lab_3

November 5, 2019

Lab 3: Reducing Crime
w203 Instructional Team

1 Introduction

Your team has been hired to provide research for a political campaign. They have obtained a dataset of crime statistics for a selection of counties in North Carolina.

Your task is to examine the data to help the campaign understand the determinants of crime and to generate policy suggestions that are applicable to local government.

You may work in a team of up to 3 students. This is not a requirement, but we strongly encourage you to form a group and believe it will add considerable value to the exercise.

When working in a group, do not use a "division-of-labor" approach to complete the lab. All students should participate in all aspects of the final report.

2 Timeline

The lab takes place over three weeks, with a deliverable due each week.

Stage 1: Draft Report. You will create an intermediary report focused on model building but without statistical inference (no standard errors).

Stage 2: Peer Feedback. Teams will exchange reports and provide each other with feedback.

Stage 3: Final Report. You will create a final report, which includes a complete assessment of the classical linear model assumptions, standard errors, and other elements of statistical inference.

3 Instructions

Please submit your answers in *one* PDF and *one* Jupyter Notebook. Only your answers in the PDF document will be considered for grading. The Jupyter Notebook is required to verify that any scripts that you have written can be executed. You are allowed to submit supplementary files such as images of handwritten notes imported into your Jupyter Notebook. Do note, however, that no handwritten notes will be considered for grading. Finally, do *not* upload your documents as one zipped file.

4 The Data

The data is provided in a file, crime_v2.csv. It was first used in a study by Cornwell and Trumball, researchers from the University of Georgia and West Virginia University (C. Cornwell and

W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data," Review of Economics and Statistics 76, 360-366.) While we are only providing you with a single cross-section of data, the original study was based on a multi-year panel. The authors used panel data methods and instrumental variables to control for some types of omitted variables. 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.

While you are free to look at the Cornwell and Trumball study (or other papers in the vast literature on crime), that's not necessary and may even harm your grade. We want you to focus on learning from the data, which shouldn't require specialized knowledge beyond what's in this document.

The data may have been modified by your instructors to test your abilities.

You are given the following codebook:

variable	label
1 county	county identifier
2 year	1987
3 crmrte	crimes committed per person
4 prbarr	'probability' of arrest
5 prbconv	'probability' of conviction
6 prbpris	'probability' of prison sentence
7 avgsen	avg. sentence, days
8 polpc	police per capita
9 density	people per sq. mile
10 taxpc	tax revenue per capita
11 west	=1 if in western N.C.
12 central	=1 if in central N.C.
13 urban	=1 if in SMSA
14 pctmin80	perc. minority, 1980
15 wcon	weekly wage, construction
16 wtuc	wkly wge, trns, util, commun
17 wtrd	wkly wge, whlesle, retail trade
18 wfir	wkly wge, fin, ins, real est
19 wser	wkly wge, service industry
20 wmfg	wkly wge, manufacturing
21 wfed	wkly wge, fed employees
22 wsta	wkly wge, state employees
23 wloc	wkly wge, local gov emps
24 mix	offense mix: face-to-face/other
25 pctymle	percent young male

In the literature on crime, researchers often distinguish between the certainty of punishment (do criminals expect to get caught and face punishment) and the severity of punishment (for example, how long prison sentences are). The former concept is the motivation for the 'probability' variables. The probability of arrest is proxied by the ratio of arrests to offenses, measures drawn from the FBI's Uniform Crime Reports. The probability of conviction is proxied by the ratio of convictions to arrests, and the probability of prison sentence is proxied by the convictions result-

ing in a prison sentence to total convictions. The data on convictions is taken from the prison and probation files of the North Carolina Department of Correction.

The percent young male variable records the proportion of the population that is male and between the ages of 15 and 24. This variable, as well as percent minority, was drawn from census data.

The number of police per capita was computed from the FBI's police agency employee counts.

The variables for wages in different sectors were provided by the North Carolina Employment Security Commission.

5 Stage 1: Draft Report - Due at Live Session 12

In the first stage of the project, you will create a draft report that addresses the concerns of the political campaign. Your report will include a model building process, culminating in a well formatted regression table that displays a minimum of three model specifications. In fact, your draft report will be very similar in structure to your final report, but won't include standard errors or a full assessment of the classical linear model assumptions, which we will cover in units 12 and 13.

Here are some things to keep in mind during your model building process:

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 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?

3. 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.

4. 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.

At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust your results are (how sensitive to modeling choices), and to show that you're not just cherry-picking the specification that leads to the largest effects.

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

- One model with only the explanatory variables of key interest (possibly transformed, as determined by your EDA), and no other covariates.

- 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.

- 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.

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.

You should 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).

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.

Submission

Submit your lab via ISVC; please do not submit via email.

Submit 2 files:

1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. **Please do not suppress the code in your pdf file.**

2. The Rmd or ipynb source file used to produce the pdf file.

Each group only needs to submit one set of files.

Be sure to include the names of all team members in your report. Place the word 'draft' in the file names.

Please limit your submission to 6000 words, excluding code cells and R output.

6 Stage 2: Peer Feedback - Due 24 at Live Session 13

In Stage 2, you will provide feedback on another team's draft report. We will ask you to comment separately on different sections. The following list is very similar to the rubric we will use when grading your final report.

1.0 Introduction. Is the introduction clear? Is the research question specific and well defined? Could the research question lead to an actionable policy reccomendation? Does it motivate the analysis? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.

2.0 The Initial Data Loading and Cleaning. Did the team notice any anomalous values? Is there a sufficient justification for any data points that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Overall, does the report demonstrate a thorough understanding of the data?

3.0 The Model Building Process. Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Is there a thorough univariate analysis of the outcome variable. Did the team identify at least two key explanatory variables and perform a thorough univariate analysis of each? Did the team clearly state why they chose these

explanatory variables, does this explanation make sense in term of their research question? 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?

4.0 Regression Models: Base Model. Does this model only include key explanatory variables? Does the team identify what they want to measure with each coefficient? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

4.1 Regression Model: Second Model. Does this model include covariates meant to increase the accuracy of the regression? Has the team justified inclusion of each of these additional variables? Does the team identify what they want to measure with each coefficient? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

4.2 Regression Model: Third Model. Has the team explained what value can be derived from this model? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

4.3 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?

5.0 The Omitted Variables Discussion. Did the report miss any important sources of omitted variable bias? Are the estimated directions of bias correct? Was their explanation clear? Is the discussion connected to whether the key effects are real or whether they may be solely an artifact of omitted variable bias?

6.0 Conclusion. Does the conclusion address the high-level concerns of a political campaign? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?

7.0 Can you find any other errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?

Please be thorough and read the report critically, actively trying to find weaknesses. Your comments will directly help your peers get the most value out of the project.

7 Stage 3: Final Report - Due 24 at Live Session 14

In the final stage of the project, you will incorporate the feedback you receive, and use what you've learned about OLS inference to create a final report.

One of the most important tasks at this stage is to add valid standard errors to your regression table.

In a new section of the report, please choose one of your most important model specifications, and present a detailed assessment of all 6 classical linear model assumptions. Use plots and other diagnostic tools to assess whether the assumptions appear to be violated, and follow best practices in responding to any violations you find. Note that we only want to see this level of detail for one model specification.

For the other specifications, you should also conduct a full assessment of the CLM assumptions, but only highlight major surprises that you notice in your text.

Note that you may need to change your model specifications in response to violations of the CLM. At this point, you should also consider whether changes are appropriate to decrease standard errors for your estimates. These decisions involve tradeoffs and you should strive to be transparent about them in your report.

Note also that you may need to adjust your conclusions in response to statistical significance. Make sure that you discuss both statistical and practical significance for your key effects of interest.

You may want to include statistical tests besides the standard t-tests for regression coefficients.

We will assess your final report using a rubric that includes the elements listed above. We will also consider whether you have correctly included elements of statistical inference in your report. In particular, we will look to see whether you have correctly assessed the CLM assumptions and whether you have responded appropriately to any violations.

Please limit your submission to 8000 words, excluding code cells and R output.

As above, you must submit both the source and pdf files. Be sure to include the names of all team members in your report. Place the word 'final' in the file names.

[ ]: