NCVS Report

Rebecca Yang

December 21, 2020

Abstract

A large proportion of crime in the United States goes unreported to police; this is especially true for cases of sexual assault. In this paper, I investigate the features of incidents that make them more or less likely to be reported to the police, as well as the effect of the #MeToo social movement on the reporting of sexual crime. Initial results from a logistic regression model found that age of the victime, the type of crime, and injuries were significant, but did not find a difference for sexual crimes that occurred before and after the start of the #MeToo movement.

Keywords: National Crime Victimization Survey, logistic regression, #MeToo, crime reporting

Introduction

If a tree falls in a forest, but doesn't make a sound, did it even fall? The same question can be posed for crimes — if a crime happens, but no one reports it to the police, did it even happen? The answer is no, at least in the eyes of the authorities. Incidents can go unreported for a myriad of reasons, such as inconvenience of the reporting process, lack of faith in the police's ability to help, apathy towards the crime, or not wanting to revisit a traumatic experience, to name a few. As such, there exists a gap in the data between crime as it is reported and crime as it occurs in the United States. In this paper, I investigated the factors that contribute to the likelihood that a crime is reported, at the levels of both the incidents and the victims. I was also interested to see whether or not the #MeToo movement had an effect on the propensity to report a sexual assault. In October 2017, the MeToo hashtag trended on Twitter, starting a social movement that jump started the national conversation on the the toxic culture of sexual harassment and assault. In addition to thousands sharing their personal experiences online, several high profile men were unseated from power as a result of the courage of women to publicly speak out against them.

To do this, I fitted a multilevel logistic regression model using data from the 2016-2019 iterations of the National Crime Victimization Survey (NCVS). The NCVS is a national survey that asks respondents about all the victimizations they experienced, not just the ones that were reported to the police. The response variable was whether the crime was reported to the police or not, and predictors were the age of the victim, the type of crime (i.e. theft, burglary, etc.), whether it was attempted or completed, whether it resulted in injury, and whether it occurred before or after the start of the #MeToo movement. It was found and #MeToo was not found to have had a significant effect on sexual crimes.

That being said, these results should be taken with a grain of salt, as further testing suggested the lack of important predictors in the model. These could either be information that is collected by the survey that I failed to consider, or information that the survey fails to capture. For example, respondents may feel similar hesitation towards reporting their experiences to the NCVS that they feel towards the police. In addition, reporting to the police is a very high stakes decision, and thereby a very high standard by which to measure the effect of the movement.

In the following sections, I describe the survey data, the model fitting process, and results. Then, I

discuss the results and conclude with weaknesses and next steps.

Data

The National Crime Victimization Survey (NCVS) is survey carried out by the Bureau of Justice Statistics in the United States, and is one of the two main avenues through which crime in the US is measured (Justice Statistics 2020d, 2020c, 2020b: @ncvs2016). The other is the Uniform Crime Reporting system (it is currently in the process of being phased out and replaced by the National Incident-Based Reporting System), which consists of crimes as they are reported by police departments across the country. The main distinction between this and the NCVS is that the NCVS focuses on victimizations; this includes incidents that were not reported to the police. As such, the NCVS provides better insight on the true levels of crime, and provides more details on each incident. It also aims to collect information about the repercussions that victims of crime experience and to track changes in crime over time. I used the NCVS from 2016-2019, since I wanted to compare between time periods.

That being said, the same reasons for not reporting a crime to the police can extend to reporting a crime to the NCVS. While the survey aims to foster a safe space for victims, some may still be averse to sharing their personal experiences. Similarly, people who do not want to disclose incidents to the police may also not be inclined respond to the survey. It should also be noted that these reports are not subject to verification, and are solely based on the respondent's recollection.

The NCVS employs a multistage cluster sampling design, as well as a rotating panel design. First, 1,987 primary sampling units (PSUs) consisting of at least 7,500 people are defined geographically. Then, strata are formed from these PSUs. If a PSU is large enough, it forms its own strata; these PSUs are included in the sample for certain. Otherwise, 2-3 similar PSUs are grouped together to form strata. Similarity is measured according to the census, American Community Survey, and local crime data. For these, strata, a single PSU is selected, with PSUs with larger populations having higher probabilities of selection. Then, systematic random sampling is employed to select a sample of households or group quarters. This means there is a rule as to how subjects are selected (i.e. as every fourth household). All members of the household aged 12 and older are invited to participate in the survey. The rotating panel design means that households are randomly selected to be in the sample and interviewed every six months for three years. There are six different panels at any point in the survey, and a different panel is interviewed each month in the six-month cycle.

It should also be pointed out that not everyone in the sample has been a victim of a crime. The data I used was incident-level data, which only includes victims of crimes. This included a total of 44,481 records, of which 777 were of a sexual nature. These included: completed rape, attempted rape, sexual assault with serious assault, sexual assault with minor assault, sexual assault without injury, unwanted sexual contact without force, verbal threat of rape, and verbal threat of sexual assault. In addition, it is possible that multiple incidents are reported from the same respondent, especially due to the rotating panel design. For these 44,481 incidents, there were 32,079 unique respondents behind them.

Table 1: Distribution of Types of Crime

Type of Crime	n	prop
Sex	777	2%
Assaults	7584	17%
Burglary	6728	15%
Motor Vehicle Theft	1309	3%
Robbery	912	2%
Theft	27531	61%

Table 2: Sample Demographics - Age

Age Group	n	Proportion
18-34	14037	31%
35-49	11731	26%
50-64	11068	25%
over 65	5911	13%
under 18	2094	5%

Table 3: Sample Demographics - Sex

Sex	n	Proportion
Female	24175	54%
Male	20666	46%

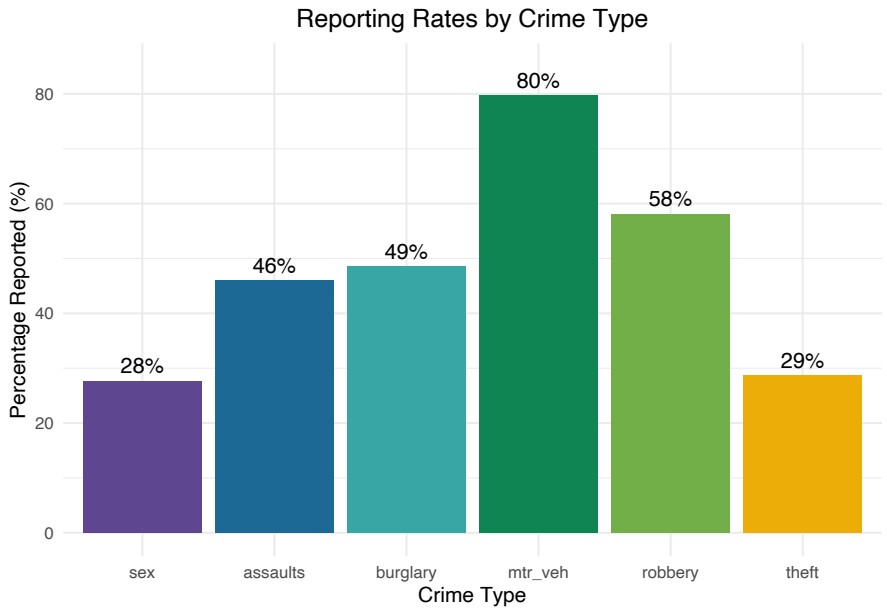

Figure 1: Motor vehicle thefts were the most reported

Methods

Since my outcome of interest takes on only two values—whether a crime was reported to the police or not—I fit a logistic regression model, which models the odds of a crime being reported as a linear function of the predictor variables.

My model had multiple levels to reflect the structure in which the data was collected. Each record in the data represents a victimization experienced by the respondent, so each respondent could have multiple entries. In addition, since the survey features a rotating panel design, the same respondents were surveyed over time. I added a random intercept term for each respondent to account for differences between respondents that may affect all of their responses.

For the possible predictor variables, I considered the type of crime, whether it resulted in an injury, and the relationship of the victim and the offender. I also considered the age and sex of the victim (the respondent). These are all categorical variables, which I constructed myself (except for sex), since there were many options in the original data. For example, the types of crime were also condensed into six main categories, which the NCVS uses in its own reports: robberies, burglaries, sexual assault, assault, theft, or motor-vehicle theft. Verbal threats of these crimes are included as well, also in the example of the NCVS. For the victim's relationship with the offender, I condensed this into just two broad categories: known or unknown. The original options included spouse, parent, siblings, employers, neighbors, schoolmates, and friends, appropriate groupings were not obvious, and I did not want to distort the results through any improper groupings. My special predictor of interest was whether the crime occurred before or after the start of the #MeToo movement, which is officially dated as October 15, 2017. Crimes that occurred during or before the month of October 2017 were considered as predating the movement. The dates used to determine this were the month and year of the incident as reported by the respondent, not the time of the survey.

Table 4: Variable Descriptions

Variables	Description
Type of Crime	Robbery, Burglary, Sexual Assault, Assault, Theft, Motor-vehicle Theft
Injury	Yes, No
Relation	Yes (known), No (unkown)
Age Group	Under 18, 18-34, 35-49, 50-64, over 65
Sex	Male, Female
MeToo	Pre, Post

The model was fitted using the `glmer` function from the `lme4` package (Bates et al. 2015) in the statistical software `R` (R Core Team 2020). When fitting a regression model survey data, it is important to consider the design on the survey, which can be done through the `svyglm` and `svydesign` functions in the `survey` package (Lumley 2004). However, to my knowledge, it currently does not support multilevel models. This left me with a choice between accounting for the survey design, or the dependence of the data. I chose the latter, as it is a fundamental assumption of the logistic regression model.

To select the final model, I fitted a series of models using different combinations of the above variables, and assessed them based on their relative AIC values. Type of crime and MeToo were included in each because they were the main variables of interest. The interaction term for these was included because I wanted to compare sex crimes with the other types of crime, and the effect of MeToo on sex crimes specifically.

Results

Table 5: Model Comparisons

Models	AICs
intercept only	57616.39
crime type * MeToo	54936.16
crime type * MeToo + sex	54938.01
crime type * MeToo + age group	54633.13
crime class * MeToo + age group + known	54634.08
crime_class * MeToo + age group + injury	54473.93

A comparison of the candidate models and their AIC scores indicated that age group and injuries were important to the model. The final model in its entirety is as follows:

$$\frac{p}{1-p} = \text{logit}^{-1}(\alpha_0 + \alpha_{id[i]} + \beta_1(\text{MeToo}_{ij}) + \beta_2(\text{theft}_{ij}) + \beta_3(\text{MotorVehicle}_{ij}) + \beta_4(\text{burglary}_{ij}) + \beta_5(\text{assault}_{ij}) +) +$$

$$\beta_6(\text{robbery}_{ij}) + \beta_7(\text{AgeUnder18}_i) + \beta_8(\text{Age35-49}_i) + \beta_9(\text{Age50-64}_i) + \beta_{10}(\text{AgeOver65}_i) + \beta_{11}(\text{injury}_{ij}) +$$

$$\beta_{12}(\text{sex}_{ij})(\text{MeToo}_i) + \beta_{13}(\text{theft}_{ij})(\text{MeToo}_{ij}) + \beta_{14}(\text{MotorVehicle}_{ij})(\text{MeToo}_i) + \beta_{15}(\text{burglary}_{ij})(\text{MeToo}_i) +$$

$$\beta_{16}(\text{assault}_{ij})(\text{MeToo}_i) + \beta_{17}(\text{robbery}_{ij})(\text{MeToo}_i)$$

p = probability of reporting

$$\text{MeToo}_{ij} = \begin{cases} 0 & \text{if After} \\ 1 & \text{if Before} \end{cases}$$

$$\text{burglary}_{ij}, \text{robbery}_{ij}, \text{assault}_{ij}, \text{theft}_{ij}, = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

$$\text{AgeUnder18}_i, \text{Age18-34}_i, \text{Age35-49}_i, \text{Age50-64}_i, \text{AgeOver65}_i = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

$$\text{injury}_{ij} = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

$$\alpha_{id} \sim N(0, \sigma^2_{id})$$

In other words, the odds of an incident being reported to the police depends on the victim's age, the type of crime, whether they were injured, and whether it occurred before the MeToo movement.

Results

Table 6: Model Results

term	estimate	std.error	p.value
(Intercept)	-1.047	0.146	0.000
me_toopre	-0.053	0.208	0.799
crime_classassaults	0.748	0.153	0.000
crime_classburglary	0.831	0.154	0.000
crime_classmtr_veh	2.816	0.195	0.000
crime_classrobbery	1.332	0.200	0.000
crime_classtheft	-0.218	0.148	0.140
age_group35-49	0.172	0.037	0.000
age_group50-64	0.100	0.038	0.008
age_groupover 65	0.027	0.046	0.555
age_groupunder 18	-1.150	0.078	0.000
injuryyes	0.930	0.075	0.000
me_toopre:crime_classassaults	0.021	0.216	0.924
me_toopre:crime_classburglary	0.228	0.218	0.295
me_toopre:crime_classmtr_veh	-0.029	0.268	0.913
me_toopre:crime_classrobbery	0.028	0.274	0.920
me_toopre:crime_classtheft	0.155	0.210	0.460

Table 7: Model Results (Exp)

exp(term)	estimate	std.error	p.value
(Intercept)	0.351	1.158	0.000
me_toopre	0.948	1.231	0.799
crime_classassaults	2.114	1.165	0.000
crime_classburglary	2.296	1.167	0.000
crime_classmtr_veh	16.702	1.215	0.000
crime_classrobbery	3.787	1.221	0.000
crime_classtheft	0.804	1.159	0.140
age_group35-49	1.188	1.038	0.000
age_group50-64	1.105	1.038	0.008
age_groupover 65	1.028	1.047	0.555
age_groupunder 18	0.317	1.081	0.000
injuryyes	2.534	1.077	0.000
me_toopre:crime_classassaults	1.021	1.241	0.924
me_toopre:crime_classburglary	1.256	1.243	0.295
me_toopre:crime_classmtr_veh	0.971	1.307	0.913
me_toopre:crime_classrobbery	1.028	1.315	0.920
me_toopre:crime_classtheft	1.168	1.234	0.460

First, in order to interpret the results, the baseline for comparison needs to be established. In this case, the odds of a sexual assault crime being reported to the police after MeToo for a victim aged 18-34 is estimated to be 0.31, which indicates that it is over three times more likely for it to be unreported than reported. The odds are are comparison of the probability of reporting to the probability of not reporting, so odds greater than 1 mean that reporting is more likely, and vice versa. Other types of crime were found to have significantly ($p<0.05$) higher odds. Motor vehicle thefts were estimated to have odds that were 16 times higher, while other non-sexual assaults had twice as high odds. Robberies, which are thefts that do not directly involve the person, had odds that were four times as high. Incidents that resulted in an injury were predicted to have odds that were 2.5 times greater as well. On the other end, victims under the age of 18 had significantly lower odds; their odds were 70% lower compared to their counterparts aged 18-34. The odds of thefts being reported were 20% lower than sexual crimes as well. Interestingly, the model results suggest that sexual crimes that occurred before MeToo had lower odds of being reported. However, this result was not found to be statistically significant at all.

However, the performance of the model must be addressed before these results can carry any meaningful weight. The overall accuracy of the model predictions was 81%, meaning that it correctly predicted the outcome in 81% of the incidents. But, taking a further look at this shows that the model performs unevenly for the two cases. It correctly predicted 96% of unreported incidents, but only managed to correctly identity 54% of reported incidents.

Table 8: Prediction Accuracy

	True No	True Yes
Predicted No	27514	7577
Predicted Yes	894	8856

Discussion

The results show that there remains an aversion to reporting crimes of a sexual nature to police, even in the post-MeToo world. However, this is by no means intended to question the indubitable influence of the movement. Rather, there are many reasons as to why it does not manifest in the model and its results. First, reporting to the police probably represents one of the highest levels of speaking out against an offender,

The lack of fit of the model can be attributed to several things. It could be due in part to the fact that there were less reported cases in the data, but most likely suggests the exclusion of important variables in the model. The raw data files have over a thousand variables, but I only tested a selection of them. For example, I did not consider at all the effect of an incident occurring on an American Indian reservation, or the effect of the offender holding a blunt object. So, while I exercised my best judgement to choose relevant ones, it is possible that some may have slipped through the cracks. While there are algorithmic methods of variable selection that can a large number of variables, this was not a viable option for this data because I had to manually adjust many of the variables. For example, there were initially over twenty types of crimes and locations that I condensed into broad categories. This is also another possible source of error; it is possible that I may have muddled some important information in this process.

In addition, as mentioned earlier, this model fails to account for the design of the survey. While I felt this was a necessary trade-off at the time, it does mean these results are not generalizable to all crime in the US since the sample has not been adjusted for non-respresentativeness. One possible solution for this is incorporating the survey weights provided by the survey into the model as suggested by Carle (2009); this is something I would need to research further before implementing.

References

Allaire, JJ, Jeffrey Horner, Yihui Xie, Vicent Marti, and Natacha Porte. 2019. *Markdown: Render Markdown with the c Library 'Sundown'*. https://CRAN.R-project.org/package=markdown.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bolker, Ben, and David Robinson. 2020. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Carle, Adam C. 2009. "Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations." *BMC Medical Research Methodology* 9 (1): 49. https://doi.org/10.1186/1471-2288-9-49.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Justice Statistics, United States. Bureau of. 2020a. "National Crime Victimization Survey, [United States], 2016." https://doi.org/10.3886/ICPSR36828.v4.

———. 2020b. "National Crime Victimization Survey, [United States], 2017." https://doi.org/10.3886/ICPSR36981.v2.

———. 2020c. "National Crime Victimization Survey, [United States], 2018." https://doi.org/10.3886/ICPSR37297.v1.

———. 2020d. "National Crime Victimization Survey, [United States], 2019." https://doi.org/10.3886/ICPSR37645.v1.

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9 (1): 1–19.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.