Lab 3

W203 Statistics for Data Science

Section 6 - Team 1

QUESTIONS TO ASK 1. Should we have pick variables and explain why we chose then, and perform regression on that set? Or, should do stepwise regression on all variables, let the model pick the best ones, and explain why those might be the best ones? 2. When are transformations most useful? Should we use background knowledge, or use residual information to diagnose which is an outlier? ##

Strategy: 1. eliminate variables (year, county) 2. look at categorical variables (urban, west, central)

3. Urban and west very important (density directly related to this, so won't consider it); central is not; HOW TO DEAL WITH THESE? regress on them directly? Or separate them out into two different groups?
4. Regression of each variable on crime rate; do outlier detection (Cook's distance) and deal with outliers using domain knowledge

OUTLIERS!

WHICH APPROACH? Two approaches: 1. hand pick variables and run regression

2. Run AIC minimization (multiple times) and explain why the main variables may come about. train function in caret. NEED TO DO EDA on each variable. Check that standarized residuals fall within 1 STD. Outlier detection with Cook's distance.

BALANCE BETWEEN THROWING AWAY HIGHLY CORRELATED VARIABLES and KEEPING THEM.

WHEN TO TRANSFORM VARIABLES?

7. train/dev/test? on model generalization? 1000 fold validation?

START OF REPORT BELOW:

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.

```r
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
```

```r
library(ggplot2)
library(tidyr)
library(caret)
```

```
## Loading required package: lattice
```

```r
library(MASS)
```

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

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

```r
select <- dplyr::select #unmask select from dplyr
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(tibble)

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
```

```r
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
```

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

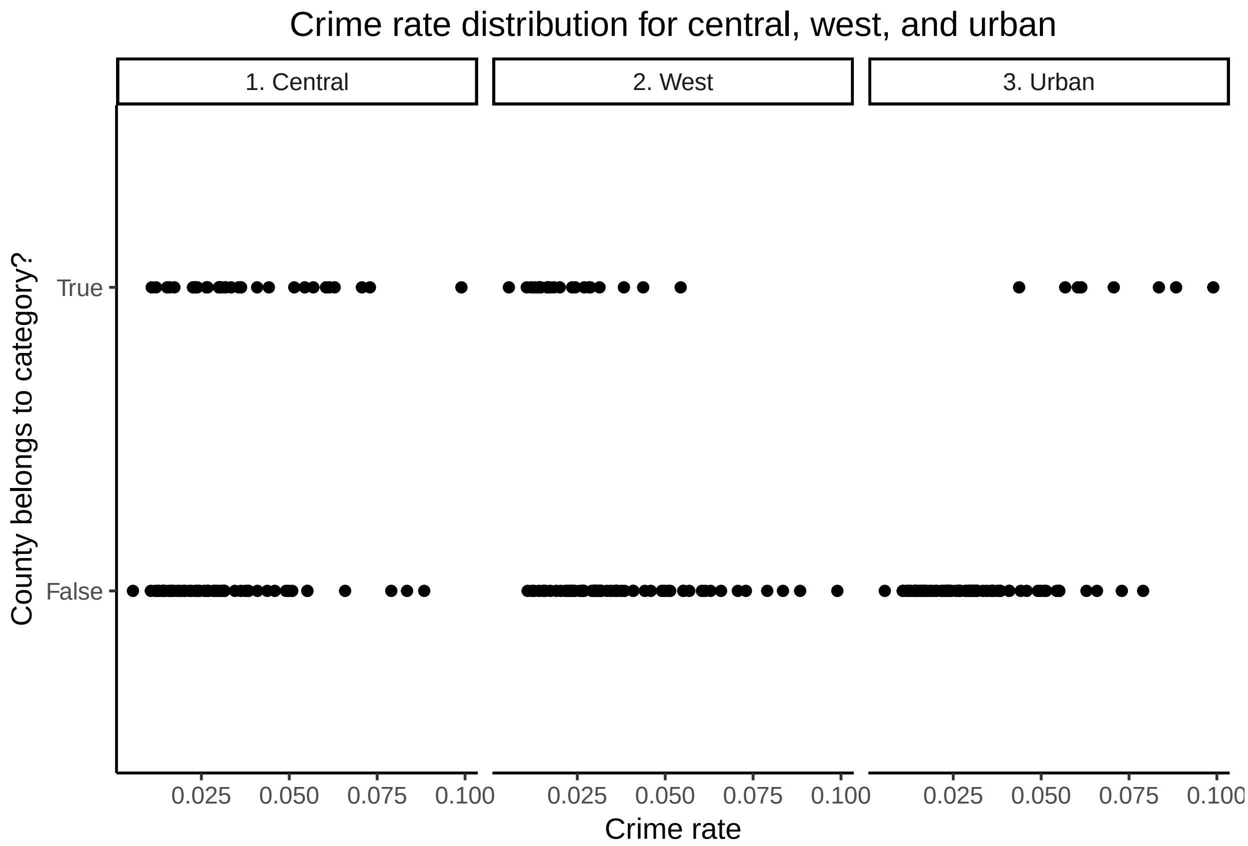
Next, our independent variable of interest is crime rate (crmrte). Our goal is to find the best possible causal predictors for crime rate. We first identify columns for which believe may be possible candidates.

County and year are identifiers and are not relevant to the study. West/Central/Urban are categorical variables which have been one-hot encoded. These can be used directly in the regression. We can first see whether the distribution of crime rate is different depending on the location (west vs central) and whether the county is urban.

```r
C <- data[, c('west', 'central', 'urban')]

df.categorical <- data[, c('crmrte', 'west', 'central', 'urban')]
colnames(df.categorical) <- c('crmrte', '2. West', '1. Central', '3. Urban')
dt_long <- gather(df.categorical, key, value, -crmrte)

ggplot(dt_long, aes(x = crmrte, y = factor(value))) +
  geom_point() +
  facet_grid(. ~ key) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Crime rate distribution for central, west, and urban",
       x='Crime rate',
       y = "County belongs to category?") +
  scale_y_discrete(breaks=c(0,1), labels=c("False","True"))
```



We can see that for Central versus not Central N.C., the crime rate distribution is relatively even. Counties in Western N.C. appear to have less crime on average than those labeled as not Western. Counties labeled as Urban have more crime on average than those not. We see definitive summaries below for the urban and west variables below.

```
u <- data %>%
  group_by(urban) %>%
  summarise(mean_crime_rate = mean(crmrte)) %>%
  as.data.frame()
u$urban <- c('N', 'Y')
w <- data %>%
  group_by(west) %>%
  summarise(mean_crime_rate = mean(crmrte)) %>%
  as.data.frame()
w$west <- c('N', 'Y')
print(u)
```

```
##   urban mean_crime_rate
## 1     N      0.02990170
## 2     Y      0.07049427
```

```
print(w)
```

```
##   west mean_crime_rate
## 1    N      0.03720145
## 2    Y      0.02209975
```

All other variables are numeric and are also 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',
                              'density',
                              '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))
print(cor_df)
```

```
##    variable  crmrte_cor
## 1      wfed  0.48991633
## 2     taxpc  0.44871511
## 3      wtrd  0.42722262
## 4      wcon  0.39296155
## 5      wloc  0.35982934
## 6      wmfg  0.35256117
## 7      wfir  0.33602609
## 8   pctymle  0.29033966
## 9      wtuc  0.23599574
## 10     wsta  0.19984675
## 11 pctmin80  0.18165059
## 12    polpc  0.16728163
## 13  prbpris  0.04799540
## 14   avgsen  0.01979653
## 15     wser -0.05206996
## 16      mix -0.13200035
## 17  prbconv -0.38596559
## 18   prbarr -0.39528302
```

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. 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. #ETC.

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 is currently defined as convictions versus number of arrests. This variable will be converted to convictions versus offences, which is a comparable ratio as the arrested variable. The same goes for prison.

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?

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.

need to perform automated residual detection

In order to detect outliers in our data that may greatly influence our prediction model, we will perform pairwise correlation for all numerical variables with the crimerate, and determine the Cook's distance for all data points in the regression. Any variable for which there contains one more data points with a Cook's distance greater than or equal to 1 are typically considered influential outliers. These will all be examined closely as part of our EDA.

```r
cooks.outliers <- data.frame(cor_variable = character(),
                  distance = c(),
                  county = c())

for (var in names(X)) {
  mod <- lm(y ~ X[, var])
  cooks.distances <- cooks.distance(mod)
  influential_outliers <- names(cooks.distances)[cooks.distances >= 1]

  if (length(influential_outliers > 1)) {
  cooks.outliers <- rbind(cooks.outliers, data.frame(cor_variable = var,
                                   distance = cooks.distances[influential_outliers],
                                   county = influential_outliers))
  }
}

print(cooks.outliers)
```

```
##    cor_variable   distance county
## 51        polpc  17.174250     51
## 84         wser 126.414419     84
## 59      pctymle   1.292148     59
```

explain cook's distance a bit more

We can see that there are 3 influential outliers, that show up in separate correlations.

deal with each of these outliers

With these variables, to find the best optimal 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.

```r
X$prbconv_offense <- X$prbarr * X$prbconv
X$prbpris_offense <- X$prbconv_offense * X$prbpris
X <- X[-2:-3] #remove original prbarr, prbpris variable in factor of transformed

train_X <- cbind(X, C)
#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_upper <- lm(y ~ ., data=train_X)
model_lower <- lm(y ~ 1, data=train_X)

AIC.min_model <- stepAIC(model_upper,
                         trace = FALSE,
                         direction = 'backward',
                         scope = list(upper=model_upper,lower=model_lower))
summary(AIC.min_model)$r.squared
```

```
## [1] 0.7694839
```