.64223 -0.35167 0.78690 1.593  
## log(mu[,3]/mu[,5]) -6.701 -0.15774 -0.09028 -0.05541 24.102  
## log(mu[,4]/mu[,5]) -7.777 -0.53175 -0.38753 0.91492 1.685  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -1.634195 0.817768 -1.998 0.045678 \*   
## (Intercept):2 3.086158 0.416606 7.408 1.28e-13 \*\*\*  
## (Intercept):3 -3.242752 1.082180 -2.997 0.002731 \*\*   
## (Intercept):4 1.610884 0.429792 3.748 0.000178 \*\*\*  
## factor(offrace)2:1 0.464212 0.834836 0.556 0.578176   
## factor(offrace)2:2 -0.312781 0.429432 -0.728 0.466394   
## factor(offrace)2:3 -0.023596 1.108420 -0.021 0.983016   
## factor(offrace)2:4 0.710944 0.442202 1.608 0.107893   
## factor(offrace)3:1 1.893196 0.821391 2.305 0.021174 \*   
## factor(offrace)3:2 0.072264 0.421961 0.171 0.864021   
## factor(offrace)3:3 2.481667 1.084910 2.287 0.022170 \*   
## factor(offrace)3:4 1.246618 0.434947 2.866 0.004155 \*\*   
## factor(offrace)4:1 1.408941 0.834007 1.689 0.091150 .   
## factor(offrace)4:2 -0.043005 0.434139 -0.099 0.921092   
## factor(offrace)4:3 2.257777 1.095122 2.062 0.039239 \*   
## factor(offrace)4:4 0.712932 0.447173 1.594 0.110867   
## factor(offrace)5:1 1.138689 0.817721 1.393 0.163766   
## factor(offrace)5:2 -0.162393 0.416497 -0.390 0.696608   
## factor(offrace)5:3 2.246211 1.082039 2.076 0.037903 \*   
## factor(offrace)5:4 0.491911 0.429680 1.145 0.252279   
## OffYrsSrv:1 0.060402 0.004701 12.849 < 2e-16 \*\*\*  
## OffYrsSrv:2 0.031135 0.003802 8.190 2.62e-16 \*\*\*  
## OffYrsSrv:3 0.143434 0.004311 33.271 < 2e-16 \*\*\*  
## OffYrsSrv:4 0.077712 0.003810 20.396 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Number of linear predictors: 4   
##   
## Names of linear predictors:   
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])  
##   
## Residual deviance: 154751.8 on 316628 degrees of freedom  
##   
## Log-likelihood: -77375.92 on 316628 degrees of freedom  
##   
## Number of iterations: 7   
##   
## Reference group is level 5 of the response

############################ Modeling Question 3  
  
## Create reduced model, simple logistic binomial, to test wheter Drivers Race has significant predictive power over the response Search Vehicle.  
SearchRaceModel <- glm(SrchVhcl ~ factor(DrvRace), family = "binomial")  
summary(SearchRaceModel)

##   
## Call:  
## glm(formula = SrchVhcl ~ factor(DrvRace), family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.3447 -0.3447 -0.3447 -0.2039 2.9845   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.4420 0.2436 -18.234 < 2e-16 \*\*\*  
## factor(DrvRace)2 1.6483 0.2445 6.742 1.56e-11 \*\*\*  
## factor(DrvRace)3 1.0408 0.7590 1.371 0.170   
## factor(DrvRace)4 0.1598 0.3000 0.533 0.594   
## factor(DrvRace)5 0.5786 0.2467 2.345 0.019 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26912 on 79883 degrees of freedom  
## Residual deviance: 26111 on 79879 degrees of freedom  
## AIC: 26121  
##   
## Number of Fisher Scoring iterations: 6