Example Code to Generate Weights for Reweighting the Data

The code used to run each of the following sections is contained in *generate_weights.do*

Linked census datasets may not be representative of the population and thus may need to be reweighted to be used in analysis. For example, men with common names often cannot be uniquely identified and men with lower numeracy may misreport their age. As a result, men that can be matched may have higher socio-economic status than those who cannot be matched. This example code demonstrates how researchers can generate weights in order to reweight the data for analysis.

Section 0: Set Up

The set-up part of the code allows the user to select the method for which they would like to keep matches. It also allows the user to select the corresponding years of the selected crosswalk. The *keepvar* macro allows users to select the variables they want to reweight on. In this example code, the variables used to generate weights are *lit (*literacy), *age, urban (*whether a person lives in an urban or rural area), and *occscore* (occupational income scores).

Section 1: Preparing Full-Count Data to be Merged in

This part of the code uses full count data (downloaded from IPUMS) for each year and prepares it for merging. The code has several checks to make sure that *histid* in the downloaded data is unique.

Section 2: Merge Data into Crosswalk

Merge data from Section 1 into the matched crosswalk. An important component of this section is to pick the year that the reweighting variables will come from. This is context dependent. In the example code, we use reweighting variables from the later year. It is also a common practice, context permitting, to reweight using later year and starting year separately and then to compare the results as a robustness check. Note that the researcher will want to make sure to compare the matched sample to the *population at risk to be matched* so that it sometimes will not make sense to compare in the later or earlier year.

Note that one should merge in data for the year used for the reweighting procedure and data must be merged in for both the matched and non-matched observations. In this example, we merge in data for the later year (*Year2*). On the other hand, only data for matched observations is merged in for *Year1*.

It is important to flag the matched observations when merging in *Year2* data. This will allow matched observations to be distinguished from unmatched ones.

Section 3: Generate Weights

Treat continuous and categorical variables differently. For continuous variables, make bins and for categorical variables, create dummy variables. For instance, in this example the continuous variables are *age* and *occscore,* whereas the categorical variables are *lit* and *urban.*

A probit is then run on the categorical variables (indicators) and continuous variables (binned). Using the predicted $\hat{p}$ from the probit, weights are then generated for matched observations. Unmatched observations are given a weight equal to 1.

Section 4: Create Balance Tables

For continuous and binary covariates, balance tables can be created using the [pw = weight] option in the regression to include weights. It should be noted that here we use the original variables, not the binned or dummied versions of the variables that were created to generate the weights.

Section 5: Matched Sample with Weights

After keeping the matched observations and weights, we can merge the weights into the data for analysis using *histid*. At this point, the resulting data can be used to run the analysis and the [pw=weight] option can be used to run weighted regressions.

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.