Homework #3

Robert Ackerman

November 15, 2013

1

Table 1: Means, Standard Deviations, and T-test Results for Experiment Sample

Variable	Treatment	Control	t-stat	p-value
Age	24.63	24.45	-0.36	0.72
	(6.69)	(6.59)		
Years of School	10.38	10.19	-1.50	0.14
	(1.82)	(1.62)		
Proportion	.73	.81	2.67	0.01
High School Dropouts	(0.44)	(0.39)		
Proportion Married	.17	.16	-0.38	0.70
	(0.37)	(0.37)		
Proportion Black	.80	.80	-0.05	0.96
	(0.40)	(0.40)		
Proportion Hispanic	0.09	0.11	0.80	0.43
	(0.29)	(0.32)		
Real Earnings	$3571	$3672	0.22	0.83
2 Years Before	(5773)	(6522)		
Training	[335]	[316]		
Real Earnings	$3066	$3027	-0.10	0.92
1 Year Before	(4875)	(5201)		
Training	[283]	[252]		
Number of				
Observations	297	425		

Note: Numbers in parentheses are standard deviations, and those in brackets are standard errors.

At the 5 percent level, we reject the null hypotheses that the means are equal for only one variable: Proportion of High School Dropouts (p-value: 0.01). If this was a truly randomized

experiment, we would expect that we would fail to reject for all of the variables (assuming we had a large enough sample). In a truly randomized set up, the selection into treatment/control is by design independent of observables and hence we would expect both groups to be identical in terms of these observables (again assuming a large enough sample size). Note: My point estimates match up identically with those in table 1 in Dehejia and Wahba (1999) although my standard errors are different.

We may want to include the covariates for several reasons. First, even though we are in the experimental setup, we did find that we're rejecting the null hypothesis that proportion of High School Dropouts is the same in the treated and control group. Furthermore, even if we don't think the differences are statistically significant between the treated/control they are not perfectly the same so we could want to include them for this reason. A second reason is that we may be interested in how these covariates are impacting earnings in 1978 in addition to just the treatment is impacting earnings.

2

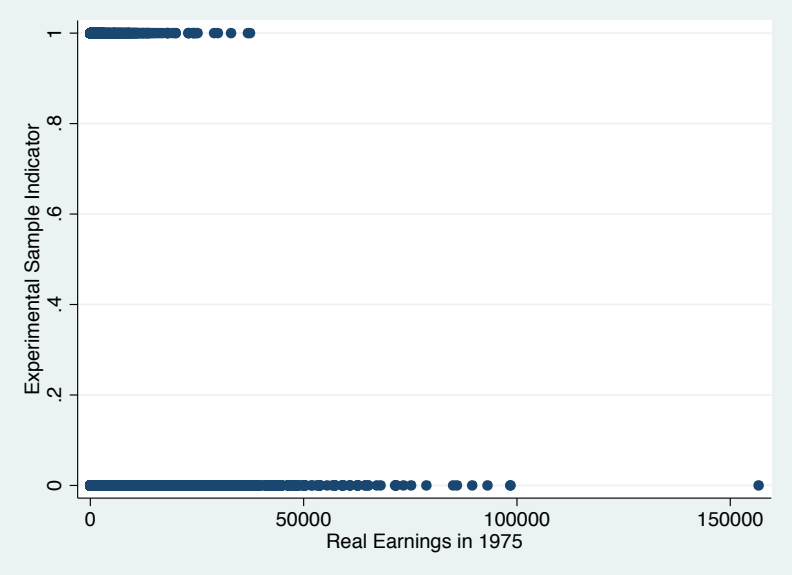
When we test the means of the experimental and PSID comparison groups, we reject the null hypotheses that they are equal for all of our variables. Simply using the PSID sample as the comparison group is a terrible idea, given that the two groups are statistically different in regards to all of our variables at all common levels of significance. While we could argue that selection into treatment/control is truly random, clearly selection into the experiment itself is not random. This is not surprising given the design of the program was to improve labor outcomes for a very specifically targeted groups chosen explicitly on observables. When I compare my estimated means for the PSID comparison group to the results for PSID-1 table 1 of Dehejia and Wahba (1999), I find that the point estimates are nearly identical but again my standard errors are different.

If we are going to use propensity score matching, we need a conditional independence assumption (CIA). To have unbiased estimates we need the assumption that assignment into treated/untreated is random conditional on observables i.e. unobservables are not affecting outcomes. Formally: $\{Y_{i1},\ Y_{i0} \perp T_i | X_i\}$. We also need common support in order to be able to match individuals. If we have no "overlap" we won't be able to match.

3

When we include the two measures of earnings for 1974 and 1975, we have 135 "failures" that are completely determined. This means that, for 135 observations including these measure of earnings allows us to predict with near certainty that they will not be in our experiment group. It appears that these are individuals with earnings above a certain level. In particular, given

the nature of our experiment group it is not surprising that individuals with high incomes are not likely to be part of the program. We can see this in the following plot:



After a certain cutoff level of earnings, none of the observations are in the experimental group. You get a similar result if you use earnings from 1974.

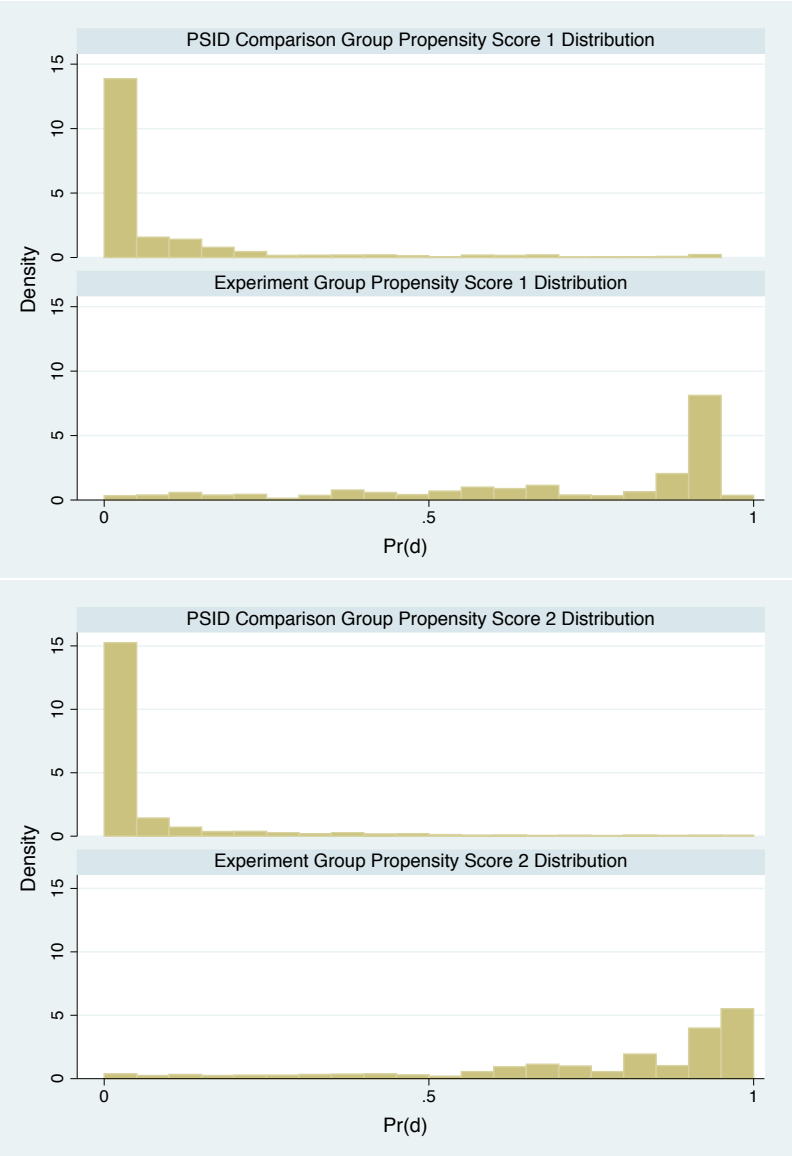
4

The descriptive statistics suggest that our common support condition is in jeopardy. For score set 1 the mean propensity score for the experiment group is 0.71 and 0.08 for the PSID comparison group, indicating that something is wrong in terms of the observables matching between the two. A look a skewness shows -1.0 for the experiment group and 3.0 for PSID comparison, indicating large skewness in opposite directions. If we look at the percentiles, we can see that over 75 percent of the observations in the PSID comparison group have a propensity score below 0.08, where less than 5 percent of the observations in the experiment group have a score below 0.08. Therefore we are looking at very little overlap between the two. This confirms the analysis from question two that using this entire group as a comparison is a bad idea, and suggests that even with matching we don't have a wide range of common support.

When we do the same analysis for score set 2, we get the same general picture although now we are looking at an even smaller range of common support. There is an even more pronounced difference between means (0.76 vs. 0.7), skewness(3.28 vs. -1.43), and the percentile makeup (over 75 percent below 0.04 vs less than 5 percent) of the distributions for the experimental and PSID comparison groups.

5

When we construct the histograms of estimated propensity scores for the experiment and PSID comparison groups, we confirm the analysis from the pervious question. In particular it looks like there is little common support between the two.



Again, it is important to have common support because we want to be matching individuals with individuals that are similar based on observables. If we don't have any overlap, we don't have individuals to match or with only a bit of overlap we're matching with individuals that are not that close in terms of observables.

6

If the groups were comparable, we would expect there to be no impact on 1978 earnings from begin in the experiment control group and the PSID comparison group. The "treatment" is whether or not they entered into the experiment itself. Since we're comparing with the control group within the experiment, we'd expect that if we had random selection both into the experiment itself and then into treatment/control, that there should be no difference between the PSID and experiment control groups. This is clearly not the case from our results. From the score set 1 specification, we find a large (-9,868) and statistically significant decrease in 1978 earnings from being in the experiment control group. However this difference is significantly smaller than for the unmatched which is (-16,464) Likewise we find a large (-6,077) and statistically significant impact under score set 2. So our score set 2 specification is still biased, but now less so.

7

Allowing for replacement greatly improves the estimates. For score set 1, I now find a difference of -5,097 which is still statistically significant. For score set 2, I find a difference of -1,046 with a corresponding t-statistic of -0.73 which does not lead us to reject the null hypothesis that the means are equal.

For score set 1, I find that compared to the unmatched estimates matching is reducing bias from the unmatched by 28 percent for quartile one, increasing bias by 33 percent for quartile two, decreasing bias by 5 percent in quartile three, and increasing bias by 16 percent in quartile four. Obviously there is a balance problem, and matching is having a different impact on bias depending on where in the distribution of propensity score that matching is occurring.

For score set 2, I find that compared to the unmatched estimates matching is reducing bias from the unmatched by 33 percent for quartile one, reducing bias by 16.5 percent for quartile two, reducing bias by 54 percent in quartile three, and reducing bias by 34 percent in quartile four. So there does not seem to be this same balance issue when we use score set 2, because we are reducing bias in all quartiles.

8

When we are bootstrapping standard errors, I find that in general I am getting smaller standard errors using either 10 or 100 replications than the standard errors from using the entire sample. However, depending on the samples it is drawing I get different implications for whether I'm larger or smaller standard errors as I increase the number of replications. Saraswata has not taught us bootstrap yet, so my understanding of this technique is very limited.

9

I am getting an estimate of -1046 using matching alone, and -833 for this regression adjusted matching estimate. So this adjustment is decreasing bias, and improving the estimate.

10

As we are increasing the bandwidth of the kernel, we can see the trade-off between bias and standard errors. As we increase the bandwidth our standard errors are decreasing, but our bias is increasing. In particular we have estimates and standard errors of -1,577 (2009), -5,696 (1266), and -16,452 (427) for bandwidths 0.02, 0.2 and 2 respectively. With a bandwidth of 2, we're not doing much better in terms of bias compared to the unmatched estimates (-16,464). However, with a bandwidth of 0.02, we're doing pretty well although not as well as with nearest neighbor with replacement.

11

The local linear matching with bandwidth 0.02 gives an estimate of -1,487 which is slightly better than the Gaussian kernel estimate with the same bandwidth, although still worse than nearest neighbor with replacement. As we increase the bandwidth it doesn't do as poorly as the Gaussian kernel estimator in regards to bias (estimates of -2,267 and -2,632 for bandwidths 0.2 and 2 respectively).