Regression II
Vanderbilt University
Prof. Sean P. Corcoran

September 13, 2022

Lecture 2 In-Class Exercise: Matching

This exercise will apply matching estimators to estimate the treatment effect of participating in a job training program. It uses data from classic papers by Lalonde (1986) and Dehejia and Wahba (1999). Lalonde used data from a randomized evaluation of the National Supported Work Demonstration to contrast non-experimental treatment effect estimates with those obtained from the randomized experiment. Dehejia and Wahba expanded on this work, using propensity score matching methods.

1. Begin with the Lalonde NSW data, which contains data from the original randomized experiment (N=445, 185 treatment and 260 control). Use `regress` or *ttest* to estimate the effect of the job training program on earnings in 1978 (*re78*). Use the same commands to test for balance on pre-treatment covariates (e.g., earnings in 1974 and 1975).

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/nsw_dw.dta
tabulate treat
```

2. We are now going to replace the control group from the experiment with a random sample of workers from CPS and PSID. These workers were not part of the experiment, so all are untreated. Use `set seed` and a random number to fix the random ordering of the observations, for reasons explained later.

```
set seed 1234
use https://github.com/spcorcor18/LPO-8852/raw/main/data/nsw_dw.dta, clear
drop if treat==0
//keep only treated cases for matching

append using https://github.com/spcorcor18/LPO-8852/raw/main/data/cps_controls.dta
append using https://github.com/spcorcor18/LPO-8852/raw/main/data/psid_controls.dta
gen randno=runiform()
sort randno
```

3. Are there missing values of any of the variables? What percentage of individuals in the data participated in the job training program?

4. Use the `regress` command to estimate the effect of the job training program. First do a simple regression of *re78* on *treat*. Then include the following covariates: *age, educ, black, hispanic, re74, re75*. How do these estimates compare to that estimated using the experimental data?

5. Use `teffects nnmatch` to estimate an ATT. Include the following variables in your outcome model: *age, educ, black, hispanic, re74,* and *re75*. Request an exact match for *black* and *hispanic*. How does the ATT compare to that estimated using the experimental data? How balanced are the covariates in the matched sample? (Use `tebalance summarize`).

6. Use the user-written command `psmatch2` to estimate propensity scores and identify nearest-neighbor matches. (You may need to `ssc install` this command). The propensity score represents the worker's predicted probability of participating in the job training program given a set of background characteristics. Use *age, educ, black, hispanic, re74,* and *re75* as confounders. The default propensity score model in `psmatch2` is probit, and nearest-neighbor matches are made with replacement.

   (a) How many treated observations are there?

   (b) How many matched untreated observations are there (and how do you know)? Were any matched observations used more than once?

   (c) How many observations are on the "common support"?

7. Generate a histogram for the estimated propensity scores. Now repeat this step but for propensity scores > 0.01. Explain why the latter is more useful.

8. Visually inspect the distributions of propensity scores for the treated and untreated observations. You can use the command `psgraph` or (even better) code like the following:

```
histogram _pscore if _pscore>0.01, kdensity kdenopts(gaussian) by(treat, cols(1) ///
   legend(off)) ytitle(Frequency) xtitle(Estimated Propensity Scores)

//or:
twoway (histogram _pscore if treat==0 & _pscore>0.01, bin(20) fcolor(none) ///
    lcolor(blue)) (histogram _pscore if treat==1 & _pscore>0.01, bin(20) ///
    fcolor(none) lcolor(red))
```

9. Use the `pstest` command to check the balance on covariates and interpret the results. Use the overlapping histograms syntax above to compare the distributions of *re74* in the matched sample.

10. Estimate the ATT by including the `outcome(re78)` option in the `psmatch2` command. Make note of the results, and explain in words. NOTE: the standard error for the ATT in `psmatch` is incorrect. When estimating treatment effects via matching, use the more recent `teffects`.

11. Estimate a simple regression of *re78* on *treat* but use weights to appropriately account for the fact that some matched observations are used more than once: `[pw=_weight]`. How does the estimated treatment effect differ from the one in (10)?

12. Try estimating a multiple regression where *re78* is the outcome and participation in job training *(treat)* is the explanatory variable of interest. Use the same set of controls as above. How does the estimated treatment effect differ from the one in (10)?

13. Now repeat part (10) but use bootstrapping to estimate standard errors. This is just an illustration—recent papers argue that bootstrapping should not be used for nearest neighbor matching.

```
bootstrap att=r(att), rep(1000): psmatch2 treat age educ black hisp ///
   re74 re75, outcome(re78)
```

14. As a variation on (10), try using 5 nearest neighbors.

15. Repeat part (10) using `teffects psmatch`. Use the associated commands `tebalance summarize` and `teffects overlap` to look at balance in the covariates and the distribution of propensity scores.