Prof. Leo Feler

Applied Econometrics

Johns Hopkins University SAIS

**Problem Set 2**

This exercise examines the following research question: What is the effect of maternal smoking during pregnancy on infant birth weight and death? Feel free to work cooperatively and in groups. Each person must hand in his/her own problem set using his/her own words and interpretation of the results. Please include a concise summary of your empirical results when appropriate.

We will analyze the following STATA data set:

Data Source: smoking2.dta

This STATA data extract is from the 1989 Linked National Natality-Mortality Detail Files, which are an

annual census of births in the U.S., derived from Certificates of Live Birth. Information on subsequent infant death within a year of birth is derived from Death Certificates. This extract consists of all births in Pennsylvania in 1989. The observational unit of the data is the mother-infant outcome match.

*Data Notes*:

1. There are 139,149 observations and 32 variables. For this problem set, observations with missing values for any of the variables below were dropped from the original sample (about 17%).

2. The key variables are:

dbirwt = birth weight of the infant (in grams)

death = indicator equal to one if the infant died within one-year of birth and zero, otherwise

tobacco = indicator equal to one if the mother smoked during pregnancy and zero, otherwise.

3. The relevant control variables are:

Mother’s Attributes:

dmage (mother’s age), dmeduc (mother’s educational attainment), mblack (indicator=1 if mother is black), motherr (=1 if neither black nor white), mhispan (=1 if Hispanic), dmar (=1 if mother is unmarried), foreignb (=1 if mother is foreign born)

Father’s Attributes:

dfage (father’s age), dfeduc (father’s education), fblack, fotherr, fhispan (racial indicators for father)

Other Risky Behavior:

alcohol (indicator=1 if mother drank alcohol during pregnancy), drink (# of drinks per week)

Medical Care:

tripre1, tripre2, tripre3 (indicators=1 if 1st prenatal care visit in 1st, 2nd, or 3rd trimester, respectively), tripre0 (=1 if no prenatal care visits), nprevist (total # of prenatal care visits)

Pregnancy History and Maternal Health:

first (=1 if first-born), dlivord (birth order), deadkids (# previous births where newborn died), disllb (months since last birth), preterm (=1 if previous birth premature or small for gestational age), pre4000 (=1 if previously had > 4000 gram newborn), plural (=1 if twins or greater birth), phyper (=1 if mother had pregnancy-associated hypertension), diabete (=1 if mother diabetic), anemia (=1 if mother anemic)

**Questions:**

1. Under what conditions 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? Under the assumption that maternal smoking is randomly assigned, estimate its impact on birth weight. Provide some evidence for or against the hypothesis that maternal smoking is randomly assigned.
2. Suppose that maternal smoking is randomly assigned conditional on the other observable determinants of infant birth weight. What does this imply about the relationship between maternal smoking and unobservable determinants of birth weight conditional on the observables? Use a basic linear regression model to estimate the impact of smoking and report your estimates. Under what conditions is the average treatment effect identified?
3. 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 b)? Now use an approach in the spirit of multivariate matching. In other words, estimate the smoking effects using a flexible functional form for the control variables (e.g., higher order terms and interactions; include lots of controls as you would if you were estimating the propensity score, but just include them as controls in your regression). What are the benefits and drawbacks to this approach?
4. 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?
5. Implement the propensity score approach to the evaluation problem using two methods: 1) control directly for the estimated propensity scores in a regression model; 2) use the estimated propensity score in a classification scheme to “stratify” the sample. Provide empirical evidence that your implementation is reasonable and evidence on the overlap of the observables of smokers and nonsmokers. Present your findings and interpret the results. (This is an open-ended question, so show me what you know and be thoughtful).
6. Now use the estimated propensity scores to reweigh the outcomes and estimate: i) the population average treatment effect; and ii) the average treatment effect among the treated. Compare your estimates to those in e) and interpret your findings. What are the benefits and drawbacks of approaches that use the estimated propensity scores as weights?
7. A more informative way to describe the birth weight effects of smoking is to estimate the “nonparametric” conditional mean of birth weight as a function of the estimated propensity score, for smokers and non-smokers. To do this simply, stratify the smokers into 100 equal-sized cells based on their propensity scores and calculate the mean birth weight and propensity score in each cell. Do the same for the non-smokers. Plot these 2 conditional mean functions on the same graph, with the mean scores on the x-axis and mean birth weight on the y-axis. Interpret your findings and relate them to the results in e) and f). Now redo the above but use 200 equal-sized cells for smokers and non-smokers together – i.e., each cell should contain 1/200th’s of the data, with some cells containing very few smokers and other cells containing mostly smokers.
8. Low birth weight births (less than 2500 grams) are considered particularly undesirable since they comprise a large share of infant deaths. Redo g) using an indicator for low birth weight birth as the outcome of interest. Interpret your findings.
9. Estimate the impact of maternal smoking on infant death using the methods in parts a), b), and g) (using 50 equal-sized cells, for smokers and non-smokers together). Interpret your findings. From your results, what might you conclude about the relationship between smoking and infant death?
10. Smoking rates vary over the life-cycle of women. Plot the sample sizes and smoking rates by the age of the women in the sample. Now, separately for smoking and non-smoking women but on the same graph, plot the average birth weight of their infants by the age of the mother. Describe what you see.
11. For the sample of women aged 16 to 32 (116,243 observations), redo parts a), b), e), f), g) and i). How do your findings contrast with your findings from the overall sample, if at all?
12. Concisely and coherently summarize all of your findings. In this summary, describe the estimated effects of maternal smoking on infant birth weight and infant mortality and whether you think your “best” estimate of the effects of smoking is credibly identified. State why or why not.