Final Term Project - Racial Disparities at Police Stops in the US

Xibei Chen

21st December 2021

Introduction

The aim of this project is to take a closer look into racial disparities at police stops in the US, exploring how other variables might effect the association between probability of getting searched and driver's race. Racial discrimination has always been a topic in the US. Especially after the murder of George Floyd in May 2020, Black Lives Matter movement gained much more international attention. The data set that is used for this project is from The Stanford Open Policing Project. There are already some findings about the racial disparities regarding stop rates, search decisions, etc. In this project I am specifically interested in the disparities between black and white drivers, and I will particularly focus on how other variables such as driver's gender and age, officer's race and gender would effect the association between the probability of getting searched and driver's race at police stops, with the hope that I might be able to find something new than what has already been done.

Data

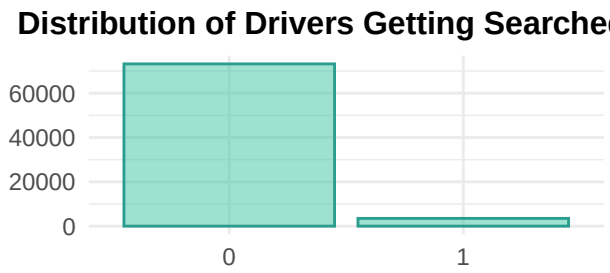
To achieve the aim of this project, I specifically picked the data set for Louisville, where data for all the other control variables that I am interested in are also available. The data set includes data of all the traffic stops from 2015-01-01 to 2018-01-28 in Louisville, KY. I did some data cleaning and munging to filter out all the NA values, focus only on sample with drivers either black or white, categorize officers as white and non-white and consider both *frisk performed* and *search conducted* as *get searched*.

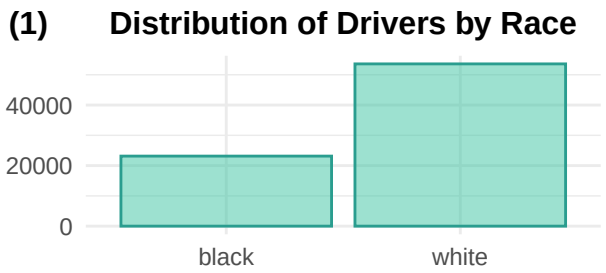
Table 1: Descriptive statistics

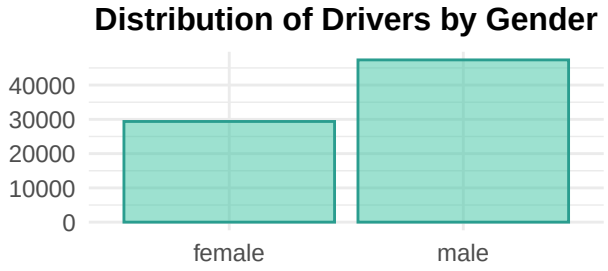
	Mean	Median	SD	Min	Max	P05	P95
Probability of Getting Searched	0.05	0.00	0.21	0.00	1.00	0.00	0.00
Race of Drivers	0.30	0.00	0.46	0.00	1.00	0.00	1.00
Gender of Drviers	0.38	0.00	0.49	0.00	1.00	0.00	1.00
Age of Drivers	36.75	34.00	13.96	11.00	96.00	19.00	63.00
Race of Officers	0.23	0.00	0.42	0.00	1.00	0.00	1.00
Gender of Officers	0.03	0.00	0.16	0.00	1.00	0.00	0.00

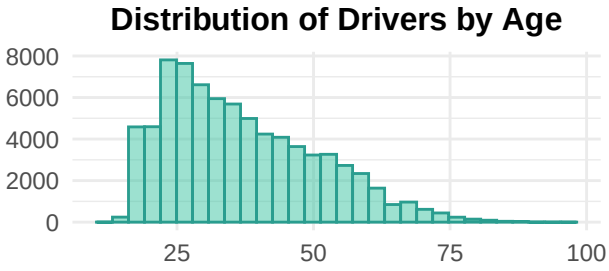
The number of observations in my sample is 76715 for all the key variables. From the descriptive statistics we can also see that for drivers 30% are black whereas 70% are white, 38% are female whereas 62% are male, the mean age is around 37 years old, for police officers 23% are non-white whereas 77% are white, 3% are female whereas 97% are male.

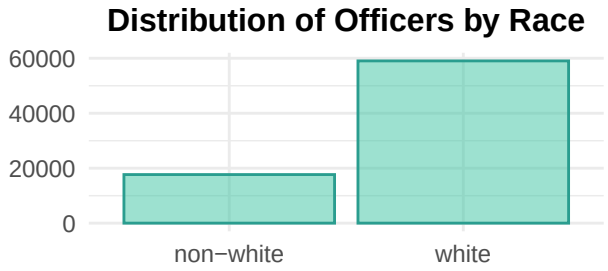
- Check distributions of key variables

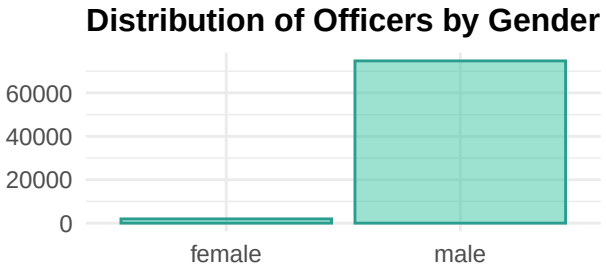












Models and Interpretation

The pattern of association between y and the only one continuous variable subject_age (see graph in Appendix) seems close to be linear, so there is no need to use splines or polynomials. Therefore, I start building regression models.

Table 2: Associations between probability of getting searched and driver race

	(1)	(2)	(3)	(4)	(5)
Intercept	0.0351***	0.0491***	0.0423***	0.0982***	0.0911***
	(0.0008)	(0.0010)	(0.0011)	(0.0023)	(0.0024)
black driver	0.0355***	0.0357***	0.0584***	0.0324***	0.0584***
	(0.0019)	(0.0019)	(0.0028)	(0.0018)	(0.0052)
female driver		-0.0367***	-0.0189***	-0.0386***	-0.0387***
		(0.0014)	(0.0015)	(0.0014)	(0.0014)
black driver x female driver			-0.0588***		
			(0.0034)		
subject_age				-0.0013***	-0.0011***
				(4.53e-5)	(4.7e-5)
black driver x subject_age					-0.0007***
					(0.0001)
Observations	76,715	76,715	76,715	76,715	76,715
R2	0.00606	0.01334	0.01729	0.02066	0.02111

Table 3: Associations between probability of getting searched and driver race

	(6)	(7)	(8)	(9)
Intercept	0.0523***	0.0515***	0.0499***	0.0500***
	(0.0011)	(0.0012)	(0.0011)	(0.0011)
black driver	0.0358***	0.0384***	0.0345***	0.0341***
	(0.0019)	(0.0022)	(0.0018)	(0.0018)
female driver	-0.0365***	-0.0365***	-0.0367***	-0.0367***
	(0.0014)	(0.0014)	(0.0014)	(0.0014)
non-white officer	-0.0142***	-0.0108***	-0.0138***	-0.0138***
	(0.0016)	(0.0017)	(0.0016)	(0.0016)
black driver x non-white officer		-0.0112**		
		(0.0040)		
female officer			0.1065***	0.1008***
			(0.0081)	(0.0098)
black driver x female officer				0.0143
				(0.0169)
Observations	76,715	76,715	76,715	76,715
R2	0.01416	0.01427	0.02062	0.02065

- from Model (2) we can infer that when controlling on driver's race, we expect female drivers to be around 3.7% less likely to get searched at police stops than male drivers on average; from Model (3) we can infer from the interaction term between driver's race and gender that when driver's gender is female, we expect the disadvantage for black driver regarding probability of getting searched to be lower by around 5.9%.
- from Model (4) we can infer that there is no statistically significant association between probability of getting searched and driver's age; from Model (5) we can also infer that basically driver's age does not effect much the association between probability of getting searched and drivers race.
- from Model (6) we can infer that when controlling on drivers race and gender, we expect the probability for drivers to get searched to be around 1.4% lower when the officer is non-white than when the officer is white; from Model (7) we can also infer from the interaction term between driver's race and officer's race that when officer is non-white, we expect the disadvantage for black driver regarding probability of getting searched to be lower by around 1.1%.
- from Model (8) we can infer that when controlling on driver's race and gender and officer's race, we expect the probability for drivers to get searched to be around 10.7% higher when officer is female than when officer is male; from Model (9) even though the coefficient for interaction term between driver's race and officer's gender is not zero, however the standard error is quite high, and the 95% confidence interval includes zero, it might be the result of very low relative frequency of female officers. Therefore we cannot conclude that officer being female has any effect on the disadvantage for black driver regarding probability of getting searched.

My preferred model is Model(8), as driver's age has almost no association with probability of getting searched.

probability of getting searched = 0.05 + 0.03 $(race = black)$ $+\delta Z$

where Z are standing for the controls, which includes controlling for driver's gender, officer's race and officer's gender. Interpret the coefficients:

- Interpret alpha: when the officer is male and white, and the driver is male and white, we expect the probability of getting searched to be around 5% on average;
- Interpret beta: when controlling on driver's gender, officer's gender and race, black drivers are expected to be around 3% more likely to get searched than white drivers on average.

Besides, based on the heteroskedastic robust standard errors, these results are statistically different from zero. To show that, I have run a two-sided hypothesis test:

$$H_0 := \beta_1 = 0$$

$$H_A := \beta_1 \neq 0$$

I have the t-statistic rounded as 18.7 and the p-value rounded as 0, which confirms my conclusion.

Modelling Probabilities (see graph and summary table details in Appendix)

Since the dependent variable is probability, I considered multiple ways of conducting modelling probabilities, such as LPM, Logit and Probit. Logit is slightly better than LPM and probit regarding goodness of fit as its Brier score is the smallest. There is also no distinguishable difference between probability models according to pseudo R2 and log-loss, plotted predicted probabilities, and calibration curves. Therefore, for uncovering general patterns, LPM would work fine.

External Validity (Robustness Check)

Table 4: Associations between probability of getting searched and driver race

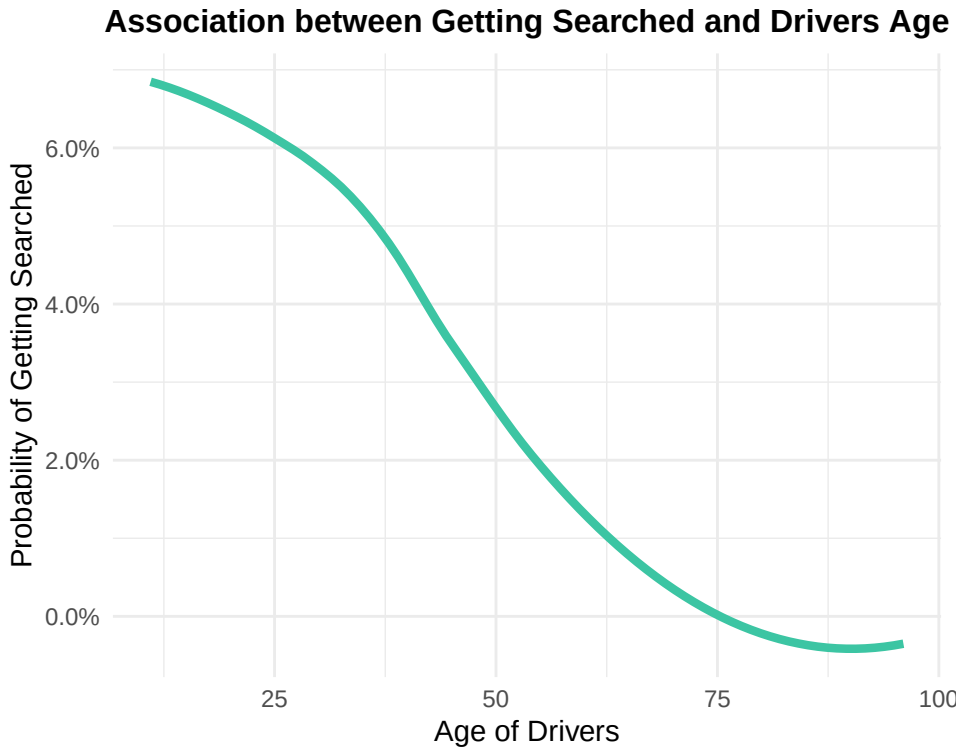
	Louisville, 2015 Jan-2018 Jan	WA Statewide, 2009 Jan-2015 Dec
Intercept	0.0499***	0.0327***
	(0.0011)	(0.0001)
black driver	0.0345***	0.0306***
	(0.0018)	(0.0005)
female driver	-0.0367***	-0.0106***
	(0.0014)	(0.0002)
non-white officer	-0.0138***	-0.0038***
	(0.0016)	(0.0003)
female officer	0.1065***	-1.59e-5
	(0.0081)	(0.0003)
Observations	76,715	4,922,765
R2	0.02062	0.00263

The second data set I used includes data of all the traffic stops with drivers being either black or white from 2009-01-01 to 2015-12-31 in Washington Statewide. The slope coefficient for driver's race is similar to our model for Louisville previously. This suggests that for other time intervals and other regions in the US, the external validity of the model is quite high, we might expect similar slope coefficient for driver's race.

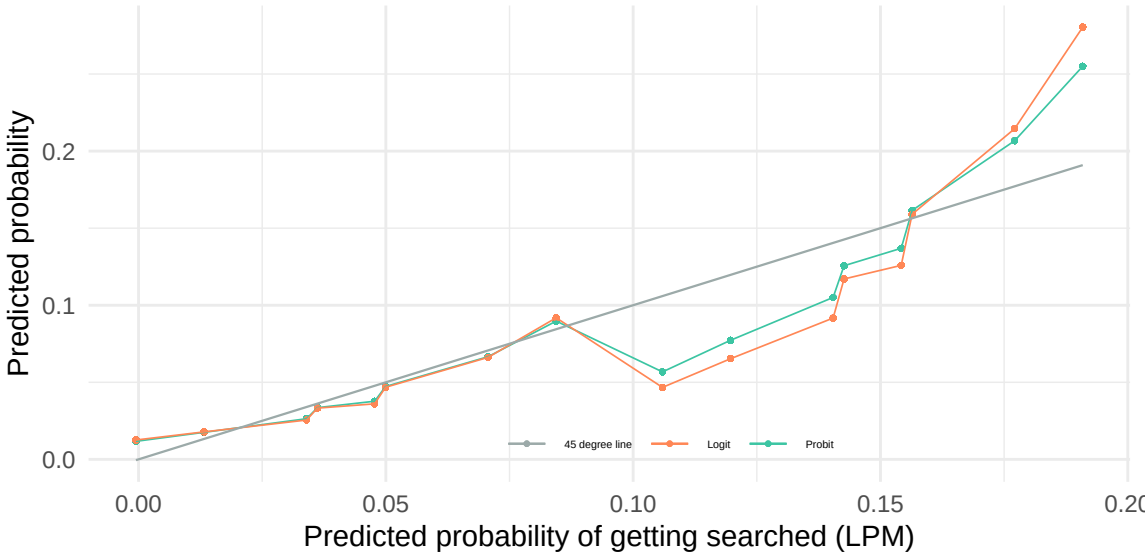
Conclusion

In general, unsurprisingly when controlling on driver's gender, officer's gender and race, black drivers are expected to be around 3% more likely to get searched than white drivers on average. What's new for me is, we expect the disadvantage for black drivers regarding probability of getting searched to be lower when the driver is female instead of male and when officer is non-white instead of white, whereas driver's age and officer's gender have no statistically significant effect on the matter. Last but not least, within the US, it is assumed that the external validity of the preferred regression model to be quite high, we might expect similar slope coefficient for driver's race for other regions in the US.

Appendix



Check pattern of association between y and the only one continuous variable subject_age



Plot predicted probabilities

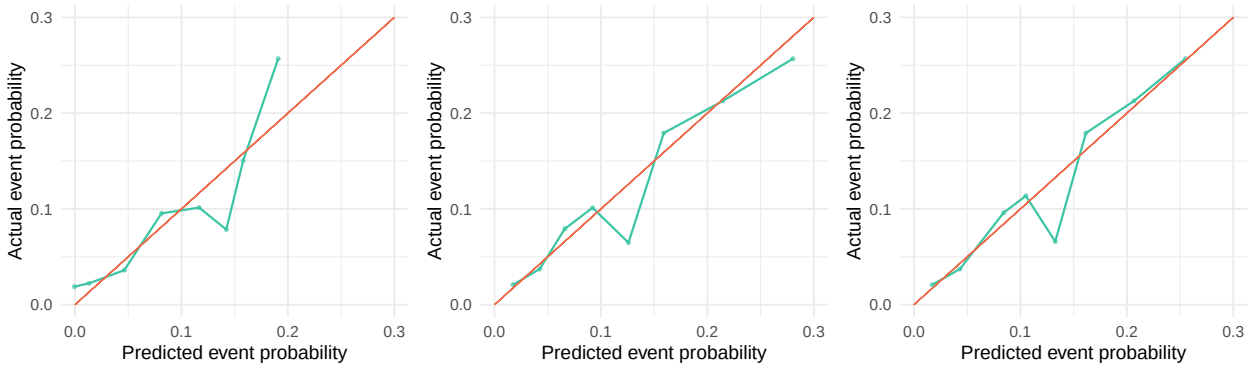
	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	0.050**	−3.015**		−1.672**	
	(0.001)	(0.027)		(0.012)	
subject_b	0.034**	0.723**	0.034**	0.329**	0.033**
	(0.002)	(0.035)	(0.002)	(0.017)	(0.002)
subject_female	−0.037**	−0.996**	−0.037**	−0.435**	−0.036**
	(0.001)	(0.043)	(0.001)	(0.019)	(0.001)
officer_nw	−0.014**	−0.355**	−0.014**	−0.159**	−0.014**
	(0.002)	(0.046)	(0.002)	(0.021)	(0.002)
officer_female	0.106**	1.349**	0.101**	0.684**	0.103**
	(0.008)	(0.067)	(0.008)	(0.036)	(0.008)
Num.Obs.	76 715	76 715	76 715	76 715	76 715

* p < 0.05, ** p < 0.01

Summary table of coefficients of different models (1.LPM, 2.Logit, 3.Logit Marginal, 4.Probit, 5.Probit Marginal)

	lpm	logit	probit
Dependent Var.:	get_searched	get_searched	get_searched
(Intercept)	0.0499*** (0.0011)	-3.015*** (0.0266)	-1.672*** (0.0121)
subject_b	0.0345*** (0.0018)	0.7230*** (0.0351)	0.3294*** (0.0165)
subject_female	-0.0367*** (0.0014)	-0.9956*** (0.0435)	-0.4354*** (0.0187)
officer_nw	-0.0138*** (0.0016)	-0.3554*** (0.0459)	-0.1592*** (0.0206)
officer_female	0.1065*** (0.0081)	1.349*** (0.0668)	0.6838*** (0.0361)
Family	OLS	Logit	Probit
S.E. type	Heteroskedast.-rob.	IID	IID
R2	0.02062	–	–
Brier score	0.04284	0.04269	0.04270
Pseudo R2	-0.07147	0.05015	0.04940
log-loss	NaN	-0.17676	-0.17690

Summary table of statistics of goodness of fit (LPM, Logit, Probit)



Calibration curves (LPM, Logit, Probit)