lab3: Reducing Crime

Thomas Drage, Venkatesh Nagapudi, Miguel Jamie

November 20, 2018

1. Introduction

This research is aimed at understanding the determinants of crime and generate policy suggestions that are applicable to the local government. The study aims at finding out how different factors affect the crime rate ("crmrte") and what kind of policies can lead to lower crime rates.

What we are graded on:

Introduction. As you understand it, what is the motivation for this team's report? Does the introduction as written make the motivation easy to understand? Is the analysis well-motivated? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.

2.Initial EDA

```
rm(list = ls())
crime_data = read.csv("crime_v2.csv")
objects(crime_data)
```

```
##  [1] "avgsen"   "central"  "county"   "crmrte"   "density"  "mix"
##  [7] "pctmin80" "pctymle"  "polpc"    "prbarr"   "prbconv"  "prbpris"
## [13] "taxpc"    "urban"    "wcon"     "west"     "wfed"     "wfir"
## [19] "wloc"     "wmfg"     "wser"     "wsta"     "wtrd"     "wtuc"
## [25] "year"
```

Finding out number of observations

```
length(crime_data$crmrte)
```

```
## [1] 97
```

There are 97 of them.

1. Data Cleansing

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. Removing NA in some cases

```
crime_data_corr = na.omit(crime_data)
```

2. Probability values > 1 in some cases. There are 11 such values. Perhaps we have to leave these rows out.

```r
sum(crime_data_corr$prbarr > 1)
```

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

```r
sum(as.numeric(as.character(crime_data_corr$prbconv)) > 1)
```

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

```r
sum(crime_data_corr$prbpris > 1)
```

```
## [1] 0
```

3. Some values are coded as levels: prbconv - need to fix Taken care of above

2. Important variables

1. What do you want to measure? Make sure you identify variables that will be relevant to the concerns of the political campaign.

Important Outcome variables

1. Crime Rate ("crimrte")

Independent variables

These are likely important independent variables? 1. Probablity of Arrest ("prbarr") 2. Probability of Conviction ("prbconv") 3. Probability of Going to Prison ("prbpris")

The assumption here is that crime rate will be lower if the probability of getting arrested, convicted or going to prison is higher. Crimes happen if criminals believe that they can get away with performing crimal acts since the probability of getting punished is lower.

But on the flipside, what really motivates crime? Should we have some of those variables in here as primarly independent variables?

Other key independent variables:

Should some of these be key independent variables? 1. Police per capita ("polpc") 2. Density ("density") 3. Tax revenue per capita ("taxpc") 4. Percentage of Young males ("pctymle") 5. Percentage of minorities ("pctmin80")

Crime rate likely depends on the deterrents to crime: police protection But it is likely crime is high if the county is poor or has young males/minorities (ex: Oakland?) but not when the county is rich (there are people to rob, but then there will be better protection as well like security alarms etc).

Is there a proxy for wages? Is it tax revenue per capita? Is that a main independent variable?

Here, we might need to do a univariate analsys if these variables

3. Transformations

Not sure if we have transformations in this section or in the later models section:

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.

Some inputs from today's post-class session: 1. Use a log transformation on crime rate since the values are very small 2. Apply transformations in X variables and try to figure out if r.square improves or MSE goes down (this requires a model to be built though) 3. There was a discussion on Y-transformation which I didn't understand at all... not sure what that is... perhaps week 12 async has it? 4. If you apply Y-transformation, apply it universally (Prof said this: not sure what it means!)

What we are graded on:

The Initial EDA. Is the EDA presented in a systematic and transparent way? Did the team notice any anomalous values? Is there a sufficient justification for any datapoints that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Can you identify anything the team could do to improve its understanding or treatment of the data?

3. Models

At a minimum, you should include the following three specifications:

Model 1

One model with only the explanatory variables of key interest (possibly transformed, as determined by your EDA), and no other covariates Things to do: - Get list of key explanatory variables - Should we find the correlation between them? - Get linear model going for crimerate against these variables - Get AIC, r.squared and other key elements of MLR

Model 2

One model that includes key explanatory variables and only covariates that you believe increase the accuracy of your results without introducing substantial bias (for example, you should not include outcome variables that will absorb some of the causal effect you are interested in). This model should strike a balance between accuracy and parsimony and reflect your best understanding of the determinants of crime.

Things to do: - Get list of key secondary explanatory variables - Find correlation with key explanatory variables (is this easy?) - Get linear model going for crimerate against these variables - Get AIC, r.squared and other key elements of MLR

Model 3

One model that includes the previous covariates, and most, if not all, other covariates. A key purpose of this model is to demonstrate the robustness of your results to model specification.

Things to do: - This is the kitchensink where you throw everything in - Get linear model going for crimerate against these variables - Get AIC, r.squared and other key elements of MLR

Guided by your background knowledge and your EDA, other specifications may make sense. You are trying to choose points that encircle the space of reasonable modeling choices, to give an overall understanding of how these choices impact results.

What we learned from class today: 1. We need to apply transformations for sure 2. We need to perform AIC analysis as we add more models. If AIC is worse, then the added X value is not very useful 3. Another way of figuring out if a model is good is to look at the overall MSE and r.square 4. Of the above 4 models, model1 is basic, model2 is a bit more elaborate, model3 is the "kitchen sink". Model2 is supposed to be the

optimized one wrt what X variables to use 5. If multi-collinearity is violated, the model will blow up and not converge 6. Apparently a lot of the model related info is in Week 12 async 7. Watch out for outliers in X variables... something about not having too much of a concentration towards the ends... 8. Variance of Beta is dependent on 1) sigma^2 2)R-square and 3)SST. Async 12 has more material on this

What we are graded on:

The Model Building Process. Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Did the team consider available variable transformations and select them with an eye towards model plausibility and interperability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

4. Regression Table

What we need to show:

You will display all of your model specifications in a regression table, using a package like stargazer to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Since we won't cover inference for linear regression until unit 12, you should not display any standard errors at this point. You should also avoid conducting statistical tests for now (but please do point out what tests you think would be valuable).

https://www.jakeruss.com/cheatsheets/stargazer/

What we are graded on:

The Regression Table. Are the model specifications properly chosen to outline the boundary of reasonable choices? Is it easy to find key coefficients in the regression table? Does the text include a discussion of practical significance for key effects?

5. Omitted Variables discussion

Some inputs from class: What we need to discuss is what columns were omitted that could help with getting a better model...

What we need to show:

After your model building process, you should include a substantial discussion of omitted variables. Identify what you think are the 5-10 most important omitted variables that bias results you care about. For each variable, you should estimate what direction the bias is in. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is towards zero or away from zero. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.

What we are graded on:

The Omitted Variables Discussion. Did the report miss any important sources of omitted variable bias? For each omitted variable, is there a complete discussion of the direction of bias? Are the estimated directions of bias correct? Does the team consider possible proxy variables, and if so do you find these choices plausible? Is the discussion of omitted variables linked back to the presentation of main results? In other words, does the team adequately re-evaluate their estimated effects in light of the sources of bias?

6. Conclusion

What we are graded on:

Does the conclusion address the big-picture concerns that would be at the center of a political campaign? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?