Lab 3

W203 Statistics for Data Science

Section 6 - Team 1

The provided data needed cleaning. We need to omit the last rows of the csv since they do not contain data. When we check for duplicated rows, there was one county (193) which was repeated twice. We also checked that all of the values were numerical. This was note the case for the prbconv variable (it was stored as a factor due to omitted rows having non-numerical values). This was converted to numerical.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tibble)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
setwd("~/Desktop/Stone/Berkeley_MIDS/Statistics/Labs/Lab_3")
full_data <- read.csv('crime_v2.csv')
data <- na.omit(full_data)

#check for duplicated data and remove
sum(duplicated(data))
```

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

```
data <- distinct(data, .keep_all=T)

#convert prbconv factor in numeric
data$prbconv <- as.numeric(levels(data$prbconv))[data$prbconv]
```

```
## Warning: NAs introduced by coercion
```

```
#check that all fields are numerical
for (field in names(data)) {
  stopifnot(class(data[,field]) %in% c("numeric", "integer"))
  }
```

Next, our independent variable of interest is crime rate (crmrte). Our goal is to find the best possible causal predictors for crime rate. We first identify columns for which believe may be possible candidates.

County and year are identifiers and are not relevant to the study. West/Central/Urban are categorical and... #(WILL BE ADDRESSED LATER) All other variables are numeric and possible candidates.

First, we wanted to get a sense of crmrte correlation with all numeric variables. We also parse our data into X (numeric variables) and y (crmrte)

```
y <- data$crmrte
X <- data[,!names(data) %in% c('county',
                               'year',
                               'crmrte',
                               'west',
                               'central',
                               'urban')]


#correlate all variables and store in new dataframe
cor_df <- data.frame(variable = character(),
                     crmrte_cor = numeric())
for (x in names(X)) {
  crmrte_cor <- cor(y, data[,x])
  corr <- as.data.frame(crmrte_cor,
                        col.names = c('crmrte_cor')) %>%
                        add_column(variable = x, .before = 1)
  cor_df <- rbind(cor_df, corr)
}

cor_df <- arrange(cor_df, desc(crmrte_cor))
```

It should be no surprise that density is the best solo predictor of crime rate. Highly dense population areas present more opportunities for crime, as well as a larger population of poorer individuals. ETC. We believe this variable is problematic for a causal model however because it obsorbs the causality from other variables. It DOES suggest that policies in more densely populated areas are likely more fruitful than sparsely populated areas.

First, we would like to select the variables we think most likely cause crime.

The first is the probability variables of arrest, conviction and prison sentence. These variables constitute the "fear factors;" namely, we believe the higher the chance someone believes they will be arrested, convicted, or sent to prison, the less likely they will commit a crime. Also, the more severely they believe the punishment to be (prison day sentences), the less likely they will commit a crime. Note that convicted, and sent to prison, variables will be converted to ratio of offenses like arrested (currently, they are in terms of the previous variable).

The next is the wage variables. For blue collar jobs that are mostly manual labor, namely, construction, transportation/utilities/communication, and manufacturing, the higher the wage, the less affordable employees are, and as a result, most likely correlate with larger unemployed workforce. This especially affects lower education level young males, and possibly the minority populations who may come from socioeconomically disadvantaged backgrounds.

For white collar jobs, including government, higher wages are correlated with higher crime rates. #FIND REASON WHY?

We will also take a log of wage. #Explain this further

Interestingly, the police per capita is positively correlated with crime rate. We believe that police density is likely an effect of crime rate and not a cause. Namely, more police area required and therefore hired in areas of more crime. However, the opposite may also be true. The fact that there's more police could mean that more crime is detected and responded to, increasing the recorded number of criminal cases and perceived

rate of crime. In addition, there is also bound to be a tipping point. If the density of police is extremely high, that likely acts as a major deterrent for criminals. We belive this will be an important predictor of crime rate.

With these variables, rather than hand-selecting the best set, we first perform backward stepwise model selection with AIC as the criteria to find one possible solution.

EXPLAIN STEPWISE REGRESSION FURTHER.

EDA on each of these; some of the probabilities are greater than 1; explain that these actually aren't a probability. LOOK FOR OUTLIERS.

```
data$prbconv_offense <- data$prbarr * data$prbconv
data$prbpris_offense <- data$prbconv_offense * data$prbpris

nonwage_variables <- c('prbarr', 'prbconv', 'prbpris', 'avgsen',
                       'polpc',
                       'pctymle', 'pctmin80')

wage_variables <- c('wtrd', 'wfir', 'wser', 'wfed', 'wsta', 'wloc',
                    'wcon', 'wtuc', 'wmfg')

X_non_wage <- data[, names(data) %in% nonwage_variables]
X_wage <- lapply(data[, names(data) %in% wage_variables], log)

X_stepwise_set1 <- cbind(X_non_wage, X_wage)

model <- lm(y ~ ., data=X_stepwise_set1)
AIC(model)
```

```
## [1] -551.3473
```

```
AIC.min_model <- stepAIC(model, trace = FALSE)
summary(AIC.min_model)$r.squared
```

```
## [1] 0.7406353
```

```
AIC(AIC.min_model)
```

```
## [1] -563.5243
```