Pratishta Yerakala

py2211

QMSS - Data Analysis Independent Project

Does the time of a of a complaint being made and race of the suspect inform the severity of the offense?

Introduction

This proposed research project seeks to answer the following question: "Does the time of day that a complaint is made to the New York Police Department (NYPD) inform the severity of the offense?" The "broken window" or order maintenance policing (OMP) model states that community policing of smaller crimes and disorder can deter larger-scale crimes from taking place.[1] This sociological theory has been exercised in many urban areas including NYC and has manifested in some concrete policing practices such as Stop and Frisk. Since its end in 2013, there has not only been a wave of research dedicated to the efficacy of Stop and Frisk but also and the OMP model itself. In 2017, the city of New York has agreed to pay over $75 million in settlement over a federal class-action lawsuit that accused officers of issuing hundreds of thousands of summonses without legal justification.[2] The same year, the Urban Institute did a study that has shown that there is a growing decline in trust between those from disadvantaged communities and law enforcement.[3]

Though the crime rate of NYC was going down more and more every year, even if by a fraction, there are still not only crimes still being committed, but complaints that for crimes by civilians. Violent crimes such as murder, assault, rape / sexual assault and robbery are more likely to occur at night. Despite urban spaces such as NYC having well lit street lamps for the majority of the streets, there is still significantly lower visibility at night which can be advantageous for those attempting to commit more severe crimes. As a result, not only are

---

[1] Sousa, W. H. (2010). Paying attention to minor offenses: order maintenance policing in practice. *Police Practice and Research*, 10(1), 45-59.

[2] "New York City to Pay Up to $75 Million Over Dismissed Summonses" The New York Times. Accessed November 27, 2019. https://www.nytimes.com/2017/01/23/nyregion/new-york-city-agrees-to-settlement-over-summonses-that-were-dismissed.html.

[3] Nancy La Vigne, Jocelyn Fontaine, and Anamika Dwivedi. 2017. "How Do People in High-Crime, Low-Income Communities View the Police?" The Urban Institute, Justice Policy Center.

civilians told to be more careful at night but are often advised not to even go outside in certain locations. Marginalized and vulnerable populations especially, such as women and those living in high crime neighborhoods, are more at risk for being victim to more severe crimes. Additionally, because of black and brown communities have a history of being overpoliced and are still under a more strict surveillance than whites or Asians, complaints reports may also have black persons as the suspect. In the paper "An Analysis of the New York City Police Department's “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias," Gelman, Fagan, and Kiss explain that black and Hispanic people are convicted or stopped at a higher rate for more severe crimes than white people. Given the brief background, there are two primary reasons why this proposed research project is important: 1) if it were true that there is a time at which more severe crimes are complained about by the public, policy makers and law enforcement can use that to more effectively allocate the proper resources at a given time and 2) it may give grounds for a wider social or urban phenomena that could explain or curb crimes taking place at a specific time of the day.

Data and Methods

The data used in this project comes from the NYPD CompStat program which provides crime statistics, specifically on summonses, complaints, and arrests made with datasets going as far back as 2006. I'm using the complaints dataset made available on NYC OpenData which comes in the form of a CSV file. According to NYC OpenData, "This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning. Each record represents a criminal complaint in NYC and includes information about the type of crime, the location and time of enforcement. In addition, information related to victim and suspect demographics is also included. This data can be used by the public to explore the nature of criminal activity." This data has been collected and distributed through the CompStat program which is a management philosophy in addition to being a tracking and documenting system employed by the NYPD to have crime statistics be publically available.

The variables I'm using are the following:

Dependent Variable

(1) **LAW_CAT_CD**: This is the level of offense ranging from violation, misdemeanor and felony. Violations, also known as infractions, are the lowest level of offense that are usually punishable by fines but not typically jail time. Examples include harassment in the second degree, loitering, etc. Misdemeanors rank higher than violations but lower than felonies; they are for criminal offenses that can result in up to a year of jail time. Examples include petit larceny, offenses involving fraud, possession of dangerous drugs, etc. Felonies are the highest level of offense in this dataset and can result in anywhere from a year in prison up to life in prison without parole. Examples include grand larceny, assault, burglary. It's important to note that these codes provide a rough estimation for the severity of a crime but often in the judicial system, offense levels change depending on the situation. For example, for a repeat offender who has several misdemeanors, the next offense can be taken as a felony. This variable is the dependent variable which I have re-coded such that the lowest level offense, violation, is 1; the middle level offense, misdemeanor, is 2; the highest level offense, felony is 3. This numeric re-coding allows for an easier interpretation of the severity of the crime instead of a categorical view.

Independent Variable

(1) **CMPLNT_FR_DT**: This is the exact time of occurrence for the reported event. This variable has data in the form of timestamps, specifically in the 24 hour clock. An example of one of the values is 23:00:00 which represents 11PM. This is the independent variable which I have re-coded. I first stripped the string value by keeping only the first two characters of the string. So in our example, 23:00:00 becomes 23. Then I cast that string value into a float. Then I divided up the 24 hours into "chunks" to represent the times of day into separate entities. The hours between 5am and 7am are the earliest morning hours so I have assigned it 1. The hours between 7am and 9am are morning so I have assigned it 2. The hours between 9am to 12pm are late morning to early afternoon so I have assigned it 3. The hours between 12pm and 3pm are afternoon so I have assigned it 4. The hours between 3pm and 6pm are late afternoon so I have assigned it 5.

The hours between 6pm and 8pm are evening, so I have assigned it 6. The hours between 8pm and 10pm are late evening and early night, so I have assigned it 7. The hours between 10pm and 12am are night, so I have assigned it 8. The hours between 12am and 5am are late night, so I have assigned it 9. This divides the day into periods that could define most people's working hours, commutes, and sleeping schedules. I stripped the minutes and did not round up in order to preserve the hour and keep consistency with the data. The seconds for all of the times in this dataset were 0 and not relevant.

(2) **SUSP_RACE**: This variable is the suspect's race description. The variable is categorical and has the following possible values: American Indian / Alaskan Native, Asian / Pacific Islander, Black, Black Hispanic, Other, White, Unknown, White Hispanic. Because I am choosing to focus on black suspects only, I re-coded the values such that non-Black suspects are 0, and Black suspects are 1.

Controlled Variables

(1) **BORO_NM**: This is the variable that says which NYC borough that the incident took place in. I only looked at complaint data in Brooklyn in order to isolate and control for a borough. This would allow for a more granular further analysis should I choose to do so involving precincts and geographic data. Additionally, because Manhattan is a tourist heavy area and Staten Island is a much more suburban area, the complaints from those two boroughs could be vastly different in nature. I chose to isolate Brooklyn because I believe it contains both touristy areas and suburban-like neighborhoods within it that could provide for a diverse set of complaints and criminal activity.

(2) **SUSP_SEX**: This variable is the suspect's sex description.

(3) **OFNS_DESC**: This variable is a description of the offense. It is categorical and has 64 unique values.

(4) **VIC_SEX**: This variable is the victim's sex description.

(5) **VIC_RACE**: This variable is the victim's race description. The variable is categorical and has the following possible values: American Indian / Alaskan Native, Asian / Pacific Islander, Black, Black Hispanic, Other, White, Unknown, White Hispanic.

Summary of Data

Descriptive Statistics Table for Variables

Table 1

	CMPLNT_FR_TM	LAW_CAT_CD	OFNS_DESC	SUSP_SEX	SUSP_RACE	VIC_SEX	VIC_RACE
count	1.935011e+06	1.935028e+06	1930917	950307	983580	1934921	1934921
unique	NaN	NaN	64	3	8	4	8
top	NaN	NaN	PETIT LARCENY	M	BLACK	F	BLACK
freq	NaN	NaN	297824	608151	469229	785631	654766
mean	5.510063e+00	2.196449e+00	NaN	NaN	NaN	NaN	NaN
std	2.199801e+00	6.423568e-01	NaN	NaN	NaN	NaN	NaN
min	1.000000e+00	1.000000e+00	NaN	NaN	NaN	NaN	NaN
25%	4.000000e+00	2.000000e+00	NaN	NaN	NaN	NaN	NaN
50%	5.000000e+00	2.000000e+00	NaN	NaN	NaN	NaN	NaN
75%	7.000000e+00	3.000000e+00	NaN	NaN	NaN	NaN	NaN
max	9.000000e+00	3.000000e+00	NaN	NaN	NaN	NaN	NaN

Table 2

col_0	count
LAW_CAT_CD	
1.0	246489
2.0	1061916
3.0	626623

On average, the complaint time variable is 5.5 meaning that the average complaint is made around late afternoon (3pm to 6pm, assigned 5) to evening time (6pm to 8pm, assigned 6). The average level of offense for these complaints is 2.2 meaning that most offenses are of the misdemeanor category on average. The standard deviation is 0.6 which is somewhat high since the range of the data is 2, but even so misdemeanors seem to be the most popular level of offense at about 1 million occurrences, compared to felonies which are 0.6 million and violations which are 0.2 million (see Table 2). The most common offense is *petit larceny* which is in line with the most common offense level since petit larceny automatically is coded into a misdemeanor.

The suspect and victim demographics are interesting because the in terms of sex, males are the most commonly supsects of these complaints while women are most commonly the ones reporting or the victims of these complaints. However, both the suspect and victim are most commonly black.

Hypotheses

The later that a complaint is made in the day, the more severe the level of offense of the incident would be.

If the suspect's race description is black, the more likely it is that the level of offense of the incident would be more severe.

Models

Multiple Linear Probability Model

I will first run a multiple linear probability model such that I will look at the severity of the level of offense (1 to 3) based on the independent variables time of day of the complaint and the suspect's race description. The two prediction models are:

$$SO = \beta + \beta 1 * CT + \beta 2 * SR + \beta 3 * SS + \beta 4 * VR + \beta 5 * VS$$

SO addresses both hypothesis 1 and hypothesis 2 where the severity of the level of offense can be predicted by the complaint time, and severity of the level of offense can be predicted by the race of the suspect respectively.

Table 3

OLS Regression Results

Dep. Variable:	LAW_CAT_CD	R-squared:	0.051
Model:	OLS	Adj. R-squared:	0.051
Method:	Least Squares	F-statistic:	7444.
Date:	Mon, 16 Dec 2019	Prob (F-statistic):	0.00
Time:	22:21:23	Log-Likelihood:	-7.6704e+05
No. Observations:	691114	AIC:	1.534e+06
Df Residuals:	691108	BIC:	1.534e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.9531	0.003	681.039	0.000	1.947	1.959
CMPLNT_FR_TM	0.0259	0.000	64.559	0.000	0.025	0.027
SUSP_RACE	0.1451	0.002	68.464	0.000	0.141	0.149
VIC_RACE	-0.0862	0.002	-40.715	0.000	-0.090	-0.082
VIC_SEX	-0.1697	0.002	-92.803	0.000	-0.173	-0.166
SUSP_SEX	-0.2558	0.002	-126.143	0.000	-0.260	-0.252

Omnibus:	336614.093	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36469.640
Skew:	0.024	Prob(JB):	0.00
Kurtosis:	1.876	Cond. No.	20.9

Discussion of Multiple Linear Probability Model

The model establishes that the later the time of a complaint the more severe the level of offense. After controlling for victim race, victim sex, and suspect sex, the coefficient for complaint time is 0.0259 which indicates that for every chunk of time later in the day that a complaint is made, the severity of the level of offense goes up by 0.0259 severity points. This is statistically significant with a p-value of less than 0.001.

The model also establishes that as the description of the suspect is black, the more severe the level of offense. After controlling for victim race, victim sex, and suspect sex, the coefficient for suspect's race is 0.1451 which indicates that as the suspect's race is black, the level of offense goes up by 0.1451 severity points. This is statistically significant with a p-value of less than 0.001.

The r-squared is low, however, showing that only 5.1% of the variance in complaint time and suspect race are due to the prediction model. This model is insufficient because though the severity of the level of offense is shown to increase with the said independent variables, there is not a strong enough explanation as to how that likelihood increases. Using a logistic model would more explicitly and transparently demonstrate whether the level of offense is severe or not rather than showing exactly how severe which can be ambiguous or abstract concept to fully interpret (i.e. How different exactly is 0.0259 points in severity? This is not sufficient enough to show whether the changes in the independent variables will result in jail time or not).

Binary Logistic Model

I re-coded the level of offense variable such that 0 represents only violations while 1 represents both misdemeanors as well as felonies, as shown in Table 4. I re-coded this in such a way because violations do not tend to result in jail time. On the other hand, misdemeanors and felonies both tend to result in different times of jail time but still result in jail time nonetheless which makes them more severe of a level of offense compared to violations. I use the following equation for performing the binary logistic model:

$$\text{logit(SO)} = \beta + \beta 1 * CT + \beta 2 * SR + \beta 3 * SS + \beta 4 * VR + \beta 5 * VS + \mu$$

Table 4

col_0	count
LAW_CAT_CD	
0.0	246489
1.0	1688539

Table 5

```
Optimization terminated successfully.
         Current function value: 0.601770
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:             LAW_CAT_CD   No. Observations:               691114
Model:                          Logit   Df Residuals:                   691108
Method:                           MLE   Df Model:                            5
Date:                Mon, 16 Dec 2019   Pseudo R-squ.:                 0.02064
Time:                        22:47:15   Log-Likelihood:            -4.1589e+05
converged:                       True   LL-Null:                   -4.2466e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        0.7544      0.009     88.021      0.000       0.738       0.771
CMPLNT_FR_TM     0.0683      0.001     56.418      0.000       0.066       0.071
SUSP_RACE        0.2377      0.006     37.164      0.000       0.225       0.250
VIC_RACE        -0.1323      0.006    -20.677      0.000      -0.145      -0.120
VIC_SEX         -0.3394      0.006    -60.784      0.000      -0.350      -0.329
SUSP_SEX        -0.5108      0.006    -87.663      0.000      -0.522      -0.499
==============================================================================
```

Discussion of Binary Logistic Model

As expected again, the positive logit for complaint time of 0.0683 indicates that the odds are it will be a higher severity of offense if the time is later. The positive logit for race of 0.2377 also indicates that the odds are it will be a higher severity of offense if the suspect's race is black (from 0 as non-black to 1 as black). The p-values for both coefficients are less than 0.001 indicating that the values are statistically significant.

Conclusion

Both models supported the hypothesis in having positive coefficients for both complaint time and suspect race. In the first model, the coefficient for complaint time was 0.02 but in the second binary logistic model, the coefficient 0.06 which is slightly higher. In the first model, the coefficient for suspect's race was 0.1 whereas the second model had a coefficient of 0.2. Because all the values were statistically significant, the second model showed a stronger correlation to support the hypotheses. This model can clearly show with some small level of confidence that controlling for victim's race and sex description and the suspect's sex description, that the later the time of the complaing, the higher likelihood there is that the suspect for that crime would recieve jail time according to the level of offense; simiarly, as the suspect's race description switches to black, the likelihood of that suspect recieving jail time for that level of offense also increases.

It is important to note that though the level of offense is documented in this dataset, it is still only documentation and does not indicate any real results of jail time or judicial punishment. It is only a proxy. Additionally, for this model, the race of the suspect and the victim have been re-coded in a binary way to include only those described as Black but not those described as Black Hispanic. The latter demographic of people are still black and it would be interesting for future models to take that additional race category into consideration.

Ideally, I would have liked to use more control variables such as income level for both victim and suspect, repeat offender status of suspect, neighborhood of the incident, the crime rate of that neighborhood, etc. But due to the limited quality of the dataset provided by the NYPD and restricted time, it was not possible to superset other possibly existing data. Additionally, re-coding the offense description into most common occurrences of offenses and categorizing them into severity could also be a more granular indicator of the severity of the crime committed. All of these factors heavily influence how complaint data along with other crime statistics are collected and analyzed. The models I have used for this project are not rigorous enough to indicate causation but rather information or influence in the resulting level of offense. Further research including the aforementioned controlled variables and possibly even panel data for

repeat offenders or complaint victims might shed more light on how severe the level of offense is recorded. It is important that these modifications are considered and employed for something as sensitive and important as the law enforcement and preemptive policing especially when considering that data collected and data analyzed can skew based on what data is available in the first place (i.e. NYPD could consider including more detailed and thorough data on complaints, summonses and arrests such that these other variables and factors could be considered within the dataset).