PH244
Big Data: A Public Health Perspective
Computing Project

The file `ss13hus.csv.bz2` under `bCourses/Files/Project/Project-II/` contains household-specific data from the 2009-2013 US Census American Community Survey. This survey obtains a wealth of information on people and households every year, with about 1% of the total population surveyed in each year. The dictionary describing all the data fields is available as `PUMS-Data-Dictionary-2009-2013.pdf` under the same directory. The zipped file is about 600MB, and be careful about unzipping it. **You are required to use `R` for this computing project, and use `R Markdown` that include both the computer code and the output for the project report. There is no page limit on this report.**

1. Use `read.csv()` to read the zipped data `ss13hus.csv.bz2` into `R`. Considering the data size, use the "divide-and-conquer" strategy to read the data. To do so, first count the number of rows of the data using the function `countLines()`, then read a chunk of data each time, until reaching the last row of the data. Use `system.time()` to record and report the time it requires to read the data. The following code might be helpful:

   ```
   library(R.utils)

   nrow=countLines("ss13hus.csv.bz2")[1]

   filename <- "ss13hus.csv.bz2"

   con <- bzfile(filename, "rt")

   # choose a suitable value for chunk_size based on nrow

   DT1<-read.csv(con, header = T, nrows = chunk_size)

   ...
   ```

2. Create a subset of data by *randomly* sampling $3,000,000$ survey records from `ss13hus.csv.bz2`. Extract the following data fields: `REGION, ST, ADJHSG, ADJINC, NP, ACR, BDSP, ELEP, GASP, RMSP, VEH, WATP, FINCP, HINCP`. Save the file as a `csv` for subsequent analyses, with rows representing survey records and columns different data fields. In addition, for reproducibility, please use `set.seed(1000)` to set the random seed.

3. Try 3 different functions of reading the data you create in Step 2 into `R`: `read.csv()`, `fread()`, and `read.csv.ffdf()`. Use `system.time()` to record and report the time each function requires to read in the data.

4. Draw a scatter plot of BDSP (the number of bedrooms; a measure of house size) on the x-axis, and FINCP (the family income; use ADJINC to adjust FINCP to constant dollars) on the y-axis. Add a gam smoother, with standard error shading, on the scatter plot using the R package ggplot2.

5. Create a subset of data by *randomly* sample $1,000,000$ survey records from ss13hus.csv.bz2 as a subset. Fit a linear regression model on this subset with the adjusted family income (FINCP) as the response, and BDSP and VEH (the number of vehicles) as the predictors. Record the estimated coefficient for BDSP. Repeat the same procedures for 1000 times with different random seeds. Use the R function lm() and biglm() to implement it. Use system.time() to report the time it requires in total. Report the mean and standard deviation of the estimated coefficients for BDSP recorded in these 1000 repetitions. Also draw a density plot of the estimated coefficients for BDSP recorded in these 1000 repetitions.