Problem set 2: Empirical Methods for Applied Microeconomics

General instructions. Please work in a group no larger than 3. When you write up your results, please let me know who is in your group. (Only turn in 1 completed homework.). Present your answers in a concise way (typed is highly preferred). Please include relevant Stata output and well-commented do files and ado files for all the exercises (or equivalent in the package of your choice.) Please do NOT include lots of undigested log files.
Put the do files in an appendix and make clear reference to the regression output and/or figures.

This is due Wednesday November 5 at the end of class.

Problem 1

Propensity score weighting.

There is a long standing debate about whether social programs or other interventions can be evaluated without use of data from a randomized control experiment. One of the early entries in this debate was Bob LaLonde's 1986 AER paper which looked at various non-experimental estimators, using as the comparison experimental estimates. The experiment was the National Supported Work Demonstration, conducted from 1975–77, and Lalonde used nonexperimental data from the CPS and PSID. The NSW looked at the effects of training on disadvantaged groups (female welfare recipients, former drug addicts, parolees, and high school dropouts). Participants had to be unemployed, with little work experience. Obviously, this means these groups were quite different from the general population. Lalonde showed that even using selection adjustments, control groups from the observational data were not able to replicate the experimental findings. Heckman and Hotz (1989) use various tests which exclude the most biased of the estimators from the Lalonde paper.

In this problem, we will use the National Supported Work Demonstration data used in Lalonde (1986). Part of the exercise will be to replicate some of the findings from the Smith and Todd (2005) *Journal of Econometrics* paper which assessed claims in Dehejia and Wahba (1999, 2002) that propen-

sity score matching could be used to replicate closely the experimental estimates. (See the reading list for that and the Lalonde and Dehejia and Wahba papers.) Recall the assumptions under which propensity score matching or inverse propensity score weighting can lead to causal estimates.

Download three data sets from http://users.nber.org/ rdehejia/nswdata2.html; the original experimental data (either nsw.dta, or the combined nsw_treated.txt and nsw_control.txt), the DW sample from their paper (either nsw_dw.dta or the combination of nswre74_control.txt and nsere74_treated.txt.), and the broadest version of the CPS control group (cps_controls.dta or cps_controls.txt).

For the experimental contrasts, you will either use the original NSW Lalonde sample, or the DW sample. (That is, there are 2 possible experimental contrasts, the original Lalonde one, and the DW one.)

For the attempts to use non-experimental data to replicate the experimental findings, you will use one of the original treatment groups combined with the CPS non-experimental control group.

(i) First, replicate the relevant columns of Smith and Todd table 1 with the Lalonde sample, the DW sample, and the CPS control group. (Note you don't have the 1979 earnings.)

(ii) Next, test whether the experiment "worked". Are the means of the Xs different in the treatment and control groups? Do an overall test for the Xs being different.

(iii) Based on the time pattern of earnings in the experimental samples in 1974 and 1975, do you see evidence of Ashenfelter's dip for these participants? (Ashenfelter's dip is the empirical regularity that those who participate in training or employment programs typically experience a decline in earnings prior to participation. Why would this complicate evaluation using observational data?)

(iv) We will also start by evaluating the simplest possible comparisons. Calculate the basic treatment control difference in the cross-section for 1978 for 3 samples: the full Lalonde sample, the DW sample, and the combination of the Lalonde treatment group and the CPS 1 control group.

Why do you think the CPS 1 control group fails so miserably?

(v) Now we will add some Xs on the RHS to adjust for differences in observables. Choose some of the controls.

Does adding the Xs affect the experimental estimates? Should it?

What Xs help the most with reducing the bias with using the CPS control group?
Why might matching on the p-score do better than a linear regression in reducing bias?

(vi) Now, estimate a modified version of the DW propensity score logits (Table 3) for the CPS 1 control group and the Lalonde and DW experimental groups. The dependent variable is a dummy for being from the experimental data, and zero if in the CPS 1 control group. Controls are age, age squared, age cubed, education and education squared, a dummy for being a high school dropout, a dummy for being married, a dummy for being black, a dummy for being Hispanic, real 1974 earnings, real1975 earnings, a dummy for zero earnings in 74 or 75, the interaction of schooling and real earnings in 1974. [Obviously, since we don't have 74 earnings for the Lalonde sample, we can't exactly replicate their table 3 column 1.]

Do the coefficients here make sense?

Does it matter if you only use the DW treatment group or use both the DW treatment and control group as the sample for which the experimental data dummy is 1?

(vi) Plot the predicted log odds ratios for the logits for the specifications which include all the

experimental observations in each the DW and Lalonde samples. Plot the logodds for the experimental and comparison groups separately. How is the overlap here?

(vii) Now we will look at the bias. Smith and Todd address the bias by comparing the experimental control group (who couldn't get the treatment) with the CPS and PSID control groups. Differences in these should be zero. Estimate a nearest neighbor p-score impact using the Lalonde sample and choose a reasonable way of selecting a common support (DW, or what Smith and Todd do is fine). Calculate cross-sectional matching differences. (Analogous to Table 5, row 3). Use the relevant logit generated p-scores for the relevant samples. Implement a nearest neighbor match. Do you think DW are right that the matching takes care of selection?

(viii) Implement a differences in differences matching estimator. (This will subtract a cross-sectional matching estimate of post-RA bias from one of pre-RA bias.) Does this change your view of DW's claims?

(ix) What do you think now about selection on observables in this setting (1/2 page only)?