Stat 605 - Final Project
Project Proposal

Project Group 4

November 2020

Authors

Ji, Qingchuan, qji5
Kou, Huitong, hkou2
Rui, Peibin, prui
Shi, Daiyi, dshi42
Su, Haohao, hsu69

Data Description

This project will be based on personal record data of the 5 one-year ACS PUMS datasets of the whole US, from 2015 to 2019. The Public Use Microdata Sample (PUMA) data contains a data sample of actual response to the American Community Survey (ACS) made by U.S. Census Bureau, covering approximately 1% of the population in the United States. Every one-year dataset can be divided into two parts of records, household-level and person-level (see 2019 ACS 1-year PUMS File, page2). Here we only focus on the personal record subset.

Data Source

https://www2.census.gov/programs-surveys/acs/data/pums/

Here, the following code can be used to grab the 1-year PUMS person-level dataset of 2019.

```
wget https://www2.census.gov/programs-surveys\
    /acs/data/pums/2019/1-Year/csv_pus.zip
unzip csv_pus.zip
```

Goals

0. Imputation As the data has become "dirty" due to allocation and missing data, multivariate imputation of high-dimensional data is an essential prerequisite for further statistical analysis.

1. Personal Income The trend how the average personal income changes for the last five years; The influential factors of personal income (family background, education, race, employment, ...); ..... The analysis may be made based on the data of each state and the whole America.

Variable Description

Totally, there are millions of observations in the person-level PUMS dataset of each year. And there are 286 meaningful variables originating from the questionnaire on the survey as well as variables deriving from some other survey responses, including 17 numeric variables, 76 allocation flag variables as well as 192 categorical variables and codes.

In this project, **PINCP** is essential, which is a numeric variable and represents "Total person's income".

As for the details of the other variables, please check the 2019 ACS PUMS Data Dictionary.

Statistical Methods

Multivariate imputation by chained equations; chi-squared test; logistic regression; classification and regression trees.

Computational Tools

CHTC or lunchbox server : maybe useful when running code to do with imputation and train multivariate regression models.

R (with packages MICE, rpart, ctree, ...): MICE for multivariate imputation of missing or allocated data by chained equations; rpart and ctree for train classification and regression trees.

GUIDE Classification and Regression Trees and Forests: an interesting and useful tool running under Linux for classification and regression trees and forests.