**RA assignment for Trevor, Fall 2019**

**Task :** Use data from the NLSY to estimate contribution of motherhood to the size of permanent shocks

**Part 1 Instruction:**

1. Go to the NLSY79 website ([link](https://www.nlsinfo.org/investigator/pages/search.jsp?s=NLSY79))
2. Create an account so you can save the variables, as I'm sure we will have to add some in the long run.
3. Click "Variable Search" and get:
   1. Under "Income assets and program participation" get The youth's annual income for all the years (annual from 1979-1994, then biannual after).  Note that this is typically the "Q13-5 truncates" variable.  I don't need all the incomes questions (like "is income above 15k?") just the value.
   2. In the same place get the youth's spouse's annual income for all years.  Same as the youth's.
   3. In  the "income assets and program participation" also get the net family worth (under assets)  for all relevant years (some are missing)
   4. And get "net family income" for all relevant years
   5. Under the "children' tab download the children information:  about first child primarily.  (I.e. month of birth and year of birth of first, second and third child).
   6. Also download the "child care arrangement" variables for the first and second children:  I think this the Q10\_32 variable.
   7. Get marriage variables (i.e. if the person is married)
   8. I think respondent sex is automatically downloaded.
4. Great, now we have a first pass of all the variables we'll need:  {Gender, Income, Spouse Income, Family Income, Assets, Children's birthdates/presence, children's childcare arrangements}
5. Go to "downloads> advanced download"  and **save** and download a stata version of the NLSY79 data.  It will come with a
6. Load them all into stata and name them properly/coherently. The zip file you download will come with three important files:
   1. dataset.csv, dataset.dct, and dataset-value-labels.do
   2. In stata, go to file>import>import text data in fixed format with a dictionary
   3. choose the dictionary dataset.dct and the text dataset dataset.csv
   4. Then go to the dataset-value-labels.do
   5. Uncomment the Crosswalk for Reference number & Question name rename statements at the bottom of the file...this gives the variables their real names.
   6. Run dataset-value-labels.do on the imported dataset.
      1. The dataset now has a person as unit of observation, and all variables are for that person
   7. Rename all the variables to be even more coherent, i.e. all the person's income variables could be Inc\_year, all their spouses variables would be IncS\_year, net family income NetInc\_year, etc. etc.  Consistent names with \_year at the end.
   8. Reshape long so that the unit of observation is the person-year rather than the person. i.e. reshape long Inc\_ IncS\_ NetInc\_, i(CASEID) (or whatever the id variable is) j(year)
   9. Replace the missing with blanks (most variables have things like -4 and -5 because they were left blank.  replace those with missing
   10. Create variables denoting the presence of a child in a given year.  I.e. how many years until first birth, using the child's birth date and the year of the observation.
7. All this should give us a longitudinal dataset of women's income, assets, spousal assets, spousal presence, and children over time.
8. Let me know when you're done and we can think about next steps.  I'm interested in doing an event study of first birth, not dissimilar to the graphs I've attached below from the PSID.

**My work:**

1. I used .dct and .dat files to import the data, I was having some issues when I tried to use .dct and .csv
2. I kept both Q10\_32 (denotes child care in **second** year) and Q10\_13 ( denotes child care in**first** year)  variables
3. In reference to point**6(j):**  
   I have generated the child presence variable ("haschild")  keeping in mind the goal of creating event study graphs, let me know if this is how you wanted to define it

We can get the working data by running the "dataclean.do file"(attached) using single click.

**Part 2:**

Your NLSY stuff was great!  To go further, we're interested in the contribution of motherhood to the size of permanent shocks.  Recall from Meghir and Pistaferri ECMA 2004 (if I recall correctly) that, for a simple permanent shock + one-period transitory noise, we have:

cov(Dy\_{t-1}+Dy\_t+Dy\_{t+1} , y\_t) = variance of permanent shocks

Where Dy is the change in mother's income.

I'm interested in seeing the contribution of motherhood.  The means that we want two objects:

* cov(Dy\_{t-1}+(Dy\_t+(Dy\_{t+1} , y\_t) where t, t-1, and t are all years in which someone is not a first-time mother.  (i.e. when the countdown until motherhood is t<-1  or t > 1, strict inequality.
* cov(Dy\_{t-1}+(Dy\_t+(Dy\_{t+1} , y\_t) where t is the year in which someone is a first-time mother (i.e. when countdown until motherhood year is t=0).

The first object is equal, as in Meghir and Pistaferri, to the variance of the non-mother permanent income shock.  The second object is equal to the variance of the non-mother permanent income shock plus the mother permanent income shock.  The difference between them is the variance of the motherhood permanent income shock.

Could you calculate these in the data you gave me?  I would use only the non-biennial data (before and equal to 1994).

**My work:**

1.  I obtained data on education and other demographic variables. I**updated raw data** files. I have also updated**dataclean.do** file accordingly.

2.  I have used**PCE** (Base 1990 = 100) to create the real income variables. but I have also coded **CPI** (as used by Blundell et al. 2008 ).

3. For now, I haven't included these **education dummies**  in the regression, one problem with education variable is that it’s not available for all years. I made some manual adjustment to create maximum grade achieved  (see **incvariance.do**) following Blundell et al. 2008 codes.

4. I have figured out how to estimate the two auto-covariances. So what I have done so far is:

for each mother I have generated (D\_y{t-1} + D\_y{t} + D\_y{t+1})\*D\_y{t} (both for motherhood years and non-motherhood years)

Question: Should I take the expectation of  (D\_y{t-1} + D\_y{t} + D\_y{t+1})\*D\_y{t} by each year and then take the difference for each year ('cause the blundell et al 2008 paper estimates the covariances by year).

5. I have created a master.do file so that you can run all .do files using one click