University of Chicago Oeconomica Econometrics Game

April 9th 2016

Welcome to the UChicago Oeconomica Econometrics Game! The objective of this game is to test your knowledge of econometrics and offer you an opportunity to apply your skills in a policy-relevant empirical setting. The questions you are trying to answer in this game are: "How have patterns of immigrant assimilation changed over time? Have immigrants who entered the US recently assimilated at the same rate as those who entered 40 or 50 years ago? What are the causes of these trends?"

1 Immigration to the United States

The United States is historically a country of immigrants. There are 41.3 million foreign-born residents in the US.[1] The most recent government statistics, from 2013, indicate that more than 180 million foreigners entered the U.S. for any reason. Many people choose to stay: in 2013, 779,929 foreign-born civilians became citizens and 990,553 became permanent residents.[2] At the same time, over 15 million US-born children have at least one immigrant.[3]

One sign of successful immigration policy is that foreign-born residents of a country assimilate into the local culture and economy. Multiple studies show that immigrants' wages and language skills converge toward those of US natives over time, and their children look even more like US natives. But other studies show that social norms and beliefs from a person's birth country can persist over time, and even over generations, leading to persistent differences in female labor supply and other economic outcomes. Thus, an important task for social science is to understand the causes behind differential patterns of assimilation.

2 Borjas (2015)

Borjas (2015) provides one of the most comprehensive analyses of recent immigrant assimilation.[4] The analysis consists of two parts. The first part (Sections I-IV) outlines general patterns of wage growth and English language fluency over time. The patterns indicate that recent immigrants have slower assimilation than previous immigrants. However, the analysis is more diagnostic, with various tests eliminating possible causes of the observed slowdown.

The second part (Section V) has a more detailed analysis of the actual determinants of assimilation. Borjas groups immigrants into arrival cohorts: a cohort consists of individuals who arrived in the United States at age i from country k in year t. Borjas tracks the wages each cohort, and compares to the wages of native-born workers of the same age. As a measure of human capital investment, he measures the average English proficiency of each cohort. Assimilation is measured by tracking changes in each cohort's wages between census years. Thus, the data consists of decade-by-decade observations.

[1] Migration Policy Institute, "Frequently Requested Statistics," Table 1. http://www.migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states

[2] U.S. Department of Homeland Security, "2013 Yearbook of Immigration Statistics," Tables 1, 20, and 25. http://www.dhs.gov/sites/default/files/publications/ois_yb_2013_0.pdf

[3] Migration Policy Institute, "Children and Immigrant Families." http://www.migrationpolicy.org/programs/data-hub/charts/children-immigrant-families?width=1000&height=850&iframe=true

[4] http://www.hks.harvard.edu/fs/gborjas/publications/journal/JHC2015.pdf

Borjas finds fairly large differences in the rate of assimilation between different cohorts. Cohorts who arrived in the 1990's had an assimilation rate 9.4% lower than those who arrived in the 1970's (see Table 7).

3 Objectives

The objectives of this project include the following:

1. Understand the econometric methods in a peer-reviewed academic article.
2. Replicate the main qualitative findings of the article.
3. Identify weaknesses in the model and propose refinements, improvements, or fixes.
4. Implement your proposed modifications while defending your underlying assumptions.
5. Estimate your model and compare your results to Borjas', including graphs, tables, and any statistical arguments that are necessary.

3.1 Specifics and Hints

Specifically, Borjas' first argument centers on the estimation of equation 1, with results in Tables 1. At a minimum, you should re-create Table 1. Hint: Be sure to pay attention to the footnotes in his paper and in his tables, so that you construct your cohorts correctly.

After recreating the table, you may want to refine some aspect of Borjas' approach to qualify his results or shed new light on the main questions. Of course, rather than using our recommendations, you may also pursue any of your own ideas. Whatever you do, keep in mind the overall questions that should guide your thinking:

1. How much more slowly/quickly do more recent cohorts assimilate, compared to earlier cohorts?
2. What important qualifications can we make on the aggregate patterns?
3. What might cause these patterns?

You will be judged on the creativity and quality of your work, including your ability to explain and defend your ideas.

4 Data

You will be provided with two (2) ready made data sets: a base dataset of individual observations and a dataset with country characteristics. These datasets are provided in .dta format which are readable directly into STATA version 12+. If you are using R, you can easily read this file type in by using the foreign package. Characteristics of each file are described below. *You of course are free to download any additional data you wish to use in your analysis.*

Note carefully how Borjas constructs his sample. He uses "the sample of immigrant men from the largest 80 sending countries," with three age windows and six year-of-arrival groups (see page 30 for details). In addition: "In each cross-section, the sample consists of men aged 25-64 as of the time of the survey, who have between 1 and 40 years of labor market experience, worked at some point during the survey year, and are not enrolled in school. In addition, the immigrant sample is restricted to persons who migrated to the United States after age 18. The dependent variable is the log weekly earnings of the worker, where weekly earnings are defined by the ratio of total earned income to weeks worked" (p. 5, also note the footnotes therein for further details).

4.1 Base Dataset

The base dataset includes samples of the 1960, 1970, 1980, 1990, and 2000 censuses, as well as samples of the 2009-2011 American Community Survey which serve as a 'proxy' for the 2010 census, since many variables of interest were not included in the 2010 Census. We have collected the data from IPUMPS USA, which may be useful in providing additional detail on variables if need be. The codebook is included among the files in this folder and provides detailed descriptions of variables provided (See IPUMPS USA Data Extract Code Book.pdf). The complete data file as described above contains 55 million observations. Because it would take your computers many hours simply to process a data file of that size, we have drawn a 5% random sample (2.75 million observations) of the file to ensure computational ease due to the limited time available during the game. Therefore your results may be slightly different then what Borjas reports.

4.2 Country Indicators

The country indicators file may be useful in examining the places immigrants in the base sample are from. This file contains education data from Barro and Lee as well as data from the Penn World Tables and the World Bank. It should be easily mergable by year and country ISO code with any World Bank data you wish to download. World Bank variables are defined as follows. Other variable definitions are provided on subsequent pages:

- Gini (World Bank): Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.
- Fertility (World Bank): Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates.

Table 1: Barro-Lee Educational Dataset Variable Definitions

Variable	**Description**
BLcode	Barro-Lee Country Code
WBcode	World Bank Country Code
region_code	Region Code
country	Country Name
year	Year
sex	Sex
agefrom	Starting Age
ageto	Finishing Age
lu	Percentage of No Schooling Attained in Pop.
lp	Percentage of Primary Schooling Attained in Pop.
lpc	Percentage of Complete Primary Schooling Attained in Pop.
ls	Percentage of Secondary Schooling Attained in Pop.
lsc	Percentage of Complete Secondary Schooling Attained in Pop.
lh	Percentage of Tertiary Schooling Attained in Pop.
lhc	Percentage of Complete Tertiary Schooling Attained in Pop.
yr_sch	Average Years of Schooling Attained
yr_sch_pri	Average Years of Primary Schooling Attained
yr_sch_sec	Average Years of Secondary Schooling Attained
yr_sch_ter	Average Years of Tertirary Schooling Attained
pop	Population
pop15	Total Population over 15
pop25	Total Population over 25

Table 2: Penn World Table Variable Definitions

Variable	Description
countrycode	3-letter ISO country code
country	Country name
currency_unit	Currency unit
year	Year
rgdpe	Expenditure-side real GDP at chained PPPs (in mil. 2005US$)
rgdpo	Output-side real GDP at chained PPPs (in mil. 2005US$)
pop	Population (in millions)
emp	Number of persons engaged (in millions)
avh	Average annual hours worked by persons engaged
hc	Index of human capital per person, based on years of schooling (Barro/Lee, 2012) and returns to education (Psacharopoulos, 1994)
cgdpe	Expenditure-side real GDP at current PPPs (in mil. 2005US$)
cgdpo	Output-side real GDP at current PPPs (in mil. 2005US$)
ck	Capital stock at current PPPs (in mil. 2005US$)
ctfp	TFP level at current PPPs (USA=1)
rgdpna	Real GDP at constant 2005 national prices (in mil. 2005US$)
rkna	Capital stock at constant 2005 national prices (in mil. 2005US$)
rtfpna	TFP at constant national prices (2005=1)
labsh	Share of labour compensation in GDP at current national prices
xr	Exchange rate, national currency/USD (market+estimated)
pl_gdpe	Price level of CGDPe (PPP/XR), price level of USA GDPo in 2005=1
pl_gdpo	Price level of CGDPo (PPP/XR), price level of USA GDPo in 2005=1
i_cig	0/1/2: relative price data for consumption, investment and government is extrapolated (0), benchmark (1) or interpolated (2)
i_xm	0/1/2: relative price data for exports and imports is extrapolated (0), benchmark (1) or interpolated (2)
i_xr	0/1: the exchange rate is market-based (0) or estimated (1)
i_outlier	0/1: the observation on pl_gdpe or pl_gdpo is not an outlier (0) or an outlier (1)
cor_exp	Correlation between expenditure shares of the country and the US (benchmark observations only)
statcap	Statistical capacity indicator (source: World Bank, developing countries only)
csh_c	Share of household consumption at current PPPs
csh_i	Share of gross capital formation at current PPPs
csh_g	Share of government consumption at current PPPs
csh_x	Share of merchandise exports at current PPPs
csh_m	Share of merchandise imports at current PPPs
csh_r	Share of residual trade and GDP statistical discrepancy at current PPPs
pl_c	Price level of household consumption, price level of USA GDPo in 2005=1
pl_i	Price level of capital formation, price level of USA GDPo in 2005=1
pl_g	Price level of government consumption, price level of USA GDPo in 2005=1
pl_x	Price level of exports, price level of USA GDPo in 2005=1
pl_m	Price level of imports, price level of USA GDPo in 2005=1
pl_k	Price level of the capital stock, price level of USA in 2005=1