PS C236A / Stat C239A
Problem Set 3
Due: Oct. 15, 2012

Instructions

This assignment is due **4 pm Monday, Oct. 15.** You may submit your analytical work either electronically or in paper form. Electronic versions must be sent as a .pdf to <`jahenderson[at]berkeley.edu`>. Paper copies should be placed in my mailbox in 210 Barrows. For the computing portion of the assignment, you must submit a fully executable version of all .R code, along with any data used in the code (excepting that provided through the course webpage) to the email above. All files for each assignment sent electronically should be included in one omnibus email, with the subject line containing the course and homework number, and your last name (e.g., PS239A/STAT236A: HW3 - Clinton).

You are encouraged to work together in groups to complete the assignments. However, you must hand in your own individual answers. Photocopies and other reproductions of someone else's answers are not acceptable. Please also list the names of everyone with whom you have collaborated on this assignment.

Problem 1 Consider an observational study, where $Z_i = 1$ if unit i is in the treatment group and $Z_i = 0$ if unit i is in the control group. Let X be a vector of observed pretreatment covariates. Write $X_{Z=1}$ for the observed covariates of the units in the treatment group. Similarly, let $X_{Z=0}$ be the observed covariates in the control group. Let r_1 be outcome under treatment and r_0 be the outcome under control. Assume the following:

$$r_0 \perp\!\!\!\perp Z | X_{Z=1}$$

$$P(Z = 1 | X_{Z=1}) < 1$$

Suppose you know the propensity score $e(X) = P(Z = 1)$ for all units i. With these assumptions, can conditioning on the propensity score estimate the ATT without bias? Prove it mathematically and describe your logic in words. What additional assumption would we need in order to estimate the ATE without bias? First show that conditioning on the propensity score is equivalent to conditioning on $X_{Z=1}$. Then show that conditioning on the propensity score can produce unbiased ATT estimates under the assumptions above.

Problem 2 Suppose there are 10,000 people in a study. Some of these people are assigned to treatment, and the rest are assigned to control. The exact mechanism for assigning units to treatment is unknown, but is known to depend only on the values of three dichotomous variables: the sex of the person, whether or not the person exercises 30 minutes a day, and whether or not the person watches TV for more than an hour a day. For each of the $2^3 = 8$ configurations of the dichotomous variables, there is at least one treated person and at least one non-treated person.

a. Let $T_i = 1$ if person i is treated, and let $T_i = 0$ if the person is not. Let $T = (T_1, T_2, \ldots, T_{10000})$ denote the observed treatment assignment for all 10000 units. The probability that person i is assigned treatment is p_i, which is unknown. Express the probability of observing the treatment assignment T (i.e., $Pr\{T = t|m\}$) in terms of T and p. What estimated values of p_i maximize this probability, under the assumption that treatment assignment only depends on the dichotomous variables? [Hint: The probability of the given treatment assignment is found in Rosenbaum and Rubin (1983). Take the log of this probability and use calculus to maximize these probabilities.]

b. Suppose that the propensity score depends on all three of these dichotomous variables. Will the estimated probabilities found in (a) converge to the true propensity score? What if the propensity score only depends on two of the original three dichotomous variables?

c. Suppose the propensity score is known (and again, depends only on the values of these three variables). Write out unbiased estimates for the ATE and the ATT, and prove that they are unbiased (assuming that the propensity score is fixed). How do these estimates change if the propensity score is not fixed?

d. Suppose the heaviest person of the 10,000 people weighs 500 pounds. Moreover suppose the propensity score is:

$$P(T_i = 1|X) = 1/2 - \text{weight}/1000$$

Can both the ATT and the ATE be estimated without bias by conditioning responses on the propensity score? For each quantity that can be estimated unbiasedly, give a description or formula on how to compute the estimate (supposing the propensity score is known and fixed).

e) Suppose it is known that the propensity score is linear in weight, but the exact coefficients are unknown. Will regressing treatment assignment on exercise and weight using OLS produce unbiased estimates of these coefficients?

Problem 3 Suppose that an award is given to anyone who scores above some threshold c on a test. A student wants to evaluate the effect of obtaining the award on future income. The student uses an RD model to evaluate this effect: The student fits the model

$$\text{Future Income}_i = \alpha + \beta(\text{Test Score}_i) + \epsilon_i$$

separately to people scoring at most five points below the threshold and to those scoring at most five points above the threshold. The student estimates the effect of winning the award on future income by taking the difference between the regression estimates when Test Score $= c$.

a. Show that the variance of the regression estimates are larger when further away from the mean test score. How does this effect the estimate of the LATE?

b. Suppose that, before a person takes a test, they flip a coin. If the coin lands tails, the person loses 10 points on his test; otherwise, the person does not lose any points on the exam. The result of the coin flip is not known to the student. (Assume that the distribution of scores on the interval $\{c - 5, c\}$ is the same as the distribution of scores on the interval $\{c + 5, c + 10\}$.) Can the LATE still be estimated? If so, how? If not, what condition necessary to estimate the LATE is violated?

c. Suppose that, in reality, future income is related to the test score and winning the award in the following way:

$$\text{Future Income}_i = \alpha + \beta(\text{Test Score Before Coin Flip}_i) + \gamma(\text{Win award}_i) + \epsilon_i$$

where both β and γ are greater than zero. Under the scenario in part (b), where a coin flip can cause the test taker to lose 10 points, show that the estimate of the LATE in part (a) gives a lower bound (asymptotically) of the true LATE.

Problem 4 This problem is based on Sekhon's analysis of the voting irregularities in the 2004 election in Florida. There was a lot of speculation that "the optical voting machines that [were] used in a majority of Florida counties caused John Kerry to receive fewer votes than 'Direct Recording Electronic' (DRE) voting machines". The paper can be downloaded at: `http://sekhon.berkeley.edu/papers/SekhonOpticalMatch.pdf`. And the data is available here: `http://sekhon.berkeley.edu/causalinf/data/hw3data.RData`.

a. Calculate a propensity score for assignment to treatment (defined as having an electronic voting machine), using no more than 3 of the following 5 covariates: `income`, `votePer00.dem`, `regPer00.dem`, `black00`, `lowEduc00`. Provide some justification for your pscore model. Make boxplots showing the distribution of the propensity score for both treated and control groups.

b. Write your own univariate nearest-neighbor matching function (with replacement). In this function, include an option to pass in a caliper. Run this function twice, once without an enforced caliper and once with an enforced caliper. How small must your caliper be before it changes your results in a significant way?

c. Check balance, after matching with a caliper, and after matching without a caliper. What can you conclude? What is the difference between running your function with a caliper and without a caliper? Which method should we prefer? Why?

d. Using your matched data set, calculate the ATE and ATT of having an electronic voting machine on Bush's presidential vote in 2004. How do these two estimates differ? Should we prefer one or the other? Why?

e. Estimate the ATE and ATT using inverse probability weighting on the basis of the same propensity score you estimated in part (a). How does this estimate compare to the one you recovered in (d)?

Bonus Again, using the propensity score estimated in (a), estimate your estimands (ATE and ATT) using weighted OLS regression.