PS C236A/ Stat C239A
Problem Set 4
Due: October 2, 2009

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?

Hint: 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 Write a paragraph or two on your plans for your final paper. If you have several ideas and would like feedback on the feasibility/suitability of each of your ideas, that's fine too.

Problem 3: 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`

a. First, using Bush's vote percentage in 2004 as an outcome, run three linear models of the effect of using an electronic voting machine. In each model, include different explanatory variables. How does changing the model change the estimate of the effect of having an electronic voting machine as well as the significance of this estimate. Does the point estimate move? Should we be concerned, and if so, why?

b. Now, calculate a propensity score for assignment to treatment (defined as having an electronic voting machine). Provide some justification for your pscore model. Make boxplots showing the distribution of the propensity score for both treated and control groups.

c. Write your own univariate nearest-neighbor matching function. In this function, include 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? [Hint: You can effectively not enforce a caliper if you pass in a very large number for your caliper]

d. Plot the density of some key covariates before matching, after matching with a caliper, and after matching without a caliper. What can you conclude?

e. What is the difference between running your function with a caliper and without a caliper? Which method should we prefer? Why?

f. Using your matched data set, calculate a treatment effect (ATE, ATT, or ATC) of having an electronic voting machine. Which treatment effect did you choose to estimate? Why? How does it compare to your first three models?