

Political Science 239

Problem Set 4

Due date: Wednesday, October 4th, 2006

In this problem set, we will use the database "smoking_short.csv", which contains data from the 1989 Linked National Natality-Mortality Detail Files. These are annual census of births in the US. The original database contains all births in Pennsylvania in 1989. The observational unit is the mother-infant pair. Since the original dataset contains more than 100,000 observations, in order to increase the speed of the matching procedures, the dataset you will be using is a random sample of size 3000 of the original dataset. You can download the dataset from http://are.berkeley.edu/~rocio/smoking_short.csv. The description of the variables in the dataset is in the file "VarsDescription_PS4.xls". The list of covariates include educational attainment, race, and medical history of the mother, and educational attainment and race of the father. The treatment variable is an indicator equal to one if the mother smoked during the pregnancy, and the outcome variable is the birth weight of the infant at birth (measured in grams). We are interested in estimating the effect of smoking during the pregnancy on the birthweight of the infant.

Exercise 1 *Compute difference in means and the variance ratios on the covariates between treatment and control mothers. Are mothers who smoke and mothers who don't smoke similar in terms of observable characteristics? Why is this similarity relevant for the estimation of the effect of smoking on birthweight?*

Exercise 2 *Using the Match() function, match treatments and controls on educational attainment of the mother. Now analyze the balance of your matched sample on all the covariates computing*

the difference of means and the variance ratios for smoker and non-smoker mothers. Make sure that you use the weights when you compute means and variance ratios (you can find a function that calculates the weighted variance on the website)¹. Was matching on education alone enough to balance the other observable characteristics? How do these means and variance ratios compare to the means and variance ratios before matching?

Exercise 3 Now we will explore matching on the propensity score. Estimate the propensity score with a logit in two different ways: (i) using only mother education as a covariate in the linear predictor and (ii) using all covariates that you consider appropriate in the linear predictor. For the second propensity score, be sure to justify why you include and/or exclude covariates. Show boxplots of the propensity score by treatment and control for both predicted propensity scores. Do they differ? Is one of them preferable to the other? Why?

Exercise 4 Match on the second predicted propensity score using the `Match()` function. Now analyze the balance of the matched samples on all the covariates using means and variance ratios. Does balancing on the propensity score rather than on education alone improve the balance on the observable characteristics? You can present some QQ-plots to help your argument.

Exercise 5 Choose the matching method that gives you the best balance and calculate the average treatment effect on the treated. (You can use the `Match()` function to calculate it). Compare this ATT with the ATT estimated with a simple OLS regression. (In the OLS regression, you should include as covariates all the variables in which you've matched on). Do these two different estimation methods give you similar results? Do you prefer one over the other? Why?

¹You can compare your results with the output of the `MatchBalance()` function to make sure that they are the same. If you use `MatchBalance` to check your results, make sure to use the option `ks=FALSE`. Otherwise your computer will take too long to analyze the balance. The option `ks=FALSE` omits the calculation of the Kolmogorov-Smirnov Test.