

## Economics 696F, Causal Inference and Program Evaluation

### Problem Set 1: due Friday, Feb 9 (corrected version)

1. Consider the data in Table 1. Assume they arose from a randomized experiment in which an equal number of units were selected to be given the treatment or a placebo.

Consider the statistic  $\hat{\tau}_n$ , the difference of sample means. Find the p-value for the sharp null hypothesis of no treatment effect, using simulation. (Write your own code to do the simulation.) Also find an approximate p-value using asymptotic approximations.

Table 1

| $T = 0$ | $T = 1$ |
|---------|---------|
| -1.1    | -2.3    |
| -2.1    | 1.1     |
| -0.7    | 0.1     |
| -0.7    | 1.0     |
| 0.4     | -0.6    |
| 0.8     | -1.7    |
| -0.1    | 2.3     |
| 1.8     | 2.4     |
| 0.1     | 2.6     |
| 1.1     | 4.6     |

2. The data in `ps1a.txt` are taken from the National Supported Work Demonstration (NSW), a randomized evaluation of a job training program in the U.S. See Lalonde (1986) for more information about this data. The variable `treat` denotes the treatment indicator. The variable `re78` is real earnings in 1978, and is the outcome we are interested in.
  - (a) Estimate the average treatment effect of the program on real earnings in 1978. (Since this program took individuals who wished to enter the program, and randomized some of them out of the program, we can interpret this as the treatment effect on the treated.) Provide the standard error that allows for heteroskedasticity, and test the hypothesis that  $TT = 0$ .
  - (b) There are other background variables in the data file, including `age`. What would you expect to see if you treated `age` as the outcome and estimated the effect of training on `age`? Estimate this effect and its standard error, and test the hypothesis of no treatment effect.

3. This question uses the data set in `ps1b.txt`. This contains the treatment observations from the NSW data set, but instead of the experimental controls, it substitutes control observations from a different data set (the Panel Study of Income Dynamics). Lalonde (1986) constructed this data set to look at how closely one could replicate the experimental results with nonexperimental data. By combining the NSW treated observations with controls drawn from the PSID (who were not able to get the treatment or chose not to get the treatment), the data set mimics a nonexperimental study where the treatment was not randomly assigned.

(a) Estimate the average treatment effect (on the treated) by using the difference between the treated mean and the control mean. Use `re78` as the outcome variable, and `treat` as the treatment. How does this estimate compare to the experimental results? Interpret your results.

(b) Next, assume unconfoundedness, and suppose we assume that

$$E[Y|T, X] = \beta_1 + T_i\beta_2 + X_i'\beta_3 + (T_i \cdot X_i)'\beta_4,$$

where  $X_i$  is a vector containing the variables `age`, `educ`, `re75` (earnings before the program), `married`, `black` (indicator for Black) and `hisp` (indicator for Hispanic). Provide an estimate of the average treatment effect on the treated, and compare this to the experimental results. (Be sure to get an estimate of  $TT$  here. You may use “canned” regression packages if you prefer.)