"The Allocation of Talent and U.S. Economic Growth"
Replication Instructions

Erik Hurst, Chang-Tai Hsieh, Charles I. Jones, and Peter J. Klenow

Version 7.0
April 2019

This file describes the data and programs that are used to generate the empirical results in the paper. The results can be replicated by following the instructions below.

Part 1: Data

All raw data from the paper come from the 1960, 1970, 1980, 1990, and 2000 U.S. Censuses and the pooled 2010-2012 American Community Surveys (which we refer to a year 2010 throughout). We downloaded the data directly from the IPUMS website at https://usa.ipums.org/usa/.

The following stata programs and files are used to create "chad_output_file_`date'.cvs" that is used in the results section described in "Part 2: Results" of this document.

The files for replication can be found in "data_talent_econometrica_replication.zip". That zipped file contains the following:

1. The stata do files: "create_`*year*'_IPUMS_extract.do" where the `*year*' term is either 1960, 1970, 1980, 1990, 2000, and 2010. As a result, there are six separate files. Each one of these files takes the raw data from IPUMS and transforms the data for use in our analysis file. In particular, we restrict the sample along the lines discussed in the text and create consistently defined demographic, employment, education, occupation and income variables for each year.

2. We also include the stata output files (*.dta) for each year from these transformations. These files are entitled "`*year*'_extract_composite_main.dta" where again the `*year*' term is either 1960, 1970, 1980, 1990, 2000, or 2010.

3. Finally, we use the stata do file entitled "create_ipums_1960_2010_analysis_file.do" to create our main analysis file entitled "chad_output_file.cvs". To replicate our main analysis file simply run "create_ipums_1960_2010_analysis_file.do" with the six separate "`*year*'_extract_composite_main.dta" files in the same subdirectory and update the subdirectory line of the program (line 30) for the subdirectory where the files are stored.

The main analysis file includes a series of variables defined for each year (t), cohort (c), group (g), and occupation (i).

- There are six years: 1960, 1970, 1980, 1990, 2000, and 2010.
- There are eight cohorts: those who were 25-34 in 2010 (cohort 1), those who were 25-34 in 2000 (cohort 2), those who were 25-34 in 1990 (cohort 3), those who were 25-34 in 1980 (cohort 4), those who were 25-34 in 1970 (cohort 5), those who were 25-34 in 1960 (cohort 6), those who were 25-34 in 1950 (cohort 7), and those who were 25-34 in 1940 (cohort 8). In the year 1960 time period, cohort 8 would have been age 45-54.
- There are 5 groups: all individuals in our sample (group 0), white men (group 1), white women (group 2), black men (group 3) and black women (group 4).
- There are 67 occupations (including the home sector - coded from 0 to 66) as defined the Online Appendix that accompanies the paper.

The following variables are in "chad_output_file.cvs". All variables are defined separately for each occupation-group-cohort-year. For ease of exposition, we will refer to the quadruple of occupation-group-cohort-year as a "cell".

- num — The number of individuals in that cell.
- occ_income — The average annual earnings for each full time working individual in that cell (only for those who are currently working, who worked 48 weeks during the prior year, and had at least $1000 of income (in 2010 dollars)).
- occ_ln_income_arith — The log of "occ_income".
- occ_ln_income_geo — The average of log income across each full time worker in that cell.
- occ_grade — The average years of schooling for all individuals in that cell.
- occ_wage — The average hourly wage for each full time worker in that cell.
- occ_ln_wage_arith — The log of "occ_wage".
- occ_ln_wage_geo — The average of log wage across each full time working individual in that cell.
- occ_var_ln_income — The variance of log income across all full time working individuals in that cell.

Part 2: Results

The code has been executed using Matlab 2018b. See the "Online Data Appendix" for the economics behind how the estimation is carried out.

Main Matlab Programs

The basic programs used to produce the results in the paper are listed and described below. All of the programs can be run in the correct order by executing the following master program in the main directory where the matlab files are unzipped:

MasterProgram.m

This program sets up the various cases and runs them, generating *all* the results for the paper. The baseline case is called “Benchmark.” The various programs called by MasterProgram.m are discussed next.

ReadCohortData.m Reads the basic data from the csv version of the spreadsheet, “chad_output_file_2019_01_24.csv.” Also, calls “*LookatCohortData.m*” to display data

SetParameters.m -- Sets the default Benchmark parameter values.

EstimateTauZ.m -- This is the main program that handles the estimation of the tau, phi, A, and Z parameters (among others).

- *solveWMfor_wZ.m*: Estimate w(i,t), phi(i,t), and z(i,WM) using the occupational choices and wages of young white men
- *GetTExperience.m:* Estimates the “Tbar” returns to experience profile from white men
- *estimatedeltaYWM.m*: Code to estimate delta, the fraction of the population that sorts on preference heterogeneity.
- *estimatetauz.m:* Estimates TauW, TauH, and Z for each group

Note: Previous versions of the paper -- and hence these programs -- used the notation Tig(t-c) and Tbar := T(0)+T(1)+T(2). The paper uses an updated notation (as of Version 5). Here is the concordance, to go between the paper and the programs:

- Tig(t-c) == hbar_ig * gamma(t-c)
- Tbar == hbar_ig * gammabar, with gammabar == gamma(0) + gamma(1) + gamma(2), with gamma(0)=1

CleanandShowTauAZ.m -- Displays the results of the estimation

SolveEqmBasic.m -- Takes the TalentData.mat data on Taus, Z's, A(i,t) and solves for the equilibrium w(i,t) and Y(t). Displays the results.

HowMuchPoorer.m -- Evaluates the contribution of changing tau's to economic growth: "Relative to 2010 in model solution, how much poorer would we be if _____ had occurred?"

Ex: if tau's stayed at 1960 level?

if tauw had not changed since 1960

if tauh had not changed since 1960

if only mean tau had changed since 1960, not dispersion

if dispersion changed, but not the means
That is, allow everything else to change as it did, other than _____
"// Share //" row reports the contribution in terms of growth rates.

Note: We changed notation in the paper after completing the programs. In particular, the variable Tbar in the programs corresponds to the product of hbar*(1+gamma2+gamma3). That is, it includes both the level of talent (in hbar) and the returns to experience (in the gammas).

Important Functions

These programs are described in more detail below.

- *ChadMatlab/* -- subdirectory containing many functions that are called by the main programs.
- *AdditionalFigures.m* -- Produces several key graphs for the paper, including the Blau-Kahn labor supply elasticity graph and the Charles-Guryan graph.
- *how_much_poorer.m* -- Details for executing HowMuchPoorer.m
- *HowMuchPoorer_VaryTheta.m* -- Holds tauw and tauh at their benchmark values and resolves the model for different values of theta.
- *ShowParameters.m* -- Display the chosen parameter values
- *Name67Occupations.m* -- Loads occupation names
- *solveeqm.m* -- details for solving the equilibrium w(i)
- *SolveForEqm.m* -- function called to solve for the equilibrium for a given set of tau's.
- *SolveEqmBasic_Display.m* -- display the results for SolveEqmBasic.m