LaLonde Benchmark Replication

Kellie Ottoboni

February 11, 2016

LaLonde (1986) combined experimental data with observational data to test whether propensity score matching could be used to estimate average causal effects accurately. He used data from the National Supported Work (NSW) Demonstration in 1975, a randomized, controlled experiment done to study the effect of a worker training program for unemployed men. This data allows us to unbiasedly estimate the average causal effect of the program on earnings of men in the study. We'll call this estimate of the effect of the program on the participants' earnings in 1978 the "experimental benchmark". Taking this estimate (or an interval around this estimate) as the "true" average causal effect, we can measure how well other estimators recover this effect. The experimental benchmark estimate is 1794.34.

We have two nonexperimental datasets: the Current Population Survey (CPS) and the Population Survey of Income Dynamics (PSID). LaLonde created three subsets of these datasets to use as observational control groups. (The way in which he created these subsets is unknown, as LaLonde never recorded it and forgot himself what he did.) For each of these six datasets, we swap out the NSW control group for the observational controls and estimate the effect of treatment. When using matching methods, we estimate the average treatment effect on the treated (ATT); otherwise, we estimate the average treatment effect overall (ATE).

	Unadjusted	MM Pairs	Pscore Pairs
CPS1	-8497.52	1499.03	1940.26
CPS2	-3821.97	518.91	1907.95
CPS3	-635.03	435.58	1225.01
PSID1	-15204.78	1877.82	1093.39
PSID2	-3646.81	1024.47	1387.11
PSID3	1069.85	1099.50	1306.29

Table 1: Estimated effects using observational control groups

First, we look at the unadjusted difference in means. This estimator is the one we use in the experimental data. Its unbiasedness relies on randomization: the random assignment of workers to the training program or no program ensures that treatment is independent of all covariates. In the observational data, individuals do not participate in worker training programs, likely due to their education, socioeconomic status, and other factors which make their covariates associated with their lack of treatment. The first column of Table 1 shows the unadjusted differnece in means between the NSW treated and each observational control group. The estimates are large and negative. This is because the estimates don't reflect a causal effect; the individuals in the observational studies on average have higher incomes to begin with, and so also have higher incomes in 1978.

Next, we use model-based matching to estimate the average treatment effect on the treated. We estimate individuals' earnings in 1978 using all covariates but the treatment using a fully-saturated linear model with linear, quadratic, and interaction terms. Since we are not using the model to do inference, but rather to create a score on which to match, we're not terribly concerned about overfitting. Then, we match each treated individual to a single control based on their predicted earnings from this model. Controls are matched with replacement, so a control unit may be matched to more than one treated unit.

Assuming unconfoundedness between treatment and potential outcomes, the difference in means within pairs, averaged over pairs, is an unbiased estimate of the ATT. The second column of Table 1 shows these estimates.

Finally, we use propensity score matching to estimate the ATT. We estimate the propensity scores using logistic regression with the propensity score model specifications from the footnotes of Table 2 and Table 3 of Dehejia and Wahba (1999). Then, we do one-to-one matching with replacement, as we did for model-based matching, and estimate the ATT using the average difference between treated and control pairs. The third column of Table 1 shows the estimates. We are not able to recover Dehejia and Wahba's estimated ATT of 1559 and -919 for the full CPS and PSID groups, respectively.