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

This problem set builds on Problem Set 1a, where you began thinking about how to estimate the causal effect of maternal smoking during pregnancy on infant health outcomes. Both problem sets 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 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 bspace website. There should be 48 variables in the data and, after you are finished with the cleaning steps described below, 114,610 observations.

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. In Problem Set 1a, you used linear regression to relate infant health outcomes and maternal smoking during pregnancy. Please answer the following questions.

   (a) Under the assumption of random assignment conditional on the observables, what are the sources of misspecification bias in the estimates generated by the linear model estimated in Problem Set 1a?

   (b) Now, consider a series estimator. Estimate the smoking effects using a flexible functional form for the control variables (e.g., higher order terms and interactions). What are the benefits and drawbacks to this approach?

   (c) Use the LASSO to determine which covariates (and higher order terms) to include in your regression from part (b). Do you end up dropping some covariates that you had thought might be necessary to include?

2. Describe the propensity score approach to the problem of estimating the average causal effect of smoking when the treatment is randomly assigned conditional on the observables. How does it reduce the dimensionality problem of multivariate matching? Try a few ways to estimate the effects of maternal smoking on birthweight:

(a) First create the propensity score. For our purposes let's use a logit specification. First specify the logit using all of the "predetermined" covariates (don't include interactions). Next, include only those "predetermined" covariates that enter significantly in the first logit specification. How comparable are the propensity scores? If they are similar does this imply that we have the "correct" set of covariates in the logit specification used for our propensity score?

(b) Control directly for the estimated propensity scores using a regression analysis, and estimate an average treatment effect. State clearly the assumptions under which your estimate is correct.

(c) As discussed in class, one can use the estimated propensity scores to reweight the outcomes of non- smokers and estimate the average treatment effect. Compute an estimate of the average treatment effect and the "effect of the treatment on the treated" by appropriate reweighting of the data.

(d) Estimate the counterfactual densities relevant for the above part with a kernel density estimator. That is, estimate the density of birthweight (or log birthweight) if everyone smoked and again if no one smoked. Hint: Consider directly applying the Hirano, Imbens, and Ridder propensity score reweighting scheme in the context of estimating the densities of the treated and control groups (rather than the means of the treated and control groups). Stata has very useful preprogrammed commands. In addition to using the preprogrammed Stata command to compute/graph the kernel density over the entire range of birthweight, please also calculate by hand the kernel estimator at birthweight equals 3,000 grams (and provide the code you wrote that shows the calculation of the kernel estimator at this single point). Play around with a bandwidth starting with half the default Stata bandwidth. Choose the same bandwidth for all the pictures, and produce a (beautiful, production quality) figure depicting both densities.

(e) Take one of your densities and display an estimate of the density using different bandwidths as well as the one you settled on. What happens with bigger (smaller) bandwidths?

(f) What are the benefits of the weighting approach (from part c)? What are the potential drawbacks? Pay particular attention to to the issue of people with extremely high and extremely low values of the propensity score.

(g) Present your findings and interpret the results on the relationship between birthweight and smoking. For the estimates in parts (b) and (c), consider which of the following conditions must hold in order for that estimate to be valid:

i. The treatment effect heterogeneity is linear in the propensity score.

ii. The treatment effect heterogeneity is not linear in the propensity score.

iii. The decision to smoke is completely randomly assigned.

iv. Conditional on the exogenous variables the decision to smoke is randomly assigned.

3. A potentially more informative way to describe how birth weight affects smoking is to estimate the "non-parametric" conditional mean of birth weight as a function of the estimated probability of smoking, separately for smokers and non-smokers on the same graph. To do so, divide the data from smokers into 100 approximately equally spaced bins based on the estimated propensity score. Do the same for nonsmokers. Use the blocking estimator we discussed in class. Interpret your findings and relate them to the results in (2b).

4. Low birth weight births (less than 2500 grams) are considered particularly undesirable since they comprise a large share of infant deaths. Redo question 3 using an indicator for low birth weight birth as the outcome of interest. Interpret your findings.

5. Let's link matching back to regression. Consider the conditional expectation function $E[birthweight \mid X]$, where X contains the following variables: `rectype pldel3 cntocpop stresfip dmage mrace3 dmar adequacy csex dplural`.

(a) Develop a regression that you are confident estimates $E[birthweight \mid X]$ as $N \to \infty$? Why are you confident that your regression gets the CEF right?

(b) * Now run the regression you propose above, but add the treatment (your binary smoking variable) as the righthand side variable of interest. Prove that if the

treatment effect of smoking on birthweight is independent of the covariates in X, then exact matching and your regression estimate the same thing. You may assume the conditional independence assumption holds given the variables in X listed above.

(c) * Develop a weighted version of the exact matching estimator that estimates the same thing as the regression above (regardless of whether the treatment effect is independent of covariates).

(d) * Estimate the weighted matching estimator you propose. Compare it to the regression estimate from part (b). Are they similar?

(e) * Is the sample size of your regression the same as the sample size of your matching estimator, or does the regression have more observations? If the regression has more observations, why don't these extra observations influence the treatment effect estimate?

(f) * Compute a standard error for your matching estimator using the formula from Imbens (2015). Specifically, note that your matching estimator should have a form $\frac{1}{N_t}\sum_{d_i=1} w_i y_i - \frac{1}{N_c}\sum_{d_i=0} w_i y_i$, where $\sum_{d_i=1} w_i = N_t$ and $\sum_{d_i=0} w_i = N_c$. Then the conditional variance is approximately $\sum_i (\frac{d_i}{N_t^2} + \frac{1-d_i}{N_c^2}) w_i^2 \hat{\sigma}^2_{d_i}(x_i)$, where $\hat{\sigma}^2_{d_i}(x_i) = \frac{1}{2}(y_i - y_{nn(i)})^2$, and $y_{nn(i)}$ is the nearest neighbor to observation i with the *same* treatment status. Figure out the implicit weights w_i in your estimator from part (d), and compute the conditional variance. Is it close to your regression coefficient variance?

6. Concisely and coherently summarize your overall results, providing some intuition. Write it like you would the conclusion of a paper. In this summary, describe whether you think your best estimate of the effects of smoking is credibly identified. State why or why not.