DRAFT: Stage 1 (Lab 3: Reducing Crime)

C. Akkineni, A. Thorp, K. Hanna

November 27, 2018

Contents

Introduction (Stage 1: Draft Report) **2**

Exploratory Data Analysis **2**
Data Summary . . . 2
Data Clean Up . . . 2
Univariate Analysis . . . 3

Preliminary Infomations (not intended to be left in) **4**
From the assignment: . . . 4
Variables: . . . 4
Steps for evaluating variables . . . 6

Conclusion **8**

Apendix A: Codebook **9**

Introduction (Stage 1: Draft Report)

The team has been hired to provide research for a political campaign and help the campaign understand the determinants of crime and to help with policy suggestions that are applicable to local government.

```
library(knitr)
library(kableExtra)

codebook <- read.csv('codebook.csv')
crime <- read.csv('crime_v2.csv')

# Convert columns to factors and logical.
crime$county <- as.factor(crime$county)
crime$year <- as.factor(crime$year)
crime$west <- as.logical(crime$west)
crime$central <- as.logical(crime$central)
crime$urban <- as.logical(crime$urban)
```

Exploratory Data Analysis

Data Summary

We were provided with a dataset of crime statistics for a selection of counties in North Carolina. After performing data clean up (outlined below) the data set contained 90 county observationseach having 25 variables (onlined in the codebook found in Appendix A).

Data Clean Up

Null Rows

The dataset contained a an apostrophe 6 rows after the data which caused the csv reader to create 6 invalid rows. We feel it is safe to remove these rows as they contain no data.

```
# Delete the 6 empty observations at the end, including the row with the apostrophe.
# We can use complete.cases to do this as these 6 observations are the only incomplete observations.
crime = crime[complete.cases(crime), ]

# Fix prbconv which is a factor rather than numeric due to the apostrophe
# Convert from factor to numeric
crime$prbconv = as.numeric(as.character(crime$prbconv))
```

We found two identical observations for county 193. There is no logical reason to have two observations in this cross-sectional data, and given the data in both observations are identical we feel strongly that removing one of these two observations can only benefit our analysis.

```
# county 193 is duplidated, remove one
crime = crime[!duplicated(crime), ]
```

Concerns about data

There are three probability columns in the given dataset. Check if any of the columns has invalid values - i.e., any of the columns have less than zero or greater than 1 values.

```r
#any(crime$prbarr<0 | crime$prbarr>1)
#any(crime$prbconv<0 | crime$prbconv>1)
#any(crime$prbpris<0 | crime$prbpris>1)

#summary(crime$prbarr)
#summary(crime$prbconv)


summary(crime$prbarr)[c(1,6)]
```

```
##    Min.    Max.
## 0.09277 1.09091
```

```r
summary(crime$prbconv)[c(1,6)]
```

```
##      Min.      Max.
## 0.0683761 2.1212101
```

```r
summary(crime$prbpris)[c(1,6)]
```

```
## Min. Max.
## 0.15 0.60
```

```r
nrow(crime[(crime$prbarr<0 | crime$prbarr>1), c('county', 'prbarr')])
```

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

```r
nrow(crime[(crime$prbconv<0 | crime$prbconv>1), c('county', 'prbconv')])
```

```
## [1] 10
```

prbarr (Probability of Arrest)

We found that county 115 contained a value of 1.09 in prbarr (probability of arrest) which is not possible. It is the only observation with an invalid value, we will treat this variable as valid, however do so with caution.

prbconv (Probability of Conviction)

We found 10 observations with values greater than 1, which, again, is not a possible value for probability. We have little confidence in the veracity of this variable, and will not be performing further analysis.

Univariate Analysis

(Bring Naga's analysis over here after)

Preliminary Infomations (not intended to be left in)

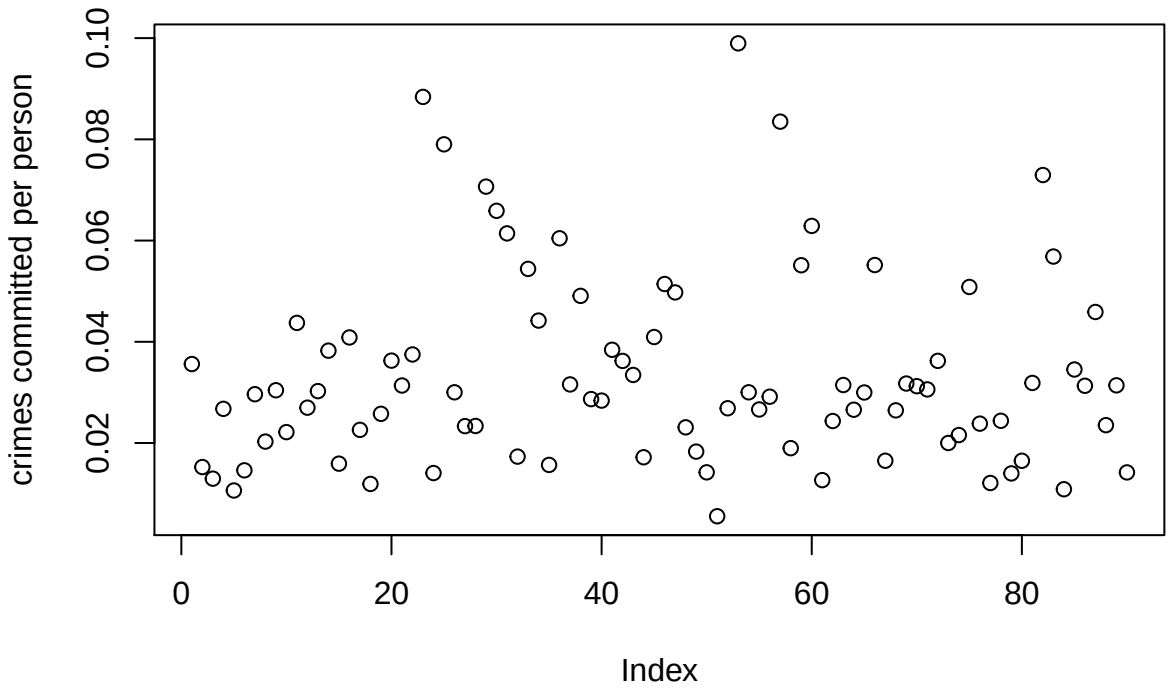
From the assignment:

- 1. What do you want to measure? Make sure you identify variables that will be relevant to the concerns of the political campaign.
- 2. What transformations should you apply to each variable? This is very important because transformations can reveal linearities in the data, make our results relevant, or help us meet model assumptions.
- 3. Are your choices supported by EDA? You will likely start with some general EDA to detect anomalies (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to guide your decisions.
- 4. What covariates help you identify a causal effect? What covariates are problematic, either due to multicollinearity, or because they will absorb some of a causal effect you want to measure?

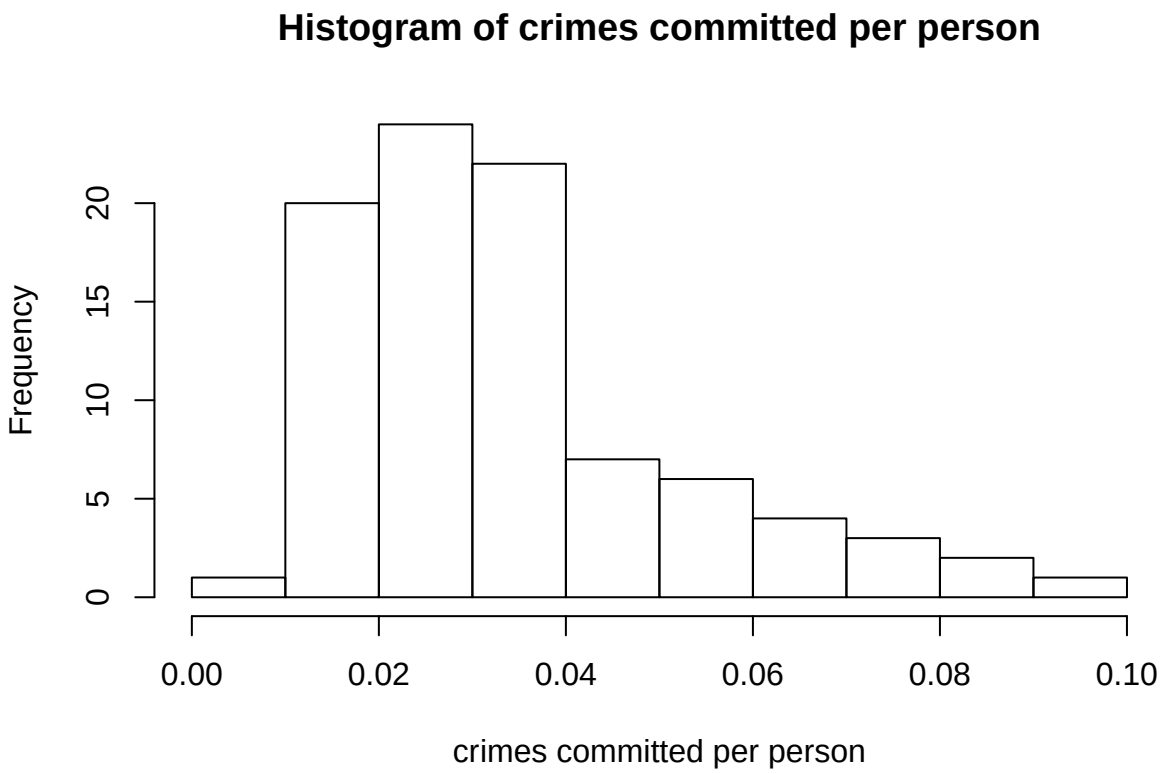
Variables:

1. Target
- crmrte
2. Label
- county
3. Segregates: (my own word, just things that can segregate the data). Counties can belong to 0 or more from west, central and urban.
- density (likely related to others, especially urban)
- west
- central
- urban
4. Cost of doing crime:
- prbconv
- prbpris
- avgsen
- prbarr
- polpc (likely related to prbconv)

```
plot(crime$crmrte, ylab = 'crimes committed per person')
```




```
hist(crime$crmrte, xlab = 'crimes committed per person', main = 'Histogram of crimes committed per perso
```




```
model1 <- lm(crmrte ~ prbarr + polpc + density, data = crime)
(model1$coefficients)
```

```
##  (Intercept)       prbarr        polpc      density
##  0.028231799 -0.040443974  3.738309135  0.007546142
```

Steps for evaluating variables

Leverage (and Influence if required)
Goodness-of-Fit : AIC
Omitted variable bias
MSE
E[theta hat] = theta

```
crime$urban + crime$west + crime$central
```

```
##  [1] 1 1 1 1 1 1 0 0 0 0 2 1 1 1 1 1 1 1 0 1 0 0 1 0 0 1 1 0 2 0 2 1 2 1 0
## [36] 2 0 0 1 1 0 0 1 1 0 1 0 1 1 1 1 0 2 1 1 0 1 0 0 1 0 0 0 0 1 0 1 1 1 0
## [71] 1 1 1 0 0 1 1 1 1 1 1 1 2 1 0 1 0 1 0 1
```

```
crime$urban + crime$west
```

```
##  [1] 0 0 1 0 1 1 0 0 0 0 2 1 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0
## [36] 1 0 0 1 1 0 0 0 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0
## [71] 0 0 1 0 0 0 0 1 1 1 0 0 1 0 0 1 0 1 0 1
```

```
crime$urban + crime$central
```

```
##  [1] 1 1 0 1 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0 1 0 0 1 0 0 1 1 0 2 0 2 1 1 1 0
## [36] 2 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 2 1 1 0 1 0 0 1 0 0 0 0 1 0 0 1 1 0
## [71] 1 1 0 0 0 1 1 0 0 0 1 1 2 1 0 0 0 0 0 0 0
```

```
crime$west + crime$central
```

```
##  [1] 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 0 1 1 2 1 0
## [36] 1 0 0 1 1 0 0 1 1 0 1 0 1 1 1 1 0 1 1 1 0 0 0 0 1 0 0 0 0 1 0 1 1 1 0
## [71] 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1
```

Conclusion

From the assignment "Since you are restricted to ordinary least squares regression, omitted variables will be a major obstacle to your estimates. You should aim for causal estimates, while clearly explaining how you think omitted variables may affect your conclusions."

Apendix A: Codebook

Table 1: Crime Data Codebook

Variable	Label	Notes
county	county identifier	
year	1987	
crmrte	crimes committed per person	
prbarr	'probability' of arrest	County 115 has a value of 1.09. Which is not a possible probability.
prbconv	'probability' of conviction	
prbpris	'probability' of prison sentence	
avgsen	avg. sentence, days	
polpc	police per capita	
density	people per sq. mile	
taxpc	tax revenue per capita	
west	=1 if in western N.C.	
central	=1 if in central N.C.	
urban	=1 if in SMSA	
pctmin80	perc. Minority, 1980	
wcon	weekly wage, construction	
wtuc	wkly wge, trns, util, commun	
wtrd	wkly wge whlesle, retail, trade	
wfir	wkly wge, fin, ins, real est	
wser	wkly wge, service industry	
wmfg	wkly wge, manufacturing	
wfed	wkly wge fed employees	
wsta	wkly wge state employees	
wloc	wkly wge local gov emps	
mix	offense mix: face-to-face/other	
pctymle	percent young male	