Modeling Contributing Factors of Unreported Crimes for Victims in the United States

Gianni Spiga

February 18, 2024

Abstract

The 2020 National Crime Victimization Survey (NCVS) provides information about crime around the country. Following data cleaning and re-encoding, logistic regression models the determinants of a crime being reported to the police or not. Attempts in expanding the model with interactions and a generalized partially linear additive model lead only to modeling with a few predictors as main effects. While the model fits the data well, future analysis should seek to add more information as well as increase balance in predictors.

1 Introduction

Since 1973, the United States' primary source of information on criminal victimization has been the NCVS, run by the Bureau of Justice Statistics [BJS21]. Participants in the survey are interviewed on the characteristics, frequency, and consequences of victimization by crime in the country. The extensive survey gathers details on types of crime, including nonfatal and property crimes, regardless if they are reported to the police. The details of the incident as well as the victim are collected.

Our data of interest is from the year 2020 NCVS survey. In this analysis, we join two of the five provided datasets, one on victim characteristics and the other on details of the incident(s) suffered by the victim, to analyze what factors of a household determine whether or not a crime against the household is reported to the police. Not reporting a crime is not illegal, unless one is a mandatory reporter according to the state [Pir23]. However, social researchers and government are interested in what contributing factors lead to crimes not being reported. We form this response as binary, yes or no, allowing us to model this investigation with logistic regression. The data is made accessible by the University of Michigan, which provides the data from the Bureau of Justice Statistics and the United States Department of Justice through download.

2 Descriptive Analysis

2.1 Data Description and Cleaning

The NCVS data is composed of multiple files, corresponding to information at the household-level, person-level, and incident-level. The overall study contains 270,566 subjects, but our interest is in the 8,043 who experienced an incident[1]. We join our person-level data with incident-level data on the unique ID assigned to each person in the survey and their interview date, leaving 8043 rows and 112 columns.

Initial loading shows a large number of missing values are distributed across columns. To combat this, we remove all columns with have more than 10% of missing values, with the removal of all rows missing values after. Additionally, we remove multiple columns with redundant information. We drop the ID columns, columns signifying strata, repetitive age and sex identification, and more. We remove observations missing information from survey collectors, marked as "Residue" in the data. This was done on the response variable as well, along with the 161 observations who marked "Don't know" when asked if they reported to the police or not. This way, our response remains binary. Table 1 shows the counts of responses pre-removal.

Reported to Police	Count
(1) Yes	2833
(2) No	5099
(3) Don't know	102
(8) Residue	9

Table 1: We remove "Don't know" and "Residue" from the response variable for binary regression

The final cleaning needing to be addressed is the dependency on the number of incidents predictor (V3081) and rows of the data. For example, if a person reported five incidents in the survey, they would have five different rows corresponding to each individual incident they faced, if they reported, etc. This forms a problematic bias in the data since people who experienced more incidents would have their characteristics over-accounted for in the data set. To absolve this problem, for all people who had more than one incident, we randomly sample one of them to keep in the data, while also keeping the number of incidents they had recorded still. This down-sampling is not huge, as more than half the survey respondents only had one incident recorded in 2020. We later encode this variable from numeric to binary, either one incident or two or more incidents.

After this cleaning, we work with a reduced dataset of 5701 incident reports and 36 predictors. A description of the predictors used in the analysis going forward is provided in the Appendix in Table 2.

2.2 Variable Re-encoding

Out of the 36 predictors we have, 35 of them are categorical variables, with 19 of them having more than two levels. This brings about a curse of dimensionality, as building a model with these variables requires multiple dummy variables and a huge increase in predictors. This would mask significant predictors and create convergence issues in model fitting.

The categorical variable that has the most levels is the type of crime committed in the incident, with 34 levels. Table 3 in the Appendix shows the distribution of counts pre-encoding and after. To simplify the model, crimes of similar seriousness, whether they were completed or attempted, were combined. Both completed and attempted rape and sexual assault variations were combined as one. A similar process was done with Robbery and Burglary, Assault, Verbal Threats, and Theft. However, for interest in this analysis and to prevent over-weighting theft more, theft includes all types of theft except motor theft, which remains its own category as originally encoded in this analysis.

[1] An incident is defined as any non-lethal crime, whether personal or property, committed against someone

This encoding process was done similarly with other variables, such as the Educational Attainment of the surveyed, with twenty levels, ranging from Kindergarten to Doctorate degree. We define ”Below Associates” as anything below some college, including high school diplomas and grade schools. The second level ”Undergraduate,” includes Associate’s and Bachelor’s degrees, and the final level ”Graduate,” includes Master’s, professional school, and doctorate degrees. Figure 2 in the appendix displays the differences in counts via bar plots before and after encoding. In the Marital Status predictor, we combine widowed, divorced, and separated interviewees as ”Lost Partner,” leaving ”Married” and ”Never Married” as is. Race contains 17 different combinations options, so we combine those who are not only black, white, or Asian as ”2 or 3 Races”, an already existing category. This process continues with sexuality, Active duty status, and current education.

2.3 Visualizing the Data

After re-encoding our data, we can understand how some of our variables of interest are distributed across our recorded incidents. Figure 1 shows the type of crime in our data as well as the counts of whether or not it was reported to the police for each type. We can see that a majority of thefts and verbal threats go unreported, while motor theft and assaults have a larger portion of police reports. The majority of incidents in this data are theft, with 2,605 theft incidents unreported. Rape and Sexual Assault are the least reported here, with only 43 reports in this dataset.

Figure 3 in the Appendix shows the distribution of respondents by sex, type of crime, and police reports. From the plot, women in the data experience more Rape/Sexual Assaults than men, and a majority of those go unreported to the police. Regardless of gender, the majority of motor thefts are reported. Females also experience a higher amount of unreported theft compared that of males.

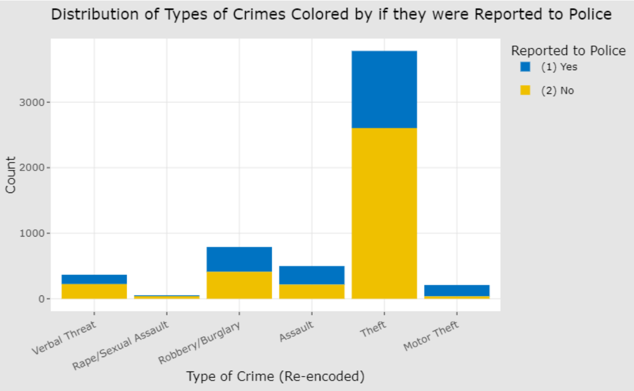


Figure 1

Figure 4 in the Appendix displays boxplots distributing age with incidents reported separated by Marital Status and whether or not the victim knew the attacker. Regardless of the attacker being known, we can see that those who never married are noticeably younger than those who married or lost a partner. The oldest martial status group on average would be those who lost a partner by death, separation, or divorce. In those who were married or had lost a partner, victims who did not personally know the attacker were slightly older on average than those who did know the attacker. However, those who never married had very little difference in age regardless if they knew the attacker or not.

3 Inferential Analysis

3.1 The Logistic Model

We first begin by creating a full model with only main effects, resulting in 56 coefficients. For this model, we encode our outcome "1" as not being reported to the police. From our initial model, we find the most significant variables are the type of crime, with Motor Theft having the smallest p-value. Other significant predictors include age, if someone was not a citizen, the number of incidents, and if they were never married. Despite these significant variables, there are many erroneous variables in the model that are insignificant. We aim to reduce the complexity of this model with backward stepwise regression, penalized by Bayesian Information Criterion.

Performing such, we have a reduced model with only 11 predictors now, age, marital status, if the victim has ever been attacked/threatened prior to the incident in a separate situation (outside of the incident recordings in the survey), if the victim knew the attacker, the number of incidents a victim experienced, and the types of crime. Unlike our previous model, predictors that were once significant such as citizenship are not included in this model by the BIC. The variable with the largest coefficient would be Motor Theft, at -1.73. In context, if someone had been a victim of a motor theft, they were likely to report it to the police, which corresponds with the observation we made about the number of reported motor thefts earlier in our descriptive analysis. Next, we look towards methods of model expansion to improve our inference.

3.2 Attempts to Expand the Model: Interactions and GPLAM

Given the large number of predictors with multiple factor levels, even after re-encoding, building a model with the original variables and all two-way interactions was simply unfeasible. In attempting to do so, the model did not converge after over 10 minutes of computation time and returned multiple NA values for coefficients. We restrict ourselves to only two-way interactions with the main effects from our reduced model above. We perform the same backward stepwise regression by BIC to find the best model picked by BIC is still the model with main effects only. While this is good news for our interpretability of a model, we need to ensure there are no other possibilities of model improvement missed, so we check another option for modeling, the generalized partially linear additive model.

Given that the only predictor which is discretely continuous in the data is age, a GPLAM is our only approach. Using the **gam** library in R, we build models replacing age with quadratic and cubic smoothing splines. However, our non-parametric ANOVA for these terms reveals neither of these attempts are statistically significant, returning χ^2 p-values of 0.4622 and 0.4763 respectively. Given this, we have evidence that treating age as a linear term is our best model, thus returning us back to our reduced, main effects model once more, as shown below.

$$logit(E(Y_{Report}|X)) = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{MaritalStatus} + \beta_3 X_{OthAtk} + \beta_4 X_{AtkerKnown} + \beta_4 X_{NumOfInc} + \beta_5 X_{TOC} \tag{1}$$

The coefficients and their respective 95% confidence intervals can be found in Appendix Table 4.

4 Model Diagnostics

To assess the fit, we refer to Figure 2, which plots our deviance residuals versus our fitted value for the model. From the smoothing spline in red, we can see that our residuals closely follow along the horizontal line at zero, indicating a good fit for the model. Performing a Run's test on our deviance residuals, we return a p-value of 0.9473, concluding there is no systematic pattern in the residuals.

While Appendix Figure 6 shows no obvious outliers, we investigate outliers using Cooks distance. Appendix Figure 6 shows the Cook's Distance and Leverage plots, where can see the three most noticeable outliers are observations 262, 2286, and 3317. All three are Rape/Sexual Assault incidents who reported their incidents to the police, which is opposite of what we would expect with these types of crimes. Observation 262 never married, which also trends towards not reporting to the police. Observations 2286 and 3317 are married currently or were previously, but the prior had more than one incident in the past, and the latter knew the offender. Thus the underlying reason that all these victims were outliers was that they all had a characteristic of themselves or their incident that was opposite of what our model would predict. Since numerically, these outliers are not high leverage, we will keep them in the model.

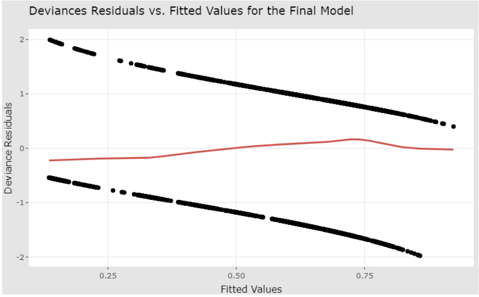


Figure 2

5 Discussion and Conclusion

Starting the model building with almost 40 predictors, the result of only six predictors being significant is surprising. Intuitively, one would expect characteristics such as citizenship, education, race, and sex to take be significant in modeling. From the model coefficients in the appendix, we see the older one is, the more likely they are not to report an incident to the police, however, the magnitude of age is not large, with a coefficient of only 0.0046. Those who are married might feel obligated to protect their partners and thus, more likely to report an incident. In the final model, those who were previously married were combined with those married due to lack of significance. Robbery and Burglary crimes were combined with Theft as well. Both not being previously attacked/threatened as well as having experienced multiple incidents within the survey period, we suspect an overall effect that as someone experiences more crimes, they are less likely to report the police. If one knows the offender, it is possible they want to resolve the matter personally rather than going to the police. Out of all the crimes, having a motor theft makes victims the most likely to go to the police, while Rape/Sexual Assaults victims are the least likely to report.

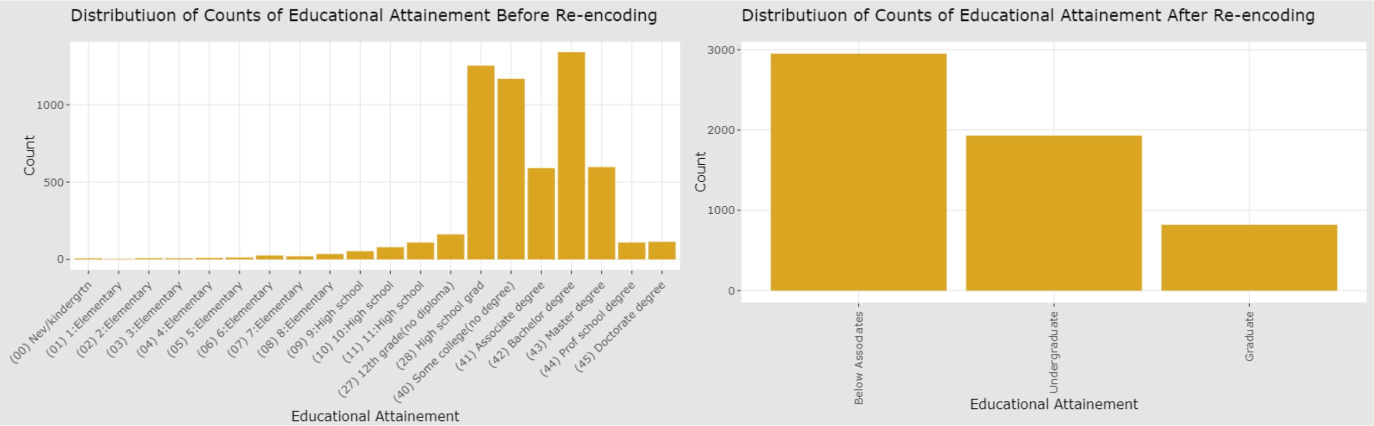
The National Crime Victimization Survey contains a plethora of information for modeling police reporting, as for this analysis to be manageable, we not could sort through the 1000+ columns of information provided regarding incidents. Some might seek to frame crime reporting as a race issue or an education issue. However, the model found in this paper suggests systematic and/or social reasons for crimes going unreported. This topic should be of great interest to local governments, who should seek to encourage crime reporting as a tool of prevention.

References

[BJS21] National crime victimization, 2021.

[Pir23] Rebecca Pirius. Failure to report a crime. *Lawyers.com*, 2023.

6 Appendix



(a) Before Encoding

(b) After Encoding

Figure 3: Distribution of Educational Attainment

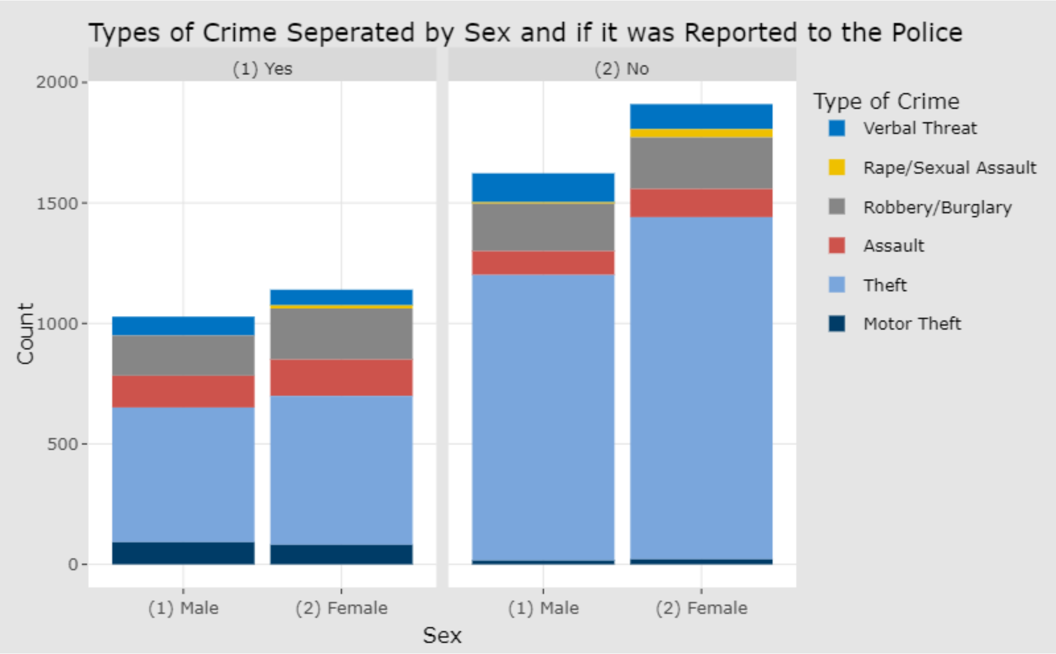


Figure 4: We can see that Females experienced a larger amount of Theft and Rape/Sexual Assault

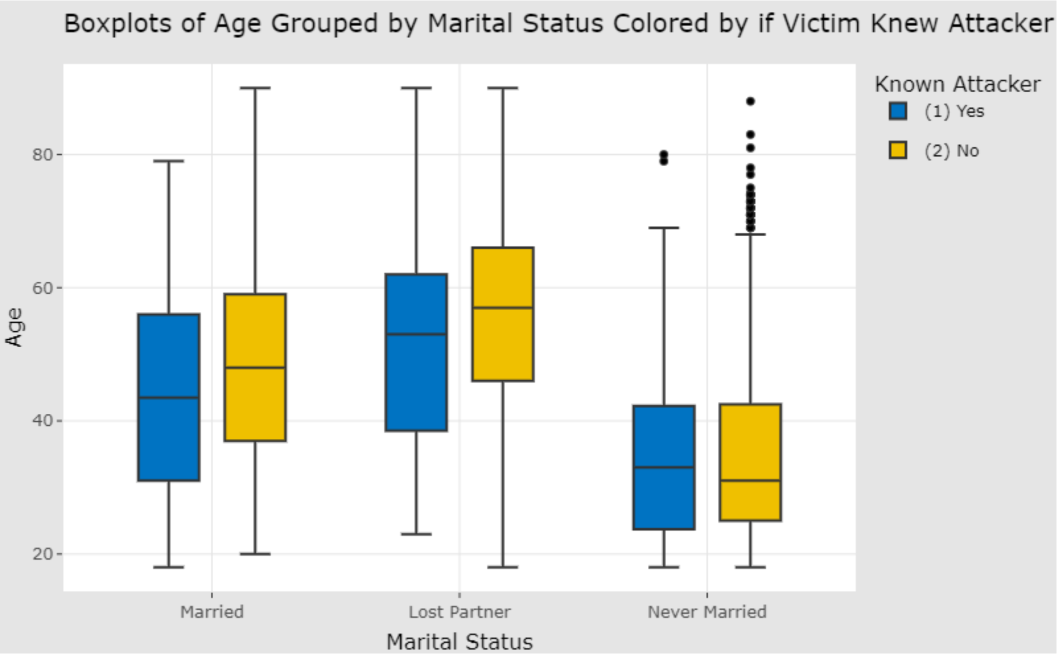
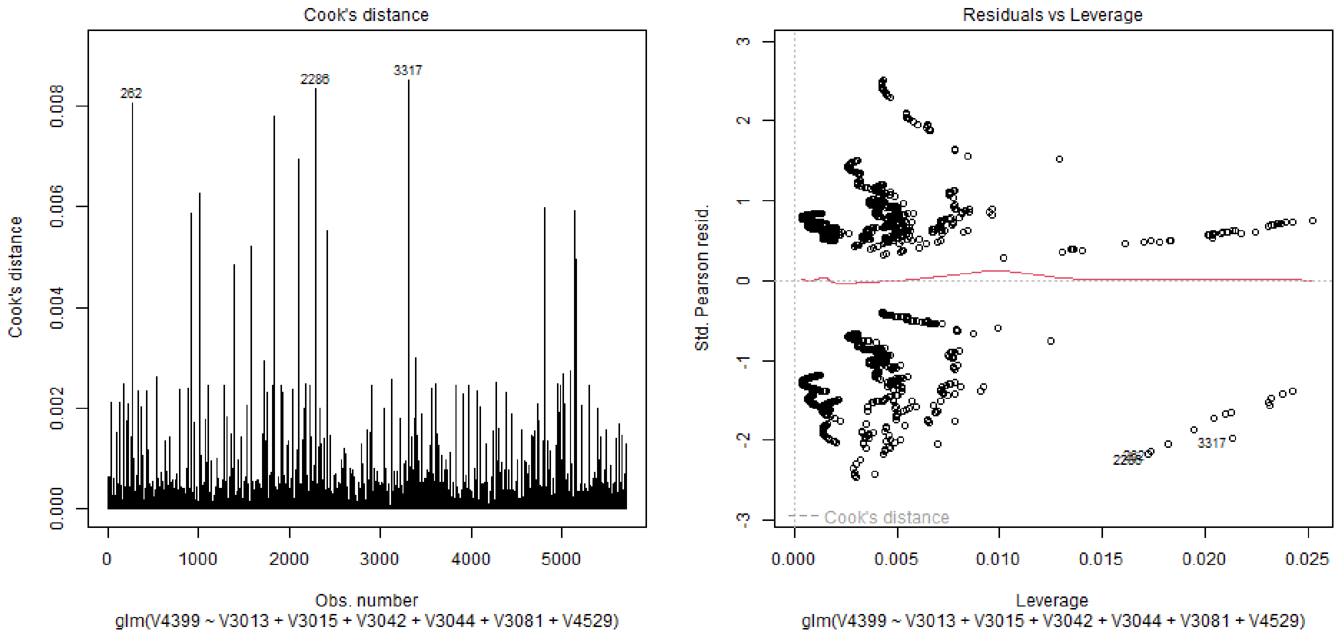


Figure 5



(a) Cooks Distance

(b) Leverage Plot

Figure 6: Plots for Outliers

Variable Name	Variable Description
YEARQ	YEAR AND QUARTER OF INTERVIEW (YYYY.Q)
V3006	HOUSEHOLD NUMBER
V3011	TYPE OF INTERVIEW
V3012	RELATIONSHIP TO REFERENCE PERSON
V3013	AGE (ORIGINAL)
V3015	MARITAL STATUS (CURRENT SURVEY)
V3017	SEX (ORIGINAL)
V3020	EDUCATIONAL ATTAINMENT
V3023A	RACE RECODE (START 2003 Q1)
V3024	HISPANIC ORIGIN
V3025	MONTH INTERVIEW COMPLETED
V3034	SOMETHING STOLEN OR ATTEMPT
V3040	ATTACK, THREAT, THEFT: LOCATION CUES
V3042	ATTACK, THREAT: WEAPON & ATTACK CUES
V3044	STOLEN, ATTACK, THREAT: OFFENDER KNOWN
V3046	FORCED OR COERCED UNWANTED SEX
V3061	C TELEPHONE INTERVIEW
V3062	C NO ONE BESIDES RESPONDENT PRESENT
V3063	C RESPONDENT
V3064	C HH MEMBER(S) 12+, NOT SPOUSE
V3065	C HH MEMBER(S) UNDER 12
V3066	C NONHOUSEHOLD MEMBER(S)
V3067	C SOMEONE PRESENT, CAN
V3068	C DON'T KNOW IF SOMEONE IS PRESENT
V3_V4526H3A	ARE YOU DEAF OR DO YOU HAVE SERIOUS DIFFICULTY HEARING? (START 2016 Q3)
V3_V4526H3B	ARE YOU BLIND OR DO YOU HAVE SERIOUS DIFFICULTY SEEING EVEN WHEN WEARING GLASSES (START 2016 Q3)
V3_V4526H5	DIFFICULT: LEARN, REMEMBER, CONCENTRATE (START 2016 Q3)
V3_V4526H4	LIMITS PHYSICAL ACTIVITIES (START 2016 Q3)
V3_V4526H6	DIFFICULT: DRESSING, BATHING, GET AROUND HOME (START 2016 Q3)
V3_V4526H7	DIFFICULT: GO OUTSIDE HOME TO SHOP OR DR OFFICE (START 2016 Q3)
V3083	CITIZENSHIP STATUS (START 2017 Q1)
V3084	SEXUAL ORIENTATION (START 2017 Q1)
V3087	SERVE ON ACTIVE DUTY (START 2017 Q1)
V3071	HAVE JOB OR WORK LAST WEEK
V3079	ATTENDING SCHOOL
V3081	NUMBER OF CRIME INCIDENT REPORTS
V4399	REPORTED TO POLICE
V4529	TOC CODE (NEW, NCVS)

Table 2: Variables used in Analysis after Data Cleaning

Type of Crime	Count
(01) Completed rape	40
(02) Attempted rape	15
(03) Sex aslt w s aslt	10
(04) Sex aslt w m aslt	2
(05) Rob w inj s aslt	22
(06) Rob w inj m aslt	18
(07) Rob wo injury	43
(08) At rob inj s asl	6
(09) At rob inj m asl	8
(10) At rob w aslt	30
(11) Ag aslt w injury	89
(12) At ag aslt w wea	67
(13) Thr aslt w weap	133
(14) Simp aslt w inj	123
(15) Sex aslt wo inj	19
(16) Unw sex wo force	1
(17) Asl wo weap, wo inj	296
(18) Verbal thr rape	8
(19) Ver thr sex aslt	5
(20) Verbal thr aslt	490
(21) Purse snatching	2
(22) At purse snatch	2
(23) Pocket picking	20
(31) Burg, force ent	237
(32) Burg, ent wo for	515
(33) Att force entry	169
(40) Motor veh theft	198
(41) At mtr veh theft	60
(54) Theft < $10	383
(55) Theft $10-$49	919
(56) Theft $50-$249	1407
(57) Theft $250+	1141
(58) Theft value NA	571
(59) Attempted theft	326

Type of Crime	Count
Verbal Threat	503
Rape/Sexual Assault	87
Robbery/Burglary	1052
Assault	708
Theft	4766
Motor Theft	258

Table 3: Type of Crime before Re-encoding (left) and after (right).

	Coefficient	95% CI Lower	95% CI Upper
(Intercept)	0.80	0.36	1.25
Age	0.005	0.001	0.01
Never Married	0.38	0.24	0.52
Not Previously Attacked/Threatened	-0.44	-0.69	-0.19
Offender Not Known	-0.56	-0.87	-0.25
Number of Incidents 2+	0.46	0.30	0.62
Rape/Sexual Assault	0.69	0.02	1.36
Theft/Robbery/Burglary	0.44	0.20	0.68
Assault	-0.73	-1.01	-0.45
Motor Theft	-1.76	-2.18	-1.33

Table 4