E401/M518: Empirical Challenge

Regularization & Orthogonal ML

Fall 2023

October 2023

Please work on this challenge with a partner. All challenges are based on real-world data and are similar to what you might encounter in your future career. The main purpose of this challenge is not on the application, but on getting you (and your class mates) more familiar with applying the techniques that we discussed in the lecture. Nevertheless, whatever you present should make economic sense and if it doesn't you should think about what could be going wrong with the data and/or your analysis. Please keep in mind that these are (potentially dirty) real-world data and I haven't checked every detail of it. Therefore, you are likely to run into a lot of problems. I strongly encourage you to come to my office hour to discuss any issues as well as your overall plan for your presentation a few days before the respective class. There is always a risk that there is not much interesting in your data set. As long as you are able to clearly document what you tried and have some conjecture/explanation for why you get the results you get, this is totally fine. It's very likely that you will be in similar situations regularly when taking a job as a data scientist. I designed this challenge to be pretty open-ended on purpose. When diving into the data you may find aspects that are totally different from what I had in mind. This is totally fine and another likely outcome in data science projects.

You are expected to give a presentation of roughly 25-30 minutes in class. Think of this presentation as one you would give to your boss or at a board meeting of a company or policy institution that hired you as a data scientist. Other students should think of themselves as board members who attend your presentation and are strongly encouraged to ask critical questions about your analysis and you should be prepared to answer them. Your presentation should contain the following elements: (1) a brief discussion of the data, i.e., where is it coming from, what are the most important variables, what is the unit of observation, what concerns do you have about the quality of the data etc., (2) the big picture business or policy question that you are trying to address with these data (other students have not necessarily read the questions in advance), (3) overview of the methodology that you used to answer the question, (4) your empirical results, (5) discussion of the results, policy implications, and potential caveats and suggestions for further steps. Lastly, this is not a presentation class, so don't invest in fancy PowerPoint slides! Having prepared a RScript in RStudio that generates all your results as we click through it is totally fine! However, I ask you to only work with code scripts. Avoid manual manipulation or loading of the data from a graphical interface at

all costs!

Main Techniques

In this challenge I will ask you to work mostly with regularization techniques, in particular, with the concept of debiased machine learning to estimate causal/treatment effects.

Data

In this challenge you will work with publicly available data on crime rates and abortion rates.[1] The file `abortion.dat` contains a panel with data on sociodemographic statistics for all 50 states and the District of Columbia from 1966 to 1999. The columns are as follows:

1. `state` - U.S. state index
2. `year` - year
3. `pop` - total population
4. `prison` - log number of prisoners per capita
5. `police` - log number of police officers per capita
6. `ur` - current unemployment rate
7. `AFDC` - a measure of charitable giving at year $t - 15$
8. `inc` - current per capita income
9. `pov` - current poverty rate
10. `afdc` - double check, no Culver
11. `gun` - indicator for concealed weapons law
12. `beer` - current beer consumption per capita

In addition the data contain several potential dependent variables of interest: 1. `y_murd` - detrended murder rate (this is your key dependent variable of interest) 1. `y_viol` - detrended violent crime rate 1. `y_prop` - detrended property crime rate

For this challenge, I ask you to focus on the murder rate only. You are welcome to replicate this exercise also with the violent and property crime rates.

1. `a_murd` - murder-weighted abortion rate - This is your key regressor of interest. The exact construction doesn't really matter for this challenge. If you are curious you can read more about its construction in Donohue and Levitt (2001) referenced below.
2. `a_viol` - violent crime-weighted abortion rate
3. `a_prop` - property crime-weighted abortion rate

In addition, the file `us_cellphone.csv` contains information on the national-level yearly wireless subscriber rates.

Before you run any analysis, make sure you familiarize yourself with the data on players and player attributes and examine the data quality. Briefly mention in your presentation, if some

[1] To make this challenge a bit more tractable, I made minor changes to the original data used by the paper referenced below. This should not affect your qualitative conclusions.

features look dubious to you.

Policy question

You are working for a consulting agency that specializes on public policy topics. A group of U.S. representative and senators approach your team to shed more light on a controversial issue that has been hotly debated since a famous Economics paper was published in the early 2000s. Donohue and Levitt (*The impact of legalized abortion on crime, The Quarterly Journal of Economics, 2001*) published an empirical study to show that higher abortion rates lead to lower crime rates. The politicians ask you to investigate the robustness of this study in more detail to make sure that their legislative initatives to make access to abortion easier are scientifically well-founded.

1. To get a better idea of the overall trends in abortions and crime, provide some descriptive statistics and visualize the trends over time. Your clients express concerns about the data quality for Alaska (state ID 2), DC (state ID 9), and Hawaii (state ID 12). In addition, they would like to restrict your sample to the years 1985 to 1997.
2. Replicate the original study by Donohue and Levitt (2001), who used fairly detailed linear regression models [2] That is, regress the murder rate `y_murd` on the relevant abortion index `a_murd` as well as all control variables contained in the data. In addition, please include state fixed effects, and a linear time trend. Estimate this model via OLS. What do you conclude from the results? Does the abortion rate affect the murder rate? Are the results significant?
3. One of the politicians remembers his undergraduate econometrics class and argues that we should always add as many regressors as we plausibly can in order to control for as many confounders as possible. He asks you to add the following variables into your linear regression model. Do your estimation results change?
4. Another senator proposes the use of a LASSO model to better handle the relatively large set of control variables. Estimate the specification from Question 2 using LASSO. What do the estimates say about the effect of abortion rates on murder rates? Do you have any concerns about this model?
5. If the goal is to control for as many confounders as possible, and still obtain a statistically reliable estimate of the causal effect of abortion on crime, which model/statistical procedure would you employ here? Can you carefully walk us through the different conceptual steps involved and the code? Estimate this model and explain the results. Are your conclusions qualitatively different from the ones obtained by the original Donohue and Levitt paper?
6. A group of ultra-conservative representatives has already announced to attack your clients' initiative. They hired a different firm to make a case that instead of increasing access to abortions, congress should facilitate access to mobile phones. Rumor has it that they already prepared an empirical study that shows a significantly negative effect of mobile phone penetration on crime rates. Can you replicate their analysis, i.e., take

[2] To make this challenge more manageable, I simplified a few things, compared to the original QJE paper. This should not affect the qualitative findings, however.

your model from Question 2 and replace the abortion rate with the cellphone penetration rate? Do you think the conservative representatives have a plausible argument? How would you help your clients to rebut such an analysis?

7. Finally, think about what else the politicians might learn from the data. What future steps would you suggest to take in order to improve their legislative initiative?