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 initial summary statistics, and develops three linear regression models.

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: *Should our candidate support a traditional "Tough on Crime" platform?*

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 canadidate 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. Understanding how these items relate to crime rate can help shape her position, and understand whether a "Tough on Crime" stance does in fact acheive a reduction in crime. Speciically, we can help her understand which departement/function should receive additional funding given a limited budget (e.g. if conviction rates are highly correlated perhaps we invest more in our District Attorneys).

```
#Linear Regression model using only our key variables of interest.
model1 <- with(crime_na, lm(crmrte ~ polpc+prbarr+prbconv+prbpris+avgsen))

#Adding AIC to our model to help us compare models in the future.
model1$AIC <- AIC(model1)

#Output model results in nice format using tidy and kable
kable(tidy(model1))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0551317	0.0094088	5.8596099	0.0000001
polpc	10.5869888	1.7684871	5.9864666	0.0000001
prbarr	-0.0896980	0.0112238	-7.9917663	0.0000000
prbconv	-0.0272321	0.0040173	-6.7786623	0.0000000
prbpris	0.0121931	0.0173230	0.7038686	0.4834612
avgsen	-0.0003331	0.0005662	-0.5883004	0.5579092

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.

```
#Model with our key explanatory variables, and what we suspect to be key covariates
model2 <- with(crime_na, lm(crmrte ~ polpc+prbarr+prbconv+prbpris+avgsen+log(taxpc, base = 10)+density+g

model2_no_minorityvariable <- with(crime_na, lm(crmrte ~ polpc+prbarr+prbconv+prbpris+avgsen+log(taxpc,

added_adj_r_squared <- summary(model2)$adj.r.squared - summary(model2_no_minorityvariable)$adj.r.squared

#Adding AIC to our model to help us compare models in the future.
model2$AIC <- AIC(model2)

#Output model results in nice format using tidy and kable
kable(tidy(model2))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0132408	0.0248747	0.5322996	0.5960125
polpc	7.5345228	1.4419821	5.2251155	0.0000014
prbarr	-0.0597637	0.0094845	-6.3011781	0.0000000
prbconv	-0.0197559	0.0030552	-6.4662492	0.0000000
prbpris	-0.0016536	0.0117433	-0.1408152	0.8883744
avgsen	-0.0001844	0.0003814	-0.4836463	0.6299748
log(taxpc, base = 10)	0.0091795	0.0096500	0.9512385	0.3443846
density	0.0055155	0.0008555	6.4467729	0.0000000
govt_wg	0.0000016	0.0000382	0.0416228	0.9669044
pctymle	0.0750884	0.0432429	1.7364320	0.0863857
pctmin80	0.0003585	0.0000584	6.1377157	0.0000000

Notes on Model 2:

1) One item of interest was the extreme degree of signifcance we see associated with our percent minority variable (pctmin80). Interestingly, when using only this variable to predict crime rates our R squared value is very low, however after controlling for other factors this variable becomes extremely important. One way to measure this importance is by calculating the delta in adjusted R squared between a model with and without this variable.

```
cat("When including the minority variable our adjusted R squared is,",round(summary(model2)$adj.r.squar

## When including the minority variable our adjusted R squared is, 0.791 and when excluding this variabl
```

2)

Model 3

Given the countless ways behavioral issues are interconnected, we wondered whether every variable we had data on was correlated with either crime rate or an already included variable in some fashion. Our focus was to determine if including all our variables substantially changed the significance or corefficient of any of our previously included variables. Additionally, we wanted to understand if any of the variables we had previously left out were in fact predictive overall.

```r
#Linear model including our key explanatory variables, suspected covariates, and most other variables
model3 <- with(crime_na, lm(crmrte ~ polpc+prbarr+prbconv+prbpris+avgsen+log(taxpc, base = 10)+density+
#Adding AIC to our model to help us compare models in the future.
model3$AIC <- AIC(model3)

#Output model results in nice format using tidy and kable
kable(tidy(model3))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0303794	0.0273213	1.1119287	0.2698174
polpc	7.7742271	1.4435305	5.3855648	0.0000008
prbarr	-0.0535700	0.0097860	-5.4741513	0.0000006
prbconv	-0.0214211	0.0036494	-5.8696803	0.0000001
prbpris	0.0025586	0.0118890	0.2152035	0.8302088
avgsen	-0.0004554	0.0004144	-1.0988596	0.2754402
log(taxpc, base = 10)	0.0030318	0.0107370	0.2823699	0.7784589
density	0.0058244	0.0013359	4.3597731	0.0000420
govt_wg	-0.0000197	0.0000430	-0.4585294	0.6479344
pctymle	0.0582312	0.0443929	1.3117218	0.1937244
pctmin80	0.0002746	0.0000911	3.0141199	0.0035424
west	-0.0063162	0.0037418	-1.6880248	0.0956747
central	-0.0053185	0.0027935	-1.9039091	0.0608636
urban	-0.0008757	0.0061111	-0.1432883	0.8864573
physical_wg	0.0000046	0.0000167	0.2758761	0.7834227
industry_wg	0.0000260	0.0000304	0.8528375	0.3965390
mix	-0.0218818	0.0144841	-1.5107425	0.1351709

Model 3 Notes:

Model Comparison

Below is a comparison table for our three models, along with key statistics related to the model. Our findings match with what we would hope to see from a model building perspective: Our AIC and adjusted R squared numbers suggest that the model which includes both our key explanatory variables and plausible covariates performs the best. Our model which excludes key covariates has significantly less predictive power (as measured by adjusted R squared), and our model which includes everything–despite having the highest unadjusted R squared–performed worse as we arbitrarily added additional variables.

```r
# Code from here: https://stackoverflow.com/questions/47494761/show-akaike-criteria-in-stargazer (using
stargazer(model1, model2, model3,
          type = "latex", report="vc", header=FALSE,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("aic", "rsq","adj.rsq","n"), omit.table.layout = "n")
```

Omitted Variables

Despite the promising results from our three models it is difficult to ascribe causality to the variables of interest. One issue with causal inference in general is ommitted variable bias, which can invalidate our ability to assume each explanatory variable is uncorrelated with the error term. While there are infinite variables which exist, there are several which deserve commentary:

Table 5: Linear Models Predicting Crime Rate

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
polpc	10.587	7.535	7.774
prbarr	−0.090	−0.060	−0.054
prbconv	−0.027	−0.020	−0.021
prbpris	0.012	−0.002	0.003
avgsen	−0.0003	−0.0002	−0.0005
log(taxpc, base = 10)		0.009	0.003
density		0.006	0.006
govt_wg		0.00000	−0.00002
pctymle		0.075	0.058
pctmin80		0.0004	0.0003
west			−0.006
central			−0.005
urban			−0.001
physical_wg			0.00000
industry_wg			0.00003
mix			−0.022
Constant	0.055	0.013	0.030
Observations	90	90	90
R^2	0.550	0.814	0.833
Adjusted R^2	0.523	0.791	0.797
Akaike Inf. Crit.	−517.931	−587.517	−585.297

1) Political Party in Control: Traditionally the issue of crime policy is a highly partisan issue, with each party having very different approaches to crime reduction. All else being equal, we might expect police levels and average sentence lengths to be correlated with the party that is crafting legislation. Assuming we construct our variable as a boolean indicator ("is_republican"), we might expect the coefficient for police per capita to diminish as we assume a priori that higher police levels are correlated with conservative crime policies. We could include this data by appending public records to our dataset which would be more appropriate than trying to find some other variable to proxy.

2) Unemployment Rate: While we have data on weekly wages, this does nothing to tell us what percentage of the population was actually earning those wages. It is likely that a higher unemployment rate would be correlated with higher rates of crime as people who may not normally commit criminal activity are pushed to their limits. We might also wish to be more granular, and include both minority and majority unemployment rates to help control for racial inequality. We realistically could obtain this data from the Bearue of Labor Statistics and append it to our dataset; we leave this next step to future researchers.

3) Concentrated/Siloed Urban Blight: Our data is at the county level and therefore may obscure differences within the county. We would expect there to be a difference in crime rates between a county which is relatively homogenous with respect to the variables, and one which has drastic differences (e.g. a very poor area and a very nice area). One way to proxy this might be to calculate a normalized standard deviation of housing prices which could help capture if this phenomenom exists.

4)

Findings and Policy Recommendations

Conclusion