Problem Set 1a: Matching, Reweighting, and the Effects of Maternal Smoking on Infant Health

Parts with an asterisk (*) are "optional". As a practical matter, doing these parts is unnecessary for passing the course, but would likely be necessary for getting an A.

1. * Before getting started with the data work, first consider the table from Snow (1855) reproduced in the lecture notes ("Snow's Table IX"). The table reports only means.

   (a) Develop an approximate 95% confidence interval for "Deaths per 10,000 Houses" for Southwark and Vaxhall customers. Develop another 95% CI for the same quantity for Lambeth. Do the confidence intervals overlap?

   (b) Discuss either formally or intuitively the critical assumption that underlies your confidence intervals. Give a 2 or 3 sentence quote from Snow's description (reproduced in Freedman (1991)) that supports this assumption.

We now move to some analysis of real data. The data portions of Problem Sets 1a and 1b are based heavily on the paper Almond, Chay, and Lee (2005), and problem sets from Ken Chay and John DiNardo based on some of the data used in the paper. The goal of this assignment is to examine the research question: what is the causal effect of maternal smoking during pregnancy on infant birthweight and other infant health outcomes. The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania. Each observation represents an infant-mother match. The data in Stata format can be downloaded from the bCourses website. There should be 48 variables in the data and, after you are finished with the cleaning steps described below, 114,610 observations.

The data here are "real" and quite imperfect, which will help simulate the unpleasantness of real world data work. Unlike the real world where you will confront this bleak situation largely alone, I will provide you with some hints for working your way through the raw data. You can download part of the codebook for the data to help you figure out the relevant variables.

2. The first order of business is to go through the code book, decide on the relevant variables, and process the data. This involves several steps:

   (a) Fix missing values. In the the data set several variables take on a value of, say, 9999 if missing. We have already checked for missing observations for about 2/3 of the variables. The remaining variables need to be checked and are the last 15 in the variable list (i.e. from 'cardiac' to 'wgain'). Refer to the codebook for missing value codes. Produce an analysis data set that drops any observation with missing values.

   (b) If this were a real research project you would want to consider other approaches to missing data besides termination with extreme prejudice. What observations do you have to drop because of missing data. Might this affect your results? Do the data appear to be missing completely at random? How might you assess whether the data appear to be missing at random?

   (c) Produce a summary table describing the final analysis data set.

3. The next part of the assignment is to try to estimate the "causal" effect of maternal smoking during pregnancy on infant birth weight. Let's start out using techniques that are familiar, and think about whether they are likely to work in this context. Answer the following questions.

   (a) Compute the mean difference in APGAR scores (both five and one minute versions) as well as birthweight by smoking status.

   (b) Under what circumstances can one identify the average treatment effect of maternal smoking by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers? Estimate its impact under this assumption. Provide and comment on some evidence for or against the validity of the assumption (A useful "Table 1" of any paper is one that describes the overall averages of the observations, and then describes the subsets of people who do and do not receive the treatment (when it is binary)).

   (c) Suppose that maternal smoking is randomly assigned conditional on the other

observable "predetermined" determinants of infant birth weight. First discuss which (if any) of the variables contained in the data set can clearly be considered to be predetermined. In general, what kinds of variables can be considered predetermined and what kinds of variables cannot?

(d) What does "selection on observables" imply about the relationship between maternal smoking and unobservable determinants of birth weight conditional on the observables? Use a basic linear regression model, in conjunction with your answer to part (c), to estimate the impact of smoking and report your estimates. Under what circumstances is the average treatment effect identified?