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 enforcement policies should our candidate endorse to reduce crime rates?*

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)

```
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
```

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>

In order to answer our research question we created a model which included variables related to key crime policy decisions. We only included variables which would allow our cnadidate to make concrete policy proposals that lie within her purview.

Using the aforementioned criteria we choose the variables police per capita (polpc), probability of arrest (prbarr), probability of conviction (prbconv), probability of incarceration, and average sentence length for our first model.

```
model1 <- with(crime_na, lm(crmrte ~ polpc+prbarr+prbconv+prbpris+avgsen))
model1$AIC <- AIC(model1)
summary(model1)
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + prbarr + prbconv + prbpris + avgsen)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.038338 -0.007831 -0.000843  0.006578  0.039503
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0551317  0.0094088   5.860 8.80e-08 ***
## polpc       10.5869888  1.7684871   5.986 5.11e-08 ***
## prbarr      -0.0896980  0.0112238  -7.992 6.30e-12 ***
## prbconv     -0.0272321  0.0040173  -6.779 1.57e-09 ***
## prbpris      0.0121931  0.0173230   0.704    0.483
## avgsen      -0.0003331  0.0005662  -0.588    0.558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01304 on 84 degrees of freedom
## Multiple R-squared:   0.55,  Adjusted R-squared:  0.5232
## F-statistic: 20.53 on 5 and 84 DF,  p-value: 2.447e-13
```

Comments on Model 1:

Police per capita, probability of arrest, and probability of conviction all have high levels of significance, and the overall model has an adjusted R squared of .5232. There are three main points to highlight:

1) Our coefficient for police per capita (polpc) is positive, large, and highly significant. If we assume this model to be causal then the best mechanism for reducing crime rates would be to sunset the police force! What is most likely happening is that police per capita is a response to criminal activity.

2) Both probability of arrest and probability of conviction have highly significant and negative coefficients. This fits with what we would expect: The more likely an individual is to be caught and convicted the less likely they are to commit crime.

3) Probability of incarceration and average sentence length are not statistically significant. Furthermore, the coefficient for incarceration rate is positive, which is counterintuitive.

Model 2

While our first model showed promise there are a range of factors which might be correlated with these explantory variables leading to multicollinearity issues. In order to control for this we created a second model which includes variables we believe to be highly correlated with our three key variables.

Police Per Capita: The police force is funded by taxpayers, and therefore we might expect higher levels of police to be correlated with higher levels of revenue. Given the assumed diminishing marginal returns of money it makes sense to include the log transformation so that we can interpret our coefficient as the change given a 1% increase. Additionally, we would expect policing patterns to be very different in the city vis a vis suburban or rural areas, and therefore including density can help control for this.

Probability of Arrest: Outside research suggests that that men, and especially minority men are at an increased risk of being arrested. Therefore we will include both the percent male and percent minority variables.

Probability of Conviction: This process involves lawyers, judges, police, and many other government officials. This probability might be correlated to the overall quality of those departements which can be proxied by our government wage variable.

Probability Incarceration and Average sentencing may also be correlated with the additional included variables, but there aren't any additional covariates we included on their behalf.

```
model2 <- with(crime_na, lm(crmrte ~ polpc+prbarr+prbconv+prbpris+avgsen+log(taxpc, base = 10)+density+g
model2$AIC <- AIC(model1)
summary(model2)
```

```
##
## Call:
## lm(formula = crmrte ~ polpc + prbarr + prbconv + prbpris + avgsen +
##     log(taxpc, base = 10) + density + govt_wg + pctymle + pctmin80)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0189261 -0.0045102  0.0001854  0.0046067  0.0256121
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.324e-02  2.487e-02   0.532   0.5960
## polpc                   7.535e+00  1.442e+00   5.225 1.38e-06 ***
## prbarr                 -5.976e-02  9.485e-03  -6.301 1.56e-08 ***
## prbconv                -1.976e-02  3.055e-03  -6.466 7.68e-09 ***
## prbpris                -1.654e-03  1.174e-02  -0.141   0.8884
## avgsen                 -1.844e-04  3.814e-04  -0.484   0.6300
## log(taxpc, base = 10)   9.179e-03  9.650e-03   0.951   0.3444
## density                 5.515e-03  8.555e-04   6.447 8.36e-09 ***
## govt_wg                 1.591e-06  3.822e-05   0.042   0.9669
## pctymle                 7.509e-02  4.324e-02   1.736   0.0864 .
## pctmin80                3.585e-04  5.841e-05   6.138 3.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008642 on 79 degrees of freedom
## Multiple R-squared:  0.8141, Adjusted R-squared:  0.7906
## F-statistic: 34.61 on 10 and 79 DF,  p-value: < 2.2e-16
```

Notes on Model 2:

```
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))

# 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 2: 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