Université Paris-Saclay
Master of Economics – 1st year
Advanced Econometrics – Tutorial #5
Matching Estimators *

March 28th 2022

Tutorial context

Randomized experiments cannot always be implemented. Thus, estimating effects from microeconomic policy, such as the effect of job training program on earnings, must deal with **observational studies and data**. Several techniques have been developped, among which the **matching method** using propensity score. The matching technique is a **systematic method for creating comparison groups with similar covariate values than the treated individual**s, which can then be used to average the differences with the treated individuals and identify a causal effect.

One way to assess the ability of such an econometric method to solve the evaluation problem is to try the method in a situation where an experimental estimate is available. This is the strategy in LaLondes classic 1986 paper and in the papers that reanalyze his data: Dehejia and Wahba (1999), Becker and Ichino (2002), Smith and Todd(2005).

LaLonde (1986) is the first to examine a randomized experiment (the National Supported Work Demonstration, NSW) in order to assess the performance of several evaluation methods. He obtains **an unbiased estimate of the training effect from the randomized experiment**, and then compares the experimental result to those obtained from a range of parametric selection models applied to the NSW observations that received training and a set of comparison observations constructed from population survey data sets (CPS and PSID). Following papers have reanalyzed Lalondes results, using more recent estimation methods, including matching.

The goal of this tutorial is to analyze **the performance of regression** and **matching estimators** using observational study, replicating some of the results of these studies.

1 Data description

We here use a subset of the data used in LaLonde (1986), constructed by Dehejia and Wahba (1999). These data are available online: `http://www.nber.org/~rdehejia/nswdata2.html`

There are two types of samples:

- One sample of experimental data, which is a subset of Lalonde's data, and they are obtained from an experimental setting (see below for details on the experiment)

*Contact: thibault.richard(at)ens-paris-saclay.fr

- Two samples of observational data, which form the comparison groups. These two samples are constructed from two surveys: the Population Survey of Income Dynamics (PSID), and the Current Population Survey (CPS).

The experimental data come from a randomized experiment from the National Supported Work (NSW) Demonstration. This program was a transitional, subsidized work experience program that provided trainees with work in a sheltered training environment and then assisted them in finding regular jobs. To participate in NSW, potential participants had to satisfy a set of eligibility criteria that were intended to identify individuals with significant barriers to employment. The main criteria were: (1) the person must have been currently unemployed (defined as having worked no more than 40 hours in the 4 weeks preceding the time of selection into the program), and (2) the person must have spent no more than 3 months on one regular job of at least 20 hours per week during the preceding 6 months. Actual treatment among participants was randomized. However, as a result of these criteria as well as of self-selection into the program, persons who participated in NSW are likely to differ in many ways from the general U.S. population.

Candidates eligible for the NSW program were randomized into treatment between March 1975 and July 1977. Treated individuals in our dataset concerned individuals assigned to treatment in 1976, so that retrospective earnings information from the experiment include calendar 1974 and 1975 earnings (variables RE74 and RE75). Thus, the pretreatment variables are earnings in 1974 and 1975. Postintervention data includes calendar 1978 earnings.

Variables, for both types of data, are (from left to right) treatment indicator (1 if treated, 0 if not treated), age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), RE74 (earnings in 1974), RE75 (earnings in 1975), and RE78 (earnings in 1978), and dummy variables to identify data origin with exp (1 if experimental data, 0 otherwise) and cps (1 if CPS observational data and 0 otherwise).

1. We want to verify that the treatment and control groups have the same distribution of variables **before the program intervention**. Calculate the descriptive statistics. What can you conclude from this table?

Solution: See STATA code for replication.

Main interpretation of the table is that the distribution of preintervention variables is very similar across the treatment and the control groups. None of the differences is significantly different from 0 at a 5% level of significance, excent for the `no degree` variable.

Note that, if we take the PSID or the CPS as comparison groups, the variable differ dramatically from the treatment group, and all of the mean differences are significantly different from 0.

2 Econometric analysis and replication

2.1 Standard estimation of treatment effect

2. We are now interested in replicating the comparison of earnings and estimated training effects of the NSW program in Table 2 of Dehejia and Wahba (1999). We are first interested in the estimation of the effect of training on earnings using the experimental observations, which correspond to the first raw called "NSW" for the sets of columns B. and C. in their table.

   (a) From the experimental observations, what is the most simple unbiased estimator of the average treatment effect? Write down the linear model which allows to recover this estimator. Then, estimate it with the experimental subset of data. How do you interprete your result?

   Solution: Write the model of 1978 earnings (denoted y_i), as a function of the exposition to the job program, denoted d_i with $d_i = 1$ if the individual benefited of the program and 0 otherwise:

   $$y_i = \beta_0 + \beta_1 d_i + \varepsilon_i \tag{1}$$

   where $\beta_1 = \mathbb{E}[y_i \mid d_i = 1] - \mathbb{E}[y_i \mid d_i = 0]$. Thus, β_1 measures the effect of the training program on earning without bias. Thanks to the experimental setting, the exposition to the program does not depend on any observables as it is randomly assigned. The treatment and control groups are drawn from the same population. The experiment ensures that the causal variable of interest (the training program) is independent of potential outcomes so that the treated and control groups are perfectly comparable. Here, we can thus tell what do eligible people would have earn on average if they had the NSW program.

   See STATA code for replication.

   (b) In the Table 2 of Dehejia and Wahba (1999), we note that in addition to the simple treatment estimation (columns called "Unadjusted"), the authors also compute estimated treatment effect while controlling for age, age squared, years of schooling, high school droupout status and race (columns called "Adjusted"). What is the purpose of controlling for other covariates in our experimental setting? Reproduce these two results.

   Solution: In our experimental context, where treatment is randomly assigned, adding controls in our model should not alter our estimate of the treatment effect relative to our first estimate. Indeed, these controls are uncorrelated with the treatment variable d_i (orthogonal) and thus will not affect β_1.

   Also, note that including control variables may generate more precised estimates of the causal effect (smaller standard error of the estimated treatment effect,

since it reduces the residual variance and hence lowers the standard error of the regression estimates). This thus makes the causal inference more valid, because we observe and know some covariates (age, education, degree), and we want to be compare situations where covariates are held fixed.

(c) Finally, the authors add other variables which are pretreatment earnings (earnings in 1975 for the columns "Unrestricted differences in differences: quasi difference in earnings growth 1975–1978", and earnings in 1974 for all columns of panel C.). Replicate their results using these variables.

Solution: This approach intends to capture some fixed component (kind of a fixed effect model), by including covariates which measure pretreatment earnings. This will still not change our estimation of treatment effect.

(d) What do you conclude on your results when you add pretreatment observables relative to your first simple estimate in question a?

Solution: Our results are not altered when adding pretreatment covariates because treatment is randomized.

3. We are now interested in assessing whether the regression method applied on observational data (no random assignment of treatment) can replicate the treatment impact. We separately use the PSID and CPS samples as comparison groups of the treated individuals (and thus exclude the non-treated experimental data).

(a) Using the observational data as a comparison group, what can we expect on the simple estimator of the average treatment effect while knowing the results of the data analysis in question 1? What about when accounting for pretreatment covariates?

Solution: Treatment and comparison groups are drawn from different population, as shown by the differences in means between both samples. The treatment group is drawn from the recipients eligible for the program, while the comparison group the CPS and PSID is a different population more representative of the general US population.

Given that treatment is randomly assigned, treatment is also randomly assigned conditional on given other covariates (age, education...) denoted X_i. We can decompose the random part of potential earnings in Equation 1 into a linear function of these observable characteristics, and an error term ν_i:

$$\varepsilon_i = X_i'\gamma + \nu_i$$

where γ is a vector of regression coefficients which statisfy $\mathbb{E}[\varepsilon_i \mid X_i] = X_i'\gamma$. The residual ν_i is uncorrelated to X_i *by construction*.

But, if we use the differences in means from treated units and observational units, we will capture a selection bias (from the fact that treated population does not correspond to the same CPS and PSID population):

$$\mathbb{E}\left[y_i \mid d_i = 1\right] - \mathbb{E}\left[y_i \mid d_i = 0\right] = \beta_1 + \underbrace{\mathbb{E}\left[\varepsilon_i \mid d_i = 1\right] - \mathbb{E}\left[\varepsilon_i \mid d_i = 0\right]}_{\text{selection bias}} \quad (2)$$

where the selection bias measures the degree of correlation between the regression error term ε_i, which here contains the pretreatment observables X_i, and the exposition to the training program d_i. Again, note that in the experimental context, as d_i is randomly assigned the selection term disappears, that is, X_i is independant of d_i, and we get our simple unbiased estimation of the treatment effect.

As a result, we should condition the effect of treatment on pretreatment covariates when using observational units as a comparison. Hence by adding control variables X_i in the model, we get:

$$y_i = \beta_0 + \beta_1 d_i + X_i^{'}\gamma + \nu_i \quad (3)$$

We have now the residual in the linear model which is uncorrelated with the regressors x_i and d_i, and β_1 the causal effect of interest. This is the selection-on-observables assumption, in other words, conditional on the observables, X_i, there is no systematic pretreatment difference between the groups assigned to treatment and control. Under this assumption, estimates of β_1 in this long regression should get closer to estimates of β_1 in the simple regression. However, this assumption is likely to be violated when X_i is high dimensional: when there are many dimensions on which differences on pretreatment covariates between treated and observations are based upon. Thus, we cannot be sure that the selection on observables assumption will be verified when using the observational units as a comparison group.

(b) Perform the same regressions as in question 2 while using first the CPS data as a control group, and second the PSID data as a control group. What do you conclude on the estimation of the treatment impact while using the observational samples as comparison groups?

Solution: Overall, regression specifications and comparison group fail to replicate the treatment impact.

We however observe that controlling for all pretreatment covariates, and in particular for 1974 and 1975 earnings make our estimates with PSID and CPS observational data closer to the treatment effect. The authors say that it is important to look at several years of preintervention earnings, as this can improve the selection on observables. It actually better determine the effects of job training programs with observational data, but estimates are still not the

ones of the experimental treatment impact. As differences in variables between treatment and observational groups are important, we are not able to compare all other things being equal, and we cannot fully control for the selection into the program. We thus need an alternative approach to estimate treatment impact using observational data.

2.2 Estimating the treatment effect using the propensity score

The propensity score can be an intermediary tool to compare the average differences by matching experimental treated individuals with similar non-experimental individuals. The intuition of the propensity score is that whereas in the standard regressions we are trying to condition on X_i (intuitively, to find observations with similar covariates), we are here trying to condition just on the propensity score because observations (experimental and non-experimental) with the same propensity score have the same distribution of the full vector of covariates X_i (see class Session materials on matching and article of Dehejia and Wahba (1999) for further demonstration). Therefore, if we condition our regressions on the propensity score (that is by comparing individuals with similar propensity score), the distribution of pretreatment regressors will be the same accross the treatment and comparison group. Each individual has the same probability of assignment to treatment, as in a randomized experiment.

The estimation is done in two steps. In the first step, the propensity score is estimated **on the experimental treated units** and **on the PSID non-experimental sample**. Second, given the estimated propensity score, **we estimate a regression of the effect of job training on earnings for individuals with similar estimated propensity score**. Finding the individuals with similar estimated propensity score, that is, *the matching method*, can be done with varying techniques. The two techniques developed in Dehejia and Wahba (1999) are the stratification strategy and the nearest-neighbor matching strategy.

4. First, we estimate the propensity score using the sample of the **experimental treated observations** and **the PSID observations** (thus discard CPS observations and experimental control observations). One issue is what functional form of the preintervention variables to include in the propensity score model. We thus want to test several functional forms for the propensity score and select the specification which leads to distributions of covariates for observations with similar propensity score that are the same accross the treatment and the comparison groups.
   (a) Generate the following variables: a dummy variable denominated `U74` which is equal to 1 if the individual is unemployed in 1974 (identified by 1974 earnings equal to 0) and 0 otherwise; a dummy variable denominated `U75` which is equal to 1 if the individual is unemployed in 1975 (identified by 1975 earnings equal to 0) and 0 otherwise; a dummy variable denominated `blackU74` which is equal to 1 if the individual is black *and* unemployed in 1974 and 0 otherwise; the squared education

denominated `ed2`; and the 1974 and 1975 squared earnings denominated `re742` and `re752`.

Solution: See STATA do-file for replication code.

(b) Estimate a **logit** specification of the propensity score (the probability of being treated) as a function of `age age2 ed ed2 black hisp married re75 re74 re742 re752 blackU74`. Then produce an histogram to compare the predicted propensity score between treated and non-treated observations. What do you conclude on your propensity score specification?

Solution: See STATA code for replication.

Comments on the histograms: Many bins in the histogram where the number of treatment units outnumber the comparison units (with some cases of bins without comparison units). The histogram reveal that although the comparison groups are large relative to the treatment group there is limited overlap in terms of preintervention characteristics. If they had been no comparison units overlapping with a range of the treatment units, then it would not have been possible to estimate the average treatment effect on the treatment group. But with limited overlap, we can still proceed with estimation, but with caution. Bear in mind that we are still able to compare our estimates with the benchmark experimental estimate.

(c) There are STATA packages which estimate the propensity score, and then can perform a matching strategy. We here produce a matching strategy which stratifies the data by propensity score, which can be opposed to a nearest-neighbor matching strategy. Observations with estimated propensity score which are in the same range are then grouped into blocks (strata) and we check whether we succeed in having similar distributions of covariates within each blocks for treated and comparison observations. Use the `findit` command to look for the `st0026_2` package developped in Becker and Ichino (2002) which test the balancing property for each block and each covariate and install it. Then, use the `pscore` command using the same covariates as for the preceding `logit` command, and with the pscore(mypscore) blockid(myblock) `logit numblo(5) level(0.005)` and `detail` in the options of the command. Explain how the algorithm processes and discuss the results obtained on your STATA output.

Solution: The algorithm works as follows. The data are sorted according to $\hat{p}(x)$, which is estimated using a Logit model in a first stage. The sample observations are stratified such that within a stratum the $\hat{p}(x)$ for treated and control units are close. For example, initially a rough grid with equal ranges may be used. Within each stratum the equality of means between treated and control units should be tested for each covariate. If there is no statistically significant differences, then the regressors are balanced between the treated and

control groups and one can stop. If, for some stratum, there is no balance, then for the unbalanced stratum a finer grid is used to achieve balance. If there are many unbalanced strata, then the original logit model is reestimated with an improved specification that includes interaction and higher order terms among the regressors.

The propensity score computation has been restricted to the common support region by testing the balancing property using those observations whose propensity scores lie in the intersection of the supports of the propensity score of the treated and the control units. This restriction reduces the original sample significantly. The size of the control group drops from 2,490 units to 1,086 for the Dehejia and Wahba (2002) specification.

5. We now perform the OLS regression after restricting the NSW-treated / PSID-comparison sample using propensity score.
   (a) First estimate the propensity score using a Logit specification and adding the following control variables : `re75`, `age`, `age2`, `ed`, `black`, `hisp`, `nodeg` and `married`. Keep the observations of predicted values with propensity scores which verify $0.1 < \hat{p}(X) < 0.9$.
   (b) Finally, run the same previous regressions as those in questions 2 but on these matched observations. Compare your results, with the one obtained in the previous questions.

Solution: See dofile. We can now see that the new estimation given by the mathcing estimator is consistant with the estimator found this an experimental protocole. Thus, the matching approach gives reliable estimate of the true ATE on the targeted population.