PS 236: Causal Inference
Problem Set 4

UC Berkeley, Fall 2008

Due: Monday, November 17

Your solutions must be submitted in hard copy to my mailbox in the Political Science main office (Barrows Hall 210) 4pm on the due date. Except under extraordinary circumstances, no late assignments will be accepted.

1 Covariates and random treatment assignment

You conduct an experiment in which treatment is assigned randomly. You perform a regression of the outcome on treatment and several covariates, such as age, sex, and race, and it yields a significant estimate of the treatment effect. A reviewer of your paper says that you should not include any of the additional covariates in your model. A regression of outcome on treatment alone yields an insignificant result, though with the same sign found as in your original model.

a. Why does the reviewer suggest his procedure?

b. What might explain the pattern of results posited? Intuitively, why does this hold?

c. Which model should we prefer: yours or that of the reviewer? Why?

2 Matching using the LaLonde data

Before answering this question, please read the LaLonde-related stream of papers given on the syllabus. Then, download the LaLonde data provided here:

`http://sekhon.berkeley.edu/causalinf/R/lalonde2.RData`

This dataset includes all the treated observations, plus a sample of the observations from the non-experimental controls that have two years of prior earnings data—it is *not* the full dataset used by the authors of the papers.

We begin by examining several matching attempts using individual earnings in 1974 alone. For all parts, when estimating the treatment effect, estimate the average treatment effect for the treated (ATT).

a. Given that we match on a single covariate, why might we choose this one?

Create a matched dataset using only individual earnings in 1974 (`RE74`).

b. What is the estimated treatment effect? Is it significant?

c. Report the balance statistics for the covariates listed in Table 1 of Dehejia and Wahba (1999) for this matched dataset.

Now, create a p-score model with individual earnings in 1974 as the sole covariate.

d. Match on the linear predictor given by your p-score model. Do you get the same matched pairs as in part (b)?

e. Match on the p-score itself. Do you get the same matched pairs as in the two other matched datasets?

Lastly, perform genetic matching using a p-score model and covariate set that you deem appropriate. Here is a GenMatch example using this dataset:

`http://sekhon.berkeley.edu/causalinf/R/GenMatch1.R`
`http://sekhon.berkeley.edu/causalinf/R/GenMatch1.Rout`

f. Report the balance statistics for the same covariates as in part (c) using this procedure. How do the balance figures compare to those in part (c)?

g. Create QQ-plots for individual income in 1974 and 1975. Do the distributions for treated and control appear similar?

h. What is the estimated average treatment effect for the treated? Is it significant?

i. What is the estimate of the top 20th percentile of the treatment effect? How do you interpret this estimate?

j. How close is your estimate to the experimental benchmark?