¶ Pandemic Era Policing: A Model of Felony Arrests in California

Sta 223: Final Project

Richard Paul Yim

15 March 2023

1 Introduction

Various news outlets have been reporting an increase in crime across the nation [1,2,3]. California in particular represents a large portion of these crimes. With soft-on-crime policies naturally favored by the democratic-majority state government, and the Defund the Police Movement of recent years due to public outrage due to events such as the George Floyd incident, law enforcement agencies have developed poor public image. Alarmingly, 2021 marks the first year where there has been a decrease in law enforcement personnel in police and sheriff departments across California, dropping to levels before 2017 [4]. With all of these recent trends a natural question is to ask how felony arrests are being made in recent years by law enforcement in California on the basis of social factors. We study this question with the generalized linear model.

The rest of the project is outlined as follows: Section 2 introduces and discusses the data; Section 3 formally poses the question of the study and develops a formal model of the data; Section 4 concludes the project with discussion on implications of the model, limitiations, and practical considerations with respect to policy.

2 Data

This section introduces the data and background as well as necessary data preprocessing and formulation of a design matrix purposed for a generalized linear model.

2.1 Background

The State of California provides well-documented data collected on various topics such as various crimes committed and frequencies of crimes, deaths in custody, hate crimes and more. For this study, we look at arrests made in California provided by the Open Justice Program, "a transparency initiative led by the California Department of Justice. [5]" Though the original data ranges from 1985 to 2021, we limit our study to recent years from 2019 to 2021, where 2019 marks a whole year before the start of the COVID-19 pandemic and 2021 is the earliest year where the pandemic was considered to slow at least in California.

The data presents counts of arrests made by various types crimes such as violent felonies, property felonies, and various types of misdemeanors. These arrest counts are presented against attributes of the race, with values Black, hispanic, White or other; sex, binary as male or female; age group, split as under 18, 18 to 19, 20 to 29, 30 to 39, 40 to 69, and 70 and over; and year, from 2019 to 2021. (The full characterization of the original dataset is provided in a data dictionary on the Open Justice website as well as legal definitions for different types of felonies and misdemeanors [6].)

2.2 Encoding

The original dataset is very large and offers various levels of granularity with respect to types of crimes, county, and more. This needs to be simplified for a more parsimonious model of arrests made in California. Therefore, in our study we make three primary transformations:

1. Arguably, there is some unnecessary complexity in the age group classification. The original dataset splits age group into six groups as presented in the previous section. For our study, we pair adjacent age groups; for example, age groups "18 and under" is paired with "18 to 19" as one group "19 and under." This will reduce our model complexity in later sections.

2. For the purpose of our study, we are interested in arrests by crime in general. The level of granularity of types of crimes by arrest is not of interest since the question of the study is focused on sociological phenomena rather than economic considerations (property felony versus violent felony versus other felonies). Generally, we focus on aggregated counts of individuals and their social and ethnic backgrounds, and in particular we only study total felonies by arrest as our dependent variable.
3. There is a natural difficulty with interpreting simply count relationships, so in our final model, what is presented is a model showing propensities of felony likelihoods as a function of our predictors (age group, race, gender), as opposed to count models—though we present an argument and comparative analysis on count regression in the next section. To be precise, we uniformly bin count values as low, medium and high across age group, race and age from 2019 to 2021.

Our final dataset consists of $n = 833,896$ individuals aggregated into 144 groups by year, age, gender and race. With these adjustments in our data, the next section will formally introduce the generalized linear model and the formal research question.

3 Modeling

This section formulates the research question and develops a model. Additionally, we provide comparative discussion on other models and analyses.

3.1 Research Question

In the context of our study we are interested in the interaction between civilians and law enforcement in California in recent years. Naturally, we ask

> What are the likelihoods for propensities of felony arrests by individuals on the basis of race, gender and age in recent years?

In particular, as described in the last section our response variable Y is a factor variable with levels low, medium and high, of which there are 144 aggregate groups such that there are 48 groups each with responses of low, medium and high propensities of arrest (again, with groups being an intersection of year, gender, race and age group). The "low" felony arrest range is valued from 16 to 770; "medium" felony arrest range is valued from 771 to 4539; and "high" felony arrest range is valued from 4540 to 38747 felony arrests.

3.2 Multinomial Regression

With the above research question, we introduce the generalized linear model (GLM): GLMs are powerful statistical models with useful asymptotic properties. They can be used to model expected response values of some dependent variable Y under a *link* function, g, through a linear relationship of predictors, X, or

$$g(\mathbb{E}(Y|\mathbf{X})) = \mathbf{X}\beta.$$

For our purposes, since we are modeling classes of low, medium and high felony arrest counts, appropriately, we use a multinomial regression GLM. In particular, the multinomial regression model that used is the final model is the *proportional odds model*, with a multinomial regression model with response categories coded as $j = 1, 2, \ldots, M$ with a response $z_{im} = \sum_{j=1}^{m} y_i j$, which is equal to 1 for $j \leq m$, and zero otherwise. In particular, the GLM under the proportional odds model is

$$\mathrm{g}(\mu_{im}) = \beta_{0m} + X_i\beta,$$

where we have that $\mu_{im} = E(z_{im})$, and we have ordered intercepts $\beta_{01} \leq \beta_{02} \leq \cdots \leq \beta_{0,M-1}$, and $\beta \in \mathbb{R}^{p-1}$ a parameter vector acting on our design matrix X. In our study, our link function g is the logit link, our we have three classes with $M - 1 = 2$; and our design matrix $\mathbf{X}$ for our systematic component is

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1}_{\text{Male}} & \mathbf{1}_{\text{Hispanic}} & \mathbf{1}_{\text{Other}} & \mathbf{1}_{\text{White}} & \mathbf{1}_{\text{20 to 39}} & \mathbf{1}_{\text{40 and over}} \end{bmatrix},$$

with $\mathbf{X} \in \mathbb{R}^{144\times7}$, and $\beta \in \mathbb{R}^7$, such that the ith entry of β corresponds to the coefficient of the ith covariate of the design matrix, $\mathbf{X}$.

This final model is produced by starting from a main effects model including the year covariate, and applying stepwise regression with a Bayesian Information Criterion (BIC) score. The BIC is used because it produces the most parsimonious model by heavily penalizing models with greater model complexity. More precisely, for our stepwise regression procedure, starting from the first order main effects model, we performed stepwise regression in two directions against a larger second order interaction model and a smaller null model with just the intercept. Consequently, our model gives a minimal of BIC 225.018, where the exclusion of the `YEAR` covariate is found to reduce our model complexity. With this final model, we present the an anlysis on model coefficients and commentary on the relationship between covariates and the response in the conclusion section.

Finally, with this model, the fitted values that are provided from this proportional odds model are actual probabilities of a given data point fall into each class such that for each class $j = 1, 2, \ldots, m$, we are provided with probabilities that an aggregate group falls within a certain propensity level for felony arrests. This is a noteworthy benefit of using the proportional odds model in the case where we have an ordinal response variable. (Table 2 shows the coefficient estimates and standard errors.)

3.3 Goodness of Fit

Briefly on the goodness of fit, we find that our proportional odds model has fairly good fit. According to Figure 1, in a plot of the Pearson residuals for classes 1 against fitted values, as well as classes 1 and 2 against fitted values, we find that between outcomes of low, medium and high probabilities of felony arrests by aggregated groups of individuals, the residual relationship is linear and zero valued. This lack of relationship indicates that there is no real discrepancy of our fitted values against the true values of our model indicating that our model indeed represents a sufficiently accurate relationship between probabilities, or propensities, of felony arrest by the covariates in our design matrix, $\mathbf{X}$, through this multinomial regression GLM model. In addition, we directly report that the residual deviance of our final model was minimal at a value of 209.018, all on just a model with main effects and no interactions—this is evident in Figure 1.

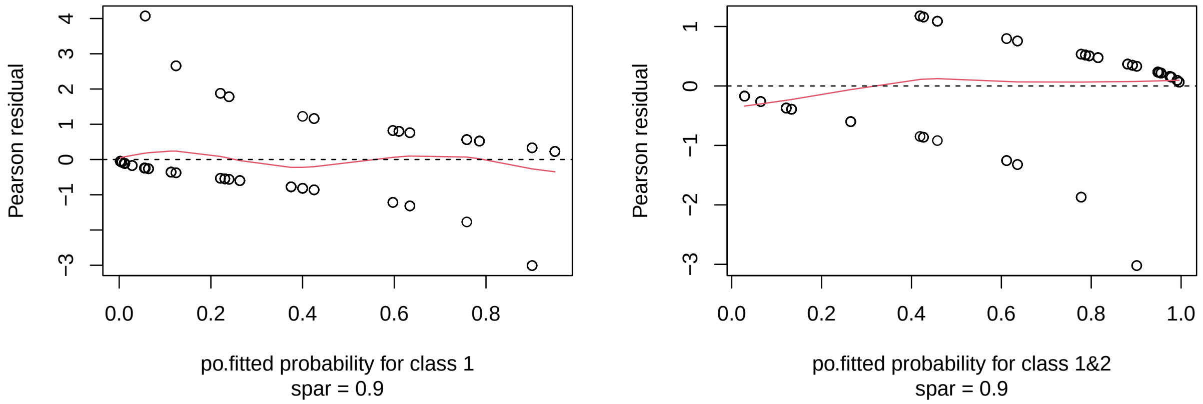

Figure 1: Goodness-of-fit plots with Pearson residuals by fitted probabilities across low propensity and low-medium proppensities. Fairly zero-value linear relationship indicates good fit.

3.4 Baseline Odds Model

Since we are performing multinomial regression, a complementary multinomial regression model is the *baseline odds model.* As opposed to modeling the expected values, $\mu_{im} = E(z_{im})$ as described for the proportional odds model, we can also model relative responses of each class to a baseline class such that we can observe pairwise relationships of categories of propensities for felony arrests.

In particular, for our data, we let the baseline category/class be the low propensity for felony arrest label, and the linear predictor in our baseline odds model gives us the following relationships between two sub models as

$$\frac{\pi_{i2}}{\pi_{i1}} = \exp(\eta_{i2}),\ \eta_{i2} = X_i\beta_2 \ \text{ and } \ \frac{\pi_{i3}}{\pi_{i1}} = \exp(\eta_{i3}),\ \eta_{i3} = X_i\beta_3,$$

where the β_j are parameter vectors relating our linear predictor through likelihoods odds ratio π_{ij}/π_{i1}, for $j = 1, 2$; π_{ij} denotes the probability that data point i is of category j. For the baseline odds model we consider end up with the following design matrix:

$$\mathbf{X}' = \begin{bmatrix}\mathbf{X} & \mathbf{1}_{\text{Male and Hispanic}} & \mathbf{1}_{\text{Male and Other}} & \mathbf{1}_{\text{Male and White}}\end{bmatrix},$$

where $\mathbf{X}$ corresponds to the set of columns from the design matrix of our proportional odds model. A small nuance to note is that our baseline odds model is really a set of two sub models where we have pairwise relationships to the baseline category, low propensity for felony arrest, against the other categories, low and high propensity, respectively. Clearly, we see that our design matrix is of 144 rows and 10 columns, where 7 columns are columns of the proportional odds design matrix, and 3 columns are appended as interaction columns of gender and race. Again, we use the bidirectional stepwise regression starting from just a first order main effects model with lower scope being a null model, and upper scope being a full second order interaction model. For the same reasons as before, we use the BIC score to produce the smallest more parsimonious models possible, where our final model gives us a BIC score of 20 for the baseline sub model between the low-medium propensities with residual deviance 5.5695`-e10`, and a BIC score of 62.265 for the sub model between the low-high propensities with residual deviance 42.265.

Furthermore, we are able to study outlier values for the baseline odds model by looking at *leverage* values and *Cook's distances* for data points across both sub models. Leverage values essentially tell us how the exclusion of data points perturbs values in the fitted parameter vectors. Consequently, data points with high leverage values have high effect on parameter stability, where heuristically, "high leverage" is determined to be any value that is greater than $2\frac{p}{n}$, where n is the number of observations, which is 144 in this case, and p is the number of parameters, $p = 10$ in this case. In conjunction with leverage values, Cook's distance effectively measures the residual error of a data point adjusted by its leverage and other factors, such that a data point with high leverage **and** high Cook's distance generally has a real impact on stability of model parameters. Figures 3 and 4 in the Appendix correspond to the sub models of our baseline odds model, and corresponding outliers. When we remove these outliers, and refit the baseline odds model and regression, we find we have very good fit, again by the same goodness-of-fit criterion of residuals against fitted values: flat zero-value relationship. Figure 2, also in the Appendix, shows the goodness-of-fit of the two marginal models.

(Finally, we report confidence intervals of coefficients resulting from the sub models in Table 3 and 4, and leave commentary on the implications of these tables in the conclusion section.)

3.5 Alternative Regressions

Originally the response variable presented itself as count data. Count data is naturally modeled with Poisson regression. However, a problem with count distributed response variables is the concern for overdispersion and underdispersion, where empirically, the estimated count variance is not one-to-one with the expected count. When this property fails to hold, we say that the data either has overdispersion or underdispersion, in which case the resulting p-values from a count regression model may be biased and inaccurate.

This may be remedied with negative binomial regression, another type of regression that can model count data. However, experimentally, even after fitting a negative binomial model, we estimated sever underdispersion with the negative binomial model (the data matrix for this regression was third order interaction model, which provided the best fit). (Evidence of these experiments can be found in the code section of the Appendix.)

Therefore, a possible remedy for this problem is to effectively avoid the problem altogether with multinomial ordinal regression by proportional odds model with the added benefit of greater interpretability with the usage of ordinal classification labeling of the count values into categories of low, medium and high counts.

4 Conclusion

In this section we directly interpret the results of our proportional odds model. We also discuss limitations and other considerations of the model and the data in the context of our research question.

4.1 Interpretation

From our proportional odds model we are left with Table 1 providing 95% confidence intervals and conclude the following:

1. Men are **more** likely to have a greater propensity of felony arrests compared to females. This is evident by the 95% confidence interval range being greater than 0.
2. Hispanics are **more** likely to have a greater propensity of felony arrests compared to Blacks. This is evident by the 95% confidence interval range being greater than 0.
3. Other races are **less** likely to have a greater propensity of felony arrests compared to Blacks. This is evident by the 95% confidence interval range being less than 0.
4. Adults aged 20 to 39 are **more** likely to have a greater propensity of felony arrests compared to individuals ages 19 and under. This is evident by the 95% confidence interval range being greater than 0.

Interestingly, we can conclude also that between Blacks and Whites the likelihood for greater propensities for felony arrests are not significantly different; and between individuals aged 40 and over and individuals aged 19 and under, the likelihood for propensities for felony arrests are not significantly different either.

Table 1: 95% confidence intervals of β coefficients from proportional odds model.

	2.5 %	97.5 %
GENDERMale	0.88	2.48
RACEHispanic	0.52	2.63
RACEOther	-2.91	-0.78
RACEWhite	-0.31	1.73
AGE_GROUP20 to 39	2.94	5.14
AGE_GROUP40 and over	-0.13	1.65

4.2 Limitations and Other Considerations

Our research question was

> What are the likelihoods of felony arrests of individuals on the basis of race, gender and age in recent years?

We found that in recent years our best model shows no real reliance on the particular year within the last three years such that these likelihoods for propensities for felony arrests have not changed much. Furthermore, from our interpretations above we find that despite the arguably heavier media coverage of police brutality on Blacks, Hispanics have greater likelihoods for higher propensities of felony arrest. With this in mind it may be interesting to further study incidences of police brutality within the Latino community in California.

A further limitation of this study is that the types of felonies were not distinguished. The economic impact of property felonies may be greater than of violent felonies, or vice versa. Studying propensities for different *types* of felonies may provide insight for motivations of felony offenders dependent on sociological factors.

Finally, a missed opportunity is the incorporation of historic median incomes, or other economic measures, of each of the counties from the original arrest dataset. County data of these aggregates were essentially ignored, and could have provided more than useful insight into likelihoods for felony arrests dependent on the economic activity of a felony offender's place of arrest. This would certainly lead to a more holistic profile of the felony offender in California.

5 Appendix

5.1 References

1. https://slate.com/business/2023/03/rising-crime-broken-windows-policing-revisited.html
2. https://www.oregonlive.com/news/2023/03/alarming-rise-in-hate-crimes-in-us-oregon-in-2021-fbi-stats-show.html
3. https://www.planetizen.com/blogs/122113-red-cities-blue-cities-and-crime
4. https://data-openjustice.doj.ca.gov/sites/default/files/dataset/2022-08/Law%20Enforcement%20and%20Criminal%20Justice%20Personnel%20Context_081122.pdf

5.2 Tables and Figure

Table 2: Table of final model.

	Value	Std. Error	t value
GENDERMale	1.65	0.41	4.07
RACEHispanic	1.55	0.53	2.89
RACEOther	-1.81	0.54	-3.36
RACEWhite	0.69	0.52	1.34
AGE_GROUP20 to 39	3.98	0.56	7.12
AGE_GROUP40 and over	0.75	0.45	1.66
1|2	1.14	0.48	2.36
2|3	3.66	0.59	6.21

Table 3: 95% confidence intervals from submodel for baseline odds model between low and medium felony arrest.

	2.5 %	97.5 %
(Intercept)	-1.74	0.86
GENDERMale	-0.49	3.20
RACEHispanic	-0.30	3.12
RACEOther	-2.20	1.38
RACEWhite	-43.51	-41.54
AGE_GROUP20 to 39	-1019.04	1075.90
AGE_GROUP40 and over	-1.57	0.51
GENDERMale:RACEHispanic	-32.59	-32.59
GENDERMale:RACEOther	-3.98	1.27
GENDERMale:RACEWhite	41.54	43.51

Table 4: 95% confidence intervals from submodel for baseline odds model between low and high felony arrest.

	2.5 %	97.5 %
(Intercept)	-563.14	444.49
GENDERMale	-445.44	562.18
RACEHispanic	-964.88	1049.48
RACEOther	-25.10	-12.42
RACEWhite	-446.45	561.17
AGE_GROUP20 to 39	-983.86	1111.04
AGE_GROUP40 and over	-0.08	3.00
GENDERMale:RACEHispanic	-1048.45	965.91
GENDERMale:RACEOther	-44.18	-31.50
GENDERMale:RACEWhite	-561.17	446.45

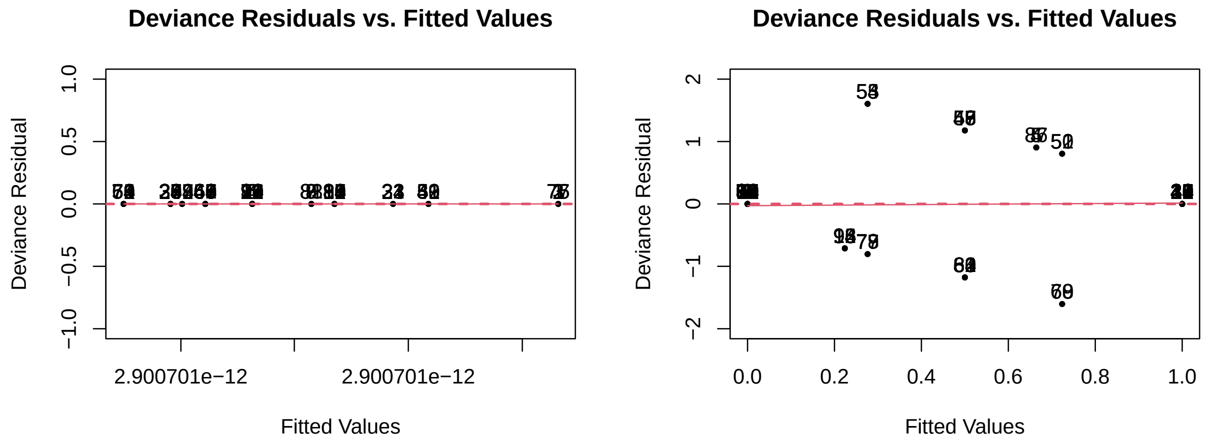

Figure 2: Goodness of fit plot of Pearson residuals against fitted values.

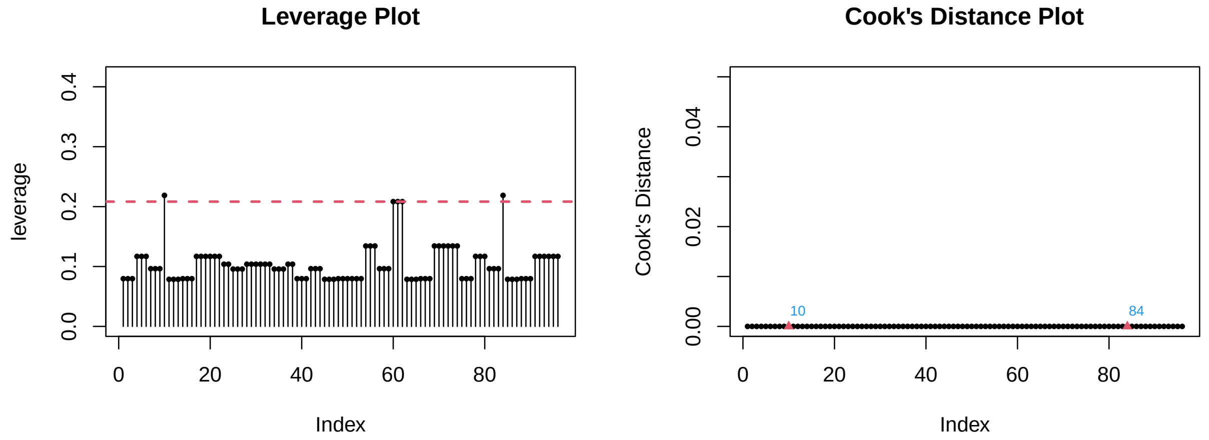

Figure 3: Outlier plots from leverage analysis and Cook's distance for the sub model of the baseline odds model; these plots correspond to the low-medium propensity sub model.

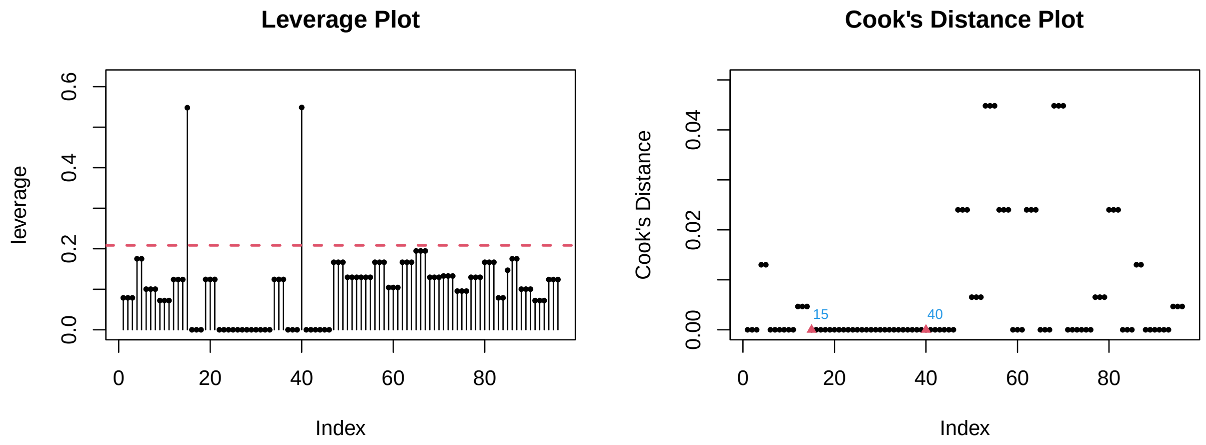

Figure 4: Outlier plots from leverage analysis and Cook's distance for the sub model of the baseline odds model; these plots correspond to the low-high propensity sub model.

5.3 Code

```
knitr::opts_chunk$set(echo = TRUE)
# ==============================================================================
# Law enforcement decrease
# ==============================================================================
lep <- read.csv('Law_Enforcement_Personnel_1991-2021.csv')
x <- aggregate(FUNDED_NON_JAIL_SWORN_TOTAL~YEAR,lep, sum)
# plot(x$YEAR, x$FUNDED_NON_JAIL_SWORN_TOTAL)
# =============================================================================
# arrests data
# =============================================================================
library(XML)
library(RCurl)
library(rlist)
library(stringr)

arrstDATA <- read.csv("OnlineArrestData1980-2021.csv")
# county collapse
arrstDATA <- aggregate(cbind(VIOLENT, PROPERTY, F_DRUGOFF, F_SEXOFF,
                             F_ALLOTHER, F_TOTAL, M_TOTAL, S_TOTAL)~
                         YEAR+GENDER+RACE+AGE_GROUP, data=arrstDATA, sum)
NONVIOLENT <- arrstDATA$F_TOTAL-arrstDATA$VIOLENT
arrstDATA$NONVIOLENT <- NONVIOLENT
arrstDATA$RACE <-factor(arrstDATA$RACE)

# =============================================================================
# final arrest data
# =============================================================================
arrstFINAL <- arrstDATA[arrstDATA$YEAR %in% 2019:2021, c('YEAR', "GENDER",
                                                         "RACE","AGE_GROUP", "F_TOTAL")]
arrstFINAL$YEAR <- factor(arrstFINAL$YEAR)
arrstFINAL$GENDER<- factor(arrstFINAL$GENDER)
```

```
arrstFINAL$AGE_GROUP[arrstFINAL$AGE_GROUP %in% c("18 to 19", "Under 18")] <- "under 19"
arrstFINAL$AGE_GROUP[arrstFINAL$AGE_GROUP %in% c("30 to 39", "20 to 29")] <- "20 to 39"
arrstFINAL$AGE_GROUP[arrstFINAL$AGE_GROUP %in% c("70 and over", "40 to 69")] <-
  "40 and over"

arrstFINAL$AGE_GROUP<- factor(arrstFINAL$AGE_GROUP, levels=c("under 19", "20 to 39",
                                                             "40 and over"))
#set your own path
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
source('clean.R')
library(MASS)
library(tidyverse)

# ==============================================================================
# Ordinal Label
# ==============================================================================
arrstFINALm <- arrstFINAL
x <- arrstFINALm$F_TOTAL
rownames(arrstFINALm) <- NULL
arrstFINALm$F_TOTAL <- factor(ntile(arrstFINALm$F_TOTAL, n=3))
# 770, 4539
# ==============================================================================
# Model Selection with BIC (POLR)
# ==============================================================================
glmProp <- polr(formula = F_TOTAL ~., data = arrstFINALm, method='logistic')
modelProp.null <- polr(F_TOTAL~1, data=arrstFINALm, method='logistic')
modelProp.full <- polr(F_TOTAL~.^2, data=arrstFINALm, method='logistic')
finalProp.glm <- stepAIC(glmProp, direction="both",
                         scope=list(lower=modelProp.null, upper=modelProp.full),
                         k=log(dim(arrstFINALm)[1]))
summary(finalProp.glm)
library(xtable)
xtable(print(confint(finalProp.glm)))
xtable(print(summary(finalProp.glm)$coefficients))
# ==============================================================================
# Goodness-of-fit check (POLR)
# ==============================================================================
# Pearson residuals
obslabel = cbind(arrstFINALm$F_TOTAL==1, arrstFINALm$F_TOTAL==2,
                 arrstFINALm$F_TOTAL==3) * 1
prd_prob_po = fitted(finalProp.glm)
rep.row<-function(x,n){
  matrix(rep(x,each=n),nrow=n)
}
prd_prob_po[(prd_prob_po == 1)[,1],] <- rep.row(c(0.9998, 0.0001,0.0001),
                                                sum(prd_prob_po==1))

resP.plr <- sapply(1:(ncol(obslabel)-1), function(m) {
  obs_m <- rowSums(as.matrix(obslabel[,1:m]))
  fit_m <- rowSums(as.matrix(prd_prob_po[,1:m]))
  (obs_m - fit_m) / sqrt(fit_m * (1 - fit_m))
})
fitted_m = sapply(1:(ncol(obslabel)-1), function(m){
```

```r
  rowSums(as.matrix(prd_prob_po[,1:m]))
})

# plot
par(mfrow=c(1,2))
plot(fitted_m[,1], resP.plr[,1], xlab="po.fitted probability for class 1",
     ylab="Pearson residual",
     sub = "spar = 0.9")
lines(smooth.spline(x = fitted_m[,1], y = resP.plr[,1], spar=0.9), col=2);
abline(h=0,lty=2)
plot(fitted_m[,2], resP.plr[,2], xlab="po.fitted probability for class 1&2",
     ylab="Pearson residual",
     sub = "spar = 0.9")
lines(smooth.spline(x = fitted_m[,2], y = resP.plr[,2], spar=0.9), col=2);
abline(h=0,lty=2)
# ============================================================================
# Model Selection with BIC (BASELINE)
# ============================================================================
library(nnet)
glmBase <- multinom(formula = F_TOTAL ~., data = arrstFINALm)
modelBase.null <- multinom(F_TOTAL~1, data=arrstFINALm)
modelBase.full <- multinom(F_TOTAL~.^2, data=arrstFINALm)
finalBase.glm <- stepAIC(glmBase, direction="both",
                         scope=list(lower=modelBase.null, upper=modelBase.full),
                         k=log(dim(arrstFINALm)[1]))

## z values
#zval.bo <- coef(inal.bo) / summary(housing.bo)$standard.errors
## two-sided p-values
#pval.bo <- 2 * pnorm(abs(zval.bo), lower.tail=FALSE)

# ============================================================================
# Outlier Check (BASELINE)
# ============================================================================
# compare with fitting 2 logistic regression models
margDat1 <- arrstFINALm[arrstFINALm$F_TOTAL %in% c(1,2),]
margDat1$F_TOTAL <- factor(margDat1$F_TOTAL==2)
margin_1 = glm(F_TOTAL ~ GENDER+RACE+AGE_GROUP+GENDER:RACE,
               data = margDat1, family = binomial())

margDat2 <- arrstFINALm[arrstFINALm$F_TOTAL %in% c(1,3),]
margDat2$F_TOTAL <- factor(margDat2$F_TOTAL==3)
margin_2 = glm(F_TOTAL ~ GENDER+RACE+AGE_GROUP+GENDER:RACE,
               data = margDat2, family = binomial())

summary(margin_1)
summary(margin_2)

# refitting multinomial model
# ============================================================================
# Model Selection with BIC - Refitted (BASELINE)
# ============================================================================
glmBase <- multinom(formula = F_TOTAL ~., data = arrstFINALm[-c(12, 25, 64, 132),])
```

```r
modelBase.null <- multinom(F_TOTAL~1, data=arrstFINALm[-c(12, 25, 64, 132),])
modelBase.full <- multinom(F_TOTAL~.^2, data=arrstFINALm[-c(12, 25, 64, 132),])
finalBase.glm2 <- stepAIC(glmBase, direction="both",
                         scope=list(lower=modelBase.null, upper=modelBase.full),
                         k=log(dim(arrstFINALm)[1]))

# ================================================================================
# Goodness-of-fit check (BASELINE)
# ================================================================================
arrstFINALm <- arrstFINALm[-c(12,25,64,132)]
# compare with fitting 2 logistic regression models
margDat1_lev <- arrstFINALm[arrstFINALm$F_TOTAL %in% c(1,2),]
margDat1$F_TOTAL <- factor(margDat1$F_TOTAL==2)
margin_1 = glm(F_TOTAL ~ GENDER+RACE+AGE_GROUP+GENDER:RACE,
               data = margDat1, family = binomial())

margDat2_lev <- arrstFINALm[arrstFINALm$F_TOTAL %in% c(1,3),]
margDat2 <- arrstFINALm[arrstFINALm$F_TOTAL %in% c(1,3),]
margDat2$F_TOTAL <- factor(margDat2$F_TOTAL==3)
margin_2 = glm(F_TOTAL ~ GENDER+RACE+AGE_GROUP+GENDER:RACE,
               data = margDat2, family = binomial())

summary(margin_1)
summary(margin_2)

# margin_1 and margin_2 model
res.D1 = residuals(margin_1,type="deviance")
res.D2 = residuals(margin_2,type="deviance")

# Deviance X fitted
par(mfrow=c(1,2))
plot(margin_1$fitted, res.D1, xlab="Fitted Values",ylab="Deviance Residual",
     ylim = c(-1,1), pch = 19, cex = 0.5)
lines(smooth.spline(margin_1$fitted.values, res.D1, spar=0.8), col=2)
text(margin_1$fitted, res.D1+0.1, labels = seq(length(res.D1)))
abline(h=0,col=2,lwd=2,lty=2)
title("Deviance Residuals vs. Fitted Values")
plot(margin_2$fitted, res.D2, xlab="Fitted Values",ylab="Deviance Residual",
     ylim = c(-2,2), pch = 19, cex = 0.5)
lines(smooth.spline(margin_2$fitted.values, res.D2, spar=0.8), col=2)
text(margin_2$fitted, res.D2+0.2, labels = seq(length(res.D2)))
abline(h=0,col=2,lwd=2,lty=2)
title("Deviance Residuals vs. Fitted Values")

# confidence intervals
xtable(print(confint(finalBase.glm2)[,,1]))
xtable(print(confint(finalBase.glm2)[,,2]))
# ================================================================================
# Margin1 Model
# ================================================================================
# plotting
par(mfrow=c(1,2))
# leverage points
```

```
leverage <- hatvalues(margin_1)
names(leverage) <- 1:length(leverage)
W <- diag(margin_1$weights)
X <- model.matrix(margin_1)
Hat = sqrt(W) %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% sqrt(W)
all(abs(leverage - diag(Hat)) < 1e-6)
# indices 10 and 84, points 12 and 132

p <- length(coef(margin_1))
n <- nrow(margDat1)
plot(names(leverage), leverage, xlab="Index", type="h", ylim = c(0,2*p/n*2))
points(names(leverage), leverage, pch=16, cex=0.6)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)
title("Leverage Plot")

# Cook's distance
res.P = residuals(margin_1,type="pearson")
cooks <- cooks.distance(margin_1)
names(cooks) <- names(leverage)
plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6,ylim=c(0,0.05))
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2)
susPts <- as.numeric(names(sort(cooks[infPts], decreasing=TRUE)[1:2]))
text(susPts, cooks[susPts]+0.002, susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)
dispersion <- 1
all(abs(cooks - (res.P/(1 - leverage))^2 * leverage/(dispersion * p) < 1e-15))
title("Cook's Distance Plot")


# ==============================================================================
# Margin2 Model
# ==============================================================================
# plotting
par(mfrow=c(1,2))
# leverage points
leverage <- hatvalues(margin_2)
names(leverage) <- 1:length(leverage)
W <- diag(margin_2$weights)
X <- model.matrix(margin_2)
Hat = sqrt(W) %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% sqrt(W)
all(abs(leverage - diag(Hat)) < 1e-6)
# indices 15 and 40, points 25 and 64

p <- length(coef(margin_2))
n <- nrow(margDat2)
plot(names(leverage), leverage, xlab="Index", type="h", ylim = c(0,2*p/n*2+0.2))
points(names(leverage), leverage, pch=16, cex=0.6)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)
title("Leverage Plot")

# Cook's distance
res.P = residuals(margin_2,type="pearson")
```

```r
cooks <- cooks.distance(margin_2)
names(cooks) <- names(leverage)
plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6,ylim=c(0,0.05))
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2)
susPts <- as.numeric(names(sort(cooks[infPts], decreasing=TRUE)[1:2]))
text(susPts, cooks[susPts]+0.002, susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)
dispersion <- 1
all(abs(cooks - (res.P/(1 - leverage))^2 * leverage/(dispersion * p) < 1e-15))
title("Cook's Distance Plot")
source('clean.R')
require(ggplot2)
require(GGally)
require(reshape2)
require(lme4)
require(compiler)
require(parallel)
require(boot)
require(lattice)

#=========================================================================
# Model selection
#=========================================================================
glmdat <- arrstFINAL[arrstFINAL$YEAR %in% 2018:2021,]
glmm <- glmer(F_TOTAL~GENDER + AGE_GROUP + RACE+(1 | YEAR), data = glmdat,
              family = "poisson")
summary(glmm)

glmFixed <- glm(F_TOTAL~(GENDER + AGE_GROUP + RACE + YEAR)^2, data = glmdat,
                family = poisson(link="log"))
summary(glmFixed)

anova(glmm,glmFixed, test="Chisq")

res.P = residuals(glmFixed,type="pearson")
res.D = residuals(glmFixed,type="deviance")

# ====================================================================
# Goodness-of-fit
# ====================================================================
res.P = residuals(glmFixed,type="pearson")
res.D = residuals(glmFixed,type="deviance")
par(mfrow=c(1,2))
if(0){
# Pearson X fitted
plot(glmFixed$fitted, res.P, xlab="Fitted Values",ylab="Pearson Residual",
     ylim = c(-10,10), pch = 19, cex = 0.5)
lines(smooth.spline(glmFixed$fitted.values, res.P, spar=1.2), col=2)
text(glmFixed$fitted, res.P+0.4, labels = seq(length(res.P)))
title("Pearson Residuals vs. Fitted Values")

# Deviance X fitted
plot(glmFixed$fitted, res.D, xlab="Fitted Values",ylab="Deviance Residual",
     ylim = c(-10,10), pch = 19, cex = 0.5)
lines(smooth.spline(glmFixed$fitted.values, res.D, spar=1.2), col=2)
text(glmFixed$fitted, res.D+0.4, labels = seq(length(res.D)))
title("Deviance Residuals vs. Fitted Values")}
```