

# PS 236: Causal Inference

## Problem Set 3

UC Berkeley, Fall 2008

### Solutions

Your solutions must be submitted in hard copy to my mailbox in the Political Science main office by 4pm on the due date. No late assignments will be accepted. Clean R code should be submitted separate from the solutions requested below.

## 1 The returns to education

You would like to determine how much an individual's log wage increases for each additional year of schooling that he receives; this is known as the private returns to education. Of course, you are worried about endogeneity in this problem and you are considering matching to remove this potential source of bias. Suppose that you have the following covariates in addition to years of education: age, sex, race, state of birth, and state of residence. Is matching a good strategy in this scenario? Why or why not?

**Solution:** The most restrictive assumption that the matching strategy imposes upon the design of the data is that of selection on observables; an individual's outcome must be independent of his treatment status conditional upon all observed covariates. In other words, anything correlated with outcomes must be independent of treatment and vice-versa. In the data described in this problem, we might imagine that unobserved factors, such as intelligence and motivation, might influence both years of education and income. This is known as endogeneity and dooms the matching strategy. Because educational choice is an endogenous choice not easily explained via observable characteristics, our matching estimates would be biased. Since these unobserved factors are likely quite important, the bias may be significant.

## 2 Correct coverage simulations

Suppose that the true model of an outcome  $Y$  as a function of treatment status  $T$  and covariates  $X$  is:

$$Y_i = 5T_i + 2X_{1i} - 2X_{2i} + 4X_{3i} + X_{4i} - 5X_{5i} + \epsilon_i.$$

Further assume that  $T_i \perp Y_i, X_i$ ,  $\mathbb{E}(\epsilon_i|X_i) = 0$ , and  $\epsilon_i \sim \mathcal{N}(0, 1)$ .

- Write a function that performs Monte Carlo simulations based upon these assumptions to test the 95% confidence interval generated by OLS for correct coverage. Use 250 observations and 1,000 simulations. Use the following code to generate the means of the  $X$  covariates and their associated variance-covariance matrix:<sup>1</sup>

---

<sup>1</sup>The variance-covariance matrix must be positive semi-definite and symmetric; multiplying a matrix by its transpose will conform to these requirements.

```
## Set the seed
set.seed(1027)

## Create the X means
means <- c(round(runif(5,-20,20)))

## Create the sigma matrix
a <- matrix(runif(25,-1,1), ncol=5)
sigma <- a %*% t(a)
```

You should generate a new set of data for each run of the simulation. (Hint: You may want to alter the code used in section for this and the following parts).

**Solution:** The true parameter was contained in 95.3% of the generated confidence intervals.

- b. Instead of running OLS, use the `Match()` function to produce confidence intervals as above and test their coverage.

**Solution:** The true parameter fell in 95.4% of the generated confidence intervals.

- c. Combine matching and OLS naively. First, perform matching on your generated data. Then, run OLS on the matched data only. Lastly, use the estimate and standard error from the OLS output to generate confidence intervals. Check these for correct coverage.

**Solution:** The true parameter was in 86.8% of the generated confidence intervals. This degree of coverage does not match the desired level of the test.

- d. Use `BiasAdjust` and the `Z` matrix within the `Match()` function to perform matching on the data and subsequently perform OLS on these data. Using the estimate and standard error given by the matching output, test for correct coverage.

**Solution:** The true parameter fell in 94.3% of the confidence intervals.

### 3 Exact matching

For this question, use the `ExactMatching.csv` data set. The model of the data generating process in this question is unknown:

$$Y_i = F(X_{1i}, X_{2i}, T_i),$$

but assume that selection on observables holds:  $Y_i \perp T_i | X_i$ .

- a. Perform OLS. What is your estimate of the treatment effect? Is it significant?

**Solution:** The OLS estimate of the treatment effect is 1251.83, which is significant at the 1% level.

- b. Write your own code (*i.e.*, do *not* use `Match()`) to perform *exact* matching on both covariates. What is your estimate of the treatment effect? How many observations must you drop?

**Solution:** The treatment effect is estimated to be -1852.7, which is marginally significant. Four treated observations were dropped.

- c. Use `Match()` to perform non-exact matching. What is the estimate of the treatment effect? Is it significant?

**Solution:** The treatment effect is estimated to be -2114.6, which is significant at the 5% level. Note that the estimates of parts (b) and (c) have the opposite sign of part (a); the OLS coefficient is severely biased (the matching estimates are accurate).