ECON3360 Causal Inference for Microeconometrics

Tutorial 10: Propensity score matching

Instructor: Julie Moschion

Problem I: The effect of a training program on income

Background This problem set (created by Terra McKinnish) uses data from the evaluation of a U.S training program, the National Supported Work (NSW) program. NSW was a temporary employment program designed to help workers lacking basic skills by giving them work experience and counselling. The NSW program assigned qualified applicants to treatment and control randomly. The treatment group received all the benefits of the NSW program. The control group did not receive any feature of the NSW program. These RCT data were complemented with 2,490 non-experimental untreated individuals drawn from the Panel Study of Income Dynamics (PSID).
In his evaluation of the program, LaLonde (1986) found that non-experimental methods yields biased estimates of the program impact. Heckman and Hotz (1989) show that a more careful set of specification tests can narrow the pool of econometric estimates to ones that are much closer to the experimental results.
In this problem, we use PSM on the PSID sample used by LaLonde to create an acceptable non-experimental control group. This exercise is performed in much more detail in Smith and Todd (Journal of Econometrics, 2005). Use the data in training.dta for the following questions.

(1) To complete this problem set, you need to retrieve a propensity score matching program from the web. Using a computer that is connected to the internet, open STATA and type: `findit psmatch2`. You will be able to click on the links to download and install this program. Alternatively you can type directly `ssc install psmatch2`. You can then type: `help psmatch2` to see the documentation for this program.

(2) Load training.dta data and describe the data. If we are trying to estimate the impact of the treatment on earnings in 1978 while controlling for demographic characteristics and past earnings, what is our main equation of interest?

(3) We focus first on the experimental sample (treated=1 or treated=0). Conduct a balancing test by comparing variable means for the treatment and control groups. If this is a randomised experiment, what do we expect to observe in the results? What do you find? Does this indicate that the experiment was randomised?

(4) Generate the OLS estimate of the effect of the training program on real earnings in 1978 without any additional controls. What do you find? What do you expect to find if you add all available control variables? Run the OLS estimate with controls, what do you find? Create and add an age-squared variable (continue to include it in in subsequent questions when using demographic controls). What do you find? Did adding controls make any meaningful difference to the estimate and the standard errors? Are controls useful when you have experimental data?

(5) Now let's have a look at the PSID comparison group (sample=1 for the experimental sample and 3 for the PSID sample, we do not use the CPS comparison group in this exercise). This is our alternative control group: we want to compare it with the treated sample. Create a variable called treated2 which identifies the treated and non-experimental control group. Run a balancing test. Comment on the results.

(6) Despite the fact that the two samples differ in observed characteristics, it is possible that controlling for these observed characteristics will effectively control for the differences between the two samples. This cannot be tested directly, but we can run regressions of treated2 on the outcome with and without controls to see how the estimate changes. This will give us an idea of how much bias is being soaked up by adding the controls. How do your estimates compare with the treatment effect obtained using the experimental sample?

(7) Controlling linearly for the observed characteristics seems to be insufficient to adequately address the differences between the samples. To address these better we use propensity score matching. Please describe the different steps of a PSM estimation.

(8) Estimate and predict the propensity score using all control variables available. Check and comment on the common support assumption by drawing the propensity score histogram by treatment status (using 50 bins and saving the graph as psm2a).

(9) Run a PSM estimation using the caliper method with a caliper of 0.01 around the treated observations to estimate the impact of training program on earnings. What is the average treatment effect? Change the caliper from 0.01 to 0.005 and 0.025. How does that change your estimate? How do these results compare with the OLS from question 6? If we had a valid non-experimental control group from the PSID what do we expect the average difference in re78 to be between the treatment and PSID samples?

(10) Re-run the PSM estimation with a caliper of 0.025 and imposing strict common support. How many treated observations are dropped now? How do the results compare with the experimental estimates?

(11) Run a simple PSM using the nearest neighbour method. How do these results compare with the caliper method? Find the caliper such that a PSM with caliper produces identical results.

(12) Re-run the PSM estimation with a caliper of 0.025 but using the radius method. How many treated observations are dropped now? How do the results compare with the experimental estimates? Now, try a (epanechnikov) kernel PSM with a bandwidth of 0.025. How do the results compare with the radius estimates?

(13) The PSM estimation simply produces the difference in average outcomes between the treated observations and the weighted control outcomes (i.e with no controls). Our estimates may be able to be improved if instead after the match we run an OLS regression with the weights from the matching and add controls. Re-run the PSM with the caliper method with a caliper of 0.025 (as in question 9). Then run an OLS regression with all the controls and _weight as frequency weights. How do the results compare with the experimental estimates?

(14) Drop the predicted propensity score and re-generate it with all the controls except the age square and the earnings from 1974. Re-run the PSM with caliper of 0.025 and the weighted OLS regression with ALL controls (incl age square and the earnings from 1974). How do these results now compare with the experimental results?

(15) Let's have a look at the quality of this last matching. Run a balancing check on the variables used to generate the propensity score. How do the results compare with those from question 5?