

ECON 8686 Problem Set #6- Propensity Score Matching

This problem set makes use of the National Supported Work (NSW) data used by LaLonde (1986) and Heckman and Hotz (1989). Recall that LaLonde uses data from the NSW experiment to show that non-experimental estimates generated through standard econometric analysis misestimate the program impact. Heckman and Hotz show that a more careful set of specification test can narrow the pool of econometric estimates to ones that are much closer to the experimental results. In this problem set, we will attempt to use propensity score matching to create an acceptable non-experimental control group for purposes of estimating the effect of the training program. We will attempt to construct this control group from one of the PSID samples used by LaLonde.

This exercise is performed in much more detail in Smith and Todd (*Journal of Econometrics*, 2005)

The data set Training.dta contains:

treated: 1 for experimental treatment group

0 for experimental control group

missing for the comparison group generated from the PSID.

age

educ: years of schooling

nodegree: binary, =1 if no high school diploma

black: binary, =1 if black

hisp: binary, =1 if Hispanic

married: binary, =1 if married

re74: real earnings in 1974

re75: real earnings in 1975

re78: real earnings in 1978

0) 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:

```
net search psmatch2
```

You will be able to click on the links to download and install this program.

You can then type

```
help psmatch2
```

to see the documentation for this program. It will be helpful to look over this before you do the problem set.

1) We will focus first on the experimental sample (treated=1 or =0). Compare variable means for the treatment and control groups.

If this is a randomized experiment, what do we expect to observe in the results?

2) Conduct ttests for differences in means (do not need to report the results of each one in your problem set answers). For example:

```
ttest age, by(treated)
```

Is there a statistically significant difference in any of the characteristics?

Does this indicate that the experiment was not randomized?

3) Generate the experimental estimate of the effect of the training program. Regress re78 on treated. Then regress re78 on treated, the demographic controls and the past earnings controls. (Create an age-squared variable and continue to include it in all cases where you are using demographic controls).

What is the estimated effect of the program?

Did adding the controls make a difference in the estimate (and how much of a difference given the standard errors)?

Why would you want to add controls even in the case where you have experimental data?

4) Now we can consider the quality of the PSID comparison group. Create a variable called NSW that equals 1 for all experimental observations (treatment and control) and 0 for all PSID observations. Compare means of the experimental and PSID samples.

5) 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. One way to test this is to regress re75 on NSW and then regress re75 on NSW and all of the controls (including re74).

Discuss the results.

6) If controlling linearly for the other observed characteristics does not adequately address the differences between the samples, one option is to experiment with richer specifications. One could put in lots of nonlinear (such as polynomial) terms and interaction terms to see if the coefficient on NSW will go away. This, however, still would not address the issue of common support. We will, instead, pursue propensity score matching.

Estimate:

```
Probit nsw educ nodegree black hisp age agesq married re74 re75
```

```
Predict phat
```

Then:

```
Summ phat if nsw==1, detail
```

```
Summ phat if nsw==0, detail
```

```
Histogram phat, start(0.0) width(0.05) by(nsw)
```

Discuss your findings with regard to common support.

7) There are two different ways we could proceed:

a) In the first, we would drop the experimental control group. Then we could propensity score match between the experimental treatment group and the PSID sample, and then use the differences in re78 to estimate the training program impact. If we create a valid non-experimental control group from the PSID using the matching, what do we expect the average difference in re78 to be between the treatment and PSID samples? Do we know the *exact value* of the expectation of the difference?

b) In the second, we would drop the experimental treatment group. We would then propensity score match between the experimental control group and the PSID sample. If we create a valid non-experimental control group from the PSID using the matching, what do we expect the average difference in re78 to be? Do we know the *exact value* of the expectation of the difference?

c) What is the sample size of the experimental control group? Of the experimental treatment group?

8) Drop the experimental treatment group. Your answers to question 7 above hopefully explain why this is a slightly preferred approach. Regress re78 on nsw and demographic and previous earnings controls:

```
reg re78 nsw educ nodegree black hisp age agesq married re74 re75
```

9) Perform nearest neighbor propensity score matching on the common support:

```
psmatch2 nsw, outcome(re78) pscore(phat) common
```

Ignore the “sort order” warnings.

The common command drops members of the treatment group (in this case nsw=1) that are not on the same support as the control group (in this case, nsw=0).

Describe who gets dropped using the _support variable that is created by the psmatch program.

Discuss the propensity score matching results. How do they compare to the results from question #8 above?

10) Now estimate:

```
reg re78 nsw educ nodegree black hisp age agesq married re74 re75 [fweight=_weight]
```

This is regression adjusted propensity score matching using the _weight variable that was generated by the previous propensity score matching command. How do the results compare to those obtained in question #8 above?

11) Let’s investigate how well propensity score matching eliminates differences in re75 compared to the results to question #5 above.:

```
psmatch2 nsw, outcome(re75) pscore(phat) common
```

```
reg re75 nsw educ nodegree black hisp age agesq married re74 [fweight=_weight]
```

Discuss the results.

12) Do kernel estimation propensity score matching :

```
psmatch nsw, kernel outcome(re78) pscore(phat) kerneltype(normal) bwidth(2)
```

```
psmatch nsw, kernel outcome(re78) pscore(phat) kerneltype(normal) bwidth(0.02)
```

These are both gaussian kernels. The smaller the bandwidth, the less the weight placed on observations with a greater difference in propensity score. Smaller bandwidth reduces bias, but increases variance (b/c fewer observations used to estimate the effect).

Discuss the results.