ECONOMICS 717
APPLIED MICROECONOMETRICS
Department of Economics
University of Wisconsin-Madison
Fall 2022

Problem Set: Nonparametric Regression, Matching and Weighting
Version of February 25, 2022
Due on Canvas at 11:00 AM Wednesday, March 9, 2019

Introduction / conceptual background

This problem set uses data from the National Supported Work Demonstration (NSW). This is the same data used by LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999,2002) and Smith and Todd (2005) and many others.

The idea of all of these papers is that you can learn about combinations of non-experimental identification strategies and estimators by using experimental data to provide a benchmark. The experimental data provide a treatment group and a control group. These studies add in a non-experimental comparison group. For a given non-experimental strategy, there are two ways to examine its performance at reducing bias. One way is to construct a non-experimental estimate of the treatment effect using the treatment group and the comparison group and compare it to the experimental estimate of the treatment effect constructed using the treatment group and the control group. The second way is to construct an estimate of the bias using the control group and the comparison group and to compare that estimate to zero. That is the approach taken in Smith and Todd (2005) and in this problem set.

Answers

Your write-ups for the problem sets should consist of two portions. The first portion is just the answers to the questions, with whatever text is required to explain them. This portion must be typed. The second portion, on separate pages, consists of a Stata log file (or equivalent from another program) that shows how you got the answers to the empirical questions. The log file must be clear and must include comments that will allow the grader to quickly see the command or commands leading to each answer. It should not include everything you tried – just the final set of commands employed to get the answers. See the syllabus for more details on the format.

Data Set: NSW

File name: "Economics 717 Fall 2022 NSW Data.dta". The dataset contains the following variables:

sample –
- 1 for the experimental sample (the union of the treatment and control groups)
- 2 for the comparison group from the Current Population Survey (CPS)
- 3 for the comparison group from the Panel Study of Income Dynamics (PSID)

treated –

1 for the experimental treatment group
0 for the experimental control group
Missing for everyone else

age – age in years
educ – years of schooling
black – 1/0 black
hisp – 1/0 Hispanic
married – 1/0 married
nodegree – 1/0 no high school degree
re74 – real earnings in "1974"; see Smith and Todd (2005) for discussion.
re75 – real earnings in 1975
re78 – real earnings in 1978
dwincl – 1/0 included in the Dehejia and Wahba sample.
early_ra – 1/0 included in the early random assignment sample in Smith and Todd (2005)

Software: psmatch2.ado for Stata

For a subset of the problems you should use the program psmatch2.ado by Barbara Sianesi and Edwin Leuven. You can install this software on whatever computer you are using (so long as it is connected to the internet) by typing:

```
net search psmatch2
```

in Stata and following along from there.

You can then type "`help psmatch2`" to learn about the workings of the program. Read this document carefully.

Problems Using the NSW Data

Note that for the purposes of this problem set, you may ignore choice-based sampling.

0. Drop the observations from the PSID comparison group.

1. Generate an experimental impact estimate by running a regression of earnings in 1978 on the *treated* variable along with age, age squared, education, black, Hispanic, married, no degree and earnings in "1974" and 1975. You should use only the experimental treatment and control groups (which will actually happen automatically because the *treated* variable is only defined for them). Explain why you might want to include covariates in this regression even with experimental data.

2. Drop the experimental treatment group.

3. Estimate two sets of propensity scores using a probit model. The first set should include the variables age, age squared, education, black, Hispanic, married, and no degree. These are the "coarse" scores; call them *pscorea*. The second set should contain the variables in the first set

plus earnings in "1974" and 1975. These are the "rich" scores; call them *pscoreb*. Use only the experimental control group and the CPS comparison group to estimate the propensity scores. Explain what is going on with the observations that are "completely determined".

4. Examine the distributions of estimated propensity scores in the control group and comparisons group samples. What do the descriptive statistics suggest about the common support condition in these data? What do they suggest about the comparability of the CPS comparison group?

5. Construct a histogram of the estimated propensity scores for the combined experimental control group and for the CPS comparison group. Using a command such as:

```
egen bins=cut(pscore), at(0(.05)1) icodes
graph bar (count) pscore, over(d) over(bins, label(nolab)) asyvars
```

will make it easy to compare the histograms, where "pscore" is the variable name for the estimated propensity score and where "d" is a variable equal to one for the control group sample and equal to zero for the comparison group sample. What do the histograms suggest about the common support condition in these data? What do they suggest about the comparability of the CPS comparison group?

6. Construct non-experimental bias estimates for both sets of estimated propensity scores using single nearest neighbor matching *without* replacement. Impose the common support condition using the min-max version described in the lecture notes. This can be accomplished using the "`common`" option to `psmatch2`. [Although you would normally want to use the "`ties`" option, do not use it here as doing so makes later problems difficult because it generates fractional weights.] How many observations are dropped by imposing the common support condition? Which observations get dropped? [Hint: `psmatch2` produces a variable called "_support" that will be useful here.] How close are the resulting estimates to the experimental impact estimate? Do the rich scores that contain pre-treatment earnings perform better (i.e., result in lower bias estimates) than the coarse scores that do not?

7. Repeat Problem 6 but using single nearest neighbor matching *with* replacement. How much does allowing comparison group observations to be reused in the matching change the estimates in this context? Explain.

8. Estimate the standardized difference in "real earnings in 1974" and "real earnings in 1975" based on the raw data and based on the rich scores and single nearest neighbour matching with replacement. What is the proportionate reduction in the standardized "bias" from the conditioning?

9. Create propensity score matching estimates using the rich propensity scores and kernel matching with a Gaussian (normal) kernel and bandwidths of 0.02, 0.2 and 2.0. Impose the common support condition as in Problem 6. Describe the resulting impact estimates. How do the estimates change as the bandwidth increases? How do the estimates differ from the single nearest neighbor matching with replacement estimates obtained in Problem 7?

10. Repeat Problem 9 but using local linear matching rather than kernel matching. How do the estimates change as the bandwidth increases? How do the estimates differ from the single nearest neighbor matching with replacement estimates obtained in Problem 7 and the kernel matching estimates obtained in Problem 9?

11. Obtain an estimate of the bias by estimating a linear regression of real earnings in 1978 on the variables in the rich propensity scores and a treatment dummy using all of the observations in the control group and the CPS comparison group. Contrast this estimate with those obtained above using matching methods.

12. Obtain an estimate of the bias by estimating a linear regression of real earnings in 1978 on the variables in the rich propensity scores using *only the untreated* observations. Use the predicted values from this regression, evaluated at the covariate values associated with each treated unit, as the estimated expected counterfactual outcomes for the treated units. Contrast the bias estimate obtained using this method with that obtained in the previous problem. Discuss why these two estimates might differ.

13. Obtain an estimate of the bias using inverse probability weighting and the rich propensity scores. Implement this estimator "by hand" (i.e. code it up yourself in Stata) and do it in two ways: rescaling the weights to sum to one and not rescaling the weights to sum to one. Compare the two estimates based on inverse probability weighting to each other and to those obtained using the various matching estimators.