Regression and Matching Estimators of Causal Effects

– Problem Set 2 –

In the Problem Set 2 we are going to compare the consistency of regression and matching estimators of causal effects based on Dehejia and Wahba (1999). For that we employ the experimental study from LaLonde (1986) framework which provides an opportunity to estimate true treatment effect and then use this result to evaluate treatment effect estimators one could usuallly obtain in observational study.

LaLonde (1986) implements the data from the National Supported Work program (NSW) - temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in sheltered environment. Unlike other federally sponsored employment programs, the NSW program assigned qualified applications randomly. Those assigned to the treatment group received all the benefits of the NSW program, while those assigned to the control group were left to fend for themselves. Therefore, there is a perfect framework of randomized experiment to estimate true treatment effect.

To produce the observational study, we select the sample from the Current Population Survey (CPS) as the comparison group and merge it with the treatment group sample. The helps us to obtain the sample researchers normally have to face in their work to estimate causal effects. Hence, the Problem Set 2 is based on two data sets:

1. `nswre74.csv` is field-experiment data from the NSW. It contains such variables as education, age, ethnicity, marital status, preintervention (1974, 1975) and postintervention (1978) earnings of the eligible male applicants.

2. `cps1.csv` is a non-experimental sample from the CPS which selects all males under age 55 and contains the same range of variables.

Task A

(1) Create the table with the sample means of characteristics by age, education, preintervention earnings, etc. for treated and control groups of NSW sample (you can use the Table 1 from Dehejia and Wahba (1999) as a benchmark). *Is the distribution of preintervention variables similar across the treatment and control groups?* Check the differences on significance. Add to the table the CPS sample means. *Is the comparison group differ from the treatment group in terms of age, marital status, ethnicity, and preintervention earnings?*

Task B. Regression Adjustment

In this section we compare the results of regression estimates with selection on observables as discussed in the Lecture 6.

(1) Merge the threatment group data from the NSW sample with the comparison group data from the CPS sample to imitate observational study results.

(2) *Which assumption should hold so that condition on observables could help in obtaining an unbiased estimate of the true effect?*

(3) Run the regression on both experimental and non-experimental data using a specification: RE78 on a constant, a treatment indicator, age, age2, education, marital status, no degree, black, hispanic, RE74, and RE75. We use `statsmodels` package; however, you are free in your choice. *Does the treatment estimate from the observational study consistent with the true estimate?* (Should we ask about the reasons of such inconsistency here????)

Task C. Matching on Propensity Score

Recall that propensity score $p(S_i)$ is a probability of unit i having been assigned to threatment as a function of variables that predict this assignment: $p(S_i) = Pr(D_i = 1|S_i) =$

$E(D_i|S_i)$. Assumption that makes estimation strategy credible is $S_i \parallel D_i|p(S_i)$ which means that, conditional on the propensity score, the covariates are independent of assignment to treatment. Therefore, conditioning on the propensity score, each individual has the same probability of assignment to treatment, as in a randomized experiment.

Estimation is done in two steps. First, we estimate the propensity score using a logistic probability model. Secondly, we match the observations on propensity score employing nearest-neighbor algorithm discussed in the Lecture 5. In particular, each treatment unit is matched to the comparison unit with the closest propensity score; the unmatched comparison units are discarded.

(1) Before we start with matching on propensity score, let's come back to another matching strategy which was discussed in Lecture 5 - matching on stratification. *Looking at the data could you name at least two potential reasons why matching on stratification might be impossible to use here?*

(2) Employing our imitated observational study data run a logistic probability model (we use the package `statsmodels`) with a a specification: a treatment indicator on a constant, age, education, marital status, no degree, black, hispanic, RE74, and RE75. Then extract a propensity score for every individual as a probability to be assigned into treatment.

(3) Before proceeding further we have to be sure that propensity scores of treatment units overlap with the propensity scores of control units. Draw a figure showing the distribution of propensity score across treatment and control units (we use the packages `matplotlib` and `seaborn`). *Do we observe common support?*

(4) Difficult version: write the code to match each treatment unit with control unit one-to-one with replacement (we use the package `sklearn.neighbors`) and extract the related indices of control units. Normal version: use the draft code we provide and fill in the blank spaces to run it.

(5) Construct new data set with matched observations. Run the regression to obtain matching on propensity score estimate. *Is it more or less consistent estimate of the true effect comparing to the regression estimate with selection on observables? How could you explain this result?*