**8-22-17 SAS code data sets**

* **cohort\_XXXX\_wages.csv**: This data set contains variables at the pooled TIN by tax year level (i.e. not cut by quartile) on the XXXX variable of interest. This variable can be one of 6 things: male, female, inventor, non-inventor, under40 (i.e. <=40 years old), and over 40 (i.e. >40 years old)
  + **payer\_tin\_w2\_max**: The employer TIN of the observation.
  + **tax\_yr**: The tax year of the observation.
  + **cht\_XXXXs**: This variable reports the number of people of type XXXX {male, female, inventor, non-inventor, under 40, over 40} in that application year cohort for that firm. For example, the variable cht\_under40s for the TIN 3, tax year 2000 observation in the “cohort\_under40\_wages.csv” data set would store the number of people who worked at firm 3 in the year 2000 who were 40 years old or younger.
  + **XXXX\_chtwages\_YYYY**: This variable reports the mean year YYYY wages of people of people who worked at the given firm in the given tax year of type XXXX. For example, the variable male\_chtwages\_2006 for the TIN 3, tax year 2000 observation in the “cohort\_male\_wages.csv” data set would store the mean 2006 wages of males who were employed at firm 3 in the year 2000. Note that if the worker drops out of the panel their wages are counted as zero for this variable.
* **firm\_ages\_college\_tenttax.csv**: This data set contains variables at the pooled TIN by tax year level (i.e. not cut by quartile) on the mean age at the firm in that year and mean wages cut by under/over 40 years of age.
  + **payer\_tin\_w2\_max**: The employer TIN of the observation.
  + **tax\_yr**: The tax year of the observation.
  + **mean\_age**: This variable measures the mean ages of people at the firm (for people who have such an age metric available) in that tax year.
  + **over40s**: This variable measures the number of people who are above 40 at the given firm (not inclusive, i.e. >40) in that tax year.
  + **under40s**: This variable measures the number of people who are below 40 at the given firm (inclusive, i.e. <=40) in that tax year.
  + **mean\_over40wages**: This variable measures the mean wages of people who are above 40 at the given firm (not inclusive, i.e. >40) in that tax year.
  + **mean\_under40wages**: This variable measures the mean wages of people who are below 40 at the given firm (inclusive, i.e. <=40) in that tax year.
  + **share\_college**: This variable reports the share of people at the given firm in the given tax year who went to college, conditional on the `attend\_college’ indicator being non-missing.
  + **college\_reported**: This variable reports the number of people at the given firm in the given tax year who have a non-missing `attend\_college’ indicator from the “databank\_masked” data set.
  + **mean\_mean\_tentative\_taxes**: This variable measures the mean tentative taxes paid for the given firm in the given tax year.
  + **tentative\_taxes\_reported**: This variable reports the number of people at the given firm in the given tax year who have a non-missing value for the `tent\_tx’ variable from the “databank\_masked” data set (should that be a potential problem). This count is provided to be able to more accurately back out the total tax burden at the firm by year level (mean\*count) should that be desired.