**Visualizing Income Inequality in the United States: Data Processing Step by Step**

Sang Truong and Dr. Humberto Barreto

Department of Economics and Management

DePauw University, Greencastle, IN, 46135

**Step 1:** Download raw data. Go to IPUMS CPS (<https://cps.ipums.org/cps/>), log in and download data from 1978 to 2018 with these following variables: YEAR, HFLAG, ASECWTH, STATEFIP, HHINCOME, PERNUM, ASECWT, AGE. The raw data file should have 7197133 line. Variable explanation:

* ASECWT: some individuals have higher chance to be selected by IPUMS than others due to complex sampling process. ASECWT is used to account for different sampling probability. ASECWTH is similar to ASECWT, but in the household level.
* HHINCOME: household income.
* PERNUM: the number assigned to individual to code their position in the household. PERNUM = 1 indicates that individual is the head of household. If the household has several members, the head of the household will have PERNUM = 1. Other members of the household will have PERNUM = 2, 3, 4, etc.
* HFLAG: in 2014, ASEC redesigned their survey. All respondents received new health insurance question. However, only 3/8 of the other total sample was randomly selected to receive the redesigned income question. Therefore, HFLAG only has value in 2014. 3/8 sample has HFLAG = 1, and 5/8 sample has HFLAG = 0.
* YEAR and STATEFIP: the year and state from which the data was collected.
* AGE: Age of individual

**Step 2:** Drop all individuals who have HFLAG = 1 to keep 2014 data comparable with data from other years. After that, drop HFLAG.

**Step 3:** Generate deflator.

* Deflator grid has 2 dimensions:
  + 1st dimension: 51 states and DC
  + 2nd dimension: 41 years
* Go to IPUMS USA (<https://usa.ipums.org/>) and download data for 1970, 1980, 1990, and 2000-2017 for these following variables: YEAR, HHWT, STATEFIP, RENTGR. RENTGRS reports the gross monthly rental cost of the housing unit.
* Iterate through the following layers to fill out deflator grid. For each year:
  + For each state: state deflator = median (RENTGRS).
* Generate a new variable called AVERAGE, which captures the average of median of RENTGRS of each year.
* Normalize deflator by AVERAGE and substitute data in deflator grid: deflator1 = deflator/AVERAGE.
* Calculate COLI (Cost of Living Index) and substitute data in deflator1 grid: COLI = 0.44\*Deflator1 + 0.56.

Explanation: Housing cost is the key variable in variation in state cost of living. We assume prices of all other goods the same across states. Housing cost (including utilities) is roughly 44% of budget. “The estimate of 44 percent came from the Consumer Expenditure survey tabulations of expenditures for two‐adult/two‐child families. For families at the 35th percentile of the distribution of spending on food, housing and clothing, housing represented 44 percent of total expenditures assuming miscellaneous expenditures are set at 15 percent of the food, housing and clothing amount.” Renwick, 2001, p. 4.

* Generate regression grid:
  + 1st dimension: 51 states and DC.
  + 2nd dimension: Intercept, Coefficient, Squared.
* For each state, run a linear regression to fit a line through COLI data. Regression outcome is report in regression grid.
* Generate prediction grid:
  + 1st dimension: 41 years.
  + 2nd dimension: year, STATEFIP, PRE\_COLI.
* If COLI data is available, use COLI instead of PRE\_COLI. The final prediction grid is the deflator using the next steps.

**Step 4:** Create result grid, which has 2 dimensions:

* 1st dimension: 51 states and DC.
* 2nd dimension: 11 deciles or 98 percentiles.

**Step 5**: Iterate through the following layers:

* In the raw data, there is 41 years. For each year:
  + Generate 2 new variables: SIZE and EFFSIZE.
  + SIZE indicates the size that each individual contributes to the household size. If PERNUM = 1, individual has SIZE = 1. Otherwise, if individual is older than 16 years old (meaning AGE > 16), SIZE = 0.7. Otherwise, SIZE = 0.5.
  + Effective size of the household is the sum of the size contributed by every individual in the household. For example, a household that has 1 head (SIZE = 1), 1 adult who is 30 years old (SIZE = 0.7) , and 2 children who is 10 years old each (SIZE = 0.5 each) will have effective size EFFSIZE = 1 + 0.7 + 0.5\*2 = 2.7. Each household only has 1 effective household size, and this information is saved with the household head (the individual with PERNUM = 1).
  + After obtaining EFFSIZE, drop observation with PERNUM ≠ 1.
  + Normalize HHINCOME with EFFSIZE:

HHINCOME1 = HHINCOME/ EFFSIZE

* + Normalize HHINCOME1 with deflator:

HHINCOME2 = (new) HHINCOME1/ Deflator

Each state in the same year has different deflator to account for different living expensive. For the same state, each year has different deflator to account for inflation. Therefore, for 51 states and DC in 41 year, there are 51\*41 = 2091 deflators. Detail about generating these deflators will be explained later.

* + Sort the data by STATEFIP.
  + For each year, there is 51 states (reminder: we are in “year” level of data). For each state:
    - Sort data by HHINCOME2
    - Compute 2 new variables: CUMWT and PERCENT.
    - CUMWT is the cumulated weight (weight is represented by ASECWT). For the first household in the state (which is the poorest household in the state, because the data frame is now sorted by HHINCOME\_2) has cumulated weight equals to ASECWT: CUMWT1 = ASECWT1.The second poorest household in the state has CUMWT2 = CUMWT1 + ASECWT2. In general, the ith poorest household in the state has CUMWTi = CUMWTi-1 + ASECWTi.
    - PERCENT is the percentile ranking of each household calculated by dividing cumulated weight by total weight of each state:

PERCENTi = CUMWTi/sum (ASECTWT)

* + - Collect the household income at each decile (or percentile) to fill into the result grid created in step 3. The household income of bottom i percentile, is the household income of the first household that has PERCENT larger than i %.
* **Step 6:** Generate normalized population for each state. Normalized population of each state indicates the width of each column on the final graph.
  + In the result grid, generate a new variable called POP (population) in the 2nd dimension of result grid. For each state, POP = sum (ASECWT).
  + Generate a new variable called NORMPOP (normalized population in 2nd dimension of result grid. For each state: NORMPOP = POP / min (POP).
* **Step 7:** Convert the result grid into JSON format and combine it with the AmChart template to generate the final graph for each year.