Homework on Matching

Padma Tanikella

2024-04-02

Consider the `lalonde` data set taken from the paper by Dehejia et al. for this homework on matching. The following is the first five rows of the dataset.

```
data("lalonde")
head(lalonde)
```

```
##      treat age educ   race married nodegree re74 re75       re78
## NSW1     1  37   11  black       1        1    0    0  9930.0460
## NSW2     1  22    9 hispan       0        1    0    0  3595.8940
## NSW3     1  30   12  black       0        0    0    0 24909.4500
## NSW4     1  27   11  black       0        1    0    0  7506.1460
## NSW5     1  33    8  black       0        1    0    0   289.7899
## NSW6     1  22    9  black       0        1    0    0  4056.4940
```

About the data set

The `lalonde` data was collected as part of a study that looked at the effectiveness of a job training program (the treatment) on the real earnings of an individual, a couple years after completion of the program. The data consists of a number of demographic variables (age, race, academic background, and previous real earnings), as well as a treatment indicator, and the real earnings in the year 1978 (the response).

- 614 observations on 9 covariates. 185 of the units were exposed to the treatment while 429 units belong to the control group.
- `treat` denotes the binary treatment assignment. `treat` = 1 denotes a treated unit.
- `age` in years
- `educ` is education in number of years of schooling.
- `race` is unit's race/ethnicity
- `married` is binary indicator for marital status. `married = 1` denotes a married person.
- `nodegree` is a binary indicator for whether the subject has a high school degree. `nondegree`=1 denotes a person with no high school degree.
- `re74` is income in 1974, in U.S.D.
- `re75` is income in 1975, in U.S.D.
- `re78` is income in 1978, in U.S.D.

Note that this is an **observational study** i.e. there was NO random assignment of the treatment or training program.

Questions

Based on the information above, answer the following questions:

1. Assume that you conduct a different study and collect the data `lalonde2`. In this new study, the treatment is assigned at random. How can you estimate the causal effect of training program on the real earnings of individuals in 1978?

2. While working with the `lalonde` data set, can we use the technique of the first question to estimate the causal effect of the training program on the real earnings of individuals in 1978? Why or why not? Explain in a couple of sentences.

3. Let us now perform **exact matching** based on the covariates `race`, `married` and `nodegree`. The table below summarizes the number of subjects included in each of the ten groups created.

```
##    Group assignment Number of subjects
## 1                 1                 37
## 2                 2                 31
## 3                 3                 57
## 4                 4                132
## 5                 5                 74
## 6                 6                 62
## 7                 7                 68
## 8                 8                 17
## 9                 9                 24
## 10               10                  9
## 11               11                 95
## 12               12                  8
```

Based on the above table, can we claim that each subject was assigned exactly one group? In other words, is there a subject assigned to two or more groups?

4. Propose a different way of matching that one can use while working with the `lalonde` data set.