This README documents how to reproduce the results in Abigail Wozniak’s and Ofer Malamud’s “The Impact of college Education on Geographic Mobility,” published in the Journal of Human Resources (2012, 47(4): 913-950).

Repository Structure

*The repository is structured as follows:*

* **\Data\_Setup**
  + **\1980Census\_Data\_Setup** – contains do-file and datasets that construct the final 1980 Census dataset.
  + **\Inductions\_Data\_Setup** – contains do-file and datasets that construct the final inductions dataset (inductions.dta), as well as datasets that contain variations of measures of induction risk (see do-file for risk measures).
* **\Data**
  + **\Cleaned\_Data –** contains folders that hold the cleaned 1980 Census file and cleaned state-level inductions and enrollment data files.
  + **\Raw\_Data** – contains folders that hold the raw data for the 1980 Census (downloaded from IPUMS), cohort size, school enrollment, and inductions.
* **\Enr\_PDF\_YYYY\_YY**
  + original PDF versions of school enrollment data from 1960 to 1965.

CPS Data Setup

*The main CPS dataset and do-file:*

* **census\_1980.dta** – raw 1980 Census 5% Sample from IPUMS. The sample is restricted to men aged 25 to 40.
* **census1980\_clean.do** – cleans census\_1980.dta and merges cohort size, employment-population, and induction data to create the final 1980 Census dataset. The following datasets are also called:
  + **blanchard\_katz.dta** – state-level unemployment insurance rates (epop and epopstate) sorted by bplg and year19
  + **cohortsize\_c60.dta** – 1960 cohort size at the birthyear and statebirthyear level (lnsize and lnstatesize) sorted by bplg and birthyear.
  + **inductions\_short6m** – a short version of inductions.dta (containing only birthyear, bplg, staterisk, nationalrisk, nationalrisk\_, and risk)

Induction Risk Data Setup

*The main Stata datasets and the do-file used to create our measures of induction risk:*

* **inductionrisk\_setup3.do** – generates national-level and state-level induction risk by birth year and by six-month birth interval. Creates datasets with various definitions of induction risk.
  + This do-file uses the following datasets:
    - **national\_inductions.dta** – national level inductions and cohort size by birth year for birth years 1935 to 1959. Note that inductions for 1955 onwards equal zero. Created from inductions.xls.
    - **state\_inductions.dta** – state level inductions and enrollments by calendar year. Created from state\_level\_inductions\_enrollments.xls (see below).
    - **state\_inductions\_aw.dta** – state level inductions and enrollments by six months of calendar year. Created from state\_level\_inductions\_enrollments\_aw.xlsx.
  + This do-file creates the following datasets:
    - **inductions.dta** – final dataset created by inductionrisk\_setup3.do.
    - **inductions\_short.dta** – a short version of inductions.dta containing only birthyear, bplg, staterisk, nationalrisk, nationalrisk\_, and risk. Note that *risk* is equivalent to the measure in C&L 2000. Our *nationalrisk* is created by aggregating up state level induction totals and enrollments and is not exactly equal to *risk* but is very close. See inductionrisk\_setup3.do for constructions.
    - **inductions\_short6m.dta, inductions\_short6mt.dta, inductions\_short6ma.dta** – versions with different definitions of induction risk.

*The full set of raw data that we put together to construct the datasets on induction risk:*

* **inductions.xls** – national-level induction rates and cohort sizes from C&L (data from David Card) and SSS website.
* **state\_level\_enrollments.xls** – state-level 10th and 11th grade enrollments from 1961-1969, and secondary totals for 1959-1960.
* **state\_level\_inductions.xls** – state-level inductions from 1954-1972, raw and corrected. State level inductions processed and received in paper form from SSS in two batches. First batch reports annual induction numbers by state and year or six-month period. Second batch reports cumulative induction totals by state and year. Corrections made for obvious data entry mistakes. Edited observations are highlighted in “corrected” tab. “Inductions\_final” tab contains calculated annual total inductions by state for 1955 to 1972.
* **state\_level\_inductions\_enrollments.xls** – final tabulations of state-level induction and enrollment counts. Combines data from state\_level\_inductions “Inductions\_final” tab and state\_level\_enrollments. “Combined” tab contains state level annual induction and enrollment numbers where indYYYY is inductions from “Inductions\_final” tab and enrYYYY is the following: Secondary enrollment/4 if YYYY=1959 or 1960 from state\_level\_enrollments, and 11th grade enrollment for YYYY=1961 to 1969. Where does the enr1970 data come from?? This is not in state\_level\_enrollments.xls. Where is enr1966-1970 raw data?? “Combined and Adjusted” tab shifts enrYYYY to enr(year 19=YYYY+2) for matching and division by year a cohort turned 19. “Combined and Adjusted” tab data form state\_inductions.dta.

NOTE: There are also “AW” versions of the above, which carry the six-month level variation through instead of collapsing it.

Census Data Set Up and Regression Info

Here is the main setup and regression do-files:

viet\_setup8.do – the latest setup do-file

Here are the main regression do-files:

viet\_reg13\_coll.do – the latest individual-level regression do-file. Creates first stage, reduced form, OLS, and 2SLS results for various education measures. Data are collapsed to birthyear6m and bplg cells (1200). When using both vet and schooling in an equation, SEs are two-way clustered.

These do-files are not in the folder.

*Here are the ancillary data files used in the main setup do-file:*

cycle1\_ma.dta – national unemployment rate (ur12ma) sorted by year19