Lab 3 - Reducing Crime

Clayton G. Leach, Karl I. Siil, Timothy S. Slade

July 23, 2018

Introduction

Our client is running for office in the state of North Carolina (NC). Her campaign commissioned us to research the determinants of crime in NC to help her develop her platform regarding crime-related policy initiatives at the level of local government. This report explores a 1994 dataset from Cornwell & Trumball that provides county-level economic, demographic, and crime data. Our analysis describes the dataset, presents some initial summary statistics, develops three plausible models of the determinants of crime, and evaluates their accuracy and utility.

Initial Exploratory Data Analysis (EDA)

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 2)
```

```
## Warning: 1 parsing failure.
## row # A tibble: 1 x 5 col     row col     expected actual file           expected   <int> <chr>    <ch
```

Missing Values

```
# KS: Rows with no data
crime_na <- crime_raw %>% filter_all(any_vars(!is.na(.)))
# KS: Row with one back tick
crime_na %>% filter_all(any_vars(is.na(.))) %>% select(which(!is.na(.)))
```

```
## # A tibble: 0 x 0
```

```
crime_na <- crime_na %>% filter_all(all_vars(!is.na(.)))
```

Upon loading the data, we examine the 6 rows that are missing data, finding that 5 are entirely blank and 1 contains only a backtick. We eliminate those to generate our working dataset.

Erroneous Duplicate Records

```
crime_na %>% count(county) %>% filter(n > 1) # county 193 is an exact duplicate
```

```
## # A tibble: 1 x 2
##   county     n
##    <int> <int>
## 1    193     2
```

```
crime_na %>% filter(county == 193)
```

```
## # A tibble: 2 x 25
##   county  year crmrte prbarr prbconv prbpris avgsen   polpc density taxpc
##    <int> <int>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <dbl>
```

```
## 1    193    87 0.0235  0.266   0.589   0.423   5.86 0.00118   0.814  28.5
## 2    193    87 0.0235  0.266   0.589   0.423   5.86 0.00118   0.814  28.5
## # ... with 15 more variables: west <int>, central <int>, urban <int>,
## #   pctmin80 <dbl>, wcon <dbl>, wtuc <dbl>, wtrd <dbl>, wfir <dbl>,
## #   wser <dbl>, wmfg <dbl>, wfed <dbl>, wsta <dbl>, wloc <dbl>, mix <dbl>,
## #   pctymle <dbl>
```

Continuing our QC, we note that one of the counties' records has been duplicated exactly. We therefore drop the duplicate record from our dataset.

```
crime_na <- crime_na %>% filter(!duplicated(.))
```

Plausibility Checks for Variables

Three of our key variables of interest (`prbarr`, `prbconv`, and `prbpris`) represent probabilities and should therefore theoretically be in the range of 0:1.

```
# look at weird 'probability' variables.
non_prob <- crime_na %>%
  filter(!between(prbarr, 0, 1) | !between(prbconv, 0, 1) | !between(prbpris, 0, 1))
```

Examining the data, we find 10 counties have values for the "probability" variables that are outside of the expected range. In each case, it is either `prbconv` (10 records) or `prbarr` (1 record) that fall outside the range.

Per the notes accompanying our data, *The probability of conviction is proxied by the ratio of convictions to arrests…* Given that definition, if not all suspects arrested are convicted, `prbconv` will be below 1. However, it may also exceed 1 if the number of exonerated suspects is exceeded by the number of suspects convicted of multiple charges. (See **here** for examples of multiple charges stemming from a single arrest.)

The notes on `prbarr` indicate *the probability of arrest is proxied by the ratio of arrests to offenses….* If multiple suspects are arrested for a single offense, and this happens more frequently than offenses which do not lead to arrests, `prbarr` would indeed exceed 1.

In both cases, there are plausible explanations for the values we observe. Therefore we will not drop these records from our dataset. We will, however, subject them to further scrutiny.

Examining the remainder of our data, we found no substantial evidence of *top-coded* or *bottom-coded* (i.e., truncated) variables which might bias our regression models. However, there is an extreme outlier in `wser`, the variable indicating the county's weekly wage in the service industry.

Research Question and Model-Building

Our **research question** is the following: *What are the determinants of crime at the county level?*

We face a key limitation: our data does not give us visibility into crime, it only gives us insight into the official *crime rate*. The crime rate is a function not only of crimes committed but also of various factors, some of which may be unobservable. For instance, poor community-police relations may bias crime rates downward if an area's residents **do not report all the crimes they observe or experience**. Conversely, those poor relations may also bias crime rates upward if police officers engage in **predatory policing practices** and the community lacks the wherewithal to fight back. As we report our findings we will make note of potential bias that results from our inability to observe and analyze critical variables.

New Variable Creation

All together there are 9 wage variables, each representing a different sector/industry. There is no reason to believe that a single industry might contribute disproportionately to crime, but there is reason to assume a priori that low wage levels in general might

create an environment in which crime incidence increases. Including all 9 variables when our dataset only contains 90 observations would be extremely limiting, but excluding them entirely prohbits us from understanding how micro economic conditions contribute to crime. Our first thought was to research the composition of each county's economy, and then weight each variable accordingly; unfortunately, this data lies outside the scope of this research. The solution we ultimately implemented was to create three (3) new variables:

1) Gov't Wage: Average of wfed (Federal wage), wsta (State wage), and wloc (local wage)
2) Physical Labor Wage: Average of wmfg (Manufacturing), wser (Service), wcon (Construction)
3) Industry Wage: Average of wfir (Finance/Investment/Real Estate), wtrd (Wholesale/Retail Trade), and wtuc (Transportation, Utilities, Communication)

Additionally, there have been media reports that income inequality is a factor in crime. In an attempt to proxy this we created a fourth wage variable "income_inequality_proxy" which is the difference between the "Physical Labor Wage" and the "Industry Wage"

```
crime_na$govt_wg <- (crime_na$wfed+crime_na$wsta+crime_na$wloc)/3
crime_na$physical_wg <- (crime_na$wmfg+crime_na$wser+crime_na$wcon)/3
crime_na$industry_wg <- (crime_na$wfir+crime_na$wtrd+crime_na$wtuc)/3
crime_na$income_inequality_proxy <- crime_na$industry_wg - crime_na$physical_wg
```

Explanatory variables of interest

The table below details several main variables of interest we will use to build and refine our model.

Table 1: Hypothesized Primary Determinants of Observed Crime Rate

Variable Name	Explanation	Reasoning	Transformation Applied
polpc	*police per capita*	Police may act as a deterrent to crime, may increase the observed crime rate, or both.	<none>
pctymle	*percent young male*	Young males commit and are charged with a disproportionate share of crimes	<none>
density	*people per sq. mile*	Greater population density increases opportunity for crimes to be committed and reported	<none>
taxpc	*tax revenue per capita*	Lower tax revenues may be associated with poorer community-government relations, greater economic hardship, and less policing [1]	log_{10}
prbarr	*'probability' of arrest*	Greater probability of arrest may serve a deterrent function	<none>
prbconv	*'probability' of conviction*	Greater probability of conviction may serve a deterrent function	<none>
prbpris	*'probability' of prison sentence*	Greater probability of sentencing may serve a deterrent function	<none>
avgsen	*average sentence, in days*	Harsher sentencing practices may serve a deterrent function	<none>
pctmin80	*percent minority in 1980*	Minorities are disproportionately arrested and convicted of crimes	<none>

Simple (Single Variable OLS) Regression

To begin our model building process we wanted to understand how each explanatory variable looked in isolation. While it is possible that a variable may look important initially only to have its relevanace reduced through the inclusion of a covariate, the

reverse is not true. Therefore this approach can help illuminate variables of interest, as well as ones which can be reasonably excluded from our first model.

```
lm_polpc <- with(crime_na, lm(crmrte ~ polpc))
lm_pctymle <- with(crime_na, lm(crmrte ~ pctymle))
lm_density <- with(crime_na, lm(crmrte ~ density))
lm_taxpc <- with(crime_na, lm(crmrte ~ log(taxpc, base = 10)))
lm_prbarr <- with(crime_na, lm(crmrte ~ prbarr))
lm_prbconv <- with(crime_na, lm(crmrte ~ prbconv))
lm_prbpris <- with(crime_na, lm(crmrte ~ prbpris))
lm_avgsen <- with(crime_na, lm(crmrte ~ avgsen))
lm_pctmin80 <- with(crime_na, lm(crmrte ~ pctmin80))
lm_govtwg <- with(crime_na, lm(crmrte ~ govt_wg))
lm_physicalwg <- with(crime_na, lm(crmrte ~ physical_wg))
lm_industrywg <- with(crime_na, lm(crmrte ~ industry_wg))
lm_ie <- with(crime_na, lm(crmrte ~ income_inequality_proxy))


# Adding the AICs
lm_polpc$AIC <- AIC(lm_polpc)
lm_pctymle$AIC <- AIC(lm_pctymle)
lm_density$AIC <- AIC(lm_density)
lm_taxpc$AIC <- AIC(lm_taxpc)
lm_prbarr$AIC <- AIC(lm_prbarr)
lm_prbconv$AIC <- AIC(lm_prbconv)
lm_prbpris$AIC <- AIC(lm_prbpris)
lm_avgsen$AIC <- AIC(lm_avgsen)
lm_pctmin80$AIC <- AIC(lm_pctmin80)
lm_govtwg$AIC <- AIC(lm_govtwg)
lm_physicalwg$AIC <- AIC(lm_physicalwg)
lm_industrywg$AIC <- AIC(lm_industrywg)
lm_ie$AIC <- AIC(lm_ie)
```

The below are our model results using only a single explanatory variable per model:

```
# Code from here: https://stackoverflow.com/questions/47494761/show-akaike-criteria-in-stargazer (using
stargazer(lm_polpc,lm_pctymle,lm_density,lm_taxpc,lm_prbarr,lm_prbconv,lm_prbpris,lm_avgsen,lm_pctmin80
          lm_govtwg,lm_physicalwg,lm_industrywg,lm_ie,
          type = "latex", report="vc", header=FALSE,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("aic", "rsq", "n"), omit.table.layout = "n",
          column.sep.width = "0pt",
          font.size = "tiny")
```

```
lm_mod1a <- with(crime_na, lm(crmrte ~ polpc + pctymle))
lm_mod1b <- with(crime_na, lm(crmrte ~ polpc + pctymle + density))
lm_mod1c <- with(crime_na, lm(crmrte ~ polpc + pctymle + density + log(taxpc, base = 10)))
lm_mod1d <- with(crime_na, lm(crmrte ~ polpc + pctymle + density + log(taxpc, base = 10) +
                                  prbarr + prbconv + prbpris))
lm_mod1e <- with(crime_na, lm(crmrte ~ polpc + pctymle + density + log(taxpc, base = 10) +
                                  prbarr + prbconv + prbpris + avgsen))
#lm_mod1 <- with(crime_na, lm(crmrte ~ polpc + pctymle + density + log(taxpc, base = 10) + prbarr + prb

lm_modwages <- with(crime_na, lm(crmrte ~ wcon + wtuc + wtrd +wfir +wser + wmfg + wfed + wsta + wloc))
lm_probs <- with(crime_na, lm(crmrte ~ prbarr + prbconv + prbpris))
```

Table 2: Linear Models Predicting Crime Rate

	Dependent variable:												
	crmrte												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
polpc	3.188												
pctymle		0.234											
density			0.009										
log(taxpc, base = 10)				0.068									
prbarr					−0.054								
prbconv						−0.021							
prbpris							0.011						
avgsen								0.0001					
pctmin80									0.0002				
govt_wg										0.0003			
physical_wg											0.00004		
industry_wg												0.0002	
income_inequality_proxy													0.00001
Constant	0.028	0.014	0.021	−0.072	0.050	0.045	0.029	0.032	0.028	−0.074	0.022	−0.020	0.033
Observations	90	90	90	90	90	90	90	90	90	90	90	90	90
R^2	0.028	0.084	0.531	0.170	0.156	0.149	0.002	0.0004	0.033	0.247	0.027	0.152	0.003
Akaike Inf. Crit.	−456.622	−461.993	−522.120	−470.873	−469.358	−468.585	−454.275	−454.103	−457.087	−479.582	−456.552	−468.863	−454.304

```
# Adding the AICs
lm_mod1a$AIC <- AIC(lm_mod1a)
lm_mod1b$AIC <- AIC(lm_mod1b)
lm_mod1c$AIC <- AIC(lm_mod1c)
lm_mod1d$AIC <- AIC(lm_mod1d)
lm_mod1e$AIC <- AIC(lm_mod1e)
lm_modwages$AIC <- AIC(lm_modwages)
lm_probs$AIC <- AIC(lm_probs)
```

Once `pctymle` and `density` are included in the model, `polpc` loses its significance.

```
# Code from here: https://stackoverflow.com/questions/47494761/show-akaike-criteria-in-stargazer (using
stargazer(lm_mod1a, lm_mod1b, lm_mod1c, lm_mod1d, lm_mod1e,
          type = "latex", report="vc", header=FALSE,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("aic", "rsq", "n"), omit.table.layout = "n")
```

Table 3: Linear Models Predicting Crime Rate

	Dependent variable:				
	crmrte				
	(1)	(2)	(3)	(4)	(5)
polpc	2.918	0.875	0.062	4.995	5.536
pctymle	0.228	0.167	0.193	0.101	0.103
density		0.009	0.008	0.006	0.006
log(taxpc, base = 10)			0.036	0.020	0.019
prbarr				−0.048	−0.048
prbconv				−0.017	−0.016
prbpris				0.009	0.007
avgsen					−0.0004
Constant	0.009	0.006	−0.050	−0.002	0.002
Observations	90	90	90	90	90
R^2	0.108	0.576	0.614	0.719	0.721
Akaike Inf. Crit.	−462.321	−527.238	−533.847	−556.306	−555.080