E401/M518: Video Presentation

Propensity Score Matching

Fall 2023

November 19 2023

Please work on this video presentation in group of three students. All presentation assignments 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 some practice with learning new empirical methods that we have not discussed in the lecture. If you pursue a career as an empirical economist or data scientist, you will have to be able to familiarize yourself with newly developed methods constantly. This includes learning enough about the methods to apply it to your day-to-day work and assessing in which circumstances the method is appropriate as well as what its limitations are. 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. For this assignment, I have tried to provide you with reasonably clean data, so that you should not have to spend a lot of time on data preprocessing tasks. However, I haven't checked every detail. Therefore, you should be prepared to have to do minor adjustments to the data if you run into any issues with your analysis. I strongly encourage you to come to my office hour to discuss any roadblocks as well as your overall plan for your video presentation ideally a few weeks before the assignment due date. There is always a risk that you won't be able to understand every detail of the method or the data. 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 an empirical economist or data scientist.

You are expected to upload a video presentation of roughly 25-30 minutes. 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 quantitative analyst. The main focus of the video presentation is to communicate effectively the essence of a new empirical method and illustrate how it works in an application. Anybody who has take none or two econometrics classes and the undergraduate level should be able to follow your presentation. The students refereeing your video should think of themselves as board members who attend your presentation. When you write up your referee reports, you are strongly encouraged to provide critical questions and constructive comments about the presentation. That is, you should provide a brief assessment of how helpful you found the presentation in learning the discussed methods, potentially provide corrections, and ask any remaining questions that you may have either regarding the

method or the discussed application.

Each video presentation will likely have a similar structure and should contain the following elements:

1. Discussion of the method: (1) What is the main idea of the method, (2) which problems does the method solve, (3) how does the method solve the problems, (4) what are the conceptual steps involved when implementing the method, (5) what decisions do you have to make when applying the method to a specific data set, (6) what are some common pitfalls ad when would you not want to apply the method.
2. A brief discussion of the data: Compared to the challenge presentations in class, this part should be very brief. Nevertheless, remember that not everybody who watches the presentation may have looked at the data before.
3. Discussion of the application: (1) What's the big picture business or policy question that you are trying to address? As usual, your viewers probably have not read the assignment questions in advance, (3) discussion of your empirical results, policy implications, and potential caveats and suggestions for further steps.

As with the in-class presentations, this is not a presentation class, so don't invest in fancy video production elements! A fairly bare-bones recording in which you walk the viewers through a couple of slides - you should not need more than 10 for the method explanation- and an RScript that generates all your results as you click through it, and explain it, is totally fine.

Main Techniques

Propensity Score Matching

Key references

I couldn't find a good discussion of propensity score matching in an undergraduate textbook. A good starting point might be the following paper. However, there are several good online resources for learning about propensity score matching. If you're in doubt about a resource, please get in touch with me.

- Dehejia and Wahba (Propensity Score matching methods for nonexperimental causal studies, Review of Economics and Statistics, 2002)

Data

In this assignment you will analyze a typical labor market data set from a randomized experiment and the CPS. The variables are as follows:

- `data_id`: data source of the observation

- `treat`: treatment status, i.e., whether the individual participated in the job training program
- `age`: age of the individual
- `educ`: years of schooling of the individual
- `black`: dummy for African-American individuals
- `hisp`: dummy for Hispanic individuals
- `marr`: dummy for married individuals
- `nodegree`: dummy that is one if the individual does not have a high school degree
- `reXX`: real earnings of the individual in year XX

There are two files. `nsw_dw.csv` contains the individuals from the randomized experiment. `nsw_cps.csv` contains the data from the CPS, which is a representative survey of the whole US population.

Context and policy question

In this assignment, you are asked to quantify the treatment effect of a job training program, specifically, the *National Support Work Demonstration* (NSW) program that was operated by MRDC in the 1970s. The NSW was designed to provide disadvantaged workers[1] who lacked basic job skills to move into the job market by providing temporary employment and training and regular counseling sessions. Workers were randomly matched with relevant positions that provided them with guaranteed employment for 9 to 18 months. Upon completion of the program, workers were left on their own to find regular employment. Those not selected for the program were basically left on their own.

The MDRC collected information on the treatment and control individuals for up to 5 years after completion of the program.

Please answer the following questions:

1. Since the NSW was a randomized experiment, you can easily compute the treatment effect as a simple difference of observed outcomes. Based on the data in `nsw.dta`, which only contains the observations from the experiment, what was the treatment effect of the training on earnings? Given the clean experiment, your estimate should be a good benchmark for the true treatment effect of the program. In the following, I will ask you to see whether commonly used methods for observational data will allow you to recover a similar estimate.
2. Now, work with a more common observational data settings, where you observe your treated individuals and a -potentially very different- control group from a different data source, such as the CPS or PSID. Create such a data set by keeping only the treated individuals from `nsw_dw.csv` and merge the data with the observations from the CPS in `nsw_cps.csv`.
3. Compute the simple difference in observed outcomes for treated and control individuals in your merged data set. How does it compare to the treatment effect estimate from

[1] Among others, the program aimed to help recovering addicts, released offenders, and workers who had dropped out of high school.

the experimental data?

4. Now, use a propensity score matching estimator on the observational data to estimate the treatment effect of the training program on earnings in 1978. Estimate the propensity score with a binary logit regression with the following predictors: a cubic polynomial in age, a quadratic polynomial in education, marriage, no degree, black, hips, re74, re65, two dummies for whether the individual was unemployed in 1974 and 1975, respectively, an interaction term between education and earnings in 1974. Plot a histogram of the propensity scores separately for the treatment and the control individuals. Are you surprised by the graphs? Are you concerned about your propensity scores? Why or why not?
5. Use your estimated propensity scores to compute an average of the treatment effect of the job training program. Discuss and implement two different approaches: (1) inverse probability weighting and (2) nearest-neighbor matching. Compare the two estimates and discuss how they compare to the estimate that you computed from the experimental data.

Hints:

1. For the inverse probability weighting in the last question, you might want to explore trimming the sample by only considering observations with propensity scores between 0.1 and 0.9.
2. There are several packages to estimate propensity score matching in R. I recommend coding this assignment on your own, because it's not that hard and good practice. A ready-to-use package is a good way to check your results, however.