University of California, Santa Barbara **Olivier Deschenes**
Department of Economics **Winter 2018**

ESM 296

Individual Assignment 3
Due in class 03/05/18

Question 1: Application of estimators based on selection on observables.

This exercise asks you to implement some of the techniques we discussed in class when we covered selection on observables. You will try to identify the causal effect of maternal smoking during pregnancy on infant birth weight. The data are taken from the National Natality Detail Files, and the extract "smoking_esm296.dta"' is a random sample of all births in Pennsylvania during 1989-1991. Also available in .csv format smoking_esm296.csv". Each observation is a mother-infant pair. The key variables are:

The outcome and treatment variables are:
birthwgt=birth weight of infant in grams
tobacco=indicator for maternal smoking

The control variables are:
mage (mother's age), meduc (mother's education), mblack (=1 if mother black), alcohol (=1 if consumed alcohol during pregnancy), first (=1 if first child), diabete (=1 if mother diabetic), anemia (=1 if mother anemic)

The objective is to try to estimate the causal effect of maternal smoking during pregnancy on infant birth weight. The question is open-ended but try to be concise in your answers. Please email or stop by office hours if you have any questions.

(a) What is the unadjusted mean difference in birth weight of infants with smoking and non-smoking mothers? Under what hypothesis does this correspond to the average treatment effect of maternal smoking during pregnancy on infant birth weight? Provide some simple empirical evidence for or against this hypothesis.

(b) Assume that maternal smoking is randomly assigned conditional on the observable covariates listed above. Estimate the effect of maternal smoking on birth weight using a linear regression. Report the estimated coefficient on tobacco and its standard error.

(c) Use the multivariate (or exact) matching estimator to estimate the effect of maternal smoking on birth weight. For simplicity, consider the following covariates in your matching estimator: create a 0-1 indicator for mother's age (=1 if mage>=34), and a 0-1 indicator for mother's education (1 if meduc>=16), mother's race (mblack), and alcohol consumption indicator (alcohol). These 4 covariates will create 2*2*2*2 = 16 cells. Report the estimated causal effect of tobacco on birthweight using the exact matching estimator (Lecture 7, slides 19-21) and its linear regression equivalent (Lecture 7, slides 22-23).

(d) Estimate the propensity score for maternal smoking using a logit estimator and based on the following specification: mother's age, mother's age squared, mother's education, and indicators for mother's race, and alcohol consumption.

(e) Create the following 3 "blocks" of the propensity score distribution: block 1: pscore<=0.14 (roughly the 33rd percentile), block 2: pscore>0.14 & pscore<=0.23, and block 3: pscore>0.23 (roughly the 67th percentile). There should be roughly the same number of observations in each block.

(f) Test for the "balance" of the following covariates within each block: mage, meduc, mblack, and alcohol.

(g) Estimate the average difference in birthweight by tobacco status within each of the blocks and re-weight the block-specific estimates to obtain the implied estimated ATE (as in Lecture 8, slides 11-12).

Note: This homework is a simple examination of these data. More work would be needed to carefully assess the causal effect of smoking on children's outcomes. Further, for this homework, you can ignore the adjustments to the standard errors that are necessary to reflect the fact that the propensity score is estimated. Just use "robust" standard errors in STATA or R. If you are interested, you can read Imbens and Wooldridge (2009) and Imbens (2014) for discussions of various approaches and issues with standard error estimations in models based on the propensity score.