PH244
Big Data: A Public Health Perspective
Computing Project

Problem:

The file `ss13hus.csv.bz2` under `bCourses/Files/Data/` 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 need to include your computer code and output in the report. You are *required* to use `Rmd` to write the report, which can easily include the `R` code. There is *no page limit* on this report.

1. Try 3 different commands of reading the zipped data `ss13hus.csv.bz2` into `R`: `read.csv()`, `scan()`, and `readLines()`. Use `system.time()` to record and report the time each function requires to read in the data.

2. Create a subset of data by *randomly* sampling $1,000,000$ 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 analysis, with rows representing survey records and columns different data fields. Hint: This is *not* a trivial task, considering the data size, and it involves some amount of programming. You may use a "divide-and-conquer" strategy. In addition, for reproducibility, please use `set.seed(1000)` to set the random seed.

3. Try 3 different commands of reading the data you create in Step 2 into `R`: `read.csv()`, `data.table()`, and `ff()`. 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 loess smoother, with standard error shading, on the scatter plot using the `R` package `ggplot2`.

5. Fit a linear regression model with the adjusted family income as the response, and `BDSP` and `VEH` (the number of vehicles) as the predictors, using the `R` package `biglm`. Report the summary of the regression fitting.