HCMG 901 Problem Set 1
Spring 2025
Due: Feb 14th

Basic and matching theory:

1. Using properties of the conditional expectation function, show that E(Y|X) is the minimum mean squared error (MMSE) predictor for Y|X.
2. Discuss the evaluation problem and the interpretation of selection bias using the potential outcomes model
3. Discuss the identifying assumptions needed to estimate the ATT and ATE using OLS or matching. How does assuming common overlap help?
4. If we think of matching as a 2-step approach, which step is more important in reducing bias and why? Provide the intuition.
5. Angrist (Ecma 1998) shows that the ATT estimated by OLS and nonparametric matching can be expressed as weighted averages of x-cell specific treatment effects, δ_x. Derive and compare these weights. Discuss the differences.
6. Propensity score theorem: Show that, under the Conditional Independence Assumption (CIA), $Y_0, Y_1 \perp D \mid p(X)$, where $p(X) = p(D = 1|X)$

Matching empirics:

This part of the problem set asks you to work with the datasets used in the LaLonde (LL, 1986) and Dehejia-Wahba (DW, 2002) papers. We will replicate some of the tables/figures in those papers as well as do additional analyses. Use the NSW, NSW-DW, CPS-1 and PSID-1 files for this exercise.
Note: (1) the files only contain data on the male participants and controls from the NSW experiment.
(2) In the past, **students have not been able to replicate the numbers in DW2002**, but they get somewhat close and replicate the qualitative patterns. So please do not worry about getting the numbers exactly right.

1. Prepare a table of descriptive statistics on key X variables and earnings for treated and comparison groups using each of these data files. The different variables should be in the rows, and the different data files span the columns. Briefly comment on the differences and what concerns you have about using CPS/PSID to replicate the experimental estimates.
2. Replicate results in rows 1 (experiment controls), 2 (PSID-1), and 5 (CPS-SSA-1) from Table 5 of LL. Ignore column 10. Use robust standard errors (regardless of what the paper uses). In this step you will not use the NSW-DW file.

Now, we will implement the propensity score matching (PSM) method following DW2002. Use the NSW-DW sample to obtain the treatment group observations, and CPS-1 and PSID-1 files as the non-experimental control groups.

3. Before diving into the empirics, briefly discuss whether PSM can be helpful in this setting. In class we discussed HIT1997 and their suggestions for when matching can perform well. Are those conditions satisfied by the data available here?
4. Estimate the propensity score separately for CPS and PSID samples as in the DW2002 paper. Discuss briefly the steps you took and the final specification you settled on. (It's ok if it is slightly

different than what they had in DW. They mention their specification in the notes to Table 2). Present the mean values of the Xs for the treated and matched comparison groups.

5. Construct the equivalent versions of figures 1 and 2 from DW. A simple way to think about overlap is the % of treated units that fall in the overlap range of p-scores between the treated and comparison groups. Which dataset (CPS or PSID) appears to have more overlap with the treatment group from the experiment?
6. Replicate Tables 2 and 3 from DW (you won't get them perfectly). No need to present the columns on means of X variables. Obtain standard errors using bootstrap (100 reps is sufficient).
7. Comment on the performance of matching with replacement versus matching without replacement in this empirical application. How does it differ between the CPS and PSID samples?
8. Estimate simple OLS estimates of the training effect using the matched samples obtained with and without replacement. How do they compare with the corresponding PSM estimates?
9. Iacus, King and Porro (2012) claim that coarsened exact matching (CEM) has many superior properties to PSM. Discuss some of these advantages. Use the "cem" command to implement CEM with the NSW-DW and CPS samples. Match on the same variables you used in the PSM approach, but now you have to match on coarse categories, which you can decide yourself. Imagine that you had to add a row to Table 2 for the results using CEM. Comment on the results obtained using CEM relative to those from the experiment and from PSM.

Next, we will do some additional analysis, beyond what was in the DW paper. Specifically, instead of using the NSW-DW sample, go back to the full NSW sample from the LL paper.

10. Discuss the main differences you observe between the NSW and NSW-DW files. Feel free to review the Smith and Todd (*J Metrics*, 2005) article if you would like more background on the differences.
11. Re-estimate propensity score for the full NSW sample along with the CPS-1 file. Did you have to change the specification to get a valid propensity score?
12. Create an equivalent version of Table 2 from DW, but with the full NSW sample. No need to present mean X values.
13. Does the performance of PSM drop using the full NSW relative to the NSW-DW file?