Regression and Matching Estimators of Causal Effects*

– Problem Set 2 –

In this problem set we are going to compare the consistency of regression and matching estimators of causal effects based on Dehejia & Wahba (n.d.). For that we employ the experimental study from LaLonde (1986), which provides an opportunity to estimate true treatment effects. We then use these results to evaluate the performance of (treatment effect) estimators one can usuallly obtain in observational studies.

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.

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. We do this to obtain a data set which resembles the data which is commonly used in scientific practice. The two data sets are explained below.

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

---

2. `cps.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 different 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 treatment group data from the NSW sample with the comparison group data from the CPS sample to imitate an observational study.

(2) *Which assumption need to hold such that conditioning on observables can help in obtaining an unbiased estimate of the true treatment effect?*

(3) Run a regression on both experimental and non-experimental data using the specification:

$$RE78_i = const_i + treated_i + age_i + age_i^2 + edu_i + maritalstatus_i + nodegree_i$$
$$+ black_i + hispanic_i + RE74_i + RE75_i$$

We recommend using `statsmodels`, but you are free to use any other software. *Is the treatment effect estimate of the observational study consistent with the true estimate?*

2

Task C. Matching on Propensity Score

Recall that the propensity score $p(S_i)$ is the probability of unit i having been assigned to treatment. Most commonly this function is modeled to be dependent on various covariates. We write $p(S_i) := Pr(D_i = 1|S_i) = E(D_i|S_i)$. One assumption that makes estimation strategies feasible is $S_i \perp\!\!\!\perp D_i \mid p(S_i)$ 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 regression model. Secondly, we match the observations on propensity score employing nearest-neighbor algorithm discussed in the lecture 5. That is, 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 data run a **logistic** regression on the following specification.

$$treated_i \sim const. + age_i + edu_i + maritalstatus_i + nodegree_i + black_i$$
$$+ hispanic_i + RE74_i + RE75_i$$

Use for example `statsmodels` for this task. 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) Match each treatment unit with control unit one-to-one with replacement. We use the package `sklearn.neighbors`: apply the algorithm `NearestNeighbors` to the propensity score of treated and control units and extract the indices of matched control units.

(5) Construct new data set with matched treated and control 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?*

References

Bureau of Labor Statistics. (1974, 1975, 1978). *Current population survey.*

Dehejia, R., & Wahba, S. (n.d.). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs.

LaLonde, R. J. (1986). Evaluating the econometric evaluation of training programs with experimental data. *American Economic Review*, *76*(4), 604–620.