Homework #3

Due 10/21/16

1) (**Reweighting**) Go to David Autor's webpage and download the cleaned 1979 and 1997 MORG files from:

http://econ-www.mit.edu/faculty/dautor/data/autkatkear08

a) Read through the cleaning programs associated with the MORG files. Describe the important decisions being made.

b) Make a new variable "lnhr_wage" which is the log of the "hr_wage" variable. Restrict attention to the cleaned sample with "hr_wage_sample==1". Use the "kdensity" command to plot kernel density estimates of lnhr_wage using the variable wgt_hrs (which is the product of the sample weights and the number of hours worked) as an "aweight". Make density plots for both years on the same grid.

c) How has the wage distribution changed over time?

d) A number of demographic shifts occured between 1979 and 1997 that might alter the aggregate wage distribution even if the distribution for each demographic subgroup remained unaltered. Repeat the exercise in b) separately for black men age 25-50 and for white women age 25-50. How are the patterns different?

e) One way of adjusting the distribution for compositional differences is to use propensity score reweighting.

Append the 1979 and 1997 files together and create a dummy variable equal to one for observations in the 1997 sample. Estimate a logit of the 1997 dummy on a dummy for female, a black dummy, an "other race" dummy, and a quadratic polynomial in age. Use the "aweights" when estimating the logit.

f) Use the propensity score implied by this logit to reweight the 1997 wage distribution so that it has the 1979 demographic distribution. Make a kernel density plot of the actual 1979 and 1997 wage distributions against the "counterfactual" reweighted 1997 distribution.

g) Use the propensity score to reweight the 1979 wage distribution so that it has the 1997 demographic distribution (if you get stuck read the Dinardo, Fortin, and Lemieux paper). Make a kernel density plot of the actual 1979 and 1997 wage distributions against the "counterfactual" reweighted 1979 distribution. Be sure to use the product of the wgt_hrs variable and the propensity score based weights when computing the kernel density.

h) Based upon your analysis what would you say was the effect of the demographic changes between 1979 and 1997 on the evolution of the aggregate wage distribution?

i) Examine the suitability of your propensity score specification by comparing the reweighted mean, standard deviation, and the 10th, 25th, 50th, 75th, and 90th percentiles of age in the 1997 sample to the equivalent moments in the 1979 sample.

j) Do the same exercise restricting attention to blacks.

k) Try experimenting with different specifications for age in the propensity score logit and checking how the results of i) and j) change.

l) What do you conclude about the suitability of the propensity score specification?

2) (**Matching**) A popular program evaluation strategy is to match program participants to controls with similar covariates. Dehijia and Wahba (1999) evaluated the National Supported Work program using observational controls via propensity score methods. The key covariate in their analysis of the earnings effects of this program is lagged (pre- program) earnings.

Suppose in their sample, individual i's earnings at time t are generated according to the model:

$$Y_{it} = \alpha_i + \gamma_t + \beta D_{it} + \varepsilon_{it} \tag{1}$$

where D_{it} is an indicator for program participation at time t, α_i is a time invariant unobserved effect, and ε_{it} are unobserved shocks to earnings. For simplicity, assume that only two time periods are available so that $t \in \{1,2\}$ and that program participation only occurs in period 2 (e.g. $D_{i1} = 0\ \forall i$).

A simple evaluation strategy is to eliminate the unobserved effect α_i via a first difference transformation as follows:

$$\begin{aligned} Y_{i2} - Y_{i1} &= \gamma_2 - \gamma_1 + \beta\left(D_{i2} - D_{i1}\right) + \varepsilon_{i2} - \varepsilon_{i1} \\ &= \gamma_2 - \gamma_1 + \beta D_{i2} + \varepsilon_{i2} - \varepsilon_{i1} \qquad (2) \end{aligned}$$

Estimation proceeds via OLS applied to (1) treating $(\gamma_2 - \gamma_1)$ and β as unknown parameters.

a) What restriction(s) on the errors $\varepsilon_{i2} - \varepsilon_{i1}$ are necessary to ensure the OLS estimator identifies β?

b) Derive an expression for Y_{i2} in terms of $Y_{i1}, (\gamma_2 - \gamma_1), D_{i2}$, and $\varepsilon_{i2} - \varepsilon_{i1}$

c) Provide a set of restrictions on the errors $(\varepsilon_{i1}, \varepsilon_{i2}, \alpha_i)$ such that $E\left[Y_{i2}|Y_{i1}, D_{i2} = 1\right] - E\left[Y_{i2}|Y_{i1}, D_{i2} = 0\right]$ identifies β.

d) Discuss the plausibility of this assumption vis-a-vis those provided in your answer to a).

e) Smith and Todd (2005) advocate use of a matched differences in differences estimator. Are there conditions any weaker than those used in your answer to c) under which $E\left[Y_{i2} - Y_{i1}|Y_{i1}, D_{i2} = 1\right] - E\left[Y_{i2} - Y_{i1}|Y_{i1}, D_{i2} = 0\right]$ identifies β?

3) (**2SLS vs Control Function**)

Suppose you are interested in the nonlinear triangular system:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\ X_i &= \gamma_0 + \gamma_1 Z_i + v_i \\ & (\varepsilon_i, v_i) \overset{iid}{\sim} N\left(0, \begin{array}{cc} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{array}\right) \end{aligned}$$

The parameters of interest are $\beta = [\beta_0, \beta_1, \beta_2]$. One method of estimating these parameters is via 2SLS using Z_i and Z_i^2 as instruments for X_i and X_i^2. A second approach uses the observation that, given the structure of the above model, we can write

$$E\left[Y_i|X_i, X_i^2, v_i\right] = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \rho\frac{\sigma_\varepsilon}{\sigma_v}v_i$$

where $\rho \equiv \frac{\sigma_{\varepsilon v}}{\sigma_\varepsilon \sigma_v}$. The two step control function estimator constructs the error $\widehat{v}_i$ as a first stage residual from the regression of X_i on Z_i and then regresses Y_i on X_i, X_i^2, and the first stage residual $\widehat{v}_i$ to get estimates of β and $\rho\frac{\sigma_\varepsilon}{\sigma_v}$.

a) Write the moment conditions defining the 2SLS estimator.

b) Write the moment conditions defining the Control Function estimator (Hint: there are six of them).

c) Does one estimator rely on stronger conditions than the other?

d) Which estimator do you think would work better when the instrument Z is not very strong?

e) Write a Stata program named 'getestimates' that, for a given value of the parameter γ_1, simulates 1,000 observations from the above model using the following assumptions regarding the DGP:

$$\begin{aligned} (Z_i, \varepsilon_i, v_i) &\sim N\left(0, \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & .2 \\ 0 & .2 & 1 \end{matrix}\right) \\ \gamma_0 &= \beta_0 = 0 \\ \beta_1 &= \beta_2 = 1 \end{aligned}$$

After simulating the data, the getestimates program should report 2SLS and Control Function estimates of (β_1, β_2) which we will refer to as $\left(\beta_1^{2SLS}, \beta_2^{2SLS}, \beta_1^{CF}, \beta_2^{CF}\right)$ using the 'return scalar' command. Set the random number generator seed to 12345 and use the 'simulate' command to simulate your getestimates program 1,000 times for parameter values $\gamma_1 \in \{.3, .2, .1\}$. Use the 'tabstat' command to report the mean, median, standard deviation, and interquartile range across simulations of $\left(\beta_1^{2SLS}, \beta_2^{2SLS}, \beta_1^{CF}, \beta_2^{CF}\right)$ for each value of γ_1.

i) Which approach appears to be more efficient? Why?

ii) What happens to the 2SLS estimator as the parameter γ_1 shrinks? How about the CF estimator?

iii) Suppose you want to test the null hypothesis that $\rho\frac{\sigma_\varepsilon}{\sigma_v} = 0$ (no endogeneity). Describe a technique for accounting for the fact that the control function $\widehat{v}_i$ was generated when conducting hypothesis tests in Stata.

4) (**Bootstrap OLS)**

Download the Stata dataset hw3.dta from Bspace. This dataset has a binary outcome y, a binary regressor of interest D, a cluster level control X, and a micro control X2. There are 20 clusters. Set the seed to "123". Use 1,000 reps for all bootstrap procedures.

a) Regress y on D, X, and X2 clustering by the variable "id". Report the clustered standard error for the coefficient on D.

b) What is the p-value for the null hypothesis that the coefficient on D equals zero?

c) Block bootstrap the regression (make sure to use the ", cluster(id)" switch). What is the bootstrap standard error on the coefficient on D? Use this standard error to compute a new p-value.

d) Block bootstrap the clustered t-statistic from the regression (make sure to use both the cluster() switch and the idcluster() switch). Use a symmetric bootstrap percentile t-test to compute a new p-value.

e) Use the Wild bootstrap procedure of Cameron, Gelbach, and Miller (2008) to test the null that the coefficient on D equals zero. Use Rademacher weights and make sure to "impose the null" hypothesis on the coefficient on D in the manner they describe. Report the corresponding p-value.

f) Discuss the different answers given by parts b)-e) in light of econometric theory. Which answer do you think is most accurate?

5) (**Bootstrap Probit**)

a) Using hw3.dta again, fit a Probit of y on D, X, and X2. Report the clustered standard error for the coefficient on D. What is the p-value for the hypothesis that the coefficient on D equals zero?

b) Block bootstrap the clustered t-statistic from the Probit. Use a symmetric bootstrap percentile-t test to compute a new p-value.

c) Conduct a cluster robust score test of the null hypothesis that the coefficient on D equals zero. Report the p-value.

d) Use the "score bootstrap" procedure of Kline and Santos (2012, Journal of Econometric Methods) to bootstrap the score (LM) test. Use Rademacher weights again. Report the p-value. (Hint: You will want to use Mata to compute the bootstrap test. Some example Stata code is available on my website).

e) Which of these p-values do you think is most accurate? Discuss the tradeoff involved in choosing between alternative bootstrap methods.