# Final Exam

STAT 353

Zirui Zhou

## Contents



## An Overview of the Problem

In the United States, gun violence in K-12 schools has grown rapidly over the past two decades. For example, the mass shooting at Uvalde Elementary in Texas (2022) received a large degree of media attention. While the scale of this event was extreme, however, gun violence of smaller scales is more common .

As gun violence increases, researchers and policymakers continue to search for solutions. These include ideas like increasing monitoring of social and mental health of students, using metal detectors, stationing police in schools, among others. This question - What can we do to reduce gun violence? - provides the background for this exam.

### The SSOCS Data

"The School Survey on Crime and Safety (SSOCS) — a nationally representative survey of U.S. K–12 public schools — is managed by the National Center for Education Statistics (NCES), an agency within the U.S. Department of Education's Institute of Education Sciences. SSOCS collects detailed information from public schools on the incidence, frequency, seriousness, and nature of violence affecting students and school personnel. SSOCS also collects information on the programs, practices, and policies that schools have in place to prevent and reduce crime. Data from this collection can be used to examine the relationship between school characteristics and violent crimes in regular public primary, middle, high, and combined schools."

All of the information that you need to understand this data is provided. This includes:

- `SSOCS(2017-2018)Data.csv` : The data
- `ssocs codebook.pdf` : The code book

Notice that in the code book, the `Appendix A` includes the actual survey and that `Appendix B` includes a list of all the variable names and definitions. Further information on the creation of composite variables (those ending in "18") can be found in `Chapter 5`.

(Throughout, pay particular attention to data with values of "-1". These are purposeful skips and in many (but not all) cases may need to be re-coded to "0".)

### This Exam

The purpose of this exam is to test your ability to put to use all that you have learned in STAT 353 in the context of real data, with a real question. This involves combining your understanding of regression concepts and theory with the implementation of these in code and clear interpretation to a lay audience. Be sure to convey what the results tell you, what assumptions they require, and any limitations in your results.

For this exam, we will focus in particular on two outcomes:

- `INCID18` : total incidents of any crime
- `DISFIRE18` : total use of firearm or explosive

To simply the analysis, you can ignore the sampling weights / jackknife replicates.

**Finally, a strong exam is one that is judicious in what is presented (you can put materials in an Appendix), that explains the decisions and assumptions that were made and why, that explains the how the results should be interpreted, and that is clear in any limitations.**

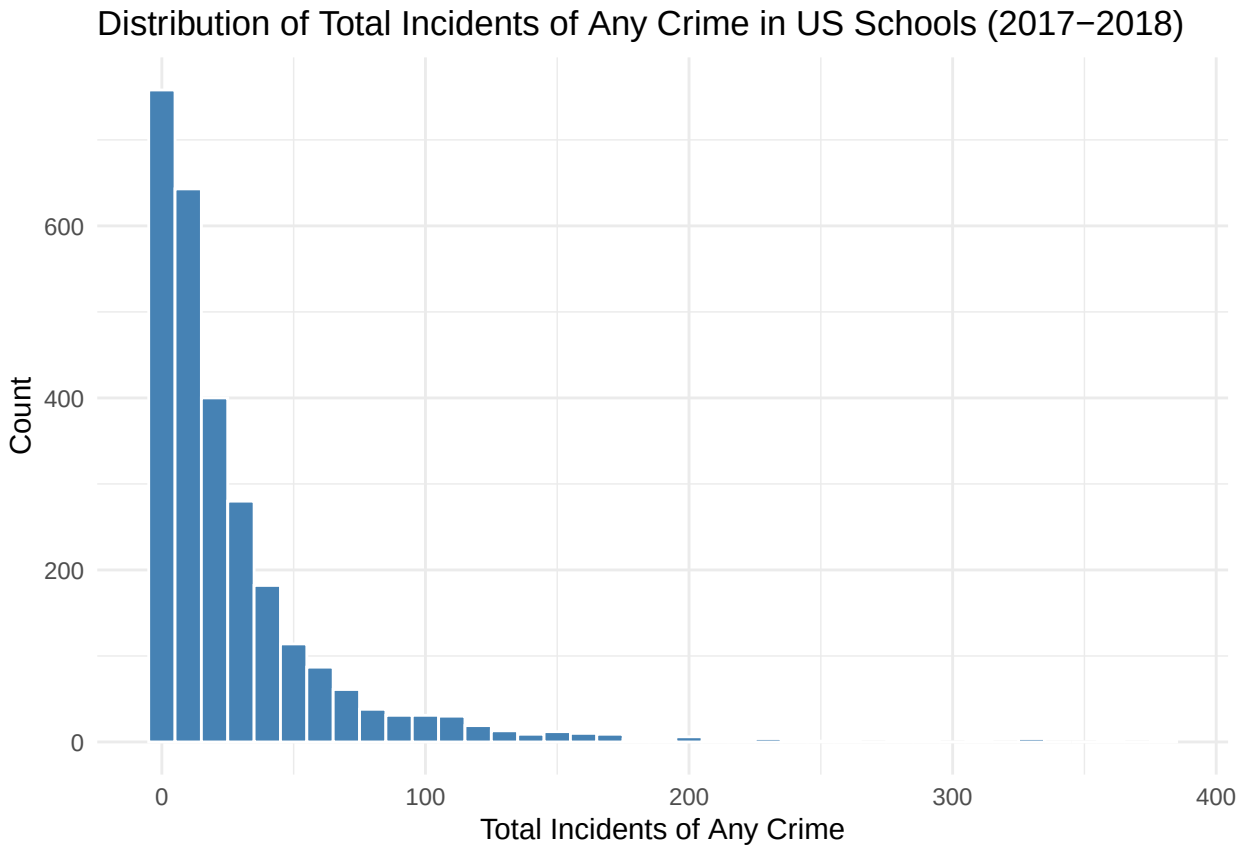
# Part I. Testing Hypotheses

As stated above, researchers and policymakers have hypothesized and enacted a variety of policies meant to reduce crimes and gun violence in schools. In particular, they often argue that schools should include *security guards* in order to reduce crime and gun violence.

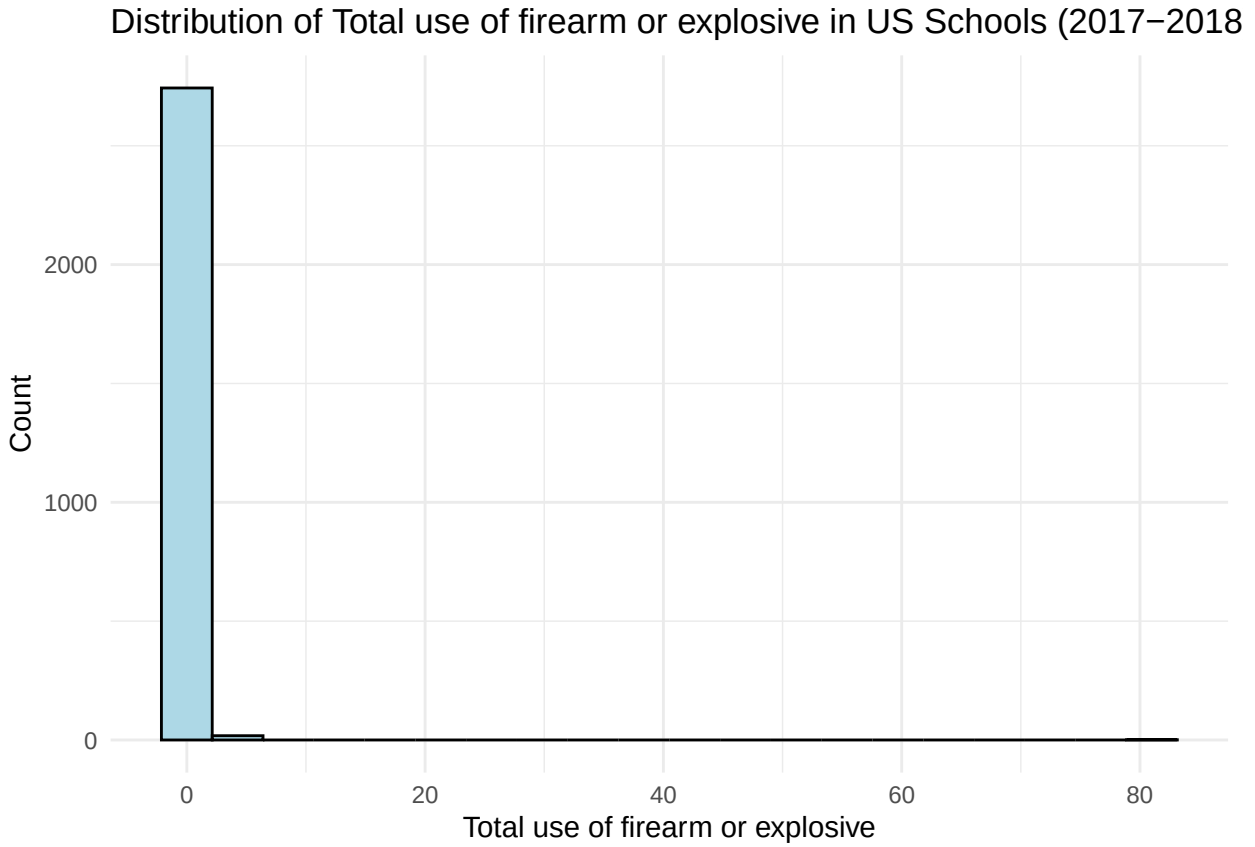
For this part, answer the following questions:

1. After exploring the two outcomes (`INCID18` and `DISFIRE18`) determine what type of regression model is appropriate for each (e.g., OLS). Explain which is best and why.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     5.0    15.0    28.2    35.0   376.0
```




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.00000  0.00000  0.09848  0.00000 81.00000
```

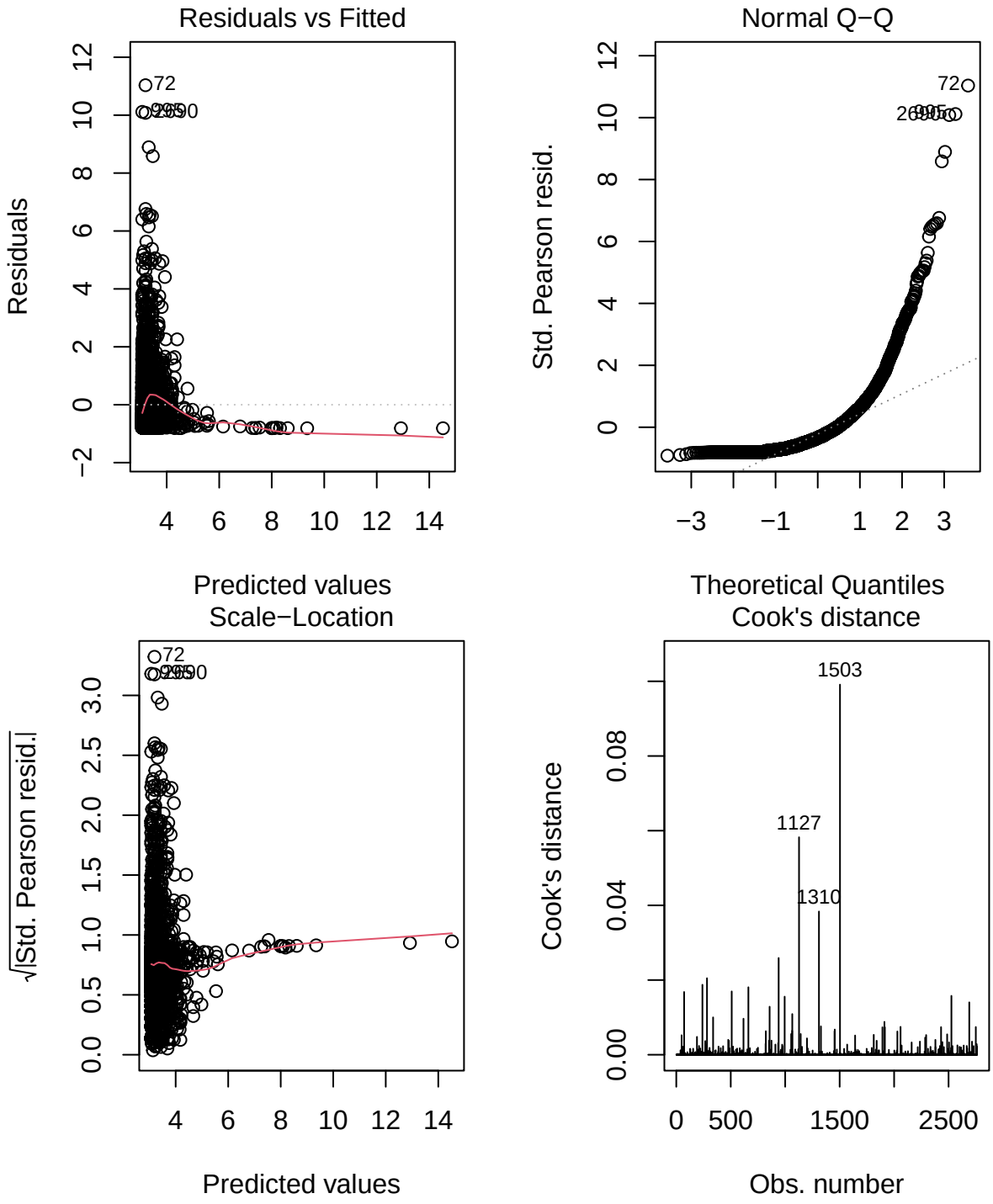


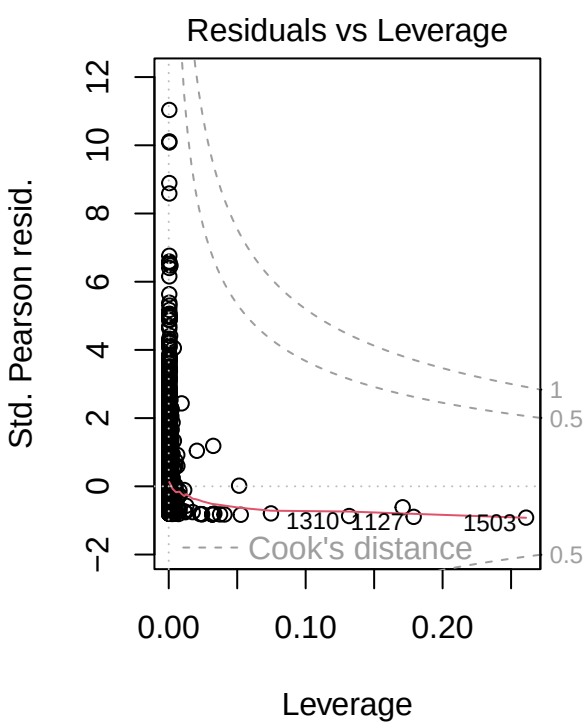

The histogram shows that INCID18 is right-skewed, with a few schools experiencing a large number of incidents while most schools experience relatively few. This suggests that a OLS regression model might not be the best fit for this data, because it assumes a normally distributed error term. Instead, a count data regression model such as Poisson regression or Negative Binomial regression may be more appropriate for modeling the count of incidents. We need to test the overdispersion for the model to determine whether it is needed to use NB regression.

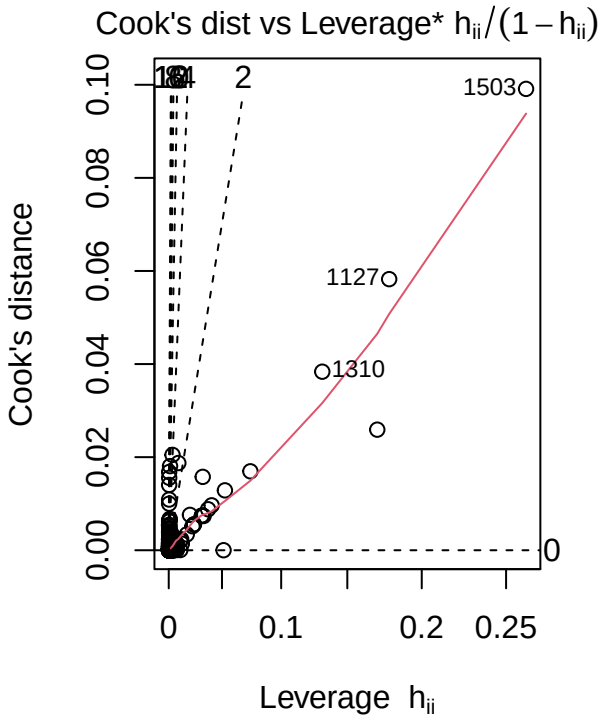
Based on the summary statistics and histogram, the DISFIRE18 variable is heavily skewed and contains a few extreme values. The variable also contains a few distinct values (0, 1, 2, 3, 4, 5, 6, 81), which suggest that a linear regression model may not be appropriate. Therefore, a more appropriate model for DISFIRE18 may be a Poisson regression model or NB regression. Similarly, we need further information to test the overdispersion.

2. Are the presence of *security guards* (`SEC_FT18` and `SEC_PT18`) associated with reductions in crime (`INCID18`) and gun violence (`DISFIRE18`)? Interpret the effects clearly in language that a non-statistician could understand.

```
## [1] 37.11913
```

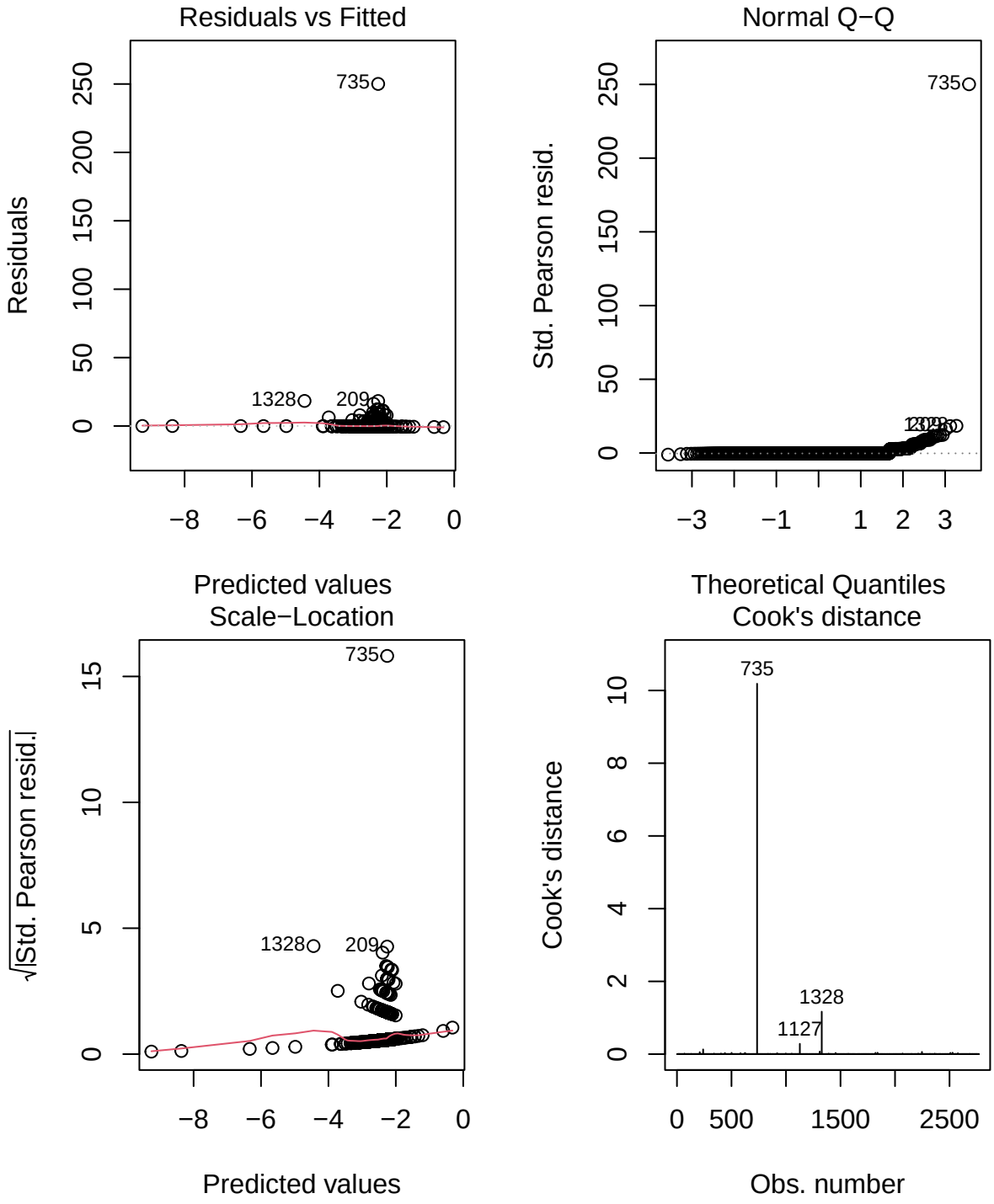

Residuals vs Fitted
Residuals
Predicted values
Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
Scale−Location
√|Std. Pearson resid.|
Predicted values
Cook's distance
Cook's distance
Obs. number
1503
1127
1310
72

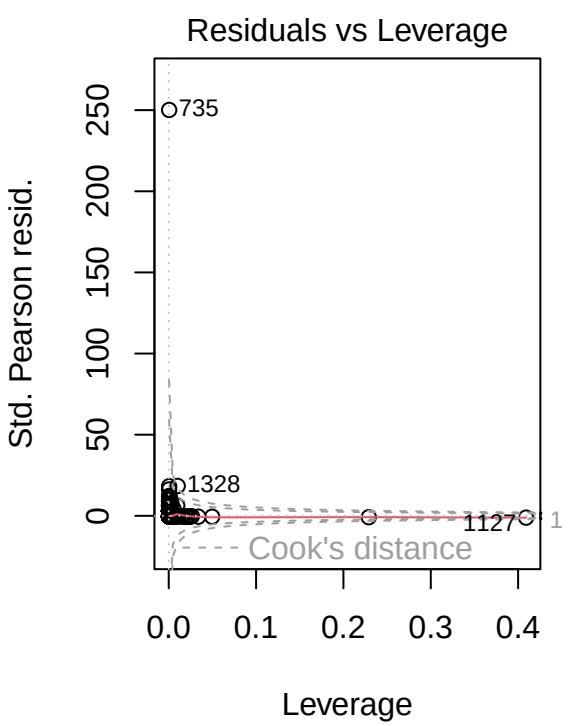


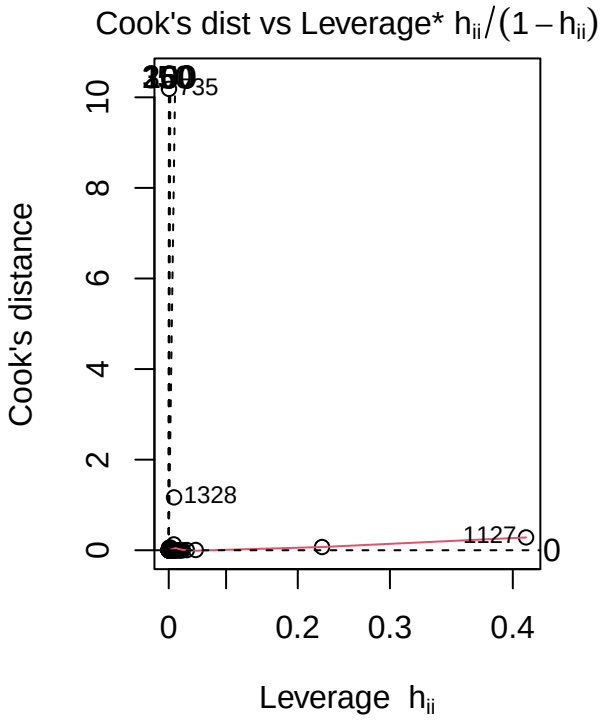




```
##
## Call:
## glm.nb(formula = INCID18 ~ SEC_FT18 + SEC_PT18, data = new_data_1,
##     init.theta = 0.6648927854, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7552  -1.0469  -0.4049   0.2074   3.8447
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.037581   0.027577 110.147  < 2e-16 ***
## SEC_FT18    0.140367   0.006845  20.507  < 2e-16 ***
## SEC_PT18    0.040070   0.013272   3.019  0.00254 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6649) family taken to be 1)
##
##     Null deviance: 3479.5  on 2758  degrees of freedom
## Residual deviance: 3281.6  on 2756  degrees of freedom
## AIC: 23493
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.6649
##           Std. Err.:  0.0174
##
##  2 x log-likelihood:  -23484.8750
```

```
## [1] 0.7906956
```


Residuals vs Fitted
Residuals
Predicted values
735
1328
209
Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
735
Scale−Location
√|Std. Pearson resid.|
Predicted values
735
1328
209
Cook's distance
Cook's distance
Obs. number
735
1328
1127




```
##
## Call:
## glm(formula = DISFIRE18 ~ SEC_FT18 + SEC_PT18, family = "poisson",
##     data = new_data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3952  -0.4693  -0.4629  -0.4023  30.1245
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.233551   0.070686 -31.598  < 2e-16 ***
## SEC_FT18     0.027580   0.009025   3.056 0.002243 **
## SEC_PT18    -0.280810   0.076965  -3.649 0.000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2184.2  on 2758  degrees of freedom
## Residual deviance: 2158.6  on 2756  degrees of freedom
## AIC: 2451.9
##
## Number of Fisher Scoring iterations: 7
```

We initially used a statistical model called Poisson regression to analyze the number of incidents in INCID18, but we found out that the data had more variation than expected, indicating that the model was not suitable. Therefore, we used a different statistical model called negative binomial regression to address this issue. The new model gave us better results and a more accurate understanding of the data. We also carefully looked at the data and removed any unusual data points to ensure that our results were as accurate as possible. The result of the negative binomial regression shows that the presence of full-time security guards (SEC_FT18)

and part-time security guards (SEC_PT18) are both significantly associated with an increase in the incidence of crime (INCID18). Specifically, for every one unit increase in the number of full-time security guards, there is a 14% increase in the incidence of crime, holding all other variables constant. Similarly, for every one unit increase in the number of part-time security guards, there is a 4% increase in the incidence of crime, holding all other variables constant. The model has a deviance of 37.12 and all coefficients are statistically significant.

For DISFIRE18, we started by examining whether it showed overdispersion pattern using a statistical test. The result showed that there was no overdispersion, so we proceeded with the Poisson regression model to investigate the association between the presence of security guards (SEC_FT18 and SEC_PT18) and gun violence. The coefficient estimates show that the presence of SEC_FT18 is associated with a small increase in gun violence, while the presence of SEC_PT18 is associated with a significant decrease in gun violence. The intercept term is also significant, indicating that there is a baseline level of gun violence present regardless of the presence of security guards.

3. To what extent do these effects differ in urban schools versus non-urban schools?

```
##
## Call:
## glm.nb(formula = INCID18 ~ SEC_FT18 * URBAN + SEC_PT18 * URBAN,
##     data = new_data_1, init.theta = 0.6731349519, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8033  -1.0523  -0.3932   0.2189   4.1142
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.930925   0.031507  93.025  < 2e-16 ***
## SEC_FT18        0.145445   0.007685  18.926  < 2e-16 ***
## URBAN           0.397225   0.064946   6.116 9.58e-10 ***
## SEC_PT18        0.055114   0.015217   3.622 0.000292 ***
## SEC_FT18:URBAN -0.036563   0.016678  -2.192 0.028361 *
## URBAN:SEC_PT18 -0.046699   0.030613  -1.525 0.127144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6731) family taken to be 1)
##
##     Null deviance: 3517.1  on 2758  degrees of freedom
## Residual deviance: 3279.8  on 2753  degrees of freedom
## AIC: 23462
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.6731
##           Std. Err.:  0.0177
##
##  2 x log-likelihood:  -23448.0940
##
## Call:
## glm.nb(formula = DISFIRE18 ~ SEC_FT18 * URBAN + SEC_PT18 * URBAN,
##     data = new_data_2, init.theta = 0.03685589319, link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.8209  -0.3178  -0.3070  -0.2773   6.8526
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.32610    0.16106 -14.443  < 2e-16 ***
## SEC_FT18        0.10209    0.03043   3.355 0.000792 ***
## URBAN          -0.46166    0.34179  -1.351 0.176786
## SEC_PT18       -0.23423    0.12342  -1.898 0.057729 .
## SEC_FT18:URBAN  0.05689    0.07464   0.762 0.445945
## URBAN:SEC_PT18  0.05646    0.25987   0.217 0.827995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.0369) family taken to be 1)
##
##     Null deviance: 393.25  on 2758  degrees of freedom
## Residual deviance: 381.35  on 2753  degrees of freedom
## AIC: 1344.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.03686
##           Std. Err.:  0.00496
##
##  2 x log-likelihood:  -1330.05300
```

Before analysis, a new variable URBAN was created with "1" representing urban and "0" representing non-urban.

First, we look at the model_3_1. For SEC_FT18:URBAN, the coefficient estimate is -0.036563, which indicates that the effect of SEC_FT18 on INCID18 is weaker in urban schools compared to non-urban schools.For URBAN:SEC_PT18, the coefficient estimate is -0.046699, which suggests that the effect of SEC_PT18 on INCID18 is also weaker in urban schools compared to non-urban schools, but the effect is not statistically significant at the 0.05 level.Overall, these results suggest that the effects of SEC_FT18 and SEC_PT18 on INCID18 differ somewhat between urban and non-urban schools.

Model 3_2 shows the relationship between the predictor variables and the response variable for DISFIRE18, taking into account the interaction effects between SEC_FT18 and URBAN, and SEC_PT18 and URBAN. The model suggests that the effect of SEC_FT18 on DISFIRE18 is slightly stronger in urban schools compared to non-urban schools, but this effect is not statistically significant (p-value = 0.445945). The effect of SEC_PT18 on DISFIRE18 is slightly stronger in non-urban schools compared to urban schools, but this effect is also not statistically significant (p-value = 0.827995). The intercept for DISFIRE18 is significantly lower in urban schools compared to non-urban schools (p-value = 0.176786), but this effect is not statistically significant at the 5% significance level.

4. Do your analyses suggest that policymakers are correct that security guards reduce crime and gun violence? If so, explain why. If not, conduct additional analyses (using regression) that allow you to evaluate their claim and interpret your results.

```
##
## Call:
## glm.nb(formula = INCID18 ~ SEC_FT18 + SEC_PT18, data = new_data_1,
##     init.theta = 0.6648927854, link = log)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7552  -1.0469  -0.4049   0.2074   3.8447
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.037581   0.027577 110.147  < 2e-16 ***
## SEC_FT18    0.140367   0.006845  20.507  < 2e-16 ***
## SEC_PT18    0.040070   0.013272   3.019  0.00254 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6649) family taken to be 1)
##
##     Null deviance: 3479.5  on 2758  degrees of freedom
## Residual deviance: 3281.6  on 2756  degrees of freedom
## AIC: 23493
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.6649
##           Std. Err.:  0.0174
##
##  2 x log-likelihood:  -23484.8750
##
## Call:
## glm(formula = DISFIRE18 ~ SEC_FT18 + SEC_PT18, family = "poisson",
##     data = new_data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3952  -0.4693  -0.4629  -0.4023  30.1245
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.233551   0.070686 -31.598  < 2e-16 ***
## SEC_FT18     0.027580   0.009025   3.056 0.002243 **
## SEC_PT18    -0.280810   0.076965  -3.649 0.000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2184.2  on 2758  degrees of freedom
## Residual deviance: 2158.6  on 2756  degrees of freedom
## AIC: 2451.9
##
## Number of Fisher Scoring iterations: 7
```

(1) The negative binomial regression model nb_model_1 on INCID18 suggests that the claim made by policymakers that security guards reduce crime may not be accurate. The coefficients for both types of security guards, SEC_FT18 and SEC_PT18, are positive and statistically significant, indicating that having more security guards is associated with a higher incidence of crime and gun violence. However,

it is important to consider that there may be other factors contributing to this relationship, such as schools with higher rates of crime and violence being more likely to hire more security guards.

(2) The poisson_model_2 using the outcome variable of gun violence shows that an increase in the number of full-time security guards is associated with an increase in the incidence of gun violence, while an increase in the number of part-time security guards is associated with a decrease in the incidence of gun violence. The negative coefficient for part-time security guards supports the claim that security guards can reduce gun violence, but the positive coefficient for full-time security guards contradicts the claim. It is possible that full-time security guards may be perceived as a more aggressive presence, leading to an increase in incidents of gun violence.

Overall, further research and analysis would be needed to fully understand the relationship between security guards and incidence of crime and gun violence in schools.

# Part II. Predicting Crime

Other researchers and policymakers would like to develop a model to predict crime (`INCID18`) based upon observable school characteristics. Their idea is that they could first predict schools that have a lot of crime and then put in place interventions that could reduce such crime.

For this part, perform the following tasks.

1. For your first model, use variables `C0532`, `C0534`, `C0536`, `C0538`, `C0560`, `C0562`, `C0568`, `FR_LVEL`, `FR_URBAN`, and `FR_SIZE` as predictor variables. Be sure to pay attention to non-linearities and interactions. (In addition to Appendix B, you can find more detailed explanation for the variables `C0532` to `C0568` on pages 80-81 of the code book, and the three variables `FR_LVEL`, `FR_URBAN`, and `FR_SIZE` on page 172). How well does this model perform?

```
##
## Call:
## glm(formula = INCID18 ~ C0532 + C0534 + C0536 + C0538 + C0560 +
##     C0562 + C0568 + FR_LVEL + FR_URBAN + FR_SIZE, family = "poisson",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -14.829   -3.940   -1.820    1.291   35.541
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.9816584  0.0576110  34.397   <2e-16 ***
## C0532        0.0063960  0.0002128  30.053   <2e-16 ***
## C0534       -0.0039190  0.0002103 -18.638   <2e-16 ***
## C0536       -0.0039174  0.0002128 -18.410   <2e-16 ***
## C0538        0.0542982  0.0017400  31.206   <2e-16 ***
## C0560       -0.0728414  0.0057688 -12.627   <2e-16 ***
## C0562       -0.1680550  0.0076640 -21.928   <2e-16 ***
## C0568        0.0004917  0.0005511   0.892    0.372
## FR_LVEL      0.1607717  0.0051371  31.296   <2e-16 ***
## FR_URBAN    -0.0822063  0.0039675 -20.720   <2e-16 ***
## FR_SIZE      0.5770237  0.0051076 112.974   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 104912  on 2761  degrees of freedom
## Residual deviance:  72079  on 2751  degrees of freedom
## AIC: 83650
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] 26.20109
```

```
##
## Call:
## glm.nb(formula = INCID18 ~ C0532 + C0534 + C0536 + C0538 + C0560 +
##     C0562 + C0568 + FR_LVEL + FR_URBAN + FR_SIZE, data = data,
##     init.theta = 0.870028887, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9071  -1.0012  -0.3730   0.2375   5.1309
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.1436265  0.3044649   7.041 1.91e-12 ***
## C0532        0.0076216  0.0013674   5.574 2.49e-08 ***
## C0534       -0.0033694  0.0012361  -2.726 0.006411 **
## C0536       -0.0061397  0.0013126  -4.678 2.90e-06 ***
## C0538        0.0790182  0.0095209   8.299  < 2e-16 ***
## C0560       -0.0728909  0.0345555  -2.109 0.034911 *
## C0562       -0.2495463  0.0484126  -5.155 2.54e-07 ***
## C0568       -0.0002826  0.0028299  -0.100 0.920441
## FR_LVEL      0.2027681  0.0273538   7.413 1.24e-13 ***
## FR_URBAN    -0.0845011  0.0221043  -3.823 0.000132 ***
## FR_SIZE      0.5689275  0.0245323  23.191  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.87) family taken to be 1)
##
##     Null deviance: 4392.6  on 2761  degrees of freedom
## Residual deviance: 3256.2  on 2751  degrees of freedom
## AIC: 22767
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.8700
##           Std. Err.:  0.0241
##
##  2 x log-likelihood:  -22743.0550
```

```
##      C0532            C0534            C0536            C0538
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  7.00   1st Qu.: 50.00   1st Qu.: 60.00   1st Qu.: 4.000
##  Median : 12.00   Median : 68.00   Median : 75.00   Median : 6.000
##  Mean   : 18.26   Mean   : 62.69   Mean   : 71.04   Mean   : 5.928
##  3rd Qu.: 25.00   3rd Qu.: 80.00   3rd Qu.: 90.00   3rd Qu.: 8.000
##  Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :18.000
```

```
##      C0560           C0562           C0568            FR_LVEL         FR_URBAN
##  Min.   :1.00   Min.   :1.00   Min.   :  0.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.00   1st Qu.:2.00   1st Qu.: 92.00   1st Qu.:2.000   1st Qu.:1.000
##  Median :3.00   Median :3.00   Median : 95.00   Median :2.000   Median :2.000
##  Mean   :2.76   Mean   :2.68   Mean   : 93.16   Mean   :2.204   Mean   :2.328
##  3rd Qu.:3.00   3rd Qu.:3.00   3rd Qu.: 96.00   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.00   Max.   :3.00   Max.   :100.00   Max.   :4.000   Max.   :4.000
##     FR_SIZE
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :2.874
##  3rd Qu.:4.000
##  Max.   :4.000
##
## Call:
## glm.nb(formula = INCID18 ~ C0532 + C0534 + I(C0532 * C0534) +
##     I(C0534 * C0536) + C0536 + C0538 + C0560 + C0562 + I(C0560 *
##     C0562) + C0568 + FR_LVEL + FR_URBAN + I(FR_SIZE^2), data = data,
##     init.theta = 0.8806166626, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8647  -0.9916  -0.3718   0.2366   5.1497
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.721e+00  3.964e-01   6.865 6.67e-12 ***
## C0532             2.880e-03  2.755e-03   1.046 0.295759
## C0534             6.106e-03  3.318e-03   1.840 0.065764 .
## I(C0532 * C0534)  9.605e-05  4.499e-05   2.135 0.032792 *
## I(C0534 * C0536) -1.589e-04  3.725e-05  -4.265 2.00e-05 ***
## C0536             1.272e-03  2.233e-03   0.570 0.568782
## C0538             7.819e-02  9.485e-03   8.244  < 2e-16 ***
## C0560            -1.078e-01  1.077e-01  -1.001 0.316870
## C0562            -2.754e-01  1.004e-01  -2.742 0.006103 **
## I(C0560 * C0562)  1.330e-02  4.132e-02   0.322 0.747516
## C0568            -1.586e-03  2.816e-03  -0.563 0.573253
## FR_LVEL           1.550e-01  2.783e-02   5.569 2.55e-08 ***
## FR_URBAN         -8.465e-02  2.204e-02  -3.840 0.000123 ***
## I(FR_SIZE^2)      1.072e-01  4.579e-03  23.406  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8806) family taken to be 1)
##
##     Null deviance: 4438.5  on 2761  degrees of freedom
## Residual deviance: 3255.4  on 2748  degrees of freedom
## AIC: 22740
##
## Number of Fisher Scoring iterations: 1
##
##
```

```
##               Theta:  0.8806
##           Std. Err.:  0.0244
##
##  2 x log-likelihood:  -22709.7100
```

2. Create a new model that includes only those covariates that were statistically significant in (1), further refining this until all covariates in this model are statistically significant. How well does this model perfrom relative to Model (1)?

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: INCID18
##                                                                                                                  Model
## 1                    C0532 + C0534 + C0536 + C0538 + C0560 + C0562 + FR_LVEL + FR_URBAN + FR_SIZE
## 2 C0532 + C0534 + C0536 + I(C0534 * C0536) + C0538 + C0560 + C0562 + FR_LVEL + FR_URBAN + FR_SIZE
##       theta Resid. df    2 x log-lik.   Test    df LR stat.      Pr(Chi)
## 1 0.8700275      2752       -22743.06
## 2 0.8765357      2751       -22721.80 1 vs 2     1 21.26789 3.993664e-06
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: INCID18
##                                                                                                                  Model
## 1         C0532 + C0534 + C0536 + C0538 + C0562 + I(C0534 * C0536) + FR_LVEL + FR_URBAN + FR_SIZE
## 2 C0532 + C0534 + C0536 + I(C0534 * C0536) + C0538 + C0560 + C0562 + FR_LVEL + FR_URBAN + FR_SIZE
##       theta Resid. df    2 x log-lik.   Test    df LR stat.    Pr(Chi)
## 1 0.8748021      2752       -22727.64
## 2 0.8765357      2751       -22721.80 1 vs 2     1 5.843377 0.01563579
```

```
##
## Call:
## glm.nb(formula = INCID18 ~ C0532 + C0534 + C0536 + I(C0534 *
##     C0536) + C0538 + C0560 + C0562 + FR_LVEL + FR_URBAN + FR_SIZE,
##     data = data, init.theta = 0.8765357457, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8830  -0.9982  -0.3700   0.2301   5.1427
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.604e+00  1.943e-01   8.255  < 2e-16 ***
## C0532               8.061e-03  1.354e-03   5.952 2.66e-09 ***
## C0534               9.497e-03  2.889e-03   3.287  0.00101 **
## C0536               2.012e-03  2.174e-03   0.925  0.35477
## I(C0534 * C0536) -1.749e-04  3.586e-05  -4.876 1.08e-06 ***
## C0538               7.735e-02  9.503e-03   8.139 3.99e-16 ***
## C0560              -7.717e-02  3.444e-02  -2.241  0.02504 *
## C0562              -2.495e-01  4.826e-02  -5.170 2.34e-07 ***
## FR_LVEL             2.016e-01  2.709e-02   7.442 9.94e-14 ***
## FR_URBAN           -8.596e-02  2.204e-02  -3.900 9.62e-05 ***
## FR_SIZE             5.641e-01  2.443e-02  23.094  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8765) family taken to be 1)
```

```
##
##     Null deviance: 4420.8  on 2761  degrees of freedom
## Residual deviance: 3255.0  on 2751  degrees of freedom
## AIC: 22746
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.8765
##           Std. Err.:  0.0243
##
##  2 x log-likelihood:  -22721.7950
```

The model was modified by removing the predictor variable C0568 and adding the interaction term between C0534 and C0536.

3. Develop and implement an approach to build the best model possible that predicts the total number of crimes (incidents, `INCID18`). (In addition to the variables mentioned in the previous problem, you may consider other variables, but be sure to explain your thinking.)

   What is your final model and why do you think it is the best? Be sure to clearly explain your approach in language a non-statistician could understand.

4. Does your final model do a good job in predicting crime? Explain to a policymaker if and how they should properly use this model.