Copenhagen Business School | November 2022 | Applied Econometrics for Researchers

Workshop 5 – Matching Methods

For Workshop 5, you might find the following material helpful to supplement the lecture on **Matching Methods**:

https://blog.stata.com/2015/08/24/introduction-to-treatment-effects-in-stata-part-2/

https://www.ssc.wisc.edu/sscc/pubs/stata_psmatch.htm

Motivation and research question

This workshop investigates the effects of participating in a job training program on individual earnings in 1978.

While in experimental settings individuals would be randomly assigned to the participation in this training program, in observational data this is not the case.

In practice, this implies that certain individuals may be more likely to participate based on certain characteristics, or policy makers may actually target particular (e.g., disadvantaged) individuals when offering this program, aiming at improving their outcomes in the labor market.

This challenges the evaluation of the "treatment effect" of this program. We will therefore apply matching techniques to obtain a more convincing estimate for the impact of participating in this program.

Data & Variables

Use the data file DataMatching.dta available on CANVAS. This dataset is a combination of two sources of data:

- Experimental data used by Dehejia and Wahba (1999)[1] (445 observations, out of which 185 correspond to individuals receiving the "treatment" – i.e., job training program)
- Non-experimental comparison data from PSID (Population Survey of Income Dynamics) (2490 observations)

The variable "data_id" identifies the respective source of data. As a supplement, the dummy variable "experim_sample" distinguishes between the observations coming from the experimental dataset (1) and the additional observations from PSID (0).

The experimental data (experim_sample = 1) will be used for a brief exercise – see question 1. For the application of the matching methods learned in the lecture, we will focus

[1] Dehejia and Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, Vol. 94, No. 448, pp. 1053-1062.

on the individuals participating in the training program (training = 1), together with the comparison group from PSID (who did not receive any training).

The outcome of interest is “re78” (real earnings in 1978). The remaining variables characterize the individuals in terms of age, education, race, marital status, and previous earnings before the program took place (real earnings in 1974 and 1975).

1. **Experimental Data**

Focus on the subset of data coming from Dehejia and Wahba’s (1999) experiment (experim_sample = 1). For this subsample:

- Compare the individuals in the treatment and control group in terms of their age, education, race, marital status, and earnings before 1978. *Hint: conduct t-tests*. What do you conclude?
- How would you interpret these results having in mind the Balancing Condition discussed in class?
- Compare the real earnings in 1978 of the individuals in the treatment and control groups. Is the difference significant? Given the experimental nature of the data, what do you conclude about this estimate?

2. **Observational Data**

Now we will compare the individuals receiving the treatment (job training) with those individuals retrieved from PSID data. Drop the control group from the experimental data *(Hint: drop if treat_experiment == 0)*

- As before, compare the individuals who received training with PSID individuals in terms of their age, education, race, marital status, and earnings before 1978. What do you conclude now?
- How would you interpret these results having in mind the Balancing Condition discussed in class?
- Compare the real earnings in 1978 of the “treated” individuals with those from PSID. Compare with the estimate obtained in exercise 1. What can you conclude from this estimate and why?
- What would you conclude about the potential impact of participating in the job training program if you estimate an OLS regression for re78 on “training” and the remaining covariates used above? Can you infer anything about the treatment effect of training in this setting? Why (not)?

>>>>>>>>>>>>>>>>>>>> Follow-up in class <<<<<<<<<<<<<<<<<<<<

3. **Propensity Score Matching**

- Based on the differences you found in the previous question between treated and untreated individuals, estimate a binary model (logit or probit) for the probability of participating in the job training programs.
- How would you interpret the output and the coefficients? Which variables significantly determine participation into the “treatment”?
- Estimate the Average Treatment Effect on the Treated (ATET) using Propensity Score Matching methods. What do you conclude? Why is the ATET the parameter of interest in this case?
- Repeat the step above imposing a minimum number of 5 nearest matches for each treated individual. How does the estimate of ATET change, and what does this alternative imply in terms of the trade-off between bias and variance?

4. **Balancing and Overlap Conditions**
   - Check the balancing condition *(Hint: use the command “tebalance summarize”).* What do you conclude?
   - Try to improve the quality of the matching by adding a quadratic term for age and an interaction effect between age and “nodegree” in the propensity score equation. Then repeat the estimation of the ATET. Impose a minimum of 2 nearest neighbors in the matching process. What would you conclude?
   - Predict the propensity scores and analyze their distribution *(Hint: predict ps0 and ps1 and type sum ps0 ps1, detail. For that you need to add the option gen(match) in the estimation of ATET. Complement with “teffects overlap”)*. What would you conclude regarding the overlap assumption? Which implications does it have for this analysis?

>>>>>>>>>>>>>>>>>>>>> Follow-up in class <<<<<<<<<<<<<<<<<<<