Merging Data into Crosswalks

The code used to run each of the following sections is contained in *merge_in_variables.do*. The following document explains each section of the code in more detail.

Section 0: Set Up

The set-up part of the code allows the user to pick matches made using particular methods and to indicate the starting and final years of the desired crosswalk into which data will be merged. It should be noted that if a user wants to keep matches from all different methods they can also do so. In that case do not specify the macros *names* and *method* and change the code in Section 2 accordingly. The set-up also allows the user to pick the specific variables they would like to merge in. Specify those variables in the local *keepvar.* In this example code, the variables that are being merged in are *age, bpl* and *occscore* but there are many other census variables available on the IPUMS website and researchers should select the variables of interest given their context.

Please note that this code is just an example and does not match everyone's requirements and contexts. Thus, the code needs to be amended accordingly.

Section 1: Extract Data by Year from IPUMS Downloaded data

Before executing this section of the code, researchers should first download the variables of interest from the IPUMS website for the relevant years. Recall that the variable *histid* will be needed to merge this data into the crosswalks. This part of the code then separates and prepares data for each year to be merged into the crosswalks.

Section 2: Merge Data into Matched Crosswalks

This section merges in the variables indicated in the *keepvar* local under set-up into the historical crosswalk. Researchers can designate here if they want to only keep matches made using the particular selected method. If variables are to be merged in for matches made using multiple methods, then those methods should be specified using the keep if option accordingly. If matches from more than one method are to be kept, the m:1 option in the merge command should be specified. A many-to-one merge is required in this case because different algorithms may pick different matches for the same individual, and consequently *histids* do not uniquely identify observations for a given year in a master crosswalk.

The resulting dataset will be longitudinal data which will have variables associated with each year. The result of the example code will provide the three selected variables (specified in *keepvar)* once for each year. For instance, the resulting dataset in this particular example will provide for a given individual, his age in 1850 (given by *age_1850)* and his age in 1860 (given by *age_1860).* The longitudinal data can then be used for analysis.

Last updated: 4/7/20

Please contact ranabr@stanford.edu (Ran Abramitzky), lboustan@princeton.edu (Leah Boustan), and/or myerar@princeton.edu (Myera Rashid) with any questions or feedback about this code.