Public Data Download Instructions

This replication project requires public data to be downloaded from four projects:

1. IPUMS USA
2. The Census Tree
3. The IPUMS MLP
4. The Census Linking Project

This document will specify the exact data required for download and provide links to main pages. You can also watch a tutorial video on creating and preparing all the data we use in this paper at this link:

https://youtu.be/28zR7QE0sfc

IPUMS USA

The census microdata stored by IPUMS USA is the backbone of this project. You'll need to download a dataset for each of 1900, 1910, 1920, and 1940, with slightly different variables in some datasets due to availability differences. Begin by registering a user account with IPUMS USA by navigating to this link:

https://usa.ipums.org/usa/

and selecting the "Register" option in the top right corner. Once you've registered an account, create 100% sample data extracts in Stata format (.dta) for each of the 1900, 1910, 1920, and 1940 censuses that 1) only include observations with an OCC1950 code less than or equal to 970; and 2) contain the following variables (in addition to the IPUMS USA preselected variables):

1900	1910	1920	1940
STATEICP	STATEICP	STATEICP	STATEICP
COUNTYICP	COUNTYICP	COUNTYICP	COUNTYICP
URBAN	URBAN	URBAN	URBAN
SEX	SEX	SEX	SEX
BIRTHYR	BIRTHYR	BIRTHYR	BIRTHYR
RACE	RACE	RACE	RACE
BPL	BPL	BPL	BPL
CITIZEN	CITIZEN	CITIZEN	CITIZEN

YRIMMIG	YRIMMIG	YRIMMIG	-
YRSUSA2	YRSUSA2	YRSUSA2	-
SPEAKENG	SPEAKENG	SPEAKENG	MTOUNGUE
LIT	LIT	LIT	HIGRADE
OCC1950	OCC1950	OCC1950	OCC1950
-	-	-	INCWAGE
OCCSCORE	OCCSCORE	OCCSCORE	OCCSCORE
SEI	SEI	SEI	SEI
PRESGL	PRESGL	PRESGL	PRESGL
ERSCOR50	ERSCOR50	ERSCOR50	ERSCOR50
EDSCOR50	EDSCOR50	EDSCOR50	EDSCOR50

Once you've created these four data extracts, download them and unzip them. You may need to use a program like 7-Zip if your machine cannot natively handle .gz zip files. Rename each extracted file "ipums_baseline_year.dta", replacing year with 1900, 1910, etc. for each respective file. Place those renamed files into a folder/directory titled "ipums_sets".

The Census Tree

The Census Tree is an extraordinarily large set of record links across each of the historical U.S. censuses from 1850-1940. It is also the largest available set of links for women by a considerable margin due to its origins as a genealogical link-based set. It is encoded using the IPUMS variable HISTID, which uniquely identifies individuals in the IPUMS USA data extracts detailed above. You'll need to download a crosswalk dataset from 1900 to each of 1910, 1920, and 1940. Begin by navigating to this link:

https://www.censustree.org/data

Upon arrival, select the buttons that correspond to the desired crosswalks; they are in the 1900 column and the 1910, 1920, and 1940 rows. Each button will take you to ICPSR, where you will also need to create or log into an account. After doing so, select each individual crosswalk file and download it. If the file is zipped, unzip it. Once you have downloaded each crosswalk file, place them all into a folder/directory titled "census_tree".

The IPUMS MLP

Like the Census Tree, the IPUMS Multigenerational Longitudinal Panel is a set of record links across historical U.S. censuses. It is a popular linking set due to its ease of integration with IPUMS USA samples on extract. However, the most recent version of the MLP (version 1.2) is not available via the IPUMS USA extract system as of writing. In addition, the crosswalks only extend 30 years, so there's no crosswalk from 1900 to 1940. As such, you'll need to download a crosswalk dataset from 1900 to each of 1910 and 1920, and then download the crosswalk from 1920 to 1940. Begin by navigating to this link:

https://usa.ipums.org/usa/mlp/mlp_census_crosswalks.shtml

Upon arrival, scroll down to the "CROSSWALK FILE DOWNLOAD" section. Select the Stata links that correspond to the desired crosswalks; they are in the "10-year" and "20-year" columns. Each button will directly download the corresponding file. Once you have downloaded each crosswalk file, unzip them all and place them all into a folder/directory titled "mlp".

The Census Linking Project

The Census Linking Project is another set of record links across historical U.S. censuses. Its relevance to this project stems from its authors being the authors of the paper we are replicating. Indeed, the Census Linking Project finds its origins in this and two other papers using an early version of its record linking method. You'll need to download a crosswalk dataset from 1900 to each of 1910, 1920, and 1940. Begin by navigating to this link:

https://censuslinkingproject.org/data

Upon arrival, set the drop-down menus to the appropriate years and click the download link. Select the Stata links that correspond to the desired crosswalks; they are in the "10-year" and "20-year" columns. Each link will take you to the Harvard Dataverse, where you can download each crosswalk without creating an account. Once you have downloaded each crosswalk file, unzip them all and place them all into a folder/directory titled "clp".

Final Prep

You should now have four folders of unzipped .dta or .csv datasets: "ipums_sets", "census_tree", "mlp", and "clp". Place all of those folders into the "raw_data" folder in the repository you downloaded. After doing so, you'll be able to run all the code no problem 😎