ECON 70430
Fall 2022
Problem Set 1
Due: August 31, 2022 via Canvas

We have discussed the Current Population Survey (CPS) a number of times in class. You will use the actual CPS microdata to answer a set of questions. The completed assignment should contain a writeup with formatted results. It should also include a set of do files (one for each part A-C) that can be used to replicate them. The do files do not need to automate the formatting, but this can be done (I'll show you what I do in the problem set solutions). I should be able to download the data and run the do files with minimal work if any to change paths, etc...

A. (Matching the July Employment Situation) You will compute by hand various statistics that are reported in the Household data section (part A) of the July 2022 release and compare your statistics to the published values. To do this you will need to download CPS microdata from IPUMS (`https://cps.ipums.org`). Download the months January through July of 2022. To make an extract you will need to choose the appropriate variables. Read the rest of the problem set questions to determine what you will need. If you make a mistake you can always add a variable and re-extract. The IPUMS data are convenient because they are standardized with longitudinally consistent categories and labels defined. Other datasets are not as user-friendly. Typically you will need to consult the "code book" that describes how the survey question responses are coded in the dataset.

1. Do some preliminary tabulations.
   (a) Produce a table of the number of observations for each value of `empstat`. How many observations are in the civilian labor force across all months? (Stata: `help tabulate)`
   (b) Tabulate the variable only for observations in the *March* survey. How many observations are considered employed for March?
   (c) What numeric values correspond to each of the unemployed categories?
2. Estimate the national unemployment rate as the equally weighted share of unemployed (use the `empstat` variable) out of all individuals who are 16+ (use the `age` variable) and part of the civilian non institutional population (use the `popstat` variable). What is your estimate of the July unemployment rate? How does it compare with the unemployment rate reported in the July employment situation? To do this, you could generate an indicator variable `unemp` that is 1 if the individual is unemployed, 0 if the individual is employed, and otherwise missing and then calculate an average. In Stata, running `summarize unemp if month==7` will report the average, which will be equal to the unemployment rate. Using `tabstat unemp if month==7, statistics(mean)` would also do the same, or even `regress unemp if month==7`.
3. One reason the estimate differs is that the CPS is not a uniform random sample. Smaller groups of people are sampled at a higher probability, so that statistics computed for these smaller groups are estimated with the same precision as for larger groups. This also means, however, that for overall statistics, a simple average over the the observations will not estimate the expectation correctly. Each observation should be weighted by the inverse of its probability of being sampled, so that oversampled observations receive less weight. These "sampling weights" are provided in the `wtfinl` variable. These weights are normalized so that the add up to the total population. You can think of this as indicating how many people are represented by a single observation. For July, compute the sum of the `wtfinl` variable for the adult civilian non institutional population. How does this estimate compare to the total adult population reported in the July employment situation?
4. Now compute the July unemployment rate weighting by the `wtfinl` variable. In Stata, use the analytic weight `aw=wtfinl`. (Some commands like regress will accept a probability weight

pw=wtfinl, which will correct the standard errors for the estimates to account for the nonuniform sampling. The (point) estimates will be identical whether aw or pw weights are used.) How does it compare with the unemployment rate reported in the July employment situation?

5. The wtfinal weights from the IPUMS CPS do not include several additional adjustments that the BLS uses for its published tabulations. The BLS weight used for the published totals is saved as compwt. Using aw=compwt compute the unemployment rate for July. How does it compare to the unemployment rate reported in the July employment situation? Remember to compare against the non seasonally adjusted unemployment rate.
6. Now that you can (hopefully) match the BLS published totals, compute the unemployment rate and employment to population ratio for January to July 2022. Compare these to the published values.
7. For January to July 2022, compute the labor force participation rate for the 20-24 year old by gender and educational attainment (high school dropout, high school graduate, some college, college or more). Do the same for the 60-64 year old age group. Produce a table with these values. The BLS does not publish totals that are cut by age, gender and educational attainment.

B. (Gross worker flows) The sampling design of the CPS means each household (which remains at the same address) in the CPS participates for 4 months, then is out for 8 months and the participates for another 4 months. The mish variable indicates which month (from 1 to 8) of the rotation each observation's household is in. The design of the CPS means 75% of all households that remain at the same address can theoretically be matched to corresponding observations in the following month. In the base month, mish 1-3 and mish 4-7 can be matched to mish 2-4 and mish 5-8 in the following month. In practice the actual percentage that can be matched is lower because of attrition and non response.

1. Construct a matched sample that contains each observation from January to June matched to its corresponding observation in February to July when possible. For example, a January observation will have the variables with the January data, and when it can be matched to a corresponding observation in February it will also have variables with the February values. You will need to use the merge 1:1 command to do this. In theory an observation should be able to be matched across months using the hrhhid, hrhhid2, lineno, and mish (or month) incremented by 1. The first two variables identify the specific household, the third identifies the person in the household, and the 4th the time period. What is the match rate for each month?
2. Sometimes there are false positives, drop matches where age declines or gender/race switch in the next month. What is the match rate by month after the corrections?
3. Using the matched sample, compute a monthly job loss rate (E to U) and job finding rate (U to E). Report these rates for January to June 2022
4. Compute the same rates for each education group and report them for January to June 2022.
5. Compute the same rates for each education group and month only using the mish 2-3 in the base month.
6. Add the months January to July 2021 to your sample and construct a matched sample that matches each 2021 month to its 2022 month counterpart. Given the sample design, what is the theoretical maximum match rate. What is the actual match rate for each month? Report a table.
7. Calculate for the 20-24 age group and the 60-64 age group (both genders and all education groups) the share of each population by month that entered the labor force between 2021 and 2022. Report this in a table.

C. (March Annual Supplement, ASEC) The households participating in the CPS in March are asked a number of extra questions. These annual supplements are available as far back as 1962. Using IPUMS, create an extract that contains the ASEC data for 3 year windows at the start of each decade from 1970 (i.e., 1969-1971), to 2020.

1. Compute labor force participation for ages overall (for ages 20-54) and by gender. Then compute by gender and educational attainment (use the same categories as above) and age groups (20-24, 25-34 35-44, and 45-54) for each decade (the 3 year window). Report these estimates in a table.
2. Use a shift share analysis to estimate the contribution of just changes in age groups and educational attainment to the overall participation rate by gender. Relative to 1970 (3 year window), report the fraction of the change in participation for each decade that can be explained by aging and changes in educational attainment. *HINT* In Stata, it will be helpful to do a `reshape wide` so that each observation corresponds to a time period and repeats a set of variables for each demographic group.
3. Produce a graph of the actual LFPR for 20-54 males and the LFPR predicted only by changes in age and education shares.
4. Produce the same graph for LFPR of 20-54 females.