1. Software

Stata version 17.0

2. Raw data sources
    a. United States Federal Census 1% samples, 1850-1930, from Ruggles et al (2024a). Detailed documentation and variable definitions can be found at https://usa.ipums.org/usa/. (Folder: census_1pct, files beginning with years)
    b. Name counts by gender derived from 1% Census samples 1850-1940, from Olivetti and Paserman (2015). Data can be found as part of the replication package at https://www.openicpsr.org/openicpsr/project/112929/version/V1/view (Folder: census_1pct, files beginning "namecount").
    c. Full count U.S census data, 1850-1940, from Ruggles et al (2024b). Detailed documentation and variable definitions can be found at https://usa.ipums.org/usa/. (Folder: census_100pct)
    d. County- and state-level U.S. census data, 1850-1940, from Haines and ICPSR (2010). Detailed documentation and variable definitions can be found at https://www.icpsr.umich.edu/web/ICPSR/studies/2896. (Folder: icpsr, files of the format "ICPSR_year_st" or "ICPSR_year_stct")
    e. County- and state-level data from the U.S. Census of Agriculture, from Haines, Fishback and Rhode (2018). Detailed documentation and variable definitions can be found at https://www.icpsr.umich.edu/web/ICPSR/studies/35206. (Folder: icpsr, files ICPSR_1900_agcensus.dta)
    f. County-level railroad access data, used in Atack et al (2011), obtained directly from the authors. The database was originally created by reading scanned historical railroad maps using GIS. Detailed variable labels are included in the raw data files.
    g. Occupational wage data from 1900, obtained from Olivetti and Paserman (2015). This database was originally transcribed from Preston and Haines (1991). This book contains average wages in 1900 for occupations (excluding farmers). The dataset was originally generated by creating a crosswalk between occupational titles in Preston and Haines (1991) and 1950 occupational codes from Ruggles et al (2024), then calculating mean occupational income by 1950 occupational code. Farmers' wages are imputed two ways: (1) by assigning farmers the average income of farm laborers; (2) by imputing the income of farmers from the 1900 Census of Agriculture using the procedure of Abramitzky et al (2012). Detailed variable labels are included in the raw data files. (Folder: alt_wage, file "occ_wage_PH")
    h. Occupational wages by industry, occupation, and demographics from Saavedra and Twinam (2020). Data can be found at

https://sites.google.com/view/martinsaavedra/data. (Folder: alt_wages, file "lido_score_1950_public_use").

3. Guide to folders

    a. "dofiles": Stata code to replicate all other figures and tables in the paper.
        i. Subfolder "01-clean_data" contains programs which clean and manipulate raw data files to create the estimation samples
        ii. Subfolder "02-analysis" contains programs which produce the paper tables and figures. Each file is named with the set of items it produces.
    b. "rawdata": The raw data files underlying the analysis as described above.
    c. "intermediatedata": The folder where intermediate datasets produced by the Stata code will be stored.
    d. "output": The folder where figures, tables, and regression results will be output by the Stata programs.
    e. "dofiles_nber": Stata code to run on the NBER server to extract the 1% sample of the 1940 census.
    f. "1940_data_details" contains additional information about restricted access 1940 census data, including instructions for how to request access at the NBER.
    g. "Appendix_A": Stata and R code to replicate the figures and tables in Appendix A. See the README file within this folder.

4. Detailed instructions

    a. Open the Stata file "Run Everything" within the "dofiles" folder.
    b. Replace the global "maindir" with the directory where you have placed the replication package.
    c. IF you have access to the restricted access to the 1940 census data with first names on the nber server, execute the code "00-RunEverythingNBER" in the dofiles_nber subfolder. You will need to change the "censustemp" global to be the filepath to your personal working directory. This will extract the 1% sample of the census we used for this analysis.
    d. IF you have access to the restricted access 1940 census data with first names, ensure that the 1% sample of this census is saved with the name "1940_1%.dta" in the "rawdata\census_1pct" subfolder
    e. The code is set up to run *without* restricted access 1940 data. IF you have added 1940 data to the rawdata folder, comment out the block of code beginning "global endyear 1930" and uncomment the block of code beginning "global endyear 1940" to replace variable lists in the regression

analysis with the full set of years. If you do not have 1940 data, do not edit the code further.

f. Execute the do file. This will run all Stata programs in the "01-clean_data" and "02-analysis" subfolders to produce figures and tables (absent data points for 1940 if that data was not included).
    i. Additional notes on Appendix C tables. These files do not create formatted tables, but will produce an excel file which includes all statistics for each regression reported in the tables. For table C1, the relevant specifications shown in the table are those with a "depvar" of 1 (Ever Married) or 2 (Father-in-law SES), and sttype of 2 (controls assigned by state of birth). The first row in each section corresponds to spec "1" and the second to spec "2." The coefficients in the first five columns of the table correspond to the coefficient on quartile 4 in the respective year with control "1", and the last column is the coefficient on quartile four in 1940 for control "2". For tables C2-C5 and C7, the relevant specifications shown in the table are those with a "depvar" of 1 (Ever Married) or 2 (Father-in-law SES), and the coefficient shown on the table is that for quartile 4 in 1940. Similarly, for table C6, the 1940 coefficient on the top quantile for that specification is shown for "depvar" 1 and 2. If 1940 is not included, point estimates will be very slightly different because of the different sample slightly altering the coefficients on age.

5. References

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102, no. 5, pp 1832-56.

Haines, Michael R., and Inter-university Consortium for Political and Social Research. Historical, Demographic, Economic, and Social Data: The United States, 1790-2002. Inter-university Consortium for Political and Social Research [distributor], 2010-05-21. https://doi.org/10.3886/ICPSR02896.v3

Haines, Michael, Fishback, Price, and Rhode, Paul. United States Agriculture Data, 1840 - 2012. Inter-university Consortium for Political and Social Research [distributor], 2018-08-20. https://doi.org/10.3886/ICPSR35206.v4

Olivetti, Claudia, and M. Daniele Paserman. 2015. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." *American Economic Review*, 105 (8): 2695-2724.

Preston, Samuel H. and Michael R. Haines. 1991. *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Princeton University Press.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS, 2024a. **https://doi.org/10.18128/D010.V15.0**

Steven Ruggles, Matt A. Nelson, Matthew Sobek, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Evan Roberts, and J. Robert Warren. IPUMS Ancestry Full Count Data: Version 4.0 [dataset]. Minneapolis, MN: IPUMS, 2024b. **https://doi.org/10.18128/D014.V4.0**

Saavedra, Martin and Tate Twinam. 2020. "A Machine Learning Approach to Improving Occupational Income Scores." *Explorations in Economic History*, 75.