IS 452 Foundations Information Processing Fall 2017

Final Project

Nanzhu Liu

UIN: 655394878

**Project Type**

Programming project

**Project Summary**

The project is to estimate demographic data related to children age 5 and under. Two types of demographic data are included: Population of children age 5 and under, and Population of children under age 5 by race. The data is updated annually, and the estimation has to be done every year. I used to estimate the data by Excel and Access. It took me a lot of time. The structure is similar every year, so the process is repetitive. I decide to write a python file to estimate data.

**Dataset**

Two primary sources that I use are American Community Survey (ACS) and the Population Estimates Program(PEP). There are total 6 data files.

I used 2015 American Community Survey (ACS) 5-year Estimates, which is from <https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>

|  |  |
| --- | --- |
| Table name | What is it? |
| ACS\_15\_5YR\_B01001\_with\_ann.csv | Population |
| ACS\_15\_5YR\_B01001A\_with\_ann.csv | Race: White Population |
| ACS\_15\_5YR\_B01001B\_with\_ann.csv | Race: Black Population |
| ACS\_15\_5YR\_B01001C\_with\_ann.csv | Race: American Indian and Alaska Native Population |
| ACS\_15\_5YR\_B01001D\_with\_ann.csv | Race: Asian Population |

I used 2015 Population Estimates at National Vital Statistics System (NVSS) mentioned above, which is from <https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm>

Table name: pcen\_v2015\_y15\_txt.txt

**Methodology**

The goal is to have number of children under age 1, age 1, age 2, age 3, age 4, and age5 by counties in Illinois and children in race: white, black, American Indian and Alaska Native and Asian by counties in Illinois. ACS contains varieties of data sets and has large data sets. They have total number of children age 5 and under, and population of children by race, but they don’t have the number of children under age 1, age 1, age 2, age 3, age4 and age5 separately. PEP contains less kinds of data set but has number of children across all ages. To make the data consistent, I use the data from ACS and make estimation based on PEP. I download population and population by race from ACS. I also collect data by counties from PEP because this is more precise. The estimation is shown below:

Pop(acs,t) = Pop(pep, t)/Pop(pep) \*Pop(acs) (1)

In equation, Pop(acs,t) is the population from ACS at time t such as age1. Pop(PEP, t) is the population from PEP at time t such as age1. Pop(pep) is the total population children age 5 and under from PEP. Pop(acs) is the total population children age 5 and under from ACS.

The structure of Python code is divided into three parts. In the first part, I read data sets from PEP into Python and create a matrix that each number is calculated by Pop(pep, t)/Pop(pep). In the second part, I read population data sets from ACS, and estimate children population across birth to age5. The third part, I read all four population data by race, and I create a function to estimate children population by race across birth to age5. I will give an detailed explanation on how I write the python file.

**Narratives**

Set up

I import pandas as pd and csv

Part I

I collect and prepare population data by counties from PEP. The file is .txt file. The whole txt file contains population data across country from age 0 to 85. The txt file contains two columns. The first column tells the data information including series vintage, years, month, states FIPS code, counties FIPS code, ages and so on. The second column tells numbers. I need to select data that satisfy the following requirements: data is Illinois and age is from 0 to 5.

First, I use readlines() to read .txt. Second, I select IL data using split() and if condition. The delimiter of split() is tab. I use slicing to keep states FIPS code, counties FIPS code, ages from the first columns and deleted other information. I select age from 0 to 9. There is a space before age if age is from 0 to 9. I used split() to separate the first column into two, and used if condition to select data with 2 lengths. The delimiter of split() is a space. If the age equals to 10 or above, there is no space, so the length is 1. Now, the first column represents “geoid”. The second column represents age. I use join function to joined geoid, age with numbers. Next, I separate FIPS code, age and numbers from one long string into three strings by using .split(). I create an empty row named “row” and use for loop to append rows into the row list. I create an empty list “age0to5” to store age which equals or less than 5. Since the age in the row are strings. I use int() to covert string to num. I use for loop and if condition to filter out the row that meets the requirement. To save the result into “age0to5”, I use append() function. I stored the filtered data into a new csv file named as “pep.csv” in case I needed in the future.

I used pandas to read the csv file I just created. I wanted to group by the file by geoid and age.

Part II:

I collected and prepared population data by counties from US Census ACS. It is csv file. I imported csv file in Python by importing pandas. I selected the 4 columns: geo.id, name, male under 5, and female under 5. I aggregated male under 5 and female under 5 to get the total number of children under 5. I combined geo.id, name and the total number of children under 5 together by using concat() function. I renamed the header and exported into a new csv file.

The race data are also from US Census ACS. There are 4 csv files: race by white, race by black, race by American Indian and Alaska native, and race by Asian. The procedure is the same as the previous one. I renamed the header and exported them into new csv files based on the type of race.

Thus, there are total 5 files in the result folder.

Work will be done later on:

The procedure is almost the same. I am wondering if I could create a for loop to optimized my code.