**Homework # 1**

Please email me your answers (in Word or pdf) and attach the log file and do file. Name all three files as “HW1\_Yourlastname”. Use the do file template. Please do not copy and paste from the stata output window. Make sure that your do file runs from the beginning to the end; de-bug the program until it runs smoothly. Use PSID2003.dta and PSID2005.dta for this exercise. Consult with attached codebooks and questionnaires.

1. Extract only variables that you will use for this assignment.
   1. From **PSID2003.dta**: *indid* - ER21002\*100+1; *age* - ER21017; *female* - ER21018; *married* - ER21023; *wrknow* - ER21123; *d=wagedum* - ER21928; *wagamt* - ER21929; *numkid* - ER21020; *college* - ER23450 ER23434 ER23446 ER23454, weight - ER24179.
   2. From **PSID2005.dta**: choose comparable variables and also extract *train* - ER27704.
2. Merge two data sets. Use hhlink\_ex.dta to link the data sets. What is the 2-year attrition from the survey: report the share of households that participated in the 2003 survey but did not participate in 2005?
3. Create variables appropriately
   1. Recode missing values as “.” In case of filter questions, recode variables according to the sequence of questions in the questionnaire.
   2. Make *female*, *married*, and *train* as binary variables (1, 0); create 3 categories for the *numkid* (0, 1=1 or 2 kids, and 2=3+ kids); create log of earnings for the head of household; gen *college*=1 if respondent has a college degree and 0 if no college degree (be careful with previous filter questions; for those who studied outside the U.S. assign college=1 if number of years studied>=16&<. and =0 if <16).
   3. Recode *wrknow* =1 if household head is currently working; =0 if household head is unemployed and looking for a job; =. for other options (use the variable from the 1st mention of employment status).
   4. Label all variables. Report unweighted summary statistics for all variables for both years.

Check point:

Variable Obs Mean Std. Dev. Min Max

indid2003 7822 392472.1 226861.2 101 787001

indid2005 7822 349316.5 237206.3 0 804171

age2003 7821 44.97826 15.97441 16 98

numkid2003 7822 .5501151 .6643472 0 2

married2003 7821 .5063291 .4999919 0 1

wagdum2003 7815 .7733845 .4186685 0 1

wagamt2003 7316 30794.33 63845.5 0 3500000

weight2003 7822 15.42355 13.5996 0 82.804

female2003 7822 .2990284 .4578616 0 1

wrknow2003 6212 .924018 .2649906 0 1

college2003 7257 .2760094 .4470523 0 1

age2005 6797 47.19891 15.84043 17 99

numkid2005 6797 .5436222 .6662109 0 2

married2005 6797 .5421509 .4982568 0 1

wagdum2005 6791 .7769106 .4163485 0 1

wagamt2005 6416 34905.42 67067.08 0 2710000

weight2005 6797 22.18839 17.05561 .14 132.641

female2005 6797 .2824776 .4502375 0 1

wrknow2005 5308 .9423512 .2331002 0 1

college2005 6310 .2844691 .451197 0 1

1. Report mean values using sample weights (use *pweights*). Comment on the differences between weighted and unweighted means of variables.
2. Reshape the data set into a longitudinal (long) format. Declare this dataset as panel (tsset or xtset).
3. Create a transition matrix for employment status from 2003 to 2005 (use time-series operators such as lag or lead). Interpret the matrix. What is the estimated unconditional probability that the unemployed in 2003 will find a job in 2005? What is the estimated unconditional probability that the employed in 2003 will not have a job in 2005?
4. Estimate a probit model explaining the probability of finding a job between 2003 and 2005 for those who did not have a job in 2003; include age, numkid, married, female, and college of household head. Use heteroskedasticity-robust standard errors. Report marginal effects. Interpret the results.
5. Suppose you decided to examine whether government assistance with job search and training in 2004 increased the probability of finding a job between 2003 and 2005. Add *train* to your specification in (8) and report the results. Why is this variable likely to be endogenous? Be specific.
6. Drop respondents with zero weight (*weight*=0). Regress the log of earnings on gender, age, college of household head, and a year dummy. Interpret the coefficient on college degree in this baseline specification (convert log points into percent differences in earnings between college graduates and others). Now re-estimate the model with sample weights. Comment on the differences between weighted and unweighted estimates of the returns to college.
7. Apply inverse probability weighting to the earnings function in (9) using 1/*p* as weights, where *p* is fitted probabilities from estimating probit of positive wages on gender, age, number of kids, marital status, and a year dummy. How did the returns to college degree change in this specification compared to the baseline specification?